QMUL-ACTIVA: 'Person Runs' detection for the TRECVID Surveillance Event Detection task

Fahad Daniyal and Andrea Cavallaro
Queen Mary University of London
Mile End Road, London E1 4NS (United Kingdom)
{fahad.daniyal,andrea.cavallaro}@eecs.qmul.ac.uk
http://www.eecs.qmul.ac.uk/~andrea/

Abstract: We discuss an event detection approach based on local feature modeling, which is based on spatio-temporal cuboids and perspective normalization (*QMUL-ACTIVA_3 / p-baseline_1*). Motion information is compared against examples of events learned from a training dataset to define a similarity measure. This similarity measure is then analyzed in both space and time to identify frames containing instances of the event of interest (a person running in an airport building). Features are analyzed locally to enable the differentiation of simultaneously occurring events in different portions of an image frame. The performance is quantified on the TRECVID 2010 surveillance event detection dataset.

Description

Event recognition in video has gained much attention in automated video surveillance, content understanding and ranking [1, 2, 3]. Events to be identified are characterized by individual actions of a target or by its interactions with the environment or other targets. A considerable amount of work has been dedicated to event detection in simple datasets (e.g. KTH [4] and Weizmann [5]), whereas more recent works focus on real-world scenarios [1, 6, 7]. These scenarios are often characterized by different level of complexity in the scene due crowdedness, clutter or unfavorable camera placement. Under such constraints several state-of-the-art methods under perform as demonstrated in various trials of [7].

Most event recognition approaches can be divided into two main steps, namely feature extraction and classification. Feature extraction involves the conversion of video data into set of representative features that describe the event we are interested in, whilst the classification steps compares these features against models generated using known

*This work was supported in part by the EU under the FP7 project APIDIS (ICT-216023)

Table 1: Results of the *Person Runs* event detection for the dry-run and the evaluation datasets. (Key. *Ref*: number of ground-truth (GT) events; *Sys*: number of events generated by the proposed method; *TP*: number of correct detections; *FP*: number of false positives; *FN*: number of missed detections, *DCR*: Detection Cost Rate).

	Ref	*Sys*	*TP*	*FP*	*FN*	*Actual DCR*	*Minimum DCR*
Dry run	63	476	34	215	29	0.531	0.307
Dry run v1	63	327	49	134	14	0.290	0.273
Final submission	107	360	36	223	71	0.737	0.628

labeled samples. The proposed approach for event recognition incorporates analysis of event candidates generated by analyzing motion vectors. For feature extraction we employ a 2D macro-block grid on pairs of frames. The resulting motion vectors are then merged into groups using contextual information. Next we perform spatio-temporal analysis to make a final decision. Details about this method can be found in [8].

To quantify the performance of the proposed method we test it on the TRECVID surveillance event detection dataset [7], for the *Person Runs* event. This dataset contains dense scenes, clutter, considerable variations in viewpoint and high variability among different instances of the same action. We will discuss the results on the *dry run* and on the *final submission* datasets.

The *dry run* dataset is composed of 54890 seconds of video @25 fps. We used 2300 instances of people running in different scene locations. The results of this evaluation, as reported by the National Institute of Standards and Technology (NIST), are given in Table 1 (top row). The proposed system had a relatively high number of detections (*Sys.* = 476), where 227 detections had a low confidence and were not included in the computation of the *Actual DCR*. However, when analyzing the results, we noticed that some sample events taken from the ground truth of the training set were not reliable due to the fact that the corresponding motion vectors were not sufficiently clear for triggering a detection. Hence as a second stage we carried out a manual verification for selecting sample events to be included in the training. Moreover the formulation of the proposed method was changed such that events that occurred simultaneously but were spatially separated could be detected as separate events. The results of these improvements (*Dry-run-v1*) are shown in Table 1 (middle row). We can see that by improving the quality of the training data and discriminating simultaneous occurrence of events of interest there was a significant increase of detected events (by 15) and a decrease of the number of false positives (by 81). The scores for *Dry run v1* were generated in-house using the TRECVID Framework for Detection Evaluations (F4DE [1]). In the final stage of testing (*final submission*) we submitted the results of the proposed method on approximately 44 hours of test data. The number of training samples from the training dataset was increased to 4753. The results of this final submission, as reported by NIST, are shown in Table 1 (bottom row). The performance compared to other systems

[1] http://www.itl.nist.gov/iad/mig/tools/

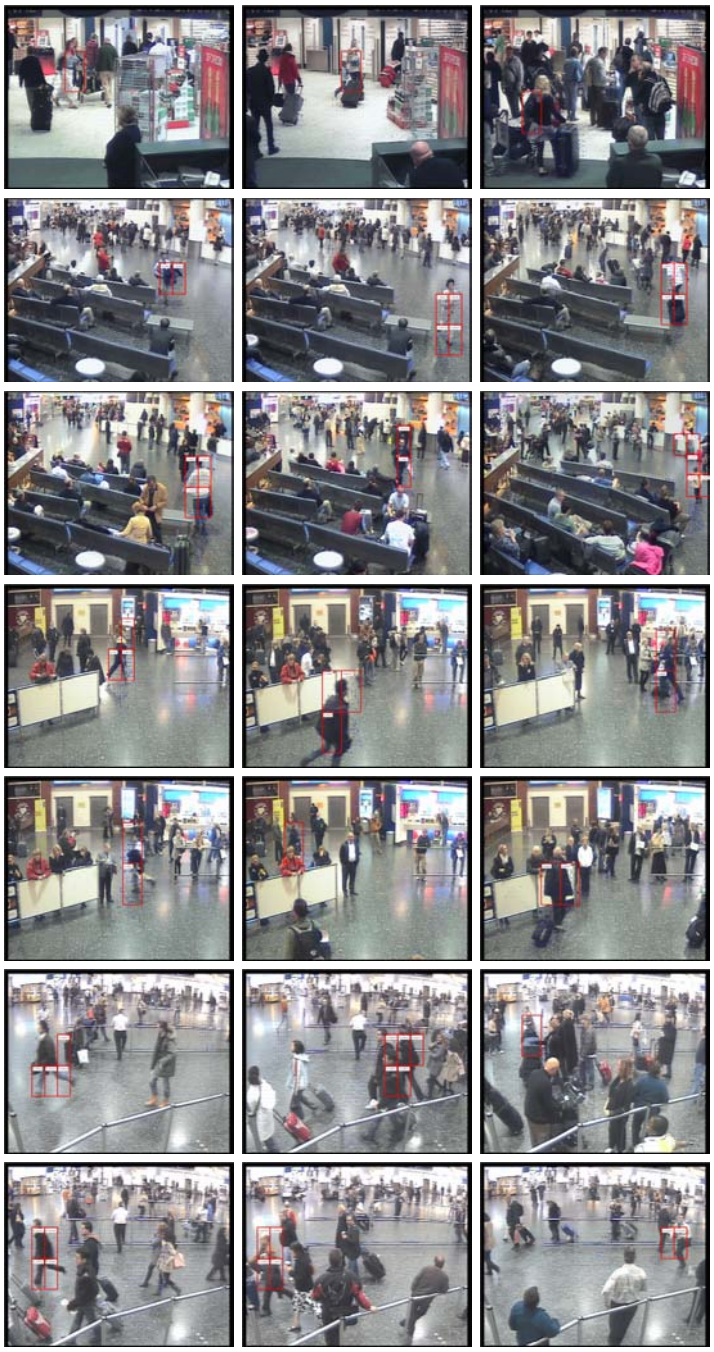

Figure 1: Examples of true positive detections for the *Person Runs* event. Row 1: camera 1; rows $2-3$: camera 2; Rows $4-5$: camera 3; rows $6-7$: camera 5. No *Person Runs* event is present in camera 4.

Table 2: Results of the *Person Runs* event detection for the final evaluation dataset by all the participants in the TRECVID 2010 evaluation campaign as reported by NIST in http://www-nlpir.nist.gov/projects/tvpubs/tv10.slides/tv10.sed.slides.pdf (the lower the DCR, the better the results).

System Id.	*Actual DCR*	*Minimum DCR*
BUPTMCPRL_2010092204 / p-baseline_1	1.614	0.991
CMU_2 / p-VCUBE_2	8.266	0.948
CRIM_1 / p-baseline_1	10.575	1.285
INRIA-WILLOW_3 / p-SYS_2	1.108	1.000
PKU-IDM_5 / p-eSur_3	1.00	0.987
QMUL-ACTIVA_3 / pbaseline_1	0.737	0.681
sfu_15 / p-pbs1_1	1.034	0.981
TJU_2 / p-TJUMM_1	7.173	0.952
TTandGT_1 / p-EVAL_1	1.002	1.002

within the TRECVID 2010 evaluation is highlighted in Table 2.

Examples of detected running people (true positives) are shown in Fig. 1. Using the proposed method we are able to detect some *difficult* events, for example in the first row (middle image) and in row 3 (last image), we are able to detect a running kid, who is partially occluded. In the second and in the third row, we are able to detect the events that happen without changing scale. However, when a person is running away from or towards the camera (Fig. 2 (a)), the magnitude of the motion vectors is not large enough to discriminate a running person from a walking person. Similarly, people that are in the far field and have similar appearance to the background do not generate any significant motion vectors (Fig. 2(b)). Finally, although the proposed method can separate events happening simultaneously, when people are running very close to each other we are at the moment unable to separate the two events (Fig. 3). This limitation can be overcome by performing object detection [9] in the frames of interest.

Summary

We presented a technique for automatically detecting people running in real-world scenarios. The proposed approach is based on motion features and on spatio-temporal modeling. We demonstrated the performance of the proposed approach in the TRECVID 2010 airport dataset with very encouraging results. As future work, we will apply the proposed method to other types of events, such as people jumping, standing up, sitting down or moving in opposite direction to the normal flow.

(a) (b)

Figure 2: Examples of missed events due to an occlusion (a) and to motion direction away from the camera (b).

Figure 3: Examples of simultaneous events that are close to each other and therefore detected as a single event.

References

[1] W. Chiu and D. Tsai. A macro-observation scheme for abnormal event detection in daily-life video sequences. *EURASIP Jrnl. on Adv. in Signal Processing*, 2010(20):20:1–20, Feb. 2010.

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1 –8, Anchorage, AK, USA, Jun. 2008.

[3] F. Daniyal, M. Taj, and A. Cavallaro. Content-aware ranking of video segments. In *Proc. of ACM/IEEE Int. Conf. Distributed Smart Cameras*, pages 1–9, Stanford, CA, USA, Sep. 2008.

[4] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *Proc. of Int. Conf. on Pattern Recognition*, pages 32–36, Cambridge, England, UK, 2004.

[5] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proc. of IEEE Int. Conf. on Computer Vision*, pages 1395–1402, Beijing, China, Oct. 2005.

[6] M. Chen, L. Mummert, P. Pillai, A. Hauptmann, and R. Sukthankar. Exploiting multi-level parallelism for low-latency activity recognition in streaming video. In *Proc. of ACM SIGMM Conf. on Multimedia Systems*, pages 1–12, Phoenix, AZ, USA, Feb. 2010.

[7] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and Trecvid. In *Proc. of ACM Int. Workshop on Multimedia Information Retrieval*, pages 321–330, Santa Barbara, CA, USA, Oct. 2006.

[8] F. Daniyal and A. Cavallaro. Abnormal behavior detection using local motion analysis. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[9] S. Karlsson, M. Taj, and A. Cavallaro. Detection and tracking of humans and faces. *EURASIP Journal on Image and Video Processing*, 2008(1):1–9, Feb. 2008.