Recognition and localization of relevant human behavior in videos

Henri Bouma[*], Gertjan Burghouts, Leo de Penning, Patrick Hanckmann, Johan-Martijn ten Hove, Sanne Korzec, Maarten Kruithof, Sander Landsmeer, Coen van Leeuwen, Sebastiaan van den Broek, Arvid Halma, Richard den Hollander, Klamer Schutte

TNO, PO Box 96864, 2509 JG The Hague, The Netherlands.

ABSTRACT

Ground surveillance is normally performed by human assets, since it requires visual intelligence. However, especially for military operations, this can be dangerous and is very resource intensive. Therefore, unmanned autonomous visual-intelligence systems are desired. In this paper, we present an improved system that can recognize actions of a human and interactions between multiple humans. Central to the new system is our agent-based architecture. The system is trained on thousands of videos and evaluated on realistic persistent surveillance data in the DARPA Mind's Eye program, with hours of videos of challenging scenes. The results show that our system is able to track the people, detect and localize events, and discriminate between different behaviors, and it performs 3.4 times better than our previous system.

Keywords: Visual intelligence, action recognition, artificial intelligence, image retrieval, computer vision.

1. INTRODUCTION

Ground surveillance is normally performed by human assets. Military forces would like to shift this mission to unmanned autonomous systems, to reduce risks and reduce man power usage. However, unmanned systems lack a capability that currently exists only in humans: visual intelligence. The Defense Advanced Research Projects Agency (DARPA) is addressing this challenge with Mind's Eye, a program aiming to develop a visual-intelligence capability for unmanned systems, in which we developed the CORTEX system.

In this paper, we present our Year-2 CORTEX system, which can recognize, localize and describe the most relevant actions of a human and interactions between multiple humans in video. The central element in our approach is an agent-based architecture, which allows activity analysis of multiple persons (agents) in a single scene. Furthermore, we use a filtering mechanism for selecting the most relevant relations and computing multi-entity properties for them in order to describe interactions effectively.

The system is trained on thousands of videos and evaluated on realistic persistent surveillance data in the DARPA Mind's Eye program, which includes hours of videos showing significant scene complexity with realistic backgrounds, different camera positions, occlusion and multiple people with a variety of clothing performing actions simultaneously. In the paper, we show that our system is able to track the people, detect and localize events, and discriminate between different behaviors, such as approach, carry and exchange.

The main contribution of this paper is that we combined several improvements and applied our advanced pipeline [14] to the Year 1 dataset of DARPA and showed that the current implementation performs 3.4 times better than the system of last year [6].

The outline of the paper is as follows. A literature overview is presented in Section 2, the CORTEX system is described in Section 3, experiments and results are shown in Section 4 and conclusions in Section 5.

[*] henri.bouma@tno.nl; phone +31 888 66 4054; http://www.tno.nl

H. Bouma, G. Burghouts, L. de Penning, P. Hanckmann, J.-M. ten Hove, S. Korzec, M. Kruithof, S. Landsmeer, C. van Leeuwen, S. van den Broek, A. Halma, R. den Hollander, K. Schutte, "Recognition and localization of relevant human behavior in videos", Proc. SPIE, vol. 8711, (2013).
http://dx.doi.org/10.1117/12.2015877

2. LITERATURE SURVEY

An extensive overview of datasets related to action recognition was presented by Liu [30] and Chaquet [17]. A review of current technologies for complex event recognition in unconstrained videos was made by Jiang [24]. Our action-recognition literature overview focusses on results obtained on the DARPA Mind’s Eye data (year 1 and year 2) and papers of fellow participants in this program. Comparisons between performance on the DARPA data and other datasets, such as IXMAS and UT-interaction, was provided already by Burghouts [15]. Recently, new results have become available on IXMAS [22][25][44][44] and UT-Interaction [49]. Our literature overview shows a wide variety of ideas for automatic action recognition – including unsupervised learning, simultaneous analysis, pose estimation and high-level representations – which are discussed in more detail below.

O’Hara e.a. [34] presented a method for unsupervised learning and recognition of human actions in video. In their experiments, the product manifolds perform better than bag-of-features for clustering video clips of actions. In another publication [33], the subspace forest was presented, designed to provide an efficient approximate nearest neighbor query of subspaces represented as points on Grassmann manifolds, and applied to action recognition. Truyen e.a. [43] also tried to avoid completely supervised training. They proposed an approach based on semi-supervised training of partially hidden discriminative models such as the conditional random field (CRF) and the maximum entropy Markov model (MEMM).

Barbu e.a. [5] proposed a method for simultaneous object detection, tracking, and event recognition. Many person and object detectors, e.g. those of Felzenszwalb, use internally a scale-space pyramid to represent all possible detections at all locations and scales in an associated frame. Instead of extracting and tracking the thresholded detections, Barbu e.a. consider tracks between all detections in the entire pyramid simultaneously by defining a distance measure between detection pyramids for adjacent frames and performing the Viterbi tracking algorithm on these pyramids. The simultaneous detection and tracking solves the missing detections in single frames, which could also be solved by lowering detection thresholds or by projecting detections forward to augment the raw detector in subsequent frames. Maximum likelihood (ML), the standard approach to using HMMs for classification, selects the event model that maximizes the likelihood of an observed event. Instead of the maximum likelihood (the standard approach to using HMMs for classification), the model with the maximum a posteriori (log) probability (MAP) can be selected. This can be computed with the Viterbi algorithm. This allows a combination of the Viterbi algorithm used for detection-based tracking with the Viterbi algorithm used for event classification. Khan e.a. [27] also proposed to analyze elements simultaneously. They proposed a probabilistic approach to simultaneously infer what action was performed, what object was used and what poses the actor went through. Natarajan [32] presented a family of graphical models that generalize the extensions of HMMs and CRFs and simultaneously model event duration, multi agent interactions and hierarchical structure.

Xu [47] showed that dynamic pose, which couples the local motion information with the human skeletal pose, is a more powerful representation of human action than skeletal pose without local motion information. Xiong [45] focused coaction discovery: the task of discovering and segmenting the common actions (coactions) between videos that may contain several actions. Agarwal [1] proposed an optimization based framework for human pose estimation, which was tested on the DARPA Mind’s Eye Year-1 dataset. The average error per frame per marker for their pose estimation is 10 pixels (L2 norm) and the average error for different markers per frame is 40 pixels (L1 norm), which is less than the dimension of an average human head in the data set (50 pixels ground truth). Banerjee e.a. [2] also uses human poses for action recognition. They described a method for activity recognition based on distribution of human poses in a video. Ramakrishna e.a. [37] presented an activity-independent method to recover the 3D configuration of a human figure from 2D locations of anatomical landmarks in a single image, leveraging a large motion capture corpus as a proxy for visual memory. Izadinia e.a. [23] evaluated the performance of a widely used tracking-by-detection and data association multi-target tracking pipeline applied to the year 2 dataset of the program. On average, their detector improves recall by 23% and their tracklet association yields an 87% decrease in the number of ID switches.

It is a challenge to obtain high-level descriptions from videos, or to combine empirical measurements with expert knowledge and bridge the gap between low-level features and high-level descriptions. Saenko [40] proposed a mid-level representations, that can bridge the gap between existing low-level models, which are incapable of capturing the structure of interactive verbs, and contemporary high-level schemes, which rely on the output of potentially brittle intermediate detectors and trackers. Sadanand [39] presented Action Bank, a high-level representation of video. This bank is comprised of many individual action detectors sampled broadly in semantic space and viewpoint space. Their representation should be semantically rich and, when paired with simple linear SVM classifiers, capable of good discriminative performance. Oltramari [35] presented the Cognitive Engine, an integrated system whose architectural characteristics and operational capabilities are designed to approximate human visual intelligence. It tries to make sense of a scene by clustering visual data: basic individual movements are interpreted as constituting a particular action, and patterns of actions are gathered into more complex activities. Kerr e.a. [26] developed a way to find parts of activities by a greedy multiple sequence alignment, and a method to transform the alignments into a Finite State Machine that accepts novel instances of activities. Ranasinghe [38] presented a solution for recognizing a set of predefined actions in video streams of variable durations, even in the presence of noise and gaps caused by occlusions or data loss. It works by using surprise-based learning (SBL) to reason on object tracks. The system autonomously learns a set of rules which captures the essential information required to disambiguate each action. Cohn [18] presented an integrated representation of moving objects that includes spatial and temporal aspects for video analysis. This integrated approach appears to give better performance on event detection after evaluation on a subset of the Year-1 dataset.

Especially the methods of Barbu e.a. [5] and Burghouts [14] deserve special attention, because of their good performance in the Year-2 evaluation of the DARPA Mind's Eye program. The first is described in their paper and the latter is described in the following sections.

3. METHOD

3.1 System overview

The CORTEX-system consists of the following components (Figure 1): Visual processing, matchmaker, event description, recognition, description, merge and anomaly detection. Each component will be detailed in the following subsections. The core idea is that visual processing delivers tracks and features. The match maker hypothesizes what agents are involved – this can be either a single person action or a person-person or person-object interaction. For each such (inter)action, the event properties are additional track-based features. Recognition and description are modules to interpret what happens at each (inter)actions, by respectively labels and short phrases. The merge module is responsible for selecting the few relevant labels and phrases and to discard irrelevant ones (e.g. walk, hold, etc.). Finally, anomaly detection aims to find those (inter)actions that did not happen recently.

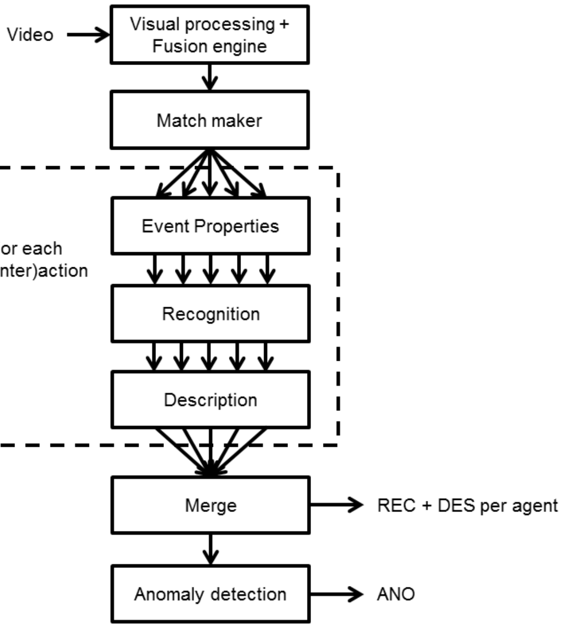

Figure 1: System architectural design.

The baseline implementation [6] was only able to classify complete video-clips; each clip was analyzed and the system reported the presence of different actions in the complete scene without space-time localization. The current novel implementation [14] allows the description of different actors (agents) in the scene separately: for each agent in the scene the actions and interactions are described. In combination with our temporal segmentation, we are able to localize the actions both in space and time. An accurate spatial localization is commonly not important for the surveillance task, but it allows the system to handle more complex and busy scenes with multiple agents and actions.

3.2 Visual features

The visual processing and fusion engine consist of person detection, moving-object detection and tracking [6][10] resulting in tracks describing possible persons spatio-temporally. Subsequently, features are computed.

Local motion is described with spatio-temporal interest point (STIP) features [28]. These features are computed at spatio-temporal interest points (a 3D Harris detector). The features comprise of histograms of gradients (HOG) and optical flow (HOF), which results in a vector of 162 features [15]. The selectivity is increased by weighting each STIP feature with a spatio-temporal saliency map that is specific to the particular video and to the action of interest [13]. In this approach, to avoid complexity the learning of the saliency and the learning of the class are not coupled. Furthermore, the spatial saliency map is extended to the spatio-temporal domain. High weights are given to relevant features both in the foreground and the background to exploit the property that some features aid in deciding that this video contains the action of interest and that other features aid in deciding that it does not contain this, in a dense computation without making distinction between foreground and background.

The weakness of STIP features is that they are only able to describe local motions and that it is not possible to effectively understand high-level relations between entities. Therefore, we also compute event properties (EP) [6]. The two main types of EP features are: single-entity (e.g. type of entity (person/car); an entity moves horizontal; a person moves his arm) and the multiple-entity and relational properties (e.g. distance between two agents decreases). There are 65 single-entity and 13 multi-entity EP features. The EP features in combination with STIP features have shown to improve the overall performance slightly [12]. However, in our current implementation, the STIP features [15] are used for recognition, and the EP features are used together with the recognition output for the description task.

3.3 Matchmaker

All objects that are detected are called entities. The entities that can perform (inter)actions are called ‘agents’ (e.g. persons, cars). The single-entity verbs are called actions (e.g., walk, run) and the multi-entity verbs are called interactions (e.g., pass, give). Each agent has a spatio-temporal region, which is used to ignore relations between agents that are far apart in the scene and limit the possible interactions [14]. Our recognition component is applied to each entity (to assess the actions) and to each pair of entities (for the interactions).

3.4 Recognition

The basic bag-of-features approach uses local spatio-temporal STIP features [28] as input and then applies a random forest to quantize the features into action histograms [9][31], and a SVM classifier with a χ^2 kernel to recognize the verbs (LibSVM).

This approach is improved in several ways. The first improvement uses a second-stage SVM classifier, which uses the correlations between present actions to obtain better results [11][12]. The second improvement uses sample selection. The weight of the positive class is set to (#pos+#neg)/#pos and the weight of the negative class to (#pos+#neg)/#neg, where #pos is the size of the positive class and #neg of the negative class. Sample selection is improved by selecting the negatives that are similar to the positive class [12]. The third improvement is related to the spatio-temporal layout. Each STIP feature does not contribute to the histogram bins by a unity value, but rather by a weight given by its spatio-temporal probability [15]. The fourth and final improvement in the recognition component uses uncertainty in the random forest. The major disadvantage of the random forest is that it makes binary decisions on the feature values, and thus are uncertainties not taken into account. Therefore, a soft-assignment random forest is used [16], where the binary decisions inside the tree nodes are substituted by a sigmoid function.

3.5 Description

For each scene a textual description is generated by a Rule-based system (RBS) [21], which encodes world knowledge about the actions. The conditions of the rules are based on input from the Event properties (EP features). The RBS connects the action with the entity or entities involved in the action and it produces a sentence that at least contain a subject and an action [14]. The rules also take into account that certain verbs are more probable in case there are more entities interacting.

3.6 Merge

In merging, all results for an agent are combined into stories describing its actions, for recognition, description and anomaly. Verbs may be rejected or given a higher priority, using the provided probability, its relative importance, and co-occurrence in time with other verbs.

For each agent, verb probabilities are provided for segments in time, for the action describing the agent in time, as well is in possible interactions with other entities. In merging, all these results are combined into a single set of stories describing its actions in time. Segments reporting the same verb in overlapping intervals are merged into larger intervals. All verbs with similar intervals are evaluated, prioritizing verbs based on the provided confidences and information level (i.e., is a certain verb more interesting or important to report than another). A maximum number of verbs is reported in these combined intervals, as well as for the whole lifetime of the agent.

3.7 Anomaly detection

A separate component was developed for detecting anomalous behavior. Anomaly detection can be used to detect threats and deviant behavior [7]. The Adaptive Outlier Distance (AOD) is used to detect outliers in sparse high dimensional data based on local distance ratios [29]. In the anomaly-detection task a set of normal instances is provided for training, after which abnormal behavior is detected in a test set.

4. EXPERIMENT AND RESULTS

4.1 Data set

The DARPA Mind's Eye dataset of Year-1 (Y1) consists of many short video clips (approx.. 10-30 sec.). It contains a development set of 3,480 clips, an evaluation set that was used in Y1 of 2,588 clips, and an evaluation set that was used in Y2 of 400 novel exemplars of the verbs. In Y2, we added the earlier evaluation set of the previous year (Y1) to our training set. Snapshots of four videos are shown in [6].

The DARPA Mind's Eye dataset of Year-2 (Y2) consists of several long video clips (shortest 2.5 min, longest 17.5 min) of persistent stare. It contains a development set of 8 clips and an evaluation set of 4 clips (approx. 10 minutes each). Four snapshots of Y2 video clips are shown in Figure 2.

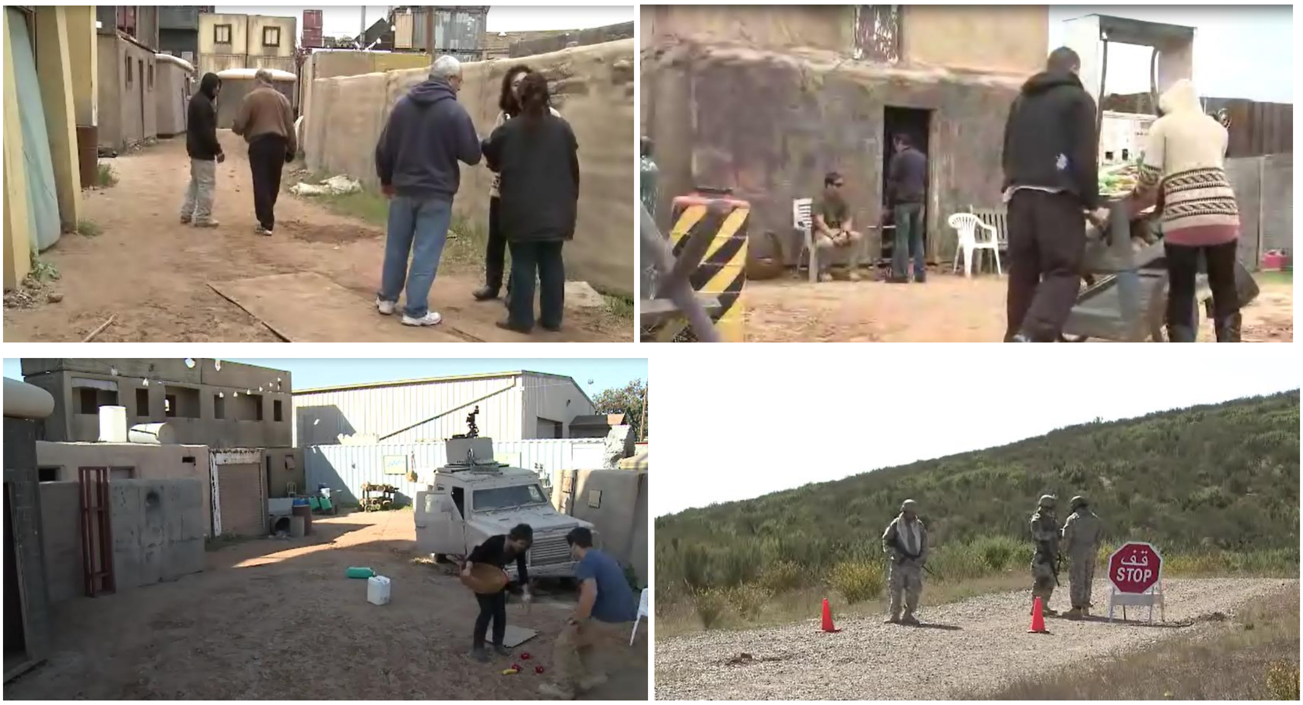

Figure 2: Four examples of persistent stare videos in the Year-2 dataset.

4.2 Experimental setup

The recognizer's performance is measured by the Matthews Correlation Coefficient,

$$\mathrm{MCC} = (TP \cdot TN - FP \cdot FN) / \mathrm{sqrt}((TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)),$$

where T=true, F=false, P=positive and N=negative. The average MCC is computed by first computing the MCC per verb, and then computing an average over all verbs. In [6], the average MCC was initially computed differently (i.e. across the whole set). Below we recomputed the 2012 performance using the average MCC instead for fair comparison with the 2013 results.

Y1 data is evaluated with an MCC on clip level, where for each clip the reported verb vector – indicating the presence / absence of 48 verbs – are compared with the ground-truth verb vector.

On Y2 data, the system response does not only consist of a verb vector per clip, but a verb-vector per segment. A segment is a part of the track, which contains space (oval in a frame), time (period) and verb information. For each verb, the video can be converted to a binary 3D voxel grid, indicating presence/absence as a mask that is used for pixel-wise comparison between segments in the ground truth and as reported by our system. Y2 data is evaluated with a pixel-wise MCC to estimate spatio-temporal localization, with a few variants:

- Spatio-temporal labeled: This test is a full 3D voxel test that sums and compares all the components for the segment-to-segment comparison, where verb labels have to be identical.
- Spatio-temporal unlabeled: This is equal to the spatio-temporal labeled test, where verb labels are ignored. This allows analysis of the quality of tracks and segments (e.g. to detect segment fragmentation).
- Temporal labeled: An alternative test is done that de-emphasizes the location accuracy. The base components are based on the temporal bounds (start and stop times) of the segment only.

For each MCC cell (segment-to-segment comparison), an overlap-threshold (e.g. MCC=0.6) is applied and only those exceeding this threshold are counted. This tests whether a given cell matches a reference cell to a certain degree of overlap in time and space.

4.3 Results on Year 1 data

The results on Y1 data for each verb are shown in Figure 3. The blue bars indicate the performance of the system from [6] and the red bars indicate the current performance. Note that a perfect score is achieved when MCC = 1.

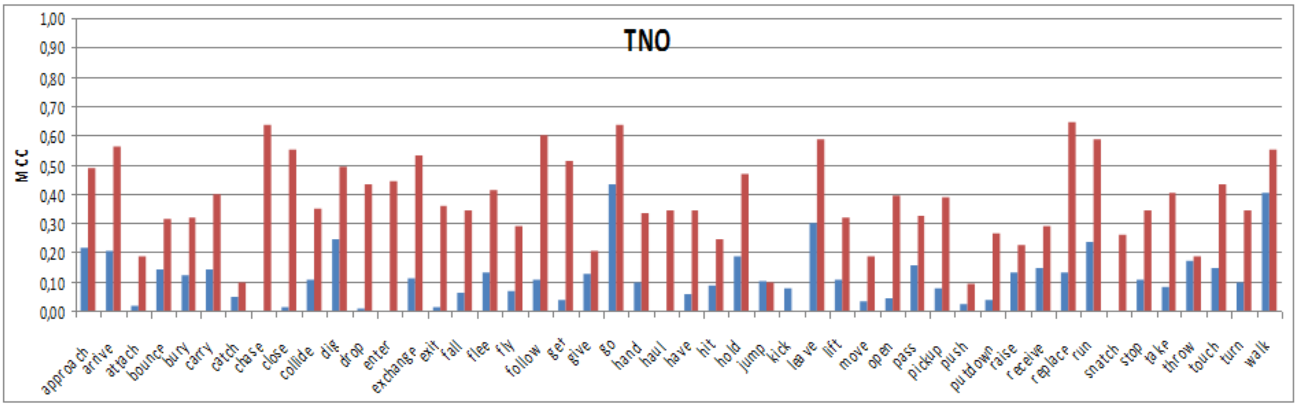

Figure 3: Results on Year-1 data per verb (blue = old results [6], red = new results).

The average MCC of our system is 0.37, and in Table 1 it is compared with the published state-of-the-art. The results show an improvement factor of 3.4 in comparison to the result of last year [6], and it is 42% better than the most recent state-of-the art results [12].

Table 1: Comparison of methods on the Year-1 dataset of DARPA.

Method	Av. MCC
Baseline: EP & STIP + single stage SVM [6]	0.11
Spatio-temporal layout [15]	0.20
STIP + two stage SVM [11]	0.23
Soft-assignment random-forest [16]	0.24
EP & STIP + two stage SVM + selective sampling [12]	0.26
Advanced pipeline described in this paper combining all of the above	**0.37**

4.4 Results on Year 2 data

The results on Y2 data are shown in Table 2. For the four test videos, we indicate the performance of our system when we demand a high degree of overlap between ground truth segments and segments reported by our system (left part of Table 2, for overlap-threshold MCC = 0.6) or when we allow less overlap (right part of Table 2, for MCC = 0.1). Human baseline for the MCC=0.6 overlap-threshold is typically >50%. The table shows that at MCC=0.6, the average score makes a huge increase (from 3% to 31%) if the position in a frame is ignored. This is an important finding, as for many surveillance tasks it is sufficient to know *when* something happens (and the location is not as important). Obviously it is harder to do both spatial and temporal localization of human activity. In a more detailed analysis (results not shown), we have seen that allowing some temporal slack may add 10% improvement and the spatial slack may add 20% improvement. For the MCC = 0.1 overlap-threshold, the performance of our system increases significantly, from 3% to 43% for spatio-temporal localization of human activity. For the identification of when an activity takes place, the performance increases from 31% to 63%.

Table 2: Results on Year-2 data. All values in the table are a percentage (%) of reference segments that could be matched with segments generated by the system for a certain overlap-threshold (MCC = 0.6 or 0.1).

Video	**Overlap threshold: MCC = 0.6**			**Overlap threshold: MCC = 0.1**		
	Spatial Temporal Label	Spatial Temporal	Temporal Label	Spatial Temporal Label	Spatial Temporal	Temporal Label
SH1_03_C1	4	16	54	44	97	54
SH1_06_C4	2	6	37	61	92	81
SH3_08_C2	4	13	19	43	92	58
CR1_11_C2	1	6	13	25	95	58
Average	3 %	10 %	31 %	43 %	94 %	63 %

5. CONCLUSIONS

In this paper, we presented an improved system that can recognize actions of a human and interactions between multiple humans. Central to the new system is our agent-based architecture. The system is trained on thousands of videos and evaluated on realistic persistent surveillance data in the DARPA Mind's Eye program, with hours of videos of challenging scenes. The results show that our system is able to track the people, detect and localize events. This work is a first attempt to discriminate between different behaviors in persistent stare camera setups, which is challenging due to the varying recording conditions and the huge amount of videos. The extensive experiments on such data show that a reasonable performance is achieved for temporal localization of human activities. Full spatio-temporal localization is still a challenge. We have demonstrated that we are able to discriminate 48 human activities in 400 novel recordings reasonably well, and it performs 3.4 times better than our previous system.

6. ACKNOWLEDGEMENT

This work is supported by DARPA (Mind's Eye program). The content of the information does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

REFERENCES

[1] Agarwal, P., Kumar, S., Ryde, J., Corso, J., Krovi, V., "An optimization based framework for human pose estimation in monocular videos," LNCS 7431, 575-586 (2012).

[2] Banerjee, P., Nevatia, R., "Pose based activity recognition using Multiple Kernel learning," ICPR, 445-448 (2012).

[3] Barbu, A., Bridge, A., Coroian, D., e.a., "Large-scale automatic labeling of video events with verbs based on event-participant interaction," arXiv 1204 3616, (2012).

[4] Barbu, A., Bridge, A., Burchill, Z., e.a., "Video in sentence out," Conf. Uncertainty in Artificial Intelligence UAI, (2012).

[5] Barbu, A., Michaux, A., Narayanaswamy, S., Siskind, J.M., "Simultaneous object detection, tracking, and event recognition," arXiv 1204 2741, (2012)

[6] Bouma, H., Hanckmann, P., Marck, J.W., Penning, L., Hollander, R., Hove, J.M., Broek, S.P., Schutte, K., Burghouts, G., "Automatic human action recognition in a scene from visual inputs," Proc. SPIE 8388, (2012).

[7] Bouma, H., Rajadell, O., Worm, D., Versloot, C., Wedemeijer, H.,"On the early detection of threats in the real world based on open-source information on the internet," Int. Conf. Information Technologies and Security ITSEC, (2012).

[8] Bouma, H., Borsboom, S., Hollander, R. den, Landsmeer, S., Worring, M., “Re-identification of persons in multi-camera surveillance under varying viewpoints and illumination,” Proc. SPIE 8359, (2012).

[9] Breiman, L., “Random forests,” Machine Learning 45(1), 5-32 (2001).

[10] Burghouts, G.J., Bouma, H., Hollander, R.J.M. den, Broek, S.P. van den, Schutte, K., “Recognition of 48 human behaviors from video,” Int. Symp. Optronics in Defense and Security OPTRO, (2012).

[11] Burghouts, G.J., Schutte, K., “Correlations between 48 human actions improve their detection,” Int. Conf. Pattern Recognition ICPR, (2012).

[12] Burghouts, G.J., Schutte, K., Bouma, H., Hollander, R.J.M., “Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos,” Machine Vision and Applications, (2013).

[13] Burghouts, G.J., Hove, J.-M. ten, Broek, S.P. van den, “Spatio-temporal saliency for retrieval of complex human actions from videos,” submitted, (2013).

[14] Burghouts, G.J., Penning, L. de, Kruithof, M., Hanckmann, P., Hove, J.M. ten, Landsmeer, J.-M., Broek, S.P. van den, Hollander, R., Leeuwen, C. van, Korzec, S., Bouma, H., Schutte, K., “A search engine for retrieval and inspection of events with 48 human actions in realistic videos,” Int. Conf. Pattern Recognition Applications and Methods ICPRAM, (2013).

[15] Burghouts, G.J., Schutte, K., “Spatio-temporal layout of human actions for improved bag-of-words action detection,” Pattern Recognition Letters, (2013).

[16] Burghouts, G.J., “Soft-assignment random-forest for discriminative representation of human actions in videos,” Int. J. Pattern Recognition and Artificial Intelligence, (2013).

[17] Chaquet, J., Carmona, E., Fernandez-Caballero, A., “A survey of video datasets for human action and activity recognition,” CVIU, (2013).

[18] Cohn, A., Renz, J., Sridhar, M., “Thinking inside the box: a comprehensive spatial representation for video analysis,” Proc. Int. Conf. Principles of Knowledge Representation and Reasoning, (2012).

[19] Fay, M.P., “Enabling imagination through story alignment,” M.Sc. thesis MIT, (2012).

[20] Geller, T., “Seeing is not enough,” Communications of the ACM 54(10), 15-16 (2011).

[21] Hanckmann, P., Schutte, K., Burghouts, G.J., “Automated textual descriptions for a wide range of video events with 48 human actions,” ECCV workshop on Video Event Categorization, Tagging and Retrieval, (2012).

[22] Huang, C., Yeh, Y., Wang, Y., “Recognizing actions across cameras by exploring the correlated subspace,” ECCV, (2012).

[23] Izadinia, H., Ramakrishna, V., Kitani, K., Huber, D., “Multi-pose multi-target tracking for activity understanding,” IEEE Workshop on Appl. of Computer Vision, 385-290 (2013).

[24] Jiang, Y., Bhattacharya, S., Chang, S., Shah, M., “High-level event recognition in unconstrained videos,” Int. J. Multimedia Information Retrieval, (2012).

[25] Junejo, I., Dexter, E., Laptev, I., Perez, P., “View-independent action recognition from temporal self-similarities,” PAMI, (2011).

[26] Kerr, W., Tran, A., Cohen, P., “Activity recognition with finite state machines,” Proc. Int. Joint Conf. Artificial Intelligence IJCAI, (2011).

[27] Khan, F.M., Singh, V., Nevatia, R., “Simultaneous inference of activity, pose and object,” IEEE Appl. Computer Vision, 281-288 (2012).

[28] Laptev, I., “On space-time interest points,” Int. J. Computer Vision IJCV 64, 107-123 (2005).

[29] Leeuwen, C. van, Halma, A., Schutte, K., “Anomalous human behavior detection: an adaptive approach,” Proc. SPIE 8745, (2013).

[30] Liu, H., Feris, R., Sun, M., “Benchmarking human activity recognition,” CVPR tutorial, (2012).

[31] Moosmann, F., Nowak, E., Jurie, F., “Randomized clustering forests for image classification,” IEEE Trans. PAMI 30(9), 1632-1646 (2006).

[32] Natarajan, P., Nevatia, R., “Hierarchical multi-channel hidden semi Markov graphical models for activity recognition,” CVIU, (2012).

[33] O’Hara, S., Draper, B., “Scalable action recognition with a subspace forest,” CVPR, 1210-1217 (2012).

[34] O’Hara, S., Liu, Y., Draper, B., “Using a product manifold distance for unsupervised action recognition,” Image and vision computing 30(3), 206-216 (2012).

[35] Oltramari, A., Lebiere, C., “Knowledge in action: Integrating cognitive architectures and ontologies,” New trends of research in ontologies and lexical resources, 135-154 (2013).

[36] Penning, H.L.H. de, Hollander, R.J.M. den, Bouma, H., Burghouts, G.J., d'Avila Garcez, A.S., "A neural-symbolic cognitive agent with a Mind's Eye," AAAI Neural-Symbolic Learning and Reasoning NeSy, (2012).

[37] Ramakrishna, V., Kanade, T., Sheikh, Y., "Reconstructing 3D human pose from 2D image landmarks," Europ. Conf. Computer Vision ECCV, (2012).

[38] Ranasinghe, N., Shen, W., "Autonomous surveillance tolerant to interference," LNCS 7429, 73-83 (2012).

[39] Sadanand, S., Corso, J., "Action bank: a high-level representation of activity in video," CVPR, (2012).

[40] Saenko, K., Packer, B., Chen, C., e.a., "Mid-level features improve recognition of interactive activities," Tech. Report UC Berkeley USA, (2012).

[41] Schutte, K., Bomhof, F., Burghouts, G., Diggelen, J. van, Hiemstra, P., Hof, J., Kraaij, W., Pasman, H., Smith, A., Versloot, C., Wit, J. de, "GOOSE: Semantic search on internet connected sensors," SPIE 8758, (2013).

[42] Song, H., Zickler, S, Althoff, T., e.a., "Sparselet models for efficient multiclass object detection," ECCV, (2012).

[43] Truyen, T., Bui, H., and Venkatesh, S., "Human activity learning and segmentation using partially hidden discriminative models," HAREM, 87-95 (2005).

[44] Wu, X., Jia, Y., "View-invariant action recognition using latent kernelized structural SVM," ECCV, (2012).

[45] Xiong, C., Corso, J., "Coaction discovery: segmentation of common actions across multiple videos," Int. Workshop Multimedial Data Mining, (2012).

[46] Xu, C., Xiong, C., Corso, J., "Streaming hierarchical video segmentation," ECCV, (2012).

[47] Xu, R., Agarwal, P., Kumar, S., e.a. "Combining skeletal pose with local motion for human activity recognition," AMDO LNCS 7378, 114-123 (2012).

[48] Yao, B., Fei-Fei, L., "Action recognition with exemplar based 2.5D graph matching," ECCV, 173-186 (2012).

[49] Yu, G., Yuan, J., Liu, Z., "Propagative Hough voting for human activity recognition," ECCV, 693-706 (2012).