SAVASA Project @ TRECVid 2013: Semantic Indexing and Interactive Surveillance Event Detection

Suzanne Little*[1], Iveel Jargalsaikhan[1], Rami Albatal[1], Cem Direkoglu[1], Noel E. O'Connor[1], Alan F. Smeaton[1], Kathy Clawson[2], Min Jing[2], Bryan Scotney[2], Hui Wang[2], Jun Li[2], Marcos Nieto[3], Juan Diego Ortega[3], Aitor Rodriguez[4], Iñigo Aramburu[4], Emmanouil Kafetzakis[5]

1. Insight Centre for Data Analytics, Dublin City University, Ireland
2. University of Ulster, United Kingdom
3. Vicomtech-IK4, Spain
4. IKUSI, Spain
5. NSCRD "Demokritos", Greece

Abstract

In this paper we describe our participation in the semantic indexing (SIN) and interactive surveillance event detection (SED) tasks at TRECVid 2013 [11]. Our work was motivated by the goals of the EU SAVASA project (Standards-based Approach to Video Archive Search and Analysis) which supports search over multiple video archives. Our aims were: to assess a standard object detection methodology (SIN); evaluate contrasting runs in automatic event detection (SED) and deploy a distributed, cloud-based search interface for the interactive component of the SED task. Results from the SIN task, underlying retrospective classifiers for the surveillance event detection and a discussion of the contrasting aims of the SAVASA user interface compared with the TRECVid task requirements are presented.

1 Introduction

The DCU-SAVASA team comprises researchers and developers from five different institutions – Dublin City University (Ireland), University of Ulster (United Kingdom), Vicomtech-IK4 (Spain), IKUSI (Spain) and NCSRD "Demokritos" (Greece). The team partipated in two tasks – Semantic Indexing (SIN) and interactive Surveillance Event Detection (iSED) – in the 2013 TRECVid benchmarking [11, 14]. This was our first time participating in SIN, motivated by object tracking requirements of the SAVASA project, and the second time for SED following on from last year's efforts [8].

Our work is motivated by the goals of the EU FP7 SAVASA project[1] (Standards-based Approach to Video Archive Search and Analysis), where (among other tasks) we contribute to the semantic annotation of CCTV footage, including person detection, object detection and tracking, semantic annotation and search. The SAVASA project aims to develop a standards-based video archive search platform that allows authorised users to query over various remote and non-interoperable video archives of CCTV footage from geographically diverse locations. At the core of the search interface is the application of algorithms for person/object detection and tracking, activity detection and scenario recognition. The project also includes research into interoperable standards for surveillance video, discussion of the legal, ethical and privacy issues and how to effectively leverage cloud computing infrastructures in these applications. Project

*Contact author: suzanne.little@dcu.ie

[1] http://www.savasa.eu

partners come from a number of different European countries and include technical and research institutions as well as end user, security and legal partners.

Our experiments consisted of two runs submitted in the SIN task to evaluate the performance of standard state-of-the-art object detection methods and the effect of extreme parameter values (section 2); three user interactive runs in the SED task (section 3.1) and two sets of runs – four from Dublin City University (DCU) (section 3.2.2) and three from the University of Ulster (UU) (section 3.2.3) – in the retrospective SED task.

2 Semantic Indexing (SIN)

In the Semantic Indexing task, we submitted three runs for main task and one run for each progress task (2014, 2015). The goal is to apply a state-of-the-art classification method (SVM with RBF-Eucludian distance kernel) on descriptors of different dimentionalities in order to evaluate the effect of varying the SVM RBF $Gamma$ parameter on the classification quality. Two descriptors were used in the submitted runs, the first is a global descriptors of 104 dimentions, the second is a Bag-of-Visual-Word descriptor of 1,000 dimensions, a third submission corresponding a late fusion of both descriptors results was also submitted. The parameter value space was explored carefully for each descriptor in the validation step, and extreme low values has led to better results for the small descriptor in opposite of the expected over-fitting.

2.1 Visual descriptors

We submitted runs based on two descriptors produced and shared by IRIM[2] research network:

- hg104: a global descriptor of 104 dimensions, corresponds to the early fusion of a normalized Gabor transform descriptor (40 dimensions), and a normalized RGB histogram (64 dimensions).
- opp_sift_dense_1000: a Bag-of-Visual-Word descriptor of 1,000 dimensions, based on dense sampling of interest points and opponent sift feature, generated using ColorDescriptor Software[17]. An opponent SIFT feature of 384 dimensions is extracted from each interest point, then a k-means clustering is applyed on a set of 535,117 opponent SIFT vector from randomly selected keyframes.

2.2 SVM-RBF Parameter tuning

In our submitted runs, we adopt a $gamma$ optimisation scheme inspired from [15] and [13] based on the calculation of distances between the descriptors at 0.9 and 0.1 quantile of all the distances and then compute the average of these two distances (denoted as $meanDist$) and apply it in the next optimisation formula:

$$\gamma = \frac{2^i}{meanDist^2} \tag{1}$$

with i is a positive integer parameter, fixed as 1 or 2 in the case of descriptors with large dimensionality, and 3 or 4 for descriptors with small dimensionality (up to few hundreds).

2.3 Results

Figure 1 shows the official TRECVid results of our two runs using opp_sift_dense_1000 and hg104 descriptors. The Bag-of-Visual-Word descriptor produced better results, but the optimisation of $gamma$ succeeded to enhance the results reported from hg104. Basing on our validation, for some concepts like Boy, Bus, Classroom, Military Airplane using large $gamma$ values can significantly improve the results obtained by applying values from conventional ranges (between 0 and 30). The average Inferred average precision obtained from opp_sift_dense_1000 is 0.1165, and from hg104 0.0722, and the fusion gives 0.1320.

[2] Indexation et Recherche d'Information Multimédia project of GDR-ISIS research network from CNRS-France.

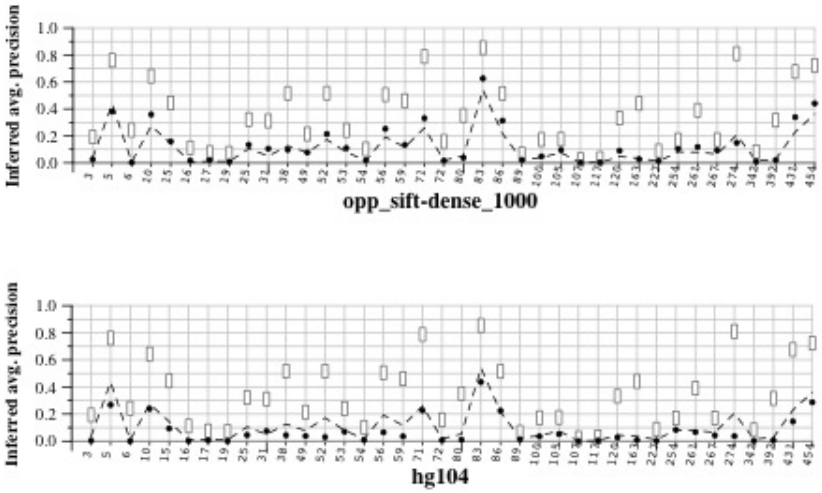


Figure 1: SIN results for opp_sift_dense_1000 and hg104

2.4 Discussion

Contrary to generally accepted practise that a small kernel size is undesirable, the preliminary results obtained from the experiments and the validation show that large values of the RBF γ parameter does not necessarily appear to lead to over-fitting. These results are clearly preliminary. However in order to better understand our results, we are planning to perform further analysis on:

- The descriptors in the dataset (distances between the descriptors in the original feature space as well as in the high RBF dimensional space).
- The relations between the positive/negative examples in the dataset and the γ values.
- The relation between the dimensionality of the descriptors and the γ values.

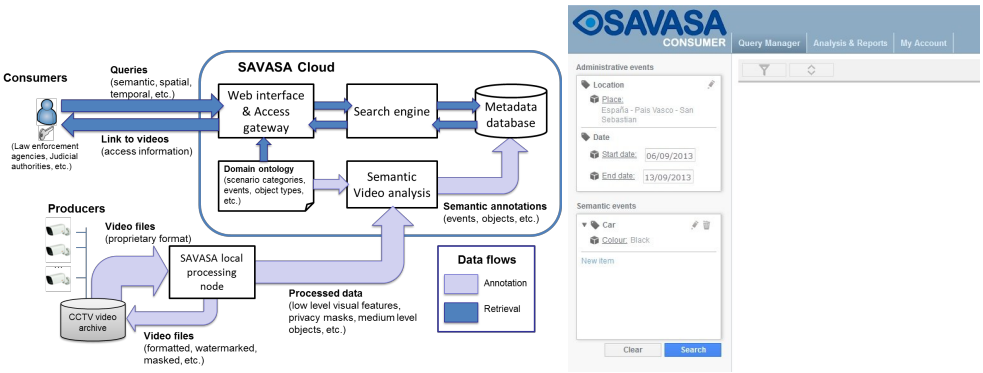
3 Interactive Surveillance Event Detection (iSED)

3.1 User Interactive Search

3.1.1 SAVASA framework and interface

The second year of the interactive element of the surveillance event detection task gave us the opportunity to trial a prototype of the savasa web-based interface running within a virtual cloud enviroment. As for last year's submission [9] the back end of the search interface was populated with the output from the retrospective runs, each acting as a separate archive for users to query. Four events – CellToEar, ObjectPut, PersonRuns or Pointing – were processed for three cameras – 1, 3 and 5. Three runs were submitted to address four objectives: gather user feedback on the search interface (*search1/search2*), stress testing of framework (*search1/search2*), value of person tracking particularly for person runs (*filtered*), use of a priori region of interest labelling to filter or sort results (*filtered*). Two runs were produced using the automatic semantic annotations (dcu-run1, uu-run1, uu-run2) based on users with more computer science or computer vision experience (*search1*), including some who had contributed to the region of interest labelling exercise, and those with less direct experience (*search2*).

Figure 2 shows the high-level SAVASA framework and a screenshot of the interface design. For the purposes of TRECVid the interface was altered to search a single location (gatwick airport) with fixed dates and a reduced set of actions to match the labelled events. The annotations produced by DCU and Ulster



(a) High-level overview of SAVASA framework (b) Screenshot of SAVASA web interface

Figure 2: SAVASA search system

were converted into RDF and stored in a Sesame database with an OWLIM front-end. The simple search engine executes SPARQL queries over the database to return an ordered list of video, startframe, endframe, confidence and source to be displayed by the search interface. To assist in the TRECVid interactive task simple animated GIFs were provided for each result. A challenge that needed to be overcome was the time required to dynamically generate and download large numbers of the GIFs. The size was reduced and a simple caching system implemented to improve the responsiveness of the search. For security and performance reasons the SAVASA framework is designed to be installed within a cloud system using a virtual private network and virtual machines. Therefore the search system and database were executed as independent services communicating via HTTP REST queries on a Linux virtual machine while the SAVASA administration and search system ran on a separate Windows virtual machine within the same private network managed by partners from NCSRD "Demokritos".

To conduct the evaluations, the user first joined the VPN then logged into the search interface using a web browser. Each user was shown how to use the system and provided with examples of the specific event to search for. They were then given up to 25 minutes to find matching video segments. Additional functionality was implemented in the interface by IKUSI to record and save lists of time-stamped results selected by the user.

Two additional sources of data were available for searching. As part of the local feature grouping approach (section 3.2.2) the region of interest for CellToEar, ObjectPut and Pointing was manually annotated for four hours of training video using the VATIC interface [19] running as an offline service. This was used to generate probability matrices based on the frequency of a pixel being part of the region for each event. Partners from Vicomtech also applied a person tracking method (described in section 3.2.1) to produce person detecting and tracking outputs for video segments from cameras 1, 3 and 5 in the test dataset. This raw data was also loaded into the RDF data store and queries provided that returned video segments based on the average amount of absolute movement of a region identified as a person over the duration they were tracked (person motion) and the *a priori* average probability of the pixels in the person region for each event (event probability). Filtering based on camera was also enabled. The experimental filtered search was produced using a simple interface that allowed the user to adjust the underlying SPARQL queries and change the level of person motion, filter based on the event probability and camera. The underlying theory was that person tracking could be used as a reliable filter to find PersonRuns events (high person motion) and that activities such as Pointing and ObjectPut had shown fairly strong connections with particular areas in the frames and, by filtering the video segments to prioritise those where people were present in the frame regions, higher numbers of results could be quickly found.

Table 1: Summary of results for interactive runs

Interactive		search1			search2			filtered		
Event	**#Targ**	**#Sys**	**#Cor**	**DCR**	**#Sys**	**#Cor**	**DCR**	**#Sys**	**#Cor**	**DCR**
CellToEar	194	10	0	1.0033	15	0	1.0049	30	2	0.9989
ObjectPut	621	7	0	1.0023	28	1	1.0072	19	1	1.0043
PersonRuns	107	3	0	1.0010	3	0	1.0010	1	0	1.0003
Pointing	1063	7	2	0.9983	18	6	0.9983	5	2	0.9991

3.1.2 Results and Discussion

The use cases for the SAVASA project are focussed around helping users to quickly find specific video footage from multiple locations based on information needs defined by location and time-date or semantic labels based around concepts such as person count or interaction, object (particularly vehicles) or semantic activities. For example, retrieve all footage showing the north entrance of the station between 2 and 3pm on October 4th where 3 or more people entered together. This reflects the senario where security personnel are responding to an incident with time-critical constraints and is closer to "known item". search. The screenshot of the SAVASA interface in figure 2b shows how location and date-time can be specified to help the user retrieve specific video.

In contrast to the use cases of SAVASA, the aim of the interactive SED task is to find as many examples of a particular action as possible in a given time frame. This is primarily an annotation task where a very successful interface would most likely provide assistance to the user to rapidly view the video segments most likely to contain the event of interest and quickly, accurately label the start and end frames. This mis-match between SAVASA and the TRECVid SED task may lead to sub-optimal performance in the task but the two are close enough that we should be able to learn a lot from participation.

Table 1 summarises the results of the three interactive runs and shows surprisingly high numbers of incorrect detections and poor performance. Even assuming that users made mistakes in identifying the four actions, this inaccuracy is very high. We speculate that this is in part due to the methodolgy for creating the video segments based on the automatic runs. Unless the start and end frames fall close to the precise values provided by the groundtruth, a search result found by the interface and verified by the user will be evaluated as incorrect. However for the purposes of helping a user to find matching video for an action this is overly exact.

The first two objectives for SAVASA in the interactive SED were to gather user feedback on the search interface and to stress test the framework. Direct and observed feedback from the users was generally positive and users were able to construct and execute queries with few difficulties. Negative feedback was mostly around the time required to download the large numbers of animated GIFs illustrating each result – which is not a requirement for the SAVASA use cases. Some bugs in the session and saving processes and issues with maintaining connections were also identified for future improvement.

In contrast to last year when thresholds for annotation were deliberately lowered to promote recall over precision, this year the annotations provided for the database were configured to achieve smaller numbers of results that were more likely to be correct. This was based on feedback from expert users last year that unexpectedly favoured accuracy over recall even in security applications [8]. Although the number of results found this year was lower than last year, users were more positive about the smaller numbers of false alarms.

The final two objectives were to use the person tracking and region of interest labelling to filter the video. The use of person tracking for PersonRun evens was clearly unsuccessful given 1 ("incorrect") segment was found. This appears to be due to the parameter assumptions required to achieve fast, accurate person tracking. PersonRuns events contain people who are moving too fast to be accurately tracked over the required minimum number of frames or are examples of a child running who is considered too short (after camera perspective adjustments) to be correctly labelled as a person. While the evaluation results (DCR) for the filtered search were not significantly different from the other searches, there is some indication that a filtered approach may be more successful in addressing the specific aims of the TRECVid interactive SED task as a larger number of matching segments were identified by the user.

3.2 Automatic Event or Action Detection

3.2.1 Person Tracking

Applying contextual information to visual object detection and tracking has shown to be of great help for increasing the accuracy and quality of the results. In this work we have focused on exploiting the perspective information of the scene to dramatically reduce the computational cost typically assocaited with traditional detection-by-classification approaches [22]. Our scheme automatizes the sliding window technique [18] by computing the optimum value of its parameters, which results on time reductions between 30% to 80% depending on the perspective while keeping similar detection rates. A fuller discussion of the person tracking method can be found in Nieto et al. [10].

3.2.2 Action Recogntion Using Local Feature Grouping

The action recognition framework is proposed based on local feature groupping. The framework consists of feature extractionand feature representation followed by classification steps. The main contribution is our feature representation method that applies clustering on the local features to group them based on spatio-temporal proximity. Each group is further described by the *Bag-of-Features (BOF) approach* and used to train an SVM classifier.

Low-level feature descriptors

We adopted Heng *et al.*'s [20] approach to extract low-level features from video data. In brief, the feature points sampled on a grid and tracked multiple scales separately to generate a set of dense trajectories. For each trajectory, four descriptors are calculated to capture the different aspects of motion trajectory. Among the existing descriptors, HOGHOF [6] has shown to give excellent results on a variety of datasets [21]. HOGHOF is calculated along the trajectories. HOG (histograms of oriented gradient) [2] captures the local appearance around the trajectories whereas HOF (histograms of optical flow) captures the local motion. Additionally, MBH (motion boundary histogram), proposed by Dalal *et al.* [3], and TD (trajectory descriptor) [20] are computed in order to represent the relative motion and trajectory shape.

Action Representation

A video segment can be understood as a cloud of local features in 3D space. In traditional *Bag-of-Features (BOF) approaches*, all local features contribute equally to represent a semantic content within the video segment. The main drawback is that noisy or unnecessary local features are being added in generating a co-occurrence histogram vector. For example, in the scenario of a pointing action, the number of local features extracted at the region where the action being performed is insignificant compared to the total number of local interest points in the video segment.

We extend the *BOF representation* by exploiting the extracted features location in time and space. The intuition is that closely localized features are more likely to correspond to a same object, and far ones are more unlikely. We apply a tree cluster to group local features based on their spatio-temporal proximity. Similar to *image pyramid*, the number of groups vary at different scales. In our experiment, we set the total scale $S = 5$ and the total number of group at the scale s is $G_s = 2^{s-1}$. Finally, all groups are separately represented by *BOF approach* that will produce in total $N = \sum_s 2^{s-1}$ *local feature group* histogram vectors to be used in training a classifier, where $s = \{1, 2, ..., S\}$.

In order to build a visual dictionary, we cluster a subset of 250,000 descriptors sampled from the training videos using k-means algorithm for each descriptor. The number of clusters is set to $k = 4000$, which has shown empirically to give good results in [6].

Table 2: Summary of results for retrospective runs

DCU		run1			run2			run3			run4		
Event	**#Targ**	**#Sys**	**#Cor**	**DCR**	**#Sys**	**#Cor**	**DCR**	**#Sys**	**#Cor**	**DCR**	**#Sys**	**#Cor**	**DCR**
CellToEar	194	21	0	1.007	51	0	1.017	80	0	1.026	146	3	1.031
ObjectPut	621	202	11	1.045	334	17	1.077	357	16	1.086	485	21	1.118
Pointing	1063	51	10	1.004	111	15	1.017	132	17	1.022	454	26	1.116

Classification

For classification, we used a non-linear support vector machine (SVM) with a Radial Basis Function (RBF) kernel. In order to represent the video frame, we utilized a temporal sliding window approach. In the experiments, we set the window size $W = 25$ frames and sliding step size $L = 10$ frames.

Each sliding window is described by N *local feature groups* according the technique described in the previous section. Since the TRECVid Dataset contains only the temporal duration of events, we manually added spatial information for 4-hours video footage by assigning bounding boxes at each frame. Thus each event is annotated by temporal bounding boxes. This information is used to label *local feature groups* separately. Regarding the prediction, we assign an event class to the video frame that has a highest vote from *local features group* belonging to its sliding window.

Results and Discussion

We submitted four runs to evaluate framework based on the local feature grouping with following parameter values:

run1 t_{min} = 1.2s, t_{max} = 3.2s, S = 4,

run2 t_{min} = 0.8s, t_{max} = 3.2s, S = 4,

run3 t_{min} = 1.2s, t_{max} = 3.2s, S = 3 and

run4 t_{min} = 1.2s, t_{max} = 3.2s (baseline: no feature grouping is performed)

where t_{min}, t_{max} is the minimum and maximum duration of event and S is the total number of scale where the feature grouping is performed. As shown in Table 2, 'run1' achieved the highest performance. It is observed the S parameter's value has a significant effect on the performance as 'run3' ,where S = 3, achieved higher DCR value compared to 'run1' and 'run2'. The 'run4' is baseline case where local features grouping is not applied and performed the worst compared to the rest. The evaluation result shows that a video segmentation based on the local feature grouping improves the performance significantly.

3.2.3 Action Recognition Using Fused Optical Flow and Moment Features

We propose a pipeline for human action recognition which incorporates: person detection; person region description using optical flow features; feature space representation via either Principal Components Analysis (PCA) or Nonnegative Matrix Factorization (NMF); state-based event classification via hidden Markov models (HMM); and classification likelihood update incorporating a priori information. Person detection is performed using the methods described in Section 3.2.1. The remainder of our pipeline is described below.

Feature Description

Low level features characterize KLT optical flow vector properties of each detected person's region of interest (ROI). After per ROI optical flow calculation, we derive features which capture motion orientation, magnitude and relative location. In addition to generating HOOF features (90 orientation bins), we follow the method of Efros [4] and view temporal flow displacements as global spatial patterns. Each ROI's

smoothed optical vector is decomposed into a number of half-wave rectified, directed channels. Specifically, the optical flow vector field F is split into two scalar fields, Fx and Fy, corresponding to the horizontal and vertical components of the flow. Fx and Fy are half-wave rectified to form 4 non-negative channels: $Fx+$, $Fx-$, $Fy+$, and $Fy-$, and blurred with a Gaussian ($sigma = 0.5$) to remove spurious motions. Finally, 2D Zernike moments of each of the new channels are calculated and concatenated, resulting in a $4n$ * 1 feature vector per frame, where n is the number of moment features per channel and $n = 6$. We append this feature vector to the HOOF feature vector, resulting in a ($4n + numerHOOFBins$) * 1 feature vector per person per frame. The complete set of ROI-based features corresponding to a single individual is regarded as a time varying sequence and used as inputs for feature representation and classification.

Feature Space Representation

PCA and NMF are investigated as alternative methods for feature space representation. PCA constitutes a common technique for data representation and reduction, which utilizes eigenvalue decomposition to generate a set of linearly uncorrelated bases ranked in descending order of variance. For PCA, the number of dimensions is 16. Unlike PCA, NMF [7] decomposes the nonnegative features into part based components, which aims to improve the low level feature presentation. Recent studies have shown positive outcomes in application of NMF to human pose recognition [16] and action recognition [12]. In the NMF model, a data matrix $\mathbf{X}$ is decomposed as the product of two matrices containing only nonnegative elements, $\mathbf{X} = \mathbf{AH}$. The optimal solution can be obtained by minimizing the distance $\|\mathbf{X} - \mathbf{AH}\|^2$ subject to the constraint $\mathbf{A}, \mathbf{H} > 0$. In this application, $\mathbf{X}$ is the fused HOOF and Moment features obtained from video sequences. The columns of $\mathbf{A}$ represents the basis vectors and $\mathbf{H}$ denotes the corresponding coefficients which are used for classification. A regularized Fixed-Point based NMF algorithm [1] was applied, the dimension of basis was empirically set as 10.

Event Classification and Probabilistic Update

We train and evaluate a multi-class classifier, specifically we perform HMM classification where the number of hidden states = 6 and the number of Gaussians under each state = 3. To generate classification models, each HMM is trained using manually annotated ground truth sequences from the TRECVid dev08 data. After classification, the log-likelihood outputs from each HMM model are used to evaluate event occurrence probabilities across the known list of classes. For a set of possible event classes Ck, $k = 1, ...G$, it is assumed that a new sequence of observation data D belongs to only one class, and that the closed set of classes C represents the complete list of possible events. In this manner event classification scores are synonymous with the conditional probabilities of D belonging to all classes, and can be computed directly from log-likelihood outputs.

Probabilistic update of event scores is achievable using *a priori* knowledge of the location of (ground truth) event occurrence. For a subset of camera scenes (specifically camera 1 and 3) and events (pointing and object put), individual event occurrences are spatially localised and cumulative event occurrence is represented as a pixel-wise probability distribution. In this manner, a heat map (sum of event occurrence per pixel across all training frames) is generated to show where each event tends to occur in each scene. After HMM classification, the posterior probability of event occurrence is viewed as a weighted combination of the HMM probability and the mean a priori probability across the sequence's ROI.

For NMF based features, four events (CellToEar, ObjectPut, PersonRuns and Pointing) are investigated. As an alternative experiment to reduce the false alarms, we manually selected 80 sequences that contain nothing related to the target events. They are considered as the "NoEvent" and used as the 5th event in the classification. The classification is performed as described above and the event is decided by the maximum log-likelihood. The events if identified as "NoEvent" are removed from the final results.

Table 3: Summary of results for retrospective runs

Ulster		run1			run2			run3		
Event	**#Targ**	**#Sys**	**#Cor**	**DCR**	**#Sys**	**#Cor**	**DCR**	**#Sys**	**#Cor**	**DCR**
CellToEar	194	279	2	1.081	–	–	–	752	2	1.236
ObjectPut	621	987	27	1.271	255	8	1.068	751	21	1.206
PersonRuns	107	477	1	1.147	–	–	–	–	–	–
Pointing	1063	867	39	1.235	25	4	1.003	765	28	1.215

Results and Discussion

We submitted three runs. Run 1 is sequence based classification with NMF, run 2 is sequence based classification with PCA and *a priori* update, and run 3 is sequence based classification with PCA. Of all UU runs, maximum performance was achieved using the HMM with a priori update. This is in concordance with previous investigations, which found that integration of a priori information could improve event classification accuracy [8]. Classification score update via priors reduced the DCR from 1.21 to 1.07 for ObjectPut and 1.22 to 1 for Pointing events (run2 versus run3). Additionally, the number of false alarms dropped from 730 to 247 for ObjectPut and from 737 to 21 for Pointing events. There exists a trade off between true positive classification and reduction of false alarms. Some true positives were missed after a priori classification update, but the overall Detection Cost Ratios were better.

The performance from Run 1 (NMF based) appears to be limited but still encouraging. Comparing to Run 2 and Run 3 (PCA based), Run 1 achieves a higher volume of correct detections for all events. For event "CellToEar" NMF also achieves a lower DCR (1.081) than the PCA method without prior update (Run 3). However it has higher false alarms for the rest events. One possible reason can be due to the preprocess and postprocess, which can play an important role in TRECVid event recognition as shown by results from Run 2. In terms of algorithm, it is reported that NMF may not always provide part based feature representations [5]. Therefore it can be challenging to apply NMF to TRECVid data, which has very noisy cluttered background. Although person detection is applied, the performance may still be affected. A more robust approach will be considered to incorporate with NMF in the future work.

4 Conclusions

In the SIN task the major challenge we faced was how to handle the processing scale required to push the performance boundaries beyond the existing standard methods. We plan to move parts of our processing and classification pipeline to a High Performance Computing (HPC) system to access greater computational power and increase our classification options. In the interactive SED task, the main challenge is the differing information needs of the SAVASA users to those evaluated. Users wish to search for specific instances of events such as person falling, loitering, vandalism, fare avoidance etc. and have limited interest in rapidly identifying simpler events. However there are ongoing challenges in the automatic machine annotations of events in surveillance video due to noise, crowded scenes and high inter-class variation for the labelled actions in the iLIDS dataset (e.g., child running around compared with someone running across the frame). In conclusion, the participation in SIN and SED by the SAVASA project has been a valuable experience in bringing together components from multiple partners, directly discussing and evaluating our system with users and evaluating individual methodologies for object detection and activity recognition.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 285621, project titled SAVASA.

References

[1] R. Zdunek A. Cichocki and S. Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893. IEEE, 2005.

[3] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision (ECCV)*, pages 428–441. Springer, 2006.

[4] A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pages 726–733, 2003.

[5] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[7] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–799, 1999.

[8] S. Little, I. Jargalsaikhan, K. Clawson, M. Nieto, H. Li, C. Direkoglu, N. E. O'Connor, A. F. Smeaton, B. Scotney, H. Wang, and J. Liu. An information retrieval approach to identifying infrequent events in surveillance video. In *Proceedings of the 3rd ACM International Conference on Multimedia Retrieval*, pages 223–230, 2013.

[9] S. Little, I. Jargalsaikhan, C. Direkoglu, N. E. O'Connor, A. F. Smeaton, K. Clawson, H. Li, M. Nieto, A. Rodriguez, P. Sanchez, K. Villarroel Peniza, A. Martínez Llorens, R. Giménez, R. Santos de la Cámara, and A. Mereu. SAVASA Project @ TRECVid 2012: Interactive Surveillance Event Detection. In *TRECVid Workshop*, 2012.

[10] M. Nieto, J. D. Ortega, A. Cortes, and S. Gaines. Perspective multiscale detection and tracking of persons. In *International Conference on Multimedia Modelling (MMM)*, 2014.

[11] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quéenot. TRECVID 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2013*. NIST, USA, 2013.

[12] I. Khan P. M. Roth, T. Mauthner and H. Bischof. Efficient human action recognition by cascaded linear classifcation. In *ICCV*, 2009.

[13] Bahjat Safadi and Georges Quénot. Descriptor optimization for multimedia indexing and retrieval. In *CBMI*, pages 1–6, 2013.

[14] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[15] Ichiro Takeuchi, Quoc V. Le, Timothy D. Sears, Alexander J. Smola, and Chris Williams. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:7–1231, 2006.

[16] C. Thurau and V. Hlavàc. Pose primitive based human action recognition in videos or still images. In *CVPR*, 2008.

[17] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[18] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, (57), 2004.

[19] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, pages 1–21.

[20] H. Wang, A. Klaser, C. Schmid, and C.L. Liu. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3169–3176, 2011.

[21] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference*, 2009.

[22] C. Zeng and H. Ma. Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. In *ICPR*, pages 2069–2072, 2010.