Automated Textual Descriptions for a Wide Range of Video Events with 48 Human Actions*

Patrick Hanckmann, Klamer Schutte, and Gertjan J. Burghouts

TNO, The Hague, The Netherlands

Abstract. Presented is a hybrid method to generate textual descriptions of video based on actions. The method includes an action classifier and a description generator. The aim for the action classifier is to detect and classify the actions in the video, such that they can be used as verbs for the description generator. The aim of the description generator is (1) to find the actors (objects or persons) in the video and connect these correctly to the verbs, such that these represent the subject, and direct and indirect objects, and (2) to generate a sentence based on the verb, subject, and direct and indirect objects. The novelty of our method is that we exploit the discriminative power of a bag-of-features action detector with the generative power of a rule-based action descriptor. Shown is that this approach outperforms a homogeneous setup with the rule-based action detector and action descriptor.

1 Introduction

This paper proposes a method to generate textual action descriptions from general videos. The action descriptions are centered around 48 verbs such as walk, bury, approach, give, etc [1].

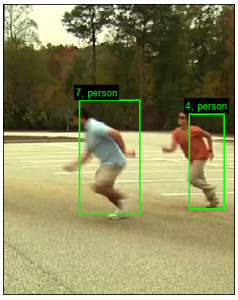
The amount of video data is increasing daily, both on the internet (e.g. YouTube) and for surveillance applications. This poses a challenge on extracting information from this huge bulk of data. In this paper, we consider the automated search for relevant event in videos. One determinant of an event's relevancy, is the action that is performed by humans in the scene. We argue that most events are characterized by multiple actions, and not a single one. A typical event is that one person approaches the other, walks up to the other person, and gives something. These actions, 'walk', 'approach', 'give' and 'receive', occur in a particular order, and are partially overlapping. Moreover, there are two persons in this event. In general, events may consist of multiple actions and performed by one or more persons. Such events are the topic of this paper. Therefore, we progress beyond single-actor datasets such as KTH [2] and Weizmann [3]. The UCF Sports [4], Hollywood2 [5] and YouTube [6] datasets are much more challenging as they involve interactions with other people and items and the recording conditions are harder. Yet they lack the realistic property of having video events which comprise multiple actions. We consider the DARPA dataset

* This work has been sponsored by DARPA, Mind's Eye program.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 372–380, 2012.

[1] in which videoclips are annotated in terms of 48 human actions, where each event consists of on average 7 actions.

In this paper, we consider the automated tagging of realistic video events. We propose a method that produces textual descriptions. The reason for this is that text is intuitive: the popularity of the Google search engine is that it enables a user to perform a text-based search. Our method produces descriptions that cover a wide range of events, they are not limited to a particular domain, and they are based on 48 generic human actions. Figure 1 illustrates the textual descriptions.



Ground truth examples:
Man flees while woman chases him.
A man and woman stand side by side,
the man begins running and the woman follows him.
One person is running and leaving.
The other person starts chasing.
Our system response:
Person 4 goes.
Person 4 leaves.
Person 4 walks.
Person 7 flees from person 4.

Fig. 1. The image shows two people who chase each other. Next to the image the ground truth provided by 3 different people is printed. Our system response provides the detected actions as verbs in a short sentence with their connected subject and object. It shows that our system response captures the essence of the action as described in the ground truth.

Prior research on creating textual descriptions from video has been focused on:

- using speech recognition to generate video subscriptions [7],
- detecting and extracting text which is present in the video [7],
- detecting patterns in a restricted environment and use the detected patterns to generate a description [8,9].

The first two options generate a description based on what can be read or heard in a video. In this paper we rather aim to deduct these description from the video data itself. The third option has only been applied in action restricted environments (e.g. video data from sports in which strict rules apply). Detecting the state of the game directly translates in a description. Behavior seen in general videos is not as structured. The proposed method in this paper includes a detector for 48 human actions and a generic descriptor that generates sentences for a wide range of events based on these actions. The approach in [10] is also generic, but there are three limitations compared to our method: (1) it has a strong focus on describing the environment and describing the subjects' emotional states, where in this paper we do not exploit emotional states as they do not occur in

the considered 48 human actions, (2) it assumes that the subject is always a person, our method generalizes subjects to both people and vehicles, and (3) we extend the set from 5 actions to 48 actions: approach, arrive, attach, bounce, bury, carry, catch, chase, close, collide, dig, drop, enter, exchange, exit, fall, flee, fly, follow, get, give, go, hand, haul, have, hit, hold, jump, kick, leave, lift, move, open, pass, pick up, push, put down, raise, receive, replace, run, snatch, stop, take, throw, touch, turn, and walk.

The contributions of our work are the action classifier and the description generator. The novelty of our work is that we take advantage of the discriminative power of 48 bag-of-features action detectors [11] to identify the subset of likely actions, and to subsequently describe them with a rule-based method that relates the actions to entities in the scene. Important aspects of our method are classification of actions, detection of actors, and connecting the actions to the relevant actors in the video. An actor can be a person or an object. The proposed method is a combination of an action classifier and a description generator.

This paper will introduce the system generating the video descriptions, including the action classifier and description generator, in section 2. In section 3 the experimental setup is discussed, followed in section 4 with the results. Finally our conclusions will be presented in section 5.

2 Method

The action classifier, and the description generator are part of our system. Our system is a video processing system using a pipeline to process the videos. It takes video data as input, and provides the action descriptions as output. An overview of the system components is depicted in figure 2. In subsection 2.1 an overview of the system is presented. The actual action classifier and description generator are described in more depth in subsections 2.2 and 2.3 respectively.

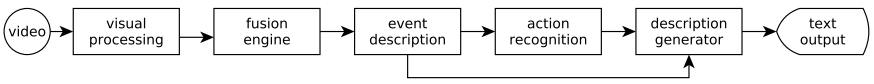


Fig. 2. The processing pipeline

2.1 System Overview

Our method is part of a larger system in which objects in the scene are detected, tracked and their features are captured. This overall system is described in [12,13] and it is summarized here. It consists of five building blocks (see figure 2): visual processing, fusion engine, event description, action classifier, and the description generator.

The **visual processing** [14] incorporates three steps. First the extraction of meaningful objects and their properties from video by (1) detection of moving

objects [15], (2) a trained object detector for specific classes like persons and cars [16,17], and (3) computation of other features (e.g. description of pose and body part movements) [18]. After detection it combines items into tracks.

The **fusion engine** [19] filters and fuses tracked objects in order to form entities. Only entities - a subset of the detected and tracked objects - are selected for further processing.

The **event description** generates a more abstract description. From the low-level object features, information at situation level [20] is created. There are three types of event properties:

1. Single-entity event properties, which describe properties of one entity (e.g. "the entity is moving fast").
2. Relations, properties about the relation between two entities (e.g. "the distance between two entities is decreasing").
3. Global properties of the scene, which present information about the scene that is not exclusively related to one or more entities (e.g. "there is only one entity present").

A belief value is assigned to each property.

The **action classifier** assigns to all 48 human actions a probability (see also [1]). We consider two types of action classifiers, which we will compare in our experiments. The first type is a discriminative bag-of-features classifier. The second type is a generative rule-based classifier. The two classifiers will be described in more detail in section 2.2. In section 2.3, we experimentally establish the best classifier to generate textual descriptions for video.

The **description generator** uses the events from the Event Description to build hypothesis about what happened in the video. The most likely hypothesis are selected based on the classified actions combined with the information from the hypothesis. The selected hypothesis connect the actions to entities and objects. If there are entities or objects that can be connected with the action, then a textual description is generated.

2.2 Action Classifier

The aim of the action classifier is to recognize the verbs that are used for the description generator. The Random-Forest Tag-Propagation (RF_TP) and multi-hypotheses Rule Based System (RBS) classifiers are considered. The choice is based on performance: the RF_TP performs best [21] and the RBS performed second best as actions classifiers in previous research [13] on the DARPA dataset under investigation (see section 3 for details).

The RF_TP classifier [11] is a rule-based classifier which learns its rules from an abundant set of decision trees (i.e. a random forest) [22]. In order to deliver a list of actions with their probability, the similarity distributions over the actions is calculated [23]. The core of the RF_TP method is that it models the probability of a verb (in a video) as a consequence of the similarities with all of the previously seen videos and the actions that are active in those. The training of the RF-TP

is described in [11]. The RF_TP outputs a vector containing a belief value for each action present in the video.

The RBS classifier is a Rule Based System. World knowledge, coded in the rules, describes the actions. There are currently 73 rules describing 48 actions. The rules are essentially a set of conditions. The conditions are based on the beliefs and relations as generated by the event description (see example 1). In the example, *E1*, *T1*, etc. are placeholders for actual entity identifiers and timestamps. As more than one entity can be present at any time, and as actions might happen multiple times by one or different entities, the RBS builds multiple hypotheses. The belief value of the action is calculated by taking the sum of the beliefs of the triggered conditions (and if the condition is not triggered, it's belief is zero), divided by the maximum possible performance: $B(hypothesis) = \frac{\sum B(conditions)}{number\ of\ conditions}$. In this way an inexact match between the rules and noisy input data is allowed. For every action, the top hypothesis is selected. For each hypothesis, a belief value is calculated. There are 73 hypotheses in total, so we have 73 beliefs. These belief values are matched to the 48 human actions. We use a simple linear mapping obtained from a least-squares fit as a linear L2 norm optimization.

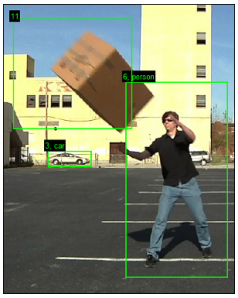
Example 1. Rule representing the catch action.

action = chase
condition(1) = object(E1) moving at time(T1)
condition(2) = object(E2) moving at time(T2)
condition(3) = object(E1) is a person
condition(4) = object(E2) is a person
condition(5) = time(T1) and time(T2) overlap
Resulting sentence = "Person E1 chases Person E2"

2.3 Description Generator

The RBS is also used to generate descriptions. It can be applied as description generator due to the generative properties and the structure of the rules which connects entities and objects to actions. When applied as a description generator the RBS builds the hypotheses and selects for every rule the hypothesis with the highest belief value. Based on the actions classified by the action classifier, a selection is made among the rules (and their top hypothesis). For every action a rule is chosen that represents that action. Then, from the list of actions the description generator selects a number of actions based on: (1) the action probability, (2) the hypothesis score generated by the RBS, and (3) if an entity or object is present in the best hypothesis (which implies that the action is connected to an entity or object). For the selected actions, the hypothesis are used to extract the subject and objects (see example 1). The actions are used as the verbs. A sentence is considered to at least contain a subject and a verb (e.g. person *E1* catches). However, the rule can also provide the direct and indirect object (e.g. person *E1* catches object *E2*). Additionally the hypothesis provides temporal information for the action, which can be used to order the actions in

time. Finally, a compact sentence is generated for each action using a template filling approach. The template provides information about what prepositions are used in combination with specific verbs, the order of the words in the sentence, and the conjugation of the verbs.



Ground truth examples:
A man catches a box that is flying through the air.
The person caught the box.
Man catches box flying through the air.
Our system response:
Entity 11 goes.
Entity 11 catches person 6.
Person 6 has.

Fig. 3. The image shows one person catching a box. Next to the image the ground truth provided by 3 different people is printed. Our system response provides the detected actions as verbs in a short sentence with their connected subject and object. It shows that our system response captures the essence of the action as described in the ground truth. However, it confuses the subject and the direct object.

3 Experimental Setup

The description generator is evaluated on 241 short videos (available at [1]). For all videos ground truth is available. The ground truth consist of 10 sentences per video, written by 10 different people. The ground truth can contain complex sentences (see the examples in figure 3, and note the confusion of the subject and object in the video) and therefore describe multiple actions. Per video at minimum 1, at maximum 10, and at average 4.9 different actions are present in the ground truth.

For each ground truth sentence we extract, using the Stanford Natural Language Parser [24,25], the verb, subject, and object(s). The subject and objects are labeled with one of the four categories: person, car, bike, or other. The description generator constructs a sentence containing a verb and subject, and (if detected) a direct and indirect object. Its subject and objects are also labeled with one of the four categories.

The experiment will compare the following action classifier - description generator combinations: RBS + RBS, and the RF_TP + RBS. Both setups of the description generator use the same event description data. The RBS + RBS uses the RBS both for action classification and description generation. The RF_TP + RBS uses the RF_TP to classify actions and the RBS to generate descriptions. For the RF_TP + RBS the rule set was optimized to gather information about the subject and objects to generate a sentence. For the RBS + RBS setup the rule set was optimized for the classification of actions.

We calculate two performance measures: a union and a percentage score. For each clip we compare the ground truth sentences to the generated sentences. The clip's *union score* is the best match for all sentence pairs (i.e. the percentage of clips where there is at least one agreement between ground truth and generated sentences); its *percentage score* is the mean match corrected for the minimum number of the amount of ground truth sentences and the amount of generated sentences (i.e. the agreement between the sets of ground truth and generated sentences). We report the average over all clips, for verbs, subjects and objects as well as an overall score (the overall score is the mean of the verb, subject and object scores).

4 Experimental Results

The performance for both the RBS + RBS and RF_TP + RBS is given in table 1. Both on union and the percentage score we see the better performance for the RF_TP + RBS compared to the RBS + RBS, supported by an increase for the descriptions' Verb, Subject and Object components.

Table 1. Performance of the description generator

RBS + RBS				
Score	Overall	Verb	Subject	Objects
union	61.6%	86.1%	52.3%	51.7%
percentage	25.4%	38.6%	18.6%	18.9%
RF_TP + RBS				
Score	Overall	Verb	Subject	Objects
union	68.3%	92.3%	62.0%	67.8%
percentage	40.4%	59.3%	30.5%	31.5%

The performance gain for the verb classification on the union score is 6.2%, thus more correct verbs have been reported by the RF_TP + RBS. For the percentage score the improvement is 20.7%, so we also have an improved accuracy of the classified verbs.

The performance on the subjects increased as well for both the union and the percentage score, with resp. 9.7% and 11.9%. Every generated sentence does at least contains a verb and a subject. The performance gain of the subject score is less than the verbs performance gain, while it would be expected to be similar or higher. Both the object and the subject score suffer from too restrictive threshold on the person, car and bike detectors leading to many entities labeled 'other'.

The performance on the objects increased for the union and the percentage score by 16.1% and 12.6%. It shows that the RF_TP + RBS is better in detecting and connecting the direct and indirect objects in a video.

The results show that the discriminative bag-of-features based RF_TP is better used as verb classifier than the RBS when creating descriptions. Although

[13] already showed that the RF_TP is a good stand alone verb classifier, here we see it also performs well when applied to a description task. Even though the RF_TP classifier is not optimized for the description generator (e.g. the set of frequently occurring actions may be different) we conclude that a dedicated action classifier improves the performance significantly.

5 Conclusions and Future Work

This paper shows that a dedicated action classifier in addition to a description generator improves the description performance significantly. Although not perfect, a percentage score of almost 60% on correctly reported verbs is quite good.

The performance on the subject and objects classification is currently low. The issue is misclassification of the actors in the video and as a result reporting "other" as classification too often. Still, we showed a significant increase in the subject and object recognition scores. This increase can be attributed to a better understanding of the scene from the description generator.

The percentage score of the current action classifier is expected to improve further if we can train the action classifier on the description ground truth. The classification of the subject and objects in the description generator should be improved by adjusting the classifiers in visual processing and by relying more on world knowledge coded in rules in the RBS. Furthermore, the number of track-breaks in the visual processing should be reduced, possibly by using multi-hypotheses tracking, as the current rule set is quite sensitive to track break errors. We expect that these latter two improvements will significantly boost the recognition performance for the subject and objects.

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. The authors acknowledge the Cortex team for their contributions.

References

1. DARPA: Hosting corpora suitable for research in visual activity recognition, in particular, the video corpora collected as part of DARPA's Mind's Eye program (2011), `http://www.visint.org`
2. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: Proc. of ICPR, pp. 32–36 (2004)
3. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. 29, 2247–2253 (2007)
4. Ali, S., Shah, M.: Floor Fields for Tracking in High Density Crowd Scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 1–14. Springer, Heidelberg (2008)
5. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR (2009)
6. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: CVPR (2009)

7. Gagnon, L.: Automatic detection of visual elements in films and description with a synthetic voice- application to video description. In: Proceedings of the 9th International Conference on Low Vision (2008)
8. Gupta, A., Srinivasan, P., Shi, J., Davis, L.: Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2012–2019 (2009)
9. Kojima, A., Tamura, T., Fukunaga, K.: Natural language description of human activities from video images based on concept hierarchy of actions. International Journal of Computer Vision 50, 171–184 (2002)
10. Khan, M.U.G., Zhang, L., Gotoh, Y.: Towards coherent natural language description of video streams. In: ICCV Workshops, pp. 664–671. IEEE (2011)
11. Burghouts, G., Bouma, H., de Hollander, R., van den Broek, S., Schutte, K.: Recognition of 48 human behaviors from video. in Int. Symp. Optronics in Defense and Security, OPTRO (2012)
12. Ditzel, M., Kester, L., van den Broek, S.: System design for distributed adaptive observation systems. In: IEEE Int. Conf. Information Fusion (2011)
13. Bouma, H., Hanckmann, P., Marck, J.-W., Penning, L., den Hollander, R., ten Hove, J.-M., van den Broek, S., Schutte, K., Burghouts, G.: Automatic human action recognition in a scene from visual inputs. In: Proc. SPIE, vol. 8388 (2012)
14. Burghouts, G., den Hollander, R., Schutte, K., Marck, J., Landsmeer, S., Breejen, E.d.: Increasing the security at vital infrastructures: automated detection of deviant behaviors. In: Proc. SPIE, vol. 8019 (2011)
15. Withagen, P., Schutte, K., Groen, F.: Probabilistic classification between foreground objects and background. In: Proc. IEEE Int. Conf. Pattern Recognition, pp. 31–34 (2004)
16. Laptev, I.: Improving object detection with boosted histograms. Image and Vision Computing, 535–544 (2009)
17. 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)
18. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: IEEE Computer Vision and Pattern Recognition (2008)
19. van den Broek, S., Hanckmann, P., Ditzel, M.: Situation and threat assessment for urban scenarios in a distributed system. In: Proc. Int. Conf. Information Fusion (2011)
20. Steinberg, A.N., Bowman, C.L.: Rethinking the JDL data fusion levels. In: NSSDF Conference Proceedings (2004)
21. Burghouts, G., Schutte, K.: Correlations between 48 human actions improve their detection. In: ICPR 2012 (2012)
22. Breiman, L.: Random forests. Machine Learning 45, 1 (2001)
23. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV (2009)
24. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 423–430 (2003)
25. The Stanford Natural Language Processing Group: The Stanford parser: A statistical parser (2003), `http://nlp.stanford.edu/software/lex-parser.shtml`