CRIM Notebook Paper - TRECVID 2011

Surveillance Event Detection

S. Foucher, *Member, IEEE,* M. Lalonde, *Member, IEEE*, L. Gagnon, *Member, IEEE*

Centre de recherche informatique de Montréal (CRIM)
405, Avenue Ogilvy, bureau 101
Montreal (Quebec) H3N1M3
{Samuel.Foucher, Marc.Lalonde, Langis.Gagnon}@crim.ca

Abstract

Approach we have tested in each of your submitted runs. For the "Object Put" event, we followed a dual foreground segmentation approach where the output difference between a short term and a long term model is used for triggering potential alerts. For Pointing, Embrace, CellToEar and PersonRuns, we applied the learning of compound spatio-temporal features based on a data mining method.

Relative contribution of each component of our approach. Our system is based on an action recognition approach which is mining spatio-temporal corners in order to detect configurations, called Compound Features, typical of an action of interest. The final detection is based on blobs around local frame-to-frame changes that are containing enough relevant compound features.

What we learned about runs/approaches and the research question(s) that motivated them. Overall, performances have improved from last year especially for PersonRuns. In addition, the training for PersonRuns was based on a standard action recognition dataset (KTH) independent of the TrecVid dataset which indicates that our implementation is behaving as expected. For Pointing, Embrace and CellToEar, results are not satisfying yet and the main reason is probably due to the fact that the training dataset derived from the development videos presents a large variability, is too noisy and too small in size in order to produce good rules. Also, given the complexity of the scenes composed of multiple action occurrences, occlusions and complex actions, the direct application of an action recognition method is a challenge. Going forward, performances could be improved if combined with other approaches such as a person tracker and also if the quality of the training set could be improved.

Introduction

This is the second year of participation for CRIM to the SED task and this time we provided results on five events by adding the Embrace and CellToEar ones. For this year, we focused primarily on improving the action recognition algorithm that used to be at the heart of our Pointing method in 2010 but was adapted this year for PersonRuns, Pointing, Embrace and CellToEar. The size of the training sets was significantly increased and was supplemented by the standard action recognition dataset KTH [4].

All the computations were performed on the "Mammouth" supercomputer located at the Center for Scientific Computing at the Université de Sherbrooke.

I – Object Put Event

The "Object Put" algorithm was unchanged from last year and is based on a very simple dual background model approach described in [1]. This year we were aiming at optimizing the detection threshold and the learning rate values for the short and long term models. The performance on Eval08 after parameter tuning is shown in Figure 1. However, this new set of parameters didn't lead to better performances compared to last year (see Section III).

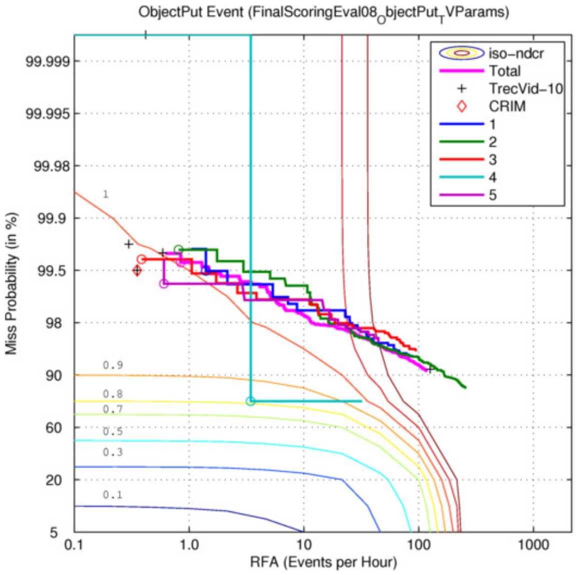


Figure 1: DET curve on Eval08 for ObjectPut.

II – Action Recognition

We pursued last year's implementation of an action recognition approach based on the hierarchical learning of Compound Features (CF) proposed recently by Gilbert *et al*. [2][3]. This method was then applied to PersonRuns, CellToEar, Pointing and Embrace. The following steps are involved:

1. Build an overcomplete set of Harris corners at various spatial scale and in the temporal domain.
2. Group corners within a 3x3x3 neighbourhood to form CF
3. CF are then encoded using information about cell position, scale and corner type to form transactions (or itemsets).
4. A data mining algorithm (APriori algorithm [5]) is applied in order to extract frequent itemsets from all the recorded level 1 transactions observed on the training set.
5. Transaction rules and associated confidence levels are derived from the frequent itemsets.
6. Rules that have fired at level 1 are used to form CF within a 6x6x6 neighbourhood at level 2.
7. The same data mining algorithm (step 4) is applied on level 2 transactions.
8. A third level of CF is formed but with a 2x2x2 spatio-temporal neighbourhood with cells extending to the limit of the image.

The last level (step 8) implicitly assumes that only one instance of a particular action is taking place at the same time which is of course rarely the case in the TrecVid data. Therefore, we slightly modified the last stage so that the spatial extent of the neighbourhood takes into account the expected size of a person given a position in the image and is derived from our camera geometric model (see [1] for details).

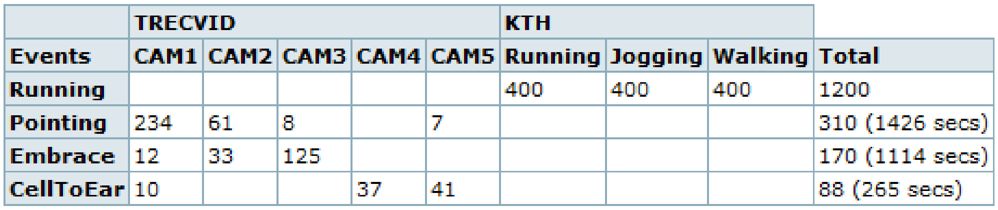
Training

Compared to last year, we greatly increased the number of training samples. In addition to samples taken from the evaluation set we added videos from the KTH dataset [4].

For Pointing, Embrace and CellToEar, we manually annotated some events in the TrecVid development set with the following guideline:

1. No action mixing (e.g. no walking and pointing).
2. No occlusions
3. The bounding box should encompass the total spatial extent of the event.
4. One single occurrence of the event during the event timeframe.

The resulting number of training samples is shown in Table 1. The negative examples were taken from the same TrecVid videos but outside the event bounding box. We also added training samples for which the video frame has been flipped horizontally; otherwise learned rules may be biased by the dominant people motion (for instance, most people move from left to right in Camera 1). So the total number of positive events for Pointing is 620 with a majority of events taken from Camera 1. For PersonRuns, because it was too difficult to build a ground truth on the TrecVid videos, we chose instead to take the KTH videos for the Running/Jogging actions as positive samples (25 persons running 4 times in 4 videos) and the Walking action as negative.



	TRECVID					KTH			
Events	**CAM1**	**CAM2**	**CAM3**	**CAM4**	**CAM5**	**Running**	**Jogging**	**Walking**	**Total**
Running						400	400	400	1200
Pointing	234	61	8		7				310 (1426 secs)
Embrace	12	33	125						170 (1114 secs)
CellToEar	10			37	41				88 (265 secs)

Table 1: Size of the training set for each event.

The resulting number of transactions at level 1 went from 1 million last year to about 30 millions for this year. The data mining algorithm (Apriori [5]) was run so that we are looking for rules with a minimum support of 5% and a minimum confidence level equal to the fraction of positive transactions if greater than 20%. The minimum support threshold is necessary in order to make sure that the derived rules are statistically significant. We also limited the rule size to 5 as recommended by [2]. For PersonRuns, we get 3886 rules at level 1, 3599 at level 2 and 147848 rules at level 3.

Detection

Another important issue is how to take a reliable decision in the presence of an event in a scene where many other actions are taking place (e.g. people walking). The original method derives a probability map from the firing rules as shown in Figure 2, in order to improve the map, we applied a Gaussian and temporal filtering; however this map is still difficult to threshold.

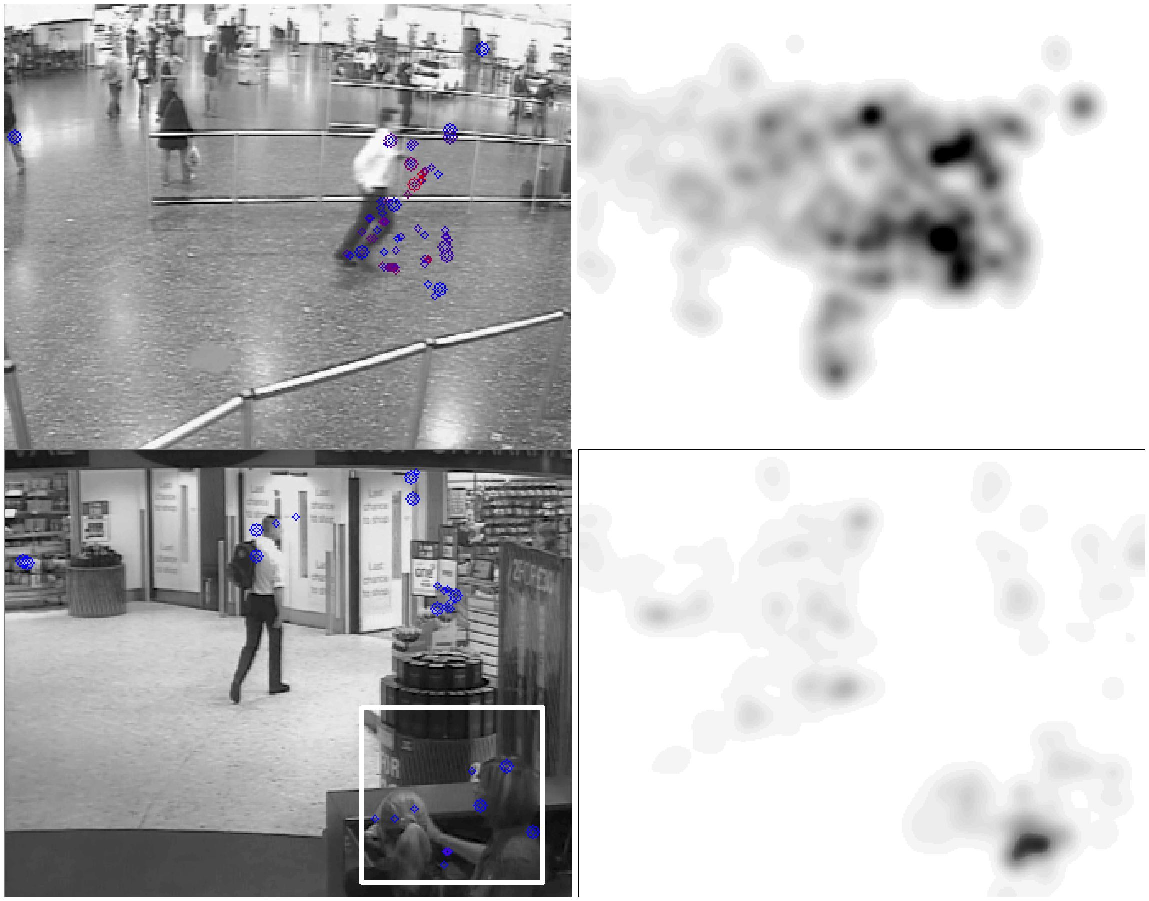


Figure 2: Rules firing for PersonRuns (top left) and corresponding probability map (top right, dark means high probability value). Rules firing for Pointing (bottom left) and corresponding probability map (top right, dark means high probability value).

Therefore, we adopted a simpler approach where we segment the frame-to-frame difference image leading to areas of motions (see Figure 3). We then compute the density of rules firing within each region of changes. The region with the higher rule density value is chosen and triggers an event if the average confidence value within this region is above a given threshold.

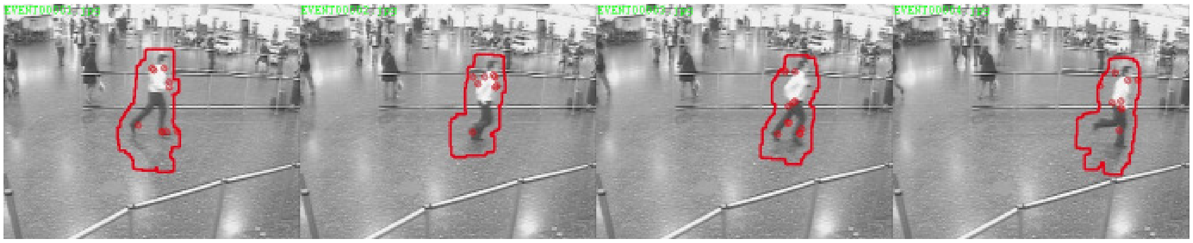


Figure 3: Action detection for PersonRuns based on a blob from a change detection.

III - Results

Results on the Development Set (Eval08)

We evaluated the performance on the Eval08 dataset (25 videos). We can see that the performance for PersonRuns (Figure 4 - left) was significantly increased when compared to last year's result on Eval09 with a global Min NDCR at 0.958 (PMiss=0.9221, RFA= 7.13). For Pointing (Figure 4 - right), the performance also increased notably with a shift of the DET curve to the left by one order of magnitude. For Embrace (Figure 5 - left), results are still one order of magnitude larger than the other teams in terms of RFA except maybe for Camera 5. Performance for CellToEar (Figure 5 - right) is similar to Embrace and close to performance levels reached by other teams last year.

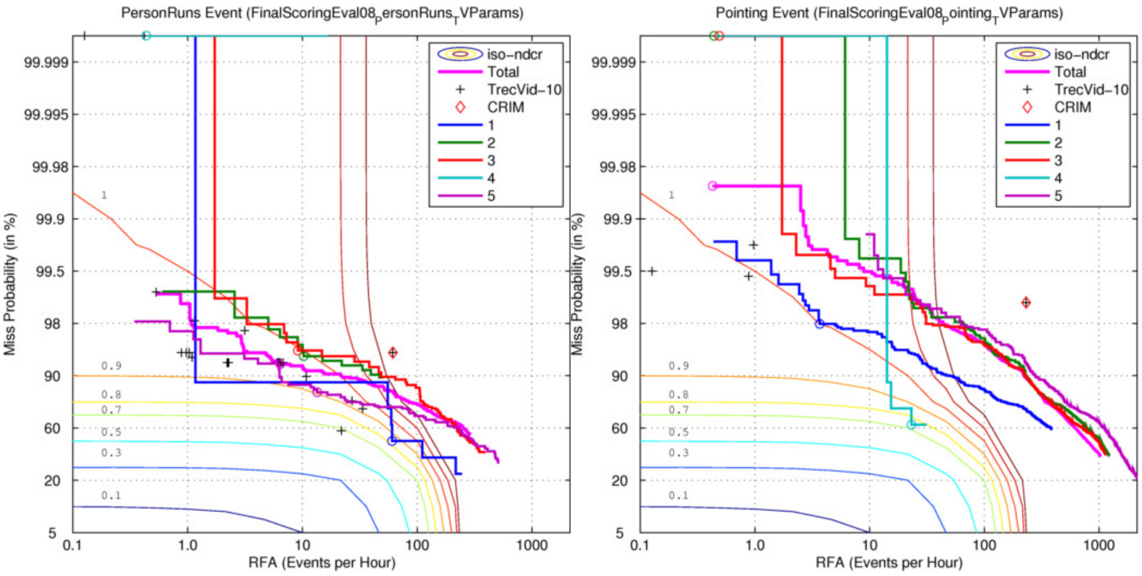


Figure 4: DET curves for PersonRuns and Pointing on the Eval08 dataset. Last year Min NDCR position for CRIM is shown as a red diamond marker.

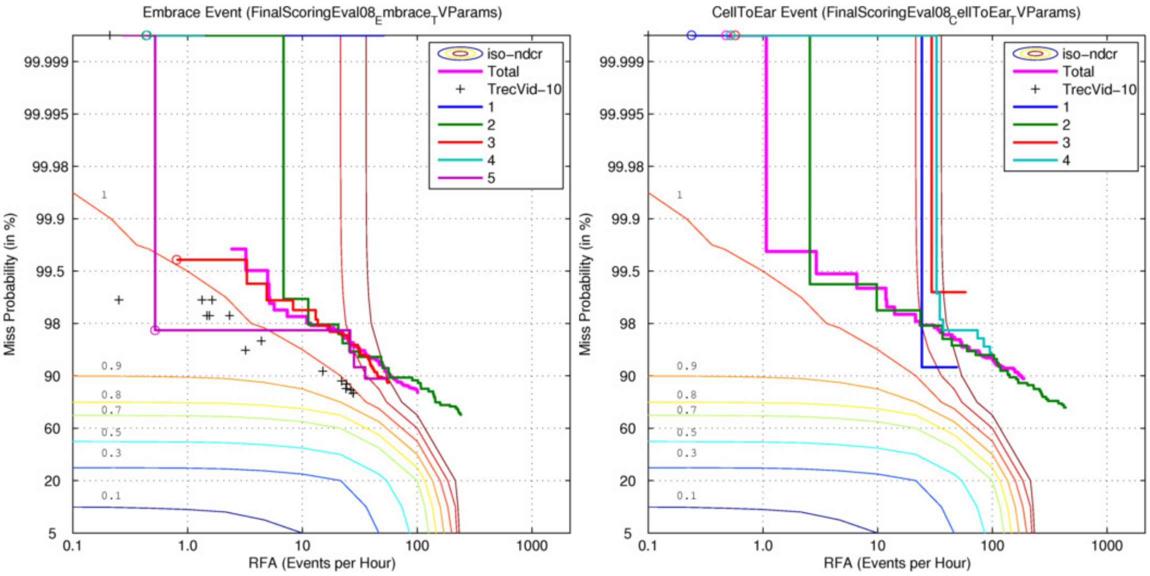


Figure 5: DET curves for Embrace (left) and CellToEar (right) on the Eval08 dataset. Last year Min NDCR position for CRIM is shown as a red diamond marker. Black crosses indicate Min NDCR positions on Eval09 by the other teams during the TrecVid-2010 competition.

Results on the Test Set (Eval09)

Results provided by NIST (see Table 2 and Figure 6 below) on Eval09 are consistent with what we observed on Eval08, the best Min NDCR was obtained for PersonRuns. However, the level of false alarms is higher for Pointing compared to Embrace and CellToEar despite the fact that it is based on a larger training set.

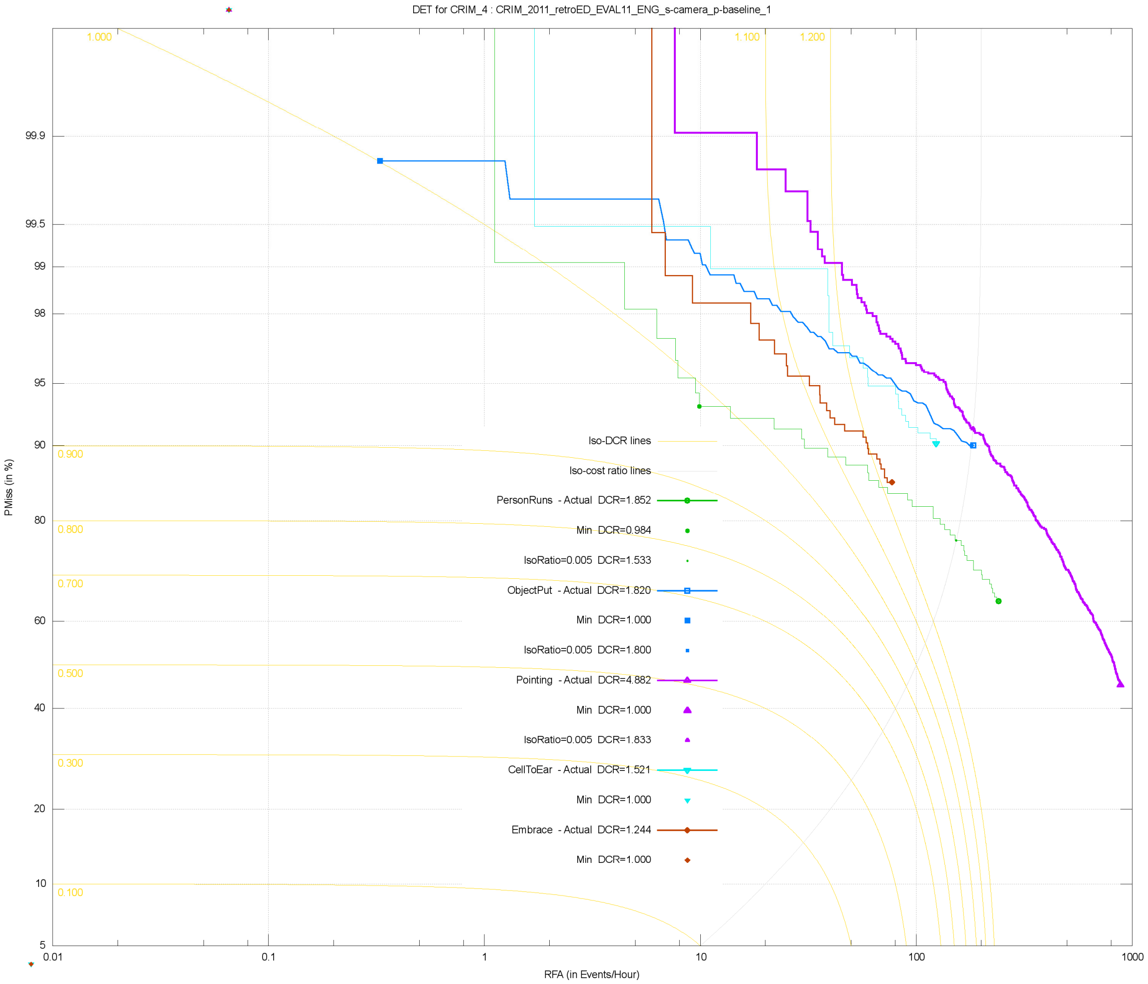
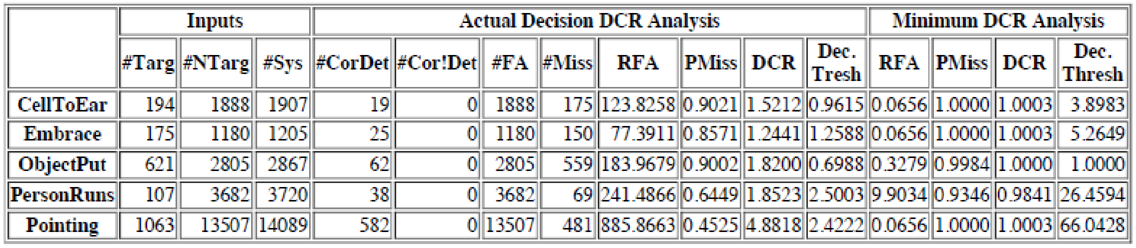


Figure 6: DET curves for PersonRuns and Pointing on the Eval09 dataset.



	Inputs			Actual Decision DCR Analysis								Minimum DCR Analysis			
	#Targ	#NTarg	#Sys	#CorDet	#Cor!Det	#FA	#Miss	RFA	PMiss	DCR	Dec. Tresh	RFA	PMiss	DCR	Dec. Thresh
CellToEar	194	1888	1907	19	0	1888	175	123.8258	0.9021	1.5212	0.9615	0.0656	1.0000	1.0003	3.8983
Embrace	175	1180	1205	25	0	1180	150	77.3911	0.8571	1.2441	1.2588	0.0656	1.0000	1.0003	5.2649
ObjectPut	621	2805	2867	62	0	2805	559	183.9679	0.9002	1.8200	0.6988	0.3279	0.9984	1.0000	1.0000
PersonRuns	107	3682	3720	38	0	3682	69	241.4866	0.6449	1.8523	2.5003	9.9034	0.9346	0.9841	26.4594
Pointing	1063	13507	14089	582	0	13507	481	885.8663	0.4525	4.8818	2.4222	0.0656	1.0000	1.0003	66.0428

Table 2: Actual Miss rate and False Alarm rates on Eval09 for each event.

Conclusions

For our second year in this SED evaluation campaign, the objective was to improve results from last year especially regarding the level of false alarms. We also provided results for two more events (CellToEar and Embrace). Results from the action recognition method are mixed due to the difficulty to form a clean training set from the TrecVid videos. Results on PersonRuns are promising but the detection step based on blobs is probably not optimal in all situations (e.g. someone running among a walking crowd).

The original action recognition method was not designed for a scene with multiple occurrences of the same action (e.g. several people walking). In particular, the last step of the method is problematic as it looks for rules in the entire image. We modified this step for a local approach with a neighborhood function of the camera geometric model. Results on the various cameras show that the performance is affected by the level of clutter and the object resolution in the scene. Cameras 1 and 4 with lesser depth of view have generally better results.

Increasing the training database will likely require the use of a 64-bit version of the APriori algorithm as we have reached the memory limit under windows 32-bit (~ 2 Gb in memory). However, the main difficulty for the training is to build a "clean" database of action units. The few samples taken from the evaluation corpus usually exhibit large variations in pose, background clutter and are usually composed of a mixture of actions (e.g. person walking and pointing). An idea could be to supplement the TrecVid samples with an in-house database in order to reinforce relevant action patterns.

The detection process needs to be refined also. We have adopted a strategy different from Gilbert *et al.* but that is probably not optimal in all situations. The current version runs on smaller frame size at about 4-5 fps, this loss of resolution does not help the detection process especially for actions in the background.

Acknowledgments

This work has been supported by MDEIE of the "Gouvernement du Québec". The authors wish to thank the RQCHP (Compute Canada) for giving access and support to their high performance computing resources.

References

[1] S. Foucher, M. Lalonde, L. Gagnon, "CRIM Notebook Paper - TRECVID 2010 Surveillance Event Detection", In NIST TREC Video Retrieval Evaluation Workshop 2010 (TRECVID 2010). Gaithesburg, MD, November 15-16, 2010.

[2] A Gilbert, J. Illingworth, R.Bowden, "Action Recognition using Mined Hierarchical Compound Features", Accepted for IEEE Trans Pattern Analysis and Machine Learning. 2010.

[3] A. Gilbert , J. Illingworth, R. Bowden, "Fast Realistic Multi-Action Recognition using Mined Dense Spatio-temporal Features", In Proc. Int. Conference Computer Vision (ICCV09), Kyoto, Japan.

[4] C. Schuldt, I. Laptev, B. Caputo, "Recognizing Human Actions: a Local SVM Approach," In Proc. of International Conference on Pattern Recognition (ICPR'04), vol. III, pp. 32–36, 2004.

[5] Christian Borgelt, "Recursion Pruning for the Apriori Algorithm", 2nd Workshop of Frequent Item Set Mining Implementations (FIMI 2004, Brighton, UK).