Automatic human action recognition in a scene from visual inputs

Henri Bouma*, Patrick Hanckmann, Jan-Willem Marck, Leo Penning, Richard den Hollander, Johan-Martijn ten Hove, Sebastiaan van den Broek, Klamer Schutte and Gertjan Burghouts.

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

ABSTRACT

Surveillance is normally performed by humans, since it requires visual intelligence. However, this can be dull and dangerous, especially for military operations. Therefore, unmanned autonomous visual-intelligence systems are desired. In this paper, we present a novel system that can recognize human actions, which are relevant to detect operationally significant activity. Central to the system is a break-down of high-level perceptual concepts (verbs) in simpler observable events. The system is trained on 3482 videos and evaluated on 2589 videos from the DARPA Mind's Eye program, with for each video human annotations indicating the presence or absence of 48 different actions. The results show that our system reaches good performance approaching the human average response.

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

1. INTRODUCTION

Ground surveillance is a mission normally performed by human assets. Military leaders would like to shift this mission to unmanned autonomous systems, removing troops from harm's way. However, unmanned systems lack a capability that currently exists only in humans: visual intelligence. The Defense Advanced Research Projects Agency (DARPA) is addressing this problem with Mind's Eye, a program aiming to develop a visual-intelligence capability for unmanned systems. DARPA has contracted several research teams to develop fundamental machine-based visual intelligence. TNO's participation in this program has led to the development of a novel system called CORTEX.

In this paper, we present the CORTEX system that can recognize and reason about the verbs and nouns, enabling a complete description of actions. Our system is inspired by human intelligence and it uses world knowledge and visual evidence to support decisions. The central element in our system is to break down high-level perceptual concepts to simpler and reusable observable cues. These cues allow to reason over the actions with several methods, including a manually generated rule-based expert system and automatically trained classification systems.

The system is trained on 3482 videos and evaluated on 2589 videos (available at visint.org), all provided by the Mind's Eye program of DARPA. The program consists of four tasks (recognition, description, gap filling and anomaly detection). This paper focuses on the recognition task. For the recognition task, a ground truth based on human annotations was provided and it contains information about the presence or absence of 48 verbs for each video.

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

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

Henri Bouma, Patrick Hanckmann, Jan-Willem Marck, Leo Penning, Richard den Hollander, Johan-Martijn ten Hove, Sebastiaan van den Broek, Klamer Schutte and Gertjan Burghouts, "Automatic human action recognition in a scene from visual inputs", Proc. SPIE, Vol. 8388, 83880L (2012); http://dx.doi.org/10.1117/12.918582

2. METHOD

2.1 System overview

The CORTEX-system consists of the following components (Figure 1): visual processing, fusion engine, event description, reasoning and reporting. A more detailed view of the architecture is shown in [8]. The aim of this system is to recognize 48 actions and to report a believe probability for each action.

Figure 1. System architectural design

2.2 Visual processing

The visual processing [7] of a scene starts with the detection of meaningful objects and their properties. The detection of objects is performed in two ways. First, moving objects are detected by background subtraction [19][22]. This provides a segmentation of all moving persons, vehicles and other objects in the scene in case they are moving during the activity. Second, a trained object detector for specific classes like persons and cars is used to detect their instances in single frames [9][15]. This enables the detection of objects without the requirement that they are moving. Many activities of persons entail e.g. small arm movement while the whole body stays at the same position. When objects have been detected, they are being tracked through the video stream. The object's position over time is an essential property that directly relates to its action. The tracker the dimensions and colors of detected objects to find new object positions in subsequent frames [21][2]. The model is updated to follow objects whose appearance changes slowly over time.

Stationary objects can perform actions that are not captured through coarse movement of the whole object. Therefore, a more detailed description of pose [10][20] and body part movement is computed. Feature descriptors of limb movement based on (space-time) interest points [1][12][14][16], skin color, structural and statistical motion descriptors and salient regions [11] are computed for all objects that have been recognized as persons (by means of the class-specific person detector).

2.3 Fusion engine

The purpose of the fusion engine [3][8] is to filter and possibly fuse the tracked objects in order to form entities. Entities should correspond to the real-world entities like a person, bike or car that contribute to the observed action in the scene. Only the entities – a subset of the detected and tracked objects – are selected for further processing. The output of the fusion engine is a container for each entity, which includes the track information and low-level visual features. To limit computation time there is a delayed execution of several features in visual processing, which are only determined where confident entities are found.

2.4 Event description

The aim of event description is to raise the level of abstraction from the low-level features towards the object or situation level that is desired to express the rules of an expert system for action recognition. The event properties, and derived rules, are our way of encoding world-knowledge about the 48 verbs. The properties are related to physical world properties and they are based on a taxonomy that positions a verb in a semantic hierarchy and makes explicit how humans assess and describe events. Three types of event properties are generated. Single-entity event properties, entity-pair event properties and global event properties. The first type of events describe properties of one entity (e.g., "the entity is moving fast"). The second type of events describe the relation between two entities (e.g., "the distance between two entities is decreasing"). The third type of events describe global properties of the scene (e.g., "there is only one entity present").

2.5 Reasoning

The reasoning component retrieves information on entities and relations present in the video clip from event description. Based on this information the component infers and describes the behavior of the entities observed in the clip and reports this to the reporting component. In order to do so, the component can be trained on the ground truth available for the observed clips.

The reasoning component contains four independent classifiers:

- RBS: Rule-Based System [3]. A set of 73 manually created rules, with each several spatio-temporal conditions on event properties, is mapped onto the set of 48 verbs. For several rules, it was not possible to define a rule, e.g., due to missing event properties. A multi-hypothesis partial matcher is designed which uses the best match per rule.
- RF-TP: Random-Forest Tag-Propagation [4]. Also a rule-based recognizer, yet here the rules are learned from an abundant set of decision trees (i.e. a random forest). The novelty of the usage of these rules is to consider the similarity distributions over them. The core of the RF-TP method is that it models the probability of a verb for the current vignette (or video clip) as a consequence of the similarities with all of the previously seen vignettes and the verbs that are active in those.
- RTRBM: Recurrent Temporal Restricted Boltzmann Machine [17][18]. A generative statistical learner that incorporates temporal relations and evidence from observations (in our case event properties).
- HUCRF: Hidden-Unit Conditional Random Field [6]. Similar to RTRBM, yet now a discriminative learner.

Two of the reasoners (RTRBM, HUCRF) use a condensed version of the event properties which is projected such that no longer can be distinguished which entity has which property due to implementation constraints.

2.6 Reporting

This component reports the results provided by the reasoning component in a pre-defined format for each task (i.e. recognition, description, gap-filling and anomaly detection). The output for the recognition task, which is the focus of this paper, consists of a vector of 48 probabilities indicating signal strength across the set of verbs for each video clip. The soft probability assignment is converted to a binary presence or absence value of a verb by applying a threshold.

3. EXPERIMENT AND RESULTS

3.1 Experiment

The system is trained on a development set (3482 video clips) and after training it is evaluated on a test set (both provided by the Mind's Eye program of DARPA). The test set consists of 2588 video clips (containing 48 verbs, 54 vignettes per verb, 10 exemplars of a verb, and some problematic exemplars were excluded by DARPA).

A ground truth based on human annotations for the recognition task contains information about the presence or absence of 48 verbs for each video. The ground truth was established by a large group of crowd-source annotators (Amazon Mechanical Turk). For every verb in every clip a yes/no answer was received to the question "Is verb X present?". The clips and questions about verb X are randomly spread over the human annotators. The annotators that gave answers that statistically deviated from others, were manually corrected by DARPA. Figure 2 shows the positive response frequency by verb class and the large variation in class size responses. Note that typically multiple verbs are present in a single clip (for example: move, walk, approach and give).

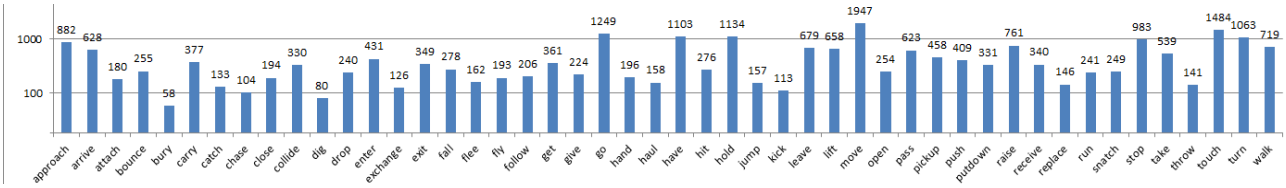

Figure 2. Positive human response frequency by verb class for the test set. Note the logarithmic scale and the variation in class size responses.

The ground truth was defined as follows:

- Ground truth. Due to a very limited number of human responses per clip, we used the mean human responses to vignettes of the same exemplar as our ground truth for all vignettes in that exemplar.

Two reference responses were computed based on ground truth to interpret the quality of the system response.

- Human average response. This is a competitive reference that indicates how well humans assess verbs. We consider the correspondence between human responses to vignettes of the same exemplar. First, we compute the mean response on each exemplar (our ground truth) and then we determined the distance from each human response to this mean response (the distance measure is discussed in Sec. 3.2). The 'Human Average' is the response that corresponds to the average distance for all humans on all vignettes within an exemplar.
- Baseline response. This is a lowerbound reference for a standard, i.e. non-varying, simple response. The simple response is the mean response of humans to the entire development set; so for every clip the same response is given. Clearly this is a lowerbound for which we achieve performance measures when only using a priori information, to compare our recognizers' performances with.

Because some of the video clips in the development set were included in the test set, the results are computed for the whole test set and also separately for seen and unseen material. We have distinguished three subsets within the test set. In increasing order of difficulty: clips we have seen before (these were also contained in the train set), unseen clips yet similar variations of the 48 behaviors (e.g., the same action under a different recording angle), and totally unseen exemplars of the 48 behaviors. For example, 'fly' can mean that a person makes wild arm movements with a kite in his hand and it can also mean that an object enters the scene and lands in a dustbin.

3.2 Results

To estimate the performance of the CORTEX-system, we applied the system to the whole test set, seen vignettes, unseen vignettes in seen exemplars and unseen exemplars, and the following average performance measures were computed per verb: precision, recall, F-measure and the Matthews Correlation Coefficient (MCC). The F-measure (or F1-score, Eq. 1) is the harmonic mean of precision and recall, so both need to be high to get a good score:

$$F = 2\frac{precision \cdot recall}{precision + recall} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \quad (1)$$

where TP is true positive, FN is false negative and FP is false positive for each verb.

The MCC (Eq. 2) is a balanced measure of correlation which can be used even if the classes are of very different sizes as in this dataset. It is in essence a correlation coefficient between the observed and predicted binary classifications and it returns +1 for a perfect prediction, 0 for a random prediction and −1 for an inverse prediction.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(FP + FN)(TN + FP)(TN + FN)}} \quad (2)$$

Several example video clips with tracked objects are shown in Figure 3, and the results are shown in Table 1, Table 2, Figure 4 and Figure 5.

Figure 3. Examples of four video clips with bounding boxes indicating the tracked objects.

Table 1. Overall results (Human performance is only available for vignettes that were seen before in the development set.) The human and baseline references are shown in italic and the best results are shown in bold.

Overall	*Human*	*Baseline*	RBS	HUCRF	RTRBM	RF-TP
F-measure	*0.578*	*0.400*	0.446	0.405	0.399	**0.563**
Precision	*0.594*	*0.406*	0.387	0.386	0.407	**0.503**
Recall	*0.573*	*0.396*	0.541	0.430	0.400	**0.647**
MCC	*0.482*	*0.288*	0.333	0.275	0.288	**0.473**

Table 2. Results for seen vignettes, unseen vignettes of seen exemplars and unseen exemplars.

		Baseline	RBS	HUCRF	RTRBM	RF-TP
Seen vignettes (1294 clips)	*F-measure*	*0.390*	0.448	0.395	0.391	**0.648**
	Precision	*0.389*	0.386	0.371	0.393	**0.573**
	Recall	*0.392*	0.549	0.426	0.396	**0.758**
	MCC	*0.277*	0.338	0.265	0.279	**0.482**
Unseen vignettes seen exemplars (383 clips)	*F-measure*	*0.429*	0.463	0.432	0.427	**0.496**
	Precision	*0.448*	0.410	0.425	0.446	**0.451**
	Recall	*0.418*	0.551	0.451	0.423	**0.557**
	MCC	*0.325*	0.352	0.310	0.323	**0.389**
Unseen exemplars (911 clips)	*F-measure*	*0.404*	**0.434**	0.409	0.399	**0.434**
	Precision	*0.415*	0.379	0.395	**0.411**	0.393
	Recall	*0.401*	**0.535**	0.436	0.403	0.500
	MCC	*0.292*	**0.315**	0.279	0.287	**0.316**

Table 1 shows that the RF-TP performs overall the best on the whole test set. It performs clearly better than the baseline, and it is close to human response. Table 2 shows that the performance of RF-TP decreases on unseen exemplars, so it is slightly overtrained. Both RF-TP and RBS perform equally well on the unseen exemplars, and still clearly better than the baseline.

Figure 4 and Figure 5 shows the F-measure and MCC for each verb on the whole test set.

Based on the results, we observe the following. With RF-TP, we showed that the visual features and event properties, on which our overall system is based, capture essential event characteristics and are discriminative.

Overall, the scores of RF-TP are similar to the human average response. Also the RBS performs clearly better than the baseline. The performance of the RTRBM and HUCRF does not exceed the baseline reference. This may indicate that automatically training temporal causal models is a hard task on this dataset, given no temporal annotation.

On seen vignettes, the RF-TP is clearly better than the baseline and performs often better than human average. On unseen vignettes and seen exemplars RF-TP still clearly performs better than the baseline. This indicates that our system is able to handle small variations in the actions. On unseen exemplars, the RBS and RF-TP seem to perform slightly better than the baseline. So even for complete different variations of a verb, the performance does not drop below the baseline. The robustness of RBS is better than that of RF-TP. Of the systems we evaluated, our it generalizes best. RBS is very well able to achieve similar performance on seen and unseen exemplars.

There seems to be a relation between the scores and the prevalence of verbs. We optimized our system for the average F-measure for all verbs. The F-measure gives a different weight to TP than to TN, which stimulates this relation. The MCC is perfectly symmetric for both and optimizing for this measure could help to improve the performance on verbs with low prevalence.

By inspecting the output of our system on various clips, we observed that the visual processing and fusion engine may deliver broken tracks, which will hamper correct reasoning. Furthermore, we observed that in the evaluated implementation, event description is not always able to detect carried items and distinguish detailed interactions correctly.

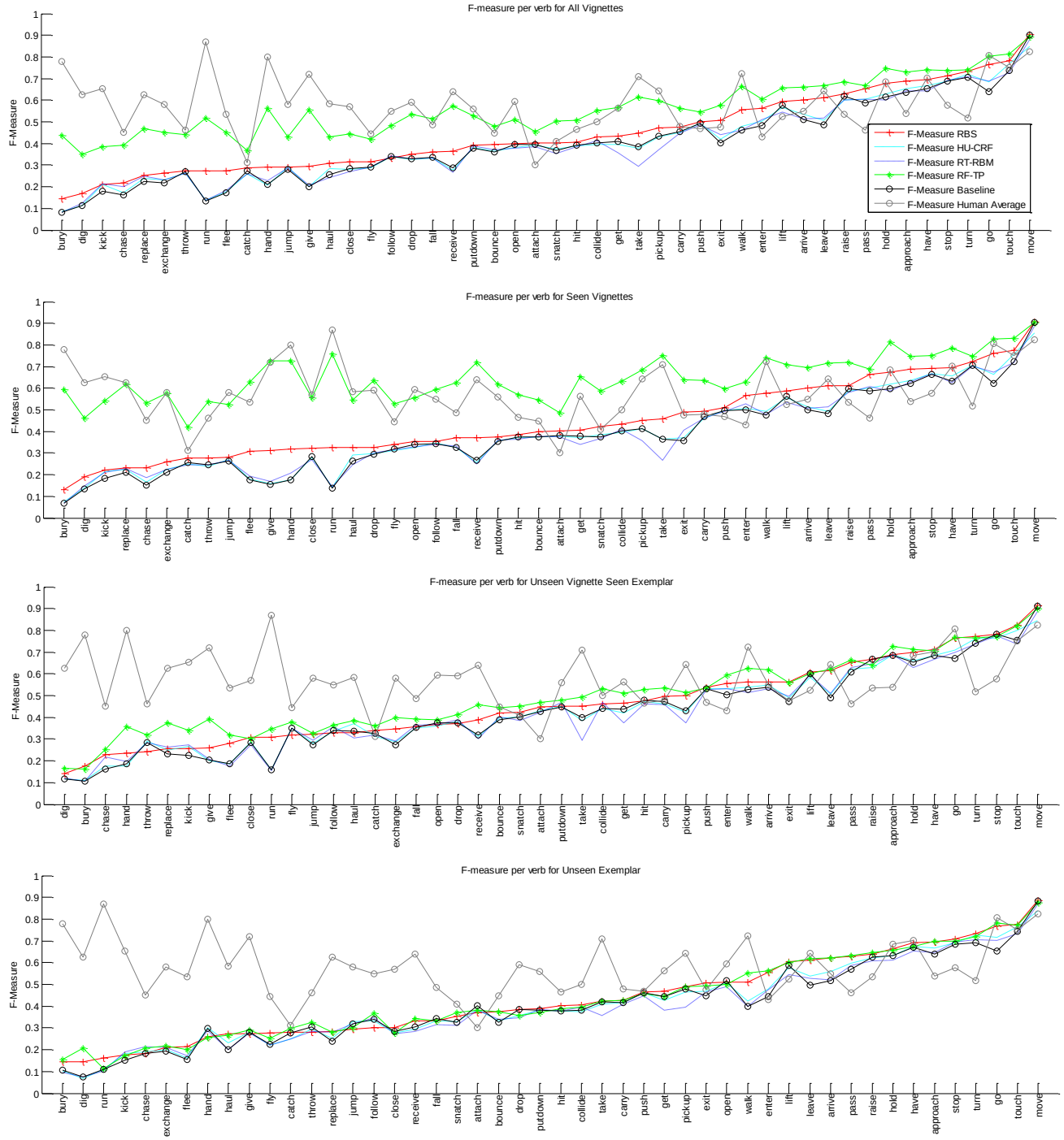

Figure 4. F-measure per verb on the complete evaluation set, seen vignettes, unseen vignettes and seen exemplars, and unseen exemplars.

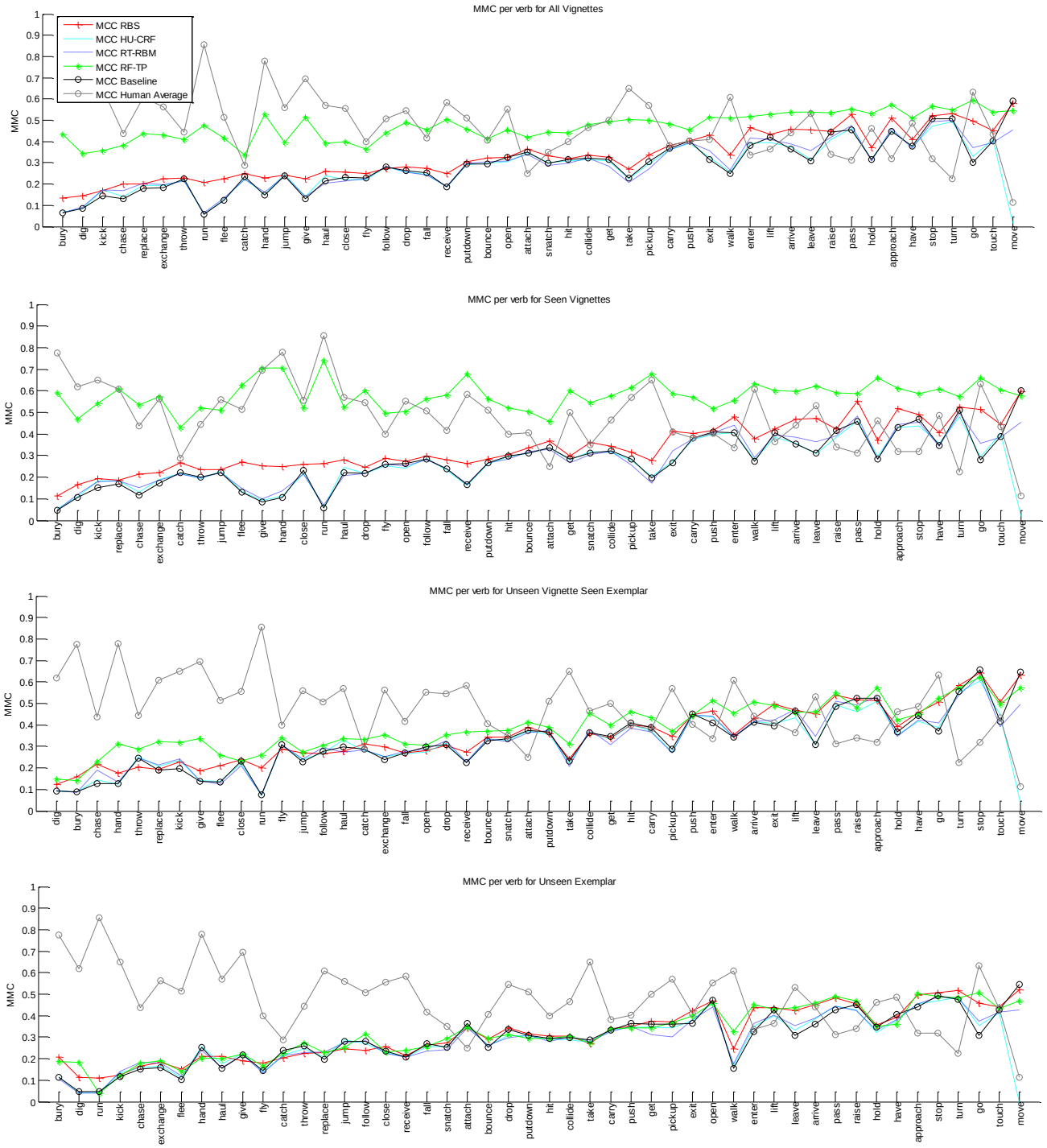

Figure 5. MCC per verb on the complete evaluation set, seen vignettes, unseen vignettes and seen exemplars, and unseen exemplars.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the CORTEX system that can recognize and reason about the verbs, enabling a more complete description of actions. The central element in our system is to break down high-level perceptual concepts to simpler and reusable observable cues. The event properties, and derived rules, are our way of encoding world-knowledge. The choice of properties is based on a taxonomy that positions a verb in a semantic hierarchy and makes explicit how humans assess and describe events. These properties allow us to reason over the actions with several methods, including a manually generated rule-based expert system (RBS) and an automatically trained random-forest tag propagator (RF-TP).

The system was trained on 3482 videos and evaluated on 2589 videos (both provided by the Mind's Eye program of DARPA). A ground truth based on human annotations contains information about the presence or absence of 48 verbs for each video.

We compared our systems response to the manual annotations. Of the recognition systems we evaluated, our RBS generalizes best and the RF-TP – although it was slightly overtrained – reaches a good overall performance approaching human average response for many verbs. We have developed a recognizer (RF-TP) that compares to human performance for seen vignettes, and has a significant potential when it becomes more robust for unseen vignettes. For example, by combining with the human-created rules of the RBS, by (manual or automatic) feature selection, or by using less leaves in the RF-TP.

The event properties can be improved by extension (e.g. detection of items that play a role in the verbs, carried items, and detailed interaction between persons) and by improvement of those properties that do not sufficiently contribute where we expected them to be valuable. At the same time, we will investigate selection of the properties to exclude features that do not perform well. The RBS could also be improved by extension, because there are verbs and verb variants for which no rules were hand-crafted yet. The trained classifiers use alternative information that was not taken into account in the rules (e.g., body parts), handle errors in input (broken tracks) and inconsistencies in annotation. So, these perform better in degenerative cases, but may also over-train.

Two of the four reasoners (RTRBM, HUCRF) use a condensed version of the event properties which is projected such that no longer can be distinguished which entity has which property. Although driven by implementation constraints, we lose selectivity and maintaining this relational information between entity and property could improve the performance. Future work will focus on recognition in an entity-based pipeline.

Temporal annotations have recently become available in the Mind's Eye program. Because there is a clearer relation between events and annotation, we expect that it will improve our recognizers.

We have found that prevalent verbs dominate the learning of event recognizers. We could become more invariant to such prevalence, either by balancing the learning set, optimizing for another performance measure, or by altering truly the recognizer or combining recognizers.

5. 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 Government, and no official endorsement should be inferred.

REFERENCES

[1] Bay, H., Ess, A., Tuytelaars, T., Gool, L. van, "SURF: Speeded up robust features", Computer Vision and Image Understanding 110(3), 346-359 (2008).

[2] Bouma, H., Borsboom, A.S., Hollander, R. den, Landsmeer, S.H., Worring, M., "Re-identification of persons in multicamera surveillance under varying viewpoints and illumination", Proc. SPIE 8359, (2012).

[3] Broek, S.P., Hanckmann, P., Ditzel, M., "Situation and threat assessment for urban scenarios in a distributed system", Proc. Int. Conf. Information Fusion, (2011).

[4] 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).

[5] Burghouts, G.J., Geusebroek, J.-M., "Performance evaluation of local colour invariants", Computer Vision and Image Understanding 113(1), 48-62 (2009).

[6] Burghouts, G.J., Marck, J.W., "Reasoning about threats: From observables to situation assessment", IEEE Trans. Systems Man and Cybernetics 41(5), 608-616 (2011).

[7] Burghouts, G.J., Hollander, R. den, Schutte, K., Marck, J.W., Landsmeer, S.H., Breejen, E. den, "Increasing the security at vital infrastructures: automated detection of deviant behaviors", Proc. SPIE 8019, (2011).

[8] Ditzel, M., Kester, L., van den Broek, S., "System design for distributed adaptive observation systems", IEEE Int. Conf. Information Fusion, (2011).

[9] Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D., "Object detection with discriminatively trained part based models", IEEE Trans. Pattern Analysis and Machine Intelligence 32(9), 1627-1645 (2010).

[10] Ferrari, V., Marin-Jimenez, M., Zisserman, A., "Progressive search space reduction for human pose estimation", IEEE Computer Vision and Pattern Recognition, (2008).

[11] Harel, J., Koch, C., Perona, P., "Graph-based visual saliency", NIPS, (2006).

[12] Harris, C. and Stephens, M., "A combined corner and edge detector", Proc. Alvey Vision Conf., 147–151 (1988).

[13] Hu, N., Bouma, H., Worring, M., "Tracking individuals in surveillance video of a high-density crowd", Proc. SPIE 8399, (2012).

[14] Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B., "Learning realistic human actions from movies", IEEE Computer Vision and Pattern Recognition, (2008).

[15] Laptev, I., "Improving object detection with boosted histograms", Image and Vision Computing 27(5), 535-544 (2009).

[16] Lowe, D.G., "Distinctive image features from scale-invariant keypoints", Int. J. Computer Vision 60(2), 91-110 (2004).

[17] Penning, H.L.H., d'Avila-Garcez, A.S., Lamb, L.C., Meyer, J.J.C., "A neural-symbolic cognitive agent for online learning and reasoning", Proc. Int. Conf. Artificial Intelligence, 1653-1658 (2011).

[18] Penning, L., "Visual intelligence using neural-symbolic learning and reasoning", Proc. Neural-Symbolic Learning and Reasoning, 34-35 (2011).

[19] Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B., "Image change detection algorithms: a systematic survey", IEEE Trans. Image Processing 14(3), 294-307 (2005).

[20] Ramanan, D., "Learning to parse images of articulated bodies", Adv. Neural Inf. Processing Systems, (2006).

[21] Withagen, P.J., Schutte, K., Groen, F.C.A., "Likelihood-based object detection and object tracking using a color histograms and EM", Proc. IEEE Int. Conf. Image Processing (1), 589-592 (2002).

[22] Withagen, P.J., Schutte, K., Groen, F.C.A., "Probabilistic classification between foreground objects and background", Proc. IEEE Int. Conf. Pattern Recognition (1), 31-34 (2004).