Event Classification using Concepts

Maaike de Boer*†, Klamer Schutte* and Wessel Kraaij*†
*TNO Technical Sciences
Email: maaike.deboer@tno.nl, klamer.schutte@tno.nl, wessel.kraaij@tno.nl

†Institute for Computing and Information Science
Email: m.deboer@cs.ru.nl, w.kraaij@cs.ru.nl

***Abstract*—The semantic gap is one of the challenges in the GOOSE project. In this paper a Semantic Event Classification (SEC) system is proposed as an initial step in tackling the semantic gap challenge in the GOOSE project. This system uses semantic text analysis, multiple feature detectors using the BoW model, SVM-based concept classifiers, event classifiers and fusion to classify if an event is present in a certain video.**

The TRECVID Multimedia Event Detection task 2013 is used to evaluate the SEC system. The results show that an initial step in bridging the semantic gap and tackling the challenges in the GOOSE project is made, but that there is room for improvement. We expect that future research in learning and defining high-level concepts and event classification will further bridge the semantic gap.

I. Introduction

The GOOSE project [1] aims to create the capability to search semantically for any relevant information within all sensor streams in the entire Internet of sensors in real time. In order to create this capability relevant sensors have to be selected, the sensor database has to be maintained, computational power has to be scaled to accommodate any number of user queries at any time, a semantic user interface must be present and information and content has to be extracted from sensor data streams [1]. The two main challenges for GOOSE are scalability and the semantic gap. Scalability means dealing with a very large number of sensors, users, queries and scenarios. The semantic gap occurs in two ways in the GOOSE project. Firstly, there must be an interpretation of the user query in order to search for the right information: *the user-sensor gap*. Secondly, the information from sensor data has to be transformed into an answer for the user: *the data-answer gap*. In this paper, we propose a Semantic Event Classification system to bridge both ways of the semantic gap. This system is derived from the TNO TRECVID MED 2013 system [2]. Differences between that system and our system is the use of lemmas in the Semantic Text Analysis, a more extensive explanation of event classifiers, the evaluation method and results.

In the following section theoretical background about the semantic gap in the domain of image retrieval is given. In the third section our proposed system is explained. The fourth section shows how we evaluated our system and the fifth section displays the results. The sixth section contains the discussion, conclusion and future research.

II. Theoretical Background

The semantic gap is a hot research topic. In the image retrieval domain the semantic gap can be 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' [3]. Another formulation is that there is no direct link between the high-level concepts and the low-level features [4]. Low-level features can be extracted from three modalities: visual, audio and text [5]. Visual features can be split up into color features, texture features, shape features, appearance features and motion features [6][7]. An audio feature is ASR and a text feature is OCR. High-level concepts are words or combination of words used in user queries.

Hare et al. [8] split the semantic gap in two sections: *the gap between descriptors and object labels* and *the gap between labeled objects and full semantics*. Both sections address the data-answer gap. Descriptors are feature vectors of an image and labeled objects are the symbolic names of combinations of descriptors. In order to bridge the gap between descriptors and object labels the system should learn which combination of descriptors represent objects and what the labels of these objects should be. This is called automatic annotation of image content [8] or image annotation [9]. Most existing approaches can be divided into the categories *classification based / discriminative* and *probabilistic modeling based / generative* [9] [10] [11]. Some use both categories [12] [13]. In the classification based methods keywords are treated as classes and classifiers are used to annotate an input image by taking the class with the highest similarity measure. Both supervised methods such as support vector machines (SVMs), Bayesian classifiers and Decision Trees and unsupervised methods such as k-means, NCut and LPC can be used [4]. Important for this category is the image feature representation. The most popular feature representation method is the bag-of-words (BoW) or bag-of-visual-words model [14]. An advantage of this approach is that we can use many machine learning techniques for learning and it is effective and more accurate than probabilistic modeling based methods. A disadvantage of this method is that it is unscalable for a huge amount of images with infinite semantics and it cannot handle missing data [9].

The probabilistic modeling based methods attempt to infer correlations or joint probabilities between images and annotations with for example a Gaussian Mixture Model, Latent Dirichlet Allocation Model, correspondence LDA or a hybrid probabilistic model [9]. An advantage of this approach is that it is flexible, because is it not unscalable for a huge amount of images and it can handle missing data. This approach also have a better explanatory power than the classification approach. A disadvantage of this approach is that performance is less effective and less accurate than the classification approach [11] [12].

In order to reduce the gap between labeled objects and full semantics object ontologies can be used to define high-level concepts [4] [8]. In object ontologies a qualitative definition of the high-level semantics is given in terms of for example color, position, size and shape.

Bridging the user-sensor gap is dependent on the type of user, search data and querying modality [15]. All content-based image retrieval systems use content-based query processing with feature extraction and feature representation as described above, assuming an image or graphics as input. Our approach is to allow text input for which we use natural language processing as a first step in bridging the user-sensor gap [15]. This gap can also be bridged by introducing relevance feedback to learn the users' intention. Bridging the user-sensor gap is sometimes seen as part of image annotation [9]. The user can give feedback about initial retrieval results after which the system can adjust weights, do query-point-movement or use machine learning techniques to create new results [4]. Several relevance models have been proposed such as Cross-Media Relevance Model, Continuous Relevance Model, Multiple Bernoulli Relevance model [8] [9].

In the next section, we propose a Semantic Event Classification system to bridge the user-sensor gap and the data-answer gap with use of a classification based image annotation method, ontologies and a text-based processing method using natural language processing.

III. Semantic Event Classification System

We propose a Semantic Event Classification system to classify events based on a textual description, because we want our users to be able to type text input as a description for their information need. The information need is in our case an event, which is defined as 'an observable occurrence that interests users' [16]. In Fig. 1 our proposed Semantic Event Classification (SEC) system is shown. It needs to bridge the user-sensor gap. This is especially necessary when no example video for a certain information need is available. For bridging the user-sensor gap we use a semantic text analysis. Furthermore, we need to bridge the data-answer gap. With concept classifiers and event classifiers we try to bridge this gap. Fusion is used to fuse the outputs from the event classifiers. The different parts of the SEC system are further explained in the next subsections.

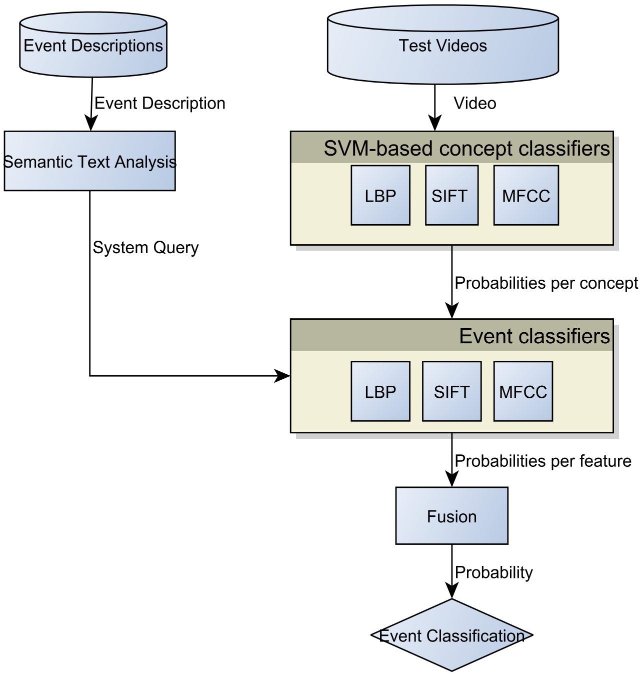

Fig. 1. Proposed Semantic Event Classification system

Event Name		Winning a race without a vehicle
Evidential Description	scene	outdoors (park, field, track, road, or stadium) or indoors (indoor track, pool, or large gymnasium)
	objects /people	runner, number worn on runner's back/front/ arm, potato sack, marker for finish line (tape stretched across road, potato sacks lying on ground), running shoes, baton, spectators, boundary markers/signs, signs supporting /encouraging a particular runner, water bottles, first aid tent
	activities	running, swimming, hopping, climbing, jumping, breaking through tape, passing a baton, spectators running a short distance with the runner, passing out water bottles to the runners
	audio	onlookers cheering, verbal or other indication of starting the race (yelling "Go!", gun shooting), narration of the race (speaking through a microphone)

TABLE I
Example of event description; from TRECVID MED 2013

A. Semantic Text Analysis

In the process *Semantic Text Analysis* we interpret the users needs. The type of user query used as input for this process is an event description, which is chosen based on our evaluation set (see section IV). This event description consists of the name of the event and a short description of the event. In this description information about the scene, object and/or people, activities and audio can be provided. An example is given in Table I. In a syntactic analysis all elements in noun phrases and verb phrases are extracted from the textual description using the Stanford Parser [17]. In this step we also retrieve whether an element is used *positive or negative*. For example 'without' and 'not' indicate that the element should have a negative relation and should therefore not be present. In the semantic analysis all elements and combination of elements, which are called textual concepts, in a specific phrase are lemmatized using the lemmatizer in the CoreNLP library of the Stanford Parser and interpreted using WordNet [18] and an OWL ontology [19]. For the interpretation we match textual concepts with a set of known concepts. The OWL ontology consists of a set of known concepts that are particularly relevant for the task (see IV.A and [2]). WordNet and the OWL ontology are also used to retrieve more known concepts by selecting *hyponyms and hypernyms* of interpreted concepts. In this step the *semantic distance* of the known concepts is set as the Lin-measure [20]. From the known concepts a system query is generated by combining the known concepts using logical operators. The AND-operator is used to combine all known concepts from the description and the OR-operator is used to connect the concepts with their known hypernyms and hyponyms. The set of concepts combined with the OR-operator is called an OR-group. Each of the concepts has information whether if it is positive or negative and its semantic distance. An example is given below.

AND (
racing(1)
OR (NOT (vehicle (1), truck (1), tractor (1), car (1), bus (1), ambulance (1), policecar (1), taxi (1), boat (1), cruiseship (1), ship (1), sailingboat (1), rowingboat (1), motorboat (1), train (1), bicycle/bike (1), motorcycle (1), airplane (1), helicopter (1)))
park (1)
field (1)
track (1)
road (1)
stadium (1)
swimmingpool (1)
runner (1)
potato (1)
finishline (1)
tape (1)
shoes (1)
spectator (1)
OR (water (1), food (0.69))
bottle (1)
sign(1)
OR (tent (1), circustent (1))
run (1)
swim (1)
cheering (1)
yelling(1)
go (1)
gun (1)
shooting (1)
person (0.3)
microphone (1))
EXAMPLE: System Query

B. SVM-based Concept Classifiers

For detection of high-level concepts in a video concept classifiers are needed. Based on event descriptions of last years of TRECVID MED, analysis of the ImageNet structure and analysis of missing concepts in each of the categories of the evidential description 546 concepts were defined. For each concept videos and images are downloaded from ImageNet [21], Google and Youtube (see Fig. 2). Both specific and general concepts from different types of concepts are used as recommended by Habibian et al. [22]. For the low-level features LBP, SIFT and MFCC using the BoW model and for each concept SVM-based classifiers are trained with a histogram-intersection kernel using the downloaded videos and images. Due to sparse training data for some objects 442 concept classifiers are used for LBP, 418 for SIFT and 86 for MFCC. MFCC needs audio information, so therefore less training data could be used. For more information about the low-level feature implementation and the feature representation see [2].

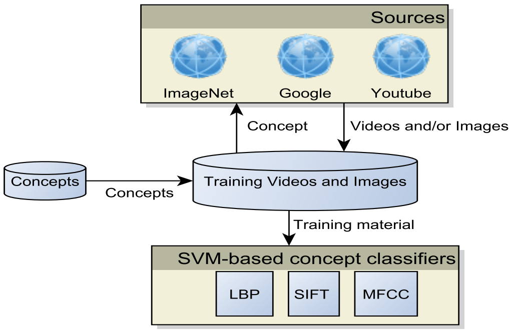

Fig. 2. Training of Concept Classifiers

C. Event Classifiers

With the system query containing the concepts and relations from an event description and output of each of the trained SVM-based concept classifiers for the complete set of videos the event classifiers can be applied. For each of the three concept classifiers a different event classifier is used. Firstly, the computed probabilities of each of the identified concepts are normalized over all videos with zero mean and unity standard deviation (NCV). Secondly, these normalized probabilities per concept are multiplied by a weighting factor (W) in order to create a concept score (S).

$$S(c,v) = \begin{cases} NCV(c,v)\cdot \text{ W(c)} & \text{if } W(c) \text{ is} \geq 0 \\ 0 & \text{other} \end{cases} \tag{1}$$

where $S(c)$ is the score of concept c for video v, $NCV(c,v)$ is the normalized probability per concept c for video v and $W(c)$ is the weight of concept c.

$$W(c) = POS(c) \cdot SD(c) \cdot DV(c) \tag{2}$$

where $W(c)$ is the weight of concept c, $POS(c)$ is positive (1) or negative (-1) concept, $SD(c)$ is the semantic distance of concept c (see III.A) and DV(c) is the detectability value of concept c.

The detectability value of a concept is an estimate of how predictive a concept is for a certain event and it is implemented as the average probability estimate of the concept classifier, measured over all events in which the concept is identified by the semantic text analysis. In our experiment the estimation of the detectability value of each concept is based on the training data from TRECVID MED 2013 [23] which consists of thirty event kits.

$$DV(c) = \frac{1}{|C|} \cdot \sum_{e=1}^{|C|} \frac{1}{V} \sum_{v=1}^{V} NCV(c, v) \tag{3}$$

where $DV(c)$ is the detectability value of concept c, $|C|$ is the amount of events in which the concept is expected (on training set), V is the total amount of videos and NCV(c, v) is normalized probability per concept c for video v.

The total score per video for a certain event is the sum of the highest concept score of each OR-group (see III.A). This total score is adjusted to a value between zero and one with the inverse function of the tangent.

D. Fusion

With total scores for all videos for each of the three event classifiers a late fusion has to be done in order to get one score for the event classification. First a *double sigmoid function* is applied to normalize the scores for the computed threshold around 0.5 [7]. We tested both the *accuracy weighted average* and the *threshold-distance weighted average* as fusion methods [7]. For the accuracy weighted average the accuracy on the training set is used as weight and for the threshold-distance weighted average the distance from threshold is used as weight [7]. Both accuracy and threshold are estimated with the same set as used to estimate the detectability value (see III.C).

IV. Evaluation

The SEC system is evaluated using the training data from the TRECVID Multimedia Event Detection task 2013. In this 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. We used the thirty training event kits, which contains ten research event kits and twenty evaluation events with 100 positive videos and 50 negative videos per event and a textual description per event. The videos from these sets have ground truth information about the event they contain or do not contain.

From the textual description only the provided event name and evidential description as shown in Table 1 are used for our semantic text analysis. The detectability value of the event classifiers is determined with the same set of events and videos as our test set.

We measure performance with the Mean Average Precision value as used by TRECVID and many other benchmarks.

$$MAP = \frac{1}{Ne} \cdot \sum_{j=1}^{Ne} \frac{1}{Vrel} \sum_{v=1}^{Vrel} pr(v, e) \tag{4}$$

where Ne is the number of events, $Vrel$ is the number of relevant videos and pr(v, e) is the precision

$$pr(v, e) = \begin{cases} r_v / rr_v & \text{if } v \text{ is retrieved and } n_i \leq Th \\ 0 & \text{other} \end{cases} \tag{5}$$

where r_v is the number of relevant videos for event e found at ranks 1 - v, rr_v is the rank of relevant and retrieved video v for event e

Because we are only using precision, the best method is to assign every video as positive. The threshold Th is thus set to 0.

V. Results

We calculated the MAP value for each of the different event classifiers (LBP, SIFT and MFCC) and the two types of fusion (threshold-weighted and accuracy-weighted). We also implemented a random system for comparison with our system. This random system gives a random number between 0 and 1 for each video and each event. We ran the system 100 times in order to calculate average performance. We calculated average performance twice and the result was a difference of 0,01%. The results are presented in Table II. Best performance on accuracy-weighted fusion was 26,45% on the event 'Winning a race without a vehicle', which was the only event with negative concepts. In this event only 10 concepts had a positive detectability value, but 'racing' and 'stadium' are high indicators for the event. The worst performance on accuracy-weighted fusion was 3,37% on the event 'Working on a metal crafts project'. In this last event 16 concepts with a positive detectability value were detected, but none had a high detectability value. This means that none of the known concepts gives a good indication for this event.

Condition	**MAP (in %)**
RANDOM	2,56
LBP	5,00
SIFT	8,06
MFCC	3,09
FUSED THR	8,07
FUSED ACC	8,19

TABLE II
Results

VI. Discussion, Conclusion and Future Research

The results show that an initial step in bridging the semantic gap and tackling the challenges in the GOOSE project is made, but that there is room for improvement. Performance on all tested conditions is better than random, but a comparison to other systems is hard to make because the official evaluation of the TRECVID MED task 2013 is done on other test sets. Furthermore, the concept classifiers and detectability value are optimized for this test set, so the results show major overfitting. In the future it would be interesting to calculate performance on both other TRECVID MED sets or other test to make a better evaluation of our system. In making further progress on bridging the semantic gap between the user query and sensor data research on the semantic representation of the concepts (AND of ORs) can be done. Based on the results quality of concepts seem more determining performance than quantity. This implies that semantic text analysis must select the concepts that are representative for a certain event. These concepts should also be known by the system and concept classifiers have to be trained on these concepts. A potential improvement may thus be to select and train (more) concepts representative for certain events.

The use of relevance feedback may also be an improvement. The bridge between sensor data and answer can be tackled by research on learning methods for both concept classifiers and event classifiers. Concept classifiers are now trained with a histogram-intersection kernel, but other methods such as taking the maximum may work better. The method of training and applying the event classifiers may also improve performance.

Acknowledgment

We would like to thank the members of the GOOSE project for providing the TNO TRECVID MED system and the technology program Adaptive Multi Sensor Networks (AMSN) and the MIST research program of the Dutch Ministry of Defense for their financial support.

References

[1] K. Schutte, F. Bomhof, G. Burghouts, J. van Diggelen, P. Hiemstra, J. van't Hof, W. Kraaij, H. Pasman, A. Smith, C. Versloot *et al.*, "GOOSE: Semantic search on internet connected sensors," in *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2013, pp. 875 806–875 806.

[2] H. Bouma, 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, "TNO at TRECVID 2013: Multimedia event detection and instance search," in *Proceedings of TRECVID 2013*, 2013.

[3] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, 2000.

[4] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

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

[6] P. Natarajan, P. Natarajan, V. Manohar, S. Wu, S. Tsakalidis, S. N. Vitaladevuni, X. Zhuang, R. Prasad, G. Ye, D. Liu *et al.*, "Bbn viser trecvid 2011 multimedia event detection system," in *NIST TRECVID Workshop*, 2011.

[7] P. Natarajan, S. Wu, S. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, and R. Prasad, "Multimodal feature fusion for robust event detection in web videos," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1298–1305.

[8] J. S. Hare, P. H. Lewis, P. G. Enser, and C. J. Sandom, "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, 2006, pp. 607 309–607 309.

[9] J. Yang and S.-j. Zhu, "Narrowing semantic gap in content-based image retrieval," in *Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM), 2012 International Conference on*. IEEE, 2012, pp. 433–438.

[10] N. Zhou, W. K. Cheung, G. Qiu, and X. Xue, "A hybrid probabilistic model for unified collaborative and content-based image tagging," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 7, pp. 1281–1294, 2011.

[11] F. Wang, "A survey on automatic image annotation and trends of the new age," *Procedia Engineering*, vol. 23, pp. 434–438, 2011.

[12] D. Ziou, T. Hamri, and S. Boutemedjet, "A hybrid probabilistic framework for content-based image retrieval with feature weighting," *Pattern Recognition*, vol. 42, no. 7, pp. 1511–1519, 2009.

[13] R. Raina, Y. Shen, A. Mccallum, and A. Y. Ng, "Classification with hybrid generative/discriminative models," in *Advances in neural information processing systems*, 2003, p. None.

[14] C.-F. Tsai, "Bag-of-words representation in image annotation: A review," *ISRN Artificial Intelligence*, vol. 2012, 2012.

[15] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

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

[17] M.-C. De Marneffe, B. MacCartney, C. D. Manning *et al.*, "Generating typed dependency parses from phrase structure parses," in *Proceedings of LREC*, vol. 6, 2006, pp. 449–454.

[18] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[19] G. Antoniou and F. Van Harmelen, "Web ontology language: Owl," in *Handbook on ontologies*. Springer, 2004, pp. 67–92.

[20] D. Lin, "An information-theoretic definition of similarity." in *ICML*, vol. 98, 1998, pp. 296–304.

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

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

[23] [Online]. Available: http://www.nist.gov/itl/iad/mig/med13.cfm