SEMANTIC MAPPING IN VIDEO RETRIEVAL

SEMANTIC MAPPING IN VIDEO RETRIEVAL

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 18 december 2017
om 12.30 uur precies

door

Maaike Heintje Trijntje de Boer

geboren op 7 november 1990
te Utrecht

Promotor:
Prof. dr. ir. W. Kraaij

Copromotor:
Dr. K. Schutte — TNO

Manuscriptcommissie:
Prof. dr. ir. A.P. de Vries — voorzitter
Prof. dr. M.A. Larson
Prof. dr. M.-F. Moens — KU Leuven, België
Prof. dr. A.F. Smeaton — Dublin City University, Verenigd Koninkrijk
Dr. C.G.M. Snoek — Universiteit van Amsterdam

SIKS Dissertation Series No. 2017-43
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems, the institute for Computing and Information Science (iCIS) of the Radboud University Nijmegen, and TNO.

Cover Design: Stefanie van den Herik ©
Printed by: ProefschriftMaken || www.proefschriftmaken.nl

ISBN 978-94-6295-759-6

An electronic version of this dissertation is available at
`http://repository.ru.nl/`.

CONTENTS

1

INTRODUCTION

1.1. USE CASES AND THESIS STATEMENT

In the modern world, networked sensor technology makes it possible to capture the world around us in real-time. For example, in the security domain cameras are an important source of information. The police in the USA is already using body cams (White, 2014; Kelsh, 2016) and the police in the Netherlands is starting to use them, surveillance cameras are present in many public places (Tokmetzis, 2013) and aerial vehicles or drones record aerial images (Reese, 2015). In case of a special event, such as a robbery or some type of violence, citizens are often recording videos with their smartphones. All these types of information can be used 1) for real time monitoring of the environment to prevent crime (*monitoring case*); and/or 2) for investigation and retrieval of crimes, for example in evidence forensics (Chamasemani et al., 2015) (*forensic case*).

In the monitoring case, security operators have to monitor many video streams to detect suspicious behavior. Some application areas of monitoring are: (public) transport (airport, railway), public places (bank, supermarket, parking lot), public events (concerts, football matches) and military (near national borders, secure perimeters) (Lee et al., 2000; Valera et al., 2005). According to Ainsworth (2002), a security operator can lose up to 95% of the activity on the screen after just 22 minutes of continuous surveillance. With the increasing amount of video data, it becomes unfeasible for a security employee to track all streams in real time to monitor the environment. Current systems are already able to assist the security operator (Vishwakarma et al., 2013; Ko, 2008). Start of the art technologies include automatically providing an estimation of the density of a crowd (Davies et al., 2005; Zhan et al., 2008), object detection and recognition (Hu et al., 2004; Uijlings et al., 2013), motion analysis / tracking (Zhang et al., 2015b; Yilmaz et al., 2006; Hu et al., 2004), person identification (over multiple video streams)(Bouma et al., 2013a; Vezzani et al., 2013) and human behavior analysis (Ko, 2008; Aggarwal et al., 2011; Vishwakarma et al., 2013), such as pickpocketing (Bouma et al., 2014; Arroyo et al., 2015), stealing from a truck (Burghouts et al., 2014) or digging up an Improvised Explosive Device (IED) (Schutte et al., 2016). In

the monitoring case, it is important that the systems 1) work in (near)-real time, i.e. can process the video streams as fast as they occur; 2) have a low false alarm rate (Bouma et al., 2014), because millions of alarms per minute are not feasible to go after; 3) have a low probability that an event is missed (high recall), because important events should not be missed. These latter two performance measures have a trade-off.

In the forensic case, the security operator, law enforcement employee or homeland security employee currently has to rewind the videos to find evidence, read the logbook transcriptions of the radio and read through a lot of paper work. Examples of evidence are the location of the suspect at a certain point of time, an illegal event, child abuse or signs of radicalisation (Mould et al., 2014). Different from an alert in the monitoring case, the forensic case often works with user queries. The security operator can either use an example image (*query-by-visual-example*) or a textual description (*query-by-keyword*) as query to explain what he/she is searching for (Snoek et al., 2008). Instead of real time video streams, the system searches through a database with videos or video segments. To allow fast look-up in the database, the videos are often indexed with the date and location information, and pre-trained concepts, such as the people, objects, or pre-trained suspicious behavior (Schutte et al., 2013). The previously mentioned techniques for monitoring can, thus, also be used in the forensic case. The alerts can be seen as pre-defined queries that produce an alert when detected. On the other hand, the forensic case can use (potentially better) detection methods that are slower than real-time and the case allows for multiple interactions with the system to gather the information, i.e. interactive search. In the presentation of the results, the monitoring and forensic case also differ. In the monitoring case few alerts should be given, whereas in the forensic case the amount of results can be bigger, dependent on the amount of time the operator has available. A security operator can scroll through the results until satisfaction. This difference also results in a different evaluation of performance. In the monitoring case the main focus is on the perceived positives (precision), i.e. few good results, whereas the forensic case mainly focuses on the actual positives (recall).

In both the monitoring case and the forensic case, the alerts and detections rely on an algorithm that is trained to detect the suspicious behavior or event. Without this training, or a visual example, current systems will not be able to find the event. It is, however, not possible to train all potential events of interest, because unexpected or unseen events will always occur. In this thesis, we use the security case as an inspiration for our research regarding unseen events. We only focus on the technical scientific challenges regarding the security case, and not on the juridical, ethical or privacy challenges. Obviously, the application of intelligent analysis of video footage in forensic or monitoring situations is a sensitive topic in the public debate as it touches the delicate balance between privacy and security. As a civilian, I do not like the idea of cameras tracking me on every corner of the street. On the other hand, with all the recent terrorist attacks I want to feel safe. This causes a tension between privacy and security in which we expect that the government takes care of our security without invading our privacy. With the upcoming technologies, this tension between privacy and security is reinforced. As an example of this tension in the Netherlands, we take

the example of the commotion about the company Exterion Media last September (2017). This company used cameras in their billboards on the train stations to count the number of people passing by and how much time people spent looking at the billboard. Although this company did not collect privacy sensitive information, they still have removed the cameras. In national and international law it is regulated that people or companies cannot just record and process everything. For example in the Netherlands there is a law (article 151c Gemeentewet) that states that camera footage of public space can only be kept for four weeks, unless illegal acts are recorded and used by the police for detection and prosecution of suspects. There is also a law 'Wet bescherming persoonsgegevens' that takes care of the usage of personal information. Within the EU the General Data Protection Regulation (GDPR) determines what is allowed and what is not allowed. Within the boundaries of the law content-based video search is a legal instrument. In this thesis, we do not search for specific instances of a specific object (instance search) or work on (re-)identification of people. We do, however, use the search for unseen events in the security case as the societal inspiration for our research.

In order to allow a user to find a previously unseen (type of) event, the system should be able to handle ad-hoc queries, i.e. queries that include concepts or events that are not pre-trained. In this thesis, we focus on textual queries, such as *Where can I find the pink Cadillac in Amsterdam?*, *Look out for a person with a red jacket, grey pants and black backpack* or *Who left the bag at camera 2 and where is the person now?*. These type of queries are applicable to both the forensic case and the monitoring case, but also to general search systems or other application domains.

Thesis Statement: *We aim to assist a user in their work on video stream data by providing a search capability that handles ad-hoc textual queries.*

Our methods to obtain such a search capability are inspired by the search engines on the World Wide Web, i.e. the Internet. Research in this field is named 'MultiMedia Information Retrieval' (*MMIR*). Currently, the most popular online search engines, or video databases with a search function, are YouTube (Burgess et al., 2013) and Google Videos (Sivic et al., 2006). Although we can use these search engines as inspiration, important differences between the videos uploaded to the Internet and the videos in the security domain are present. First, the uploaded videos often contain metadata in the form of a title and textual description of the content of the video. This information is highly valuable, because it allows for the same search mechanisms that are used in the retrieval of text documents on the Internet, i.e. document retrieval. In the security domain, the video streams typically have no textual information about the content besides the date and location of the stream. Without textual information, we have to focus on 'Content-Based' (multimedia) Information Retrieval (*CBIR*) (Kato, 1992). Second, the uploaded videos are often specifically produced or edited. The videos in the surveillance domain are typically not edited. Third, the characteristics of the use case within the security domain and the 'YouTube' domain are different. For the majority of YouTube-style videos, the primary use case is entertainment. The primary access function is through social channels or recommender systems. Some videos 'go viral' over the internet and other videos are never retrieved. Within the se-

curity domain, the focus is on alerting or forensic search. This means that search effectiveness (especially recall) and relevance have a completely different meaning, i.e. they aim at another point on the ROC curve. The security domain, therefore, can be inspired by the search capabilities obtained from document retrieval, but the main differences have to be taken into account in the creation of the search capability.

In the next section, we provide a short high level overview of the state of the art regarding the current search capabilities for ad-hoc queries. Section 1.2 introduces the research questions, whereas Section 1.3 discusses the scientific methodology. Section 1.4 contains the main contributions of this thesis, and Section 1.5 consists of an overview of the structure the thesis.

1.2. OVERVIEW OF CONTENT BASED VIDEO INDEXING AND RETRIEVAL

In our goal to implement a search capability that handles ad-hoc queries, we envision several common elements in CBIR that have to be in place. Figure 1.1 provides an overview of the visual search system components that are important for ad-hoc queries. This figure is only intended as an example of how such a system would work, because in the literature many different systems are proposed. The important components are Indexing, Query Interpretation, Feedback Interpretation and Scoring & Ranking (indicated by the blocks). Each of the components can be interpreted as a function that converts an input to an output. The system on the one hand obtains video data from one or multiple sensors, such as a camera, drone or body cam. This video data is processed and indexed in the component *Indexing*. The pre-trained detectors (D) are applied to the video (v) and the result is an indexed video ($\vec{i}_i$), which is a vector with for each detector a score indicating the whether the concept is present in the video or not. This indexed video is stored in the video database (V). In the monitoring case, the video database might also be filled, but the score is then directly forwarded to the Scoring & Ranking component to check whether the video triggers an alert.

The user can search the video database for relevant information. The user has to provide a query to the system using an interface (q_u). This query is interpreted in the component *Query Interpretation*. This component uses information from a Concept Bank that contains the labels of the pre-trained detectors (L), i.e. the words that are used to index the videos with. The Query Interpretation component outputs a system query ($\vec{q_s}$), which is, in this case, a (sparse) vector with a weight for each of the concepts in the Concept Bank. In a query that is not ad-hoc (but pre-trained) this component will only output a weight of one for the pre-trained concept. This can be either implicit through a neural network or explicit, that is in the query. The system representation of the query is provided to the component *Scoring & Ranking*. This component uses a similarity measure to combine the system query with each of the indexed videos from the video database. Depending on the case, the component outputs a ranked list of results or an alert. The user can provide feedback on these videos (for example relevant / not relevant / not sure), or on the set of concepts (for example add / delete / adjust weight) (f_u). This information is used, together with

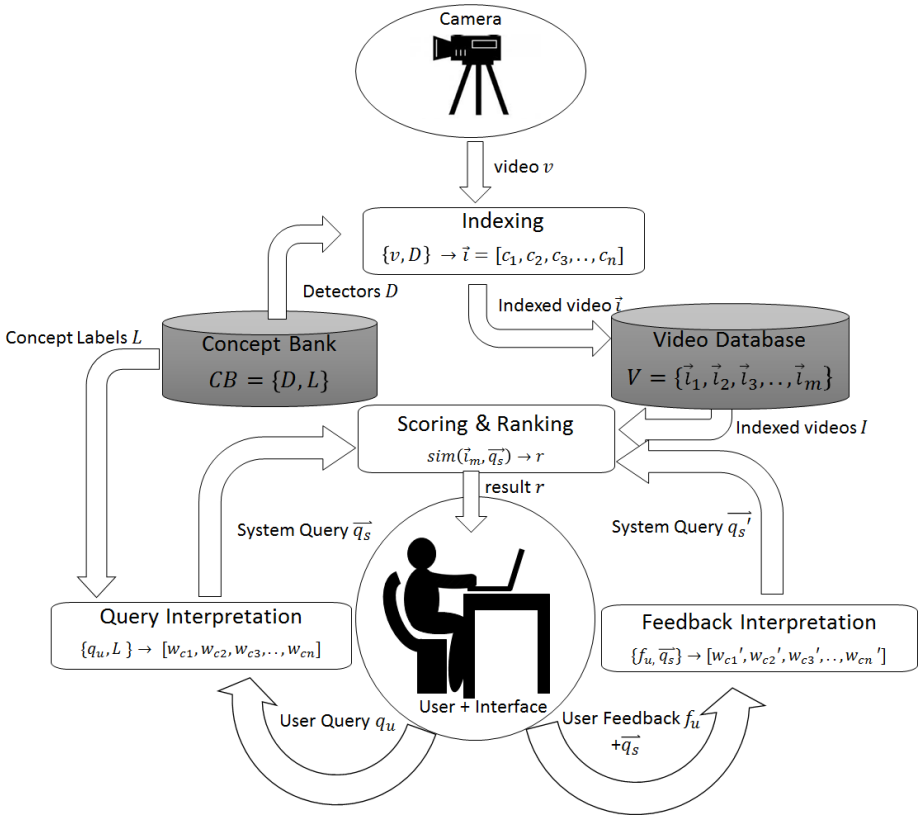


Figure 1.1: Overview of Visual Search System Components: CB is the Concept Bank that contains a set of n detectors D and their respective labels L. V is the video database that contains the set of indexed videos $\overrightarrow{I}$. Video v is indexed in $Indexing$, using D, in an indexed video $\vec{i}$. This index (descriptor) is a vector of concept detector scores c. The user can pose a query q_u, which is interpreted together with L in $QueryInterpretation$ into a system query $\overrightarrow{q_s}$. In $Scoring\&Ranking$ this $\overrightarrow{q_s}$ is used together with $\overrightarrow{I}$ to produce the result r, which is a ranking score. The user can provide feedback f_u, which results through $FeedbackInterpretation$ in a new system query q_s' and new results.

the original system query, in the component *Feedback Interpretation*, in which the system query is updated to create a better result.

Each of the previously described components embodies a full research field. We will not explain all state of the art in each of the research fields, but rather provide just enough information to understand the choices and challenges present in this thesis.

1.2.1. INDEXING

An overview of typical components in Indexing is shown in Figure 1.2. Some key references related to this Indexing can be found in Smeulders et al. (2000), Snoek et al. (2008), Lew et al. (2006), Liu et al. (2007), Hu et al. (2011), and Zhang et al. (2015c). In general, videos are split into shots, which are the “consecutive sequence of frames captured by a camera action that takes place between start and stop operations" (Hu et al., 2011). Shot boundaries are detected by a shift or change of the camera by us-

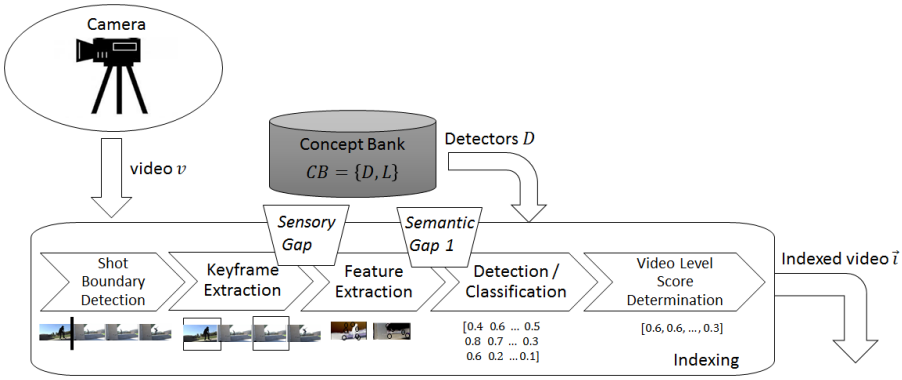


Figure 1.2: Components within indexing: Shot Boundary Detection, Keyframe Extraction, Feature Extraction, Detection / Classification and Video Level Score Determination.

ing for example threshold-based or statistical-based methods (Hu et al., 2011). Additionally the shots are split into keyframes, defined as a single image in a sequence of frames that represent an important point in the sequence. The selection of keyframes should contain salient points and little redundancy. Several methods used to extract keyframes are: sequential comparison-based, global comparison-based, reference frame-based, clustering-based, curve simplication-based, and object/event-based (Truong et al., 2007). An often used strategy is to pick one keyframe in every two seconds of video. Additionally, features are extracted from the keyframes. Features are distinctive characteristics of an image, such as colors, edges, textures and shapes (Snoek et al., 2008). One of the challenges in the feature extraction is the *sensory gap* (Smeulders et al., 2000). This is the gap between a concept in the world and the information in a digital recording of that concept. This implies that the extracted features should be invariant to for example illumination, rotation, scale, translation and viewpoint. Some common features used on images include Scale-Invariant Feature Transform (SIFT) (Lowe, 2004), Speeded Up Robust Feature (SURF) (Bay et al., 2008), Histogram of Oriented Gradients (HOG) (Dalal et al., 2005), Opponent-SIFT (Tuytelaars et al., 2008; Van De Sande et al., 2010), color histogram (Novak et al., 1992; Wang et al., 2010b) and Local Binary Pattern (LBP) (Ojala et al., 2002). Some features used on shot level include Spatio-Temporal Interest Points (STIP) (Laptev et al., 2003), Motion Boundary Histograms (MBH) (Dalal et al., 2006) and Histogram of Optical Flow (HOF) (Laptev et al., 2008). Currently, Convolutional Neural Networks (CNN) are used to extract the features (Krizhevsky et al., 2012; Sharif Razavian et al., 2014; Jia et al., 2014; Karpathy et al., 2014). Examples of these CNNs are AlexNet (Krizhevsky et al., 2012), VGG(16/19) (Simonyan et al., 2014), GoogleNet (Szegedy et al., 2015), Inception V3 (Szegedy et al., 2016) and ResNet (He et al., 2016). The extracted features from the image (or shot) are often named descriptors. These descriptors are then converted into a feature representation. This representation is often a Bag of (Visual) Words (BoW), in which each descriptor is assigned to the closest entry of a visual vocabulary / codebook. This codebook is previously learned on a large dataset. The histogram of BoW, a Fisher vector representation (Perronnin et al., 2010) or Vector of

Locally Aggregated Descriptors (VLAD) pooling (Jegou et al., 2012) is used as a representation of the image. With this representation a classification to a concept can be done, using the pre-trained concept detectors from the Concept Bank. The concept detectors are trained using image representations as input and the concept classes as output. A combination of large datasets, such as ImageNet (Deng et al., 2009), PASCAL VOC (Everingham et al., 2015) and TRECVID SIN (Over et al., 2015), are often used to create the Concept Bank. In the past decades, machine learning methods, such as SVMs, Bayesian Classifiers and Random Forests (Liu et al., 2007; Jiang et al., 2012) were used, but currently deep learning techniques are the common state of the art (Karpathy et al., 2014; Schmidhuber, 2015; Jiang et al., 2017). These deep learning techniques often process some of the previous steps, such as the feature extraction, implicit in the network. The deep learning techniques either take an image or a video as input and directly output the classification. The output score for the classification is a vector with a length that is equal to the number of concepts and each item has a value between zero and one, resembling a confidence that the concept is present in the image. If the classification was applied on keyframe or shot level, some kind of pooling, such as average or max pooling (Wang et al., 2010a; Zhang et al., 2015c), is applied for the video level score determination step to create the index video score ($\vec{i}$).

One of the major challenges in multimedia information retrieval, related to the indexing, is the *semantic gap* and it is defined as the gap between the abstraction level of the pixels in a video and the semantic interpretation of the pixels (Smeulders et al., 2000). This gap can be split into two parts (Hare et al., 2006): the gap between descriptors and object labels (semantic gap 1) and the gap between object labels (concepts) and full semantics (semantic gap 2). As explained in the previous paragraph, descriptors are the feature vectors of an image and object labels or concepts are the symbolic names for the objects in the image. Full semantics is the meaning of the words in the query or even the intent of the user. The first gap is also referred to as automatic image annotation and is part of the indexing of the video. One of the challenges with this gap is the context dependency on the training examples, i.e. only training the concept *fire* in the context of a campfire might not detect a house on fire. The concept labels should, thus, be disambiguous and precise (*campfire*). The second gap is related to the query interpretation (see next section).

An advantage of indexing a video is that the videos are searchable through the pretrained concepts available in the Concept Bank, and not all low-level features have to be stored in the database (which requires more space and search time). Because of big annotated datasets such as ImageNet (Russakovsky et al., 2015), PASCAL VOC (Everingham et al., 2015) and TRECVID SIN (Over et al., 2015), the improvement of hardware and computing power (i.e. GPU) and the new techniques within the field of deep learning, the detection accuracy of concepts in video and images has significantly improved in the past decade (Deng, 2014; Awad et al., 2016b).

1.2.2. QUERY INTERPRETATION

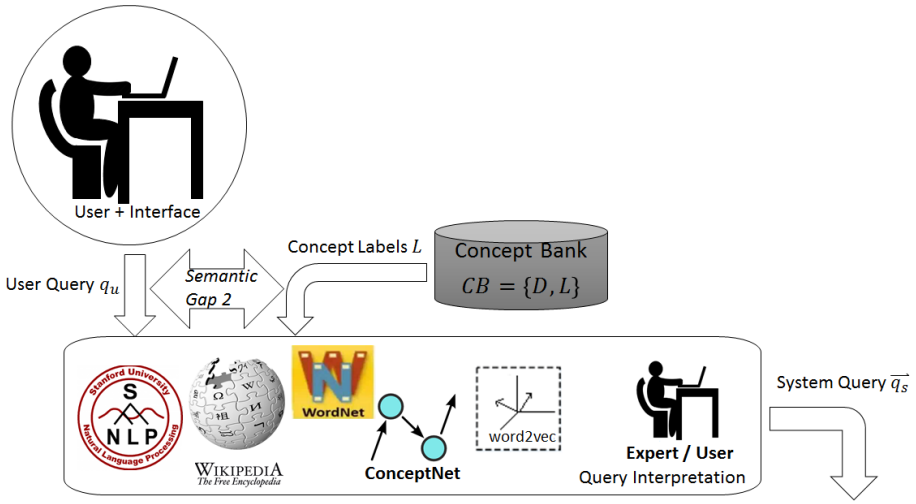


Figure 1.3: Examples of tools used within Query Interpretation: knowledge bases such as Wikipedia, WordNet and ConceptNet, machine learning such as the semantic embedding method word2vec, manual mapping through a expert or user.

The second component is the Query Interpretation (*QI*), of which an illustrative figure is shown in Figure 1.3. Query Interpretation is often used in text retrieval, to map the user query to words present in the documents (Zhao et al., 2012). Some of the approaches include linguistic analysis (i.e. stemming, ontology browsing and syntactic parsing), corpus-specific techniques (concept terms and term clustering), query-specific techniques (distribution difference analysis, model-based Automatic Query Expansion, document summarization), search log analysis (related queries, query-document relationships) and web data techniques (anchor texts, Wikipedia) (Carpineto et al., 2012). One of the challenges in this mapping is the *vocabulary mismatch*. This is the (semantic) mismatch between the concepts people use to formulate their query and the concepts that are used to index the document, image or video with. In the image or video domain the vocabulary of trained concepts is much smaller than the vocabulary of for example the English language. One reason is that not all words in a vocabulary are visually presentable and another reason is that it is unfeasible to train concept detectors for all words in the English vocabulary. A related challenge, specifically for the image / video retrieval, is the semantic gap. As explained in the previous section, the gap between object labels (concepts) and full semantics (semantic gap 2) is related to the query interpretation: to capture the meaning or intent of the user query in terms of the available object labels, or concepts, which we define as *query-to-concept mapping*. As an example, the user query *extinguishing a fire* needs at least the concepts *fire, fire extinguisher* and *fire (wo)man* to be able to match the semantics of the information need underlying the query.

The query-to-concept mapping, also named concept selection, is often done using one of the following three categories: ontologies, machine learning or manual mapping (Liu et al., 2007). Ontologies are conceptual representations of the world.

Ontologies or knowledge bases can be created by expert (*expert knowledge base*) or created by the public (*common knowledge base*). Expert knowledge bases provide good performance, but dedicated expert effort is needed in the creation of such a knowledge base. Some early work on expert knowledge bases and reasoning in the field of event recognition is explained in Ballan et al. (2011). One current expert ontology for events is EventNet (Ye et al., 2015). Common knowledge bases, such as Wikipedia (Milne et al., 2013) and WordNet (Miller, 1995), are freely available and often used in the video retrieval community (Neo et al., 2006; Yan et al., 2015; Tzelepis et al., 2016), but might not contain the specific information that is needed. The query-to-concept mapping in common knowledge bases is often done by using the most similar or related concepts to events found in the knowledge base. An overview of the type of methods to find similar or related concepts can be found in Natsev et al. (2007).

Machine learning techniques can be used to automatically select the proper concepts. Examples are graphical models, such as hidden Markov models (Dalton et al., 2013), and statistical methods, such as co-occurrence statistics (Mensink et al., 2014) and a skip-gram model (Chang et al., 2015). One group of current state of the art models is *word2vec*, which produce semantic embeddings. These models either use skip-grams or continuous bag of words (CBOW) to create neural word embeddings using a shallow neural network that is trained on a huge dataset, such as Wikipedia, Gigawords, Google News or Twitter. Each word vector is trained to maximize the log probability of neighboring words, resulting in a good performance in associations, such as *king - man + woman = queen*.

Currently, often a manual mapping from the user query to a set of concepts is used in benchmarks such as TRECVID Multimedia Event Detection (Awad et al., 2016a). This *manual mapping* can either be done by an expert or the user that entered the query. Allowing the user to map the query to the concepts in the Concept Bank implies that the user should learn the concepts in the Concept Bank, and asks for additional effort. The expert mapping is, similar to the expert knowledge base, probably good performing, but unfeasible with many user queries.

1.2.3. FEEDBACK INTERPRETATION

Feedback interpretation can be divided into Relevance Feedback (*RF*) and active learning approaches (Snoek et al., 2008). Relevance feedback displays the most relevant results on top and uses the positively and negatively annotated results to update the system query in an iterative process. Relevance feedback is typically used in an ad-hoc case, such as our use case. Active learning displays the most uncertain results, i.e. the results that are most informative for the system, to the user to quickly learn an optimal model. Active learning is typically used to train (or improve) concepts. In this section, we will only focus on relevance feedback. An illustration is shown in Figure 1.4.

The use of relevance feedback stems from the dynamic nature of information seeking (Ruthven et al., 2003): information needs can be continuously changing and be unique to each user. Relevance feedback can be done in different ways: implicit, explicit and blind/pseudo. In *implicit* relevance feedback, implicit information, such

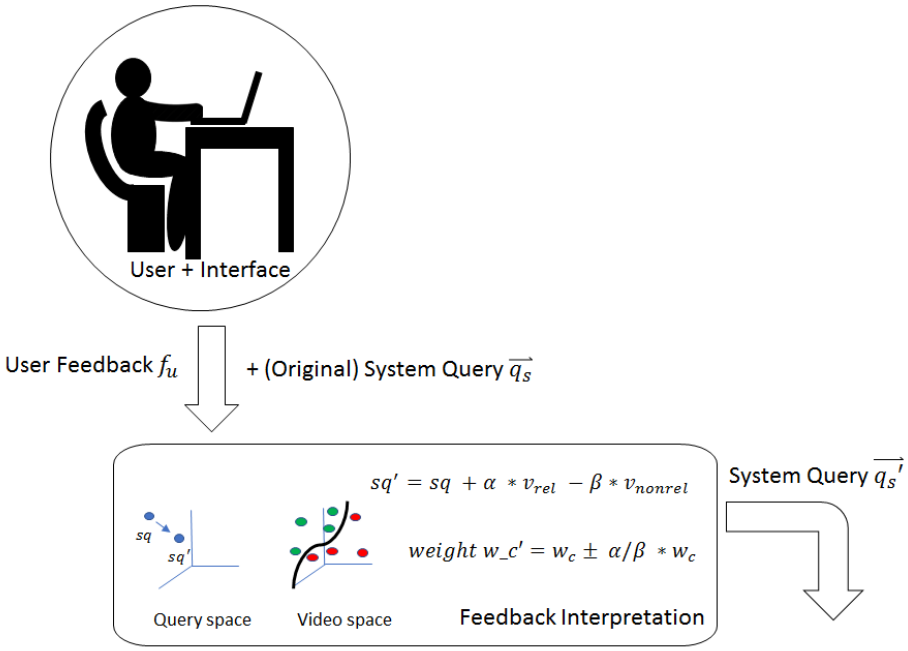


Figure 1.4: Examples of algorithms used in Feedback Interpretation: changing the concept space in for example query point modification, changing the video space in for example cluster-based methods, changing the weights in re-weighting and using Rocchio to change the system query.

as user clicks or dwell time, is used. The advantage of this method is that you do not have to bother the user, but the inference of the results is much harder. In *explicit* relevance feedback, the user explicitly indicates if a certain item is relevant or not relevant. This can be done using a binary scale or a gradual scale. The advantage of this method is that you have a clear indication of the relevance and a higher performance, but the disadvantage is that you have to bother the user. This user might not have time or motivation to give such feedback. In *blind- or pseudo*-relevance feedback, the manual user part is automated. In this automation, we assume that the first k ranked items are relevant. This assumption is not without a risk, because in the case of rare events or new query domains, bad retrieval systems or ambiguous queries this assumption might not hold. Human relevance feedback (implicit and explicit) has been known to provide major improvements in precision for information retrieval system. Dalton et al. (2013) have shown that — in the domain of video retrieval — pseudo-relevance feedback can increase Mean Average Precision (MAP) up to 25%, whereas with human judgments this number can grow up to 55%. Of course the effectiveness of pseudo-relevance feedback critically depends on the assumption that the collection contains at least a reasonable number of relevant results and that the first retrieval pass is able to pick up a good fraction of those in the top k. It is clear that relevance feedback, when applied correctly, can help the user in better finding results.

One of the most well-known and applied relevance feedback algorithms that has its origins in text retrieval is the Rocchio algorithm (Rocchio, 1971). This algorithm is used in state of the art video and text retrieval systems that use for example Query Point Modification (QPM) to move the query representation in the vector space and re-weighting in which the terms in the query are re-weighted (Rocha et al., 2015; Jiang et al., 2014b; Tsai et al., 2015; Kaliciak et al., 2013). Often a document is represented as a vector with a real-valued component, e.g. tf-idf weight (see next section), for each word. The Rocchio algorithm works on a vector space model in which the query drifts away from the negatively annotated documents and converges to the positively annotated documents. The Rocchio algorithm is effective in relevance feedback, fast to use and easy to implement. The disadvantages of the method are that its α and β parameters have to be tuned and it cannot handle multimodal classes properly.

Other state of the art approaches, such as feature-, navigation-pattern, and cluster-based approaches, in image retrieval are explained by Zhou et al. (2003) and Patil (2012). Some vector space models use k-Nearest Neighbor methods, such as in the studies by Gia et al. (2004) and Deselaers et al. (2008). Other methods use decision trees, SVMs, or multi-instant approaches are explained in Crucianu et al. (2004).

1.2.4. SCORING & RANKING

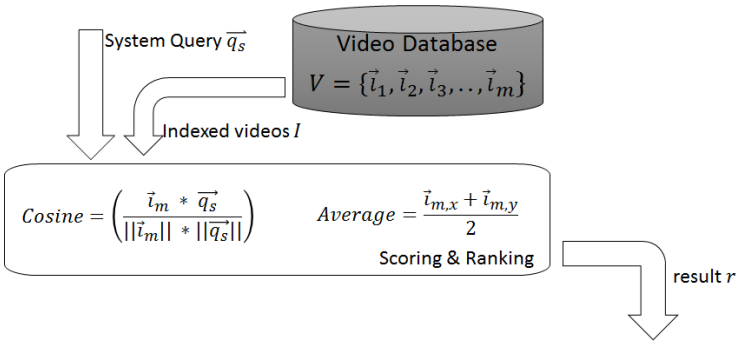


Figure 1.5: Algorithm (cosine similarity) and often used Average Fusion in Scoring & Ranking component

The final component is Scoring & Ranking, illustrated by Figure 1.5. This component deals with the retrieval models and fusion. Examples of models are set-theoretic, algebraic and probabilistic models (Baeza-Yates et al., 1999). The *set-theoretic* or Boolean model represents the system query with AND, OR and NOT operators. The video or document is indexed with concepts in a binary fashion (present or not). The Boolean model returns the videos or documents that contain the concepts that need to be (not) present according to the query. This type of model can be extended towards fuzzy sets, in which the binary separation is less strict. The Boolean models are not often used, because they have no gradation or weighting and do not allow partial matching. The *algebraic* or vector space model, which is the model of our choice, computes the distance between the system query and each of the videos or documents. This model allows gradation, or ranking of the

videos, as well as weighting. In text retrieval the most basic weight is a TF-IDF value, which represents how often a word in the query occurs in a certain document normalized by the number of documents the word occurs in. This allows specific words to have a higher weight compared to often occuring words, such as stopwords. The distance between the query and the video or document is often calculated through a cosine similarity (Salton et al., 1988; Salton et al., 1997). Latent Semantic Indexing and Neural Network approaches are also algebraic models that allow implicit weights and more sophisticated distance measures. *Probabilistic* models calculate the probability that a video or document is relevant and the probability that it is not relevant. The ratio of these probabilites often determines the ranking. This probability can be determined with for example language models, Bayesian Networks or measures such as BM25. Both vector space and probabilistic models can perform well, but their retrieval performance is dependent on the validity of the assumptions, such as the independence assumption.

Often multimedia information retrieval systems do not only rely on one source of information. The Concept Bank can for example contain concepts trained on different datasets, as well as visual and motion information. These different types of information (visual and motion) are defined as a modality or data source. In the Scoring & Ranking component, this information should be fused. Atrey et al. (2010) give an overview of the multimodal fusion methods in multimedia analysis. Firstly, a distinction between early fusion on feature level and late fusion on decision level is made. The advantage of early fusion is that correlations between multiple features can be used, but it can be hard to create a meaningful combined feature vector. Lan et al. (2012) add that early fusion techniques suffer from the curse of dimensionality and require much training data. According to Atrey et al. (2010), the advantage of late fusion is that the most suitable method for a single modality can be applied and it is more flexible and robust to features that have a negative influence compared to early fusion. A disadvantage is that the correlation between modalities cannot be fully exploited.

Besides the level of fusion, the method of fusion is also important. Two of the methods explained in Atrey et al. (2010) are rule-based methods and classification-based methods. Examples of rule-based methods are linear weighted fusion and manually defined rules. In the linear weighted fusion some form of normalization and weighting is used to combine different modalities. In general, the rule-based methods are computationally inexpensive and easy to implement, but the assignment of appropriate weights remains an issue. This method is often used in late fusion. Oh et al. (2014) further split late fusion into a blind method with fixed rules, such as geometric mean, a normalization method with assumptions on score distributions and a learning method that needs training data to set an appropriate weight. The difficulty of assigning appropriate weights made us focus on blind methods with fixed rules for the integration of different information sources.

According to Xu et al. (1992), three types of classifier outputs can be used in fusion: 1) abstract level: single class label; 2) rank level: ordered sequence of candidate classes; 3) measurement level: candidate classes with confidence scores. According to Tulyakov et al. (2008), voting techniques such as majority voting and borda count

are the methods most used for the abstract and rank level classifiers, whereas sum, product and max-rules are the elementary combinations on measurement level.

1.3. RESEARCH QUESTIONS

In this thesis, we cannot focus on all components to create the search capability. Given our aim that the search capability should handle ad-hoc textual queries, the major challenge is the Query Interpretation. We assume that we have no training examples for the ad-hoc queries, and, thus, we should represent the query in terms of the concepts used to index the video (*query-to-concept mapping*). Our main research question is:

Main Research Question: *How can we improve visual search effectiveness by semantic query-to-concept mapping?*

Our main research question touches some of the central problems of Artificial Intelligence, such as reasoning, knowledge representation, natural language processing and perception. Both the vocabulary mismatch and the semantic gap are present in the semantic query-to-concept mapping. We zoom in on cases in which the videos are already indexed with concepts, disregarding the challenges in processing the video, training concept detectors and storing the videos in a scalable way. Because of the decision to index the video with concepts instead of non-semantic descriptors or features, we, however, use a scalable solution that has the complexity of $O(n * m)$, where n is the number of concept detectors and m is the number of videos. Because we assume that the number of concepts is much lower than the number of features, our query-to-concept mapping enables a scalable system.

Additionally, we have two assumptions. Our first assumption is the *open-world assumption*. This means that it is not the case that our system has complete world knowledge. This means that not all queries are known at design time of the system, which is the case in the ad-hoc search task. Because not all queries are known not enough concepts can be trained to cover all queries, hence we have a vocabulary mismatch. Although current state of the art systems are able to detect an increasingly large number of concepts, this number still falls far behind the near infinite number of possible (textual) queries that a system needs to be able to handle. In order to handle the ad-hoc queries with a vocabulary mismatch, we assume that *our query can be decomposed to smaller pieces* (decompositionality). For a certain unseen event 'birthday party', we assume that we can use related concepts, such as a group of people, a cake and decorations and relations between those concepts to capture the essence of that unseen event. The system query can then be formalized as a vector of concepts in which each concept has a weight (Equation 1.1).

$$\overrightarrow{q_s} = [w_{c1}, w_{c2}, w_{c3}, .. w_{cn}] \tag{1.1}$$

As a working hypothesis, we use a sparse linear combination (inproduct) of these concepts as our retrieval model (Equation 1.2). This is based on the fact that it is hard to determine a good non-linear retrieval model without training examples.

$$r_v = \vec{q_s} \cdot \vec{i_v} \tag{1.2}$$

Given the main research question and our assumptions, we have formulated some in depth research questions that help to answer the main research question. Figure 1.6 places the research questions in the visual search system that was introduced in the previous section. Each research question is addressed in one chapter of this thesis, which relates to one peer reviewed paper. Additional published papers are mentioned in this section and a full list of publications is included on page 161.

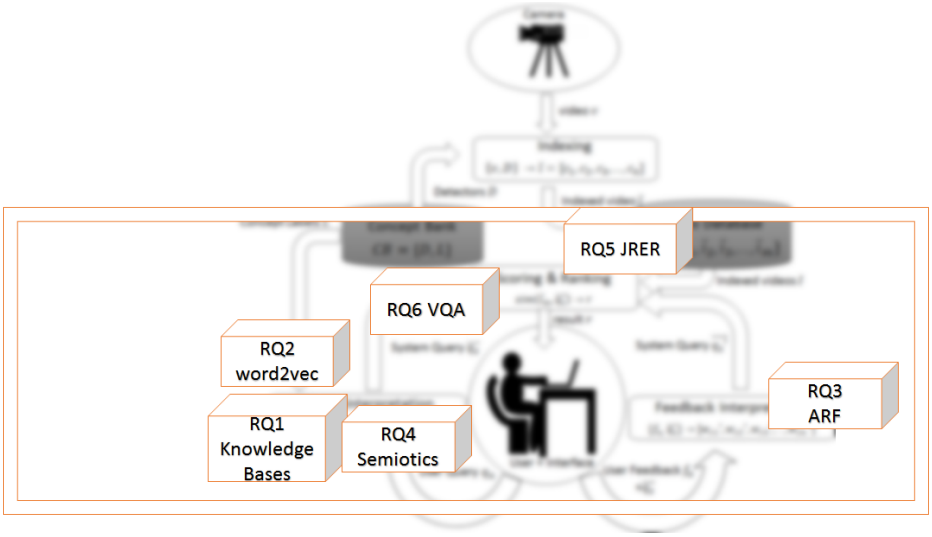


Figure 1.6: Visualization of the Research Questions in the Visual Search System

The first research questions are related to the method that is used to apply for the query-to-concept mapping. The query-to-concept mapping is currently done using any of the following three methods: using knowledge bases, using a machine learning approach, such as an semantic embedding, or using a manual mapping. The first research question, named KnowledgeBases, explores the incorporation of knowledge bases as a method in the Query Interpretation:

RQ1 KnowledgeBases: *How can we incorporate knowledge bases in query-to-concept mapping and which of the current knowledge bases is most applicable for this purpose?* (Chapter 2)

Based on the literature, we choose to compare three knowledge bases. We distinguish common knowledge bases such as Wikipedia and expert knowledge bases such as a manually created ontology. We use the text retrieval method TF-IDF on the text from Wikipedia, the graphical structure / ontology from ConceptNet to exploit the strength of the relations between words and a manually created ontology

as our in-domain knowledge base. This research is published in Boer et al. (2015b). We also explored directions with the knowledge base WordNet, such as natural language processing with WordNet (Boer et al., 2013; Bouma et al., 2013b) and TFIDF with WordNet (Ngo et al., 2014; Lu et al., 2016b). The result of Lu et al. (2016b) is mentioned in chapter 3, but the methods are not included in this thesis.

Knowledge bases are, however, not the only way to perform the query-to-concept mapping. In 2013, Mikolov et al. (2013) introduced word2vec, which is a method based on semantic embeddings. This method extracts knowledge from large text corpora, without the explicit modelling and structuring that is needed for the knowledge bases. The disadvantage of this method is that it can only model that two words are used in the same context, but not what this specific relation is. In the research question word2vec, we investigate whether word2vec is a better alternative in query-to-concept mapping:

RQ2 word2vec: *How can we use semantic word embedding in query-to-concept mapping, and how does the mapping depend on the concepts in the Concept Bank?* (Chapter 3)

We explore the semantic embedding based on word2vec and different types of concepts in the Concept Bank, such as low-level objects or scenes, mid-level actions and high-level events, and how these types of concepts influence the mapping and effectiveness of the whole search capability in terms of Mean Average Precision. This research is published in Boer et al. (2017b). An initial exploration of word2vec is also published in Zhang et al. (2015a).

The third type of mapping is the manual mapping. A disadvantage of a manual mapping is that the user should know all concepts in the Concept Bank to be able to select the relevant concepts. This is often not feasible, and thus the third research question explores how we can involve the user in the query-to-concept mapping without the manual mapping:

RQ3 ARF: *How can we involve the user to optimize semantic mapping and retrieval performance for a visual search capability?* (Chapter 4)

We propose and compare methods that use the user feedback on both concept level and video level. For the video level relevance feedback, we base our method on literature in text retrieval. We propose an Adaptive Relevance Feedback (ARF) algorithm that uses the Rocchio algorithm (Rocchio, 1971) in the video retrieval domain by using the concept space. We compare our algorithm to a k-Nearest Neighbour (k-NN) method (Gia et al., 2004; Deselaers et al., 2008) that is perceived as state of the art in video retrieval, and several algorithms on concept level, such as Query Point Modification and re-weighting. This research is published in Boer et al. (2017a). Additional experiments on ARF are published in Pingen et al. (2017). These experiments include different modes of feedback, such as optimal, random and pseudo relevance feedback, and the usability of our video retrieval system named AVES. An exploration

on personal re-ranking of results using k-NN for queries with subjective adjectives, such as *dangerous animal*, can be found in Schavemaker et al. (2015).

The success of the query-to-concept mapping is not only dependent on the model, Concept Bank and user feedback, but the type of query might have influence on the performance. Some type of queries might have a higher performance, because they have a better fit to the type of semantic structure that is assumed in the mapping. As an example, for some queries a mapping from 'man' in the query to 'woman' as a concept might not hurt performance, whereas in other queries it might. In this research question, we explore different types of semantic structures in a query:

RQ4 Semiotics: *To what extent can semantic structures increase understanding of the query?* (Chapter 5)

We create a set of queries with different objects and spatial relations, and provide ground truth for all concepts and images in a created Toy and Office-Supplies Objects (TOSO) dataset. We compare performance of methods that only use specific semiotic structures and all structures on both concept level (semantic query-to-concept mapping) and image (retrieval) level. This research is published in Boer et al. (2015c). Initial papers on this topic can be found in Boer et al. (2015a) and Schutte et al. (2015a).

Our fifth research question is not directly used in the Query Interpretation or query-to-concept mapping, but focuses on fusion. Most concept detectors are not trained on one dataset. In this research question, we look into fusion of trained classifier outputs based on their score distribution and dependency:

RQ5 JRER: *Can we design a more effective score fusion method that is motivated by explicit assumptions about the distribution of classifier output values and the dependency between input sources?* (Chapter 6)

We introduce a novel blind late fusion method named Joint Ratio Exterme Ratio (JRER) to combine information from multiple modalities. This method is based on state of the art methods, such as the average fusion or joint probability. This research is published in Boer et al. (2016b).

Other research questions on the query-to-concept mapping have focused on an explicit mapping of the query to specific concepts, but currently many deep learning methods used in image understanding do not use an explicit mapping. Our last research question aims to explore these implicit mappings:

RQ6 VQA: *What are the possibilities of implicit query-to-concept mapping in terms of visual search effectiveness?* (Chapter 7)

We explore the state of the art deep learning network named DPPnet (Noh et al., 2016) in a benchmark that involves a query and an image as input and a textual answer as output. We improve upon this method by adding concept detectors to the

network and postprocessing the answers to filter on the right type of answer. This research is published in Boer et al. (2016a).

1.4. RESEARCH METHODS AND COLLABORATION

As an experimental testbed, we would like to use a large statistical dataset, which is more tuned to a retrieval case (forensic) than a real-time case with alerts (monitoring). Our methods can, however, be used for video streaming data. A security task would be in scope for this thesis, as this is used as an inspiration for this research. Although many countries, such as the US, have big research programs for the security domain, such as DARPA (military domain)[1] and IARPA (homeland security domain)[2], we could not find a dataset that met our requirements. The known benchmarks for the security domain, such as the Performance Evaluation of Tracking and Surveillance (PETS) (Ferryman et al., 2009) and the i-Lids data that is used in the TRECVID Surveillance Event Detection (SED) task (Over et al., 2015), contain (pre-defined) actions such as picking up a phone, a group formation or speeding up. These events need a specifically trained action classifier, because the combination of a person and a phone might not be sufficient to retrieve the specific action of picking up a phone. In this thesis, we focus on high-level events, which can be described in terms of objects and actions between those objects, i.e. meets our requirement of decompositionality. To our knowledge, no large annotated dataset is available for complex events in the security domain.

Within the multimedia information retrieval field, many standard test collections, or international benchmarks, are available. With the increase of performance on existing benchmarks, new more complex benchmarks were created. The advances in object detection through benchmarks, such as PASCAL VOC (Everingham et al., 2015) and ImageNet (Russakovsky et al., 2015), have led to harder datasets such as MSCOCO (Lin et al., 2014) or VisualQA (Antol et al., 2015). These datasets have led to advances in image understanding. The earlier datasets were often contructed and therefore less generalizable, whereas MSCOCO and VisualQA contain more real world data, which allows researchers to measure whether the tested systems are also generalizable or applicable for real world applications. Within the event retrieval domain, the advances in the TRECVID MED benchmark on the detection of events using 100 examples and 10 examples have led to the introduction of a task without any training examples.

Because of the disadvantages of creating a new large annotated dataset, we choose to use the TRECVID Multimedia Event Detection (*MED*) task in the majority of our work (Awad et al., 2016a). The data set that is released with the task contains videos of different types of quality, which resembles the surveillance domain with static cameras, body cams and mobile phone videos. The MED task consists of a train set of forty events and a test set of twenty events, varying from social events such as *tailgating* to procedural events such as *changing a vehicle tire*. These events can typically be described using multiple concepts, such as objects, actions and

[1] https://www.darpa.mil/
[2] https://www.iarpa.gov/

scenes. Although these events have been inspired by cases from the homeland security domain, the transfer of the developed knowledge in this thesis to a security case is not a part or goal of this thesis. The events have the same magnitude of visual dissimilarity, i.e. every instance of an event looks different, but all are positive for the event, and there is a high imbalance of positive examples in the videos compared to the number of videos in the database (20 positives on 27.000 videos). The evaluation metric of the TRECVID MED is the Mean Average Precision, which focuses on the rank of the positive examples, i.e. penalizes systems for the positive examples that are not retrieved. This metric is not uncommon in a forensic case, in which a combination of true positives, false positives, true negatives and/or false negatives is often used (Schütze, 2008). In the monitoring case the metric is often based on false alarms and misses, such as Normalized Detection Cost Rate (NDCR), Area under the Curve (AUC), or ROC curve (Bouma et al., 2014; Over et al., 2015). The dataset and evaluation metric are, thus, suitable as an experimental testbed for our research question.

The TRECVID MED benchmark is of such a size that it takes a large research team to complete the full task. After a TNO-only participation in 2013, we intensively collaborated with the VIREO team from the City University of Hong Kong in 2014 and 2015, from where we received all the concepts in the Concept Bank and the detections on all videos. This allowed us to focus on the query-to-concept mapping, without the burden of the indexing component.

1.5. MAIN CONTRIBUTIONS

The main contribution of this thesis can be summarized as:

- We create a smart automatic query-to-concept mapping method named incremental word2vec (**i-w2v**) (Chapter 3). This i-w2v method uses word2vec trained on GoogleNews items as a semantic embedding model and incrementally adds concepts to the set of selected concepts for a query in order to deal with query drift. In combination with a state of the art video event retrieval pipeline, we achieve top performance on the TRECVID MED benchmark regarding the zero-example task (MED14Test results).

- We propose a feedback interpretation method named **ARF** that not only achieves high retrieval performance, but is also theoretically founded through the Rocchio algorithm (Rocchio, 1971) from the text retrieval field (Chapter 4). This algorithm is adjusted to the event retrieval domain in a way that the weights for the concepts are changed based on the positive and negative annotations on videos. The ARF method has higher visual search effectiveness compared to k-NN based methods on video level annotations and methods based on concept level annotations.

- We introduce several blind late fusion methods based on a combination of state of the art methods (Chapter 6). Especially the **JRER** method achieves high performance in cases with reliable detectors, i.e. enough training examples.

1.6. THESIS OUTLINE

In the remainder of this thesis, we follow the structure and order of the research questions in which one research question is addressed in one chapter. All chapters are based on a peer-reviewed and published paper. This means that the chapters can be read separately. In **Chapter 2** we explore different knowledge bases that can be used to create a query-to-concept mapping. In **Chapter 3** we propose a novel semantic embedding method named i-w2v that improves upon word2vec that achieves state of the art performance. We also compare vocabularies / Concept Banks with different types of concepts in terms of complexity, i.e. low- (object, scene), mid- (action) and high-level (event). **Chapter 4** is dedicated to user feedback. We analyze different feedback interpretation methods, both on concept level and video level. Additionally, we propose a novel method named ARF based on the well-known Rocchio algorithm that improves upon state of the art performance. **Chapter 5** is related to semiotics. We use different type of queries and different type of semantic structures, such as synonyms and functional related words (man vs. woman), to calculate the performance on concept level and image level. **Chapter 6** is focused on fusion. We introduce novel blind late fusion methods based on state of the art methods, such as average and product. In **Chapter 7** we explore how and whether we can use an implicit query-to-concept mapping in a visual question answering task. **Chapter 8** contains the conclusion, limitations and future work. Additionally, the thesis contains a glossary with a definition of the technical terms used in this thesis, a summary (in English and Dutch), the acknowledgements, CV and list of publications.

2

KNOWLEDGE BASED QUERY EXPANSION IN MULTIMEDIA EVENT DETECTION

Edited from: **Maaike de Boer, Klamer Schutte and Wessel Kraaij** (2016) *Knowledge Based Query Expansion in Complex Multimedia Event Detection*. In: Multimedia Tools and Applications (MTAP), volume 75, pp. 9025 - 9043.

For our aim to improve visual search effectiveness, we use a semantic query-to-concept mapping. In order to achieve a good mapping, we need to interpret the query and understand the relation of the query to the concepts. This requires world knowledge. One of the sources that can provide such knowledge are knowledge bases. In this chapter, we propose and compare methodologies to use knowledge bases as a resource for content-based information retrieval of complex events. This chapter is related to the first research question **RQ1 KnowledgeBases**. *We distinguish common knowledge bases such as Wikipedia and expert knowledge bases such as a manually created ontology. We use text retrieval methods, such as TFIDF, to map the query to the most related concepts using Wikipedia through query expansion. We compare that method with a method that exploits the graphical structure of the common knowledge base ConceptNet. Additionally, we use an expert knowledge base on the TRECVID MED events for our comparison. Results on the TRECVID MED test set of 2014 show that using a knowledge base improves performance if a vocabulary mismatch exists, i.e. if the main noun in the query has no direct match to a concept in the Concept Bank. Additionally, the expert knowledge base does not necessarily outperform the approaches using common knowledge bases. From the common knowledge bases, ConceptNet performs slightly better compared to Wikipedia.*

2.1. INTRODUCTION

Retrieving relevant videos for your information need is most often been done by typing a short query in a video search engine such as YouTube (Burgess et al., 2013). Typically, such visual search engines use metadata information such as tags provided with the video, but the information within the video itself can also be extracted by making use of concept detectors. Concepts that can be detected include objects, scenes and actions (Jiang et al., 2012). Concept detectors are trained by exploiting the commonality between a large number of training images. One of the challenges in content-based visual information retrieval is the semantic gap, which is defined as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" (Smeulders et al., 2000). The importance of bridging the semantic gap is reflected by the emergence of benchmarks such as TRECVID (Over et al., 2004) and ImageCLEF (Caputo et al., 2014).

The semantic gap can be split in two sections (Hare et al., 2006): the gap between descriptors and object labels and the gap between object labels and full semantics. Descriptors are feature vectors of an image and object labels are the symbolic names for the objects in the image. Full semantics is the meaning of the words in the query or even the information need of the user. The first gap is also referred to as *automatic image annotation* and progress is made rapidly (Snoek et al., 2010; Russakovsky et al., 2015). For the purpose of this chapter the second gap is considered.

In the second semantic gap, the challenge is to represent the user intent in terms of the available object labels, which are provided by the concept detectors. State-of-the-art methods used to bridge this second semantic gap include query expansion using knowledge bases (Hoque et al., 2013) and relevance feedback (Patil et al., 2011). Relevance feedback is a method that uses feedback from the user, such as explicit relevance judgments or user clicks, to optimize results. Relevance feedback is a powerful tool, but it requires an iterative result ranking process and dedicated algorithms (Patil et al., 2011), which is outside the scope of this chapter. Another disadvantage of relevance feedback is that the system does not know why a video is not relevant.

Knowledge bases, on the other hand, are interpretable for both systems and humans. Knowledge bases can add more relevant words to the short user query to represent the user intent in a better way. This larger user query contains more words and, thus, more potential to match the object labels. Both common knowledge bases such as WordNet (Liu, 2002) or Wikipedia (Hassan et al., 2011) and expert knowledge bases created by an expert can be used (Bagdanov et al., 2007; Tu et al., 2014). Common knowledge bases are easy to access and do not require a lot of dedicated effort to construct, but they might not have sufficient specific information and they can be noisy due to disambiguation problems. The lack of sufficient specific information implies that no additional relevant concept detectors can be selected and the noise can cause the selection of irrelevant concept detectors. Expert knowledge bases may have sufficient specific information and are less noisy, but it requires a lot of dedicated effort to create them.

Our research focuses on which type of knowledge base is best to use in the domain of complex or high-level events, defined as "long-term spatially and temporally dynamic object interactions that happen under certain scene settings" (Jiang et al., 2012). Examples of complex events are *birthday party*, *doing homework* and *doing a bike trick*. In this chapter, only textual information is used as input for the system, which is referred to as the zero-example case. In this situation it is unfeasible to create a dedicated detector for each possible word and we, therefore, have to bridge the semantic gap between the pre-determined labels assigned to the image and the full semantics of the event. Complex events cannot be captured by a single object, scene or action description and, therefore, complex events have a large semantic gap.

In our experiments, we use the Test Set of TRECVID 2014 Multimedia Event Detection (MED) task (Over et al., 2013) to compare retrieval performance on the complex event query, ConceptNet 5 (Speer et al., 2012) and Wikipedia as common knowledge bases and the textual description provided with the TRECVID task to determine which type of knowledge base is best to use. ConceptNet and Wikipedia are chosen, because both are easy accessible and provide information about complex events. We expect that query expansion has a positive effect on performance, especially if the main noun of the query cannot be detected with the available concept detectors. Because common knowledge bases are not tailored, expert knowledge bases might be able to outperform common knowledge. No difference in performance of ConceptNet and Wikipedia is expected. Fusion, on the other hand, is expected to increase performance, because not all knowledge bases will provide the same information.

In the next section, related work about the query expansion using knowledge bases and complex event detection is reviewed. The third section contains information about the method with the TRECVID MED task and design of the experiment. Section 2.4 consists of the results and the last section contains the discussion, conclusions and future work.

2.2. RELATED WORK

2.2.1. QUERY EXPANSION USING KNOWLEDGE BASES

One of the challenges in keyword search is that the user uses different words in the query than the descriptors used for indexing (Bodner et al., 1996). Another challenge is that users often provide a short, vague or ill-formed query (Bodner et al., 1996). In order to find relevant results, the query has to be expanded with relevant, related words, such as synonyms. Computers have no knowledge of our world or language themselves and, therefore, cannot use this information in the way humans do. In order to automatically expand the query without requiring the user to reformulate the query, computer systems should have access to world knowledge and language knowledge. One way to provide this knowledge is to use a knowledge base (Bodner et al., 1996). Two types of knowledge bases exist: common knowledge bases and expert knowledge bases. In Bodner et al. (1996) these are called *General World Knowledge Base* and *Domain Specific Knowledge Base*, respectively. Both types of knowledge bases are accessible on the Internet because of the Semantic Web and Linked Open Data (LOD) initiative (Sheth et al., 2005; Baeza-Yates et al., 2008). The Semantic Web

is about exposure of structured information on the Web and the LOD is about linking the structured information. This information is often structured using an ontology, which is a formal way to represent knowledge with descriptions of concepts and relations. An advantage of using ontologies is that they provide a formal framework for supporting explicit, specific and machine-processable knowledge and provide inference and reasoning to infer implicit knowledge (Ballan et al., 2011). Several standards such as OWL (Web Ontology Language) are easy accessible. A disadvantage of an ontology is that the knowledge has to be inserted in the framework manually.

COMMON KNOWLEDGE BASES

Many common knowledge bases are available on the Internet and this section can, therefore, not include all available common knowledge bases. Many comparisons between common knowledge bases are available including Mascardi et al. (2007) and Zon (2014). The Linked Open Data initiative gave rise to using existing common knowledge bases in order to expand your own common knowledge base. One example is ConceptNet 5, which is a knowledge representation project in which a semantic graph with general human knowledge is build. This general human knowledge is collected using other knowledge bases, such as Wikipedia and WordNet, and experts and volunteers playing a game called *Verbosity* (Von Ahn et al., 2006). Some of the relations extracted using this game are *RelatedTo*, *IsA*, *partOf*, *HasA*, *UsedFor*, *CapableOf*, *AtLocation*, *Causes*, *HasSubEvent*, *HasProperty*, *MotivatedByGoal*, *ObstructedBy*, *CreatedBy*, *Synonym* and *DefinedAs*. The strength of the relation is determined by the number and reliability of the sources asserting the fact. As of April 2012, ConceptNet contains 3.9 million concepts and 12.5 million links between concepts (Speer et al., 2012). Experiments on the previous version of ConceptNet, which is ConceptNet 4, indicated that the knowledge base is helpful in expanding difficult queries (Kotov et al., 2012).

Besides factual knowledge, the common knowledge base Wikipedia contains encyclopedic information. Wikipedia is a free multi-lingual online encyclopedia edited by a large number of volunteers. Wikipedia contains over 4.8 English million articles. Both information on Wikipedia pages and links between the pages are often used (Voss, 2005). An open source tool kit for accessing and using Wikipedia is available (Milne et al., 2013) and many other common knowledge bases include information or links from Wikipedia, such as YAGO2 (Hoffart et al., 2013) and ConceptNet (Speer et al., 2012).

Besides encyclopedic and factual knowledge bases, WordNet is a hierarchical dictionary containing lexical relations between words, such as synonyms, hyponyms, hypernyms and antonyms (Miller, 1995). It also provides all possible meanings of the word, which are called *synsets*, together with a short definition and usage examples. WordNet contains over 155,000 words and over 206,900 word-sense pairs. WordNet is often used to expand a query with similar words (Carpineto et al., 2012) and several similarity measures can be used (Pedersen et al., 2004). Most similarity measures use path-based algorithms.

The common knowledge base sources described above are easy to access, provide enough data for statistical analysis and do not require a lot of human effort to get results, but they might not have sufficient specific information or they might be noisy.

Query expansion using these knowledge bases can also suffer from query drift, which means that the focus of the search topic shifts due to a wrong expansion (Carpineto et al., 2012). Query expansion using common knowledge bases most often moves the query to the most popular meaning.

EXPERT KNOWLEDGE BASES

Besides many common knowledge bases, many expert knowledge bases exist such as in the field of geography (Vatant et al., 2012) and medicine (Pisanelli, 2004), but also in applications in multimedia (Naphade et al., 2006), video surveillance (Francois et al., 2005), bank attacks (Georis et al., 2004) and soccer (Bagdanov et al., 2007). Expert knowledge bases are domain-specific, because disambiguation, jargon and structure of concepts and relations is unfeasible in the open domain. Expert knowledge bases are complete and have good performance in information retrieval tasks, but dedicated expert effort in creation of the ontology is a big disadvantage.

2.2.2. COMPLEX EVENT DETECTION

Complex or high-level events are defined as "long-term spatially and temporally dynamic object interactions that happen under certain scene settings" (Jiang et al., 2012) or "something happening at a given time and in a given location" (Ballan et al., 2011). Research regarding complex event detection and the semantic gap increased with the benchmark TRECVID. Complex events cannot be captured by a single object, scene, movement or action. Research mainly focused on what features and concept detectors to use (Naphade et al., 2006; Habibian et al., 2013) and how to fuse results of these concept detectors (Natarajan et al., 2011). The standard approach for event detection is a statistical approach to learn a discriminative model from visual examples. This is an effective way, but it is not applicable for cases in which no or few examples are available and the models cannot give interpretation or understanding of the semantics in the event. If few examples are available, the web is a powerful tool to get more examples (Mazloom et al., 2013a; Ma et al., 2012).

On the web, common knowledge bases can be accessed for query expansion in complex event detection. WordNet (Miller, 1995) is for example used to translate the query words in visual concepts (Natsev et al., 2007). Wikipedia is often successfully used to expand a query in image and video retrieval (Leong et al., 2011; Hoque et al., 2013). A challenge with these methods is that common knowledge sources use text and many words are not 'picturable'. These words cannot be captured in a picture and are often abstract, such as *science*, *knowledge* and *government*. One approach to deal with this challenge is to use Flickr. Both Leong et al. (2011) and Chen et al. (2014) use Flickr to find 'picturable' words by using the co-occurrence of tags provided with the images resulting from a query. ConceptNet (Speer et al., 2012) has high potential, but it has not yet shown significant improvement of performance in finding a known item (Zon, 2014).

Expert knowledge bases are not often used in complex event detection. Two examples are the Large-Scale Concept Ontology for Mulitimedia (LSCOM) that contains a lexicon of 1000 concepts describing the broadcast news videos (Naphade et al., 2006) and the multimedia ontology in soccer video domain (Bagdanov et al.,

2007). The multimedia ontology consists of an ontology defining the soccer domain, an ontology defining the video structure and a visual prototype that links both ontologies. This visual prototype aims to bridge the semantic gap by translating the values of the descriptors in an instance of the video structure ontology to the semantics in the soccer ontology. This ontology is able to detect high-level events such as *scored goal*. Natsev et al. (2007) show that in the TRECVID topic domain manual ontologies work on average better than automatic, which uses WordNet and synonymy match, and no query expansion. To our knowledge, the only expert knowledge base for complex events is used in Boer et al. (2013) and this knowledge base is not publicly available.

2.3. EXPERIMENTS

In our experiments, we compare three types of expansion methods in the field of complex event detection. The first expansion method is considered as our baseline and only uses the complex event query, which has one to four words, to detect the event. The second expansion method uses query expansion with a common knowledge base. We compare two common knowledge bases: ConceptNet 5 and Wikipedia. Both knowledge bases contain information about events, whereas many other knowledge bases only contain information about objects or facts. As opposed to our previous paper (Bouma et al., 2013b), WordNet is not used as a common knowledge base, but it is used in another way (see Section 2.3.2). The third expansion method uses query expansion with an expert knowledge base. To our knowledge, no expert knowledge base for our high-level complex events is available and we, therefore, use the textual description provided with the TRECVID Multimedia Event Detection (MED) task as expert knowledge source.

2.3.1. TASK

The open and international TRECVID benchmark aims to promote progress in the field of content-based video retrieval by providing a large video collection and uniform evaluation procedures (Smeaton et al., 2006). Its Multimedia Event Detection (MED) task was introduced in 2010. In the MED task, participants develop an automated system that determines whether an event is present in a video clip by computing the event probability for each video. The goal of the task is to assemble core detection technologies in a system that can search in videos for user-defined events (Over et al., 2013).

In this research, two sets of TRECVID MED 2014 are used. The first set is called the Research Set and contains approximately 10.000 videos, which have a text snippet describing the video. The Research Set also has ground truth data for five events. The other set is the Test Set with more than 27.000 videos and ground truth data for twenty events. For each of the twenty events in the Test Set and the five events in the Research Set a textual description containing the event name, definition, explanation and an evidential description of the scene, objects, activities and audio is used.

The standard performance measure for the MED task is the Mean Average Precision (Hauptmann et al., 2004). Performance on the official evaluation of 2013 and

2014 show that complex event detection is still a challenge. In the case with no training examples, which is the representative case for this research, Mean Average Precision is below ten procent.

2.3.2. DESIGN

This section describes the design of the experiment, which is also shown in Figure 2.1.

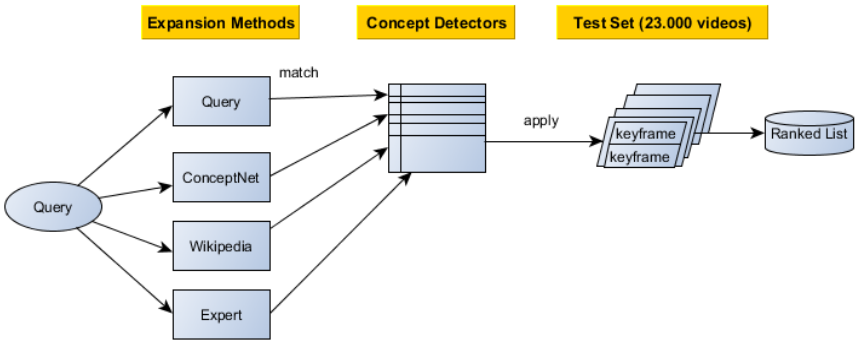


Figure 2.1: Design

In the experiments, twenty complex events, such as *dog show*, *felling a tree* and *tailgating*, are evaluated. In this evaluation, a ranked list of all videos is created using the score of a video:

$$S_{e,v,em} = \sum_{c \in CD} \left(\frac{A_{c,e,em} \cdot W_{c,em}}{\sum_{c \in CD} A_{c,e,em} \cdot W_{c,em}} \cdot \max_{k \in v} [CD_k] \right) \tag{2.1}$$

, where $S_{e,v,em}$ is the score of video v for event e in expansion method em, c is a concept, CD is the set of concept detectors, $A_{c,e,em}$ is a binary variable denoting the availability of concept c in event query e in expansion method em, $W_{c,em}$ is the weight of the concept c in expansion method em, and CD_k is the concept detector value in keyframe k.

The name of an event is used as an input for the expansion methods. Each of the expansion methods creates a list of weighted words. These weighted words are matched against the available concept detector labels. Our set of concept detectors is limited to less than 2000, so a gap between the words from the expansion methods and the concept detector labels exists. The matching step is, therefore, a filtering step. The value of $A_{c,e,em}$ is one for the selected concept detectors and zero for the concept detectors that are not selected. In this way, only the values of the selected concept detectors are considered in the score. Additionally, the sum of the weights of the expansion method is one because of the division. The following sections describe this design in further detail.

EXPANSION METHODS

Complex Event Query

The baseline method only uses the complex event query. The query is split into nouns and verbs with equal weights with a total sum of one. In the complex event query, nouns can be compound nouns such as *dog show*. If the compound noun cannot be matched against the available concept detectors (see next subsection named 'Concept Detectors'), this noun is split into the separate words, in this case *dog* and *show*. This is also shown in the following formula:

$$W_{c,ceq} = \frac{1}{\sum N_c} \tag{2.2}$$

, where N_c is the number of concepts in the query

The weight of these words is the previous weight, in this example 1.0, divided by the number of new words, which is two and, thus, results in a weight of 0.5 for *dog* and and 0.5 for *show*. Negative concepts are not taken into account, which means that the word *vehicle* is not matched against the available concept detectors in the event *winning a race without a vehicle*.

ConceptNet

ConceptNet 5 (Speer et al., 2012) is used to expand the query. Because ConceptNet contains more knowledge about objects and activities compared to events, this expansion method is used to expand the nouns and verbs in the query that have no matching concept detector label. If no label was found in the query, we search for the whole event. For example, in the event *dog show* a concept detector with the label *dog* is present in our collection of concept detectors, but no label is present for *show*. ConceptNet (version 5.3) is automatically accessed through the REST API. All words with the relation *RelatedTo*, *IsA*, *partOf*, *MemberOf*, *HasA*, *UsedFor*, *CapableOf*, *AtLocation*, *Causes*, *HasSubEvent*, *CreatedBy*, *Synonym* or *DefinedAs* to the searched word are selected. The words with a synonym relation to the searched word are also searched through the REST API. An example is shown in Figure 2.2.

The weight of the words is determined by the weight of the edge ($score_{rel}$) between the found word and the word in the complex event query. The weight is often a value between zero and thirty and is adjusted to a value that is typically between zero and one using:

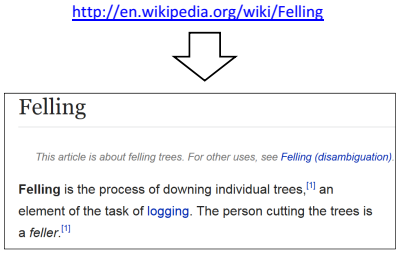
$$W_{c,cn} = (\frac{score_{rel}}{30})^3 \tag{2.3}$$

The triple power of the scoring was found by training on the five events in the Research Set. In order to deal with query drift towards the expanded word, the weighted sum of the newly found words is adjusted to the weight of the word searched for. In the event *dog show*, both *dog* and *show* have a weight of 0.5. The sum of the weights of the expanded words of *show* is, thus, 0.5. If the expanded words for *show* are *concert* (0.8), *popcorn* (0.3) and *stage* (0.5), the adjusted weights are 0.25, 0.09375 and 0.15625, respectively.

Figure 2.2: Example of ConceptNet expansion method

Wikipedia

Compared to ConceptNet, Wikipedia has more information about complex events. For each event, we automatically search for the corresponding Wikipedia page through the REST API (on October 13, 2014) and manually disambiguate to select the correct Wikipedia page. From this page all text above the table of contents, which we consider as the general definition part, is scraped. All nouns and verbs are selected using the Stanford Core NLP parser (Manning et al., 2014). The weight is calculated using TFIDF. The term frequency (*TF*) is calculated by counting the amount of times a word is present in the text (*f(t,d)*). The inverse document frequency (*IDF*) is calculated by counting the number of documents the word appears in ($\log \frac{N}{1+|\{d \in D: t \in d\}|}$). The document set is a set of 5798 Wikipedia pages (collected on July 9, 2014). These Wikipedia pages are selected by taking all nouns, verbs, combined nouns and adjective-noun pairs from the text snippets of videos in the Research set. The term frequency is multiplied with the inverse document



Word	W_wiki
Felling	$(1.0)^2$
Be	$(0.11597262497682248)^2$
Process	$(0.16042157500279314)^2$
Down	$(0.4544441499244059)^2$
Tree	$(0.5174730350807129)^2$
Element	$(0.20527196659390032)^2$
Task	$(0.3106030421086125)^2$
Log	$(0.3988084423028006)^2$
Person	$(0.18149404465480518)^2$
Cut	$(0.2218497575644838)^2$
Feller	$(0.793225159887144)^2$

Figure 2.3: Example of Wikipedia expansion method

frequency to obtain the TFIDF. The TFIDF is divided by the highest possible IDF value and squared. This squaring is added because training on the five events in the Research Set increased performance using these steps. This leads to:

$$W_{c,wiki} = \frac{f(t,d) \cdot \log \frac{N}{1+|\{d \in D: t \in d\}|}}{log \frac{N}{1}}^2 \tag{2.4}$$

, where t is the term, d is the current document, $f(t,d)$ is the frequency of t in d, D is the document set, N is the total number of documents and $|\{d \in D : t \in d\}|$ is the number of documents in which t appears. An example can be found in Figure 2.3.

Expert

The textual description provided with the TRECVID Multimedia Event Detection (MED) task is used as expert knowledge. This description consists of the event name, definition, explanation and an evidential description of the scene, objects, activities and audio. An example of a description is shown in Figure 2.4. From all different parts the nouns and verbs are manually extracted. The Stanford Core NLP parser (Manning et al., 2014) is not used, because the text also contained separate words instead of sentences. This causes problems for the parser. An addition to the selection of these words is that words within brackets or enumerations with an *or* were clustered. This cluster is used to indicate that only one of these concepts has to be available in the video. In texts in which a noun or verb is placed before or after a

negation, such as .. *is not considered positive for this event* and *without a vehicle*, are not taken into account. The determination of the weight is equal to the weight determination in the Wikipedia expansion method (W_{wiki}). From the clustered word the compound noun is chosen for determination of term frequency and inverse document frequency. In case no compound word was present the first word is chosen. An example is shown in Figure 2.4.

Event name	Felling a tree
Definition	One or more people fell a tree
Explication	Felling is the process of cutting down an individual tree
	transforming its position from vertical to horizontal. Felling a tree
	can be done by hand or with a motorized machine. If done by hand, it
	usually involves a tool such as a saw, chainsaw, or axe. A
	tree-felling machine, known as a feller buncher, can also be
	used. Felling is part of the logging process, but can also be done to
	single trees in non-logging contexts. possibly climbing the tree or
	accessing upper parts of the tree from a cherry-picker bucket and then
	cutting branches from the tree before felling it, possibly cutting a
	horizontal wedge from the tree's trunk to cause the tree to fall in a
	desired direction, cutting horizontally through the trunk of the tree
	with saw(s) or ax(es), using wedges or rope(s) to prevent the tree
	from falling in some particular direction (such as onto a house)
Evidential description	
scene	outdoors, with one or more trees
objects/people	persons in work clothing, hand saws or chain saws, axes, metal wedges,
	tree-felling machines
activities	sawing, chopping, operating tree felling machine
audio	chainsaw motor, sounds of chopping, sawing, tree falling

Figure 2.4: Example of a textual description in TRECVID MED

CONCEPT DETECTORS

The list of weighted words from the methods is matched to a set of 1818 concept detector labels, which are (compound) nouns or verbs. This comparison is done in two ways. The first way is to compare the word to the concept detector label. The exact matches are selected. The words without an exact match are compared using WordNet (Miller, 1995). Both the word and the concept detectors are compared in WordNet using the Wu-Palmer (WUP) similarity (Wu et al., 1994). Concept detectors with a similarity of 1.0 to a word are selected, which means that both point to the same synset, such as with *foot* and *feet* or *project* and *task*. The only two exceptions are *fight* for *engagement* and *hide* for *fell*. These matches are not taken into account. The selected concept detectors get the weight of the words. If multiple words point to the same concept detector, such as with synonyms, the weights are added. If one word points to multiple concept detectors, such as *dog* from one collection within the set and *dogs* from another collection, the weight is equally divided over the concept detectors. At the end of the matching process the weight of a concept detector is divided by the total amount of weights in order to create a sum of the weights equal to 1.0.

The set of 1818 concept detectors consists of three collections. The first collection consists of 1000 concept detectors and is trained on a subset of the ImageNet dataset

with 1.26 million training images as used in ILSVRC-2012 (Deng et al., 2009). The second collection has 346 concept detectors, which are trained on the dataset from the TRECVID Semantic Indexing task of 2014. The final collection contains 472 concept detectors and is trained and gathered from the Research Set of TRECVID MED (Ngo et al., 2014). The last collection originally contained 497 concept detectors, but the detectors directly trained on the high-level events are removed. In this way we can test the elements in the query and query expansion instead of just the (rather good) accuracy of these concept detectors. More details on the concept detectors can be found in Ngo et al. (2014).

VIDEOS

The test set of TRECVID MED 2014 consists of more than 27.000 videos. From each video one keyframe per two seconds is extracted. Each concept detector is applied to each keyframe. As a result for each keyframe for each concept detector a value between zero and one is available, which represents the confidence score. The highest confidence score over all keyframes for one video is selected. This score is multiplied by the weight of the concept detector, which was originally coming from the methods. The weighted sum of the concept detector values represents an estimation of the presence of the event in the video. This estimation is used to rank all videos and place the videos in a list in descending order.

2.4. RESULTS

Results on the Test Set of TRECVID MED 2014 for each of the twenty events are split up in four tables: Table 2.1, 2.2, 2.3 and 2.4. Bold digits indicate the highest performance in the row and italic digits indicate random performance. Table 2.1 shows average performance of the events in which all nouns and verbs in the event query have matching concept detectors. ConceptNet is only used to expand words that could not be matched and, thus, no average performance is available in Table 2.1 for ConceptNet. Wikipedia has no performance if no page containing the event could be found. Table 2.2 contains performance of the events in which the main noun of the event query, which is the second noun in compound words, is matched to a concept detector. If no additional words could be found by ConceptNet, performance of ConceptNet is equal to performance of the query. Table 2.3 shows performance of the events in which at least one word in the event query (not the main noun) could be matched to a concept detector. Table 2.4 contains information about events in which no word in the event query could be matched to a concept detector. The Mean Average Precision on all twenty events is 0.03865, 0.06143 (0.06220 without same as Query and 0.03047 without *beekeeping*), 0.03024 (0.02042 with random) and 0.03262 for the Query method, ConceptNet, Wikipedia and Expert knowledge, respectively.

Table 2.1: Average Precision: matching query

Event Name	Query	ConceptNet	Wikipedia	Expert
cleaning appliance	**0.10228**			0.01055
town hall meeting	**0.03800**		0.01568	0.00866
rock climbing	**0.13932**		0.01936	0.01957
fixing musical instrument	0.04245			**0.04954**
MEAN	**0.08051**		0.01752	0.02208

Table 2.2: Average Precision: one matching main noun in query

Event Name	Query	ConceptNet	Wikipedia	Expert
non-motorized vehicle repair	0.02016	0.02016		**0.02915**
renovating home	**0.01568**	**0.01568**		0.01261
winning race	**0.04048**	0.01228	0.01181	0.00695
felling tree	**0.04057**	0.01145	0.01461	0.00656
parking vehicle	**0.10675**	0.00321	0.00390	0.00404
tuning musical instrument	0.01496	0.02436	0.02235	**0.05572**
MEAN	**0.03977**	0.01452	0.01317	0.01917

Table 2.3: Average Precision: one match (not main noun) in query

Event Name	Query	ConceptNet	Wikipedia	Expert
attempting bike trick	0.07117	0.02361		**0.07486**
working metal craft project	**0.04621**	0.00336		0.03865
horse riding competition	0.07655	0.02766	**0.11451**	0.01017
playing fetch	0.00264	**0.01519**	0.00936	0.00275
dog show	0.00901	0.05339	0.00943	**0.05362**
MEAN	0.04111	0.02464	**0.04443**	0.03601

Table 2.4: Average Precision: no matching word in query

Event Name	Query	ConceptNet	Wikipedia	Expert
giving direction location	*0.00095*	**0.00324**		0.00321
marriage proposal	*0.00219*	0.00203	0.00324	**0.00414**
beekeeping	*0.00116*	**0.64970**	0.15346	0.23404
wedding shower	*0.00121*	**0.03929**	0.01301	0.02594
tailgating	*0.00133*	0.00199	**0.00244**	0.00169
MEAN	0.00137	**0.13925**	0.04304	0.05380

2.4.1. QUERY EXPANSION VS. NO QUERY EXPANSION

Comparing average performance of our baseline, which is presented as *Query* in the tables, to each of the other columns shows that query expansion does not always improve performance. Mean Average Precision on all twenty events show the highest value for the method in which no query expansion is used (ConceptNet without *beekeeping*). Table 2.1 shows that if all nouns and verbs in the query could be matched to a concept detector, average performance is highest for the query. The events *town hall meeting* and *rock climbing* have significantly higher performance for the query compared to the expansion methods. Table 2.2 shows the same trend as Table 2.1, but the exception is *tuning musical instrument*. Table 2.3 shows a mixed performance and in Table 2.4 performance of the baseline is random and, thus, query expansion methods perform better.

2.4.2. EXPERT KNOWLEDGE VS. COMMON KNOWLEDGE

The average results regarding common knowledge (ConceptNet without *beekeeping*) and expert knowledge show no clear preference for either method. Comparing the separate results, the performance using expert knowledge is clearly higher in the events *non-motorized vehicle repair*, *tuning musical instrument*, *attempting bike trick* and *working metal craft project*. For the other fourteen events, the common knowledge bases perform equally good or better than expert knowledge.

2.4.3. CONCEPTNET VS. WIKIPEDIA

Common knowledge bases ConceptNet (without *beekeeping*) and Wikipedia have comparable Mean Average Precision values. Wikipedia has a higher average precision in Table 2.3 and ConceptNet has a higher average precision in Table 2.4. Comparing the different events in Table 2.3 and 2.4, Wikipedia performs better than ConceptNet in *horse riding competition*, *marriage proposal* and *tailgating*.

2.4.4. LATE FUSION

In this section, we present the result of late fusion, because we expect that late fusion will help to exploit complementary information provided in the different expansion methods. In late fusion, the scores of the videos ($S_{e,v,em}$, see Equation 2.1) of the different expansion methods are combined using four different fusion techniques.

The first fusion technique is the arithmetic mean in which the fused score is calculated by:

$$Fa_{e,v} = \frac{1}{EM} \sum_{em \in EM} S_{e,v,em} \tag{2.5}$$

, where $Fa_{e,v}$ is the fused score for video v and event e, EM is the set of expansion methods and $S_{e,v,em}$ is the score for video v and event e in expansion method em

The geometric mean is used as a second fusion technique:

$$Fg_{e,v} = \prod_{em \in EM} S_{e,v,em} \tag{2.6}$$

, where $Fg_{e,v}$ is the fused score for video v and event e, EM is the set of expansion methods and $S_{e,v,em}$ is the score for video v and event e in expansion method em

As a third fusion technique, the highest value for a video is taken:

$$Fm_{e,v} = \max_{em \in EM} S_{e,v,em} \tag{2.7}$$

, where $Fm_{e,v}$ is the fused score for video v and event e, EM is the set of expansion methods and $S_{e,v,em}$ is the score for video v and event e in expansion method em

The last fusion technique is a weighted mean, in which

$$Fw_{e,v} = \frac{1}{\sum_{em \in EM} W_{em}} \cdot \sum_{em \in EM} (W_{em} \cdot S_{e,v,em}) \tag{2.8}$$

, where $Fw_{e,v}$ is the fused score for video v and event e, EM is the set of methods, W_{em} is the weight for expansion method em and $S_{e,v,em}$ is the score for video v and event e in expansion method em

Table 2.5: Mean Average Precision with Late Fusion

Fused Part	MAP before fusion	MAP after fusion	MAP increase (%)
Per event	0.08103	0.09198	13.5
Events Table 2.1	0.08051	0.08051	0.0
Events Table 2.2	0.03977	0.04030	1.3
Events Table 2.3	0.05014	0.06130	22.3
Events Table 2.4	0.13925	0.13925	0.0

The fusion score of each combination of two, three and four expansion methods are calculated. In the weighted mean, both values 0.25 and 0.75 are examined as W_{em} for the expansion methods. The results of the fusion optimized per event and optimized per part is shown in Table 2.5. Results show that Mean Average Precision optimized per event improves from 0.08103 to 0.09198 (+13.5%) with fusion. Because we are working with the zero-example case, this is our upper boundary. Mean Average Precision optimized per part increases from 0.07538 to 0.07833 (+ 3.9%) overall. In the column *MAP before fusion* in Table 2.5, the Mean Average Precision of the query method is used for the events of Table 2.1 and 2.2. Wikipedia is used for the events in Table 2.3 and if Wikipedia has no result, the query method is used. The results on ConceptNet are used for the events of Table 2.4. In the fusion of these parts, no single fusion method could outperform the query in complete matched query (events from Table 2.1) and ConceptNet in the events from Table 2.4. For the matching main nouns (Table 2.2) a fusion with the maximum of the query, Wikipedia and the Expert provides highest performance. This fusion method improves 22.7% on the event *tuning musical instrument* and less than 1.0% on the other events. In the matching without the main nouns (Table 2.3) a weighted mean of the query (0.25), Wikipedia (0.75) and the Expert (0.25) provides highest performance. This fusion method improves on the events *attempting bike trick* (7.8%), *working metal crafts project* (136.5%) and *dog show* (46.4%).

2.5. DISCUSSION, CONCLUSION AND FUTURE WORK

In our experiments, the Test Set of TRECVID 2014 Multimedia Event Detection (MED) task (Over et al., 2013) was used to compare the effectiveness of our retrieval system for complex event queries. We compared ConceptNet 5 (Speer et al., 2012) and Wikipedia as common knowledge bases and the textual description provided with the TRECVID task to determine which type of knowledge base is best to use.

Results comparing the baseline with the query expansion methods show that the complex event query not necessarily performs worse than methods using query expansion. These results, however, do not imply that knowledge bases should not be used. It is important to know in which cases a knowledge base can add information and in which cases the complex event query is enough. The results clearly show that if all query terms are found, additional information does not improve performance. This is also the case in most of the events in which the main noun is found. On the other hand, query expansion is beneficial to use in the other events, which confirms our expectations. This brings us to the first conclusion: *1) Query Expansion can improve performance compared to using no query expansion in the case that the main noun of the query could not be matched to a concept detector.*

A result that does not meet our expectations is that query expansion using expert knowledge is not better than query expansion using common knowledge bases. In the events in which no word could be matched to the query, the expert only performs best in *marriage proposal*, whereas the common knowledge bases perform best in the other four events. In the events in which one match in the query is found, expert knowledge and common knowledge both perform best in two of the five events. The second conclusion, therefore, is: *2) Query expansion using expert knowledge is not necessarily better than query expansion using common knowledge.*

Another interesting result is in the comparison of ConceptNet and Wikipedia. The results in Table 2.3 and 2.4 show that Wikipedia only performs better than ConceptNet in *horse riding competition*, *marriage proposal* and *tailgating*. In *horse riding competition*, ConceptNet is used to search for *competition*. This word is general and, therefore, more general words for competitions are found. In Wikipedia, *horse riding competition* is used and one of the key words for the event (*vault*) is found. In *marriage proposal*, ConceptNet has less information than Wikipedia and, therefore, Wikipedia has better performance. In *tailgating*, ConceptNet and Wikipedia contain complementary information. Wikipedia has more information on sports and food, while ConceptNet has more information about the car. Two events in which ConceptNet clearly outperforms all other methods are *beekeeping* and *wedding shower*. Wikipedia and Expert both find *bee* and *apiary*, but other concepts suppress the weight of *apiary*, which decreases performance. In *wedding shower*, the same problem occurs. The concept *party* seems to provide the best information and a low weight of this concept decreases performance. Weighting is, thus, an important part in the expansion methods. In general, we can conclude that, in this configuration, *3) ConceptNet performs slightly better than Wikipedia.*

The last result section shows the results of late fusion. With the twenty events, it is not yet clear which fusion method performs best in which cases. Several events show

highest performance using geometric mean, but in the separation of the parts the geometric mean does not have highest performance over a part. Furthermore, some fusion methods improve performance in one event, but decrease performance drastically in other events. For Table 2.2 the best method per part is a weighted mean. In the events of Table 2.2, *horse riding competition* has a high performance in the Wikipedia method. In order to not lose this result in the mean, Wikipedia has a weight of 0.75 and the query and expert have a weight of 0.25. ConceptNet, apparently, provides no complementary information and is, therefore, not increasing performance. For Table 2.3 the best fusion method is an arithmetic mean. The event *working metal crafts projects*, for example, has information in expert about a workshop and kinds of tools and the query has *metal*. Adding this information provides slightly better performance compared to taking a product or taking the maximum. In general, we can conclude that: *4) Late fusion can slightly improve performance*.

To conclude, query expansion is beneficial, especially in events of which the main noun of the query could not be matched to a concept detector. Common knowledge bases do not always perform worse than expert knowledge, which provides options for automatic query expansion from the Web in complex event detection.

The experiments conducted in this chapter have some limitations. First, research is only conducted on twenty complex events, which is a very small set. The conclusions on the comparison between the common knowledge bases can, therefore, be different in a larger or different set of complex events. In a larger set of complex events the specific situations in which any of the methods is preferred over the others can be determined in a better way. Second, less than 2000 concept detectors are used. Many words in the query and, especially, the query expansion methods could not be matched to concept detectors. Third, the weight determination of ConceptNet, Wikipedia and the expert expansion method is trained on the Research Set with only five events. This number of events is not enough to train on and the weighting is, therefore, not optimal. Fourth, the fusion methods as well as the weights in the weighted mean are not fully explored.

In the future, we want to compare the kind of information available in the expert knowledge and in common knowledge in order to determine what kind of information provides the increase in performance in complex event detection. This can be combined with the further exploration of fusion methods. Other common knowledge bases, such as YAGO2 and Flickr, are possibly worth integrating in our system. Another interesting option is to examine the use of (pseudo-)relevance feedback. This feedback can also be combined with, for example, common knowledge sources.

3

SEMANTIC REASONING IN ZERO EXAMPLE VIDEO EVENT RETRIEVAL

Edited from: **Maaike de Boer Yi-Jie Lu, Hao Zhang, Klamer Schutte,Chong-Wah Ngo and Wessel Kraaij** (2017) *Semantic Reasoning in Zero Example Video Event Retrieval.* In: Transactions on Multimedia Computing, Communications, and Applications, DOI:10.1145/3131288.

Apart from knowledge bases that are used in chapter 2, semantic embeddings can be used to map the query to a set of concepts. One of the recently introduced semantic embedding methods is word2vec. This method uses large text corpora to create a representation in which words that often occur in the same context have a small distance. This representation makes it possible to find the concepts that are closest to a user query. Instead of comparing the concepts separately to the query, we propose a method to create a set of concepts that is jointly closest to a user query. In our experiments on the TRECVID MED test set of 2014, we show that the method has a better performance in terms of Mean Average Precision, and it is more robust to query drift and cut-off parameter tuning compared to a method that uses the top n concepts that are closest to the user query.

Additionally, we experiment with different Concept Banks / vocabularies. We focus on the complexity of the concepts, in which we use a separation in low-, mid- and high-level concepts. Low-level concepts are basic components of an image, such as objects or scenes. Mid-level concepts are basic actions, activities or interactions, for example running or swimming. High-level concepts are complex activities that include interactions between people and/or objects, such as a birthday party or a robbery. Whereas in chapter 2 we only used low- and mid-level concepts, in this chapter we include high-level concepts. We show that high-level concepts are important in a task that involves high-level events. This chapter is related to **RQ2 word2vec**.

3.1. INTRODUCTION

The domain of content-based video information retrieval has gradually evolved in the previous 20 years. It started as a discipline mostly relying on textual and spoken information in news videos, and moved towards richer multimedia analysis leveraging video, audio and text modalities. The last 10-15 years have shown impressive progress in image classification, yielding larger and larger concept vocabularies. In 2011, the TRECVID MED task defined a testbed for even deeper machine understanding of digital video by creating a challenge to detect high level or complex events, defined as “long-term spatially and temporally dynamic object interactions" (Jiang et al., 2012). Examples of high-level events are social events (*tailgating party*) and procedural events (*cleaning an appliance*) (Jiang et al., 2012). Given the extreme difficulty of the MED task, in early years of TRECVID system development was facilitated by providing a set of example videos for the event, making this essentially a supervised video classification task. In the last few years, the MED task has stepped up towards its real challenge: retrieving relevant video clips given —only —a precise textual description of a complex event. In TRECVID MED context, this task is referred to as the zero example case, since no visual examples are provided (Over et al., 2015). The problem of detecting multimedia events is different from the TRECVID datasets from 2005 to 2008 (Kennedy et al., 2006; Smeaton et al., 2006). The TRECVID MED events contain complex and generic high-level events, such as *winning a race without a vehicle*. This query is generic because it is referring to a wide variety of races, including running, swimming, jumping and crawling. The query is also significantly more complex than the entity-based queries, e.g. *emergency vehicle in motion*, used in multimedia research ten years ago, because the number of relevant concepts is higher and the relationship between the concepts plays an important role. Not only should the awareness of a race be captured, but also the winning of a race and the absence of a vehicle in the race (although vehicles could be present on the parking lot near the race or at the side of the street in a marathon).

In our paper, we describe the challenges of building an effective system for zero example complex event retrieval in video. The main issue in zero example video event retrieval is that state of the art machine learning techniques cannot be used, because no training examples are available. A common approach is to use a set of pre-trained classifiers and try to map the event to a set of classifiers. Within this approach two challenges exist: what set of pre-trained classifiers to use (*Vocabulary challenge*) and how to map the event to a set of classifiers (*Concept Selection challenge*).

The Vocabulary Challenge deals with the determination of a good set of concepts to pre-train and put in the vocabulary. This vocabulary is built with pre-trained concept detectors on off-the-shelf datasets. Whereas five to ten years ago fewer than a 1000 pre-trained concepts were available, previous work (Hauptmann et al., 2007a; Aly et al., 2012) was focused on simulations to show how many concepts are actually needed to achieve a reasonable performance. Currently, many datasets with a large vocabulary of pre-trained concepts (Deng et al., 2009; Karpathy et al., 2014; Zhou et al., 2014; Jiang et al., 2017) are available and we can therefore use actual pre-trained concepts in real datasets instead of simulations. Not all concepts are, however, nec-

essary or useful for a certain test case. For example, the ImageNet dataset (Deng et al., 2009) contains many classes of dog breeds. These concepts are not useful in test cases that only include people and scenes. This implies that it is crucial to at least pre-train and apply those concepts that are valuable for the unseen test case. Some recommendations on how to build a good vocabulary are already available (Habibian et al., 2013).

In this chapter, we show the importance of high-level concepts, defined as "complex activities that involve people interacting with other people and/or objects under certain scene" (Chen et al., 2014), because a combination of objects and actions often cannot capture the full semantics of a high-level event. We do not claim that we are the first to use high-level concepts, but we show the difference in performance for different types of concepts.

The Concept Selection challenge embeds the problem that the system has no prior knowledge about the events, so in many cases no precise visual concept detectors are available. Commonly, this challenge is approached by mapping the event to a set of classifiers by optimizing the match between the User Query (*UQ*) and the System Query (*SQ*). Within the TRECVID community, this is also referred to as Semantic Query Generation (Over et al., 2015). Here the User Query is a textual description of the event and the System Query is a combination of concepts present in our vocabulary. In this chapter, we will refer to the term *concept* as the label or name of the concept itself and to *concept detectors* as pre-trained classifiers. In this challenge, we build upon the existing word2vec models (Mikolov et al., 2013; Pennington et al., 2014), which use semantic embeddings. The main novelty of our method is that it accurately selects the proper concepts without the problem of query drift, in which the selected concepts create a drift towards one facet of the query (Carpineto et al., 2012).

The main contributions of this chapter can be summarized as follows:

- We show the importance of high-level concepts in defining a good vocabulary of pre-trained concept classifiers in the case of search queries that contain high-level events.
- We introduce an incremental word2vec method (*i-w2v*) for concept selection that is more robust to query drift and cut-off parameter tuning.

The next section contains related work. We focus on our two challenges. The third section explains our Semantic Event Retrieval System that includes our novelties in both challenges. The fourth section presents the experiments conducted on the international benchmark TRECVID Multimedia Event Detection (Over et al., 2015) and the results are included in the fifth section. The sixth section contains a discussion and the final section provides the conclusion.

3.2. RELATED WORK

In this section we only focus on the Vocabulary challenge and the Concept Selection challenge in zero example video event retrieval.

3.2.1. VOCABULARY

Concept vocabularies are designed as a representation layer for a specific purpose, such as indexing descriptors for video clips, shots or frames. Ideally, concept vocabularies consist of unambiguous precise descriptors of entities, activities, scenes, objects and ideas. Different vocabularies are developed for different purposes. Combining different vocabularies often results in vagueness and ambiguity, such as polysemy and homonymy. We will focus on two properties of concepts: *level of complexity* and *level of granularity*. In the level of complexity, three levels can be differentiated. First, *low-level* concepts are the basic components in images or videos, such as objects. Second, *mid-level* concepts are basic actions, activities or interactions. Actions or activities are a "sequence of movements" (Chen et al., 2014) and can be performed by one entity, such as people or objects. Interactions are actions between two or more entities. Third, *high-level* concepts are "complex activities that involve people interacting with other people and/or objects under certain scene" (Chen et al., 2014). The key difference between mid-level and high-level concepts is that a high-level concept contains multiple actions and interactions evolving over time (Chen et al., 2014), such as the difference between the action *horse riding* and the event *horse riding competition*. Furthermore, concepts can have different levels of granularity, also referred to as specificity. Examples are animal (*general*), dog and chihuhua (*specific*).

The importance of the level of granularity in a vocabulary was already indicated by Hauptmann et al. (2007a) and Habibian et al. (2013). Both argue that in video event recognition a mixture of both general and specific concepts achieves higher performance compared to using only general or specific concepts. Interestingly, both papers state that the general concepts achieve in general higher performance compared to the specific concepts, because specific concepts only occur in a few videos, and many general concepts can be distinctive enough to recognize an event. The importance of the level of complexity is not yet introduced, but Habibian et al. (2013) recommend to use a vocabulary that contains concepts of the following categories: object, action, scene, people, animal and attribute. Using our definitions an action is comparable to a mid-level concept and the concepts from the other categories are low-level concepts. Another work of these authors introduces primary concepts and bi-concepts (Habibian et al., 2014a).

Other recommendations from Habibian et al. (2013) are 1) use a vocabulary with at least 200 concepts; and 2) do not use too many concepts of one type, such as animals or people. Additionally, they argue that it is better to include more concepts than to improve the quality of the individual concepts, which is also concluded by Jiang et al. (2015). Previous research of Aly et al. (2012) indicated that few concepts (100) with a simulated detector performance of only 60% is already sufficient to achieve reasonable Mean Average Precision performance (20%). Hauptmann et al. (2007b) argue that 3000 concepts are needed for a Mean Average Precision of 65%. We follow this recommendation and focus on extending the vocabulary instead of

improving performance of concept detectors.

In addition to the type of concepts, Jiang et al. (2015) report the influence of training with different datasets on performance for the events in the TRECVID Multimedia Event Detection task. The dataset with the highest performance is Sports (Karpathy et al., 2014), followed in descending order by the 1000 concepts from ImageNet (Deng et al., 2009), the Internet Archive Created Commons (IACC) dataset (Over et al., 2014), the big Yahoo Flickr Creative Common dataset (YFCC100M) (Thomee et al., 2015) and the Do It Yourself (DIY) dataset (Yu et al., 2014). We use the concepts of their top two performing datasets in our vocabulary. Furthermore, one of their recommendations is to train concept detectors on large datasets, both in terms of training examples as well as number of concepts. We take this recommendation into account and focus on large datasets.

3.2.2. CONCEPT SELECTION

Many different techniques are used in Concept Selection. Liu et al. (2007) present five categories in which concepts can be selected, of which we use three as a guideline to give an overview of the different methods used in the recent years. The first category is making use of an ontology. These ontologies or knowledge bases can be created by expert (*expert knowledge base*) or created by the public (*common knowledge base*). Expert knowledge bases provide good performance, but dedicated expert effort is needed in the creation of such a knowledge base. Some early work on expert knowledge bases and reasoning in the field of event recognition is explained in Ballan et al. (2011). One current expert ontology for events is EventNet (Ye et al., 2015). Common knowledge bases, such as Wikipedia (Milne et al., 2013) and WordNet (Miller, 1995), are freely available and often used in the video event retrieval community (Neo et al., 2006; Yan et al., 2015; Tzelepis et al., 2016), but might not contain the specific information that is needed. A comparison of performance between an expert knowledge base and two common knowledge bases, which are Wikipedia and ConceptNet, is given in Boer et al. (2015b). Concept selection in common knowledge bases is often done by using the most similar or related concepts to events found in the knowledge base. An overview of the type of methods to find similar or related concepts can be found in Natsev et al. (2007). The number of selected concepts and the similarity measures used differ per paper and no conclusive result on which method works best is found.

The second category is making use of machine learning techniques. Machine learning techniques can be used to automatically select the proper concepts. These techniques are used more often in tasks with example videos, because many models need training examples. In the zero example video event retrieval, graphical models such as hidden Markov models (Dalton et al., 2013) are used. More often statistical methods are used, such as co-occurrence statistics (Mensink et al., 2014) and a skip-gram model (Chang et al., 2015). One group of current state of the art models is word2vec, which produce semantic embeddings. These models either use skip-grams or continuous bag of words (CBOW) to create neural word embeddings using a shallow neural network that is trained on a huge dataset, such as Wikipedia, Gigawords, Google News or Twitter. Each word vector is trained to maximize the log prob-

ability of neighboring words, resulting in a good performance in associations, such as *king - man + woman = queen*. Two often used models are the skip-gram model with negative sampling (SGNS) (Mikolov et al., 2013), which has relations to the pointwise mutual information (Levy et al., 2014), and the Glove model (Pennington et al., 2014), which uses a factorization of the log-count matrix. Although Pennington et al. (2014) claimed to have performance superior to SGNS, this is called into question by Levy et al. (2015) and Goldberg[1]. The advantage of word2vec over other semantic embedding methods is that the latent variables are transparent, because the words are represented in vector space with only a few hundred dimensions. Examples of other semantic embedding methods are Wu et al. (2014) with their common lexicon layer, Habibian et al. (2014b) with VideoStory and Jain et al. (2015) with the embedding of text, actions and objects to classify actions.

The third category is making use of relevance feedback. User clicks or explicit relevance judgements from users can be used to optimize the results. A review of relevance feedback in content based image retrieval can be found in Patil et al. (2011). In concept selection using relevance feedback often an initial set of concepts is chosen using the ontology, machine learning techniques or one of the other techniques and a user is asked to remove the irrelevant concepts and/or to adjust the importance of concepts (Jiang et al., 2015; Chang et al., 2015). A second option is to refine the text query instead of removing concepts (Xu et al., 2015). A third option is to use weakly labelled data (Chang et al., 2016) to dynamically change the weights of the selected concepts. Besides user interaction, pseudo-relevance feedback can be used. In pseudo-relevance feedback we assume that the top videos are relevant for the query (Jiang et al., 2014a; Jiang et al., 2014b). Although this method by the CMU team has top performance in TRECVID MED 2014, pseudo-relevance feedback is a high risk for rare events. In our experiments, we focus on the first run of the video event retrieval system and, therefore, do not include pseudo-relevance feedback. We, however, compare our method with a method that uses a user to create the System Query.

In addition to the different categories from Liu et al. (2007), Jiang et al. (2015) found that a sensible strategy for concept selection might be to incorporate more relevant concepts with a reasonable quality. They state that automatic query generation or concept selection is still very challenging and combining different mapping algorithms and applying manual examination might be the best strategy so far. Huurnink et al. (2008) propose a method to asses the automatic concept selection methods and compare that method to a human assessment. Mazloom et al. (2013b) show that an informative subset of the vocabulary can achieve higher performance compared to just using all concepts of the vocabulary in a setting of video event retrieval with examples. This strategy is also used in our previous work (Lu et al., 2016b) that uses evidential pooling of the concepts in the video.

[1] On the importance of comparing apples to apples: a case study using the GloVe model, Yoav Goldberg, 10 August 2014

3.3. SEMANTIC EVENT RETRIEVAL SYSTEM

In our Semantic Event Retrieval System we use five large external datasets to form our vocabulary, which is explained in the following subsection. Our vocabulary is used in our concept selection method to transform the user query (*UQ*) into a System Query (*SQ*), as explained in the second subsection. UQ is a fixed textual description of an event, for which we only use the name of the event. SQ is a list of concepts (c) and their associated similarities (c_s). The constraints on our SQ are: *sparsity*, *non-negativity* and *linear weighted sum*. Regarding sparsity, we use an informative subset of concepts, as recommended by Mazloom et al. (2013b) and similar to our previous findings, resulting in a sparse set of concepts in SQ. No negative similarities are used, because in our findings this decreases performance. For example, in the event *winning a race without a vehicle* using a negative similarity for the concept *vehicle* decreases performance, because in some videos of this event a parking lot with vehicles is present at the beginning of the video. The linear weighted sum is used to combine the concepts in our SQ to create the event score for a certain video ($S_{e,v}$). The concept detector score per video ($c_{d,v}$) is the concept detector score (d) belonging to a video (v).

The formula to create the event score is shown in Equation 3.1.

$$S_{e,v} = \sum_{c \in SQ} c_s \cdot c_{d,v}, \tag{3.1}$$

, where c is the concept, V is the vocabulary, c_s is the similarity of concept c, $c_{d,v}$ is the concept detector score for concept c over video v. The event scores can be used to order the videos and calculate performance.

3.3.1. VOCABULARY

While creating the vocabulary, we follow the recommendations of Habibian et al. (2013), which are to use a large and diverse vocabulary, and use the top two performing datasets from Jiang et al. (2015), i.e. Sport and ImageNet. Furthermore, we aim for a set of datasets that not only contains low- and mid-level concepts, but also high-level concepts. Figure 3.1 shows our interpretation of the different datasets on the level of complexity.

The two low-level datasets are ImgNet (Deng et al., 2009) and Places (Zhou et al., 2014). ImgNet, which is an abbreviation for ImageNet, contains low-level objects and for our vocabulary the standard subset of 1000 objects is used. The Places dataset does not contain objects, but scenes or places. We have one dataset that contains both low- and mid-level concepts: SIN (Over et al., 2015). These concepts have been developed for the TRECVID Semantic Indexing Task of 2015. We also included one dataset that contains both mid-level and high-level concepts: Sport (Karpathy et al., 2014). This is a dataset that contains one million sports videos, classified into 487 categories. Our high-level dataset is the Fudan Columbia Video dataset (Jiang et al., 2017), which contains 239 classes within eleven high-level groups, such as art and cooking&health.

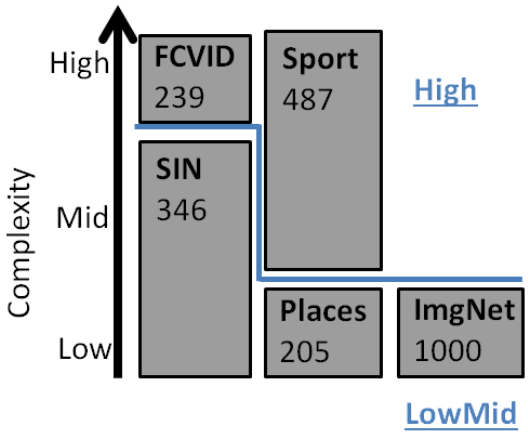


Figure 3.1: The level of complexity for the five datasets used in this chapter
The number under each dataset indicates the number of concepts in the dataset.

Table 3.1: Overview Datasets

Name	**#Concepts**	**Structure**	**Dataset**
FCVID	239	DCNN+SVM	Fudan-Columbia (Jiang et al., 2017)
SIN	346	DCNN	TRECVID SIN (Over et al., 2015)
Sport	487	3D-CNN (Tran et al., 2015)	Sports-1M (Karpathy et al., 2014)
Places	205	DCNN	MIT Places (Zhou et al., 2014)
ImgNet	1000	DCNN (Krizhevsky et al., 2012)	ImageNet (Deng et al., 2009)

Table 3.1 shows additional information on the datasets, such as the number of concepts, the reference to the publication of the dataset and the structure used to train the concept detectors. Training of the concepts is done by using one of the states of the art Deep Convolutional Neural Network (DCNN) architectures. The original DCNN architecture of Krizhevsky et al. (2012), named AlexNet, is used for ImgNet. The output of the eighth layer of the DCNN network trained on the ILSVRC-2012 (Deng et al., 2009) is used as concept detector score per keyframe. This DCNN architecture is fine-tuned for both SIN and Places. The concept detector scores per keyframe are max pooled to obtain the score per video. The keyframes are extracted at the rate of one keyframe per two seconds.

The two high-level datasets are annotated on video level instead of keyframe level and are, therefore, trained in a slightly different way. FCVID also uses the same DCNN architecture, but the seventh layer of the network is used as an input for an SVM. This SVM is trained on the videos within the dataset on video level instead of keyframe level. The Sport dataset is trained with the 3D CNN network of Tran et al. (2015).

3.3.2. CONCEPT SELECTION (I-W2V)

Our incremental word2vec method (*i-w2v*) starts with a vector containing the words in the User Query (UQ). In our experiments, the UQ is the name of an event, such as [‘parking’, ‘vehicle’]. On the other hand, we have a vocabulary with concepts. These concepts can also be represented as a vector, such as the concept [‘police’, ‘car’]. In the function *sim(c,UQ)*, we use the Gensim code[2], which is an implementation of the SGNS model (Mikolov et al., 2013), to calculate the cosine similarity between UQ and each of the concepts in the vocabulary. This similarity is stored in c_s. We sort the concepts in the vocabulary based on this similarity. We discard the concepts with a similarity less than 80% of the highest similarity. This cut-off is used to decrease the possibility of introducing noise. Subsequently, we try whether a combination of concepts will increase the similarity to take care of the query drift. Where other methods might choose the top five as the selected concepts, we only use the concepts that increase the similarity. In the multidimensional word2vec space, one facet might have a vector into one direction towards UQ, whereas another facet might have a vector into another direction. Using both concepts will move the vector more towards the vector of UQ and increase the cosine similarity. We start with using the concept with the highest similarity in a concept vector. We iteratively add concepts (in order of their similarity) to this concept vector and each time compare the cosine similarity of the new vector to UQ. If the similarity is higher with the concept than without, we retain the concept in the concept vector. In the case of the event *parking a vehicle*, the first concept is *vehicle*. All types of vehicle, such as *police car* or *crane vehicle* are not added to the concept list as the concept list with the police car added, such as [‘vehicle’, ‘police’, ‘car’] does not increase the cosine similarity to UQ. The concept *parking lot*, which was not in the top five concepts, is included, because the facet *vehicle* and the facet *parking (lot)* together increase the similarity to the event *parking a vehicle*. Similarly, the tenth concept *parking meter* is not included as it covers the same facet as *parking lot*. The output of the Concept Selection method is the list of selected concepts and their original cosine similarity c_s to UQ. This concept selection method has a complexity of $O(n)$ in which n is the number of concepts, because we have to calculate the similarity between the query and each of the concepts. This method is faster than look-up time of the video in the database, which makes it applicable for real-time systems.

Table 3.2 shows that our method is robust to a range of cut-offs, both percentages and a fixed similarity threshold of 0.1, on the vocabulary using pre-trained concepts from all datasets mentioned in the previous section (referred to as the All vocabulary). The average number of concepts remaining after applying our algorithm is also included in Table 3.2. The novelty in our method is to only add the concepts that improve the similarity to the full event. To our knowledge, current word2vec models did not yet look into solutions to a possible query drift in this way.

[2] https://radimrehurek.com/gensim/models/word2vec.html

Table 3.2: MAP performance for different cut-off points in i-w2v algorithm (All vocabulary on MED14Test) *Cut-off means discard all concepts that have a similarity lower than the cut-off value compared to the concept with the highest similarity.*

Cut-off	MAP	Average Number of Concepts
none	0.136	9.4 ± 13.4
25%	0.136	9.3 ± 13.4
50%	0.137	7.2 ± 12.1
75%	0.141	3.8 ± 6.3
80%	0.142	3.0 ± 4.6
85%	0.142	2.3 ± 2.4
90%	0.142	1.9 ± 1.3
0.1	0.142	2.9 ± 5.3

3.4. EXPERIMENTS

In our experiments, we use the MED2014Test Set of the TRECVID Multimedia Event Detection Pre-specified Zero-Example task of 2015 (Over et al., 2015). The MED2014Test contains more than 27,000 videos and has ground truth information for twenty events. The evaluation metric is Mean Average Precision (Over et al., 2015). All video scores are sorted in descending order and the rank of the positive videos is used in the evaluation. The next sections explain our experiments on the Vocabulary Challenge and Concept Selection challenge.

3.4.1. VOCABULARY

In the experiments on the Vocabulary challenge, we compare performance of vocabularies that consist of 1) only one dataset; 2) only low- and mid-level concepts (*LowMid*); 3) only high-level concepts (*High*); 4) low-, mid- and high-level concepts (*All*). The datasets used in the LowMid, High and All vocabularies are visualized in Figure 3.1 two pages back.

According to the literature, combining resources generally improves robustness and performance and therefore we hypothesize that 1) *All* outperforms all other vocabularies. Our intuition is that the high-level concepts play an important role in the detection of high-level events and thus we hypothesize that 2) *High* outperforms *LowMid* and 3) Sport and FCVID outperform the other single datasets.

The Concept Selection method used for the experiments on the Vocabulary Challenge is not our proposed Concept Selection method, but the best number of concepts over all events (*top-k*) using the original word2vec method. This number is determined by experiments on the MED2014 TEST with a varying number of selected concepts, from one to twenty. This number therefore displays the best possible k over all events for these twenty events and is thus not influenced by the proposed Concept Selection method, enabling an independent experiment on the vocabularies.

3.4.2. CONCEPT SELECTION

In the experiments on the Concept Selection challenge, we compare performance of our proposed Concept Selection method (*i-w2v*) to the original word2vec method (*top-k*), a knowledge-based method (*CN*), a method using manually selected concepts and weights (*manual*) and the currently known state of the art methods describing their performance on MED14Test. Relating back to the related work, CN is selected as a method from the first category (*ontology*). The i-w2v method falls within the second category (*machine learning*), and the manual method falls within the third category (*relevance feedback*). We hypothesize that 1) i-w2v outperforms CN and 2) manual outperforms both CN and i-w2v. This second hypothesis is based on the finding of Jiang et al. (2015) that automatic Concept Selection is still a challenge.

In the *CN* method, UQ (event name) is first compared to the concepts in the vocabulary. If a concept completely matches UQ, this concept is put in SQ. If no concept completely matches UQ, ConceptNet is used to expand UQ. In this expansion, ConceptNet 5.3 is automatically accessed through the REST API and all words with the relation *RelatedTo*, *IsA*, *partOf*, *MemberOf*, *HasA*, *UsedFor*, *CapableOf*, *AtLocation*, *Causes*, *HasSubEvent*, *CreatedBy*, *Synonym* or *DefinedAs* to UQ are selected, split into words by removing the underscore and compared to the lemmatized set of concepts in the vocabulary. The matching concepts are put in the SQ. The value for c_w is determined by the following equation:

$$c_w = (\frac{score_{rel}}{30})^3 \tag{3.2}$$

This equation is based on the experiments in the previous chapter, where we explain that the scores are often between zero and thirty, which would create a value between zero and one. The third power is based on previous experiments and has some ground in Spagnola et al. (2011), because they explain that ConceptNet uses the third root of the score of the edges to calculate the final score.

If the query expansion directly to UQ still gives no related concepts, the separate words in UQ are compared to the concepts. The words with a matching concept are put in SQ and the other words are expanded through ConceptNet. In order to avoid query drift, the sum of the weights of the expanded words should be the same as the weight of a matched concept. If for example UQ contains of two words, each set of concepts that represent one word should have a weight of 0.5.

In the *manual* method a human researcher had to select the relevant concepts and weights for those concepts for each event. The researcher was presented the event description provided within the TRECVID MED (Over et al., 2015) benchmark, access to the internet to search for examples for the event and knowledge sources such as Wikipedia or the dictionary and the list of concepts. In order to help the human researcher, the ranked list from our i-w2v method (without similarities) was provided to show a list that is somewhat ordered in terms of relevance to the event. This human researcher is a non-native fluent English speaker with a West-European background. The human researcher was instructed to create a diverse and concise list of concepts, to prevent query drift and adding too much noise. The human researcher had to provide weights for the concepts that summed up to one.

3.5. RESULTS

3.5.1. VOCABULARY

The results of the Average Precision performance of the different vocabularies are shown in Table 3.3. The bold number indicates the highest performance per event per vocabulary, both from the vocabularies that contain a single dataset and the vocabularies with concepts from multiple datasets.

Comparing performance of All to the other datasets, we clearly see that on average the combination of all resources is better than using a subselection of the resources, which is consistent with our first hypothesis. Additionally, LowMid and High both have a performance which is on average higher than any of the single dataset vocabularies in that category.

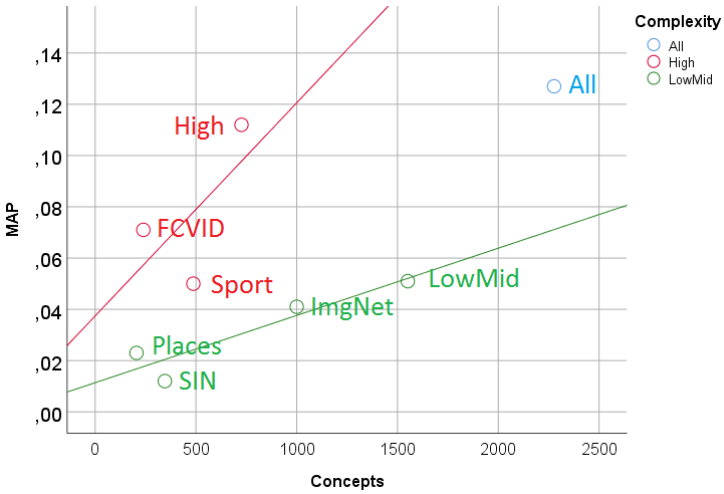


Figure 3.2: Correlation between Number of Concepts and MAP for different complexities

Furthermore, the high-level concepts are important in these experiments, because High outperforms LowMid and the high-level datasets Sports and FCVID outperform Places and SIN. Besides the complexity of the datasets, the number of concepts could also be a factor. A higher number of concepts increases the possibility that the event can be captured within these concepts. This factor can be further verified by the plot in Figure 3.2.

In this plot, the correlation between the number of concepts for each of the complexities is shown. LowMid has a high correlation, whereas High has not (R^2 LowMid = 0.867 and R^2 High = 0.412) between number of concepts and MAP. The plot clearly shows that High performs better than LowMid with the same number of concepts.

Please note that these results could also be explained by that the high level concepts are trained in a domain more like TRECVID MED compared to the domain in which the low level concepts are trained. This domain shift could decrease the performance of the low level concepts compared to the high level concepts.

Table 3.3: Average Precision per vocabulary item using top-k word2vec concept selection (k is optimal determined on MED2014TEST)
Bold is highest in row and group.

	ImgNet (1)	**Places (1)**	**SIN (1)**	**Sport (1)**	**FCVID (1)**	**LowMid (2)**	**High (1)**	**All (1)**
AttemptBikeTrick	0.061	0.002	0.050	0.003	**0.062**	**0.078**	0.062	0.062
CleanAppliance	0.011	0.011	0.009	0.006	**0.062**	0.009	**0.062**	**0.062**
DogShow	0.013	0.011	0.011	**0.766**	0.006	0.005	**0.766**	**0.766**
GiveDirection	**0.006**	0.001	0.002	0.002	0.001	**0.006**	0.002	**0.006**
MarriageProposal	0.002	0.002	0.003	0.002	**0.010**	0.003	**0.010**	**0.010**
RenovateHome	**0.003**	0.002	0.002	0.002	0.001	**0.002**	0.001	0.002
RockClimbing	**0.003**	0.004	0.005	**0.128**	0.065	0.003	**0.128**	**0.128**
TownHallMeeting	0.001	0.008	0.015	0.001	**0.148**	0.015	**0.148**	**0.148**
WinRace	0.006	0.005	0.006	0.010	**0.011**	0.005	**0.010**	0.005
WorkMetalCraftsProject	0.003	0.003	0.002	0.001	**0.005**	0.003	**0.005**	**0.005**
Beekeeping	**0.620**	0.013	0.007	0.011	0.262	**0.620**	0.262	**0.620**
WeddingShower	0.002	0.002	0.002	0.003	**0.005**	0.002	**0.005**	0.002
VehicleRepair	0.002	0.003	0.006	**0.007**	0.001	0.002	0.001	**0.006**
FixMusicalInstrument	0.021	0.024	0.001	0.002	**0.147**	0.004	**0.147**	**0.147**
HorseRidingCompetition	0.022	**0.224**	0.071	0.044	0.098	**0.224**	0.098	0.044
FellingTree	0.002	0.052	0.019	0.012	**0.026**	0.002	**0.026**	**0.026**
ParkingVehicle	0.003	0.023	0.022	0.002	**0.217**	0.023	**0.217**	**0.217**
PlayingFetch	0.001	0.001	**0.001**	0.001	0.001	**0.001**	0.001	0.001
Tailgating	0.004	0.010	0.002	0.002	**0.232**	0.006	**0.232**	**0.232**
TuneMusicalInstrument	0.035	**0.052**	0.004	0.001	0.052	0.008	**0.052**	**0.052**
MAP	0.041	0.023	0.012	0.050	**0.071**	0.051	0.112	**0.127**

3.5.2. CONCEPT SELECTION

The previous section shows the top-k performance for different vocabularies, whereas in this section we compare the Concept Selection methods. The Average Precision performance results for our Concept Selection experiments are shown in Table 3.4. The bold number indicates the highest performance per event per vocabulary. The italic numbers for the CN method indicate random performance, because no concepts are selected. In the All vocabulary, for some events performance of all concept selection methods is equal, indicating that a complete match between the event and a concept in the vocabulary is found. In each of the methods a complete match will result in only selecting that concept. These events are, therefore, displayed on top of the table and separated from the 'interesting' events on the bottom of the table.

Additionally, we compare our best performance against state of the art performance reported on the same dataset in Table 3.5. Performance of CN, top-k and i-w2v on the All vocabulary is shown. This performance is directly comparable to EventPool, because the same vocabularies are used. The vocabularies used by Chang et al. (2016) and Jiang et al. (2015) are comparable in size and type of concepts. In Bor, PCF and DCC (Chang et al., 2016) semantic concepts are discovered using weakly labelling the TRECVID MED research set using word2vec vectors. Bor uses Borda Rank to aggregate the weights on the concepts. PFC uses a pair-comparison framework. DCC uses a dynamic composition to determine the appropriate weights. Fu is the AND-OR method proposed by Habibian et al. (2014a) to create an AND-OR graph of the concepts, but applied to the vocabulary of Chang et al. (2016). The vocabulary of Habibian et al. (2014a) was composed of 138 concepts. These concepts were automatically extracted from the TRECVID MED research set. Jiang et al. (2015) uses an average fusion of the mapping algorithms that use exact word matching, Wordnet, Pointwise Mutual Information and word embeddings. Table 3.5 shows a gain in MAP of 1% compared to state of the art methods.

Comparing the Concept Selection methods, manual is the best overall Concept Selection method, as expected given our hypothesis. The largest differences between manual and i-w2v and CN are in *VehicleRepair* and *HorseRidingCompetition* in High and All. Table 3.6 shows the different concepts and similarities for *VehicleRepair* in All. The concept *assemble bike* has high performance, because this is the only concept that differs between i-w2v / top-k and manual. In the High vocabulary, performance for this event drops, because the concept *vehicle* is no longer within the vocabulary. This same phenomenon happens in the event *Beekeeping* with the concept *apairy*. The main difference in performance in *HorseRidingCompetition* is that the human researcher was able to select all types of horse riding competitions, whereas CN only selected *dressage* and i-w2v only selected the concept *horse racing* in High and *horse racing* and *horse* in All. The difference between High and All with manual in this event is due to the concept *horse race course*.

Table 3.4: Average Precision on MED2014TEST for proposed i-w2v, top-k, ontology-based CN and manual concept selection
Top part are events with direct matches to a concept. Bold is highest value in row and group.

	LowMid				**High**				**All**			
	i-w2v	top-k	CN	manual	i-w2v	top-k	CN	manual	i-w2v	top-k	CN	manual
AttemptBikeTrick	**0.103**	0.078	0.021	0.08	0.062	0.062	0.062	0.062	0.062	0.062	0.062	0.062
CleanAppliance	0.014	0.009	0.005	**0.021**	0.062	0.062	0.062	0.062	0.062	0.062	0.062	0.062
DogShow	**0.021**	0.005	0.011	0.011	0.766	0.766	0.766	0.766	0.766	0.766	0.766	0.766
MarriageProposal	0.003	0.003	*0.001*	**0.005**	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
RockClimbing	0.006	0.003	0.002	**0.025**	0.309	0.128	0.309	0.309	0.309	0.128	0.309	0.309
TownHallMeeting	0.012	0.015	0.007	**0.023**	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148
FixMusicalInstrument	0.025	0.004	0.009	**0.057**	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147
Tailgating	0.007	0.006	0.002	**0.010**	0.232	0.232	0.232	0.232	0.232	0.232	0.232	0.232
MAP (direct)	0.024	0.015	0.008	**0.029**	0.217	0.194	0.217	0.217	0.217	0.194	0.217	0.217
GiveDirection	**0.006**	**0.006**	0.002	0.004	0.002	0.002	0.001	**0.004**	0.006	0.006	0.002	**0.008**
RenovateHome	0.003	0.002	**0.017**	0.003	0.001	0.001	0.002	**0.015**	0.002	0.003	**0.015**	0.008
WinRace	0.006	0.005	0.007	**0.035**	0.043	0.01	0.007	**0.093**	0.021	0.005	0.011	**0.093**
WorkMetalCraftsProject	0.003	0.003	0.001	**0.016**	0.005	0.005	0.001	**0.007**	0.005	0.005	0.001	**0.008**
Beekeeping	0.62	0.62	0.65	**0.694**	**0.262**	**0.262**	**0.262**	**0.262**	0.62	0.62	0.666	**0.714**
WeddingShower	0.002	0.002	0.002	**0.005**	**0.005**	**0.005**	0.002	**0.005**	0.005	0.002	0.002	**0.005**
VehicleRepair	**0.006**	0.002	0.003	**0.006**	0.006	0.001	0.003	**0.162**	0.006	0.006	0.005	**0.284**
HorseRidingCompetition	0.182	**0.224**	0.015	0.183	0.098	0.098	0.096	**0.261**	0.119	0.044	0.096	**0.288**
FellingTree	**0.031**	0.002	0.006	0.015	0.024	0.026	0.001	**0.033**	**0.042**	0.026	0.008	0.015
ParkingVehicle	0.026	0.023	0.022	**0.031**	**0.217**	**0.217**	0.001	**0.217**	**0.220**	0.217	0.013	0.216
PlayingFetch	0.001	0.001	**0.012**	0.004	0.001	0.001	0.022	**0.023**	0.001	0.001	0.02	**0.023**
TuneMusicalInstrument	**0.050**	0.008	0.012	0.046	**0.052**	**0.052**	*0.001*	**0.052**	**0.052**	**0.052**	0.012	**0.052**
MAP (no direct matches)	0.078	0.075	0.062	**0.087**	0.06	0.057	0.033	**0.095**	0.092	0.082	0.071	**0.143**
MAP (all)	0.056	0.051	0.04	**0.064**	0.123	0.112	0.107	**0.144**	0.142	0.127	0.129	**0.173**

Table 3.5: Comparison to State of the Art (MAP reported on MED2014TEST)

Method	**MAP**
AND-OR(Habibian et al., 2014a)	0.064
Bor (Chang et al., 2016)	0.102
Fu (Chang et al., 2016; Habibian et al., 2014a)	0.111
PCF (Chang et al., 2016)	0.114
AutoSQGSys (Jiang et al., 2015)	0.115
top-k (All)	0.127
EventPool (Lu et al., 2016b)	0.129
CN (All)	0.129
DCC (Chang et al., 2016)	0.134
i-w2v (All)	**0.142**

Table 3.6: Comparison for VehicleRepair in All

i-w2v / top-k		**CN**		**manual**	
c	c_s	c	c_s	c	c_s
vehicle	0.760	vehicle	0.500	vehicle	0.5
		band aid	0.095	assemble bike	0.5
		highway	0.095		
		apartments	0.095		
		boating	0.095		
		shop	0.095		
		casting fishing	0.024		

Following our hypothesis, i-w2v outperforms CN in all vocabularies. I-w2v even outperforms manual in some events, of which *FellingTree* is the most interesting. Table 3.7 shows the concepts and similarities of the different methods for the event *FellingTree* in All. In i-w2v, the concept *tree farm* causes the high performance, whereas *chain saw* decreases performance compared to only using the concept *fruit tree pruning*. In CN, the wrong expansion from *felling* to *falling* to all concepts, except for *trees*, causes the low performance. Please note that the human researcher has the highest performance in High. The selected concepts for manual in High are *forest* and *fruit tree pruning*.

Comparing i-w2v to top-k, the i-w2v method outperforms the top-k in all vocabularies. In the High vocabulary, performance of the event *Rock Climbing* in top-k is slightly lower compared to the other direct matches, because in top-k the first occuring direct match is used instead of all direct matches. Using all direct matches for this event would improve MAP performance in All to 0.136.

Table 3.7: Comparison for FellingTree in All

i-w2v		**CN**		**manual**		**top-k**	
c	c_s	c	c_s	c	c_s	c	c_s
fruit - tree pruning	0.720	trees	0.500	trees	0.5	fruit - tree pruning	0.720
tree frog	0.686	cliff	0.186	chain saw	0.5		
tree farm	0.678	painting	0.106				
		skateboarding	0.085				
		climbing	0.040				
		windows	0.040				
		head	0.002				
		running	0.001				
		building	7×10^{-6}				

Table 3.8: Comparison for RenovateHome in LowMid

i-w2v		**CN**		**manual**		**top-k**	
c	c_s	c	c_s	c	c_s	c	c_s
apartment - building- outdoor	0.542	apartments	0.113	apartment- building- outdoor	0.25	apartment- building- outdoor	0.542
building	0.526	city	0.102	apartments	0.25		
home office	0.475	person	0.083	construction site	0.5		
apartments	0.466	wardrobe	0.065				
church- building	0.465	sofa	0.065				
	0.465						
building- facade	0.452	tabby cat	0.065				
	0.452	tabby cat	0.065				
mobile- home	0.437	closet	0.065				
		bedroom	0.065				
		comfort	0.065				
		dogs	0.065				
		building	0.058				
		pillow	0.047				
		refrigerator	0.047				
		furniture	0.047				
		pantry	0.047				

Interestingly, CN outperforms both i-w2v and manual in the events *Renovate-Home* in LowMid and All and *PlayingFetch* in LowMid. Table 3.8 shows the concepts and similarities of the different methods for the event *RenovateHome* in LowMid. In the event *PlayingFetch* in LowMid the addition of concepts, such as *throwing, ball* and *stick* (manual), decreases performance compared to only using the concept *dog* (CN).

3.6. DISCUSSION

Regarding the Vocabulary challenge, the results of the experiments show that a combination of multiple datasets improves performance. Although state of the art already tends to add as many datasets as possible to their vocabulary, we show that including high-level concepts is important in video event retrieval. The results on the Vocabulary challenge show that using only the High vocabulary is better than using the LowMid vocabulary. The All vocabulary with both LowMid and High is also better than the LowMid. The correlation graph in Figure 3.2 shows that All is in the middle between LowMid and High. This observation makes us wonder if a combination of a LowMid and High vocabulary is indeed a good way to go, or if we should focus on a High vocabulary with more concepts. On one hand, the LowMid concepts are useful when no close matches of the High level concepts are present. On the other hand, the High level concepts can capture more than the combination of the LowMid level concepts. A related point is whether the high-level concepts can improve performance on lower level concept queries, such as *horse riding*. Will the high-level concept *horse riding competition*, possibly together with other events that include horse riding, improve performance on this query? In our opinion a concept on the same level of complexity as the query will provide the best performance, i.e. the query *horse riding* will achieve a higher retrieval performance with the matching concept *horse riding* compared to the concept *horse riding competition*, assuming both concept detectors perform accurately. In this example, the higher-level concept *horse riding competition* only includes a limited set of the query, resulting in a high precision but low recall situation. A lower-level concept, such as *horse* would include a set that is too broad, resulting in a high recall and low precision situation.

Regarding i-w2v, performance is higher compared to current state of the art zero shot methods without re-training or re-ranking. I-w2v can be combined to the event pooling method from Lu et al. (2016b) and the DCC method of Chang et al. (2016) to achieve an additional performance gain. The increase in performance compared to top-k does not seem significant, but when increasing the number of concepts, the possibility of query drift is high. Current top-k strategy is to add only the one most relevant concept. With a direct or near direct match between the event and the concepts, this is a reasonable strategy. In other tasks or with other events, this strategy is not optimal and a different number of k should be taken. Instead of optimizing the number k for each task, our strategy does not need this optimization. I-w2v is also able to combine concepts which cover different facets of the event, whereas other methods might only use the raw cosine similarity. Additionally, i-w2v does not seem that sensitive to the cut-off point, as shown in Table 3.2.

Our proposed i-w2v method approaches the manual method. An advantage of the manual method is that human knowledge is richer than the knowledge in current knowledge bases or in word2vec, but the disadvantage is that 1) it requires a human to interpret all queries, which seems unfeasible in real-world applications; 2) it is hard for a human to indicate the proper weight. CN and w2v can automatically assign weights, but these weights are based on textual similarity. W2v learns from the context in which words appear, but the context does not indicate if the words are similar because they have an antonym (cat vs. dog), hyponym (chihuhua vs. dog),

hypernym (animal vs. dog) or other type of relation. Knowledge bases such as ConceptNet have such relations, but for events little or no information is present. Because word2vec works as a vector model, the combination of multiple words in an user query gives better results than a combination of the different words searched in one of the knowledge bases. The method can, however, still be improved, because concepts with one directly matching word, such as *tree* in the concept *tree frog* for the event *FellingTree* and *home* in *home theater* for the event *RenovateHome*, sometimes retrieve a similarity that can be argued to be too high. But our word2vec method does not suffer from query drift and it approaches human performance, especially in a vocabulary that contains high-level concepts. In future work, an option could be to combine our method with the manual method using either relevance feedback or a hybrid method containing i-w2v and a knowledge base.

3.7. CONCLUSION

In this chapter, we presented our Semantic Event Retrieval System that 1) includes high-level concepts and 2) uses a novel method in Concept Selection (*i-w2v*) based on semantic embeddings. Our experiments on the international TRECVID Multimedia Event Detection benchmark show that a vocabulary including high-level concepts can improve performance on the retrieval of complex and generic high-level events in videos, indicating the importance of high-level concepts in a vocabulary. Second, we show that our proposed Concept Selection method outperforms state of the art.

4

IMPROVING VIDEO EVENT RETRIEVAL BY USER FEEDBACK

Edited from: **Maaike de Boer, Geert Pingen, Douwe Knook, Klamer Schutte and Wessel Kraaij** (2017) *Improving Video Event Retrieval by User Feedback* In: Multimedia Tools and Applications, volume 76, number 21, pp. 22361-22381.
* Experiments have been conducted by Geert Pingen and Douwe Knook under supervision of Maaike de Boer

In previous chapters, we have shown that we are able to achieve state of the art performance in a zero example case using automatic query-to-concept mapping, but the current methods do not work flawlessly. In this chapter, we investigate how we can exploit the users to improve performance. This chapter is related to research question **RQ3 ARF**. *We explore relevance feedback methods on concept level and on video level. On concept level, we use re-weighting and Query Point Modification as well as a method that changes the semantic space the concepts are represented in. This semantic space is the word2vec space used in the previous chapter. On video level, we propose an Adaptive Relevance Feedback (*ARF*) method, which is based on the classical Rocchio relevance feedback method from the field of text retrieval. The other video level method is a state of the art k-Nearest Neighbor method. Results on the TRECVID MED 2014 train set show that user feedback improves performance compared to no feedback. Feedback on video level provides a higher performance gain compared to the concept level. A possible reason for this higher gain is that feedback on video level can capture both information on the semantic relevance of a certain concept as well as the accuracy of the concept detector. Feedback on concept level can only provide the former. Our proposed ARF method outperforms all other methods, i.e. methods on concept level, the state of the art k-NN method and manually selected concepts.*

4.1. INTRODUCTION

Current video search systems, such as YouTube (Burgess et al., 2013), mostly rely on the keywords typed with the uploaded videos. In the field of content-based video retrieval, systems retrieve videos using the content of the video within keyframes of the video. Typically *concept detectors* are trained to index videos with the concepts present. One of the constraining factors in concept-based video retrieval systems is the limited number of concepts a system can be trained to detect. While current state-of-the-art systems are able to detect an increasingly large number of concepts (*i.e.* thousands), this number still falls far behind the near infinite number of possible (textual) queries general-purpose heterogeneous video search systems need to be able to handle (Boer et al., 2015a). One of the challenging areas within the concept-based video retrieval is that of event retrieval. Events can be defined as complex queries that consist of a multitude of concepts, such as objects, actions and scenes. One example of an event query is *Attempting a bike trick*. This query can be represented by more general concepts such as *bike trick*, *attempt* and *flipping bike*. Creating an automatic representation of a query can, however, include non-relevant or less representative concept detectors and, thus, decrease retrieval performance. Furthermore, the meaning of a concept is different in different contexts, and therefore the quality of a concept detector might differ in the context in which it is applied.

One approach to improve performance when less or non-relevant detectors are selected is the use of *relevance feedback*. With relevance feedback the (estimated) behavior of the user with the system is used to improve the system. This method is well accepted and commonly used in text retrieval. In video retrieval the trend is to either use click behavior or to use pseudo-relevance feedback (Zhou et al., 2003; Patil, 2012; Jiang et al., 2015), in which we assume that the first x videos are relevant. In this chapter, we focus on explicit user feedback, both on the retrieved videos and on selected concepts that represent a query. We compare which relevance feedback level can provide the highest performance gain. Furthermore, we propose a novel method on video level. Our Adaptive Relevance Feedback (*ARF*) is inspired by the Rocchio algorithm (Rocchio, 1971) often applied in the field of text retrieval. Whereas state of the art relevance feedback algorithms on video level use the annotated videos to create a novel model based on nearest neighbor or SVM type of algorithms (Gia et al., 2004; Deselaers et al., 2008), we use the videos to approximate the proper weights of the selected concepts in our query representation. The advantage of changing the weights is that this method is able to benefit from just a few positive and negative annotations, compared to newly trained models.

We compare the results of our ARF algorithms on the MEDTRAIN set of the TRECVID benchmark (Over et al., 2015) against traditional relevance feedback approaches, such as approaches on concept level such as re-weighting and QPM, and a k-NN based method. Results show that 1) relevance feedback on both concept and video level improves performance compared to using no relevance feedback; 2) relevance feedback on video level obtains higher performance compared to relevance feedback on concept level; 3) our proposed ARF method on video level outperforms a state of the art k-NN method, all methods on concept level and even manual selected concepts.

4.2. RELATED WORK

The use of relevance feedback stems from the dynamic nature of information seeking (Ruthven et al., 2003): information needs can be continuously changing and be unique to each user. Relevance feedback can be done in different ways: *implicit*, *explicit* and *blind/pseudo*. In implicit relevance feedback, implicit information, such as user clicks or dwell time, is used. The advantage of this method is that you do not have to bother the user, but the inference of the results is much harder. Because we focus on a subset of users in which we expect less queries, we expect that implicit feedback will provide a smaller gain compared to explicit user feedback. In explicit relevance feedback, the user explicitly indicates if a certain item is relevant or not relevant. This can be done using a binary scale or a graded scale. The advantage of this method is that you have a clear indication of the relevance and a higher performance, but the disadvantage is that you have to bother the user. This user might not have time or motivation to give such feedback. In blind- or pseudo-relevance feedback, the manual user part is automated. In this automation, we assume that the first x ranked items are relevant. This assumption is not without a risk, because in the case of rare events or new query domains, bad retrieval systems or ambiguous queries this assumption might not hold. Human relevance feedback (implicit and explicit) has been known to provide major improvements in precision for information retrieval system. Dalton et al. (2013) have shown that —in the domain of video retrieval —pseudo-relevance feedback can increase Mean Average Precision (MAP) up to 25%, whereas with human judgments this number can grow up to 55%. Of course the effectiveness of pseudo relevance feedback critially depends on the assumption that the collection contains at least a reasonable number of relevant results and that the first retrieval pass is able to pick up a good fraction of those in the top x. It is clear that relevance feedback, when applied correctly, can help the user in better finding results.

One of the most well-known and applied relevance feedback algorithms that has its origins in text retrieval is the Rocchio algorithm (Rocchio, 1971). This algorithm is used in state of the art video and text retrieval systems that use for example Query Point Modification (QPM) to move the query representation in the vector space and re-weighting in which the terms in the query are re-weighted (Rocha et al., 2015; Jiang et al., 2014b; Tsai et al., 2015; Kaliciak et al., 2013). Often a document is represented as a vector with a real-valued component (e.g. tf-idf weight) for each term. The Rocchio algorithm works on a vector space model in which the query drifts away from the negatively annotated documents and converges to the positively annotated documents. The Rocchio algorithm is effective in relevance feedback, fast to use and easy to implement. The disadvantages of the method are that an α and β parameter have to be tuned and it cannot handle multimodal classes properly.

Other state of the art approaches, such as feature-, navigation-pattern, and cluster-based approaches, in image retrieval are explained by Zhou et al. (2003) and Patil (2012). Often the system will actively select the documents that achieve the maximal information gain (Tong et al., 2001). Some vector space models use k-Nearest Neighbor methods, such as in the studies by Gia et al. (2004) and Deselaers et al. (2008). K-NN based methods are shown to be effective, and are

non-parametric, but run time is slower and it can be very inaccurate when the training set is small. Other methods use decision trees, SVMs, or multi-instant approaches and are explained in Crucianu et al. (2004). A disadvantage of those other methods that they need sufficient annotations to work properly. SVMs are often used (Xu et al., 2015; Yang et al., 2010; Tao et al., 2008), but according to Wang et al. (2016b), SVM-based RF approaches have two major drawbacks: 1) multiple feedback interactions are necessary because of the poor adaptability, flexibility and robustness of the original visual features; 2) positive and negative samples are treated equally, whereas the positive and negative examples provided by the relevance feedback often have distinctive properties, such as that the positive examples are close to each other whereas negative examples are arbitrarily distributed. Within the pseudo-relevance feedback, this second point is taken by Jiang et al. (Jiang et al., 2014a; Jiang et al., 2014b; Jiang et al., 2015), who use an unsupervised learning approach in which the 'easy' samples are used to learn first and then the 'harder' examples are iteratively added.

4.3. VIDEO EVENT RETRIEVAL SYSTEM

Our Video Event Retrieval System is inspired by state of the art video event retrieval systems without training examples (Zhang et al., 2015a; Jiang et al., 2014b). The pipeline of our system is shown in Figure 4.1. In our system a user can enter a textual query (*Event Query*) into the search engine. This query is represented by a combination of concepts in the module *Query Interpretation* using the *word2vec model* and the *Concept Bank*. This combination of concepts is propagated back to the user to obtain *relevance feedback* on concept level and the *top n concepts* are used as an OR query in the *Scoring+Ranking* module. This module retrieves the videos in the database, sums the evidence from individual concepts and ranks the results in descending order of estimated relevance. These results are presented back to the user and the user can provide *relevance feedback* on video level. These modules are explained in more depth in the next subsections.

4.3.1. QUERY INTERPRETATION

The Event Query is translated to a system query (video concept representation) using a word2vec model, which is commonly used in video retrieval (Elhoseiny et al., 2016; Jiang et al., 2015; Snoek et al., 2015; Norouzi et al., 2013). A word2vec model uses a shallow neural network that is trained on a huge dataset, such as Wikipedia, Gigawords, Google News or Twitter, to create semantic word embeddings. The Word2Vec models operate on the hypothesis that words with similar meanings occur in similar contexts (Goldberg et al., 2014), resulting in a good performance in associations, such as *king - man + woman = queen*. We use a model that is pre-trained on Google News[1]. The embedding of each word is expressed in a 300-dimensional feature vector. This model is used because it shows better results compared to the other pre-trained word2vec models, such as the Wikipedia models. We do not re-train the network, because this did not increase performance in our experiments. Using the word2vec

[1] https://code.google.com/archive/p/word2vec/

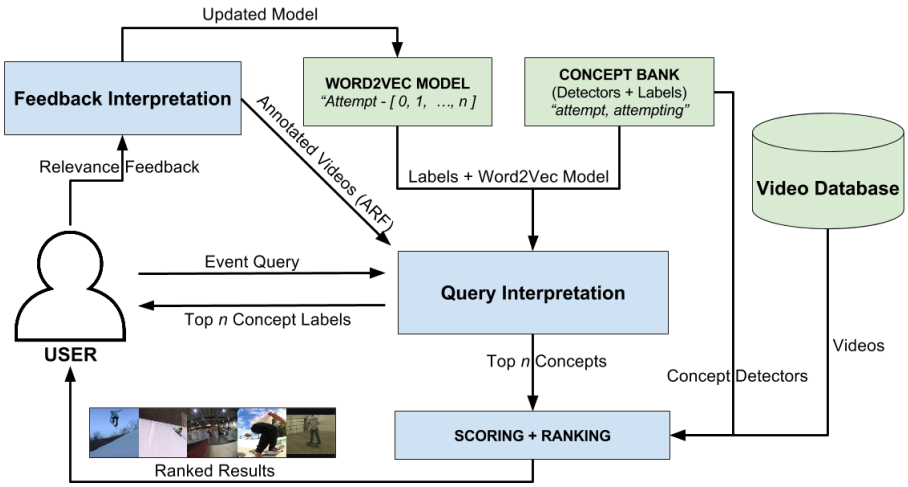


Figure 4.1: Pipeline of our Semantic Video Event Search System

model, we calculate the distance between the event query and each of the concepts that can be detected. The concepts that can be detected are obtained from the Concept Bank (explained in the next subsection). In the word2vec model we calculate the vector of the event query by mean pooling the vectors representing the words in the event query, without using the vectors of stopwords, such as 'a'. If a word in the event query is not in the vocabulary of the word2vec model, we discard this word as well. As shown by Lev et al. (2015), mean pooling is a simple pooling method that performs well. The words in the labels of the concepts in our Concept Bank are mean pooled as well and compared to the vector representing the event query. The cosine similarity, which is a robust similarity measure in this semantic space (Lev et al., 2015), is used to calculate the distance between the event query and each of the concepts that can be detected independently, as explained in:

$$w_d = \frac{\overrightarrow{q} * \overrightarrow{v_d}}{||\overrightarrow{q}|| * ||\overrightarrow{v_d}||} \tag{4.1}$$

, where $\overrightarrow{q}$ is the 300 dimensional mean word2vec vector for the event query, $\overrightarrow{v_d}$ is the mean vector for detector d.

This distance is used to determine the combination of concepts that represent the query. In our experiments, we use the top n concepts with the highest similarity measure (w_d), based on initial experiments. These concepts are used for the 1) relevance feedback and 2) scoring.

CONCEPT BANK

The Concept Bank contains labels and detectors that are trained on different datasets using Deep Convolutional Neural Networks (DCNN). We use the eight layers of the DCNN network trained on the ILSVRC-2012 (Deng et al., 2009), as is often used in

this field (Jiang et al., 2015; Snoek et al., 2015; Zhang et al., 2015a). We finetune the architecture on the data in the dataset for SIN (Over et al., 2015), Places (Zhou et al., 2014) and TRECVID MED (Over et al., 2015) to obtain more concepts (2048) than the 1000 objects used in the ILSVRC-2012. The concepts from the TRECVID MED are manually annotated on the Research set, comparable to Natarajan et al. (2011) and Zhang et al. (2015a). We purposely did not use higher level concept detectors, such as those available in the FCVID (Jiang et al., 2017) or Sports (Karpathy et al., 2014) dataset, to obtain more interesting experiments using relevance feedback. We, therefore, do not aim at highest possible initial ranking, but at a gain with the use of relevance feedback. We believe this is applicable to real world cases, because relevant high level concepts are not always present.

4.3.2. SCORING AND RANKING

For the scoring, we need the video scores of the top n concept detectors, obtained from the Query Interpretation module, from our database. The pre-trained concept detectors are applied on each of the videos in our database. Because the network is trained on images, we extract 1 keyframe per 2 seconds uniformly from a video. We use max pooling over these keyframes to obtain a concept detector score per video. Furthermore, we use the average concept detector scores on a background set to normalize the detector scores on the videos in our database.

The scoring function is defined as:

$$s_v = \sum_{d \in D} w_d * (s_{v,d} - b_d) \tag{4.2}$$

, where w_d is described in Equation 4.1 and represents the cosine similarity between the event query vector q and the detector vector v_d, b_d is the average background score of detector d and $s_{v,d}$ is the score for detector d on video v. The videos are returned to the user in descending order of their overall score s_v.

4.3.3. FEEDBACK INTERPRETATION - ADAPTIVE RELEVANCE FEEDBACK (ARF)

Feedback can be obtained on concept level and on video level. We propose an algorithm on video level for explicit relevance feedback, but implementations on concept level are available in our system as well (explained in the experiments).

Our *Adaptive Relevance Feedback* algorithm (ARF) is inspired by the Rocchio algorithm (Rocchio, 1971). Different from traditional algorithms on video level (Crucianu et al., 2004; Zhou et al., 2003; Patil et al., 2011), we use relevance feedback to update the weights for our concept detectors instead of training a new model based on (few) annotations. We choose to update the weights to make our algorithm more robust to few or noisy annotations. In k-NN methods, noisy annotations can have a high impact on ranking performance. By taking into account the initial concept detector cosine distance to the query, the proposed algorithm is more robust to this type of relevance feedback.

The weights are updated using the following formula:

$$\begin{aligned} w'_d &= w_d + (\alpha \cdot m_R) - (\beta \cdot m_{NR}) \\ m_R &= \frac{\sum_{v \in R} s_{v,d} - b_d}{|R|} \\ m_{NR} &= \frac{\sum_{v \in NR} s_{v,d} - b_d}{|NR|} \end{aligned} \tag{4.3}$$

, where v is the considered video, d is the detector, R is the set of relevant videos, NR is the set of non-relevant videos, $s_{v,d}$ is the score for concept detector d for video v, w_d is word2vec cosine similarity between the query vector $\overrightarrow{q}$ and the detector vector $\overrightarrow{v_d}$, b_d is the average background score of detector d, and α and β are Rocchio weighting parameters for the relevant and non-relevant examples respectively.

The adjusted detector weight, w'_d, is then plugged back into the scoring function, where we substitute the original word2vec score for the adjusted weight. This results in new scores, s'_v, for each video v, which is used to create an updated ranked list of videos.

4.4. EXPERIMENTS

In our experiments, we evaluate our proposed methods in an international video retrieval benchmark and compare performance to state of the art.

4.4.1. EXPERIMENTAL SET-UP

We use the MEDTRAIN data set from the TRECVID Multimedia Event Detection (MED) benchmark (Over et al., 2015). This data set contains 5594 videos of user-generated content. The MEDTEST set is often used in other papers to report performance on, but the MEDTRAIN contains relevance judgments for forty events (i.e. queries), whereas MEDTEST contains judgments for only twenty events. Although we purposely did not use higher level concept detector datasets to obtain our concepts, some concepts caused a (near-)perfect performance because of a direct match between an event and the concept. We, therefore, excluded eight of the forty events[2]. These events are not interesting for the user feedback experiments.

The number of concepts n for ARF is chosen to be 30. Our baseline experiments showed highest performance for $n = 5$ as shown in Figure 4.2, but our experiments showed that a higher performance gain can be achieved by using more concepts. Furthermore, the α of our ARF algorithm is set to 1.0 and the β is set to 0.5, which is in line with text-information retrieval (Rocchio, 1971). Visualizations of these results can be found in Figure 4.3.

[2] excluded events are *Wedding ceremony*; *Birthday party*; *Making a sandwich*; *Hiking*; *Dog show*; *Town hall meeting*; *Beekeeping*; *Tuning a musical instrument*

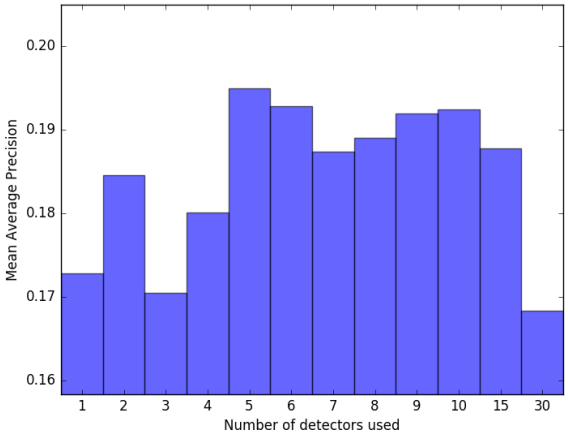


Figure 4.2: MAP per number of concept detectors over all events

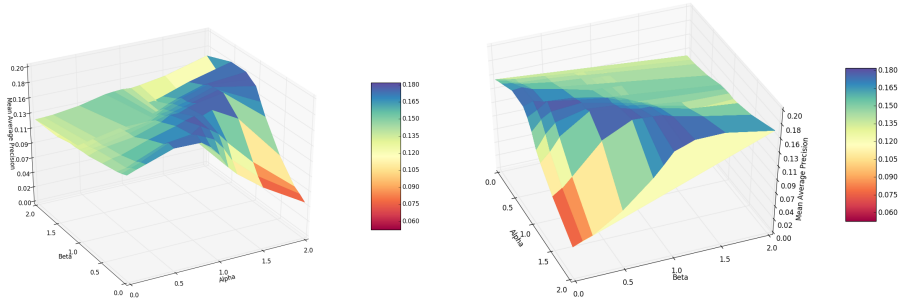


Figure 4.3: MAP^* relative to α and β values

EVALUATION

Mean Average Precision (MAP) (Over et al., 2015), which is the official performance measure in the TRECVID MED task, is used to measure performance. With relevance feedback on video level, the positively annotated videos will remain on the top of the list and, thus, increase MAP. It is, however, also interesting to know whether the algorithm is able to retrieve new relevant videos. This is why we introduce a variant of the MAP. *MAP** calculates MAP disregarding the videos that have been viewed already by the user. We assume that a user has viewed all videos up to the last annotated video.

Additionally, we calculate robustness of our proposed method compared to the best state of the art method on that level by the robustness index (RI) (Sakai et al., 2005) and the concept level methods against the initial ranking using:

$$RI = \frac{|Z_P| - |Z_N|}{|Z|} \tag{4.4}$$

, where $|Z_P|$ - $|Z_N|$ is the number of queries in which the first method has higher performance compared to the second method, and $|Z|$ is the total number of queries.

USER INTERFACE

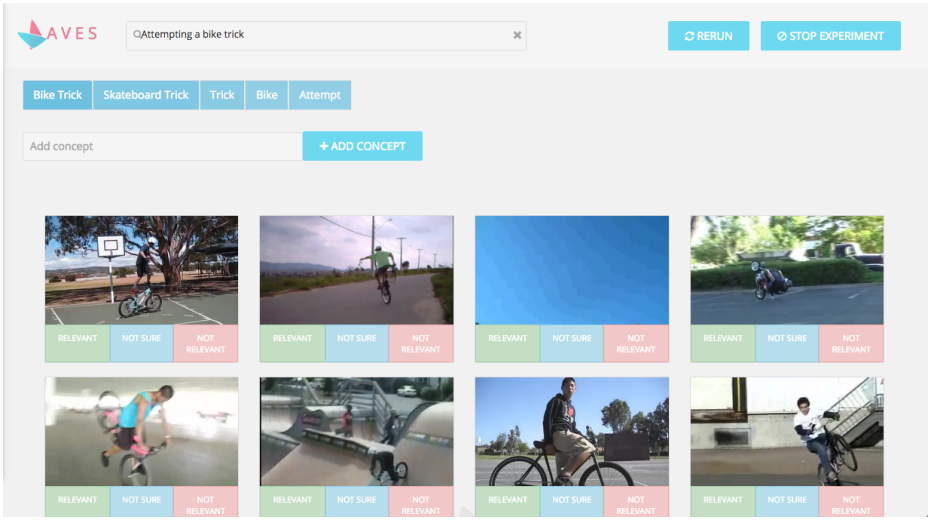


Figure 4.4: Screenshot of our User Interface for the event *Attempting a bike trick*

To provide the user with a quick and efficient way of viewing the concepts and the videos in the experiment, we designed a User Interface (UI). A screenshot of the UI is presented in Figure 4.4. For the videos, we aim to show a small subset of the keyframes instead of the whole video. For each video, the 5 keyframes (based on state of the art in current search engines, such as Bing[3]) with the highest scores over the top n detectors, based on their word2vec scores, are selected. A single frame is shown initially for each video in a container, under which we presented the relevance selection tools. When a user moves the mouse over the container, a new frame appears based on the relative position of the mouse in the container. This means that the first frame would be visible when the user was hovering in the first 20% of the container, the second frame when the mouse position was detected in the next 20%, and so on. This enabled our users to get a quick overview of the relevant parts of the video, without having to spend minutes watching each video. For the feedback on the concepts, the videos were not presented but a list of the top 15 concept detectors was shown. This is further explained in the next section.

4.4.2. RELEVANCE FEEDBACK ON CONCEPTS

Fifteen participants (12 male; 3 female; μ age= 24.87; σ age= 3.739) were asked to volunteer in providing relevance feedback. The majority of the participants were non native but fluent English speakers with an education level of Bachelors or higher. The participants were presented with a list of the 32 events on several pieces of paper with the top 15 concepts (in English) per event as provided by the initial system. They were asked to evaluate these concepts and provide relevance judgments by marking

[3]www.bing.com/videos

the non-relevant concepts for each of the events. On average, participants marked 6.2 out of 15 concept detectors as non-relevant (σ = 1.494). The average number of detectors marked as non-relevant differed greatly per event (minimum 0.5 to maximum 11.7) and per user (minimum 3.7 to maximum 8.7). A Fleiss' Kappa test was performed to determine user agreement in the flagging of non-relevant concepts, which resulted in κ = 0.514. According to the Landis and Koch scale (Landis et al., 1977), this indicates a *moderate agreement* among users.

4.4.3. RELEVANCE FEEDBACK ON VIDEOS

For the relevance feedback on videos, a group of ten male participants (μ age= 26.3, σ age= 1.567) with mainly non native but fluent English speakers and an education level of Bachelors or higher without dyslexia, colour-blindness, concentration problems, or RSI problems, voluntarily participated in an experiment. The task of the participants was to select relevant and non-relevant videos in our UI. 24 results were shown initially, and more could automatically be loaded by scrolling to the bottom of the page. The experiment consisted of two conditions, which correspond to the re-ranking results by ARF and the k-NN method named RS (next subsection). In each of the conditions, 16 queries, randomly assigned using a Latin rectangle (Cochran et al., 1957), were presented to the user using our UI, after which they performed relevance feedback on the retrieved videos.

4.4.4. BASELINE METHODS

We compare our ARF algorithm with several baselines, which are presented in the next subsections. The SVM-based methods are not included in this chapter, because preliminary experiments showed that on average performance is poor due to limited number of positive samples.

NO FEEDBACK

The No Feedback method is the system without the relevance feedback module. The number of concepts n is chosen to be 5, based on the results reported in Figure 4.2.

MANUAL

An expert familiar with the TRECVID MED events, the Concept Bank and data set was asked to select a set of relevant concepts and their weights for each event. The number of selected concepts varies among the events.

CONCEPT LEVEL - ALTERWEIGHTS

As a re-weighting method, we alter the weights of the concept detectors following an approach inspired by the Rocchio algorithm (Rocchio, 1971). The weights of the relevant detectors are increased, whereas the weights of the irrelevant detectors are decreased following Equation 4.5. The values for γ and δ are the best values based on our experiments on the same data to provide upper bound performance. This method is different from ARF, because this method works on the relevance feedback on concept level and not on video level. The number of concepts n for all concept level based experiments is set to 15, because previous experiments showed that a

higher number of concepts in relevance feedback can achieve higher performance gain compared to using only the top 5 (often positive) concepts.

$$w'_d = \begin{cases} w_d + \gamma * w_d, & \text{if d is relevant.} \\ w_d - \delta * w_d, & \text{otherwise.} \end{cases} \tag{4.5}$$

, where $\gamma = 0.4$ and $\delta = 0.9$.

CONCEPT LEVEL - QUERYSPACE

As a QPM method, we change the semantic space of the query using Rocchio's algorithm. Using the vector representations of both the relevant and non-relevant detectors provided by concept level relevance feedback, we update the initial query vector $\overrightarrow{q}$ that is used to calculate the cosine similarity w_d (Equation 4.1 in Section 4.3.1) according to Equation 4.6. Again, the values for ϵ and ζ are the best values based on our experiments on the same data to provide optimal performance.

$$\overrightarrow{q}' = \overrightarrow{q} + \epsilon * \left(\frac{1}{|C_r|} \sum_{d \in C_r} \overrightarrow{v_d} \right) - \zeta * \left(\frac{1}{|C_{nr}|} \sum_{d \in C_{nr}} \overrightarrow{v_d} \right) \tag{4.6}$$

, where $\overrightarrow{q}'$ is the modified query vector, C_r and C_{nr} are the set of relevant and non-relevant concept detectors, respectively and $\overrightarrow{v_d}$ is the word2vec vector representation of detector d, $\epsilon = 0.6$ and $\zeta = 0.7$.

CONCEPT LEVEL - DETECTORSPACE

Instead of changing the query space, we can also change the semantic space. We change the concept detector labels by moving the mean pooled vector of the relevant concepts toward the mean pooled vector of the event query, whereas we move the non-relevant concepts away from the event query with the following equation:

$$\overrightarrow{v_d}' = \overrightarrow{v_d} + \eta * \theta_d * (\overrightarrow{q} - \overrightarrow{v_d}) \tag{4.7}$$

, where $\overrightarrow{v_d}'$ is the new vector for detector d, $\overrightarrow{v_d}$ is the old vector of detector d and $\overrightarrow{q}$ is the event query vector, $\eta = 0.1$, θ is described as:

$$\theta_d = \begin{cases} -1, & \text{if } d \in C_{nr} \\ 1, & \text{otherwise} \end{cases} \tag{4.8}$$

, where d is the detector, C_{nr} is the set of non-relevant concept detectors.

This new vector is used to calculate the new cosine similarity w_d, which is used in the determination of the relevant concepts and the scoring function (Equation 4.2 in Section 4.3.2). This method changes the concepts in the space and, therefore, this method can change performance on other events, whereas in the other methods the performance on only one query is improved. This method, however, introduces different results for different order of events. In our experiments, we choose the average performance over 2 runs of 32 events over all 15 users.

VIDEO LEVEL - RS

The final baseline is a k-NN based relevance feedback algorithm named *Relevance Score* (RS). The RS algorithm is well-performing in image retrieval (Gia et al., 2004; Deselaers et al., 2008) and the relevance score $relevance(v)$ of a video v calculated as

$$relevance(v) = \left(1 + \frac{dR(v)}{dNR(v)}\right)^{-1} \tag{4.9}$$

, where dR is the dissimilarity, measured as Euclidean distance, from the nearest video in relevant video set R, dNR is the dissimilarity from the nearest video in non-relevant video set NR. The video set is ordered such that the videos with the highest *relevance score* are listed first and MAP is calculated on this list.

4.5. RESULTS

4.5.1. MAP AND MAP*

The MAP results on all methods are displayed in Table 4.1. The results show superior performance for our ARF method. All relevance feedback methods outperform the No Feedback run, except DetectorSpace.

Table 4.1: MAP and Standard Deviation over all users and all events on MEDTRAIN dataset

	Method	**MAP** (μ)	**Standard Deviation** (σ)
Baseline	No Feedback	0.19	0.15
	Manual	0.23	0.18
Concept Level	AlterWeights	0.21	0.16
	QuerySpace	0.20	0.16
	DetectorSpace	0.19	0.15
Video Level	RS	0.20	0.17
	ARF	**0.24**	0.17

The standard deviation is relatively high, because we average over all events (some have almost random performance near zero and some have a very good performance near one). This comparison is, however, not completely fair, because annotations on video level will keep the positively annotated videos on the top of the ranked list. One method to overcome this problem is to discard the videos which the users have already seen (MAP^*). We assume that all videos displayed before the last video are seen. The results in MAP^* over all video level methods, including the initial method without these videos, is presented in Table 4.2.

Table 4.2: MAP^* scores and standard deviations on video level on MEDTRAIN dataset

Algorithm	**MAP*** (μ)	**Standard Deviation** (σ)
No Feedback	0.13	0.01
RS	0.11	0.02
ARF	**0.15**	0.02

These results show that RS performs worse compared to No Feedback, because this method might move in the wrong direction when few positive examples are annotated. A Shapiro-Wilk test showed that the precision score distributions do not deviate significantly from a normal distribution at $p > 0.05$ ($p = 0.813; p = 0.947; p = 0.381$, for No Feedback, RS, and ARF, respectively). A statistically significant difference between groups was determined by a one-way ANOVA ($F(2,27) = 18.972$, $p < 0.0005$). A post-hoc Tukey's HSD test was performed to verify intergroup differences. The means of all algorithms differed significantly at $p < 0.05$ ($p = 0.006; p = 0.01; p < 0.0005$, for No Feedback-RS, No Feedback-ARF, and RS-ARF, respectively).

4.5.2. ROBUSTNESS

The robustness index (RI) on concept level, compared to No Feedback, is $RI = 0.125$ for AlterWeights (better in 18 events), $RI = -0.375$ for QuerySpace (better in 9 events) and $RI = -0.0625$ for DetectorSpace (better in 15 events). Interestingly, QuerySpace has higher performance compared to DetectorSpace, although RI is lower. One reason is that in some events DetectorSpace has moved a concept in a wrong direction by which it is not able to retrieve that concept anymore, resulting in a lower MAP.

The RI on video level is calculated by comparing RS to ARF. The RI for ARF is $RI = 0.4375$ and for RS it is $RI = -0.25$. The bar plot is shown in Figure 4.5. Compared to No Feedback ARF improves ranking in 23 of the events, and RS in 12 of the events.

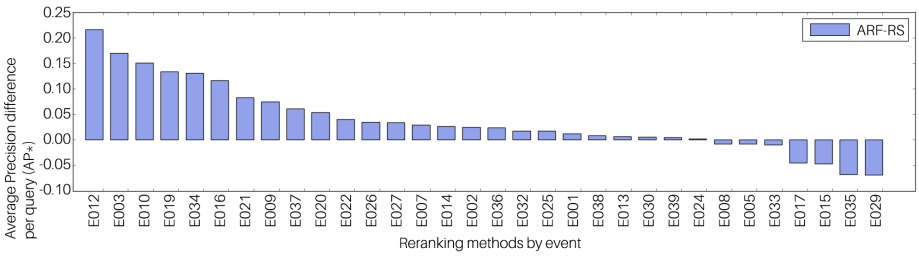


Figure 4.5: Average precision difference (AP^*) per event

Giving an example of results of the methods, Figure 4.6 shows the different results from the video level methods.

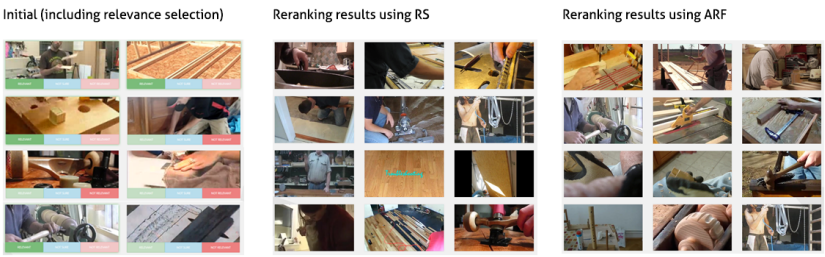


Figure 4.6: Example of returned results for the query *Working on a woodworking project*. The initial result set on the left also shows relevance selection

Table 4.3 shows the weights of the top 5 concepts for the baseline and the best method for the concept level and video level for the event *Attempting a board trick*. These results show that the manual annotator is able to capture all type of board tricks, such as skateboard, surfboard and snowboard tricks. AlterWeights does not have the general concept *attempt* or two board concepts as the No Feedback, but added the concepts *flipping* (highly relevant) and *board game* (semantically discussable relevant). ARF also has the concept *board game*, even on top of the list. This indicates that the detector has relevance for this event. The concept *attempt* is moved to the bottom of the list.

Table 4.3: Comparison Concepts and weights for the event *Attempting a board trick*

No Feedback (0.19)		**Manual (0.31)**		**AlterWeights (0.20)**		**ARF (0.25)**	
c	c_s	c	c_s	c	c_s	c	c_s
attempt	0.65	skateboardtrick	0.33	trick	0.88	board game	0.72
trick	0.63	surf	0.33	board2	0.81	skateboardtrick	0.54
board1	0.58	snowboard	0.33	skateboardtrick	0.76	board1	0.43
board2	0.58			board game	0.74	trick	0.38
skateboardtrick	0.54			flipping	0.44	attempt	0.13

4.6. CONCLUSIONS AND FUTURE WORK

Results show that 1) relevance feedback on both concept and video level improves performance compared to using no relevance feedback; 2) relevance feedback on video level obtains higher performance compared to relevance feedback on concept level; 3) our proposed ARF method on video level outperforms a state of the art k-NN method, and all methods on concept level and even manual selected concepts.

Our results are, however, bound to few events and few users. For the concept level method, we also use an indirect performance metric, because we obtain performance on video level. We, thus, do not take into account that relevant concepts can have poorly performing detectors. We believe that these experiments clearly show that although concept level user feedback can improve performance upon the initial ranking, video level user feedback is more valuable. One reason might be that this feedback can provide information on both the semantic relevance of the concept and the accuracy of the concept detector. In future work it might be interesting to investigate if we can distinguish whether the concept detector is not accurate or whether the concept is not semantically related based on the video level feedback.

5

QUERY INTERPRETATION—AN APPLICATION OF SEMIOTICS IN IMAGE RETRIEVAL

Based on: **Maaike H.T. de Boer, Paul Brandt, Maya Sappelli, Laura M. Daniele, Klamer Schutte, Wessel Kraaij** (2015) *Query Interpretation —an Application of Semiotics in Image Retrieval.* In: International Journal On Advances in Software, volume 8, number 3 and 4, pp 435 - 449.

* The translations from semiotic structures to ConceptNet relations have been performed by Paul Brandt and are not included in this chapter. The translations and related work on semiotics are available in the journal paper stated above

Previous chapters have focused on the events in the TRECVID MED task. Users of our aimed search capability will, however, not only search for events. They might want to ask specific questions, such as find the pink Cadillac in Amsterdam. *Some of these queries might suffer from the semantic gap or the vocabulary mismatch. The Concept Bank might for example contain a car, but not a Cadillac. In this chapter, we focus on the influence of the type of query in the query-to-concept mapping. This chapter is related to* **RQ4 Semiotics**. *We explore to what extent semiotic structures contribute to the semantic interpretation of user queries. Semiotics is about how humans interpret signs, and we use its text analysis structures to guide the query-to-concept mapping. Examples are paradigms, which signify functional contrasts in words, such as 'man' and 'woman', and syntagms, which signify positional contrasts of words in sentences, such as 'the ship that banked' and 'the bank that shipped'. These semiotic structures can be related to certain relations present in ConceptNet, such as* MemberOf *for paradigms and* CapableOf *or* Causes *for syntagms. In our experiments, we show that semiotic structures can contribute to a significantly higher semantic interpretation of user queries and significantly higher image retrieval performance, measured in quality and effectiveness and compared to a baseline with only synonym expansions.*

5.1. INTRODUCTION

More and more sensors connected through the Internet are becoming essential to give us support in our daily life. In such a global sensor environment, it is important to provide smart access to sensor data, enabling users to search semantically in this data in a meaningful and, at the same time, easy and intuitive manner. For visual data, an impediment to this achievement is "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation", coined by Smeulders et al. (2000) as the semantic gap in content based image retrieval (CBIR). Towards this aim, we developed a search engine that combines CBIR, Human Media Interaction and Semantic Modelling techniques in one single application: 'Google®for sensors' or 'GOOSE' for short. An overview paper of the GOOSE application is given in Schutte et al. (2015a) and Schutte et al. (2013). This application is able to retrieve visual data from multiple and heterogeneous sources and sensors, and responds to the assumption that the semantic gap consists of two parts (Enser et al., 2007): the first part addressing the realm where raw image pixels are transformed into generic objects to which labels are applied to represent their content; the second part addressing the realm of semantic heterogeneity, representing the semantic distance between the object labelling and the formulation by the end-user of a query that is meant to carve out a part in reality that situates that object. The GOOSE approach to closing the first part addresses image classification and quick image concept learning, presented in Bouma et al. (2015), and fast re-ranking of visual search results, presented in Schavemaker et al. (2015). This chapter addresses the second part of the semantic gap and builds on our earlier work on applying semantic reasoning in image retrieval (Boer et al., 2015a) that is realized through query parsing, concept expansion, and mapping it to labels associated with certain classifiers. Query concepts that do not match any classifier's label are expanded using an external knowledge base, in this case ConceptNet (Speer et al., 2012), to find alternative concepts that are semantically similar to the original query concepts but do match with a classifier label. Whereas in Boer et al. (2015a) we only used 'IsA' and 'Causes' relations in the query expansion, in this work we address additional types of relations in order to improve the matching rate between query concepts and classifier labels. However, a drawback of considering additional relations is that the algorithms for their semantic interpretation become tightly coupled to the particular external knowledge base of choice, rendering them less applicable for other knowledge bases. To overcome this limitation and keep our semantic interpretation generically applicable to other knowledge bases, we introduce the use of semiotics that provides guidance to how humans interpret signs and how the abstract relationships between them apply. Due to its universal application, a semiotic approach not only provides us with the flexibility to use different knowledge bases than ConceptNet, but it is also independent from domain-specific terminologies, vocabularies and reasoning. By defining a simple mapping from the specific relationships of the knowledge base of choice, e.g., ConceptNet, onto semiotic structures, the semantic interpretation algorithms can latch onto the semiotic structures only. The resulting transparency between the

semantic interpretation algorithms at the one hand, and at the other hand abstracting from the specifics of

i the relationships that are available in the external knowledge base, and

ii the domain-specific vocabularies, bring about the required general applicability of our solution.

Summarizing, we seek to improve the matching rate between query concepts and classifier labels, by

1. considering more, if not all, relations that are available in a knowledge base; while remaining
2. as independent to the external knowledgebase as possible; and
3. as computationally lean as possible.

We formulate our research question as

To what extent can semiotic structures contribute to the semantic interpretation of user queries?

In order to answer our research question, we conducted an experiment on our TOSO dataset (Schutte et al., 2015b), which contains 145 test images and 51 trained classifiers. For evaluation purposes we furthermore defined 100 user queries. We annotated these user queries with their ground truth for both parts of the semantic gap:

i the ground truth for semantic matching, identifying the classifier labels that are meant to be found for each user query , and

ii the ground truth for the image retrieval, identifying the images that are meant to be found.

The queries, annotations and images are available at DOI: 10.13140/RG.2.1.3688.9049. For different types of semiotic structures we calculated the effectiveness and quality in terms of different types of F-measure for both semantic matching and image retrieval. From the results of these experiments, we can conclude that applying semiotic relations in query expansion over an external, generic knowledge base, contributes to a high quality match between query concepts and classifier labels. It also significantly improves image retrieval performance compared to a baseline with only synonym expansions. Some relations that are present in ConceptNet could not be assigned to the applied semiotic structures; inclusion of these relations in the semantic analysis provided for higher effectiveness at the cost of losing loose coupling between these relations and the algorithms that implement the semantic analysis. The main contribution of this chapter is a generic approach to the expansion of user queries using general-purpose knowledge bases, and how semiotics can guide this expansion independently from the specific knowledge base being used. This chapter is structured as follows: Section 5.2 describes related work on query expansion and

semiotics; Section 5.3 provides an overview of the generic semantic interpretation system; Section 5.4 describes the experiment that has been performed with the application, followed by a presentation and discussion of their results in Sections 5.5 and 5.6, respectively. We conclude our work, including indications for future work, in Section 5.7.

5.2. RELATED WORK

In this section we discuss related work in CBIR about the first part of the semantic gap, i.e., automatic classifier annotation, as well as the second part of the semantic gap, i.e., some efforts related to query expansion using semantic relations. Finally, we discuss related work on computational semiotics.

5.2.1. AUTOMATIC IMAGE ANNOTATION

Most of the effort in applying semantics in CBIR is aimed at training classifiers using large sources of visual knowledge, such as ImageNet (Deng et al., 2009) and Visipedia (Perona, 2010). The trained classifiers are subsequently annotated with one or more labels that should describe their meaning. However, these annotations are often subjective, e.g., influenced by the domain of application and not accurate from a semantic point of view. Consequently, users that apply these classifiers need to have prior knowledge about the context of use of the annotations. In order to overcome this issue and facilitate the use of classifiers without the need of training, various efforts in the literature focus on improving the annotations. These efforts mainly apply domain-specific ontologies as basis for annotation, such as the ontologies in Bai et al. (2007) and Bagdanov et al. (2007) that are used to annotate soccer games, or for the purpose of action recognition in a video surveillance scenario (Oltramari et al., 2012). Although these approaches provide more intuitive semantics that require less prior knowledge from the user, they are tailored to specific domains and cannot be re-used for general-purpose applications.

5.2.2. RELATION-BASED QUERY EXPANSION

Several systems proposed in the literature address query expansion exploiting relations with terms that are semantically similar to the concepts in the query (Erozel et al., 2008; Chen et al., 2013; Boer et al., 2015b). The system in Erozel et al. (2008) facilitates natural language querying of video archive databases. The query processing is realized using a link parser (Sleator et al., 1995) based on a light-parsing algorithm that builds relations between pairs of concepts, rather than constructing constituents in a tree-like hierarchy. This is sufficient for the specific kind of concept groups considered in the system (Erozel et al., 2008), but is limitative for more complex queries. The Never Ending Image Learner (NEIL) proposed in Chen et al. (2013) is a massive visual knowledge base fed by a crawler that runs 24 hour a day to extract semantic content from images on the Web in terms of *objects*, *scenes*, *attributes* and their*relations*. The longer NEIL runs, the more relations between concepts detected in the images it learns. Analogously to our approach, NEIL is a general-purpose system and is based on learning new concepts and relations that are then used to aug-

ment the knowledge of the system. Although NEIL considers an interesting set of semantic relations, such as taxonomy (*IsA*), partonomy (*Wheel is part of Car*), attribute associations (*Round_shape is attribute of Apple* and *Sheep is White*), and location relations (*Bus is found in Bus_depot*), most of the relations learned so far are of the basic type 'IsA' or 'LooksSimilarTo'. Furthermore, in Boer et al. (2015b) knowledge bases ConceptNet and Wikipedia, and an expert knowledge base are compared for semantic matching in the context of multimedia event detection. Results show that query expansion can improve performance in multimedia event detection, and that the expert knowledge base is the most suitable for this purpose. When comparing Wikipedia and ConceptNet, ConceptNet performs slightly better than Wikipedia in this field. In their comparison, the authors only considered query expansion using the ConceptNet 'IsA' relation.

5.2.3. SEMIOTICS IN CBIR

Although text analysis is its primary field of application, recently semiotics gained the interest in the field of ICT. The application of semiotics in computer science is best illustrated with the emergence of computational semiotics, where a clear starting point for its definition is the fact that signs and sign systems are central to computing: manipulation of symbols applies to everything that happens in computer science, from user interfaces to programming and conceptual modelling alike. In relation to CBIR, many studies, summarized by Enser et al. (2007), accept the existence of 'semantic layers' in images. Every layer provides for another abstraction and aggregation of the things that are being denoted. The referenced studies address these layers as distinct realms, and act accordingly by constraining themselves to one layer. However, semioticians address these layers as a whole, and study it as a process to which they refer as *unlimited semiosis*. We are inspired by that approach and therefore part of our work considers unlimited semiosis as algorithmic foundation when addressing these layers. Application of semiotics in CBIR and especially about user query interpretation is very limited, and the following two studies represent, to the best of our knowledge, good examples of its main focus. Yoon (2006) has investigated the association between denotative (literal, definitional) and connotative (societal, cultural) sense-making of image meta-data in support of image retrieval. This approach is similar to ours in that it is based on semiotic structures to bridge the semantic gap. Although the results are promising, it cannot be applied in our generic context due to the domain-specific foundations that are implicit to connotations. Closely related to it, Hartley (2004) studies how semiotics can account for image features that characterize an audio, visual or audio-visual object, in order to facilitate visual content description or annotation. Their model integrates low-level image features such as color and texture together with high-level denotative and connotative descriptions. This approach differs with ours in that they do not make a distinction between the two cascading parts of the semantic gap, but instead take an integrated approach.

5.3. GENERIC SEMANTIC REASONING SYSTEM

Figure 5.1 shows an overview of the semantic reasoning parts of the GOOSE system in which green and blue parts represent the components that realize the semantic reasoning, yellow parts represent the components dedicated to the image classification task and the white parts represent external components. The image classification task, which is elaborated in Bouma et al. (2015), captures the semantics of visual data by translating the pixels from an image into a content description (which could be a single term), referred to as *annotated images*.

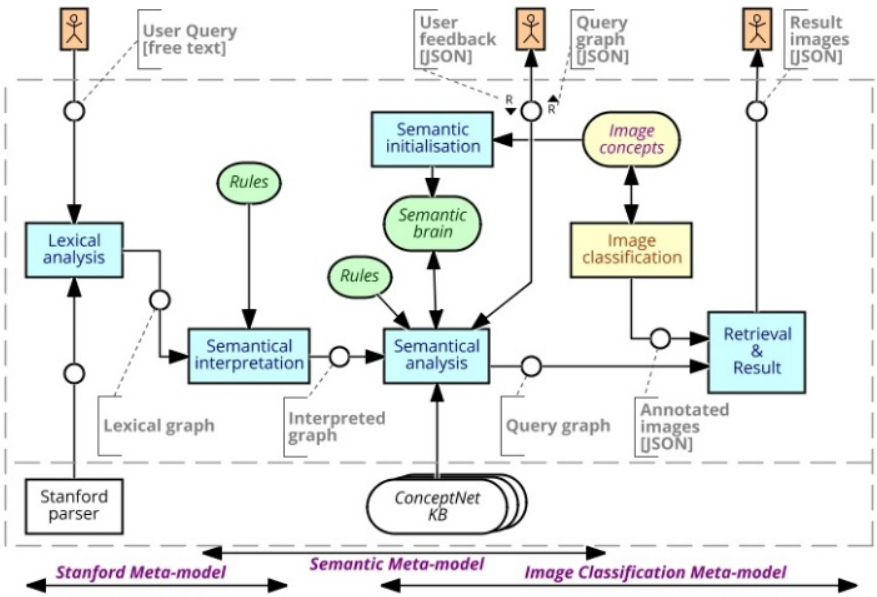


Figure 5.1: System overview

The semantic reasoning starts with a user query in natural language. The query is processed by four modules, while a fifth module takes care of initializing the system and learning new concepts. In the first stage, the query is sent to the *Lexical Analysis* module that parses it using the Stanford Parser (De Marneffe et al., 2006). The Stanford Parser returns a lexical graph, which is used as input to the *Semantic Interpretation* module. In this module, a set of rules is used to transform the lexical elements of the Stanford meta-model into semantic elements of the intermediary ontology that represents objects, attributes, actions, scenes and relations. The interpreted graph is sent to the *Semantic Analysis* module that matches the graph nodes against the available image concepts. If there is no exact match, the query is expanded using an external knowledge base, i.e., ConceptNet, to find a close match. The interpretation resulting from the Semantic Analysis is presented as a query graph to the user. The query graph is also used as input for the *Retrieval and Result* module, which provides the final result to the user. In the following subsections the complete process is de-

scribed in detail using the sample query *find a red bus below a brown animal*. In this particular query, its positional part, e.g., *below*, should be understood from the viewpoint of the user posing the query, i.e., the relative positions of the 'red bus' and the 'brown animal' as shown in the user's screen.

5.3.1. SEMANTIC INITIALIZATION

This module provides an initial semantic capability by populating the Semantic Brain, which holds all *image concepts* that are known to the system. Image concepts are represented as instances of the meta-model (discussed in section 5.3.3), and refer to those things that the image classification task is capable of detecting. This component also handles updates to the Semantic Brain following from new or modified image classification capabilities and semantic concepts.

5.3.2. LEXICAL ANALYSIS

In the Lexical Analysis module, the user query is lexically analyzed using the Typed Dependency parser (englishPCFG) of Stanford University (De Marneffe et al., 2006). Before parsing the query, all tokens in the query are converted to lower case. In the example of *find a red bus below a brown animal*, the resulting directed graph from the Lexical Analysis is shown in Figure 5.2.

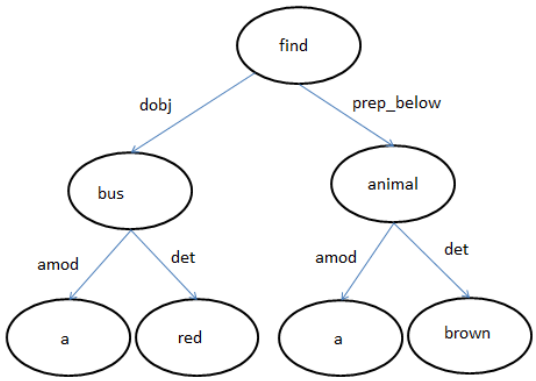


Figure 5.2: Lexical Graph

5.3.3. SEMANTIC INTERPRETATION

Since GOOSE is positioned as a generic platform, its semantics should not depend on, or be optimized for, the specifics of one single domain of application. Instead, we apply a generic ontological commitment by defining a semantic meta-model, shown in Figure 5.3, which distinguishes objects that might

i bear attributes (*a yellow car*),

ii take part in actions (*a moving car*),

iii occur in a scene (*outside*), and

iv have relations with other objects, in particular ontological relations (*a vehicle subsumes a car*), spatial relations (*an animal in front of a bus*), and temporal relations (*a bus halts after driving*).

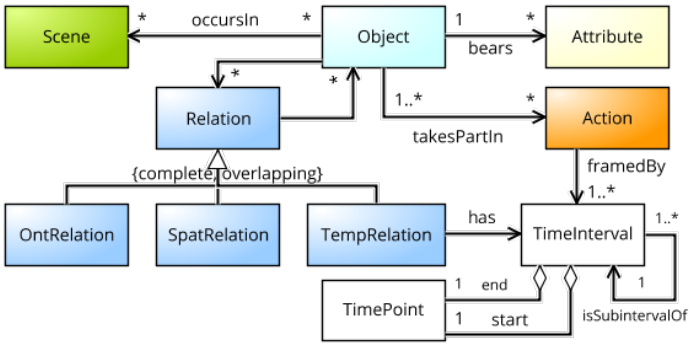


Figure 5.3: Semantic meta-model

In the Semantic Interpretation module, a set of rules is used to transform the elements from the lexical graph into *objects*, *attributes*, *actions*,*scenes* and *relations*, according to the semantic meta-model in Figure 5.3. These rules include the following examples:

- Derive *cardinality* from a *determiner* (*det* in Figure 5.2), e.g., *the* in a noun in the singular form indicates a cardinality of 1, while *a/an* indicates at least 1;
- Derive *attributes* from *adjectival modifiers* (*amod* in Figure 5.2), i.e., adjectival phrases that modify the meaning of a noun;
- Derive *actions* from *nominal subjects* and *direct objects* (*nsubj* and *dobj* in Figure 5.2), i.e., the subject and object of a verb, respectively;
- Actions that represent the query command, such as *find*, *is*, *show* and *have*, are replaced on top of the tree by the subject of the sentence.

The output of the Semantic Interpretation for the sample query *find a red bus below a brown animal* is shown in Figure 5.4.

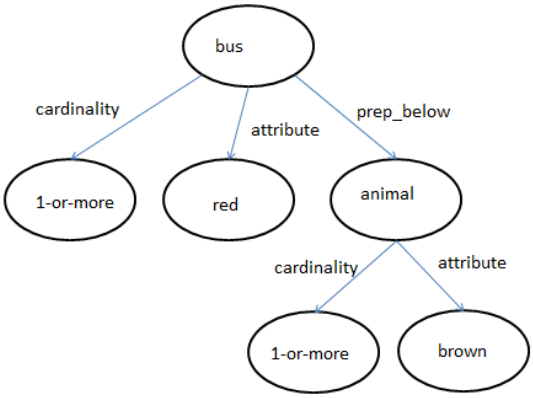


Figure 5.4: Interpreted Graph

5.3.4. SEMANTIC ANALYSIS

The purpose of the Semantic Analysis is to align the elements from the interpreted graph, which are the query concepts, with the image concepts that are available in the Semantic Brain. For those objects, actions, scenes or attributes from the graph that do not have a syntactical identical counterpart ('exact match') in the Semantic Brain, and hence cannot be recognized by the image classification component, the query concepts are expanded into *alternative concepts* using an external general-purpose knowledge base. We use the external knowledge base ConceptNet to find these alternative concepts. The alternative concepts are dependent on the semiotic structure that is used (further explained in the Experiment section). An example of this method is shown in Figure 5.5. Unlimited semiosis expand to more abstract and more specific concepts. Paradigms expand to disjoint concepts and syntagms expand to functionally related concepts. Our principle of genericity and loose coupling, however, facilitates the use of other or even more knowledge bases without the need to adapt the semantic analysis algorithms.

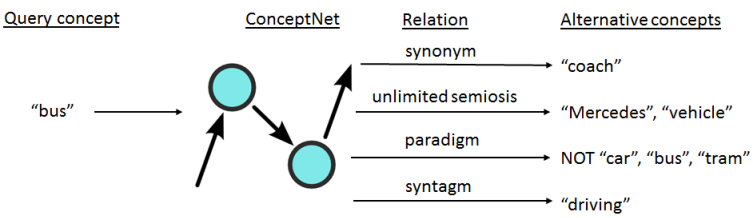


Figure 5.5: Example of Semantic Analysis

5.3.5. RETRIEVAL AND RESULT

This module retrieves the images that, according to the classifiers, contain concepts that carry an identical label as the query concepts (or alternative concepts). Furthermore, the cardinality, attribute and spatial relations should match with the query. If the image contains too many instances, the image is still included. The spatial relations are determined by the edges of the bounding box. Because our bounding boxes are not accurate, we use a relaxed version of the prepositions. The upper left edge of the bounding box has the values [0,0]. In the preposition *left of*, the left edge of the bounding box of the right object should be right of the left edge of the bounding box of the left object, denoted as:

Left of: a.min.x <b.min.x
Right of: a.max.x >b.max.x
On top of: a.min.y <b.min.y
Below: a.max.y >b.max.y
And: a *and* b

5.4. EXPERIMENT

In order to answer our research question **To what extent can semiotic structures contribute to the semantic interpretation of user queries?** we conducted an experiment. In this experiment we measure effectiveness and quality of different semiotic structures on the level of both semantic matching and image retrieval. The variable of the experiment is therefore represented by the differences in query expansion strategy, their core being the semiotic structures. These semiotic structures include unlimited semiosis (more abstract and more specific concepts), paradigms (disjoint concepts, i.e. 'man' vs 'woman') and syntagms (functionally related concepts, such as 'car' or 'bike' for the concept 'driving'). The experiment context is defined by our TOSO dataset and 100 manually defined queries. More information on the TOSO dataset can be found in subsection 5.4.1. The type of queries can be found in subsection 5.4.2. The experiment variations and its baseline are explained in subsection 5.4.3. The design of the experiment if presented in subsection 5.4.4, and its evaluation is explained in subsection 5.4.5.

5.4.1. DATASET

The TOSO dataset (Schutte et al., 2015b) consists of 145 images of toys and office supplies placed on a table top. In these images multiple objects can be present in several orientations as well as objects of the same type with different colors. In Figure 5.6 a sample of the dataset has been depicted. Examples of these objects are different types of cars, a bus, an airplane, a boat, a bus stop, a traffic light, different types of traffic signs, barbies with different colored dresses, different colored plants, a water bottle, a screwdriver, a hamburger and a helmet. For this dataset 40 relevant object classifiers, trained on table top images, are available as well as 11 attribute classifiers, which are colors. The object classifiers are trained with a recurrent deep convolutional neural network that uses a second stage classifier (Bouma et al., 2015). The colors are extracted using Van De Weijer et al. (2009).

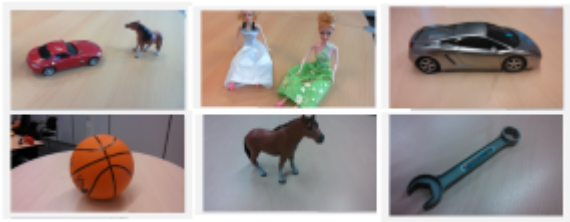


Figure 5.6: A sample of the TOSO dataset

5.4.2. QUERIES

In this experiment, we created 100 queries. In the definition of the queries we used our prior knowledge of the available classifiers by intentionally choosing interesting expansions, for example their synonyms or hypernyms. This was done by searching online thesauri, independently from our ConceptNet example. In this way, we created a set of queries that does not have direct matches to the available classifiers, but for which the use of semiotic structures could be helpful. These queries are divided into five equal groups based on their semiotic or semantic structure as follows:

1. Synonym: synonyms of our labels;
   find the auto (classifier label: car);

2. Unlimited semiosis: hyponyms or hypernyms, i.e. parents or children of a label or suspected part of relations;
   find the Mercedes (classifier label: car)
   find the animal (classifier label: giraffe)
   find the leaf (classifier label: plant)

3. Paradigm: excluding brothers and sisters in the graph (man vs. woman), restrictions to objects by color and/or spatial relations;
   find the air vehicle (as opposed to land vehicle, e.g., car, bus, tram);
   find the red sign on the right of the yellow car;

4. Syntagm: actions and properties related to our labels;
   find the things landing (classifier label: airplane);
   find the expensive things (classifier labels: airplane, car);

5. Other: words which have a less clear or vague relation with a classifier label:
   find the flower pot (classifier label: plant)
   find the traffic jam (classifier label: cars)

For each of the queries, we established a semantic ground truth as well as an image ground truth. The semantic ground truth was established by manually annotating for each classifier label in our classifier set whether it is irrelevant (0) or relevant (1) to the query. In our annotation, a classifier is *relevant* if (i) a classifier label

is syntactically similar to a concept in the query, or (ii) a classifier label represents a synonym of a query concept. For the image ground truth we used the 145 test images from the TOSO dataset. An external annotator established the ground truth by annotating, for each query and for each image, whether the image was irrelevant (0) or relevant (1) to the query. Establishing relevancy was left to the annotator's judgement. For both the semantic and image annotations, the instructions indicated that all cases of doubt should be annotated as relevant (1).

5.4.3. EXPERIMENTAL VARIABLE

In the experiment, we compare the following query expansion methods:

1. SYNONYM (baseline)
2. UNLIMITED SEMIOSIS
3. PARADIGM
4. SYNTAGM
5. ALL

In the first method, which represents our baseline, we use the basic expansion over specific relations that are found in ConceptNet: *Synonym* and *DefinedAs*. In the other methods we use the baseline relations as well as their specific semiotic relations. The translation between a certain semiotic relation and the ConceptNet relations is explained in our journal paper. In the second method (UNLIMITED SEMIOSIS), we use the following relations *IsA*, *hasSubEvent*, *PartOf* and *HasA* from ConceptNet. The 'IsA' and 'PartOf' relations are directed towards more abstract concepts, whereas the 'HasSubEvent' and 'HasA' relations are directed towards the more specific concepts. In the third method (PARADIGM), we consider *MemberOf* and *DerivedFrom* as paradigmatic relations. In the fourth method (SYNTAGM), we use *CapableOf*, *UsedFor*, *CreatedBy* to reflect transitions from objects to actions; *Causes* to reflect a transition from object to action; *hasProperty* from object to property. In the fifth method (ALL) all possible relations from ConceptNet, excluding *TranslationOf* and *Antonym*, are applied for query expansion.

5.4.4. EXPERIMENT DESIGN

The design of the experiment is based on the hypothesis that a query will be served best by a query expansion strategy that shares its semiotic structures, e.g., the SYNTAGM expansion method will find most mappings for syntagm queries and perform worse for other queries. Each expansion method from the previous section, 5.4.3, will apply its one single expansion strategy over all query groups from section 5.4.2; different methods will therefore perform differently, i.e., result in different mapping counts. In order to test our hypothesis, we designed and ran two evaluation cases. The first evaluation case addresses the part of the semantic gap that is about *semantic matching*. This case shows the impact of using semiotic structures on the effectiveness and quality of the mapping from the query to the classifier labels. The second

evaluation case addresses the part of the semantic gap that is about *image retrieval*. This case shows the impact of semiotic structures on the effectiveness and quality of a full general-purpose image search engine.

5.4.5. EVALUATION CRITERIA

In our evaluations we calculate the effectiveness and quality in terms of different types of F-measure for each query from 5.4.2. The following provides more detail for each evaluation case.

SEMANTIC MATCHING

In order to show the result of the expansion method on the mapping from the query to the classifier labels, we compare the result of each of the methods against the ground truth. This result is a list of classifier labels that are found by searching ConceptNet using the relations that are characteristic for the subject expansion method. In the evaluation we use two kind of metrics, corresponding to quality and effectiveness. The typical metric for quality is using precision, denoted P_{sg}, which takes into account the number of true positives, i.e, found and annotated as relevant labels, and the total number of found labels, i.e., true positives and false positives, denoted as TP and FP, respectively:

$$P_{sg} = \frac{1}{n} * \sum_{q=1}^{n} \frac{TP_{sg}}{TP_{sg} + FP_{sg}} \tag{5.1}$$

, where n denotes the total number of queries.

The typical metric for measuring effectiveness is recall, denoted R_{sg}, which takes into account the number of correctly found labels, i.e. true positives and the total number of relevant labels, i.e., true positives and false negatives, the latter denoted as FN:

$$R_{sg} = \frac{1}{n} * \sum_{q=1}^{n} \frac{TP_{sg}}{TP_{sg} + FN_{sg}} \tag{5.2}$$

, where n denotes the total number of queries.

Precision and recall are always an interplay, so we decided to not use precision and recall separately, but combine them by means of applying the F-measure. Since different applications can value the precision and recall of the semantic matching differently, the Fβ-measure can be used to express that one should attach β times as much value to the recall results of the semantic matching than to its precision results. By using the Fβ-measure as our primary means of evaluation, we can show the impact of the experiment results on three classes of applications, i.e., high quality applications that value precision over recall, high effectiveness applications that value recall over precision, and neutral applications that value precision equally important as recall. The Fβ-measure is defined as:

$$F\beta = (1 + \beta^2) * \frac{P_{sg} * R_{sg}}{(\beta^2 * P_{sg}) + R_{sg}} \tag{5.3}$$

For high quality applications, we put 10 times more emphasis on the precision and choose to use $\beta = 0$. For neutral applications we use the basic F-measure, i.e., $\beta = 1$ and for high effectiveness applications, we value recall 10 times more than precision and use $\beta = 10$. Naturally, these choices for β are made in order to show relative trends as opposed to an absolute judgement.

IMAGE RETRIEVAL

The annotations are used in a similar way as on the level of the semantic matching. Again, F-score with $\beta = 0.1$ is used for high quality applications, $\beta = 1$ for neutral applications and $\beta = 10$ for high effectiveness applications.

5.5. RESULTS

In this section, we show the results of our experiment. The sections have the same structure as section 5.4.5, so the first section explains the results about the semantic matching and the second section is about the results of the image retrieval. For each of the evaluations, the assumption of normality was violated, as indicated by significant Kolmogorov-Smirnov statistics. We present non-parametric Friedman-tests and Wilcoxon Signed-Ranks Tests to compare the different methods.

5.5.1. SEMANTIC MATCHING

HIGH PRECISION SYSTEM ($\beta = 0.1$)

Graph 5.7 shows the F-score for the high precision system for each of the methods for each type of query group with the confidence interval of 95%. For two queries, both in group 4, no relevant annotation was available, so in group 4 analysis is done with 18 queries instead of 20 and in total 98 queries were analyzed.

A Friedman test showed a statistically significant difference among the methods ($\chi^2(4)=57.938$, $p < .001$). Wilcoxon Signed-Ranks Test were used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a .01 level of significance (.05/5 conditions). The results can be found in Figure 5.8.

The order of overall performance is thus SYNONYM + PARADIGM >SYNTAGM >UNLIMITED SEMIOSIS >ALL, all significant differences. For group 1 no significant differences between SYNONYM and the other methods are found. For group 2 significant differences between UNLIMITED SEMIOSIS and SYNONYM (Z=-3.550,p<.001), SYNTAGM (Z=-3.432,p=.001) and PARADIGM (Z=-3.651,p<.001) are found. For group 3 no significant differences between PARADIGM and the other methods are found. For group 4 significant differences between SYNTAGM and SYNONYM (Z=-2.670, p=.008) and PARADIGM (Z=-2.670,p=.008) are found. For group 5 significant differences between ALL and SYNONYM (Z=-3.053,p=.002), SYNTAGM (Z=-2.833,p=.005) and PARADIGM (Z=-3.053,p=.002) are found.

NEUTRAL SYSTEM ($\beta = 1$)

Graph 5.9 shows the F-score for the neutral system for each of the methods for each type of query group with the confidence interval of 95%.

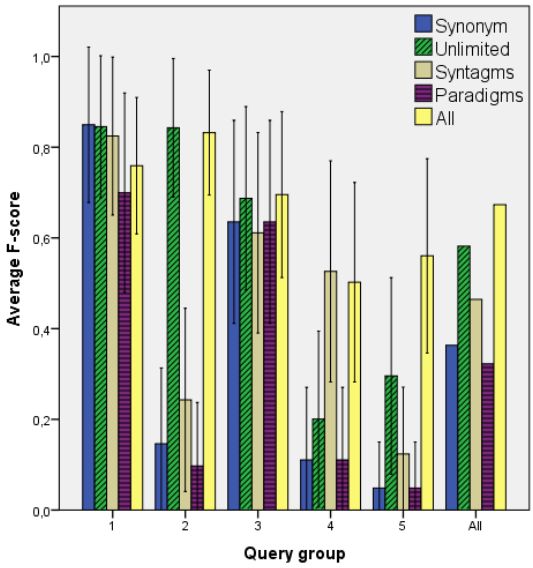


Figure 5.7: F-score Semantic Graph for High Quality

	SYNO	UNL	SYNT	PARA	ALL
SYNO		Z=-4.458, p<0.001*	Z=-3.128, p=0.02*	Z=-1.890, p=0.059	Z=-5.224, p<0.001*
UNL	Z=-4.458, p<0.001*		Z=-2.401, p=0.016*	Z=-4.859, p<0.001*	Z=-2.162, p=0.031*
SYNT	Z=-3.128, p=0.02*	Z=-2.401, P=0.016*		Z=-3.635, p<0.001*	Z=-3.985, p<0.001*
PARA	Z=-1.890, p=0.059	Z=-4.859, p<0.001*	Z=-3.635, p<0.001*		Z=-5.635, p<0.001*
ALL	Z=-5.224, p<0.001*	Z=-2.162, p=0.031*	Z=-3.985, p<0.001*	Z=-5.635, p<0.001*	

Figure 5.8: F-score All Semantic Graph Wilcoxon for High Quality

A Friedman test showed a statistically significant difference among the methods ($\chi^2(4)$=98.571, p < .001). Wilcoxon Signed-Ranks Test were used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a .01 level of significance (.05/5 conditions). The results can be found in Figure 5.10. The order of overall performance is thus equal to the performance for the high quality system. The same significant differences are found for query group 1, 2 and 5. For group 3 significant differences between PARADIGM and UNLIMITED SEMIOSIS

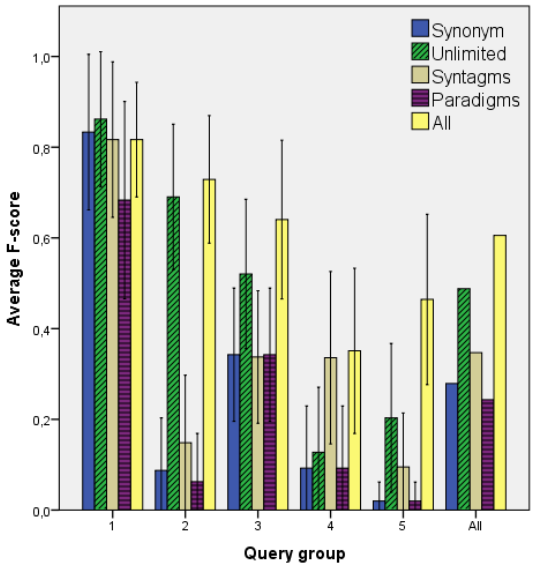


Figure 5.9: F-score Semantic Graph for Neutral

	SYNO	UNL	SYNT	PARA	ALL
SYNO		Z=-5.050, p<0.001*	Z=-3.043, p=0.02*	Z=-1.890, p=0.059	Z=-6.049, p<0.001*
UNL	Z=-5.050, p<0.001*		Z=-3.276, p=0.001*	Z=-5.366, p<0.001*	Z=-3.535, p<0.001*
SYNT	Z=-3.043, p=0.02*	Z=-3.276, p=0.001*		Z=-3.576, p<0.001*	Z=-5.329, p<0.001*
PARA	Z=-1.890, p=0.059	Z=-5.366, p<0.001*	Z=-3.576, p<0.001*		Z=-6.434, p<0.001*
ALL	Z=-6.049, p<0.001*	Z=-3.535, p<0.001*	Z=-5.329, p<0.001*	Z=-6.434, p<0.001*	

Figure 5.10: F-score All Semantic Graph Wilcoxon for Neutral

(Z=-2.805,p=.005) and ALL (Z=-3.237,p=.001) exist, as well as significant differences between UNLIMITED SEMIOSIS and SYNONYM (Z=-2.805,p=.005), PARADIGM (Z=-2.805,p=.005) and SYNTAGM (Z=-2.926,p=.003). For group 4 an additional significant difference between SYNTAGM and UNLIMITED SEMIOSIS (Z=-2.603,p=.009) is found.

HIGH EFFECTIVENESS SYSTEM (β = 10)

Graph 5.11 shows the F-score for the high effectiveness system for each of the methods for each type of query group with the confidence interval of 95%.

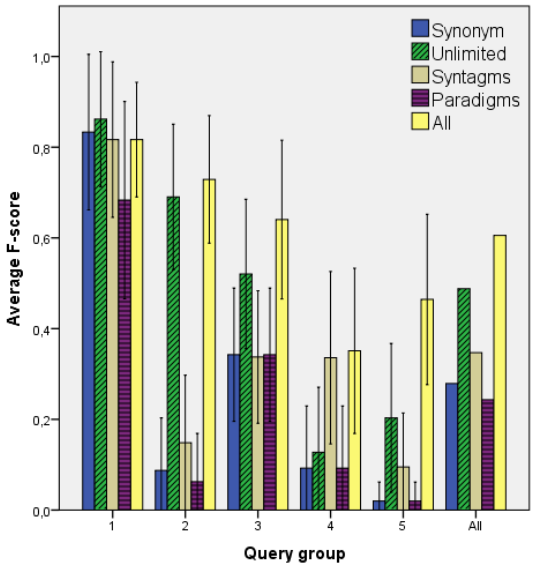


Figure 5.11: F-score Semantic Graph for High Effectiveness

	SYNO	UNL	SYNT	PARA	ALL
SYNO		Z=-5.230, p<0.001*	Z=-3.241, p=0.001*	Z=-1.890, p=0.059	Z=-6.664, p<0.001*
UNL	Z=-5.230, p<0.001*		Z=-3.524, p<0.001*	Z=-5.521, p<0.001*	Z=-4.815, p<0.001*
SYNT	Z=-3.241, p=0.001*	Z=-3.524, p<0.001*		Z=-3.716, p<0.001*	Z=-6.083, p<0.001*
PARA	Z=-1.890, p=0.059	Z=-5.521, p<0.001*	Z=-3.716, p<0.001*		Z=-6.896, p<0.001*
ALL	Z=-6.664, p<0.001*	Z=-4.815, p<0.001*	Z=-6.083, p<0.001*	Z=-6.896, p<0.001*	

Figure 5.12: F-score All Semantic Graph Wilcoxon for High Effectiveness

A Friedman test showed a statistically significant difference among the methods (χ^2 (4) =108.197, p< .001). Wilcoxon Signed-Ranks Test were used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a .01 level of significance (.05/5 conditions). The results can be found in Figure 5.12. The order of overall performance is thus equal to the performance for both the high quality and neutral system. The same significance values are found as for the neutral system, except for the significant difference between PARADIGM and ALL in group 3. This difference is no longer significant for the high effectiveness system.

5.5.2. IMAGE RETRIEVAL

HIGH PRECISION SYSTEM ($\beta = 0.1$)

Graph 5.13 shows the F-score for the high quality system for each of the methods for each type of query group with the confidence interval of 95%. For 14 queries of which one in group 1, three in group 4 and ten in group 5, no relevant annotation was available. In total 86 queries are analyzed.

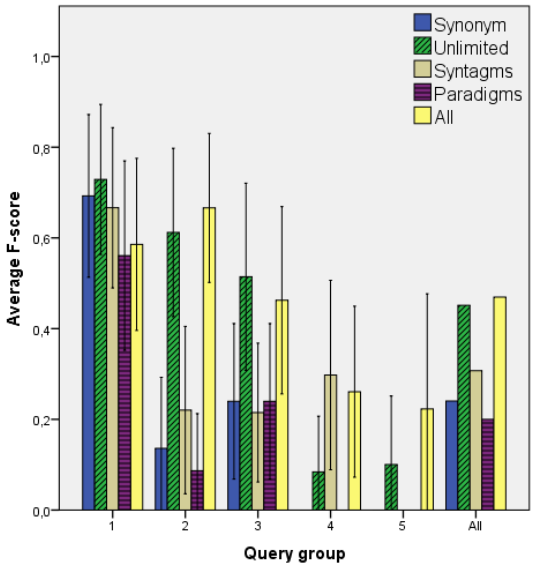


Figure 5.13: F-score Image Retrieval for High Quality

	SYNO	UNL	SYNT	PARA	ALL
SYNO		Z=-4.684, p<0.001*	Z=-2.511, P=0.012*	Z=-1.826, p=0.068	Z=-4.108, p<0.001*
UNL	Z=-4.684, p<0.001*		Z=-3.060, p=0.002*	Z=-5.012, p<0.001*	Z=-0.686, p=0.493
SYNT	Z=-2.511, p=0.012*	Z=-3.060, p=0.002*		Z=-3.155, p=0.002*	Z=-3.250, p=0.001*
PARA	Z=-1.826, p=0.068	Z=-5.012, p<0.001*	Z=-3.155, p=0.002*		Z=-4.722, p<0.001*
ALL	Z=-4.108, p<0.001*	Z=-0.686, p=0.493	Z=-3.250, p=0.001*	Z=-4.722, p<0.001*	

Figure 5.14: F-score All Image Retrieval Wilcoxon for High Quality

A Friedman test showed a statistically significant difference among the methods ($\chi^2(4)$ =58.891, p < .001). Wilcoxon Signed-Ranks Test were used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a .01 level of significance (.05/5 conditions). The results can be found in Figure 5.14. The order of overall performance is thus SYNONYM + PARADIGM >SYNTAGM >UNLIMITED SEMIOSIS + ALL, all significant differences. For group 1 no significant differences between SYNONYM and the other methods are found. For group 2 significant differences between UNLIMITED SEMIOSIS and SYNONYM (Z=-3.294,p=.001), SYNTAGM (Z=-2.982,p=.003) and PARADIGM (Z=-3.413,p=.001) are found. For group 3 significant differences between PARADIGM and UNLIMITED SEMIOSIS (Z=-2.701,p=.007) exist, as well as significant differences between UNLIMITED SEMIOSIS and SYNONYM (Z=-2.701,p=.007), PARADIGM (Z=-2.701,p=.007) and SYNTAGM (Z=-2.845, p=.004). For group 4 no significant differences between SYNTAGM and the other methods are found and for group 5 no significant differences were found.

NEUTRAL SYSTEM (β = 1)

Graph 5.15 shows the F-score for the neutral system for each of the methods for each type of query group with the confidence interval of 95%.

A Friedman test showed a statistically significant difference among the methods ($\chi^2(4)$ =71.047, p < .001). Wilcoxon Signed-Ranks Test were used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a .01 level of significance (.05/5 conditions). The results can be found in Table 5.16. The same significant differences between conditions for all and the different query groups can be found as for the high quality system.

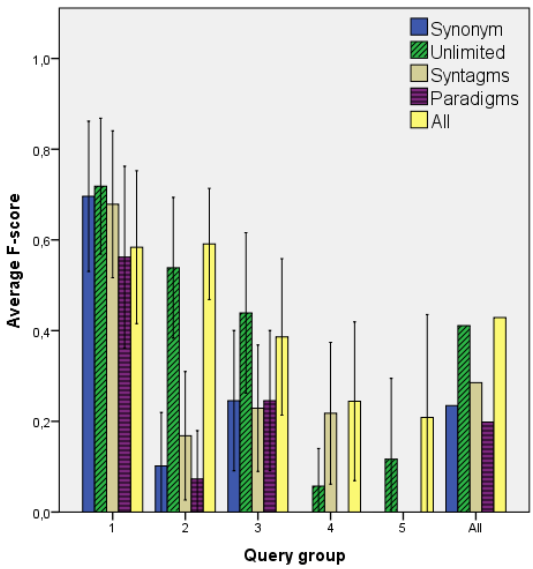


Figure 5.15: F-score Image Retrieval for Neutral

	SYNO	UNL	SYNT	PARA	ALL
SYNO		F=-4.723, p<0.001*	F=-2.354, p=0.019*	F=1.826, p=0.068	F=-4.019, p<0.001*
UNL	F=-4.723, p<0.001*		F=-3.350, p=0.001*	F=-5.045, p<0.001*	F=-0.800, p=0.424
SYNT	F=-2.354, p=0.019*	F=-3.350, p=0.001*		F=-3.052, p=0.002*	F=-3.554, p<0.001*
PARA	F=1.826, p=0.068	F=-5.045, p<0.001*	F=-3.052, p=0.002*		F=-4.704, p<0.001*
ALL	F=-4.019, p<0.001*	F=-0.800, p=0.424	F=-3.554, p<0.001*	F=-4.704, p<0.001*	

Figure 5.16: F-score All Image Retrieval Wilcoxon for Neutral

HIGH EFFECTIVENESS SYSTEM ($\beta = 10$)

Graph 5.17 shows the F-score for the high effectiveness system for each of the methods for each type of query group with the confidence interval of 95%.

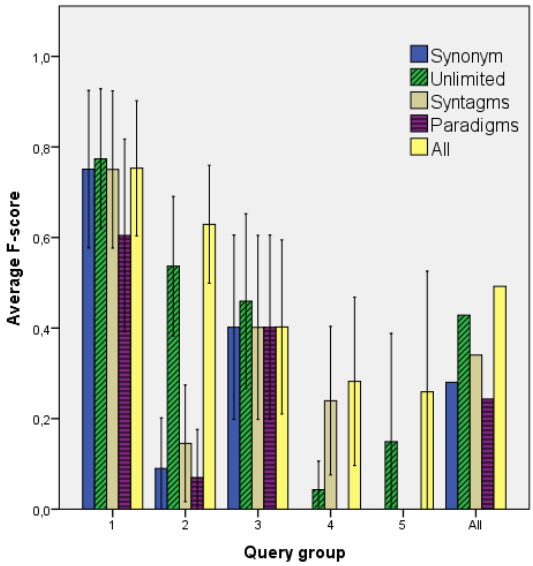


Figure 5.17: F-score Image Retrieval for High Effectiveness

A Friedman test showed a statistically significant difference among the methods ($\chi^2(4)$ =67.386, p < .001). Wilcoxon Signed-Ranks Test were used to follow up this finding. A Bonferroni correction was applied and all effects are reported at a .01 level of significance (.05/5 conditions). The results can be found in Figure 5.18. The same

	SYNO	UNL	SYNT	PARA	ALL
SYNO		F=-4.586, p<0.001*	F=-2.825, p=0.005*	F=-1.826, p=0.068	F=-4.775, p<0.001*
UNL	F=-4.586, p<0.001*		F=-2.654, p=0.008*	F=-4.930, p<0.001*	Z=-2.257, p=0.024*
SYNT	F=-2.825, p=0.005*	F=-2.654, p=0.008*		Z=-3.362, p=0.001*	Z=-4.184, p<0.001*
PARA	F=-1.826, p=0.068	F=-4.930, p<0.001*	Z=-3.362, p=0.001*		Z=-5.196, p<0.001*
ALL	F=-4.775, p<0.001*	Z=-2.257, p=0.024*	Z=-4.184, p<0.001*	Z=-5.196, p<0.001*	

Figure 5.18: F-score All Image Retrieval Wilcoxon for High Effectiveness

significant differences between conditions for all and the different query groups can be found as for the high quality and neutral system, except that we have no longer significant differences in group 3.

5.6. DICUSSION

In the discussion we reflect on the experimental results regarding the use of semiotic structure to close the semantic gap in CBIR applications, both its semantic matching and its image retrieval parts. Additionally, we discuss the limitations of this research.

5.6.1. SEMANTIC MATCHING

Results on the semantic matching show that the use of the ALL method for query expansion, e.g., taking into account each and every type of relation that is available in the knowledge base, gives best overall performance independent of the type of application you want to use, i.e., high quality, neutral or high effectiveness. This effect is mainly rooted in the *RelatedTo* relation. This relation does not reflect any semiotic structure and was hence excluded from the other methods, however does lead to alternative concepts that appear relevant to the original query concept. Examples of such relations produce expansions such as 'key fob' to 'key ring' and 'aircraft' to 'airplane', which can fuel a debate whether their relation with the query concept would not be better expressed as *synonym*. The method UNLIMITED SEMIOSIS gives the second best overall performance for all types of applications, both significantly lower than ALL and significantly higher than the other methods. As expected, this method turns the external knowledge base into a directed graph that expresses levels of aggregation, and therefore finds more abstract or more specific concepts compared to the baseline. The method SYNTAGMS has the third overall performance. This semiotic type is particularly good due to the fourth group of queries where a translation to another syntagmatic type is due for success, i.e., from *action* to *object* or from *attribute* to *object*. PARADIGMS have equal performance, or even slightly worse, com-

pared to BASELINE. Whereas our hypothesis was that it would only exclude irrelevant concepts, it also excluded some concepts that were annotated as relevant, such as 'soccer_ball' for 'ball' and 'motorbike' for 'motorcycle'. Whether these relations should be excluded or included is based on the type of application that it is used for. Alternatives that are found by this semiotic type, and that are causing a decrease in performance, are all found as *synonym* as well. Hence, a strategy might be to include all relations that have the synonym relation independent of the presence of the PARADIGMS.

A closer look at the results for the different query groups as introduced in section 5.4.2 show the following. For the first query group (*baseline*) no significant differences are found over the different expansion strategies. Surprisingly, no SYNONYM relation between 'key fob'and 'keyring', and between 'aircraft' and 'airplane' is available in ConceptNet. The first is present through a *RelatedTo* relation and the second through an *IsA* relation. Furthermore, no ConceptNet entry for 'camping bus' is present, so no expansions are found, dropping performance. Performance for PARADIGMS is lowest, because expansions for 'automobile' and 'beefburger', which are both considered relevant according to our ground truth, are paradigmatically excluded, dropping performance. SYNTAGMS is slightly lower than SYNONYM, because of an, as irrelevantly annotated, relation between 'shit' and 'cow', which is a debatable choice. UNLIMITED SEMIOSIS finds a relation between 'football' and 'skateboard', which slightly decreases performance. ALL has found several good alternatives, but also irrelevant ones, which is nicely visible in Figure 5.7 (relatively low F-score) and Figure 5.11 (relatively high F-score).

For the second query group (*unlimited semiosis*) the hypothesis was that UNLIMITED SEMIOSIS performs best. Significant differences between all other expansion methods, except ALL, are found. As the different graphs show, UNLIMITED SEMIOSIS has a better quality (Figure 5.7), whereas ALL has a higher effectiveness (Figure 5.9). The ALL method finds additional relevant concepts for 'animal' ('cow'), 'tool' ('screwdriver') and 'door' ('bus'), but irrelevant concepts for 'vehicle' ('tag') and 'leaf' ('pig'). The other methods find very little concepts and, therefore, performance is low.

The third query group (*paradigms*) is a group that expresses restrictions, such as spatial relation and color. No significant differences in performance are found for high precision applications, but for neutral and high effectiveness applications performance of UNLIMITED SEMIOSIS and ALL methods are significantly higher than PARADIGMS. This is due to relevant expansions for 'animal', 'flower', 'vehicle', 'hat', 'Mercedes' and 'Range Rover'. No results for 'air vehicle', 'water vehicle' and 'land vehicle' are found.

For the fourth query group (*syntagms*) the hypothesis was that the SYNTAGMS method performs best. As with the second query group and the UNLIMITED SEMIOSIS method, we see a high-quality for the SYNTAGMS method, but a high recall for the ALL method. The other methods have low performance, because they remain in the same syntagmatic part of the graph, whereas this query group requires a transition to other syntagmatic alternatives. The main difference between SYNTAGMS and ALL is rooted in 'riding', 'stopping' and 'fast' in favor of ALL and 'landing' in favor of SYNTAGMS.

Finally, the fifth query group (*others*) are queries that have a very loose relation with the classifiers. Results show a significant difference between the ALL method and the other methods, as expected, but not compared to the UNLIMITED SEMIOSIS method. Many concepts in this group can only be found by ALL, but for 'headgear', 'tomato', 'farm' and 'wool' concepts are also found by the UNLIMITED SEMIOSIS method.

In the context of this case of the experiment, we can conclude that the type of query and the type of application prescribe the type of semiotic methods to consider. For applications that value effectiveness, the ALL method will be a good choice. Contrarily, for applications that require high quality, the UNLIMITED SEMIOSIS method would be a better choice, assuming that its queries do not require transitions between syntagmatic concepts (group 4), or are vaguely related to classifiers (group 5). Another good option for high quality applications would be to combine the SYNTAGMS and UNLIMITED SEMIOSIS methods. Finally, although in theory the PARADIGMS method should improve results for high quality applications, results indicate that it needs a more careful approach.

5.6.2. IMAGE RETRIEVAL

Results on the image retrieval case show the impact of the semiotic structures on both parts of the semantic gap, and therefore the system as a whole. The general trend is that performance for this case is lower than for the semantic matching case. This originates from the fact that our classifiers do not perform very well. For instance, in query groups 4 (*syntagms*) and 5 (*other*) some expansion methods show no performance at all, which implies that despite the presence of relevant ground truth for them, none of the queries produce image results. The largest difference in overall performance between both cases is that the methods for UNLIMITED SEMIOSIS and ALL are no longer significantly different (Figures 5.13 - 5.17). This is an indication that by adding irrelevant concepts (by the ALL method) more irrelevant images are produced, which might hurt more than adding less relevant concepts (by the UNLIMITED SEMIOSIS method) that produces less irrelevant images. This result even holds for high effectiveness applications.

A closer look at the results for the different query groups as introduced in section 5.4.2 show the following. For the first group (*Synonyms*) not much difference is found over the various methods. Only the ALL methods drops a little more than the SYNTAGMS method, because the expansion by the ALL method from 'motorbike' to concepts 'horse' and 'helmet' really hurts performance as both are not synonyms while any image with either a horse, a helmet or a motorcycle will still be retrieved. As indicated above, the performance of the image retrieval case is lower than the semantic matching case. In this query group that is exemplified by the fact that although our classifiers for 'boat', 'motorcycle' and 'turd' are performing flawless, 'car', 'bus', 'traffic light' and 'turnscrew' perform less optimal (F0.1~80%), whilst the classifiers for 'airplane', 'helmet' and 'football' can only be graded acceptable (F0.1~60%).

In the second query group (*unlimited semiosis*), the UNLIMITED SEMIOSIS method performs best. Significant differences between all other expansion methods, except ALL, are found. Differently from the results in the Semantic

Matching, the UNLIMITED SEMIOSIS method is not better than ALL for high quality applications.

Results from the third query group (*paradigms*) interestingly show that the UNLIMITED SEMIOSIS method is slightly, but not significantly, better than its counterpart ALL, even for high recall applications. This is, however, not only because of irrelevant expansions by the ALL method. In this group many paradigmatic restrictions are specified by the queries, specifically about color, and colors cause a large decrease in performance in image retrieval. For example, the ALL method produces a semantic match between 'silver' and 'gray', indicating that gray cars are relevant. Unfortunately, in the image retrieval part silver cars are not detected as silver, but mainly as black. This is because many of the cars have black windows. Another example shows that green traffic lights are never detected, because the main color of the traffic light is black, irrespective of the light that is lit. In fact, this represents a typical example for unlimited semiosis where the semantic value of 'green' refers to an abstraction level that is far above the specific level that is indicated by the minimal part of the object that actually represents the green lit light. After all, we are not searching for a completely green traffic light. Besides the color classifiers, also other classifiers perform suboptimal, which has a negative effect on the results: when a classifier is not able to detect the relevant concept in a relevant image, no difference between the methods can be registered.

Image retrieval results in the fourth group (*syntagms*) show similar results as in the semantic matching case: a slightly higher quality for the SYNTAGMS method and a higher effectiveness for the ALL method. These differences are, however, not significant any more. This is also the case for the fifth query group (*others*): no differences compared to semantic graph results, while the results are not significant anymore.

Overall we can thus conclude that for high quality applications, the ALL method potentially hurts performance. Already for neutral applications, the UNLIMITED SEMIOSIS method, or a combined application of the SYNTAGMS and UNLIMITED SEMIOSIS methods might be a better choice than the ALL method. Additionally, this conclusion might prove stronger when taking into account the end user of the system whom might judge the results from the ALL method far worse than the results from the semiotic methods: in a retrieval system with many irrelevant results, as with application of the ALL method, it would be hard to find the relevant results amongst them, whilst the less, but more relevant results of the UNLIMITED SEMIOSIS method will be much easier to detect by the end user.

5.6.3. LIMITATIONS OF EXPERIMENT

One of the limitations of these experiments is that our dataset is really small. With only 51 classifiers, the probability that any of the words in ConceptNet matches our classifier labels is, therefore, much lower. One single true positive, then, has a major impact on score whilst the many false positives that happen to have no match do not add to the score balance. This might be the reason that the ALL method is performing better than we expected. A second limitation is performance of the classifiers. As explained in previous subsection, our color classifiers as well as some object classifiers are suboptimal. In order to profit from improvements in the semantic reasoning part

of the system, good classifiers are needed. This argument also holds in reverse: on optimizing classifiers, overall little will be gained unless the improvements in this part of the semantic gap is matched with an equal improvement in the semantic matching part of the semantic gap.

An algorithm performs only as good as the quality of the data it is provided with. Especially when the focus is on generic semantic knowledge, a third limitation is the knowledge base of choice. ConceptNet has a lot of different types of relations and, therefore, connections between concepts exist that have different relations than expected, i.e., impacting accuracy, or no relations are available at all where one would expect their occurrence, impacting completeness. Although we experienced major improvements of version 5.3 over 5.2, e.g., corrections from erroneous relationships, several flaws in our experiment find their root in debatable concept relations from ConceptNet, or absent concepts. Another lesson learned from ConceptNet is the use of underscored words. Underscored words represent complex concepts that are represented by composition of two or more words by applying underscores, e.g., 'woman_wardrobe' or 'red_traffic_light'. Humans easily recognize their (syntagmatic) structure, but putting such understanding into (semiotic) rules is another matter completely. We therefore decided to abandon their use altogether, in order to stay away from potentially incorrect expansion results from factually correct data such as *CapableOf(camper, shoe_away_bear)* and *PartOf(dress, woman_wardrobe)*.

Finally, we have designed the experiment to score against two ground truths, one for the semantic matching and one for the image retrieval. They have the 100 queries in common, and since we only have 20 queries for each query group (Section 5.4.2) they also share their susceptibility to annotation-induced performance variations. We acknowledge this weakness in our experiment, especially since each annotation is performed by one individual each.

5.7. CONCLUSION AND FUTURE WORK

In conclusion, applying semiotic relations in query expansion over an external, generic knowledge base, contributes to a higher quality semantic match between query concepts and classifier labels, and also significantly improves image retrieval performance compared to a baseline with only synonym expansions. The type of query and the type of application prescribe the type of semiotic methods that should be considered for semantic matching. The indiscriminate use of all available relations that are present in the external knowledge base potentially hurts performance of the image retrieval part. The same approach for the semantic matching surprisingly outperformed the dedicated semiotic methods, although we have strong reasons to believe this effect is rooted in coincidental flaws in the knowledge base of choice. The experiment results also confirmed that the semantic gap that is experienced within CBIR consists of two cascading parts, and that little is gained overall when improvements address one part only. Finally, although multiple relations from the external knowledge base have been mapped onto one single semiotic method that at best approximates the semantics of the underlying relations, it is above doubt that semiotic coherence emerges in the otherwise non-semiotic semantic network that the exter-

nal knowledge base represents. We have shown that this semiotic coherence can be employed to improve the semantic capability of a software system.

In future research, it is advisable to explore the effectiveness of these semiotic structures on other knowledge bases, containing either generic or domain-specific knowledge, in order to further evaluate the true genericity of this semiotic approach. Specifically related to ConceptNet it may be worthwhile to investigate appropriate (semiotic) ways to handle complex concepts (underscored words) in order to disclose their knowledge and improve query expansion.

Inclusion of more classifiers, including better color classifiers, and more classifier types, such as action classifiers and object relation classifiers, will improve the significance of the outcome of the experiments as well as the applicability of the expansion methods. Furthermore, it would be interesting to conduct research into the influence of other semiotic structures, such as the semiotic square about contradictions, expressing relations that are also available in external databases, e.g., negated concepts and antonyms.

Additionally, it would be beneficial to measure image retrieval performance using relevance feedback from an end user on the found classifier labels by ConceptNet. For instance, our use of paradigms is completely unaware of the intentions of the end user and therefore might wrongly exclude a specific set of paradigmatic concepts. This can be easily corrected by adding context of use through relevance feedback.

6

BLIND LATE FUSION IN MULTIMEDIA EVENT RETRIEVAL

Edited from: **Maaike de Boer, Klamer Schutte, Hao Zhang, Yi-Lie Lu, Chong-Wah Ngo and Wessel Kraaij** (2016) *Blind Late Fusion in Multimedia Event Retrieval.* In: Int. J. on Multimedia Information Retrieval, volume 5, pp. 203-217.

This chapter relates to the research question **RQ5 JRER**. *In this chapter, we move from the query-to-concept mapping to multiple sources of information, and how to combine them. Often multimedia information retrieval systems not only rely on one source of information. In chapter 3, we used concepts trained on different datasets, as well as visual and motion information. These different types of information (visual and motion) are defined as a modality or data source. Previous research has shown that integration of different data sources can improve performance compared to only using one source. The specific fusion method that improves performance mostly is dependent on the assumptions about the data sources. We focus on blind late fusion, in which the weights of the different modalities are not trained (blind) and classifier scores are used instead of the features from the different modalities (late). In this chapter, we produce datasets with different distributions and dependencies to explore the influence of these on the performance of state of the art blind late fusion methods. We introduce several new blind late fusion methods based on inversions and ratios of state of the art blind fusion methods. Results show that five of the newly introduced blind late fusion methods have superior performance over the current state of the art methods in a case with enough training examples. The elegance of our proposed methods, especially JRER, is that it does not rely on a specific independence assumption as many fusion methods do.*

6.1. INTRODUCTION

The domain of content-based video information retrieval has gradually evolved in the last twenty years, from a discipline mostly relying on textual and spoken information in news videos, towards a richer multimedia analysis leveraging video, audio and text modalities. In 2011, the TRECVID Multimedia Event Detection (*MED*) task (Over et al., 2015) defined a testbed for machine understanding of digital video, by creating a challenge to detect high level or complex events, defined as "long-term spatially and temporally dynamic object interactions" (Jiang et al., 2012). The videos in this testbed are selected from the Heterogeneous Audio Visual Internet Corpus (HAVIC) (Strassel et al., 2012), a heterogeneous set of internet videos with a large variation of quality and duration. The MED task is part of the NIST TRECVID benchmark in which systems from over the world get evaluated on a yearly basis. Besides the yearly evaluation, a test set and a train set consisting of ground truth information for twenty events are available for research purposes.

This task is known to be extremely challenging because one of the key ideas is that these events are too complex to be grasped by a single channel of sensory input, which is named a *modality* or *data source*, resulting in a challenge to fuse information from multiple (data) sources. In most work reported on the TRECVID MED benchmark some type of fusion is used. As combining information from multiple modalities, such as visual and audio, in an early fusion makes little sense, *late* fusion is often the better choice. As only little training data is available in TRECVID MED *blind* fusion is commonly used, in which the weights of the different modalities are not trained. The fusion methods are, however, often empirically determined on the test set for which annotations are available rather than theoretically grounded. This often results in the choice that in this task the average fusion is the selected method to fuse information from different modalities (Myers et al., 2014; Oh et al., 2014).

In this chapter, we take a step back and consider applicability of several blind late fusion methods. Additionally, we introduce several novel methods based on the current methods by application of the inverse and their associated ratio. We focus on the integration of two modalities: vision (v) and motion (m). We provide simulations to give insight into which situations which methods work well. We consider 1) the underlying distributions of the features (Gaussian or uniform) for both the positive (relevant) and negative (irrelevant) examples and 2) the relation between the sources (dependent or independent). The goal of the simulations is to predict which of the fusion methods works best in the TRECVID MED. The choices within the simulations are, therefore, inspired by the TRECVID MED benchmark in which we have a case with 100 positive training examples (*100Ex*) and a case with only 10 positive training examples (*10Ex*). Furthermore, we use an non-linear SVM to train on the simulated data and provide a confidence score similar to the SVM output scores used on the TRECVID MED data. This non-linearity of the SVM is the reason to model the distributions in feature space and not in confidence score space. We evaluate performance using the Mean Average Precision (Over et al., 2015), which is a rank based performance measure.

The contributions of this chapter can be summarized by:

1. We provide both experimental results on an international benchmarked dataset and simulated data that provide insights in which blind fusion methods are good in which situations. To the best of our knowledge this has not been done before with a ranked-based performance metric.

2. We introduce novel blind late fusion methods based on current state of the art methods using the inverse and ratio of these methods and combining them.

3. We show that several of these introduced fusion methods outperform current state of the art on the simulated data as well as on the TRECVID MED dataset.

4. We recommend to use the introduced fusion method JRER as a new state of the art method in cases with sufficient training examples.

In the next section, we give a short overview of fusion methods in multimedia analysis and multimedia event retrieval. Section 6.3 explains the state of the art fusion methods as well as the proposed fusion methods. Section 6.4 contains the experiments and results on both the simulated data as well as the TRECVID MED, for both 100Ex and 10Ex. Section 6.5 consists of a short discussion and the final section provides conclusions.

6.2. RELATED WORK

Atrey et al. (2010) give an overview of the multimodal fusion methods in multimedia analysis. Firstly, a distinction between early fusion on feature level and late fusion on decision level is made. The advantage of early fusion is that correlations between multiple features can be used, but it can be hard to create a meaningful combined feature vector. Lan et al. (2012) add that early fusion techniques suffer from the curse of dimensionality and require much training data. According to Atrey et al. (2010), the advantage of late fusion is that the most suitable method for a single modality can be applied and it is more flexible and robust to features that have a negative influence compared to early fusion. A disadvantage is that the correlation between modalities cannot be fully exploited. These advantages with respect to the disadvantages inspired us to focus on late fusion methods in our research.

Besides the level of fusion, the method of fusion is also important. Two of the methods explained in Atrey et al. (2010) are rule-based methods and classification-based methods. Examples of rule-based methods are linear weighted fusion and manually defined rules. In the linear weighted fusion some form of normalization and weighting is used to combine different modalities. In general, the rule-based methods are computationally inexpensive and easy to implement, but the assignment of appropriate weights remains an issue. This method is often used in late fusion. Oh et al. (2014) further split late fusion into a blind method with fixed rules, such as geometric mean, a normalization method with assumptions on score distributions and a learning method which needs training data to set an appropriate weight. The difficulty of assigning appropriate weights made us focus on blind methods with fixed rules for the integration of different information sources.

According to Xu et al. (1992), three types of classifier outputs can be used in fusion: 1) abstract level: single class label; 2) rank level: ordered sequence of candidate classes; 3) measurement level: candidate classes with confidence scores. According to Tulyakov et al. (2008), voting techniques such as majority voting and borda count are the methods most used for the abstract and rank level classifiers, whereas sum, product and max-rules are the elementary combinations on measurement level.

In Multimedia Event Retrieval, several fusion methods have been explored. Mc Donald et al. (2005) compared fusion techniques on measurement level and on rank level. They show that fusion on measurement level achieves higher performance compared to fusion on rank level, even though the Mean Average Precision is a rank-based performance metric. Lan et al. (2012) propose a method that uses both early and late fusion by combining single feature classifiers, category-based classifiers and complete-feature set classifiers. Natarajan et al. (2012) combine features with p-norm Multiple Kernel Learning as early fusion method and use the double sigmoid function to normalize the scores. For late fusion a combination of Bayesian models is used. Xiong et al. (2015) combine a spatial detection map and a holistic deep representation using a deep neural network. Wilkins et al. (2006) and Zheng et al. (2015) propose late fusion methods based on score distributions. Similar to our experiments, Myers et al. (2014) compare the following fusion methods on the TRECVID MED 2012 Test set: arithmetic mean, geometric mean, mean average precision-weighted fusion, weighted mean root, conditional mixture model, sparse mixture model, SVMlight, distance from threshold and bin accuracy weighting. They use the visual, motion and speech information in the fusion. They conclude that the simple fusion methods geometric mean and arithmetic mean perform as well or even better than their complex fusion methods. This conclusion is also drawn by Oh et al. (2014) with their Local Expert Forest learning algorithm experimented on the TRECVID MED 2011 Test set.

6.3. BLIND LATE FUSION

In this section, we describe four state of the art late fusion methods and propose several novel methods as extensions on these state of the art methods. Each of these methods has their strengths and weaknesses, of which we provide some insights. For these insights, we focus on 1) the relation between the data points in the positive (relevant) examples (*Pos*) and the negative (irrelevant) examples (*Neg*), i.e. a Gaussian (normal) distribution or an uniform distribution, and 2) the relation between the data sources, which can be dependent or independent. As mentioned in the introduction, we focus on two sources, i.e. v for visual and m for motion. In the equations of the different methods, we only use two sources because of readability reasons, but all methods are trivially extendable to more than two sources. Each source produces a score, which is a confidence score between zero and one resulting from an SVM. Platt scaling is used to produce this score (Platt, 1999) and thus this score should relate to a probability, as denoted in the following equation.

$$S_v \hat{=} P(e|v) \tag{6.1}$$

, where given visual feature v, the probability score for an event e is denoted as S_v.

Whereas all state of the art methods have a fused score between zero and one, the novel extensions of these scores do not necessarily have a fused score between zero and one. In case a value of zero or one causes an equation to have an undefined answer, a value slightly higher than zero or slightly lower than one is used. Please note that the underlying data (feature) distributions and number of training examples influence the confidence score of one source and the dependency of the sources influences the similarity between the confidence scores of the sources.

6.3.1. STATE OF THE ART

First, the Joint Probabiliy (*JP*) is a theoretically found late fusion method and a good method in video retrieval (Tamrakar et al., 2012; Oh et al., 2014) and related to Naive Bayes classification (Lewis, 1998). Joint Probability (*JP*) is the square of the geometric mean and thus provides the same ranking results. As described by Kraaij et al. (2002), the joint probability *JP* can be derived from $P(e|v,m)$ under the assumption that the representations v and m are statistically independent and conditional independent given the event e. We expect that JP is a proper late fusion method in case the *positive* examples of the data sources are *independent* for each event. This has a consequence that if either of the sources has a low score, we become less certain of the event being true. This is also visible in the contour map in Figure 6.1. The lines, starting from the origin, represent the scores in assending order and the lines are convex.

$$JP = S_v \times S_m \tag{6.2}$$

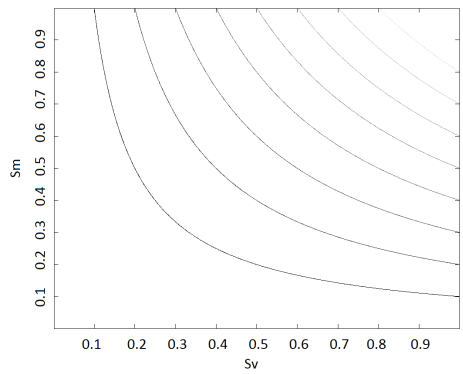


Figure 6.1: Contour Map for Joint Probability (JP)

Second, average fusion (*Av*), also called arithmetic mean, is known to be a strong fusion method in multimedia retrieval (Myers et al., 2014; Oh et al., 2014; Kittler et al., 1998). Theoretically the joint probability can be written as the sum by adding the assumption that the posterior class probabilities do not deviate greatly from the prior probability (Kittler et al., 1998). In practice this method is less sensitive to estimation

errors (Kittler et al., 1998). We expect that Av outperforms JP with the *independent* sources in case the classifiers produce more estimation errors. The contour map is shown in Figure 6.2. The map shows similar straight lines, which are neither concave nor convex.

$$Av = \frac{S_v + S_m}{2} \tag{6.3}$$

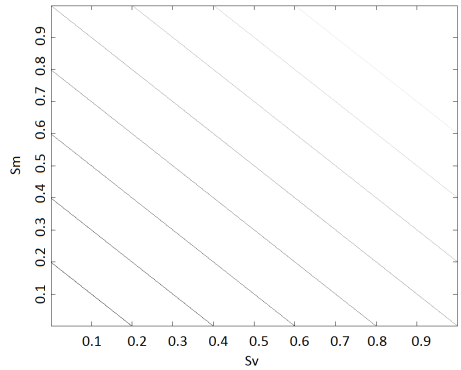


Figure 6.2: Contour Map for Average (Av)

Third, harmonic mean (H), is often used in information retrieval (Ravana et al., 2009; Van Rijsbergen, 1979). It is known as a method robust to (positive) outliers. This method should thus produce better results when the negative examples are producing high confidence scores, which would happen with a *uniform distribution of the negative examples* that overlaps with the positive examples. The contour map is shown in Figure 6.3. This map shows a similar map compared to JP, but with a more vigorous decrease in fused value (towards the bottom left corner).

$$H = \frac{2}{\frac{1}{S_v} + \frac{1}{S_m}} \tag{6.4}$$

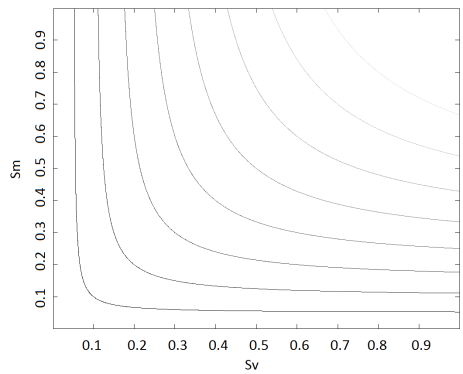


Figure 6.3: Contour Map for Harmonic (H)

Fourth, max fusion (*Max*) is one of the extreme cases in which we only rely on the reliability of the *high confidence scores*. The contour map is shown in Figure 6.4. This map shows different behavior compared to the other state of the art methods, because the lines are not rounded off as in JP and H and the lines are concave.

$$Max = \begin{cases} S_v, & \text{if } S_v > S_m \\ S_m, & \text{otherwise} \end{cases} \tag{6.5}$$

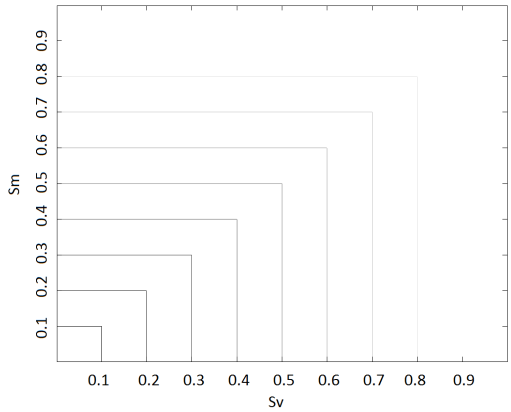


Figure 6.4: Contour Map for Max

6.3.2. INVERSE

As extension to the state of the art methods we introduce the inverse, with a general formula:

$$I(fn(S_v, S_m)) = 1 - fn(1 - S_v, 1 - S_m) \tag{6.6}$$

In probability theory, this function is based on the assumption that a low confidence score indicates the probability that the event is not present, whereas the previous section was based on the assumption that a high confidence scores indicates that the event is present. This implies that instead of relying on the values in the upper right part of the contour map, the inverse relies on the scores in the bottom left part of the contour map.

An easy example of the inverse is the inverse of the max: *Min*. This method relies on the reliability of the *low confidence scores*. The contour map, shown in Figure 6.5, nicely shows that Min is indeed the inverse of Max.

$$Min = \begin{cases} S_v, & \text{if } S_v < S_m \\ S_m, & \text{otherwise} \end{cases} \tag{6.7}$$

Another interesting inverse is the inverse of the average, which is equal to itself. This implies that we expect that Av is a proper late fusion method in case both *positive and negative* examples are *independent*.

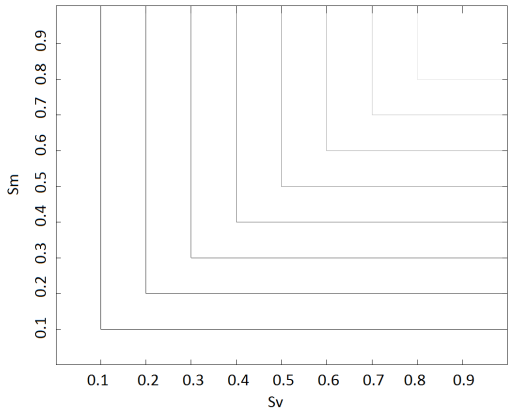


Figure 6.5: Contour Map for Min

The inverse of the joint probability, Inverse Joint Probability (*IJP*), equates to:

$$IJP = 1 - (1 - S_v) \times (1 - S_m) \tag{6.8}$$

Please note that IJP can also be written as combination of Av and JP $((S_v + S_m) - (S_v * S_m))$. Following our line of reasoning, we assume this method should be a good method to use when the *negative examples* of the sources are *independent*. The contour map is shown in Figure 6.6.

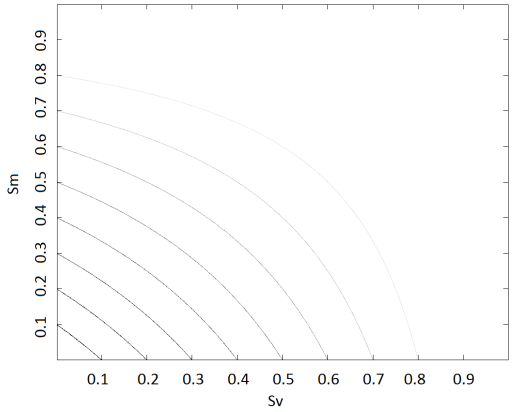


Figure 6.6: Contour Map for Inverse Joint Probability (IJP)

The inverse of the harmonic mean is equates to:

$$IH = 1 - \frac{2}{\frac{1}{1-S_v} + \frac{1}{1-S_m}} \tag{6.9}$$

Where the harmonic mean should produce good results in a uniform distribution of negative examples, the inverse should produce good results in a *uniform* distribution of the *positive examples*. This situation does not produce high performance, but this method is included for completeness. The contour map is shown in Figure 6.7.

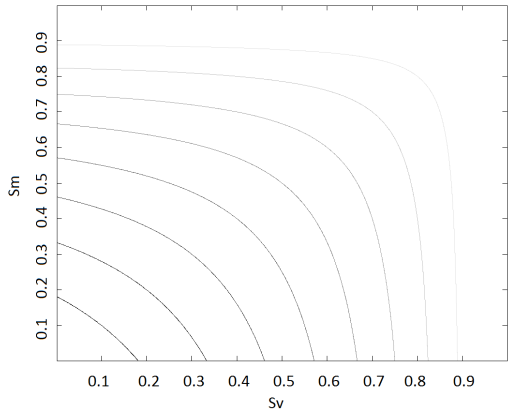


Figure 6.7: Contour Map for Inverse Harmonic (IH)

6.3.3. RATIO

We introduce the ratio as a combination of a method and its inverse with the general formula:

$$(fn(S_v, S_m))R = \frac{fn(S_v, S_m)}{1 - I(fn(S_v, S_m))} \tag{6.10}$$

This ratio is inspired by the likelihood from probability theory. These methods are based on the assumptions that both a high confidence score relates to a high probability that the event is present and a low confidence score relates to a low probability that the event is not present. We do not use the ratio vice versa, because we expect the SVM to comply to the former assumption about the higher confidence scores, whereas we are less certain about the compliance to the second assumption.

The ratio of the joint probability, Joint Ratio (*JR*), can be written as:

$$\begin{aligned} JR &= \frac{JP}{1 - IJP} \\ &= \frac{S_v}{1 - S_v} \times \frac{S_m}{1 - S_m} \end{aligned} \tag{6.11}$$

This method is the ratio between the probability that the event is present and the probability that the event is not present, which writes to $\frac{P(x,y)}{P(\overline{x},\overline{y})}$. This method is already introduced by Cremer et al. (2001) and is also known as the odds ratio (without

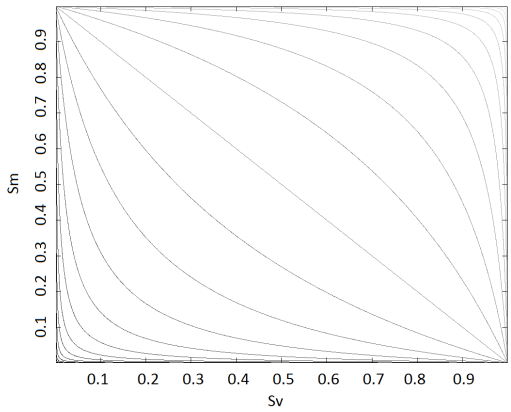


Figure 6.8: Contour Map for Joint Ratio (JR)

the log) in the Binary Independency Model (Mladenić, 1998; Yu et al., 1976; Robertson et al., 1976). We expect this method to work well when the two sources are *independent* for the *positive* as well as the *negative* examples. The contour map of JR is shown in Figure 6.8. The figure shows that for the low values JR tends to behave as JP, for the middle values JR behaves as Av and for the very high values JR behaves as IJP.

The ratio of the harmonics, *HR*, is not expected to work well, because an uniform distribution of both positive and negative examples will result in random performance. The contour map of HR is shown in Figure 6.9 and shows the behavior of H with the lower values and IH with the higher values. Comparable to JR, the middle values tend towards Av.

$$\begin{aligned} HR &= \frac{H}{1-IH} \\ &= \frac{\frac{2}{\frac{1}{S_v}+\frac{1}{S_m}}}{\frac{2}{\frac{1}{1-S_v}+\frac{1}{1-S_m}}} \\ &= \frac{\frac{1}{1-S_v}+\frac{1}{1-S_m}}{\frac{1}{S_v}+\frac{1}{S_m}} \end{aligned} \tag{6.12}$$

The ratio of the max and min can also be taken, resulting in the Extreme Ratio (*ER*). This ratio might be more robust compared to the Max, because it also uses information from the other source. The contour map of ER is shown in Figure 6.10 and as expected the behavior is most similar to Max for the lower values and most similar to Min for the higher values, but the edges of the contour do no longer have an angle of 90°. Because this method is not based on the independence assumption, we expect this method to work well for *dependent* sources.

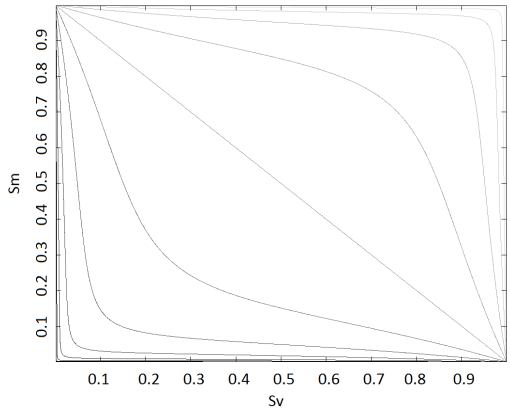


Figure 6.9: Contour Map for Harmonic Ratio (HR)

$$ER = \frac{Max(S_m, S_v)}{1 - Min(S_m, S_v)} \tag{6.13}$$

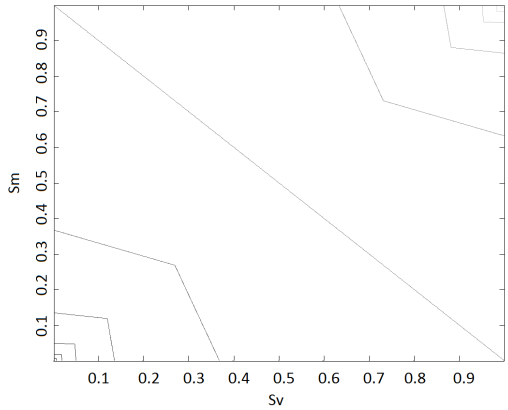


Figure 6.10: Contour Map for Extreme Ratio (ER)

6.3.4. COMBINING RATIOS

Finally, we introduce a combined ratio in which we aim at robust fusion by multiplying ratios. First, we combine JR and ER. In a combination of ratios, multiplication is a natural choice, as adding or dividing ratios makes little sense. The contour map of JRER is shown in two parts in Figure 6.11.

$$JRER = JR \times ER \tag{6.14}$$

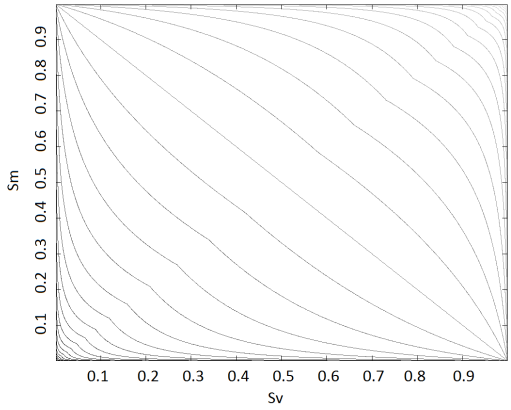


Figure 6.11: Contour Map for JRER

Although we expect HR to be not a very good method, it is interesting to see how a combination of all ratios performs. The contour map of Full is shown in Figure 6.12.

$$Full = JR \times ER \times HR \tag{6.15}$$

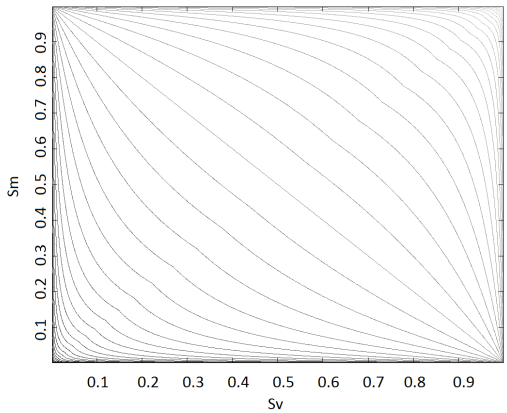


Figure 6.12: Contour Maps for Full

6.4. EXPERIMENTS

In our experiments, we use simulations and a real world multimedia event retrieval dataset named TRECVID MED (Over et al., 2015) to find out in which fusion method performs best under which circumstances. The parameters used in the simulations are inspired by the TRECVID MED case, but with the simulations we are able to generalize the results from TRECVID MED to more than twenty events and to different type of distributions of the data. In both the simulations and the TRECVID MED dataset, we use a case in which sufficient training examples are available to train classifiers (*100Ex*) and a case in which only few training examples are available (*10Ex*). In TRECVID MED we have ground truth information for 20 events where in the simulations we repeat the experiments 1000 times. The Percentage Mean Average Precision (*%MAP*) (Over et al., 2015), which considers ranking of positive examples, is used to measure the overall performance as it is the standard metric used in TRECVID MED. In addition, we added the amount of times in which the method was the best method (*#best*) and the amount of times the method had at least 95% MAP compared to the best performing method (*good*).

6.4.1. SIMULATIONS

In our simulations, we compare performance of the different fusion methods explained in Section 6.3. We use a Monte Carlo like method to produce data sets with different distributions and dependencies. Different from Ma et al. (2013) and Terrades et al. (2009) we do not create the distributions and dependencies on the classifier output, but in the feature space. As indicated by Ma et al. (2013) “analysis on dependencies between classifiers or/and features shows that statistics of classifier scores cannot truly reflect the dependency characteristics in feature level". We simplify the problem by using features with only one dimension. A second dimension with only zeros is added to the feature vector to properly train the SVM. The generation of the features for both sources is configured using a covariance matrix. This covariance matrix is build from the formula:

$$\Sigma = \begin{bmatrix} Var(X) & Covar(X,Y) \\ Covar(Y,X) & Var(Y) \end{bmatrix} = R(SS)R^{-1} = \begin{bmatrix} cos(\alpha) & -sin(\alpha) \\ sin(\alpha) & cos(\alpha) \end{bmatrix} \cdot \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \cdot \begin{bmatrix} cos(\alpha) & sin(\alpha) \\ -sin(\alpha) & cos(\alpha) \end{bmatrix} \tag{6.16}$$

, where R is the rotation matrix and S is the scaling matrix.

We use different configurations of the parameters α (degree of rotation), a (variance in source 1) and b (variance in source 2) to simulate the key factors for our fusion methods. For example no rotation, i.e. α is 0°or 90°can be seen as independent variables, resulting in a formula

$$\Sigma = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \tag{6.17}$$

A rotation of 45°can be seen as dependent variables, resulting in a formula

$$\Sigma = \frac{1}{2} * \begin{bmatrix} a+b & a-b \\ a-b & a+b \end{bmatrix} \tag{6.18}$$

To give more insight in the resulting values from these types of covariance matrices, we added two figures to illustrate in Figure 6.13. Furthermore, we added an example of uniform in Figure 6.14 to provide the full picture.

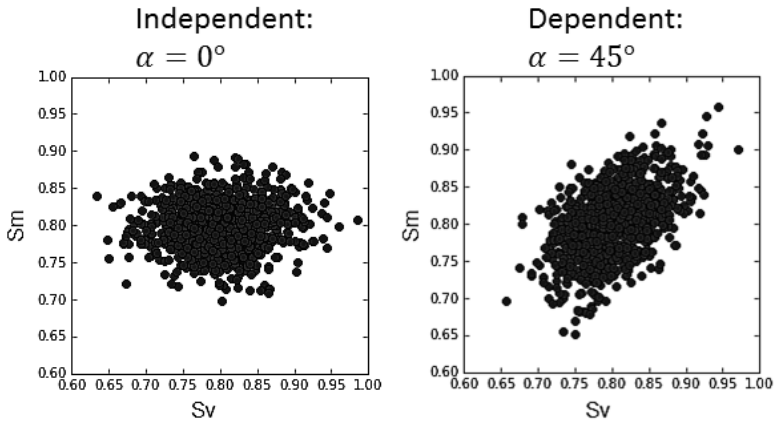


Figure 6.13: An Example of Independent Positives; μ = 0.8, a = 0.003 , b = 0.001, #datapoints = 1000

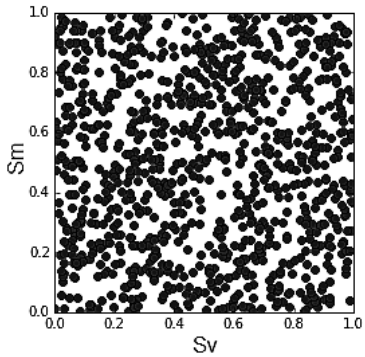


Figure 6.14: An Example of Uniform, #datapoints = 1000

Using these covariance matrices, we create positive examples with a mean of 0.8 and negative examples with a mean of 0.3. The variance of the positive examples is randomly picked between 0.001 and 0.1 for a and b separately, whereas the negative examples always have a bigger variance than the positive examples, resulting in a variance between a and 0.1 for source 1 and between b and 0.1 for source 2. In case we use a uniform distribution, we randomly pick a number between zero and one. In the training phase, we generate 100 positive examples and 100 negative examples for each source in 100Ex, whereas we generate 10 positives and 10 negatives for each source in the 10Ex case. In the testing phase, we generate 100 positive examples and 900 negative examples for each source. This unbalance in positive and negative examples resembles real world data in which few positive samples are present. In training we use a balanced number of positive and negative examples, because this

causes no problems with class weighting. We train an SVM with an RBF kernel on each source and fuse the scores of the SVMs applied to the test data using the late fusion methods. This process is repeated 1000 times to report statistically representative results.

6.4.2. HYPOTHESIS

Given these methods, we summarize our hypothesis in which cases which method should perform best in Table 6.1 based on Section 6.3. For example JP is expected to be the best method with *positive examples* and *independent* sources, meaning that it will appear in the first row. Both Av and JR are expected to be good methods with independent positive and negative examples and will overrule JP in the first column of the first row. Having both a uniform foreground and background does not make sense, as this would result in random performance. Min and Max are not present in this table, as they only display the extreme cases which we do not expect to be the best method with reasonable classifiers.

Table 6.1: Hypothesis best method to use when based on distribution of positive and negative examples and dependency between sources

Factors	**Negative Gaussian Independent (Neg Indep.)**	**Negative Uniform (Neg Unif.)**	**Negative Gaussian Dependent (Neg Dep.)**
Positive Gaussian Independent (Pos Indep.)	Av/JR	JP	JP
Positive Uniform (Pos Unif.)	IJP	x	IH / Full
Positive Gaussian Dependent (Pos Dep.)	IJP	H	JRER

RESULTS

The results of the experiments in terms of %MAP, best (the amount of times in which the method was the best method) and good (the amount of times the method had at least 95% MAP compared to the best performing method) are presented in Table 6.4 and 6.5 at the end of this chapter. The bold digit in a row indicates the highest value. A summary of the results is displayed in Table 6.2. Please note that in these simulation experiments we have shown quite extreme cases of dependency, independency and uniform results by choosing an α of 0°and 45°, and a value between zero and one for uniform. When different mean, variance and alpha values are chosen, the results of the methods will become different in the sense that the results will converge towards each other. Although these extreme cases might not be present in real world datasets, the results show the extreme differences between the blind fusion methods.

The results in the overview show that our hypothesis is only partially confirmed.

Table 6.2: Results best method to use when based on distribution of positive and negative examples and dependency between sources

Factors	**Negative Gaussian Independent (Neg Indep.)**	**Negative Uniform (Neg Unif.)**	**Negative Gaussian Dependent (Neg Dep.)**
Positive Gaussian Independent (Pos Indep.)	Av	Av	Av
Positive Uniform (Pos Unif.)	IJP	x	JRER/Full
Positive Gaussian Dependent (Pos Dep.)	JRER	JRER	JRER

The results, however, show a nice consistence. When the positive examples are independent, the best method to use is Av, whereas if the positive examples are dependent the best method is JRER. One sidenote on the independent positive examples is that in additional experiments we observed that with a lower variance JP tends to win more often compared to Av, which is in line with our hypothesis.

Furthermore, looking into the results on the 100Ex, we observe that in all cases it is beneficial to use fusion, no matter which method. In case of independent or dependent positive examples and uniform negatives or dependent negatives one single classifier (c1, c2) never outperforms any of the fusion methods, as indicated by a 'Best' value of zero. With an independent relation between the positive examples (Pos Indep.) Av is the best method. JP, ER and JRER also perform reasonably well in these cases. In Pos Unif.- Neg Indep. IJP is the best method, as expected based on our intuition. For Pos Unif. - Neg Dep. JRER and Full are the best methods, closely followed by JR and Av. With the dependent positive examples (Pos Dep.), JRER is the best method.

For the 10Ex simulations, the fusion methods are always better than the single classifier in the independent and dependent positive situations, but not in the uniform positive case. In many of the cases the results of 10Ex are similar to 100Ex, but in general Av performs slightly better with dependent positives compared to the 100Ex.

6.4.3. TRECVID MED

From the TRECVID MED benchmark (Over et al., 2015) we use the Train set with 100 examples in 100Ex and with 10 examples in 10Ex as positive examples and a Background set with 5000 videos as negative examples to train our SVMs on. We use the MED14Test to test the performance of the fusion methods. The MED14Test contains more than 27,000 videos and has ground truth information for twenty events. The extraction of the visual and motion features is explained in the next subsections.

VISUAL

For the visual information, we represent each video by a bag of keyframes. These keyframes are uniformly sampled as one frame per two seconds. For each of the keyframes, we apply pre-trained neural networks to obtain concepts which we use as our features. These neural networks are current state of the art in this domain and often used in this task. Each of the networks is trained on different images and specializes on different type of concepts. For example, one network is specialized in scenes, whereas another is in type of sports and another in objects. In total, we use information from six different neural networks. Table 6.3 gives an overview of the datasets used to pre-train the visual features and the type of learning used. The value behind the underscore indicates the number of concepts.

Table 6.3: Visual Features

Name	Structure	Dataset
FCVID_239	DCNN+SVM	Fudan-Columbia (Jiang et al., 2017)
SIN_346	DCNN	TRECVID SIN (Over et al., 2015)
Sport_487	3D-CNN (Tran et al., 2015)	Sports-1M (Karpathy et al., 2014)
Places_205	DCNN	MIT Places (Zhou et al., 2014)
ImgNet_1000	DCNN (Krizhevsky et al., 2012)	ImageNet (Deng et al., 2009)

In general, the DCNN (Deep Convolutional Neural Network) of Krizhevsky et al. (2012) is used for classification. This neural network, also called Alexnet, won the famous object recognition task named the ImageNet challenge in 2012. For the concepts present in the Semantic Indexing Task (SIN_346), Scene concepts (Places_205) and Research Events (RC_497) this network is fine-tuned. This means that the weights in the neural network trained on ImageNet are adjusted to better suit the concepts available in these datasets.

The RC_497 is a set of concepts selected from the MED'14 Research Collection dataset (Strassel et al., 2012) for which at most 200 positive keyframes per concept are manually annotated. This method is comparable to Natarajan et al. (2011) and Zhang et al. (2015a). For the DCNN network-based methods, the feature vector on video level is achieved by extracting the eighth (pre-final) layer of the network and average pool the results over the keyframes. This is done separately for each network. For the Fudan-Columbia dataset, we also use the DCNN network, but we extract the features from the seventh layer. These features are used to train an SVM to learn the video level responses on this dataset. This strategy is used, because in the paper published with this dataset (Jiang et al., 2017) the training of the SVM had a better performance compared to using the DCNN. The concepts from the final dataset are trained using a 3D CNN network (Tran et al., 2015). This network has the additional dimension of temporality and can, therefore, provide higher score on the Sports dataset.

The extracted visual features from each of the datasets are concatenated into one feature vector and trained with a Chi-Square SVM. The trained classifier is used on the test data to create a visual score per video S_v. This score is typically a confidence score in the range zero to one.

MOTION

For the motion features, we use state of the art Improved Dense Trajectories following Wang et al. (2013). The difference between visual and motion features is that motion features track feature points in each frame through a video, whereas visual features are used to identify concepts in each keyframe separately. Improved Dense Trajectory (*IDT*) uses histograms of oriented gradients (HOG), histograms of oriented flow (HOF) and motion boundary histograms (MBH). The dimensionality of these features are first reduced by using principle component analysis (PCA) and subsequently encoded using a Fisher vector with a pre-trained Gaussian Mixture Model (k=256).

Classification is done using a linear SVM, resulting in a motion score per video S_m. Similar to the visual score, this score is typically a confidence score in the range zero to one.

TRECVID MED CORRELATION

Using the results from the simulations, we define a hypothesis about which blind late fusion method is expected to perform best on the TRECVID MED events. We use the ground truth data from the TRECVID MED to determine the dependency between the sources. We do not use the feature scores here, because the features are high dimensional and thus not easily visualizable. We use the Pearson correlation coefficient to calculate this dependency using the values from the covariance matrix:

$$\rho_{X,Y} = \frac{CoVar(X,Y)}{\sqrt{Var(X)} \times \sqrt{Var(Y)}} \tag{6.19}$$

A value of 0 means no linear relation between both sources, a value of 1 means a positive relation and a value of -1 means a negative relation. The Pearson correlation coefficients on the MEDTEST for 100Ex are on average 0.54 ($min = 0.3; max = 0.79$) for the positive examples and 0.38 ($min = 0.15; max = 0.62$) for the negative examples. The mean classifier score is 0.69 for S_v and 0.45 for S_m. For 10Ex the ρ is on average 0.54 ($min = 0.21; max = 0.72$) for the positive examples and 0.27 ($min = 0.11; max = 0.91$) for the negative examples. The mean classifier score is 0.32 for S_v and 0.13 for S_m. The classifiers are in none of the events negatively correlated. According to the interpretation schema proposed by Mukaka (2012), the positive examples are regarded moderate positive correlated, whereas the negative examples are low positive (0.3 to 0.5) or negligible (0.0 - 0.3) correlated. Based on these coefficients, the positive examples seem dependent, whereas the negative examples are slightly dependent. Our hypothesis is, based on previous simulation results, that JRER will perform best.

Comparing the Pearson correlation coefficient from the TRECVID MED to the simulations Pos Dep. - Neg Dep., the average coefficient in the simulations is slightly lower (0.30), because negative correlations are included. A scatter plot of the positive examples in TRECVID MED 100Ex is shown in Figure 6.15. Despite the fact that the simulated data (presented in Figure 6.13) and TRECVID data have different score distributions for S_v and S_m, both bivariate score distributions exhibit a comparable level of dependence. We conjecture that the comparable dependence levels are a

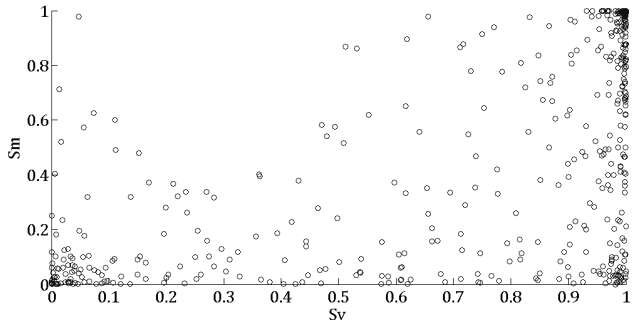


Figure 6.15: Scores for S_v and S_m for all positive examples of all 20 events in TRECVID MED 100Ex; $\rho = 0.54$

possible explanation why JRER performs best for both the simulated and MED fusion experiments. Further research is necessary to validate this conjecture.

When we create a subspace of these simulations in which no negative correlations are present (by selecting cases with b < a) and the positives and negatives with the same variance, ρ is on average 0.46 for the positives and 0.31 for the negatives in the 100Ex and 0.47 for the positives and 0.35 for the negatives in the 10Ex. These values are comparable to the values in TRECVID MED. The results of these simulations show the same trend compared to the results in Tables 6.4 and 6.5 in the row Pos Dep. - Neg Dep., indicating good performance of JR, HR, JRER and Full.

RESULTS

Table 6.6 and 6.7 show the Percentage Mean Average Precision for both visual (*Vis*) and motion information (*Mot*) as well as the fusion methods between both sources of information. Bold indicates the highest value in a row. The results on 100Ex show that JRER has the highest performance. The newly introduced IJP, JR, HR, JRER and Full all perform better than the current state of the art blind late fusion methods. In 10Ex Av performs best, although JR, JRER and Full also have 'good' performance often.

6.5. DISCUSSION

The results in the previous section show that in 100Ex JRER is the best method, both in the simulations and the TRECVID MED dataset. Although the differences in performance seem minor, JRER is slightly better compared to Av over 1000 runs in the simulations. In a competitive task such as TRECVID MED this increase of 3% on the twenty events can mean the difference in winning or placing second. We see no reason to not apply JRER in a blind late fusion setting with enough training examples instead of Av in case the features are dependent. Not only JRER is a good method, but all five newly introduced ratio methods perform better than Av in the 100Ex case. The reason for the superior performance is that the ratio methods use the probability that the event is present and the probability that the event is not present, whereas Av only uses the probability that the event is present.

In 10Ex the newly introduced methods also have 'good' performance, but Av is slightly better. One of the reasons for the success of Av in 10Ex is probably the Platt algorithm. This algorithm outputs the posterior probability, in which the prior is already included. As explained in Section 6.3, an additional assumption of Av is that the posterior does not deviate greately from the prior. The Platt algorithm has included this prior and, therefore, average is a good match. In the 100Ex the prior in the Platt algorithm has not much influence, because enough training examples are available to properly calibrate the confidence scores. Furthermore, the Platt algorithm produces lower confidence scores for the 10Ex and our proposed ratio methods flip behavior at 0.5. Our proposed ratio methods will, therefore, not differ from the method that is dominant on the lower end of the contour map. To improve upon Av in the 10Ex case, one could place the flip point of the ratio lower than 0.5, or try to normalize the scores in a way that the average classification score for positive examples is above 0.5.

Additionally, our simulations on the 10Ex might not fully explain the behavior in the TRECVID MED, because the 10Ex is much harder in a multi-dimensional space compared to the 1-dimensional simulation space.

6.6. CONCLUSIONS

We conclude that in this chapter we showed in which situation based on the distribution and dependency of data sources which blind late fusion method is theoretically and empirically the proper method to use. We not only used a simulation to produce different situations, but also provided results on an international benchmark to ground that our simulation results can also be produced in real world datasets. Five of the newly introduced blind late fusion methods showed superior performance over the current state of the art methods in a case with enough training examples. Especially the method named JRER seems a good and robust blind late fusion method.

Table 6.4: Performance of the Late Fusion methods for different simulated distributions on a 100Ex case

		c1	c2	JP	Av	H	Max	Min	IJP	IH	JR	HR	ER	JRER	Full
Pos Indep.	%MAP	62.10	60.99	86.23	**86.76**	85.48	68.47	83.83	82.04	72.79	85.89	83.96	86.14	86.39	85.89
Neg Indep.	Best	10	6	85	165	10	2	13	2	0	167	131	77	**198**	134
	Good	57	45	952	**991**	867	63	595	566	92	871	743	943	923	874
Pos Indep.	%MAP	60.90	60.98	84.67	**85.15**	83.91	65.18	82.30	79.69	69.70	83.93	81.79	84.59	84.48	83.94
Neg Unif.	Best	0	0	138	159	7	2	12	15	6	146	99	96	**202**	118
	Good	9	12	946	**983**	860	38	617	464	69	852	700	933	896	855
Pos Indep.	%MAP	60.98	60.98	85.21	**85.61**	84.49	64.88	82.93	79.84	69.41	84.36	82.14	85.07	84.97	84.36
Neg Dep.	Best	0	0	150	176	5	0	10	15	5	166	104	107	**188**	74
	Good	8	9	953	**986**	884	26	653	417	53	846	675	944	900	847
Pos Unif.	%MAP	25.65	25.06	33.56	34.67	32.54	31.37	28.16	**35.44**	34.63	35.01	34.98	33.94	34.88	35.00
Neg Indep.	Best	23	18	11	74	15	40	8	220	**232**	39	30	222	44	24
	Good	45	39	527	675	395	216	54	**803**	578	716	710	560	704	720
Pos Unif.	%MAP	24.37	24.58	32.86	33.72	31.96	28.88	28.03	33.74	32.31	33.74	33.47	33.32	**33.76**	**33.76**
Neg Dep.	Best	10	14	29	120	23	140	15	162	93	51	44	**216**	53	30
	Good	23	26	613	**730**	437	226	54	635	385	696	577	679	723	708
Pos Dep.	%MAP	60.07	61.25	84.31	84.74	83.59	67.84	81.82	81.70	72.76	85.35	84.25	83.98	**85.42**	85.34
Neg Indep.	Best	0	2	123	44	72	0	77	10	1	205	**163**	15	160	128
	Good	24	37	836	932	727	23	535	586	49	937	844	840	**964**	942
Pos Dep.	%MAP	60.55	60.49	83.52	84.09	82.76	66.14	81.05	80.72	71.25	84.41	83.11	83.43	**84.56**	84.42
Neg Unif.	Best	0	0	98	60	71	0	98	2	2	**186**	171	27	157	128
	Good	4	9	867	952	727	11	512	587	37	934	804	895	**964**	938
Pos Dep.	%MAP	61.14	61.00	83.81	84.29	83.08	66.37	81.46	80.93	71.47	84.59	83.35	83.66	**84.71**	84.60
Neg Dep.	Best	0	0	118	56	61	1	99	5	0	**168**	164	30	157	141
	Good	12	7	866	**964**	766	23	556	556	54	928	813	895	960	929

Table 6.5: Performance of the Late Fusion methods for simulated distributions on a 10Ex case

		c1	**c2**	**JP**	**Av**	**H**	**Max**	**Min**	**IJP**	**IH**	**JR**	**HR**	**ER**	**JRER**	**Full**
Pos Indep.	%MAP	59.20	58.95	78.20	**78.62**	77.56	65.49	74.59	76.29	70.60	78.28	76.97	78.19	78.59	78.31
Neg Indep.	Best	90	84	94	**135**	98	0	43	14	3	81	85	111	117	45
	Good	145	127	**822**	812	800	104	522	515	199	739	592	**822**	773	746
Pos Indep.	%MAP	56.89	57.13	76.84	**77.09**	76.27	61.47	73.34	73.65	67.14	76.21	74.52	76.81	76.71	76.27
Neg Unif.	Best	58	63	132	128	105	3	60	16	5	60	75	**146**	98	51
	Good	63	72	**833**	814	806	63	521	423	124	680	538	826	755	695
Pos Indep.	%MAP	56.45	56.61	75.52	**75.76**	74.98	61.12	72.13	72.71	66.67	75.04	73.51	75.50	75.47	75.10
Neg Dep.	Best	65	65	111	134	89	7	78	14	10	62	71	**154**	88	52
	Good	75	78	**795**	789	781	79	513	439	137	688	539	787	739	695
Pos Unif.	%MAP	**12.30**	11.84	10.86	11.02	10.72	11.92	10.39	11.25	11.41	11.10	11.16	10.91	11.05	11.09
Neg Indep.	Best	**455**	427	2	2	10	58	11	6	10	3	1	13	0	2
	Good	**614**	601	84	81	74	329	109	83	103	84	85	84	81	85
Pos Unif.	%MAP	**11.83**	11.80	10.79	10.93	10.68	11.62	10.45	11.12	11.21	11.01	11.06	10.83	10.98	11.01
Neg Dep.	Best	**418**	405	3	3	17	85	33	3	6	2	0	18	6	1
	Good	**574**	566	93	91	92	348	121	87	100	86	88	95	88	88
Pos Dep.	%MAP	57.40	57.37	76.89	**77.09**	76.34	62.78	73.30	74.51	68.64	76.82	75.59	76.63	77.08	76.85
Neg Indep.	Best	66	66	106	65	**144**	0	126	14	4	70	85	85	108	61
	Good	95	106	**806**	792	764	64	502	464	148	705	568	796	761	715
Pos Dep.	%MAP	57.38	57.68	76.24	76.56	75.67	62.47	72.97	74.25	68.20	76.46	75.38	76.13	**76.69**	76.50
Neg Unif.	Best	62	70	103	96	**136**	0	107	8	2	84	81	101	81	69
	Good	78	88	820	807	790	68	528	453	138	716	568	**823**	773	726
Pos Dep.	%MAP	57.41	58.46	76.10	76.35	75.55	62.77	72.91	74.07	68.46	76.22	75.16	75.94	**76.42**	76.25
Neg Dep.	Best	52	70	104	99	**144**	1	109	8	0	100	72	98	85	58
	Good	61	80	**826**	814	785	41	529	471	144	728	586	821	776	741

Table 6.6: %MAP integrating visual and motion features in MED14Test 100Ex

Event	Vis	Mot	JP	Av	H	Max	Min	IJP	IH	JR	HR	ER	JRER	Full
BikeTrick	**36.43**	8.98	16.39	16.33	16.15	29.04	16.22	23.98	29.30	23.72	30.07	16.58	22.57	23.78
CleanAppl	16.25	15.06	23.56	24.58	22.97	21.40	20.99	25.26	21.77	28.91	28.53	23.34	**29.21**	28.90
DogShow	93.13	64.23	80.34	87.53	76.82	94.06	75.06	**94.15**	93.70	90.36	90.02	87.52	90.44	90.31
GiveDir	10.06	9.44	16.05	16.27	15.55	12.44	14.54	13.10	12.62	13.85	15.21	**16.31**	13.36	13.72
MarriageProp	2.19	12.97	14.34	13.22	**14.57**	6.27	13.84	9.58	7.00	12.48	11.68	11.71	12.32	12.46
RenovHome	6.82	9.95	13.63	**14.06**	12.93	10.37	12.48	10.26	9.01	11.02	9.86	13.92	11.33	11.02
RockClimb	12.65	18.54	20.55	**21.05**	19.87	15.95	19.54	19.31	15.96	16.30	16.15	20.74	16.35	16.34
TownHallMeet	40.25	34.24	**46.34**	46.17	45.74	38.51	45.29	43.71	39.78	44.95	42.28	45.82	46.22	44.82
WinRace	20.65	16.56	20.62	**21.67**	19.77	20.33	19.15	20.97	20.42	21.17	20.64	21.82	21.11	21.20
MetalCrafts	18.07	22.43	19.72	19.49	19.58	19.15	19.76	**23.25**	19.28	23.24	21.06	19.29	22.28	23.21
Beekeeping	70.50	58.78	77.58	78.17	75.92	74.13	74.97	77.02	74.92	80.69	79.70	77.79	80.39	**80.73**
WeddingShower	23.39	18.46	27.34	28.65	26.65	27.41	24.21	**30.32**	28.46	29.93	27.90	27.77	29.64	29.86
VehicleRepair	45.35	37.29	**60.63**	60.24	59.93	44.89	57.95	53.34	46.22	59.04	56.22	59.43	59.95	59.00
FixMusInstr	47.61	39.45	51.34	52.07	50.24	47.49	46.78	**56.54**	48.81	55.21	52.52	48.53	54.94	54.92
HorseRidComp	**49.91**	32.03	40.20	42.10	38.50	49.43	36.26	45.09	49.03	43.44	46.05	40.94	42.91	43.58
FellingTree	21.12	20.99	32.80	33.53	32.38	25.90	27.06	31.71	26.42	**33.49**	31.09	30.44	33.23	33.56
ParkVehicle	30.13	29.69	35.36	36.26	34.78	35.41	35.06	39.59	35.79	41.30	39.21	37.10	**41.54**	41.50
PlayFetch	**18.27**	6.25	13.73	13.99	13.12	13.82	12.86	17.98	17.22	17.10	17.25	13.78	17.33	17.12
Tailgate	44.54	36.36	52.26	**53.76**	51.65	45.62	50.38	51.65	46.66	53.17	50.65	53.07	53.06	53.10
TuneMusInstr	13.94	18.80	27.02	26.65	27.14	18.64	25.37	22.87	18.97	27.18	24.86	26.21	**29.32**	27.07
MAP	31.06	25.53	34.49	35.29	33.71	32.51	32.39	35.48	33.07	36.33	35.55	34.61	**36.37**	36.31
Best	3	0	2	**4**	1	0	0	**4**	0	1	0	1	3	1
Good	2	1	9	9	6	2	3	9	2	**13**	5	7	**13**	12

Table 6.7: %MAP integrating visual and motion features in MED14Test 10Ex

Event	Vis	Mot	JP	Av	H	Max	Min	IJP	IH	JR	HR	ER	JRER	Full
BikeTrick	20.58	5.15	12.49	13.74	10.96	18.73	10.13	**20.86**	15.74	17.42	16.90	13.73	17.63	17.47
CleanAppl	**7.44**	1.20	3.68	3.91	3.38	7.02	3.03	7.13	7.12	7.31	7.40	3.81	4.32	7.31
DogShow	71.66	36.84	65.93	71.45	61.08	69.09	60.75	72.33	71.64	77.68	**77.77**	70.36	76.94	77.68
GiveDir	5.71	3.10	8.32	8.06	8.52	5.00	**8.60**	6.48	5.85	7.35	7.49	7.51	7.19	7.31
MarriageProp	0.40	0.54	0.74	0.50	0.88	0.42	**1.17**	0.49	0.48	0.71	0.75	0.43	0.64	0.69
RenovHome	4.00	4.35	8.42	6.64	**8.87**	4.49	8.20	4.97	4.87	5.92	5.90	6.19	5.85	5.87
RockClimb	13.76	20.01	21.48	**25.19**	20.98	19.48	20.77	20.91	20.36	20.22	18.99	22.90	21.47	20.18
TownHallMeet	23.97	8.01	19.44	**24.33**	14.76	23.70	12.38	23.93	23.93	23.37	22.43	23.76	23.14	23.57
WinRace	19.38	14.73	22.43	23.85	21.20	**24.14**	19.24	23.59	21.37	22.27	21.34	23.82	23.15	22.32
MetalCraft	8.45	12.66	16.77	**16.94**	16.41	7.94	16.24	11.89	8.30	11.84	11.47	16.83	11.99	11.82
Beekeeping	51.85	37.27	**56.21**	55.67	55.68	50.98	55.90	53.55	51.73	55.76	54.70	54.75	55.54	55.66
WeddingShower	15.33	5.45	13.89	15.06	13.53	13.48	12.33	14.45	13.92	16.16	15.73	12.92	**16.27**	16.18
VehicleRepair	45.93	9.14	44.47	**46.41**	40.58	43.62	38.77	45.65	44.48	45.10	43.65	45.20	45.46	45.38
FixMusInstr	**29.23**	3.54	20.91	24.04	16.88	20.69	12.27	20.68	20.25	18.79	17.34	24.39	19.68	18.91
HorseRidComp	39.57	28.57	40.30	42.87	36.20	39.19	33.93	39.55	39.09	37.10	35.74	**43.82**	37.85	37.02
FellingTree	9.46	15.44	17.57	16.18	16.79	14.89	15.76	14.86	15.35	17.95	17.27	16.00	17.61	**18.02**
ParkVehicle	15.52	20.89	31.29	29.77	**31.80**	16.73	31.19	17.61	17.32	22.90	23.74	29.35	22.46	22.93
PlayFetch	2.10	1.47	2.41	2.13	**2.47**	2.05	2.30	2.11	2.10	2.25	2.14	2.07	2.28	2.27
Tailgate	36.07	12.86	35.23	38.34	29.05	35.13	30.25	37.38	36.61	**42.73**	40.02	36.78	42.68	42.00
TuneMusInstr	13.02	3.41	22.20	19.78	**23.67**	11.47	23.40	12.90	12.53	17.80	17.50	19.11	19.38	17.63
MAP	21.67	12.23	23.21	**24.24**	21.68	21.41	20.83	22.57	21.65	23.53	22.91	23.69	23.57	23.51
Best	1	0	1	**4**	**4**	1	2	1	0	1	1	1	1	1
Good	6	0	6	7	7	2	6	6	3	**8**	5	6	**8**	**8**

7

COUNTING IN VISUAL QUESTION ANSWERING

Edited from: **Maaike de Boer, Steven Reitsma and Klamer Schutte** (2016) *Counting in Visual Question Answering.* In: Proc. of Dutch Belgian Information Retrieval Conference, 2016.
* Experiments have been conducted by Steven Reitsma under supervision of the first author

Previous chapters on query-to-concept mapping have focused on an explicit mapping and a linear weighted sum of the concepts. This chapter explores the possibilities of an implicit mapping and techniques beyond the linear weighted sum. We use the Visual Question Answering task in which a good answer should be formulated based on a query and an image. This task is, thus, not a retrieval task, such as the TRECVID MED task or the TOSO dataset, but an image understanding task. In this chapter, we explore **RQ6 VQA**. *This chapter is only a first exploration of the research question and it is meant as a broadening of the scope that was created by our assumptions. We focus on questions that require counting of objects in the image. We build upon the well performing DPPnet method by training concept detectors. These detectors are used in addition to the visual features. Additionally, we use a postprocessing technique to output the right type of answer to each type of question. Both the concept detectors and the postprocessing slightly improve performance and are usable on current state of the art methods in the field of image understanding.*

7.1. INTRODUCTION

One of the most common forms of visual question answering is one where a system answers natural language questions posed by human users about images (Wu et al., 2017). In practice this could take recent developments, such as Google Now, Siri and Cortana, a step further by not only being able to answer questions on general topics that are searchable on the web, but also on the local user context using e.g. a smartphone camera. This could be especially useful for visually impaired users, who can take a picture using their smartphone and ask their device questions about the local scene, such as *where is an empty seat in this train?* or *is there a pedestrian crossing here?*.

The VisualQA task (Antol et al., 2015) aims to enable research in visual question answering and is set up by VirginiaTech and Microsoft Research after the release of the Microsoft Common Objects in Context (*MSCOCO*) dataset (Lin et al., 2014). The MSCOCO dataset consists of more than 250,000 images. In the VisualQA task three questions for each image are posed together with 10 human answers to each question. The types of answers can be categorized into three major categories: *closed (yes / no)*, *numerical answers* and *categorical answers*.

In this chapter, we focus on questions with numerical answers. Current well performing methods such as DPPnet (Noh et al., 2016) achieve low performance on this type of question compared to the questions with closed and categorical answers. We propose to count the number of object instances using concept detectors with object segmentations. In addition, we introduce a postprocessing method to enforce providing an answer that is in the right category.

Results show that the use of concept detectors improves performance. Post-processing slightly improves performance further.

7.2. RELATED WORK

According to Wu et al. (2017), visual question answering solutions can be put into four categories: joint embedding, attention, compositional and knowledge bases. We focus on the first and biggest category. Approaches in this category use deep learning networks for both the image and the question and combine these in a classifier such as another neural network to predict the most probable answer. This is used as the baseline for the VisualQA task (Antol et al., 2015) and in the DPPnet system (Noh et al., 2016). DPPnet uses the state of the art VGGnet network (Simonyan et al., 2014), trained on the ImageNet images in the ILSVC-2012 dataset (Deng et al., 2009) to understand the image. This pre-trained model is finetuned using the MSCOCO dataset (Lin et al., 2014) in order to create a network tailored to the VisualQA task. Instead of the 1000 concepts from the final layer, the 4096 features in the pre-final layer are used. To create an embedding for the question, a recurrent neural network with Gated Recurrent Units (GRU) is used. The question model is pre-initialized using the *skip-thought* vector model (Kiros et al., 2015) which is trained on the BookCorpus dataset (Zhu et al., 2015), containing 74 million sentences. To generate an answer DPPnet uses a dynamic parameter layer to combine the image and question features. The image features are used as input for this layer and the weights are determined by

the question features using a hashing function (Chen et al., 2015). Recently, the Multimodal Compact Bilinear Pooling (MCB) (Fukui et al., 2016) further improved performance. This model combines the joint embedding with attention. The winning submissions in the VisualQA challenge, linked to the VisualQA task[1], combine joint embedding with an attention model to focus on a specific part of the image.

7.3. METHOD

In our method, we build upon DPPnet (Noh et al., 2016). We use *concept detector* activations in addition to the other features as input for the dynamic parameter layer. To train the concept detectors we use masked images of the ground truth annotations of the MSCOCO dataset. The concept detectors can be applied on either the full image, where the non-softmaxed activations are used as features, or they are used on the object proposals and the activations are summed. Additionally, we add *postprocessing* repair to enforce the correct answer type.

7.3.1. CONCEPT DETECTION

We train concept detectors using the ground truth annotation of the MSCOCO dataset (Lin et al., 2014) for each of the 80 classes. These 80 classes are tailored to the test set, whereas the 1000 ImageNet concepts are not and thus we expect better performance for these classes. We use a pre-trained GoogLeNet model (Szegedy et al., 2015) based on the Inception architecture. GoogLeNet was chosen for its high accuracy and the fact that it uses 12 times fewer parameters and thus fewer VRAM than the next-best ImageNet submission. From the ground truth annotations we use a masked version of each separate segmentation with a black background. This masked segmentation is fed through the convolutional neural network to obtain its features, similarly to the normal process for the unmasked images. A fixed number of segmentations is chosen (25 in our experiments) and if an image has fewer segmentations, the concatenated feature vector is zero-padded.

The network is trained on the segmentations using gradient descent with Nesterov momentum (Sutskever et al., 2013) for 25 epochs. For the first 10 epochs, the weights of the convolutional layers are locked to prevent the noisy gradients from the randomly initialized fully-connected layers from changing the pre-trained weights too much. Cross-entropy loss is used and Top 1 accuracy is used for validation. The segmentation masks are stretched to use the entire 224×224 image space (aspect ratio is retained), which improves validation accuracy from 57% to 87%. This removes scale variance and reduces overfitting. Furthermore, segmentations that have a surface smaller than 500 pixels are removed from training as they provide no meaningful information. The biases in the first convolutional layer are set to 0 to ensure the black background causes no activations. Finally, since the class balance is skewed —the most prevalent class occurs 185,316 times, while the rarest class occurs only 135 times —the amount of data per epoch is limited to 5000 per class. Note that if a class has more than 5000 samples, each epoch different data is shown to the network. Effectively, this means samples in underrepresented classes will be shown to

[1] http://visualqa.org/challenge.html

the network more than samples in large classes.

These concept detectors can be used on either the full image or the object proposals within the image. Using the full image, we expect a deterioration of the activation with fewer objects (i.e. less pixels firing on the object) for which the network can map activations to number of objects. When we use object proposals, we expect that the activation will be high for objects that are in that proposal and summing over the proposals will resemble actual counting.

The object proposals can be obtained in several ways. First, ground truth segmentations as present in the MSCOCO dataset could be used. These segmentations are, however, not available in many datasets, so automatic object proposals can be obtained using *Edgebox* (Zitnick et al., 2014) or *Deepbox* (Kuo et al., 2015) (with non-maximum supression), which are state of the art segmentation methods. Edgebox generates bounding box object proposals, which makes it especially suited for objects that are rectangular. Often, the Edgebox strategy generates hundreds of object candidates. The algorithm scores and sorts these according to the number of contours that are wholly contained within the image. Deepbox (Kuo et al., 2015) uses a different scoring metric: it trains a convolutional neural network that re-ranks the proposals from Edgebox. Using Deepbox, the same recall is achieved with four times fewer proposals.

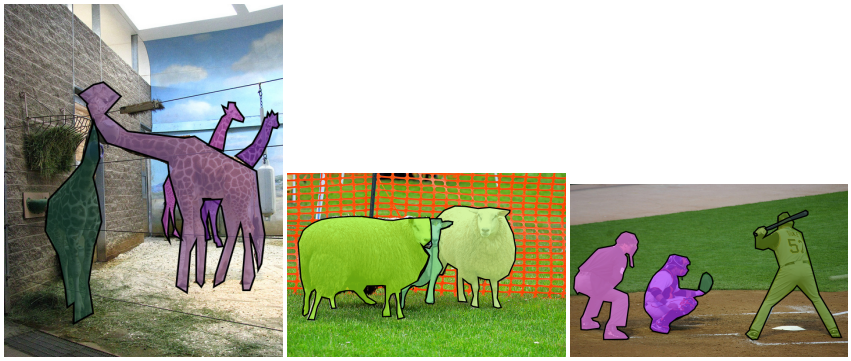
7.3.2. POSTPROCESSING REPAIR

For some questions, such as *how many...* questions, we know that the answer should be numerical. Often, the network will predict other answers as well, such as the string equivalents of the numeric digits, e.g. *one* instead of *1*. For questions that start with *are there...*, *does this...*, and so on, we expect as an answer *yes, no* or a word that exists in the question. For example, the question *does this image contain a cat?* always has to be answered by either *yes* or *no*, while the question *is there a cat or a dog in this image?* should be answered with either *cat*, *dog*, *yes* or *no*. Using a simple rule-based program, questions that start with *how many* always get the numerical answer that generates the highest softmax response in the network and the questions that have a closed answer are processed as explained above.

7.4. RESULTS

Table 7.1: Evaluation of segmentation methods on val2014

Method	All	Yes/No	Number	Other
DPPnet (downsized)	51.94	78.34	33.66	36.77
Ground truth annotations	**52.29**	78.34	**36.46**	36.77
Ground truth annotations (regression)	52.23	**78.40**	34.84	**37.01**
Edgebox	52.08	78.34	34.97	36.77
Deepbox	52.16	78.34	35.36	36.77



(a) *How many giraffes are in this picture?*
DPPnet: 2
Ground truth annotations: 4
Edgeboxes: 2
Deepboxes: 3

(b) *How many sheep are in this picture?*
DPPnet: 1
Ground truth annotations: 3
Edgeboxes: 2
Deepboxes: 2

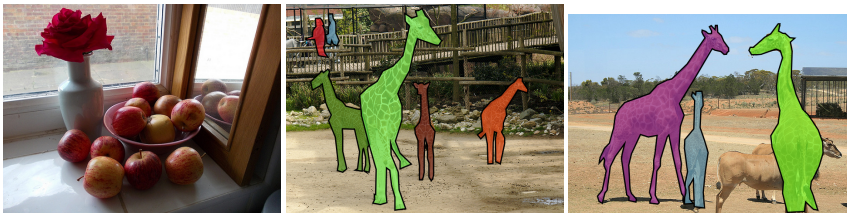
(c) *How many players can you see?*
DPPnet: 2
Ground truth annotations: 3
Edgeboxes: 3
Deepboxes: 3

Figure 7.1: Comparison of segmentation methods
Overlays are the ground truth proposals (in different colors).

In our experiments, we first investigate the optimal performance gain using segmentations. We use the validation set to test the ground truth annotations (which are not available in the test set), the Deepbox and Edgebox methods. These methods do not yet use the concept detectors, but have as input a vector of $25 \times 4096 + 4096$, which resembles the output of the pre-final layer of the VGGnet on each of the 25 object proposals and the full image. Furthermore, we test a classification and regression approach. Classification is similar to DPPnet, using a softmax over all answers present in the training set and for *regression* we have one output node outputting a number. Because the use of finetuning and the large dynamic parameter layer requires at least 12GB of VRAM (i.e. a GTX Titan X or Tesla M40), we remove the finetuning and the large dynamic parameter layer to make the network fit into 6GB of VRAM, enabling its use on less hardware. In the downsized network, the hash size is decreased from 40000 to 10240 and the number of linear units in the dense part of the network is decreased from 2000 to 1024. Afterwards, we use the finetuned and the full network on our best run to make a submission in the VisualQA challenge. The results of these experiments are shown in Table 7.1. These results show that object proposals can increase performance by 3%. Classification works slightly better than regression in the overall and numerical categories. The ground truth annotations obviously achieve the highest performance, but the bounding boxes by Deepbox gain 2% performance and are slightly better than those produced by the Edgebox system. To gain some insight, we show a few images with the answers to a numerical question for the different methods in Figure 7.1. In 7.1(a) and 7.1(b), we see that automatic segmentation techniques such as Edgeboxes and Deepboxes have trouble segmenting the objects, while in 7.1(c) this seems to be no problem. This makes sense when realizing that Edgeboxes and Deepboxes create rectangular object proposals, which

Table 7.2: Results on test-dev2015

Method	All	Yes/No	Number	Other
LSTM Question + Image (baseline from Antol et al. (Antol et al., 2015))	53.74	78.94	35.24	36.42
DPPnet (finetuned, no downsizing from Noh et al. (Noh et al., 2016))	57.22	80.71	37.24	41.69
DPPnet (downsized)	56.11	79.88	36.81	40.18
Concept detectors on Deepbox segm. (downsized)	56.13	79.99	36.87	40.16
Concept detectors on full image (downsized)	56.34	79.99	37.31	40.45
Concept detectors on full image (downsized, +pp repair)	56.45	80.03	37.46	40.66
Concept detectors on full image (finetuned, no downsizing, +pp repair)	**58.01**	**80.89**	**38.03**	**42.44**



(a) *How many apples are in the picture?*
DPPnet: 3
Concept detectors: 8
Ground truth: 11

(b) *How many giraffes can you see?*
DPPnet: 2
Concept detectors: 4
Ground truth: 4

(c) *How many giraffes are there?*
DPPnet: 2
Concept detectors: 4
Ground truth: 3

Figure 7.2: Ground truth annotations and questions for some images in the MSCOCO dataset

suits 7.1(c) very well, but 7.1(a) and 7.1(b) less so.

Based on these results, we continue with the classification method and Deepbox, because the ground truth annotations are not available in the test set. We now use the concept detector activation sum over all 25 object proposals and concatenate this vector to the 4096 vector of the original pre-final layer. In the full image runs, the concept detector activations over the whole image are concatenated with the 4096 vector. Results are shown in Table 7.2. Interestingly, using the full image is better than using the object proposals. As indicated before the deterioration of the concept detectors scores might be more useful for the network compared to the top 25 rectangular object proposals. The postprocessing repair slightly improves performance. In Figure 7.2 we can see some example questions and images with the answers given by the regular DPPnet and the DPPnet with concept detector information. In 7.2(a), the bias of DPPnet towards more often occuring answers can be seen. Using concept detectors, the answer is closer to the truth, but still not correct. In 7.2(b) and 7.2(c), we view the disadvantage of building a scale invariant system. The concept detector activations for both images are almost equal, caused by the larger objects in 7.2(c) compared to 7.2(b). In the context of the VQA challenge, using our method scores 5th place out of 30 on numerical answers, tested on the test-standard dataset split. Although the difference in performance caused by the concept detectors seems small, it is a real improvement due to the volume of the dataset.

7.5. CONCLUSIONS

Using concept detectors to count in a visual question answering task improves performance with respect to only using the regular image features. The postprocessing further improves performance. Object proposals intuitively should increase performance, but with current state of the art methods, performance on the full image is higher. The top performing methods, such as Multimodal Compact Bilinear Pooling (MCB) (Fukui et al., 2016), could use the described concept detectors and postprocessing to potentially improve their system.

8

CONCLUSION

In this thesis, we have explored ways to improve semantic query-to-concept mapping to achieve a visual search capability for ad-hoc queries. We have evaluated different existing techniques for query expansion from the text retrieval domain, such as knowledge bases and word2vec, improved upon them, investigated how we can incorporate users in the process, which strategies are appropriate for which type of queries and how we can combine different sources of information together. In the following sections, we reflect on all research questions, moving from the specific research questions to the main research question. We continue with the limitations of the studies conducted in this thesis. In the final section, we move towards the impact of this thesis on the work of the analyst and visual search capabilities in general.

8.1. KNOWLEDGE BASES (RQ1 KNOWLEDGEBASES)

RQ1 KnowledgeBases: *How can we incorporate knowledge bases in query-to-concept mapping and which of the current knowledge bases is most applicable for this purpose?* (Chapter 2)

Knowledge bases can be incorporated in query-to-concept mapping in several ways. In chapter 2, we explored the incorporation of the general knowledge bases ConceptNet and Wikipedia. For ConceptNet, we explored the graphical structure and used the weights of the edges between the concepts to determine the relevance of a concept for a certain query. For Wikipedia, we used the information on the Wikipedia page that was strongly related to the query to find relevant concepts. Based on the experiments in chapter 2, we see a slightly better effectiveness for the mapping based on ConceptNet. Both knowledge bases suffer from the fact that they are not complete and, thus, do not cover all aspects of the event. The solution to the insufficient coverage problem is an expert knowledge base. In a closed domain, all knowledge can be modelled in a way that the knowledge base is complete. The downside of these models is that they take a significant amount of time to create and maintain. In our experiments, the expert knowledge base did not necessarily outperform the general

knowledge bases. Another option to deal with the insufficient coverage problem is to fuse the results from the knowledge bases. In chapter 2, we show that a late fusion improves Mean Average Precision performance by 13.9%. For the purpose of searching an event, ConceptNet seems the most applicable single knowledge base, but we are also aware of the recently published EventNet (Ye et al., 2015), which we did not incorporate in our research. A combination of multiple knowledge bases seems the key to significantly improve query-to-concept mapping by increasing the coverage of event facets.

8.2. SEMANTIC EMBEDDINGS (RQ2 WORD2VEC)

RQ2 word2vec: *How can we use semantic word embedding in query-to-concept mapping, and how does the mapping depend on the concepts in the Concept Bank?* (Chapter 3)

In chapter 3, we explained that semantic word embedding methods, such as word2vec, are game changers in text retrieval. A common way of using word2vec in a mapping is to find the related k items with the smallest semantic distance to the query. Our incremental w2v (*i-w2v*) method, introduced in chapter 3, builds a mapped set of concepts in a way that the (cosine) similarity of this set of concepts is close to the original query. Only concepts that increase the similarity to the query are added to the set of concepts. This incremental method is more robust to query drift, because concepts pointing in the same direction in vector space are not added if they do not increase the similarity to the query vector. Our method improves visual search effectiveness in terms of 12% MAP (based on the full vocabulary) compared to the state of the art word2vec method (Mikolov et al., 2013).

The most effective query-to-concept mapping in terms of MAP highly depends on the concepts in the Concept Bank. In chapter 3, we show that high-level concepts are important for the retrieval of high-level events. Low-level concepts, however, are also necessary, because if none of the high-level concepts is semantically close to the event, a combination of low-level concepts might provide a sufficiently good match. Obviously, in a closed or semi-closed domain, we should aim to create a vocabulary that contains all possible relevant concepts. This makes selecting the relevant concept(s) an easy task by direct matching, synonym searching and getting the closest match in case of typos in the query. Mapping techniques based on knowledge bases or semantic embeddings will not find different concepts and provide the same performance, as indicated in chapter 2 and 3. Many domains are, however, not closed. It is unknown what queries a user will ask and what the system should be able to handle, i.e. our open world assumption. In that case, we need a vocabulary that is as rich as possible and a good query-to-concept mapping technique, such as i-w2v.

8.3. FEEDBACK INTERPRETATION (RQ3 ARF)

RQ3 ARF: *How can we involve the user to optimize semantic mapping and retrieval performance for a visual search capability?* (Chapter 4)

Previous methods have focused on automatically mapping the query to a set of concepts. The results in chapter 2 and 3 show that a manual selection of the concepts improves visual search effectiveness compared to an automatic mapping. A manual mapping, however, involves expert knowledge on all concepts in the Concept Bank. Other methods to involve the user in the mapping and retrieval process is through user feedback. In chapter 4, we explore whether feedback on concept level or on video result list level can improve visual search effectiveness. We propose a novel algorithm on video level and compare performance to state of the art methods from both levels. Our method on video level did not create a novel classifier based on the relevant and non-relevant videos as many state of the art methods do, but it uses the binary feedback on individual videos in the result list to change the weights of the selected concepts. This method achieves better performance, both objective and subjective, and is more robust compared to training a new classifier based on the relevance feedback.

Both the manual mapping method and the relevance feedback on video level method can be applied in specific application domains. Whereas in the general search engine case, users will not be eager to learn all (relevant) concepts in the Concept Bank, users in for example the security domain might be able to learn all concepts and achieve performance that is comparable to the manual mapping method. By comparing results from experiments with the manual mapping method and experiments in chapter 4 on the relevance feedback method using ARF, the latter yields a significantly better retrieval effectiveness on the 32 events from the TRECVID MED task. We can, thus, conclude that an automatic query-to-concept mapping with user feedback on video level is a good way to involve the user in the event retrieval task.

8.4. SEMANTIC STRUCTURES (RQ4 SEMIOTICS)

RQ4 Semiotics: *To what extent can semantic structures increase understanding of the query?* (Chapter 5)

A user can provide queries to the system. Some queries might be about high-level events, such as in the TRECVID MED case. Other queries might contain objects, attributes and actions, such as *Look out for a person with a red jacket, grey pants and black backpack*. Instead of placing the whole query into a semantic embedding, a syntactic analysis can guide us in determining the nouns, verbs and adjectives. If no direct match is available for these words, a guided automatic query expansion using knowledge bases should increase precision of the mapping. In chapter 5, we explore whether certain types of semantic structures improve semantic mapping on certain type of queries. These queries have no direct match on at least one of the words and the goal is to find the concept related to that word. This relation is inspired by a semantic structure, such as a synonym (*car is vehicle*), unlimited semiosis (*Mercedes is*

car), paradigm (*man is NOT woman*) and syntagm (*landing thing is airplane*). The results show that the semantic structures related to that type of query improve performance in terms of F-score compared to only using synonyms. We, however, also show that using all types of semantic structures can often provide as good an understanding, in terms of mapping quality, as only using the specific semantic structure (i.e. synonym). This is a risk, because 'wrong' concepts can also be matched, but results show that in our dataset this does not hurt retrieval performance. In the TRECVID MED task, we have already shown that the same is true (Chapter 3). Using word embeddings with many, unknown relations (i-w2v) provides better retrieval performance compared to the knowledge bases. The specific semantic relations are, thus, not essential in providing a sufficient query-to-concept mapping. Because our Concept Bank is limited, a vaguely related concept might provide a better retrieval compared to having no concepts in the mapping or a mapping in which only part of the query is covered. On the other hand, in a system in which precision or quality has high value (for example the monitoring case), specific semantic relations should be considered.

Semantic structures can, thus, increase understanding of the query in terms of effectiveness on query-to-concept mapping, but this mapping does not necessarily translate into a better visual search effectiveness. Based on the results in chapter 2 and 5, visual search effectiveness is higher in case all semantic structures are used compared to only using the appropriate structure.

8.5. FUSION (RQ5 JRER)

RQ5 JRER: *Can we design a more effective score fusion method that is motivated by explicit assumptions about the distribution of classifier output values and the dependency between input sources?* (Chapter 6)

In chapter 6, we use simulations and the international benchmark TRECVID MED to get some insights into why certain methods, such as average fusion, work better compared to other methods. We propose several novel methods based on the inverse (probability that something is not true) and the ratio (probability that something is true divided by the probability that something is not true) of state of the art methods and show that these novel methods improve performance in cases with sufficient training examples. The elegance of our proposed methods, especially JRER, is that it does not rely on a specific independence assumption as many fusion methods do.

8.6. VISUAL QUESTION ANSWERING (RQ6 VQA)

RQ6 VQA: *What are the possibilities of implicit query-to-concept mapping in terms of visual search effectiveness?* (Chapter 7)

In chapter 7, we explore the possibilities of implicit query-to-concept mapping using the Visual Question Answering task. We adopt the state of the art deep learning architecture named DPPnet (Noh et al., 2016) in order to handle concept detec-

tors. We show that implicit query-to-concept mapping is possible and it achieves a decent performance in question and image understanding through answering questions about images. Although performance is decent, a downside of the implicit mapping is that it is harder to get insight in why the system provided wrong answers.

This research question is, however, not completely explored. Novel directions such as Bayesian deep learning (Wang et al., 2016a), attention models (Kim et al., 2016; Lu et al., 2016a), and adversial deep learning methods (GANs) (Chen et al., 2016), could provide both higher performance and some insight in the mapping.

8.7. SEMANTIC QUERY-TO-CONCEPT MAPPING (MAIN RESEARCH QUESTION)

Main Research Question: *How can we improve visual search effectiveness by semantic query-to-concept mapping?*

Based on the methods explored in this thesis, we can improve visual search effectiveness by using a combination of i) query-to-concept mapping based on semantic word embeddings (+12%), ii) exploiting user feedback (+26%) and iii) fusion of different modalities (data sources) (+17%). The results in chapter 2 and 3 show that an automatic mapping can be achieved through knowledge bases or a semantic embedding. Our proposed incremental word2vec (*i-w2v*) method improves effectiveness by 12% in terms of MAP compared to the state of the art word2vec method and the knowledge based techniques (Table 3.4). This improvement is, however, dependent on the availability of the concepts in the Concept Bank: without concepts related to or occurring in the event, we cannot detect the event. An additional improvement can be achieved by incorporating the user in the process. Our proposed Adaptive Relevance Feedback (*ARF*) method, proposed in chapter 4, improves the query-to-concept mapping and visual search effectiveness by 26% MAP compared to no feedback and 20% MAP compared to state of the art (Table 4.1). Our method changes the weights of the relevant concepts instead of training a novel model based on the annotations, and is, thus, more robust to few (positive) annotations. Additionally, we show in chapter 6 that average fusion of different sources is robust with just a few training examples (Table 6.5 and 6.7), whereas our JRER fusion method is effective with dependent sources with sufficient training examples to create a good model for the data source (Table 6.4 and 6.6).

8.8. LIMITATIONS

Although the results in this thesis look promising, the experiments conducted in this thesis have some limitations. One of the major limitations is the use of the different datasets. First, the TRECVID MED dataset. Although this dataset contains events and videos that best match our requirements, it only contains up to forty events. Although we do not expect the conclusions of the experiments to be different for different queries, a small dataset does not provide enough data to properly conduct statistical analysis on to further strengthen our conclusions nor to estimate and exploit

the event level statistics. Second, the TOSO dataset. This dataset contains a sufficient number of queries (100), but has a smaller Concept Bank (51 concepts). The probability that we find a matching concept in the Concept Bank for any of the relations is low. One true positive has a major impact on both precision and recall, where many false positives that have no concept in the Concept Bank have no impact on the performance. This could have influenced the result that using all semantic structures is better than using specific semantic structures.

A second limitation is our working hypothesis to linearly combine the concepts. We have noticed that the weights for the concepts are highly influencing the performance. Because we did not have sufficient training examples of events, we had little tuning possibilities on the setting of the weights for the different methods, such as the knowledge bases or i-w2v. Because none of the methods is tuned, we do not expect that the conclusions about the order of the performance of the methods are different, but a higher performance might be possible for the methods.

A third limitation is within our proposed incremental word2vec model. We used a model that is trained on the textual information on GoogleNews. This implies that the contextual textual information is modelled, which might be different from the contextual visual information, i.e. some words might occur in the same kind of texts, but not in the same kind of images. As an alternative to our experiments on the word2vec embedding based on text only, recent semantic embeddings that use both words and videos, such as VideoStory (Habibian et al., 2014b) and Word2VisualVec (Dong et al., 2016), can potentially improve performance.

A fourth limitation is related to our proposed ARF method. ARF is dependent on the initially chosen set of concepts. If one of the relevant concepts has a bad detector, the user feedback will potentially decrease the weight and, thus, the importance of the concept. This decrease in weight could result in a query drift towards the relevant concepts that might only cover few facets of the query. Because feedback on video level does not provide insight in whether the concept detector is not performing well, or the concept is not relevant, it is not easy to determine whether this query drift is wanted or not.

A final limitation is related to our fusion methods. In the experiments conducted, we assume that both data sources are equally important, and the fusion is blind. A direction for future work is that JRER could be optimized using a weighted fusion based on the performance of the different sources. Distribution-based or rank-based methods could be explored.

8.9. POTENTIAL IMPACT AND FUTURE WORK

In this thesis, we improve the visual search effectiveness by semantic query-to- concept mapping. But how does this change the world, and how will it be used in the future? Although the goal of this thesis is not to transfer the knowledge obtained in this thesis to the security domain, we will sketch the potential impact it might have to application domains in which professionals (such as analysts or operators) use a visual search system. Additionally, we look into the possibilities within general visual search engines, such as YouTube.

Our research has potential impact for the visual search capability in our application domain. Analysts do not have to memorize all available concepts and can use natural language queries to retrieve relevant videos. Based on this thesis queries on a higher level, such as events, and specific queries can, in theory, be interpreted in a good manner. Because our methods are transparent, the analyst can interact with the system both on concept level and video level. We expect that analysts are more willing to provide feedback, because this feedback will be valuable in the future. Additionally, our feedback interpretation, query interpretation and fusion methods are scalable towards the increasing number of videos and concepts.

Although the potential impact of this thesis in our application domain is high, the modus operandi of analysts will probably not directly change because of the potential impact. In the near future, we expect that a search capability is possible. With current advances in deep learning, we expect that the (near real-time) indexing of the increasing number of video streams will be possible. With the advances in indexing and the results of this thesis, we can provide a visual search system that can be used as a baseline. Analysts, however, need systems with a high search effectiveness. This high effectiveness can only be achieved in cases in which the relevant concepts are available. Our methods can improve performance, but we are highly dependent on the availability and reliability of the pre-trained concepts. Although deep learning is a highly evolving field, good datasets for specific application domains such as security are still sparse. An option could be to use a set of pre-trained concepts and apply incremental learning mechanisms to train new concepts when needed. An example of such a situations would be in a situation in which no (near-)direct match is available. In this case, the system can provide the analyst a choice to either select a less related concept or train a novel concept detector.

A first step for future work is to collaborate with the analysts to optimally match their needs. We have to verify whether our assumptions hold in their use cases, whether they indeed have many ad-hoc queries, and how they would like to work with this search capability. This first step should be done by connecting all components and build a good user interface as a prototype. A general user interface as those used in general search engines might not be the best way for an analyst to do their job. With a proper user interface, prototype and in domain dataset, we could measure the performance within the application domain. A first direction for improving performance is the feedback interpretation component, for example by making ARF more robust to wrongly selected concept detectors by identifying whether the concept is not relevant or the detector is not performing well. Another direction is to work on ambiguity. In this thesis, we have not explicitly experimented with ambiguous queries or concepts, but it is important to be able to deal with ambiguity. A third direction is the phrasing of the query. We have now assumed that analysts would like to enter a natural language query, but we have both experimented with full queries and 'keywords' (events). Another option could be structured queries with logical combinations, such as AND and ORs between words, but also temporal relations such as BEFORE or spatial relations such as IN. A fourth direction is the scoring and ranking, particularly the alternatives to a linear weighted sum and a blind fusion of data sources, such as non-linear combinations

and weighted fusion. A fifth direction is in the indexing part. We have not fully exploited the spatial and temporal information in the videos, the training and combination of subevents, or the transfer from the concept detectors trained in another domain into our event retrieval domain. As a final direction the implicit mapping could be further explored, for example through generative models (Chen et al., 2016) or multi-task learned models, such as in the recent multi-task deep reinforcement learning architecture of Google (Oh et al., 2017).

With these directions of research, it is still not clear whether the performance will be good enough for analysts to work with the system. Although we have increased performance significantly from below 10% MAP up to 20%, we cannot predict whether these previously mentioned directions will reach the desired level of performance that is probably up to 90%. We might have to reach to other innovative, and not yet explored directions. These directions will probably not include a set of concepts (such as with deep learning), or even concepts in general.

Our work can also be applied outside a specific application domain. We foresee that future visual search capabilities for general purpose will rely on deep learning techniques. General purpose visual search engines, such as YouTube, have a massive amount of data, which can be used as training data. Currently, the majority of the general visual search engines do not offer content-based search, but we expect this type of search in the future. Our exploration on the Visual QA challenge in chapter 7 shows the possibilities for a closed loop system using deep learning and that field is still evolving rapidly. Because the general purpose systems value high performance over transparency, deep learning is likely to be used in this field. We cannot predict whether these systems will use pre-trained concepts, or use one big neural network. On the one hand, many of the current closed loop systems do not need explicit concepts. On the other hand, the field of explainable AI is growing, in which the concepts might have a role in explaining what is happening. But even if our query-to-concept mapping (i-w2v) has no use in the closed loop systems, we do foresee a role for our ARF and JRER method. First, our ARF method might be applicable as relevance feedback module for the deep learning model, for example to update weights in one of the layers. Second, a weighted version of the JRER method can be used as a method to combine outputs of several deep learning models. The JRER method is even applicable in other sciences that have multiple data sources. The methods proposed in this thesis are, thus, broadly usable.

BIBLIOGRAPHY

Aggarwal, J. K. and M. S. Ryoo (2011). "Human activity analysis: A review". In: *ACM Computing Surveys (CSUR)* 43.3, p. 16.

Ainsworth, T. (2002). "Buyer beware". In: *Security Oz* 19, pp. 18–26.

Aly, R., D. Hiemstra, F. de Jong, and P. M. Apers (2012). "Simulating the future of concept-based video retrieval under improved detector performance". In: *Multimedia Tools and Applications* 60.1, pp. 203–231.

Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C Lawrence Zitnick, and D. Parikh (2015). "Vqa: Visual question answering". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433.

Arroyo, R., J. J. Yebes, L. M. Bergasa, I. G. Daza, and J. Almazán (2015). "Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls". In: *Expert Systems with Applications* 42.21, pp. 7991–8005.

Atrey, P. K., M. A. Hossain, A. El Saddik, and M. S. Kankanhalli (2010). "Multimodal fusion for multimedia analysis: a survey". In: *Multimedia systems* 16.6, pp. 345–379.

Awad, G., J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quénot, M. Eskevich, R. Aly, and R. Ordelman (2016a). "Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking". In: *Proceedings of TRECVID*.

Awad, G., C. G. Snoek, A. F. Smeaton, and G. Quénot (2016b). "TRECVid semantic indexing of video: a 6-year retrospective". In: *ITE Transactions on Media Technology and Applications* 4.3, pp. 187–208.

Baeza-Yates, R., B. Ribeiro-Neto, et al. (1999). *Modern information retrieval*. Vol. 463. ACM press New York.

Baeza-Yates, R., M. Ciaramita, P. Mika, and H. Zaragoza (2008). "Towards semantic search". In: *Natural Language and Information Systems*. Springer, pp. 4–11.

Bagdanov, A. D., M. Bertini, A. Del Bimbo, G. Serra, and C. Torniai (2007). "Semantic annotation and retrieval of video events using multimedia ontologies". In: *International Conference on Semantic Computing*. IEEE, pp. 713–720.

Bai, L., S. Lao, G. J. Jones, and A. F. Smeaton (2007). "Video semantic content analysis based on ontology". In: *Machine Vision and Image Processing Conference, 2007. IMVIP 2007. International*. IEEE, pp. 117–124.

Ballan, L., M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra (2011). "Event detection and recognition for semantic annotation of video". In: *Multimedia Tools and Applications* 51.1, pp. 279–302.

Bay, H., A. Ess, T. Tuytelaars, and L. Van Gool (2008). "Speeded-up robust features (SURF)". In: *Computer vision and image understanding* 110.3, pp. 346–359.

Bodner, R. C. and F. Song (1996). "Knowledge-based approaches to query expansion in information retrieval". In: Springer Berlin Heidelberg, pp. 146–158.

Boer, M de, S Reitsma, and K Schutte (2016a). "Counting in Visual Question Answering". In: *DIR2016*, pp. 1–4.

Boer, M. de, K. Schutte, and W. Kraaij (2013). "Event Classification using Concepts". In: *ICT-Open*, pp. 39–42.

Boer, M. de, L. Daniele, P. Brandt, and M. Sappelli (2015a). "Applying semantic reasoning in image retrieval". In: *Proc. ALLDATA*, pp. 69 –74.

Boer, M. de, K. Schutte, and W. Kraaij (2015b). "Knowledge based query expansion in complex multimedia event detection". In: *Multimedia Tools and Applications*, pp. 1–19.

Boer, M. de, G. Pingen, D. Knook, K. Schutte, and W. Kraaij (2017a). "Improving video event retrieval by user feedback". In: *Multimedia Tools and Applications*, pp. 1–21.

Boer, M. H. de, P. Brandt, M. Sappelli, L. M. Daniele, K. Schutte, and W. Kraaij (2015c). "Query Interpretation–an Application of Semiotics in Image Retrieval". In: *International Journal On Advances in Software*. vol 3 & 4 8, pp. 435 –449.

Boer, M. H. de, K. Schutte, H. Zhang, Y.-J. Lu, C.-W. Ngo, and W. Kraaij (2016b). "Blind late fusion in multimedia event retrieval". In: *International journal of multimedia information retrieval* 5.4, pp. 203–217.

Boer, M. H. de, Y.-J. Lu, H. Zhang, K. Schutte, C.-W. Ngo, and W. Kraaij (2017b). "Semantic Reasoning in Zero Example Video Event Retrieval". In: *to appear*.

Bouma, H., J. Baan, S. Landsmeer, C. Kruszynski, G. van Antwerpen, and J. Dijk (2013a). "Real-time tracking and fast retrieval of persons in multiple surveillance cameras of a shopping mall". In: *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 87560A–87560A.

Bouma, H., G. Azzopardi, M. Spitters, J. de Wit, C. Versloot, R. van der Zon, P. Eendebak, J. Baan, J.-M. ten Hove, A. van Eekeren, F. ter Haar, R. den Hollander, J. van Huis, M. de Boer, G. van Antwerpen, J. Broekhuijsen, L. Daniele, P. Brandt, J. Schavemaker, W. Kraaij, and K. Schutte (2013b). "TNO at TRECVID 2013: Multimedia Event Detection and Instance Search". In: *Proceedings of TRECVID 2013*.

Bouma, H., J. Baan, G. J. Burghouts, P. T. Eendebak, J. R. van Huis, J. Dijk, and J. H. van Rest (2014). "Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall". In: *SPIE Security+ Defence*. International Society for Optics and Photonics, 92530F–92530F.

Bouma, H., P. T. Eendebak, K. Schutte, G. Azzopardi, and G. J. Burghouts (2015). "Incremental concept learning with few training examples and hierarchical classification". In: *SPIE Security+ Defence*. International Society for Optics and Photonics, 96520E–96520E.

Burgess, J. and J. Green (2013). *YouTube: Online video and participatory culture*. ISBN-13: 978-0745644790. John Wiley & Sons.

Burghouts, G., K Schutte, R.-M. ten Hove, S. van den Broek, J Baan, O Rajadell, J. van Huis, J van Rest, P Hanckmann, H Bouma, et al. (2014). "Instantaneous threat detection based on a semantic representation of activities, zones and trajectories". In: *Signal, Image and Video Processing* 8.1, pp. 191–200.

Caputo, B., H. Müller, J. Martinez-Gomez, M. Villegas, B. Acar, N. Patricia, N. Marvasti, S. Üsküdarlı, R. Paredes, M. Cazorla, et al. (2014). "ImageCLEF 2014:

Overview and analysis of the results". In: *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*. Springer, pp. 192–211.

Carpineto, C. and G. Romano (2012). "A survey of automatic query expansion in information retrieval". In: *ACM Computing Surveys (CSUR)* 44.1, pp. 1.

Chamasemani, F. F., L. S. Affendey, N. Mustapha, and F. Khalid (2015). "A Framework for Automatic Video Surveillance Indexing and Retrieval". In: *Research Journal of Applied Sciences, Engineering and Technology* 10.11, pp. 1316–1321.

Chang, X., Y. Yang, A. G. Hauptmann, E. P. Xing, and Y. Yu (2015). "Semantic Concept Discovery for Large-Scale Zero-Shot Event Detection." In: *IJCAI*. Vol. 2. 5.4, p. 6.

Chang, X., Y. Yang, G. Long, C. Zhang, and A. G. Hauptmann (2016). "Dynamic Concept Composition for Zero-Example Event Detection." In: *AAAI*, pp. 3464–3470.

Chen, J., Y. Cui, G. Ye, D. Liu, and S.-F. Chang (2014). "Event-Driven Semantic Concept Discovery by Exploiting Weakly Tagged Internet Images". In: *Proceedings of International Conference on Multimedia Retrieval*. ACM, p. 1.

Chen, W., J. T. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen (2015). "Compressing Neural Networks with the Hashing Trick". In: *Proceedings of The 32nd International Conference on Machine Learning*, 2285–2294.

Chen, X., Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel (2016). "Infogan: Interpretable representation learning by information maximizing generative adversarial nets". In: *Advances in Neural Information Processing Systems*, pp. 2172–2180.

Chen, X., A. Shrivastava, and A. Gupta (2013). "Neil: Extracting visual knowledge from web data". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1409–1416.

Cochran, W. G. and G. M. Cox (1957). "Experimental designs". In: *Wiley*.

Cremer, F., K. Schutte, J. G. Schavemaker, and E. den Breejen (2001). "A comparison of decision-level sensor-fusion methods for anti-personnel landmine detection". In: *Information fusion* 2.3, pp. 187–208.

Crucianu, M., M. Ferecatu, and N. Boujemaa (2004). "Relevance feedback for image retrieval: a short survey". In: *Report of the DELOS2 European Network of Excellence (FP6)*.

Dalal, N. and B. Triggs (2005). "Histograms of oriented gradients for human detection". In: *CVPR*. Vol. 1. IEEE, pp. 886–893.

Dalal, N., B. Triggs, and C. Schmid (2006). "Human detection using oriented histograms of flow and appearance". In: *European Conf. on computer vision*. Springer, pp. 428–441.

Dalton, J., J. Allan, and P. Mirajkar (2013). "Zero-shot video retrieval using content and concepts". In: *Proc. of the 22nd ACM Int. Conf. Information & Knowledge Management*. ACM, pp. 1857–1860.

Davies, A. C. and S. A. Velastin (2005). "A progress review of intelligent CCTV surveillance systems". In: *Proc. IEEE IDAACS*, pp. 417–423.

De Marneffe, M.-C., B. MacCartney, C. D. Manning, et al. (2006). "Generating typed dependency parses from phrase structure parses". In: *Proceedings of LREC*. Vol. 6. 2006. Genoa, pp. 449–454.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 248–255.

Deng, L. (2014). "A tutorial survey of architectures, algorithms, and applications for deep learning". In: *APSIPA Transactions on Signal and Information Processing* 3.

Deselaers, T., R. Paredes, E. Vidal, and H. Ney (2008). "Learning weighted distances for relevance feedback in image retrieval". In: *19th Int. Conf. on Pattern Recognition, 2008, ICPR 2008*. IEEE, pp. 1–4.

Dong, J., X. Li, and C. G. Snoek (2016). "Word2VisualVec: Cross-media retrieval by visual feature prediction". In: *arXiv preprint arXiv:1604.06838*.

Elhoseiny, M., J. Liu, H. Cheng, H. S. Sawhney, and A. M. Elgammal (2016). "Zero-Shot Event Detection by Multimodal Distributional Semantic Embedding of Videos." In: *AAAI*, pp. 3478–3486.

Enser, P. G., C. J. Sandom, J. S. Hare, and P. H. Lewis (2007). "Facing the reality of semantic image retrieval". In: *Journal of Documentation* 63.4, pp. 465–481.

Erozel, G., N. K. Cicekli, and I. Cicekli (2008). "Natural language querying for video databases". In: *Information Sciences* 178.12, pp. 2534–2552.

Everingham, M., S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman (2015). "The pascal visual object classes challenge: A retrospective". In: *International Journal of Computer Vision* 111.1, pp. 98–136.

Ferryman, J and A Shahrokni (2009). "Pets2009: Dataset and challenge". In: *IEEE International Workshop on PETS-Winter*. IEEE, pp. 1–6.

Francois, A. R., R. Nevatia, J. Hobbs, R. C. Bolles, and J. R. Smith (2005). "VERL: an ontology framework for representing and annotating video events". In: *MultiMedia, IEEE,* 12.4, pp. 76–86.

Fukui, A., D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach (2016). "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding". In: *arXiv preprint arXiv:1606.01847*.

Georis, B., M Maziere, F Bremond, and M. Thonnat (2004). "A video interpretation platform applied to bank agency monitoring". In: *IEEE Intelligent Surveillance Systems (IDSS-04)*, pp. 46–50.

Gia, G., F. Roli, et al. (2004). "Instance-based relevance feedback for image retrieval". In: *Advances in neural information processing systems*, pp. 489–496.

Goldberg, Y. and O. Levy (2014). "word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method". In: *arXiv preprint arXiv:1402.3722*.

Habibian, A., K. E. van de Sande, and C. G. Snoek (2013). "Recommendations for video event recognition using concept vocabularies". In: *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, pp. 89–96.

Habibian, A., T. Mensink, and C. G. Snoek (2014a). "Composite concept discovery for zero-shot video event detection". In: *Proc. of Int. Conf. on Multimedia Retrieval*. ACM, p. 17.

Habibian, A., T. Mensink, and C. G. Snoek (2014b). "Videostory: A new multimedia embedding for few-example recognition and translation of events". In: *Proc of Int. Conf. on Multimedia*. ACM, pp. 17–26.

Hare, J. S., P. H. Lewis, P. G. Enser, and C. J. Sandom (2006). "Mind the gap: another look at the problem of the semantic gap in image retrieval". In: *Electronic Imaging 2006*. International Society for Optics and Photonics, pp. 607309–607309.

Hartley, E. (2004). "Bound together: Signs and features in multimedia content representation". In: *Cosign Conference*.

Hassan, S. and R. Mihalcea (2011). "Semantic Relatedness Using Salient Semantic Analysis." In: *Proceedings of AAAI Confences on Artificial Intelligence*, pp. 884–889.

Hauptmann, A., R. Yan, W.-H. Lin, M. Christel, and H. Wactlar (2007a). "Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news". In: *IEEE Trans. on Multimedia,* 9.5, pp. 958–966.

Hauptmann, A., R. Yan, and W.-H. Lin (2007b). "How many high-level concepts will fill the semantic gap in news video retrieval?" In: *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, pp. 627–634.

Hauptmann, A. G. and M. G. Christel (2004). "Successful approaches in the TREC video retrieval evaluations". In: *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, pp. 668–675.

He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hoffart, J., F. M. Suchanek, K. Berberich, and G. Weikum (2013). "YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia". In: *Artificial Intelligence* 194, pp. 28–61.

Hoque, E., O. Hoeber, G. Strong, and M. Gong (2013). "Combining conceptual query expansion and visual search results exploration for Web image retrieval". In: *Journal of Ambient Intelligence and Humanized Computing* 4.3, pp. 389–400.

Hu, W., T. Tan, L. Wang, and S. Maybank (2004). "A survey on visual surveillance of object motion and behaviors". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34.3, pp. 334–352.

Hu, W., N. Xie, L. Li, X. Zeng, and S. Maybank (2011). "A survey on visual content-based video indexing and retrieval". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41.6, pp. 797–819.

Huurnink, B., K. Hofmann, and M. De Rijke (2008). "Assessing concept selection for video retrieval". In: *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, pp. 459–466.

Jain, M., J. C. van Gemert, T. Mensink, and C. G. Snoek (2015). "Objects2action: Classifying and localizing actions without any video example". In: *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 4588–4596.

Jegou, H., F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid (2012). "Aggregating local image descriptors into compact codes". In: *Trans. on Pattern Analysis and Machine Intelligence* 34.9, pp. 1704–1716.

Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell (2014). "Caffe: Convolutional architecture for fast feature embedding". In: *Proc. of Int. Conf. on Multimedia*. ACM, pp. 675–678.

Jiang, L., D. Meng, T. Mitamura, and A. G. Hauptmann (2014a). "Easy samples first: Self-paced reranking for zero-example multimedia search". In: *Proc. of the ACM Int. Conf. on Multimedia*. ACM, pp. 547–556.

Jiang, L., T. Mitamura, S.-I. Yu, and A. G. Hauptmann (2014b). "Zero-example event search using multimodal pseudo relevance feedback". In: *Proceedings of International Conference on Multimedia Retrieval*. ACM, p. 297.

Jiang, L., S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann (2015). "Bridging the ultimate semantic gap: A semantic search engine for internet videos". In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, pp. 27–34.

Jiang, Y.-G., S. Bhattacharya, S.-F. Chang, and M. Shah (2012). "High-level event recognition in unconstrained videos". In: *Int. Journal of Multimedia Information Retrieval*, pp. 1–29.

Jiang, Y.-G., Z. Wu, J. Wang, X. Xue, and S.-F. Chang (2017). "Exploiting feature and class relationships in video categorization with regularized deep neural networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kaliciak, L., D. Song, N. Wiratunga, and J. Pan (2013). "Combining visual and textual systems within the context of user feedback". In: *Advances in Multimedia Modeling*. Springer, pp. 445–455.

Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei (2014). "Large-scale Video Classification with Convolutional Neural Networks". In: *CVPR*, pp. 1725–1732.

Kato, T. (1992). "Database architecture for content-based image retrieval". In: *SPIE/IS&T 1992 symposium on electronic imaging: science and technology*. International Society for Optics and Photonics, pp. 112–123.

Kelsh, C. (2016). *Do body cameras change how police interact with the public?* URL: `https://journalistsresource.org/studies/government/criminal-justice/body-cameras-police-interact-with-public`.

Kennedy, L. and A. Hauptmann (2006). "LSCOM lexicon definitions and annotations (version 1.0)". In:

Kim, J.-H., S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang (2016). "Multimodal residual learning for visual qa". In: *Advances in Neural Information Processing Systems*, pp. 361–369.

Kiros, R., Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler (2015). "Skip-thought vectors". In: *Advances in Neural Information Processing Systems*, pp. 3276–3284.

Kittler, J., M. Hatef, R. P. Duin, and J. Matas (1998). "On combining classifiers". In: *IEEE transactions on pattern analysis and machine intelligence* 20.3, pp. 226–239.

Ko, T. (2008). "A survey on behavior analysis in video surveillance for homeland security applications". In: *Applied Imagery Pattern Recognition Workshop, 2008*. IEEE, pp. 1–8.

Kotov, A. and C. Zhai (2012). "Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries". In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, pp. 403–412.

Kraaij, W., T. Westerveld, and D. Hiemstra (2002). "The importance of prior probabilities for entry page search". In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 27–34.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*, pp. 1097–1105.

Kuo, W., B. Hariharan, and J. Malik (2015). "Deepbox: Learning objectness with convolutional networks". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2479–2487.

Lan, Z.-z., L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann (2012). "Double fusion for multimedia event detection". In: *Advances in Multimedia Modeling*. Springer, pp. 173 –185.

Landis, J. R. and G. G. Koch (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, pp. 159–174.

Laptev, I. and T. Lindeberg (2003). "Space-time interest points". In: *9th International Conference on Computer Vision, Nice, France*. IEEE, pp. 432 –439.

Laptev, I., M. Marszalek, C. Schmid, and B. Rozenfeld (2008). "Learning realistic human actions from movies". In: *Computer Vision and Pattern Recognition*. IEEE, pp. 1–8.

Lee, L, R Romano, and G Stein (2000). "Introduction to the special section on video surveillance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8, p. 745.

Leong, C. W., S. Hassan, M. E. Ruiz, and R. Mihalcea (2011). "Improving query expansion for image retrieval via saliency and picturability". In: *Multilingual and Multimodal Information Access Evaluation*. Springer, pp. 137–142.

Lev, G., B. Klein, and L. Wolf (2015). "In defense of word embedding for generic text representation". In: *Natural Language Processing and Information Systems*. Sprin- ger, pp. 35–50.

Levy, O. and Y. Goldberg (2014). "Neural word embedding as implicit matrix factorization". In: *Advances in Neural Information Processing Systems*, pp. 2177–2185.

Levy, O., Y. Goldberg, and I. Dagan (2015). "Improving distributional similarity with lessons learned from word embeddings". In: *Trans. of the Association for Computational Linguistics* 3, pp. 211–225.

Lew, M. S., N. Sebe, C. Djeraba, and R. Jain (2006). "Content-based multimedia information retrieval: State of the art and challenges". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 2.1, pp. 1–19.

Lewis, D. D. (1998). "Naive (Bayes) at forty: The independence assumption in information retrieval". In: *European conference on machine learning*. Springer, pp. 4–15.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick (2014). "Microsoft coco: Common objects in context". In: *European Conference on Computer Vision*. Springer, pp. 740–755.

Liu, X. H. (2002). "Semantic understanding and commonsense reasoning in an adaptive photo agent". PhD thesis. Massachusetts Institute of Technology.

Liu, Y., D. Zhang, G. Lu, and W.-Y. Ma (2007). "A survey of content-based image retrieval with high-level semantics". In: *Pattern Recognition* 40.1, pp. 262–282.

Lowe, D. G. (2004). "Distinctive image features from scale-invariant keypoints". In: *Int. J. of computer vision* 60.2, pp. 91–110.

Lu, J., J. Yang, D. Batra, and D. Parikh (2016a). "Hierarchical question-image co-attention for visual question answering". In: *Advances In Neural Information Processing Systems*, pp. 289–297.

Lu, Y.-J., H. Zhang, M. de Boer, and C.-W. Ngo (2016b). "Event detection with zero example: select the right and suppress the wrong concepts". In: *Proc. of the 2016 ACM on Int. Conf. on Multimedia Retrieval*. ACM, pp. 127–134.

Ma, A. J., P. C. Yuen, and J.-H. Lai (2013). "Linear dependency modeling for classifier fusion and feature combination". In: *IEEE transactions on pattern analysis and machine intelligence* 35.5, pp. 1135–1148.

Ma, Z., Y. Yang, Y. Cai, N. Sebe, and A. G. Hauptmann (2012). "Knowledge adaptation for ad hoc multimedia event detection with few exemplars". In: *Proceedings of the 20th ACM international conference on Multimedia*. ACM, pp. 469–478.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014). "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

Mascardi, V., V. Cordì, and P. Rosso (2007). "A Comparison of Upper Ontologies." In: *WOA*, pp. 55–64.

Mazloom, M., A. Habibian, and C. G. Snoek (2013a). "Querying for Video Events by Semantic Signatures from Few Examples". In: *MM'13*, pp. 609–612.

Mazloom, M., E. Gavves, K. van de Sande, and C. Snoek (2013b). "Searching informative concept banks for video event detection". In: *Proc. of the 3rd Int. Conf. on Multimedia Retrieval*. ACM, pp. 255–262.

Mc Donald, K. and A. F. Smeaton (2005). "A comparison of score, rank and probability-based fusion methods for video shot retrieval". In: *International Conference on Image and Video Retrieval*. Springer, pp. 61–70.

Mensink, T., E. Gavves, and C. G. Snoek (2014). "COSTA: Co-occurrence statistics for zero-shot classification". In: *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 2441–2448.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Miller, G. A. (1995). "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11, pp. 39–41.

Milne, D. and I. H. Witten (2013). "An open-source toolkit for mining Wikipedia". In: *Artificial Intelligence* 194, pp. 222–239.

Mladenić, D. (1998). “Feature subset selection in text-learning”. In: *European Conference on Machine Learning*. Springer, pp. 95–100.

Mould, N., J. L. Regens, C. J. Jensen III, and D. N. Edger (2014). “Video surveillance and counterterrorism: the application of suspicious activity recognition in visual surveillance systems to counterterrorism”. In: *Journal of Policing, Intelligence and Counter Terrorism* 9.2, pp. 151–175.

Mukaka, M. (2012). “A guide to appropriate use of Correlation coefficient in medical research”. In: *Malawi Medical Journal* 24.3, pp. 69–71.

Myers, G. K., R. Nallapati, J. van Hout, S. Pancoast, R. Nevatia, C. Sun, A. Habibian, D. C. Koelma, K. E. van de Sande, A. W. Smeulders, et al. (2014). “Evaluating multimedia features and fusion for example-based event detection”. In: *Machine Vision and Applications* 25.1, pp. 17–32.

Naphade, M., J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis (2006). “Large-scale concept ontology for multimedia”. In: *MultiMedia, IEEE* 13.3, pp. 86–91.

Natarajan, P., P. Natarajan, V. Manohar, S. Wu, S. Tsakalidis, S. N. Vitaladevuni, X. Zhuang, R. Prasad, G. Ye, D. Liu, et al. (2011). “Bbn viser trecvid 2011 multimedia event detection system”. In: *NIST TRECVID Workshop*.

Natarajan, P., S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, and R. Prasad (2012). “Multimodal feature fusion for robust event detection in web videos”. In: *CVPR*. IEEE, pp. 1298–1305.

Natsev, A. P., A. Haubold, J. Tešić, L. Xie, and R. Yan (2007). “Semantic concept-based query expansion and re-ranking for multimedia retrieval”. In: *Proc. of the 15th Int. Conf. on Multimedia*. ACM, pp. 991–1000.

Neo, S.-Y., J. Zhao, M.-Y. Kan, and T.-S. Chua (2006). “Video retrieval using high level features: Exploiting query matching and confidence-based weighting”. In: *International Conference on Image and Video Retrieval*. Springer, pp. 143–152.

Ngo, C.-W., Y.-J. Lu, H. Zhang, T. Yao, C.-C. Tan, L. Pang, M. de Boer, J. Schavemaker, K. Schutte, and W. Kraaij (2014). “VIREO-TNO @ TRECVID 2014: Multimedia Event Detection and Recounting (MED and MER)”. In: *Proceedings of TRECVID 2014*.

Noh, H., P. Hongsuck Seo, and B. Han (2016). “Image question answering using convolutional neural network with dynamic parameter prediction”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 30–38.

Norouzi, M., T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean (2013). “Zero-shot learning by convex combination of semantic embeddings”. In: *arXiv preprint arXiv:1312.5650*.

Novak, C. L. and S. A. Shafer (1992). “Anatomy of a color histogram”. In: *Computer Vision and Pattern Recognition*. IEEE, pp. 599–605.

Oh, J., S. Singh, H. Lee, and P. Kohli (2017). “Zero-Shot Task Generalization with Multi-Task Deep Reinforcement Learning”. In: *arXiv preprint arXiv:1706.05064*.

Oh, S., S. McCloskey, I. Kim, A. Vahdat, K. J. Cannons, H. Hajimirsadeghi, G. Mori, A. A. Perera, M. Pandey, and J. J. Corso (2014). “Multimedia event detection with multimodal feature fusion and temporal concept localization”. In: *Machine vision and applications* 25.1, pp. 49–69.

Ojala, T., M. Pietikainen, and T. Maenpaa (2002). "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.7, pp. 971–987.

Oltramari, A. and C. Lebiere (2012). "Using ontologies in a cognitive-grounded system: automatic action recognition in video surveillance". In: *Proceedings of the 7th International Conference on Semantic Technology for Intelligence, Defense, and Security*. Citeseer.

Over, P., C. Leung, H. Ip, and M. Grubinger (2004). "Multimedia retrieval benchmarks". In: *Multimedia* 11.2, pp. 80–84.

Over, P., G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quot (2013). "TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics". In: *Proceedings of TRECVID 2013*. NIST, USA.

Over, P., G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quot (2014). "TRECVID 2014 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics". In: *Proc. of TRECVID 2014*. NIST, USA.

Over, P., G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, G. Quot, and R. Ordelman (2015). "TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics". In: *Proc. of TRECVID 2015*. NIST, USA.

Patil, P. B. and M. B. Kokare (2011). "Relevance Feedback in Content Based Image Retrieval: A Review." In: *Journal of Applied Computer Science & Mathematics* 10.10, pp. 40–47.

Patil, S. (2012). "A comprehensive review of recent relevance feedback techniques in CBIR". In: *Int. Journal of Engineering Research & Technology (IJERT)* 1.6.

Pedersen, T., S. Patwardhan, and J. Michelizzi (2004). "WordNet:: Similarity: measuring the relatedness of concepts". In: *Demonstration Papers at HLT-NAACL 2004*. Association for Computational Linguistics, pp. 38–41.

Pennington, J., R. Socher, and C. D. Manning (2014). "Glove: Global Vectors for Word Representation." In: *EMNLP*. Vol. 14, pp. 1532–1543.

Perona, P. (2010). "Vision of a Visipedia". In: *Proceedings of the IEEE* 98.8, pp. 1526–1534.

Perronnin, F., J. Sánchez, and T. Mensink (2010). "Improving the fisher kernel for large-scale image classification". In: *Computer Vision–ECCV 2010*, pp. 143–156.

Pingen, G., M. de Boer, and R. Aly (2017). "Rocchio-Based Relevance Feedback in Video Event Retrieval". In: *International Conference on Multimedia Modeling*. Sprin- ger, pp. 318–330.

Pisanelli, D (2004). "Biodynamic ontology: applying BFO in the biomedical domain". In: *Ontologies in medicine* 102, pp. 20.

Platt, J. et al. (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3, pp. 61–74.

Ravana, S. D. and A. Moffat (2009). "Score aggregation techniques in retrieval experimentation". In: *Proceedings of the Twentieth Australasian Conference on Australasian Database-Volume 92*. Australian Computer Society, Inc., pp. 57–66.

Reese, H. (2015). *Police are now using drones to apprehend suspects and administer non-lethal force: A police chief weighs in.* URL: `http://www.techrepublic.`

com/article/police-are-now-using-drones-to-apprehend-suspects-and-administer-non-lethal-force-a-police-chief/.

Robertson, S. E. and K. S. Jones (1976). "Relevance weighting of search terms". In: *Journal of the American Society for Information science* 27.3, pp. 129–146.

Rocchio, J. J. (1971). "Relevance feedback in information retrieval". In:

Rocha, R., P. T. Saito, and P. H. Bugatti (2015). "A Novel Framework for Content-Based Image Retrieval Through Relevance Feedback Optimization". In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, pp. 281–289.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. (2015). "Imagenet large scale visual recognition challenge". In: *International Journal of Computer Vision* 115.3, pp. 211–252.

Ruthven, I. and M. Lalmas (2003). "A survey on the use of relevance feedback for information access systems". In: *The Knowledge Engineering Review* 18.02, pp. 95–145.

Sakai, T., T. Manabe, and M. Koyama (2005). "Flexible pseudo-relevance feedback via selective sampling". In: *ACM Transactions on Asian Language Information Processing (TALIP)* 4.2, pp. 111–135.

Salton, G. and C. Buckley (1988). "Term-weighting approaches in automatic text retrieval". In: *Information Processing & Management* 24.5, pp. 513–523.

Salton, G., A. Singhal, M. Mitra, and C. Buckley (1997). "Automatic text structuring and summarization". In: *Information Processing & Management* 33.2, pp. 193–207.

Schavemaker, J., M. Spitters, G. Koot, and M. de Boer (2015). "Fast re-ranking of visual search results by example selection". In: *International Conference on Computer Analysis of Images and Patterns*. Springer, pp. 387–398.

Schmidhuber, J. (2015). "Deep learning in neural networks: An overview". In: *Neural Networks* 61, pp. 85–117.

Schutte, K., F. Bomhof, G. Burghouts, J. van Diggelen, P. Hiemstra, J. van't Hof, W. Kraaij, H. Pasman, A. Smith, C. Versloot, et al. (2013). "GOOSE: semantic search on internet connected sensors". In: *SPIE Defense, Security, and Sensing*. Int. Society for Optics and Photonics, pp. 875806–875806.

Schutte, K., H. Bouma, J. Schavemaker, L. Daniele, M. Sappelli, G. Koot, P. Eendebak, G. Azzopardi, M. Spitters, M. de Boer, et al. (2015a). "Interactive detection of incrementally learned concepts in images with ranking and semantic query interpretation". In: *13th Int. Workshop on Content-Based Multimedia Indexing (CBMI), 2015*. IEEE, pp. 1–4.

Schutte, K., H. Bouma, J. Schavemaker, L. Daniele, M. Sapelli, G. Koot, P. Eendebak, G. Azzopardi, M. Spitters, and M. de Boer (2015b). *TOSO dataset*.

Schutte, K., G. Burghouts, N. van der Stap, V. Westerwoudt, H. Bouma, M. Kruithof, J. Baan, and J.-M. ten Hove (2016). "Long-term behavior understanding based on the expert-based combination of short-term observations in high-resolution CCTV". In: *SPIE Security+ Defence*. International Society for Optics and Photonics, 99950P–99950P.

Schütze, H. (2008). “Introduction to Information Retrieval”. In: *Proceedings of the international communication of association for computing machinery conference*.

Sharif Razavian, A., H. Azizpour, J. Sullivan, and S. Carlsson (2014). “CNN features off-the-shelf: an astounding baseline for recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813.

Sheth, A., C. Ramakrishnan, and C. Thomas (2005). “Semantics for the semantic web: The implicit, the formal and the powerful”. In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 1.1, pp. 1–18.

Simonyan, K. and A. Zisserman (2014). “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556*.

Sivic, J. and A. Zisserman (2006). “Video Google: Efficient visual search of videos”. In: *Toward category-level object recognition*. Springer, pp. 127–144.

Sleator, D. D. and D. Temperley (1995). “Parsing English with a link grammar”. In: *arXiv preprint cmp-lg/9508004*.

Smeaton, A. F., P. Over, and W. Kraaij (2006). “Evaluation campaigns and TRECVid”. In: *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, pp. 321–330.

Smeulders, A. W., M. Worring, S. Santini, A. Gupta, and R. Jain (2000). “Content-based image retrieval at the end of the early years”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.12, pp. 1349–1380.

Snoek, C. G. and M. Worring (2008). “Concept-based video retrieval”. In: *Foundations and Trends in Information Retrieval* 2.4, pp. 215–322.

Snoek, C. G. and A. W. Smeulders (2010). “Visual-concept search solved?” In: *IEEE Computer* 43.6, pp. 76–78.

Snoek, C. G., S. Cappallo, D. Fontijne, D. Julian, D. C. Koelma, P. Mettes, K. Sande, A. Sarah, H. Stokman, R. B. Towal, et al. (2015). “Qualcomm Research and University of Amsterdam at TRECVID 2015: Recognizing Concepts, Objects, and Events in Video”. In: *TRECVID 2015*.

Spagnola, S. and C. Lagoze (2011). “Edge dependent pathway scoring for calculating semantic similarity in ConceptNet”. In: *Proc. of the Ninth Int. Conf. on Computational Semantics*. Association for Computational Linguistics, pp. 385–389.

Speer, R. and C. Havasi (2012). “Representing General Relational Knowledge in ConceptNet 5.” In: *LREC*, pp. 3679–3686.

Strassel, S., A. Morris, J. G. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel (2012). “Creating HAVIC: Heterogeneous Audio Visual Internet Collection.” In: *LREC*. Citeseer, pp. 2573–2577.

Sutskever, I., J. Martens, G. Dahl, and G. Hinton (2013). “On the importance of initialization and momentum in deep learning”. In: *Proceedings of the 30th international conference on machine learning (ICML-13)*, pp. 1139–1147.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich (2015). “Going deeper with convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2016). “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826.

Tamrakar, A., S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney (2012). "Evaluation of low-level features and their combinations for complex event detection in open source videos". In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 3681–3688.

Tao, D., X. Tang, and X. Li (2008). "Which components are important for interactive image searching?" In: *Circuits and Systems for Video Technology, IEEE Transactions on* 18.1, pp. 3–11.

Terrades, O. R., E. Valveny, and S. Tabbone (2009). "Optimal classifier fusion in a non-bayesian probabilistic framework". In: *IEEE transactions on pattern analysis and machine intelligence* 31.9, pp. 1630–1644.

Thomee, B., D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li (2015). "The new data and new challenges in multimedia research". In: *arXiv preprint arXiv:1503.01817*.

Tokmetzis, D. (2013). *Hoeveel camera's hangen er in Nederland?* Dutch. URL: `http://sargasso.nl/cameratoezicht-in-nederland-hoeveel-cameras-zijn-er-eigenlijk/`.

Tong, S. and E. Chang (2001). "Support vector machine active learning for image retrieval". In: *Proc. of the 9th ACM Int. Conf. on Multimedia*. ACM, pp. 107–118.

Tran, D., L. Bourdev, R. Fergus, L. Torresani, and M. Paluri (2015). "Learning spatiotemporal features with 3d convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497.

Truong, B. T. and S. Venkatesh (2007). "Video abstraction: A systematic review and classification". In: *ACM transactions on multimedia computing, communications, and applications (TOMM)* 3.1, p. 3.

Tsai, C.-F., Y.-H. Hu, and Z.-Y. Chen (2015). "Factors affecting rocchio-based pseudorelevance feedback in image retrieval". In: *Journal of the Association for Information Science and Technology* 66.1, pp. 40–57.

Tu, K., M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu (2014). "Joint video and text parsing for understanding events and answering queries". In: *MultiMedia, IEEE* 21.2, pp. 42–70.

Tulyakov, S., S. Jaeger, V. Govindaraju, and D. Doermann (2008). "Review of classifier combination methods". In: *Machine Learning in Document Analysis and Recognition*. Springer, pp. 361–386.

Tuytelaars, T., K. Mikolajczyk, et al. (2008). "Local invariant feature detectors: a survey". In: *Foundations and trends® in computer graphics and vision* 3.3, pp. 177–280.

Tzelepis, C., D. Galanopoulos, V. Mezaris, and I. Patras (2016). "Learning to detect video events from zero or very few video examples". In: *Image and vision Computing* 53, pp. 35–44.

Uijlings, J. R., K. E. Van De Sande, T. Gevers, and A. W. Smeulders (2013). "Selective search for object recognition". In: *International journal of computer vision* 104.2, pp. 154–171.

Valera, M. and S. A. Velastin (2005). "Intelligent distributed surveillance systems: a review". In: *IEE Proceedings-Vision, Image and Signal Processing* 152.2, pp. 192–204.

Van De Sande, K., T. Gevers, and C. Snoek (2010). "Evaluating color descriptors for object and scene recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 32.9, pp. 1582–1596.

Van De Weijer, J., C. Schmid, J. Verbeek, and D. Larlus (2009). "Learning color names for real-world applications". In: *IEEE Transactions on Image Processing* 18.7, pp. 1512 –1523.

Van Rijsbergen, C. (1979). "Information retrieval". In:

Vatant, B. and M. Wick (2012). *Geonames ontology*. URL: `http://www.geonames.org/ontology/`.

Vezzani, R., D. Baltieri, and R. Cucchiara (2013). "People reidentification in surveillance and forensics: A survey". In: *ACM Computing Surveys (CSUR)* 46.2, p. 29.

Vishwakarma, S. and A. Agrawal (2013). "A survey on activity recognition and behavior understanding in video surveillance". In: *The Visual Computer* 29.10, pp. 983–1009.

Von Ahn, L., M. Kedia, and M. Blum (2006). "Verbosity: a game for collecting common-sense facts". In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, pp. 75–78.

Voss, J. (2005). "Measuring wikipedia". In: *Proceedings of 10th International Conference of the International Society for Scientometrics and Informetrics*.

Wang, H. and D.-Y. Yeung (2016a). "Towards Bayesian deep learning: A framework and some existing methods". In: *IEEE Transactions on Knowledge and Data Engineering* 28.12, pp. 3395–3408.

Wang, H. and C. Schmid (2013). "Action recognition with improved trajectories". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551–3558.

Wang, J., J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong (2010a). "Locality-constrained linear coding for image classification". In: *CVPR*. IEEE, pp. 3360–3367.

Wang, X.-Y., J.-F. Wu, and H.-Y. Yang (2010b). "Robust image retrieval based on color histogram of local feature regions". In: *Multimedia Tools and Applications* 49.2, pp. 323–345.

Wang, X.-Y., L.-L. Liang, W.-Y. Li, D.-M. Li, and H.-Y. Yang (2016b). "A new SVM-based relevance feedback image retrieval using probabilistic feature and weighted kernel function". In: *Journal of Visual Communication and Image Representation* 38, pp. 256–275.

White, M. D. (2014). "Police officer body-worn cameras: Assessing the evidence". In: *Washington, DC: Office of Community Oriented Policing Services*.

Wilkins, P., P. Ferguson, and A. F. Smeaton (2006). "Using score distributions for query-time fusion in multimedia retrieval". In: *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, pp. 51–60.

Wu, Q., D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel (2017). "Visual question answering: A survey of methods and datasets". In: *Computer Vision and Image Understanding*.

Wu, S., S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan (2014). "Zero-shot event detection using multi-modal fusion of weakly supervised concepts". In: *Conf. on Computer Vision and Pattern Recognition*. IEEE, pp. 2665–2672.

Wu, Z. and M. Palmer (1994). "Verbs semantics and lexical selection". In: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 133–138.

Xiong, Y., K. Zhu, D. Lin, and X. Tang (2015). "Recognize complex events from static images by fusing deep channels". In: *Proc. CVPR*, pp. 1600–1609.

Xu, L., A. Krzyzak, and C. Y. Suen (1992). "Methods of combining multiple classifiers and their applications to handwriting recognition". In: *IEEE transactions on systems, man, and cybernetics* 22.3, pp. 418–435.

Xu, S., H. Li, X. Chang, S.-I. Yu, X. Du, X. Li, L. Jiang, Z. Mao, Z. Lan, S. Burger, et al. (2015). "Incremental Multimodal Query Construction for Video Search". In: *Proc. of the 5th ACM on Int. Conf. on Multimedia Retrieval*. ACM, pp. 675–678.

Yan, Y., Y. Yang, H. Shen, D. Meng, G. Liu, A. G. Hauptmann, and N. Sebe (2015). "Complex Event Detection via Event Oriented Dictionary Learning." In: *AAAI*, pp. 3841–3847.

Yang, L. and A. Hanjalic (2010). "Supervised reranking for web image search". In: *Proc. of the Int. Conf. on Multimedia*. ACM, pp. 183–192.

Ye, G., Y. Li, H. Xu, D. Liu, and S.-F. Chang (2015). "Eventnet: A large scale structured concept library for complex event detection in video". In: *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*. ACM, pp. 471–480.

Yilmaz, A., O. Javed, and M. Shah (2006). "Object tracking: A survey". In: *Acm computing surveys (CSUR)* 38.4, p. 13.

Yoon, J. (2006). "Improving recall of browsing sets in image retrieval from a semiotics perspective". PhD thesis. University of North Texas.

Yu, C. T. and G. Salton (1976). "Precision weighting- an effective automatic indexing method". In: *Journal of the ACM (JACM)* 23.1, pp. 76–88.

Yu, S.-I., L. Jiang, and A. Hauptmann (2014). "Instructional videos for unsupervised harvesting and learning of action examples". In: *Proc. of the ACM Int. Conf. on Multimedia*. ACM, pp. 825–828.

Zhan, B., D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu (2008). "Crowd analysis: a survey". In: *Machine Vision and Applications* 19.5-6, pp. 345–357.

Zhang, H., Y.-J. Lu, M. de Boer, F. ter Haar, Z. Qiu, K. Schutte, W. Kraaij, and C.-W. Ngo (2015a). "VIREO-TNO @ TRECVID 2015: Multimedia Event Detection". In: *Proc. of TRECVID 2015*.

Zhang, S., C. Wang, S.-C. Chan, X. Wei, and C.-H. Ho (2015b). "New object detection, tracking, and recognition approaches for video surveillance over camera network". In: *IEEE Sensors Journal* 15.5, pp. 2679–2691.

Zhang, Y., J. Chen, X. Huang, and Y. Wang (2015c). "A Probabilistic Analysis of Sparse Coded Feature Pooling and Its Application for Image Retrieval". In: *PloS one* 10.7, e0131721.

Zhao, L. and J. Callan (2012). "Automatic term mismatch diagnosis for selective query expansion". In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 515–524.

Zheng, L., S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian (2015). "Query-Adaptive Late Fusion for Image Search and Person Re-identification". In: *Computer Vision and Pattern Recognition*. Vol. 1.

Zhou, B., A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva (2014). “Learning deep features for scene recognition using places database”. In: *Advances in Neural Information Processing Systems*, pp. 487–495.

Zhou, X. S. and T. S. Huang (2003). “Relevance feedback in image retrieval: A comprehensive review”. In: *Multimedia systems* 8.6, pp. 536–544.

Zhu, Y., R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler (2015). “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 19–27.

Zitnick, C. L. and P. Dollár (2014). “Edge boxes: Locating object proposals from edges”. In: *Computer Vision–ECCV 2014*. Springer, pp. 391–405.

Zon, R. van der (2014). “A knowledge base approach for semantic interpretation and decomposition in concept based video retrieval”. MA thesis. TU Delft.

GLOSSARY

Blind Late Fusion Integrating information from data sources after classification without using pre-trained weights
Cardinality Number of elements
Classifier Function that uses characteristics of the data to identify which class the data belongs to
Concept Abstract idea representing the characteristics it represents, often the textual label of the concept detector
Concept Detector Classifier that can detect a certain concept
Concept Detector score Score representing the estimation of the probability that a certain concept is present in an image, keyframe or video
ConceptBank A set of concepts and their detectors, related to vocabulary
Concept Selection A process in which a specific set of concepts are selected from the ConceptBank
Convolutional Neural Network Type of (feedforward) neural network often used in image classification, based on convoluting part of the image. This is one of the architectures used in deep learning
Data Source A source that collects (raw) data that has not been processed into valuable information, related to modality
Deep Learning A branch in machine learning that uses deep networks to model data
Descriptor Feature vector describing an image
Determiner A word that is often used before a noun to provide context about the noun, such as 'the' or 'a'
Event (High-level) Long-term spatially and temporally dynamic object interactions that happen under certain scene settings (Jiang et al., 2012)
Feature A distinctive characteristic, such as color or edge
Frame Digital image that is sent to display image rendering devices
Keyframe A single image in a sequence of images / frames that represents an important point in the sequence. This is related to the compression of the raw video
Knowledge Base Organized repository for information
Object Label Symbolic name for the objects in the image
Ontology Explicit specification of the conceptualization of a domain, involving for example objects, properties and their relations
Max Pooling Aggregation strategy that summarizes the response across a set of responses by taking the maximum (spatially or temporally)
Mean Average Precision Evaluation metric that takes a ranked list of outputs and calculates performance based on the rank of the positive instances in the list
Modality The channel by which signs are transmitted, related to data source. Example modalities are speech, vision and motion

Paradigm A group of words that concern substitution and signification of functional contrasts
Query Representation of information need
Query Expansion Process of reformulating a query to improve retrieval, often by adding (weighted) query terms
Relevance Feedback Feedback on a certain result of a system
(lexical) Semantics Meaning of the words
Semantic Gap (video retrieval) The gap between the abstraction level of the pixels in a video and the semantic interpretation of the pixels (Smeulders et al., 2000)
Semantic Matching Identification of semantically related information
Semiotics Study of sign processes and meaning communication
Shot Series of frames
Skip-gram Model Model that uses contexts to predict a window of context words based on the current word
Synonym A word that has the same or nearly the same meaning as another word
Syntagm A group of words that concern substitution and signification of positional contrasts
System Query Representation of an information need interpretable by systems
TRECVID MED Multimedia Event Detection task within the TRECVID benchmark
User Query Representation of an information need of a user
User Feedback Relevance feedback provided by a user
Vocabulary A set of known words, related to ConceptBank
Vocabulary Mismatch Phenomenon that two people or systems name the same concept differently
Word2Vec Group of models that produce word embeddings
Word Embedding Vector representations of words, exploiting word distributions of the context of a word
Zero Shot TRECVID MED Task performed without having received training examples of that task
10Ex TRECVID MED Task with 10 training examples
100Ex TRECVID MED Task with 100 training examples

SUMMARY

In the modern world, networked sensor technology makes it possible to capture the world around us in real-time. In the security domain cameras are an important source of information. Cameras in public places, bodycams, drones and recordings with smart phones are used for real time monitoring of the environment to prevent crime (*monitoring case*); and/or for investigation and retrieval of crimes, for example in evidence forensics (*forensic case*). In both cases it is required to quickly obtain the right information, without having to manually search through the data. Currently, many algorithms are available to index a video with some pre-trained concepts, such as people, objects and actions. These algorithms require a representative and large enough set of examples (training data) to recognize the concept. This training data is, however, not always present.

In this thesis, we aim to assist an analyst in their work on video stream data by providing a search capability that handles ad-hoc textual queries, i.e. queries that include concepts or events that are not pre-trained. We use the security domain as inspiration for our work, but the analyst can be working in any application domain that uses video stream data, or even indexed data. Additionally, we do only consider the technical aspects of the search capability and not the legal, ethical or privacy issues related to video stream data. We focus on the retrieval of high-level events, such as birthday parties. We assume that these events can be composed of smaller pre-trained concepts, such as a group of people, a cake and decorations and relations between those concepts, to capture the essence of that unseen event (decompositionality assumption). Additionally, we hold the open world assumption, which means that the system does not have complete world knowledge. Although current state of the art systems are able to detect an increasingly large number of concepts, this number still falls far behind the near infinite number of possible (textual) queries that a system needs to be able to handle.

In our aim to assist the analyst, we focus on the improvement of the visual search effectiveness (e.g. performance) by a semantic query-to-concept mapping: the mapping from the user query to the set of pre-trained concepts. We use the TRECVID Multimedia Event Detection benchmark, as it contains high-level events inspired by the security domain. In this thesis, we show that the main improvements can be achieved by using a combination of i) query-to-concept mapping based on semantic word embeddings (+12%), ii) exploiting user feedback (+26%) and iii) fusion of different modalities (data sources) (+17%).

First, we propose an incremental word2vec (**i-w2v**) method, which uses word2vec trained on GoogleNews items as a semantic embedding model and incrementally adds concepts to the set of selected concepts for a query in order to deal with query drift. This method improves performance in terms of MAP compared to the state of the art word2vec method and knowledge based techniques. In combination with a

state of the art video event retrieval pipeline, we achieve top performance on the TRECVID MED benchmark regarding the zero-example task (MED14Test results). This improvement is, however, dependent on the availability of the concepts in the Concept Bank: without concepts related to or occurring in the event, we cannot detect the event. We, thus, need a properly composed Concept Bank to properly index videos.

Second, we propose an Adaptive Relevance Feedback interpretation method named **ARF** that not only achieves high retrieval performance, but is also theoretically founded through the Rocchio algorithm from the text retrieval field. This algorithm is adjusted to the event retrieval domain in a way that the weights for the concepts are changed based on the positive and negative annotations on videos. The ARF method has higher visual search effectiveness compared to k-NN based methods on video level annotations and methods based on concept level annotations.

Third, we propose blind late fusion methods that are based on state of the art methods, such as average fusion or fusion based on probabilities. Especially the combination of a Joint Ratio (ratio of probabilities) and Extreme Ratio (ratio of minimum and maximum) method (**JRER**) achieves high performance in cases with reliable detectors, i.e. enough training examples. This method is not only applicable to the video retrieval field, but also in sensor fusion in general.

Although future work can be done in the direction of implicit query-to-concept mapping through deep learning methods, smartly combining the concepts and the usage of spatial and temporal information, we have shown that our proposed methods can improve the visual search effectiveness by a semantic query-to-concept mapping which brings us a step closer to a search capability that handles ad-hoc textual queries for analysts.

SAMENVATTING

We krijgen steeds meer toegang tot informatie door de opkomst van nieuwe sensoren. Zo kunnen we onze kamertemperatuur op afstand instellen, onze fysieke activiteiten real time volgen met de Fitbit App en onze omgeving filmen met een camera, go pro, bodycam, drone of mobiele telefoon. In deze thesis richten we ons op de sensoren die video of beeldmateriaal produceren. Net als bij andere sensoren, is er steeds meer materiaal aanwezig. Voor een persoonlijke collectie is het waarschijnlijk nog mogelijk om handmatig de data te doorzoeken en in mapjes te stoppen, maar voor het beeldmateriaal dat binnenkomt bij beveiligingsbedrijven of de politie is dit niet meer haalbaar. Naast de camera's in publieke gebieden, zoals in winkelcentra en op straat, krijgen beveiligers mogelijk ook nog beelden van drones, body cams gedragen door de politie of mobiele opnames via burgers. Deze beelden kunnen gebruikt worden om een bepaald gebied te monitoren of als forensisch bewijs in een rechtszaak. In beide gevallen is het gewenst om zo snel mogelijk de juiste informatie te verkrijgen, zonder handmatig door alle beelden heen te moeten. Op dit moment zijn er al een aantal methoden om bepaalde verdachte gebeurtenissen, zoals zakkenrollen, het stelen van een voertuig of het opgraven van een verdacht voorwerp, te herkennen. Het ontwikkelen van detectiemethoden voor gebeurtenissen die volgens een vast patroon verlopen is haalbaar als er voldoende trainingsmateriaal is. Het is moeilijk om een set detectoren voor alle mogelijke complexe gebeurtenissen te maken. We voorzien dat er een zoekmogelijkheid nodig is om dit soort gebeurtenissen ook snel terug te kunnen vinden. We gebruiken het veiligheidsdomein als inspiratie voor dit onderzoek, maar we zullen alleen ingaan op de technische aspecten van een zoekmogelijkheid voor gebeurtenissen, en niet de juridische, ethische of privacy aspecten. De zoekmogelijkheid is ook niet alleen voor beveiligers relevant, maar voor allerlei soorten analisten die een dergelijke zoekmogelijkheid kunnen gebruiken in hun werk. Om zo'n zoekmogelijkheid te bewerkstelligen gebruiken we inspiratie uit het vakgebied genaamd *(Multimedia) Information Retrieval*, waaronder de huidige zoekmachines zoals Google en YouTube vallen. Deze zoekmachines 'indexeren' documenten en beelden met bepaalde woorden (*concepten*) om zo snel via die woorden resultaten te kunnen produceren. De uitdaging met beeldmateriaal is dat de concepten niet altijd juist herkend worden (bijvoorbeeld door een lage beeldkwaliteit), én dat er veel minder concepten zijn dan woorden in de taal.

In deze thesis onderzoeken we hoe we een gebruikersvraag beter kunnen omzetten naar een set van concepten die herkend kunnen worden. Daarbij maken we voornamelijk gebruik van gebruikersvragen die te maken hebben met gebeurtenissen. De gebeurtenissen, die afkomstig zijn van een internationale dataset genaamd TRECVID MED, zijn over het algemeen niet een concept zelf, maar kunnen in theorie wel beschreven worden als een combinatie van concepten. We zijn begonnen met het vergelijken van methoden die gebruikmaken van kennisbronnen, zoals Wikipe-

dia. Per kennisbron hebben we een methode voorgesteld, gebaseerd op de literatuur, om een set van concepten te selecteren. Zo gebruiken we TFIDF voor Wikipedia en de sterkte van de link tussen de woorden in de kennisgraaf voor ConceptNet. Uit het onderzoek bleek dat de kennisbronnen niet volledig zijn op het gebied van gebeurtenissen, en een combinatie van de resultaten het beste resultaat oplevert.

Naast het gebruik van verschillende kennisbronnen hebben we gekeken naar een ander type methode om automatisch een set van concepten te vinden. Dit is een recent geïntroduceerde methode genaamd word2vec, die gebruik maakt van (in dit geval) alle artikelen in GoogleNews om een semantische ruimte te creëren. In deze ruimte hebben de woorden die in dezelfde context gebruikt worden een kleine afstand. Het voordeel van de word2vec ruimte is dat je sommetjes kunt doen met taal, zoals 'koning – man + vrouw = koningin'. Een standaard manier om een set van concepten te vinden is om de afstand tussen de gebruikersvraag en elk van de concepten uit te rekenen en een top x aantal concepten te kiezen als de set van concepten. In deze thesis hebben we een methode ontwikkeld om, met behulp van de semantische ruimte waarin je sommetjes kunt doen, de set van concepten te vinden die het dichtstbij de gebruikersvraag ligt. We laten zien dat deze methode beter werkt dan de kennisbronnen of het handmatig selecteren van de set van concepten (+12%). Daarbij benadrukken we wel dat het belangrijk is om een zo goed mogelijke lijst van concepten te hebben om uit te kiezen (*vocabulary / Concept Bank*). Een voorbeeld is dat als je geen vuur kunt herkennen, het lastig is om de gebeurtenis 'het blussen van een vuurtje' te herkennen. Voor het herkennen van gebeurtenissen heb je niet alleen objecten, scenes (*low-level concepts*) en (inter)acties nodig (*mid-level concepts*), maar ook gebeurtenissen zelf of ingewikkeldere acties (*high-level concepts*). Een combinatie van de concepten 'paard' en 'rijden' geeft namelijk mogelijk niet hetzelfde resultaat als 'paardrijden'.

Naast automatische methoden om een set van concepten te vinden, kan een gebruiker ook helpen om een beter resultaat te krijgen. We vergelijken methoden waarbij de gebruiker de initiële set van concepten moet aanpassen en methoden waarbij de gebruiker feedback moet geven of een video wel of niet relevant is. In het aanpassen van de concepten vergelijken we algoritmen die de gewichten met een bepaalde waarde aanpast (*re-weighting*), de locatie van de gebruikersvraag in de semantische ruimte aanpast (*Query Point Modification*) en de locatie van de concepten in de semantische ruimte aanpast (*Detector Space*). In de feedback op de video's stellen we een nieuwe methode voor, genaamd *Adaptive Relevance Feedback*, die de gewichten van de concepten aanpast volgens het bekende Rocchio algoritme dat gebruikmaakt van de scores van de concepten op de positief en negatief geannoteerde video's. Die methode werkt beter dan de gebruikelijke 'k-nearest neighbor' methode die kijkt naar de afstand tussen huidige video en de dichtstbijzijnde positieve en negatieve video om de uiteindelijke relevantie van een video te bepalen. Uit het onderzoek blijkt dat de gebruikersinformatie de prestatie van het systeem altijd verbetert. Onze ARF methode verbetert het systeem meer dan de methoden die de concepten aanpassen of de k-NN methode, waarbij de verbetering 26% is ten opzichte van geen feedback en 20% ten opzichte van de andere methoden.

Met deze grote verbeteringen zullen analisten al beter gebeurtenissen kunnen te-

rugvinden dan voor het werk van deze thesis. Maar mogelijk zoeken gebruikers niet alleen naar grote gebeurtenissen, maar ook naar bijvoorbeeld een specifieke persoon met een bepaalde kleur jas of een ander signalement. Het kan zijn dat bepaalde type concepten bij gebeurtenissen geen invloed hebben op de prestatie, maar wel invloed hebben bij het zoeken naar een specifiek signalement, zoals dat je bij het zoeken naar een vrouw het concept 'man' niet acceptabel vindt. In de word2vec methode representeert de afstand tussen woorden 'een' relatie, maar het is niet expliciet welke relatie. In dit onderzoek willen we weten of dit schadelijk kan zijn voor de prestatie van het zoeksysteem als er niet gezocht wordt naar gebeurtenissen. In dit gedeelte van het onderzoek gebruiken we een gecreëerde dataset die bestaat uit speelgoed en kantoorartikelen. We gebruiken de kennisbron ConceptNet om bepaalde semantische structuren te representeren en onderzoeken of bepaalde type gebruikersvragen alleen bepaalde semantische structuren zouden moeten toestaan. Uit het onderzoek blijkt dat de set van concepten wel beter is als een bepaald type semantische structuur gebruikt wordt, maar voor het vinden van de juiste plaatjes dit niet altijd het geval is. Hierdoor kunnen we concluderen dat het beter is om alle mogelijke relaties te gebruiken, zoals die aanwezig zijn in de word2vec semantische ruimte.

Een laatste methode die de prestatie zou kunnen verbeteren is fusie. In veel vakgebieden blijkt dat het combineren van verschillende methoden de prestatie kan verbeteren. In deze thesis stellen wij een aantal relatief simpele fusiemethoden voor die onderlegd zijn vanuit de bestaande fusiemethoden die op beslisniveau werken (*late*) en geen trainingsvoorbeelden hebben om gewichten toe te kennen (*blind*), zoals het gemiddelde en product. In simulaties en experimenten met de TRECVID MED dataset laten we zien dat deze fusiemethoden tot wel 17% prestatieverbetering (MAP) kunnen opleveren.

In het laatste hoofdstuk kijken we naar andere methoden dan een expliciete set van concepten om een gebruikersvraag te beantwoorden. We gebruiken een bestaande methode die met neurale netwerken een vraag en plaatje kan analyseren en een antwoord kan genereren. Daaraan voegen we een objectdetectiemethode toe, evenals een postprocessing methode die ervoor zorgt dat je het juiste type antwoord op een vraag geeft, zoals een cijfer op een numerieke vraag ('hoeveel').

Een belangrijke vraag is nu natuurlijk in hoeverre deze nieuwe inzichten gebruikt zullen worden in de toekomst. Het succes van de toepassing van de nieuwe word2vec methode en de fusie methode is afhankelijk van veel factoren. Ten eerste behandelt deze thesis maar een klein deel van de zoekmogelijkheid die nodig is voor een analist. Zo moet er niet alleen een goede lijst van getrainde concepten zijn om mee te kunnen matchen, maar de gegevens moeten ook goed en veilig opgeslagen worden, een gebruikersinterface moet uitnodigen ermee te werken en, misschien wel het belangrijkste, de modus operandi van de analist zal moeten veranderen. Maar zelfs als analisten niet bereid zijn deze inzichten mee te nemen, zijn ze ook bruikbaar buiten dat domein. Ook al maakt deep learning nu zijn opkomst in multimedia retrieval, het gebruik van transparante methoden (*Explainable AI*) is nu ook een punt van focus voor de grote bedrijven in zoeksystemen, of zelfs binnen TNO. Daarnaast is de fusiemethode, eventueel uitgebreid met een wegingsfunctie voor de verschillende databronnen, zeker waardevol in veel domeinen.

DANKWOORD (ACKNOWLEDGEMENTS)

Dit boekje is het resultaat van een reis in de wereld van het onderzoek. Voor mij was het een geweldige reis, waarin ik mijn weg heb kunnen vinden naar een niche binnen het vakgebied waar mijn interesses liggen, namelijk de combinatie van AI met taalkunde / semantiek. Ik heb mijn nieuwsgierigheid naar hoe de wereld in elkaar zit, en hoe je computers dat kunt leren, kunnen uitbuiten om uiteindelijk te eindigen met dit boekje. Maar dat heb ik natuurlijk niet alleen gedaan, en daarom wil ik een aantal mensen bedanken, beginnend met jou, als lezer (van in ieder geval het dankwoord), voor het nemen van de moeite om het boekje open te slaan en er wat in te lezen.

Als eerste wil ik mijn promotor Wessel Kraaij bedanken voor de kans om te kunnen promoveren en mij kennis te laten maken met jouw vele connecties in het vakgebied. Ondanks je drukke agenda en het wisselen van universiteit, heb je toch geprobeerd mij op een zo goed mogelijke manier te begeleiden, en met succes.

Naast Wessel heb ik veel hulp gehad van mijn begeleider Klamer Schutte. Bedankt dat ik altijd mocht binnenlopen en je altijd tijd vrij kon maken voor overleg. Je hebt me niet alleen heel erg geholpen met het vinden van mijn pad in het onderzoek, maar ook het ontwikkelen van een kritische blik en het zuiver formuleren van mijn gedachten.

Mijn paranimfen en mede-PhD studenten van Wessel, Maya en Saskia, verdienen ook zeker een bedankje. Ik vond het heel fijn om de eerste twee jaar van mijn PhD tijd met jullie te kunnen brainstormen, advies in te winnen, en te kunnen praten over alle sores van het doen van een PhD. Maya, ik vergeet nooit de eerste dag dat we elkaar ontmoetten en erachter kwamen dat we wel heel veel gemeen hebben. Dat was een heel fijn begin van mijn PhD tijd, en ik ben heel blij je als vriendin en collega te hebben.

I would like to express my special thanks to Chong-Wah Ngo, who gave me the opportunity to join the VIREO team at the City University at Hong Kong during the summer of 2015. I really enjoyed working with you, not only during my time in Hong Kong but also in the time afterwards: thank you for the time you put into taking my papers to the next level. I will remember the 'celebration' drinks and nice dinners you throw. Thank you for flying all the way to the Netherlands to be a committee member at my defense. I would also like to thank James and Zhang Hao for the good collaboration on the TRECVID MED task and all your efforts to make me comfortable in Hong Kong, not only in translating the incomprehensible Mandarin sounds to English, but also in helping to find a room and getting nice food. JingJing, a special thanks to you too. Thank you for the nice conversations, your openness and your help in Hong Kong. I really enjoyed meeting you at the conferences afterwards as well. Ad-

ditionally, I would like to thank the other members of the VIREO group, such as Lei Pang, and the members from the swimming team: I really enjoyed meeting you.

Additionally, I would like to thank the members of my reading- and promotion commission for investing valuable time to read my thesis.

Mijn collega's op al mijn werklocaties wil ik ook graag bedanken. Ten eerste de collega's bij Intelligent Imaging voor het delen van hun kennis en kunde op het gebied van computer vision, maar ook de gezellige wandeling door de duinen tijdens de lunch. Ten tweede mijn collega's bij Data Science voor hun warme welkom. In het bijzonder Paul: bedankt voor de filosofische discussies en goede gesprekken. Ten derde wil ik graag mijn collega's uit Nijmegen (iCIS en IFL) bedanken voor de gezellige lunches, thee en wetenschappelijke discussies. In het bijzonder wil ik nog even noemen: Max, Elena, Gabriel, Jacopo, Ruifei en Simone.

Vanuit de onderzoeksschool SIKS wil ik ook graag twee mensen in het bijzonder bedanken: Myriam en Niels. Bedankt voor de goede gesprekken. Daarnaast wil ik graag nog mijn stagiairs / afstudeerders bedanken: Geert, Camille, Steven, Douwe, Laurens en Vangelis. Jullie hebben bijgedragen aan de resultaten van deze thesis.

Naast werk gerelateerde bedankjes, zou ik ook graag mijn vriend(inn)en willen bedanken. Ten eerste wil ik Marianne bedanken. Ik ken je al vanaf de middelbare school en ook al hebben we een hele andere studierichting gekozen, ik kan altijd bij je terecht. Bedankt voor het aanhoren van mijn verhalen, de leuke uitjes (Costa Rica in het bijzonder), je positiviteit, je eeuwige support en adviezen, en natuurlijk je hulp bij het bedenken van de omslag (ook dank aan Bram en Peter daarvoor). Ten tweede wil ik Myrthe bedanken. Onze paden liepen redelijk gelijk, met dezelfde bachelor en master, afstuderen bij TNO en een PhD (maar jij dan in Delft). Ik wil je bedanken voor de gezellige woensdagavonden en de gesprekken over het doen van een PhD. Ik hoop dat onze paden hierna parallel blijven lopen. Ten derde wil ik Roos bedanken. Bedankt dat je er nog altijd bent, om spelletjes te spelen, te sporten, of gewoon op vakantie te gaan naar bijvoorbeeld Ijsland. Ten vierde wil ik Bart (en Rob en Astrid) bedanken: jullie zijn een grote support geweest. Als laatste wil ik nog de zwemmers van de Duinkikkers bedanken, niet alleen voor de soms nodige afleiding, maar ook jullie interesse in mijn onderzoek en goede gesprekken.

Als laatste gaat mijn dank uit naar mijn familie, in het speciaal mijn ouders en broertje: bedankt voor jullie oneindige support en vertrouwen in mijn kunnen. Zonder jullie was dit boekje er misschien niet geweest.

Maaike de Boer
Bilthoven, Mei 2017

CURRICULUM VITÆ

Maaike de Boer was born November 7th, 1990 in Utrecht. She finished secondary school (Gymnasium) in 2008 at 'Het Nieuwe Lyceum' in Bilthoven, with a focus on Nature and Health, extended with Economics and Latin.

Maaike received her BSc (with a minor in Linguistics) in (Cognitive) Artificial Intelligence from the University of Utrecht in 2011. Her bachelor thesis was named 'Feature Analysis of Containers' and the research focused on how different kind of common containers, such as a jar or a bottle, are named in different languages. A feature graph that is used to visualize the language invariant space shows that a partitioning of the different categories is present. The result supports the theory of the universal conceptual space, which hypothesizes that the conceptual space is universal, i.e. the 'real' world is perceived similar by all people but the space partitioned differently in different languages.

Maaike continued her education with the master Artificial Intelligence at the same university and she obtained her MSc degree (cum laude) in 2013. Her master thesis, which was conducted in collaboration with TNO, concerned a decision support system for intelligence analysts. This support system uses trust models to help decide which agents are trusted enough to receive questions, provide information about the reliability of the sources and advise on decisions based on information from possibly unreliable sources.

Immediately after her graduation in 2013, Maaike continued as AiO (PhD student) at TNO, in collaboration with the Radboud University. She has conduced her research included in this thesis in the projects named Google for Sensors (Goose) and the ERP Making Sense of Big Data (MSoBD). In the summer of 2015, Maaike joined the VIREO team at the City University in Hong Kong to work on the TRECVID Multimedia Event Detection benchmark.

Currently, Maaike continues to work as a researcher at TNO.

LIST OF PUBLICATIONS

This dissertation is based on the following publications:

1. **Maaike de Boer**, Klamer Schutte and Wessel Kraaij (2016). "Knowledge based query expansion in complex multimedia event detection". *Multimedia Tools and Applications*, vol. 75, pp. 9025 - 9043.

2. **Maaike de Boer**, Yi-Jie Lu, Chong-Wah Ngo, Klamer Schutte, Wessel Kraaij, Zhang Hao (2017). "Semantic Reasoning in Zero Example Video Event Retrieval". *Transactions on Multimedia Computing, Communications, and Applications*, DOI:10.1145/3131288.

3. **Maaike de Boer**, Geert Pingen, Douwe Knook, Klamer Schutte and Wessel Kraaij (2017). "Improving Video Event Retrieval by User Feedback". *Multimedia Tools and Applications*, vol. 76, number 21, pp. 22361-22381.

4. **Maaike de Boer**, Paul Brandt, Maya Sappelli, Laura Daniele, Klamer Schutte and Wessel Kraaij (2015). "Query Interpretation - an Application of Semiotics in Image Retrieval". *International Journal On Advances in Software*, vol. 8, number 3 and 4, 2015. pp. 435 - 449.

5. **Maaike de Boer**, Klamer Schutte, Hao Zhang, Yi-Jie Lu, Chong-Wah Ngo and Wessel Kraaij (2016). "Blind late fusion in multimedia event retrieval". *International Journal of Multimedia Information Retrieval*, vol. 5, pp. 203 - 217.

6. **Maaike de Boer**, Steven Reitsma and Klamer Schutte (2016). "Counting in Visual Question Answering". *Proc. of 15th Dutch-Belgian Information Retrieval (DIR) Workshop*.

Other research that is used as inspiration for this dissertation is:

1. **Maaike de Boer**, Camille Escher and Klamer Schutte (2017). "Modelling Temporal Structures in Video Event Retrieval using an AND-OR graph". *Proc. MMEDIA 2017: The Ninth International Conference on Advances in Multimedia*, ISBN: 978-1-61208-548-7, pp. 85-88, IARIA.

2. Maya Sappelli, **Maaike de Boer**, Selmar Smit and Freek Bomhof (2017). "A Vision on Prescriptive Analytics". *Proc. ALLDATA 2017: The Third International Conference on Big Data, Small Data, Linked Data and Open Data*, ISBN:978-1-61208-552-4, pp. 45-50, IARIA.

3. Geert Pingen, **Maaike de Boer** and Robin Aly (2017). "Rocchio-Based Relevance Feedback in Video Event Retrieval". *International Conference on Multimedia Modelling (MMM)*, pp. 318-330, Springer.

4. Yi-Jie Lu, Hao Zhang, **Maaike de Boer** and Chong-Wah Ngo (2016). "Event Detection with Zero Example: Select the Right and Suppress the Wrong Concepts". *Proc. of 2016 ACM on International Conference on Multimedia Retrieval (ICMR)* , pp. 127-134, ACM.

5. Maya Sappelli, Gabriella Pasi, Suzan Verberne, **Maaike de Boer** and Wessel Kraaij (2016). "Assessing e-mail intent and tasks in e-mail messages". *Information Sciences* ,vol 358, pp 1-17.

6. Zhang Hao, Yi-Jie Lu, **Maaike de Boer**, Frank ter Haar, Zhaofan Qiu, Klamer Schutte, Wessel Kraaij and Chong-Wah Ngo (2015). "VIREO-TNO @ TRECVID 2015: Multimedia Event Detection". *Proc. TRECVID 2015*.

7. John Schavemaker, Martijn Spitters, Gijs Koot and **Maaike de Boer** (2015). "Fast re-ranking of visual search results by example selection". *Proc. 16th International Conference on Computer Analysis and Patterns (CAIP)*, pp. 387-398, Springer.

8. Klamer Schutte, Henri Bouma, John Schavemaker, Laura Daniele, Maya Sappelli, Gijs Koot, Pieter Eendebak, George Azzopardi, Martijn Spitters, **Maaike de Boer**, Maarten Kruithof and Paul Brandt (2015). "Interactive detection of incrementally learned concepts in images with ranking and semantic query interpretation". *Proc. Content-Based Multimedia Indexing Workshop (CBMI)*, pp. 1-4, IEEE.

9. **Maaike de Boer**, Laura Daniele, Paul Brandt and Maya Sappelli (2015). "Applying Semantic Reasoning in Image Retrieval". *Proc. ALLDATA 2015: The First International Conference on Big Data, Small Data, Linked Data and Open Data*, ISBN:978-1-61208-445-9, pp. 69-74, IARIA. *Best Paper Award*.

10. Chong-Wah Ngo, Yi-Jie Lu, Hao Zhang, Ting Yao, Chun-Chet Tan, Lei Pang, **Maaike de Boer**, John Schavemaker, Klamer Schutte and Wessel Kraaij (2014). "VIREO-TNO @ TRECVID 2014: Multimedia Event Detection and Recounting (MED and MER)". *Proc. TRECVID 2014*.

11. **Maaike de Boer**, Klamer Schutte and Wessel Kraaij (2013). "Event Classification using Concepts". *Proceedings ICT.OPEN 2013*, ISBN/EAN: 978-90-73461-84-0, pp. 38-42.

12. Henri Bouma, George Azzopardi, Martijn Spitters, Joost de Wit, Corné Versloot, Remco van der Zon, Pieter Eendebak, Jan Baan, Johan-Martijn ten Hove, Adam van Eekeren, Frank ter Haar, Richard den Hollander, Jasper van Huis, **Maaike de Boer**, Gert van Antwerpen, Jeroen Broekhuijsen, Laura Daniele, Paul Brandt, John Schavemaker, Wessel Kraaij and Klamer Schutte (2013). "TNO at TRECVID 2013: Multimedia event detection and instance search". *Proc. TRECVID 2013*.

SIKS DISSERTATIONS

2011

2011-01 Botond Cseke (RUN), *Variational Algorithms for Bayesian Inference in Latent Gaussian Models.*
2011-02 Nick Tinnemeier (UU), *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language.*
2011-03 Jan Martijn van der Werf (TUE), *Compositional Design and Verification of Component-Based Information Systems.*
2011-04 Hado van Hasselt (UU), *Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference.*
2011-05 Base van der Raadt (VU), *Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.*
2011-06 Yiwen Wang (TUE), *Semantically-Enhanced Recommendations in Cultural Heritage.*
2011-07 Yujia Cao (UT), *Multimodal Information Presentation for High Load Human Computer Interaction.*
2011-08 Nieske Vergunst (UU), *BDI-based Generation of Robust Task-Oriented Dialogues.*
2011-09 Tim de Jong (OU), *Contextualised Mobile Media for Learning.*
2011-10 Bart Bogaert (UvT), *Cloud Content Contention.*
2011-11 Dhaval Vyas (UT), *Designing for Awareness: An Experience-focused HCI Perspective.*
2011-12 Carmen Bratosin (TUE), *Grid Architecture for Distributed Process Mining.*
2011-13 Xiaoyu Mao (UvT), *Airport under Control. Multiagent Scheduling for Airport Ground Handling.*
2011-14 Milan Lovric (EUR), *Behavioral Finance and Agent-Based Artificial Markets.*
2011-15 Marijn Koolen (UvA), *The Meaning of Structure: the Value of Link Evidence for Information Retrieval.*
2011-16 Maarten Schadd (UM), *Selective Search in Games of Different Complexity.*
2011-17 Jiyin He (UVA), *Exploring Topic Structure: Coherence, Diversity and Relatedness.*
2011-18 Mark Ponsen (UM), *Strategic Decision-Making in complex games.*
2011-19 Ellen Rusman (OU), *The Mind ' s Eye on Personal Profiles.*
2011-20 Qing Gu (VU), *Guiding service-oriented software engineering - A view-based approach.*
2011-21 Linda Terlouw (TUD), *Modularization and Specification of Service-Oriented Systems.*
2011-22 Junte Zhang (UVA), *System Evaluation of Archival Description and Access.*
2011-23 Wouter Weerkamp (UVA), *Finding People and their Utterances in Social Media.*
2011-24 Herwin van Welbergen (UT), *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior.*
2011-25 Syed Waqar ul Qounain Jaffry (VU)), *Analysis and Validation of Models for Trust Dynamics.*
2011-26 Matthijs Aart Pontier (VU), *Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots.*
2011-27 Aniel Bhulai (VU), *Dynamic website optimization through autonomous management of design patterns.*
2011-28 Rianne Kaptein (UVA), *Effective Focused Retrieval by Exploiting Query Context and Document Structure.*
2011-29 Faisal Kamiran (TUE), *Discrimination-aware Classification.*
2011-30 Egon van den Broek (UT), *Affective Signal Processing (ASP): Unraveling the mystery of emotions.*
2011-31 Ludo Waltman (EUR), *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality.*
2011-32 Nees-Jan van Eck (EUR), *Methodological Advances in Bibliometric Mapping of Science.*
2011-33 Tom van der Weide (UU), *Arguing to Motivate Decisions.*
2011-34 Paolo Turrini (UU), *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations.*
2011-35 Maaike Harbers (UU), *Explaining Agent Behavior in Virtual Training.*
2011-36 Erik van der Spek (UU), *Experiments in serious game design: a cognitive approach.*

2011-37 Adriana Burlutiu (RUN), *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference.*
2011-38 Nyree Lemmens (UM), *Bee-inspired Distributed Optimization.*
2011-39 Joost Westra (UU), *Organizing Adaptation using Agents in Serious Games.*
2011-40 Viktor Clerc (VU), *Architectural Knowledge Management in Global Software Development.*
2011-41 Luan Ibraimi (UT), *Cryptographically Enforced Distributed Data Access Control.*
2011-42 Michal Sindlar (UU), *Explaining Behavior through Mental State Attribution.*
2011-43 Henk van der Schuur (UU), *Process Improvement through Software Operation Knowledge.*
2011-44 Boris Reuderink (UT), *Robust Brain-Computer Interfaces.*
2011-45 Herman Stehouwer (UvT), *Statistical Language Models for Alternative Sequence Selection.*
2011-46 Beibei Hu (TUD), *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work.*
2011-47 Azizi Bin Ab Aziz (VU), *Exploring Computational Models for Intelligent Support of Persons with Depression.*
2011-48 Mark Ter Maat (UT), *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent.*
2011-49 Andreea Niculescu (UT), *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality.*

2012
2012-01 Terry Kakeeto (UvT), *Relationship Marketing for SMEs in Uganda.*
2012-02 Muhammad Umair (VU), *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models.*
2012-03 Adam Vanya (VU), *Supporting Architecture Evolution by Mining Software Repositories.*
2012-04 Jurriaan Souer (UU), *Development of Content Management System-based Web Applications.*
2012-05 Marijn Plomp (UU), *Maturing Interorganisational Information Systems.*
2012-06 Wolfgang Reinhardt (OU), *Awareness Support for Knowledge Workers in Research Networks.*
2012-07 Rianne van Lambalgen (VU), *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions.*
2012-08 Gerben de Vries (UVA), *Kernel Methods for Vessel Trajectories.*
2012-09 Ricardo Neisse (UT), *Trust and Privacy Management Support for Context-Aware Service Platforms.*
2012-10 David Smits (TUE), *Towards a Generic Distributed Adaptive Hypermedia Environment.*
2012-11 J.C.B. Rantham Prabhakara (TUE), *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics.*
2012-12 Kees van der Sluijs (TUE), *Model Driven Design and Data Integration in Semantic Web Information Systems.*
2012-13 Suleman Shahid (UvT), *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions.*
2012-14 Evgeny Knutov (TUE), *Generic Adaptation Framework for Unifying Adaptive Web-based Systems.*
2012-15 Natalie van der Wal (VU), *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes..*
2012-16 Fiemke Both (VU), *Helping people by understanding them - Ambient Agents supporting task execution and depression treatment.*
2012-17 Amal Elgammal (UvT), *Towards a Comprehensive Framework for Business Process Compliance.*
2012-18 Eltjo Poort (VU), *Improving Solution Architecting Practices.*
2012-19 Helen Schonenberg (TUE), *What's Next? Operational Support for Business Process Execution.*
2012-20 Ali Bahramisharif (RUN), *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing.*
2012-21 Roberto Cornacchia (TUD), *Querying Sparse Matrices for Information Retrieval.*
2012-22 Thijs Vis (UvT), *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?.*
2012-23 Christian Muehl (UT), *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction.*
2012-24 Laurens van der Werff (UT), *Evaluation of Noisy Transcripts for Spoken Document Retrieval.*
2012-25 Silja Eckartz (UT), *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application.*
2012-26 Emile de Maat (UVA), *Making Sense of Legal Text.*

2012-27 Hayrettin Gurkok (UT), *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games.*
2012-28 Nancy Pascall (UvT), *Engendering Technology Empowering Women.*
2012-29 Almer Tigelaar (UT), *Peer-to-Peer Information Retrieval.*
2012-30 Alina Pommeranz (TUD), *Designing Human-Centered Systems for Reflective Decision Making.*
2012-31 Emily Bagarukayo (RUN), *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure.*
2012-32 Wietske Visser (TUD), *Qualitative multi-criteria preference representation and reasoning.*
2012-33 Rory Sie (OUN), *Coalitions in Cooperation Networks (COCOON).*
2012-34 Pavol Jancura (RUN), *Evolutionary analysis in PPI networks and applications.*
2012-35 Evert Haasdijk (VU), *Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics.*
2012-36 Denis Ssebugwawo (RUN), *Analysis and Evaluation of Collaborative Modeling Processes.*
2012-37 Agnes Nakakawa (RUN), *A Collaboration Process for Enterprise Architecture Creation.*
2012-38 Selmar Smit (VU), *Parameter Tuning and Scientific Testing in Evolutionary Algorithms.*
2012-39 Hassan Fatemi (UT), *Risk-aware design of value and coordination networks.*
2012-40 Agus Gunawan (UvT), *Information Access for SMEs in Indonesia.*
2012-41 Sebastian Kelle (OU), *Game Design Patterns for Learning.*
2012-42 Dominique Verpoorten (OU), *Reflection Amplifiers in self-regulated Learning.*
2012-43 Withdrawn, .
2012-44 Anna Tordai (VU), *On Combining Alignment Techniques.*
2012-45 Benedikt Kratz (UvT), *A Model and Language for Business-aware Transactions.*
2012-46 Simon Carter (UVA), *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation.*
2012-47 Manos Tsagkias (UVA), *Mining Social Media: Tracking Content and Predicting Behavior.*
2012-48 Jorn Bakker (TUE), *Handling Abrupt Changes in Evolving Time-series Data.*
2012-49 Michael Kaisers (UM), *Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions.*
2012-50 Steven van Kervel (TUD), *Ontologogy driven Enterprise Information Systems Engineering.*
2012-51 Jeroen de Jong (TUD), *Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching.*

2013

2013-01 Viorel Milea (EUR), *News Analytics for Financial Decision Support.*
2013-02 Erietta Liarou (CWI), *MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing.*
2013-03 Szymon Klarman (VU), *Reasoning with Contexts in Description Logics.*
2013-04 Chetan Yadati (TUD), *Coordinating autonomous planning and scheduling.*
2013-05 Dulce Pumareja (UT), *Groupware Requirements Evolutions Patterns.*
2013-06 Romulo Goncalves(CWI), *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience.*
2013-07 Giel van Lankveld (UvT), *Quantifying Individual Player Differences.*
2013-08 Robbert-Jan Merk (VU), *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators.*
2013-09 Fabio Gori (RUN), *Metagenomic Data Analysis: Computational Methods and Applications.*
2013-10 Jeewanie Jayasinghe Arachchige (UvT), *A Unified Modeling Framework for Service Design..*
2013-11 Evangelos Pournaras (TUD), *Multi-level Reconfigurable Self-organization in Overlay Services.*
2013-12 Marian Razavian (VU), *Knowledge-driven Migration to Services.*
2013-13 Mohammad Safiri (UT), *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly.*
2013-14 Jafar Tanha (UVA), *Ensemble Approaches to Semi-Supervised Learning Learning.*
2013-15 Daniel Hennes (UM), *Multiagent Learning - Dynamic Games and Applications.*
2013-16 Eric Kok (UU), *Exploring the practical benefits of argumentation in multi-agent deliberation.*
2013-17 Koen Kok (VU), *The PowerMatcher: Smart Coordination for the Smart Electricity Grid.*
2013-18 Jeroen Janssens (UvT), *Outlier Selection and One-Class Classification.*
2013-19 Renze Steenhuizen (TUD), *Coordinated Multi-Agent Planning and Scheduling.*

2013-20 Katja Hofmann (UvA), *Fast and Reliable Online Learning to Rank for Information Retrieval.*
2013-21 Sander Wubben (UvT), *Text-to-text generation by monolingual machine translation.*
2013-22 Tom Claassen (RUN), *Causal Discovery and Logic.*
2013-23 Patricio de Alencar Silva (UvT), *Value Activity Monitoring.*
2013-24 Haitham Bou Ammar (UM), *Automated Transfer in Reinforcement Learning.*
2013-25 Agnieszka Anna Latoszek-Berendsen (UM), *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System.*
2013-26 Alireza Zarghami (UT), *Architectural Support for Dynamic Homecare Service Provisioning.*
2013-27 Mohammad Huq (UT), *Inference-based Framework Managing Data Provenance.*
2013-28 Frans van der Sluis (UT), *When Complexity becomes Interesting: An Inquiry into the Information eXperience.*
2013-29 Iwan de Kok (UT), *Listening Heads.*
2013-30 Joyce Nakatumba (TUE), *Resource-Aware Business Process Management: Analysis and Support.*
2013-31 Dinh Khoa Nguyen (UvT), *Blueprint Model and Language for Engineering Cloud Applications.*
2013-32 Kamakshi Rajagopal (OUN), *Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development.*
2013-33 Qi Gao (TUD), *User Modeling and Personalization in the Microblogging Sphere.*
2013-34 Kien Tjin-Kam-Jet (UT), *Distributed Deep Web Search.*
2013-35 Abdallah El Ali (UvA), *Minimal Mobile Human Computer Interaction.*
2013-36 Than Lam Hoang (TUe), *Pattern Mining in Data Streams.*
2013-37 Dirk Börner (OUN), *Ambient Learning Displays.*
2013-38 Eelco den Heijer (VU), *Autonomous Evolutionary Art.*
2013-39 Joop de Jong (TUD), *A Method for Enterprise Ontology based Design of Enterprise Information Systems.*
2013-40 Pim Nijssen (UM), *Monte-Carlo Tree Search for Multi-Player Games.*
2013-41 Jochem Liem (UVA), *Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning.*
2013-42 Léon Planken (TUD), *Algorithms for Simple Temporal Reasoning.*
2013-43 Marc Bron (UVA), *Exploration and Contextualization through Interaction and Concepts.*

2014

2014-01 Nicola Barile (UU), *Studies in Learning Monotone Models from Data.*
2014-02 Fiona Tuliyano (RUN), *Combining System Dynamics with a Domain Modeling Method.*
2014-03 Sergio Raul Duarte Torres (UT), *Information Retrieval for Children: Search Behavior and Solutions.*
2014-04 Hanna Jochmann-Mannak (UT), *Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation.*
2014-05 Jurriaan van Reijsen (UU), *Knowledge Perspectives on Advancing Dynamic Capability.*
2014-06 Damian Tamburri (VU), *Supporting Networked Software Development.*
2014-07 Arya Adriansyah (TUE), *Aligning Observed and Modeled Behavior.*
2014-08 Samur Araujo (TUD), *Data Integration over Distributed and Heterogeneous Data Endpoints.*
2014-09 Philip Jackson (UvT), *Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language.*
2014-10 Ivan Salvador Razo Zapata (VU), *Service Value Networks.*
2014-11 Janneke van der Zwaan (TUD), *An Empathic Virtual Buddy for Social Support.*
2014-12 Willem van Willigen (VU), *Look Ma, No Hands: Aspects of Autonomous Vehicle Control.*
2014-13 Arlette van Wissen (VU), *Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains.*
2014-14 Yangyang Shi (TUD), *Language Models With Meta-information.*
2014-15 Natalya Mogles (VU), *Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare.*
2014-16 Krystyna Milian (VU), *Supporting trial recruitment and design by automatically interpreting eligibility criteria.*
2014-17 Kathrin Dentler (VU), *Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability.*
2014-18 Mattijs Ghijsen (VU), *Methods and Models for the Design and Study of Dynamic Agent*

Organizations.
2014-19 Vincius Ramos (TUE), *Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support.*
2014-20 Mena Habib (UT), *Named Entity Extraction and Disambiguation for Informal Text: The Missing Link.*
2014-21 Kassidy Clark (TUD), *Negotiation and Monitoring in Open Environments.*
2014-22 Marieke Peeters (UU), *Personalized Educational Games - Developing agent-supported scenario-based training.*
2014-23 Eleftherios Sidirourgos (UvA/CWI), *Space Efficient Indexes for the Big Data Era.*
2014-24 Davide Ceolin (VU), *Trusting Semi-structured Web Data.*
2014-25 Martijn Lappenschaar (RUN), *New network models for the analysis of disease interaction.*
2014-26 Tim Baarslag (TUD), *What to Bid and When to Stop.*
2014-27 Rui Jorge Almeida (EUR), *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty.*
2014-28 Anna Chmielowiec (VU), *Decentralized k-Clique Matching.*
2014-29 Jaap Kabbedijk (UU), *Variability in Multi-Tenant Enterprise Software.*
2014-30 Peter de Cock (UvT), *Anticipating Criminal Behaviour.*
2014-31 Leo van Moergestel (UU), *Agent Technology in Agile Multiparallel Manufacturing and Product Support.*
2014-32 Naser Ayat (UvA), *On Entity Resolution in Probabilistic Data.*
2014-33 Tesfa Tegegne (RUN), *Service Discovery in eHealth.*
2014-34 Christina Manteli (VU), *The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems..*
2014-35 Joost van Ooijen (UU), *Cognitive Agents in Virtual Worlds: A Middleware Design Approach.*
2014-36 Joos Buijs (TUE), *Flexible Evolutionary Algorithms for Mining Structured Process Models.*
2014-37 Maral Dadvar (UT), *Experts and Machines United Against Cyberbullying.*
2014-38 Danny Plass-Oude Bos (UT), *Making brain-computer interfaces better: improving usability through post-processing..*
2014-39 Jasmina Maric (UvT), *Web Communities, Immigration, and Social Capital.*
2014-40 Walter Omona (RUN), *A Framework for Knowledge Management Using ICT in Higher Education.*
2014-41 Frederic Hogenboom (EUR), *Automated Detection of Financial Events in News Text.*
2014-42 Carsten Eijckhof (CWI/TUD), *Contextual Multidimensional Relevance Models.*
2014-43 Kevin Vlaanderen (UU), *Supporting Process Improvement using Method Increments.*
2014-44 Paulien Meesters (UvT), *Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden..*
2014-45 Birgit Schmitz (OUN), *Mobile Games for Learning: A Pattern-Based Approach.*
2014-46 Ke Tao (TUD), *Social Web Data Analytics: Relevance, Redundancy, Diversity.*
2014-47 Shangsong Liang (UVA), *Fusion and Diversification in Information Retrieval.*

2015

2015-01 Niels Netten (UvA), *Machine Learning for Relevance of Information in Crisis Response.*
2015-02 Faiza Bukhsh (UvT), *Smart auditing: Innovative Compliance Checking in Customs Controls.*
2015-03 Twan van Laarhoven (RUN), *Machine learning for network data.*
2015-04 Howard Spoelstra (OUN), *Collaborations in Open Learning Environments.*
2015-05 Christoph Bösch (UT), *Cryptographically Enforced Search Pattern Hiding.*
2015-06 Farideh Heidari (TUD), *Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes.*
2015-07 Maria-Hendrike Peetz (UvA), *Time-Aware Online Reputation Analysis.*
2015-08 Jie Jiang (TUD), *Organizational Compliance: An agent-based model for designing and evaluating organizational interactions.*
2015-09 Randy Klaassen (UT), *HCI Perspectives on Behavior Change Support Systems.*
2015-10 Henry Hermans (OUN), *OpenU: design of an integrated system to support lifelong learning.*
2015-11 Yongming Luo (TUE), *Designing algorithms for big graph datasets: A study of computing bisimulation and joins.*
2015-12 Julie M. Birkholz (VU), *Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks.*

2015-13 Giuseppe Procaccianti (VU), *Energy-Efficient Software.*
2015-14 Bart van Straalen (UT), *A cognitive approach to modeling bad news conversations.*
2015-15 Klaas Andries de Graaf (VU), *Ontology-based Software Architecture Documentation.*
2015-16 Changyun Wei (UT), *Cognitive Coordination for Cooperative Multi-Robot Teamwork.*
2015-17 André van Cleeff (UT), *Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs.*
2015-18 Holger Pirk (CWI), *Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories.*
2015-19 Bernardo Tabuenca (OUN), *Ubiquitous Technology for Lifelong Learners.*
2015-20 Loïs Vanhée (UU), *Using Culture and Values to Support Flexible Coordination.*
2015-21 Sibren Fetter (OUN), *Using Peer-Support to Expand and Stabilize Online Learning.*
2015-23 Luit Gazendam (VU), *Cataloguer Support in Cultural Heritage.*
2015-24 Richard Berendsen (UVA), *Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation.*
2015-25 Steven Woudenberg (UU), *Bayesian Tools for Early Disease Detection.*
2015-26 Alexander Hogenboom (EUR), *Sentiment Analysis of Text Guided by Semantics and Structure.*
2015-27 Sándor Héman (CWI), *Updating compressed column-stores.*
2015-28 Janet Bagorogoza (TiU), *Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO.*
2015-29 Hendrik Baier (UM), *Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains.*
2015-30 Kiavash Bahreini (OUN), *Real-time Multimodal Emotion Recognition in E-Learning.*
2015-31 Yakup Koç (TUD), *On Robustness of Power Grids.*
2015-32 Jerome Gard (UL), *Corporate Venture Management in SMEs.*
2015-33 Frederik Schadd (UM), *Ontology Mapping with Auxiliary Resources.*
2015-34 Victor de Graaff (UT), *Geosocial Recommender Systems.*
2015-35 Junchao Xu (TUD), *Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction.*

2016

2016-01 Syed Saiden Abbas (RUN), *Recognition of Shapes by Humans and Machines.*
2016-02 Michiel Christiaan Meulendijk (UU), *Optimizing medication reviews through decision support: prescribing a better pill to swallow.*
2016-03 Maya Sappelli (RUN), *Knowledge Work in Context: User Centered Knowledge Worker Support.*
2016- 04 Laurens Rietveld (VU), *Publishing and Consuming Linked Data.*
2016- 05 Evgeny Sherkhonov (UVA), *Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers.*
2016-06 Michel Wilson (TUD), *Robust scheduling in an uncertain environment.*
2016-07 Jeroen de Man (VU), *Measuring and modeling negative emotions for virtual training.*
2016-08 Matje van de Camp (TiU), *A Link to the Past: Constructing Historical Social Networks from Unstructured Data.*
2016-09 Archana Nottamkandath (VU), *Trusting Crowdsourced Information on Cultural Artefacts.*
2016-10 George Karafotias (VUA), *Parameter Control for Evolutionary Algorithms.*
2016-11 Anne Schuth (UVA), *Search Engines that Learn from Their Users.*
2016-12 Max Knobbout (UU), *Logics for Modelling and Verifying Normative Multi-Agent Systems.*
2016-13 Nana Baah Gyan (VU), *The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach.*
2016-14 Ravi Khadka (UU), *Revisiting Legacy Software System Modernization.*
2016-15 Steffen Michels (RUN), *Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments.*
2016-16 Guangliang Li (UVA), *Socially Intelligent Autonomous Agents that Learn from Human Reward.*
2016-17 Berend Weel (VU), *Towards Embodied Evolution of Robot Organisms.*
2016-18 Albert Meroño Peñuela (VU), *Refining Statistical Data on the Web.*
2016-19 Julia Efremova (Tu/e), *Mining Social Structures from Genealogical Data.*
2016-20 Daan Odijk (UVA), *Context & Semantics in News & Web Search.*
2016-21 Alejandro Moreno Célleri (UT), *From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground.*

2016-22 Grace Lewis (VU), *Software Architecture Strategies for Cyber-Foraging Systems.*
2016-23 Fei Cai (UVA), *Query Auto Completion in Information Retrieval.*
2016-24 Brend Wanders (UT), *Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach.*
2016-25 Julia Kiseleva (TU/e), *Using Contextual Information to Understand Searching and Browsing Behavior.*
2016-26 Dilhan Thilakarathne (VU), *In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains.*
2016-27 Wen Li (TUD), *Understanding Geo-spatial Information on Social Media.*
2016-28 Mingxin Zhang (TUD), *Large-scale Agent-based Social Simulation - A study on epidemic prediction and control.*
2016-29 Nicolas Höning (TUD), *Peak reduction in decentralised electricity systems - Markets and prices for flexible planning.*
2016-30 Ruud Mattheij (UvT), *The Eyes Have It.*
2016-31 Mohammad Khelghati (UT), *Deep web content monitoring.*
2016-32 Eelco Vriezekolk (UT), *Assessing Telecommunication Service Availability Risks for Crisis Organisations.*
2016-33 Peter Bloem (UVA), *Single Sample Statistics, exercises in learning from just one example.*
2016-34 Dennis Schunselaar (TUE), *Configurable Process Trees: Elicitation, Analysis, and Enactment.*
2016-35 Zhaochun Ren (UVA), *Monitoring Social Media: Summarization, Classification and Recommendation.*
2016-36 Daphne Karreman (UT), *Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies.*
2016-37 Giovanni Sileno (UvA), *Aligning Law and Action - a conceptual and computational inquiry.*
2016-38 Andrea Minuto (UT), *Materials that Matter - Smart Materials meet Art & Interaction Design.*
2016-39 Merijn Bruijnes (UT), *Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect.*
2016-40 Christian Detweiler (TUD), *Accounting for Values in Design.*
2016-41 Thomas King (TUD), *Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance.*
2016-42 Spyros Martzoukos (UVA), *Combinatorial and Compositional Aspects of Bilingual Aligned Corpora.*
2016-43 Saskia Koldijk (RUN), *Context-Aware Support for Stress Self-Management: From Theory to Practice.*
2016-44 Thibault Sellam (UVA), *Automatic Assistants for Database Exploration.*
2016-45 Bram van de Laar (UT), *Experiencing Brain-Computer Interface Control.*
2016-46 Jorge Gallego Perez (UT), *Robots to Make you Happy.*
2016-47 Christina Weber (UL), *Real-time foresight - Preparedness for dynamic innovation networks.*
2016-48 Tanja Buttler (TUD), *Collecting Lessons Learned.*
2016-49 Gleb Polevoy (TUD), *Participation and Interaction in Projects. A Game-Theoretic Analysis.*
2016-50 Yan Wang (UVT), *The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains.*

2017

2017-01 Jan-Jaap Oerlemans (UL), *Investigating Cybercrime.*
2017-02 Sjoerd Timmer (UU), *Designing and Understanding Forensic Bayesian Networks using Argumentation.*
2017-03 Daniël Harold Telgen (UU), *Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines.*
2017-04 Mrunal Gawade (CWI), *Multi-core Parallelism in a Column-store.*
2017-05 Mahdieh Shadi (UVA), *Collaboration Behavior.*
2017-06 Damir Vandic (EUR), *Intelligent Information Systems for Web Product Search.*
2017-07 Roel Bertens (UU), *Insight in Information: from Abstract to Anomaly.*
2017-08 Rob Konijn (VU), *Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery.*

2017-09 Dong Nguyen (UT), *Text as Social and Cultural Data: A Computational Perspective on Variation in Text.*
2017-10 Robby van Delden (UT), *(Steering) Interactive Play Behavior.*
2017-11 Florian Kunneman (RUN), *Modelling patterns of time and emotion in Twitter #anticipointment.*
2017-12 Sander Leemans (TUE), *Robust Process Mining with Guarantees.*
2017-13 Gijs Huisman (UT), *Social Touch Technology - Extending the reach of social touch through haptic technology.*
2017-14 Shoshannah Tekofsky (UvT), *You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior.*
2017-15 Peter Berck (RUN), *Memory-Based Text Correction.*
2017-16 Aleksandr Chuklin (UVA), *Understanding and Modeling Users of Modern Search Engines.*
2017-17 Daniel Dimov (UL), *Crowdsourced Online Dispute Resolution.*
2017-18 Ridho Reinanda (UVA), *Entity Associations for Search.*
2017-19 Jeroen Vuurens (UT), *Proximity of Terms, Texts and Semantic Vectors in Information Retrieval.*
2017-20 Mohammadbashir Sedighi (TUD), *Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility.*
2017-21 Jeroen Linssen (UT), *Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds).*
2017-22 Sara Magliacane (VU), *Logics for causal inference under uncertainty.*
2017-23 David Graus (UVA), *Entities of Interest — Discovery in Digital Traces.*
2017-24 Chang Wang (TUD), *Use of Affordances for Efficient Robot Learning.*
2017-25 Veruska Zamborlini (VU), *Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search.*
2017-26 Merel Jung (UT), *Socially intelligent robots that understand and respond to human touch.*
2017-27 Michiel Joosse (UT), *Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors.*
2017-28 John Klein (VU), *Architecture Practices for Complex Contexts.*
2017-29 Adel Alhuraibi (UvT), *From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT".*
2017-30 Wilma Latuny (UvT), *The Power of Facial Expressions.*
2017-31 Ben Ruijl (UL), *Advances in computational methods for QFT calculations.*
2017-32 Thaer Samar (RUN), *Access to and Retrievability of Content in Web Archives.*
2017-33 Brigit van Loggem (OU), *Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity.*
2017-34 Maren Scheffel (OU), *The Evaluation Framework for Learning Analytics .*
2017-35 Martine de Vos (VU), *Interpreting natural science spreadsheets .*
2017-36 Yuanhao Guo (UL), *Shape Analysis for Phenotype Characterisation from High-throughput Imaging .*
2017-37 Alejandro Montes Garcia (TUE), *WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy .*
2017-38 Alex Kayal (TUD), *Normative Social Applications .*
2017-39 Sara Ahmadi (RUN), *Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR .*
2017-40 Altaf Hussain Abro (VUA), *Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems .*
2017-41 Adnan Manzoor (VUA), *Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle.*
2017-42 Elena Sokolova (RUN), *Causal discovery from mixed and missing data with applications on ADHD datasets.*
2017-43 Maaike de Boer (RUN), *Semantic Mapping in Video Retrieval.*