An Information Retrieval Approach to Identifying Infrequent Events in Surveillance Video

Suzanne Little*
Iveel Jargalsaikhan
Cem Direkoglu
Noel E. O'Connor
Alan F. Smeaton
CLARITY, Centre for Sensor Web Technologies
Dublin City University, Ireland

Kathy Clawson
Hao Li
Jun Liu
Bryan Scotney
Hui Wang
University of Ulster
Belfast, United Kingdom

Marcos Nieto
Vicomtech-IK4
San Sebastian, Spain

ABSTRACT

This paper presents work on integrating multiple computer vision-based approaches to surveillance video analysis to support user retrieval of video segments showing human activities. Applied computer vision using real-world surveillance video data is an extremely challenging research problem, independently of any information retrieval (IR) issues. Here we describe the issues faced in developing both generic and specific analysis tools and how they were integrated for use in the new TRECVid interactive surveillance event detection task. We present an interaction paradigm and discuss the outcomes from face-to-face end user trials and the resulting feedback on the system from both professionals, who manage surveillance video, and computer vision or machine learning experts. We propose an information retrieval approach to finding events in surveillance video rather than solely relying on traditional annotation using specifically trained classifiers.

Categories and Subject Descriptors

H.1.1 [**Systems and Information Theory**]: Information Theory; I.2.10 [**Vision and Scene Understanding**]: Video analysis

Keywords

surveillance event detection, video analysis

1. VIDEO SURVEILLANCE EVENT DETECTION

Efficiently and reliably finding complex events of interest in video surveillance footage presents several challenges. These include: the volume of data to be processed, especially in relation to the frequency of event occurrence; the low resolution and high noise of the video including activity occlusion and multiple co-occurring events; the low inter-class variance between many events and, finally, users' misleading expectations of the computer's ability to accurately and confidently identify people, objects and activities. In this paper we present the outcomes from our experiences in applying multiple computer-vision based approaches to support interactive user retrieval of video segments from surveillance video with feedback from end users including professionals from major companies in transport and event stadium surveillance and security.

By surveillance video we are generally considering fixed CCTV cameras placed for security observation purposes where video archives will be accessed post-event for investigative or judicial reasons. Events of interest may include (but are not limited to): unauthorised access, accident, anti-social behaviour, abandoned luggage, unauthorised photography/filming, sabotage or equipment tampering, etc. Such complex events can be identified with varying but generally low levels of success using computer vision techniques. Simpler analysis tasks include person detection, gesture modelling and object detection that can then be used by semantic classifiers for the more complex events.

In our research we aim to develop a standards-based video archive search platform that allows authorised users to perform semantic queries over various remote and non-interoperable video archives of CCTV footage from geographically diverse locations. At the core of the semantic search interface is the output 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.

To facilitate these aims, we took part in the TRECVid Interactive Surveillance Event Detection task 2012 [16]. TRECVid is an annual benchmarking exercise sponsored by the US National Institute of Standards and Technology (NIST) with the aim of stimulating video information retrieval research and improving the performance of systems using large, challenging, realistic and noisy datasets for real world problems. Surveillance Event Detection using CCTV footage has been a TRECVid task for the previous five years but, due in part to the lack of significant improvements in detection rates, was changed this year to include an interactive element. Previously, a set of test (unannotated) videos would be processed by one or more event classifiers and the ordered list of possible matches would be evaluated to determine the system's performance. This year a human computer interface could be used to find matching video segments in the test set with a 25 minute search time limit per user per event class.

*Contact author: suzanne.little@dcu.ie

Our approach was to combine individual methods for video analysis and annotation and provide a dashboard style search interface that enabled the user to view results for various algorithms and filter them by factors such as confidence, level of motion, camera, number of people etc. The use of a search interface changes the problem from a pure annotation task, where the objective for each individual classifier is to maximise precision and recall for one or more events over all videos, to an information retrieval task, where fuzzy, difficult to predict factors such as user understanding and patience, interface aesthetics, minimising false alarms and optimising for high precision in the top ranked results have the greatest influence on the success of the system.

This paper describes the process of bringing our independent algorithms together, the analysis of the surveillance video characteristics and the discussions we held with professional end user partners to establish the challenges to be faced in developing systems to support usable, practical search systems in this field. The next section analyses the characteristics of the TRECVid dataset, describes user profiles from our interactive retrieval evaluations and discusses the implications of the specific characteristics of surveillance event detection, as listed in this section, on an information retrieval (IR) user model. Section 3 presents, at a high level, the main computer vision approaches that were applied to analyse and identify video segments to show to the user and some preliminary evaluation scores relating to their performance. Finally the conclusions and future work from our collaboration are presented.

2. TRECVID INTERACTIVE EVALUATIONS

2.1 Dataset characteristics

As listed in the first section, there are three main challenges associated with visual events in surveillance videos: low event frequency, noisy and low-resolution data and difficult to describe events. The video dataset used in the TRECVid task comprises 145 hours of footage, captured over a number of days from five fixed cameras in an airport, that has been previously annotated with ten events (PersonRuns, Pointing, CellToEar, ObjectPut, Embrace, PeopleMeet, PeopleSplitUp, ElevatorNoEntry, OpposingFlow, TakePicture) of which three (ElevatorNoEntry, OpposingFlow, TakePicture) are not currently used. 100 hours is used for training and 45 is reserved for testing of which 15 hours were used this year. The video is provided in MPEG-2 at 25 frames per second, 720×576 frame resolution. For our submissions we used two cameras (CAM1, CAM3) and focussed on three events – ObjectPut, PersonRuns and Pointing. These events were performed by a single person, had a reasonable number of training examples in the chosen videos and were gesture-based.

Table 1 summarises the frequency and duration of events in the training dataset (100 hours) calculated from the manual annotations provided in ViPeR format [7] by the TRECVid organisers. Median duration is given as there are a few very long segments in the provided annotations that we think are incorrect and therefore overly inflate the mean.

Many approaches to event detection that use motion information apply a sliding temporal window to segment the continuous CCTV footage [1, 17]. The size of this window is critical; note median ObjectPut duration is 10 frames compared with PersonRuns at 67 frames or PeopleSplitUp at 167 frames. The most common configuration is a sliding window size of 15 frames with steps of between 4 and 10 frames. Subsampling of the videos either by resizing the frames to a lower resolution or by only extracting features from a subset of frames (e.g., every second frame) is also often applied to reduce the computational effort. The cumulative effect of these decisions sets the sampling density of the video. For example, Figure 1 illustrates a sliding temporal window of size 15 frames with a step size of 10 frames. If applied over 1 minute of video (at 25 fps) containing 1500 frames this would produce 120 video segments (15 frames in length) to be analysed.

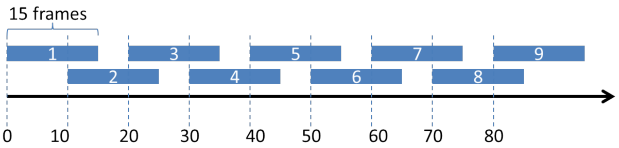

Figure 1: Illustration of temporal sliding window over continuous video footage

60% of ObjectPut and 32% of Pointing events take place in less than 15 frames. PersonRuns events have much longer durations with less than half a percent (4 samples) under this threshold. In contrast, events with duration greater than that of the sliding window will be described by a sequence of partial events spanning multiple windows that need to be classified and fused to identify the event's start and end points. Smoothing the degree of varation across a sequence of windows by using a smaller step size would potentially improve the accuracy but would increase the computational load.

Subsampling of videos, by selecting every 2nd or 3rd frame, can also improve computational efficiency, especially where expensive descriptors are being calculated. However, if the duration of events is less than 3 frames (and Table 1 shows some are) important information is lost. Some of our approaches (described in section 3) used every 2nd frame for person tracking or to calculate motion trajectories for example as it improved responsiveness. Of the three events we investigated, over 4% of the ObjectPut events in the training set had a duration of 2 frames or less while for Pointing and PersonRuns it was just under 1%. This means that some relevant event segments will never be detected or shown to the user – reducing the maximum possible recall for the final system. Large window sizes are also likely to mute the important features for an event of very short duration.

The spatial location and relative size of the person involved in the event is also useful to consider. The region of interest (ROI) where the actual event activity takes place is not provided in the original annotations. We performed manual annotations on a subset of training videos from cameras 1 and 3 for ObjectPut and Pointing events on every second frame annotated with the event. Figure 2 shows four example heat maps indicating the normalised frequency of each frame pixel being part of the event. These graphics show clear hot spots for some events and differences in the location of events within the same camera area. Section 3.2 also discusses how this prior probability can be exploited to improve classifier confidence.

Figure 3 shows the pixel area of the event regions as a percentage of the total frame size (720×576) based on the approximate region of interest dimensions from our manual annotations. The area of the region of interest for these events is rarely more than 12% of the total frame. While there is certainly some connection to the camera configuration – note approximate similarity of ROI sizes between the different events on CAM1 but less so on CAM3 (a wider viewing area) – it's clear that these are difficult even for people to visually identify due to the typically short duration and the small activity area. The heat maps show that while there are some spots where the events most frequently occur, there's still variation

Table 1: Characteristics of the 7 main events in TRECVid Dev08/Eval08 training video dataset

Event	Frequency (#)			Duration (# frames (secs))			
	All	CAM1	CAM3	CAM1		CAM3	
				median	min/max	median	min/max
CellToEar	828	40	284	16.5 (0.66s)	5/429	17 (0.68s)	1/123
Embrace	940	27	629	66 (2.64s)	27/636	71 (2.84s)	1/2034
ObjectPut	**3177**	**706**	**903**	**11 (0.44s)**	**1/625**	**9 (0.36s)**	**1/419**
PeopleMeet	2719	813	906	65 (2.6s)	1/1176	106.5 (4.26s)	1/3236
PeopleSplitUp	1571	762	235	179 (7.16s)	1/7169	135 (5.4s)	1/4287
PersonRuns	**673**	**25**	**218**	**54 (2.16s)**	**14/268**	**67.5 (2.7s)**	**18/276**
Pointing	**4097**	**926**	**1106**	**24 (0.96s)**	**1/2717**	**21.5 (0.86s)**	**1/360**

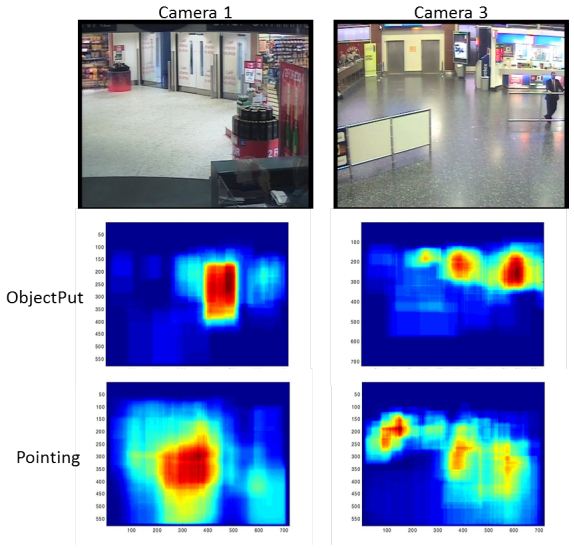

Figure 2: Heat maps showing frequency of pixel involvement in event

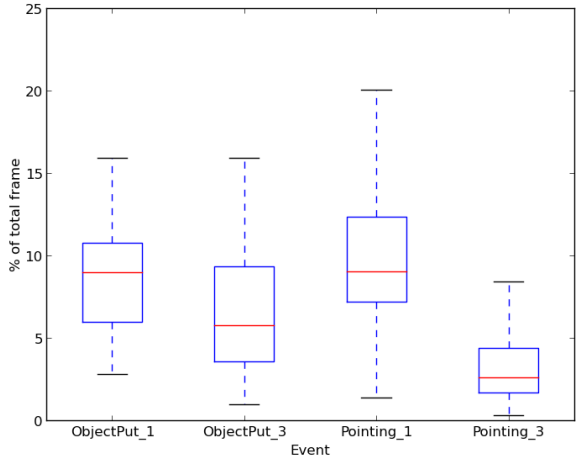

Figure 3: Event region size per camera as % of total viewing frame

both within and between events and cameras. The challenge is assisting the viewer to rapidly and accurately scan suggested matches. More analysis over all events and cameras is required to determine any significant trends.

Event frequency for the TRECVid dataset, a genuine sample collected from a real life situation, shows that the occurance of events ranges from just under 7 events per hour (PersonRuns) up to almost 41 events in an hour (ObjectPut). However, the events annotated for the TRECVid task are somewhat artificial – a point made by our users during the evaluations (see discussion in section 2.4) – and therefore frequency for true events of interest in the surveillance space is likely to be even less than that demonstrated here.

2.2 User profiles

Based on our visits to end users, we identified two main categories of user – *professionals* and *computer experts*. In this section we describe the characteristics of each user category. Eight users (four in each class) from four institutions completed the evaluation. All users were given an introduction to TRECVid and a demonstration of the search interface. They were given time to familiarise themselves with the interface and the three events using video loaded from the training dataset.

Professionals generally worked directly in managing surveillance video systems or archives in a major company in transport or stadium surveillance. These users had no previous experience of TRECVid, evaluation workshops or of how research in information retrieval fields is evaluated. They view and search surveillance videos as part of their everyday tasks and therefore have high knowledge of surveillance video characteristics (resolution, view point etc.) and understood how difficult and time consuming current methods are for identifying events of interest. They gave minimal importance to completing the task and, in some cases, found the specific events used by TRECVid (e.g., Pointing) to be of little practical interest. The discussions held with these users during the evaluation was enlightening with respect to identifying their expectations of what multimedia information retreival systems were capable of and what events were likely to be useful for their work.

Our *computer experts* had experience in programming, often specifically in machine learning or computer vision. They understood the purpose of the evaluation exercise and treated it more competitively. At the conclusion of the time period spent on each task a notification dialogue would tell the user how many segments they had selected and the amount of search time they had spent. This was often ignored by the professionals but noted by the computer experts. Computer experts were generally more forgiving of obvious mis-classifications and displayed more willingness to try different settings, camera filters and looked at results from all of the different algorithms.

Table 2 shows the search time spent by each user type – quantifying the main difference between the two classes. Note that PersonRuns had fewer segments identified by the underlying classifiers. Our prediction was that computer experts with superior understanding of how computer vision and search systems worked would have

Table 2: Average search time (seconds) per event class for each class of user

	computer expert	professional	all
ObjectPut	1109	238	673
PersonRuns	187	95	141
Pointing	1020	184	602

an advantage in navigating and applying the various options and, having more patience with underperforming algorithms, would find more correct results.

2.3 Interaction paradigm

We've discussed the characteristics of the TRECVid surveillance video dataset and looked at the relative frequency, varying duration and area of the event's region. We've also described the two main classes of user we used for the TRECVid evaluation task. In putting together a system to support these users in finding as many instances as possible of events in the surveillance video test set, decisions and trade-offs need to be made to determine where in the interaction paradigm the processing burden for various tasks is best placed.

Option 1 is to optimise the underlying classification systems for recall and allow the user to act as a filter, removing false alarms. The system tries to reduce the amount of video the user needs to view while avoiding losing relevant segments. The motivation is that segments can't be selected by the user as matches if they have not been annotated by at least one of the underlying classification options. Providing low-level filtering options (amount of activity, number of people etc.) also contributes to this paradigm by giving the user more precise control. This is the option we chose to use for the TRECVid interface.

Maximising recall in an interactive system would seem to be an attractive option as conventional thinking among information retrieval developers is that recall is paramount in applications such as security, science, health etc. where consequences for missing information are higher. However, direct feedback from the professional users indicates that false alarms (incorrect suggestions) are highly detrimental, reducing user confidence in the system performance and limiting its practical utility. In discussion users commented that in many scenarios they were willing to miss some examples rather than having to deal with too many results. Further research is required here to establish the limits of this thinking and the specific conditions where recall can be sacrificed for higher user confidence and more satisfactory interactions.

Option 2 would be to optimise for precision on very specific objects, events, cameras and configurations so that each classifier is highly specialised but (hopefully) more reliable. The user would then act as a form of late fusion, choosing the different classifiers based on detecting component objects or events. The interface would need to be carefully designed to support the professional user by using familiar, consistent terminology and some training would be beneficial to ensure users understand the different options available to them.

The trade-off between precision and recall is influenced by the user characteristics and their requirements. The dataset characteristics also bring two complicating factors when considering the design of the interaction. The first relates to the event duration and small relative region size. For the TRECVid task we chose to display results ordered by classifier confidence in a grid using looping animated gifs downsampled to half-size and showing every 3rd frame to exploit the users' ability to quickly scan and identify correct results – particular the experienced professionals. As discussed in the section 2.1, some events have durations less than 2 frames and are very difficult to identify from a small, downsampled gif. In addition, the mean region of interest size shows that activities in the TRECVid dataset tended to occur in a small region and were much harder to spot than anticipated. This trade-off for 'human processing' is equivalent to the traditional one for computer vision classifiers between ensuring good coverage through dense feature sampling and reducing computational complexity.

The second complicating factor of the dataset was the question of how to consistently segment the videos. In using an information retrieval style interface that displays lists of order results, some method was required to define start and end times of the events. Each classifier used a different approach to determining the length of video segments including a temporal sliding window, fixed trajectory lengths, aggregating confidence values across frame-based sequences to produce segments and using person-based tracking. For TRECVid, it is very important that the start and end frame values are accurate as they are used automatically to assess the correctness of each result.

Overall, the effect of event duration and relative size of event area to frame size makes it hard for a user to easily spot events when shown an out-of-context video segment. Improvements are likely by using fast-browsing or summarisation (without frame-based downsampling) rather than requiring the user to filter video segments. Based on the dataset and user characteristics and our experiences with the professional users in TRECVid, we propose that an interactive user information retrieval approach supported by a structure of highly specialised semantic classifiers is likely to be better for surveillance video due to the infrequency of events, user dislike of false alarms and the general difficulty of automatically identifying events. The next section presents the evaluation results for both user classes from the interactive runs submitted to TRECVid using a simple, retrieval interface with classifiers optimising for recall.

2.4 Preliminary results and feedback

Table 3 shows the evaluation results for the interactive runs submitted to TRECVid. These comprised the set of all segments identified by users in either the computer expert or professional class ordered by the frequency of the segments selection and the normalised confidence score of the underlying classifier. #Sys is the number of video segments identified for the event by the system, #Cor is the number of these segments that are correct. RFA is the rate of false alarms defined as the number of incorrectly identified segments (#Sys-#Cor) per hour. PMiss is the number of missed detections over the total number of observed events for the target class (#Targ). DCR is Detection Cost Ratio which measures performance in terms of the cost per unit time and is calculated as the product of PMiss and RFA. Therefore a perfect system will receive a DCR of 0. The video segments in the results list, defined by start and end frame numbers, are matched with the ground truth annotations using the Hungarian Solution to the Bipartite Graph [13].

The most surprising feature of the interactive results was the high number of false alarms. Our prediction was that having a user act as a result filter would greatly reduce the incorrect annotations. One possibility is that the high RFA is due to the sensitivity of the event alignment method used in the evaluation. It is likely that the event segments found by the classifiers have a shorter duration or inexact overlap with the groundtruth event segments. This would be interesting to explore further, as for surveillance video retrieval tasks it

Table 3: summary of results for interactive runs

user, event	#Targ	#Sys	#Cor	RFA	PMiss	DCR
P ObjectPut	621	48	3	2.951	0.995	1.010
C ObjectPut	621	64	3	4.000	0.994	1.015
P PersonRuns	107	10	2	0.525	0.981	0.984
C PersonRuns	107	14	2	0.787	0.981	0.985
P Pointing	1063	25	12	0.853	0.989	0.993
C Pointing	1063	100	25	4.919	0.976	1.001

is less critical to find the exact boundaries of the event but rather to identify that an event has occured, particularly if the task is interactive.

Our user evaluations provided us with the opportunity to meet with professional end users from major companies in the security and surveillance field and discuss their requirements for event retrieval. Many of the users we met with were unaware of TRECVid and information retrieval research in general. The particular activities annotated in the dataset were not of interest to our users who were often confused as to why we had annotated events such as Pointing.

We expected that the computer experts would produce better results due to a clearer understanding of the task and a willingness to spend more time searching for relevant segments. Contrary to our expectations the professionals results were slightly better. This is most likely due to the smaller number of correct results in the segments found by the underlying classifiers and available to the users while searching. The computer experts were also more likely to find more segments and hence have a higher false alarm rate.

The evaluation metric is based on how closely the start and end frames of the segment match with those identified by the humans who created the ground truth. The various methods for segmenting the video will have different levels of precision in determining the exact start and end point. It is not clear if the surprising number of false alarms, even in results chosen by our users, is due to overly stringent restrictions on the start/end numbers by the matching algorithm. For interactive retrieval of video for surveillance tasks this level of precision may not be necessary.

3. COMPUTER VISION TECHNIQUES FOR MULTIMEDIA IR

At the backend of our interface we applied a number of computer vision techniques to provide semantic annotations indicating which segments of the test videos were likely to contain an event. This section gives high-level descriptions of these components, including the evaluation results from processing the TRECVid dataset without the user intervention, and outlines some future directions.

3.1 Person tracking

All events in TRECVid and most of general interest in surveillance involve people and therefore person detection and tracking is a key component. We were interested in using this to identify regions of interest for other classifiers to analyse and to produce raw statistics about video scenes such as the number of people (crowded), the degree of activity or motion. In this section we describe how existing techniques were adapted to handle the type of footage available in the TRECVid dataset.

For detection, we used HOG descriptors [3] and a pre-trained person detector which yields a "sparse" set of detections in time, i.e. there are a lot of misdetections. False negatives can be solved using tracking approaches, which are anyway needed to provide time coherence to detections, so that we can reconstruct the trajectory of objects.

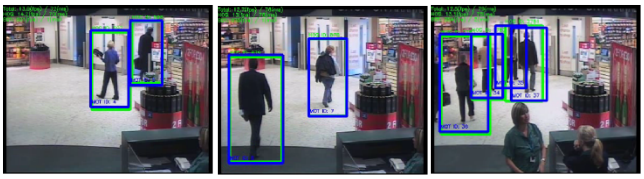

Figure 4: Examples of Person Tracking

For the tracking, we have implemented a Rao-Blackwellized Data Association Particle Filter (RBDAPF) [9]. This type of filter has been proven to provide good multiple object tracking results even in the presence of "sparse" detections as the ones we have in these sequences, and can be tuned to handle occlusions. The Rao-Blackwellization can be understood as splitting the problem into linear/Gaussian and non-linear/non-Gaussian parts. The linear part can be solved with Kalman Filters, while the non-linear one must be solved with approximation methods like particle filters. In our case, the linear part is the position and size of a bounding box that models the persons. The non-linear part refers to the data association that is the process of generating a matrix that links detections (the HOG ones, for instance), with objects or clutter. The association process can be strongly non-linear so that sampling approaches can be used. In our case we have implemented ancestral sampling [6].

The control of input/output of new persons is handled thanks to the use of the data association filter that classifies detections according to the existing objects, removes objects that remain undetected for a sufficiently long time, and creates new objects when detections not associated to previous objects appear repeatedly.

Preliminary results indicate that this approach is able to detect and track up to four or five simultaneous persons whose full body is clearly seen in the scene. With more than five persons we have found that in these types of images multiple occlusions happen and the full-body detector does not provide good detection results.

High-levels of occlusion and very crowded scenes remain a challenge. Figure 3 shows the pixel area of the event region relative to the total frame size is often only between 4 and 12% and for some configurations may be less than 5%. The relative size of people will be even smaller and the camera configuration means that often only the head/shoulders is consistently visable. Future plans are to apply this work to real data from project partners, to use research datasets such as CUHK occlusion[1] to improve the ability to track occluded persons and to examine applying fluid dynamics models (such as [8]) to recognise other crowd-based behaviour and events.

3.2 Region-based activity recognition (rb1)

The motivation behind region-based activity recognition was to use the output from person tracking to segment the frames and identify likely regions for further analysis. This was trialled on a subset of the training data and used as input for the manual region of interest annotation activities but due to difficulties with accurate tracking in crowded scenes it was found to be insufficient to apply unsupervised on the test videos at this stage. Therefore a frame-based approach employing a fixed grid was applied. Comparison systems were implemented that use prior-probability based on the region of interest data to improve overall detection scores. This section de-

[1] http://www.ee.cuhk.edu.hk/~xgwang/CUHK_pedestrian.html

scribes the two frame-based methodologies for event recognition: using Optical Flow features with a Hidden Markov Model (HMM) classifier (rb1c) compared with dense SIFT features processed with a Bag of Words (BoW) and applied with an SVM classifier (rb1a).

For the Optical Flow features we computed a normalised histogram of oriented optical flow (90 bins, equally spaced) where the magnitude of each bin corresponded to the sum of the magitude of the optical flow. To describe these we used 2D Zernike Moments [24](p689) of Efros Descriptor Images [10] as follows. We calculate the optical flow vector field F for each frame and split F into two scalar fields, Fx and Fy, corresponding to the horizontal and vertical components of the flow. Then we half-wave rectify Fx and Fy into 4 non-negative channels: $Fx+$ $Fx-$ $Fy+$ and $Fy-$. Finally, we blur with a Gaussian to remove spurious motions. These channels, known as Efros descriptors [10] may be regarded as distinct images. As features, we calculate 2D Zernike moments of each of the new channels (16 features per channel, per frame), resulting in a 64×1 feature vector per frame. These features are concatenated into a single vector (154×1) and the feature space is reduced to 16×1 using principal components decomposition.

As a constrasting approach, we also extracted dense SIFT features [2] around spatial interest points and clustered these features to create a visual bag of words. For classification model learning and test set evaluation, histograms of visual words are used as input features to an SVM with radial basis function (RBF).

To generate classification models, we trained our HMM and SVM using ground truth events from the TRECVid training dataset. Specifically, we utilised the manually identified regions of interest within each frame as a unique event instance. To apply frame-based classification to the TRECVid test data, we adopt a grid based approach. Each frame is divided into 36 equally-sized regions and each region is evaluated separately. After classification using both methods, we threshold to retain only the top *n* video segments (ranked by confidence), and link these segments temporally to derive start and end times. The final confidence score is the mean confidence across the linked time period. We deliberately set *n* high to allow high false positive rates, in the hope that this will allow a higher proportion of true positives to be captured and therefore be available through the user interface.

Given that a fixed grid was used to segment the test video frames, we were also interested in using the prior probability information about the location of our events to improve the final confidence measure. We represented each heat map (shown in figure 2) as a pixel-level probability distribution whose sum is 1. Based on this the probability of an event occuring within a single grid region was calculated as the sum of probabilites from within that grid. The result of this is a 6×6 matrix containing update weights that we applied to adjust the final classification score (SVM classifier – rb1b; HMM classifier – rb1d).

The next steps for this approach are to incorporate the improvements in person tracking and exploiting the tracking to move from a frame-based, fixed-grid method to motion history images (MHI) [5]. The use of prior probability displays some promise (see section 3.4) so we are interested in completing more events and further testing its usefulness.

3.3 Motion trajectory (mt1)

Person-based activity recognition using motion trajectory is a common approach [21, 17, 1] based on identifying and describing patterns of movement. The types of events we were interested in identifying in the TRECVid dataset involved temporal actions by a single person, therefore classifiers using motion trajectory descriptions was a clear avenue of investigation.

To represent motion, we used salience point trajectory as a low-level feature and described it using four different descriptors. First, in order to extract the motion trajectory, we applied a background subtraction algorithm [12] to detect foreground regions. This processing helps to reduce computational complexity and increases the accuracy of point tracking by reducing the search area. Salience points [19] are located within the foreground regions by a Harris Corner Detector and are tracked over video sequences using the Kanade-Lucas-Tomasi (KLT) algorithm [15]. In the experiments, we have observed that longer salience point trajectories are likely to be erroneous. Thus we empirically set the maximum trajectory length to be 15 frames.

We adopted Heng *et al.*'s [20] approach to describe the trajectory features. For each trajectory, we calculated four descriptors to capture the different aspects of motion trajectory. Among the existing descriptors, HOGHOF [14] has shown to give excellent results on a variety of datasets [21]. Therefore we computed HOGHOF along our trajectories. HOG (histograms of oriented gradient) [3] 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.* [4], and TD (trajectory descriptor) [20] are computed in order to represent the relative motion and trajectory shape.

In order to represent the video scene, we have built a Bag-of-Features (BoF) model based on our four descriptors. This requires the construction of a visual vocabulary. In our experiments, we cluster a subset of 250,000 descriptors sampled from the training videos with the k-means algorithm for each descriptor. The number of clusters is set to $k = 4000$, which has shown empirically to give good results in [14]. The BoF representation then assigns each descriptor to the closest vocabulary word in Euclidean distance and computes the co-occurrence histogram over the video sub-sequence.

For classification, we used a non-linear support vector machine (SVM) with a Radial Basis Function (RBF) kernel. Using the cross-validation technique, we empirically found the parameters of cost (32) and gamma (1×10^{-5}) of the kernel. In order to represent the video frame, we utilized a temporal sliding window approach. In the experiments, we set the window size to 25 frames and sliding step size to 8 frames.

Here we have implemented an action detection algorithm that is based on sparse motion trajectory. Since trajectory is suitable for representing gesture-like movements, we mainly focused on building a classifier for Pointing events using this method. Although we have not considered any spatial association between the extracted trajectories' descriptors, this approach performed reasonably well on the challenging TRECVid SED dataset. In the future, we would like to explore an alternative way to represent the video scene rather than the Bag-of-Features (BoF) approach that ignores the spatial information.

3.4 Evaluation

Table 4 shows the outcome of the automatic classification runs submitted to TRECVid 2012. Section 2.4 defines the metrics (RFA, PMiss, DCR).

The purpose of the different runs for the 'rb1' configuration was to evaluate and compare two frame-based supervised-learning techniques for event classification (HMM with optical flow features and SVM with BoW), considering whether or not *a priori* information could increase accuracy and reliability of event classification. Across the 'rb1' experiments, the SVM incorporating *a priori* information was slightly more successful.

One factor worth mentioning is the high RFA across all exper-

Table 4: Sumary of results for automatic runs

run, event	#Targ	#Sys	#CorDet	RFA	PMiss	DCR
mt1, ObjectPut	621	9	0	0.59027	1.000	1.0030
mt1, PersonRuns	107	20	3	1.11496	0.972	0.9775
mt1, Pointing	1063	136	16	7.87029	0.985	1.0243
rb1a, ObjectPut	621	457	11	29.25123	0.982	1.1285
rb1a, Pointing	1063	981	50	61.06030	0.953	1.2583
rb1b, ObjectPut	621	308	3	20.00364	0.995	1.0952
rb1b, Pointing	1063	950	57	58.56805	0.946	1.2392
rb1c, ObjectPut	621	730	24	46.30352	0.961	1.1929
rb1c, Pointing	1063	2174	96	136.28712	0.910	1.5911
rb1d, ObjectPut	621	876	22	56.0102	0.965	1.2246
rb1d, Pointing	1063	1286	56	80.67043	0.947	1.3507

iments. During temporal linking and thresholding, our threshold was set intentionally to overestimate event occurrence. Future objectives include the reduction of RFA, by modification of threshold values and generation of enhanced / improved event models for classification. It is envisaged that we calculate non motion-based descriptors for event classification, and combine these with our existing optical flow feature sets. The inclusion of the user gave a 1% improvement in the performance over the fully automatic classifiers. As discussed in section 2.4, this is considerably less than we expected given the generous thresholding and the ability of the user to determine an event match.

One consideration that we realised after submission was that the ObjectPut definition used by the manual annotators sets the start frame as when the person has released the object. We didn't take this into account when defining the training segment. Therefore assumptions about the ability to detect the motion of a person's downward gesture before releasing the object were incorrect. This illustrates the effect that decisions about temporal segmentation of the video can have on performance.

The retrieval of events in surveillance video is a very challenging task and the low absolute values for the DCR metric reflect this. Our numbers are reasonable for the TRECVid challenge where we ranked in the second quartile for almost all of our runs.

4. CONCLUSIONS AND FUTURE WORK

Future work will continue on improving the accuracy and responsiveness of the underlying computer vision systems and is likely to focus more on developing specific components to work in collusion with each other. We aim to complete the annotation of region of interest and analysis of dataset characteristics and have held discussions with fellow TRECVid participants to collaboratively perform manual annotations. The team from the City College of New York Media [23] have also developed an automatic method for generating heat maps similar to those we have produced. Comparison of techniques using these approaches would be very interesting.

Also of interest is using more formal early fusion techniques. Currently all output is presented equally to the user which caused some inconsistencies as different methods for calculating confidence values and different thresholds were used. We intend to apply fusion [22] and hierarchical modelling techniques [18, 11] to enable classification and tracking of both low-level (person/object) and more complex events as required by our users.

We began this work with the aim of bringing together disparate computer vision methods to support event retrieval in surveillance video, initially for the TRECVid 2012 challenge. Through the process of analysis the dataset characteristics, charting the decisions and tradeoffs during implementation and exploring our assumptions with professional users who manage CCTV collections in their daily work, we have established better understanding of how to manage the challenges presented by this type of data.

A surprising outcome was changing our expectations for a surveillance retrieval system based on feedback from professional users. They found it difficult to see the relevance in the events we were exploring even after explaining the purpose of TRECVid and that actions such as 'Pointing' were examples. It also changed our assumptions about the trade-off between precision and recall. Users have told us that false alarms (common when aiming to maximise system recall) are more irritating than we had expected.

An information retrieval approach for finding events in surveillance video will need to be responsive to users' needs in different scenarios and, given the data characteristics and the difficulty of confidently and accurately finding many of the events of interest, a hierarchical, tool-kit based approach is likely to be the most effective. This would combine a number of generic (e.g., person tracking) and highly specific (e.g., 'whole arm non-aggressive pointing') computer vision based classifiers to enable semi-customised services based on the needs of surveillance professionals. Our future work will be to build upon the feedback gathered from our professional users and develop collections of semantic classifiers to address their requirements for video retrieval in the surveillance domain.

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.

5. REFERENCES

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3), Apr. 2011.

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pages 401–408, 2007.

[3] 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, 2005.

[4] 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, 2006.

[5] J. Davis. Hierarchical motion history images for recognizing human motion. In *Proceedings of the IEEE Workshop on Detection and Recognition of Events in Video*, pages 39–46, 2001.

[6] C. R. del Blanco, F. Jaureguizar, and N. Garcia. An advanced Bayesian model for the visual tracking of multiple interacting objects. *EURASIP Journal on Advances in Signal Processing*, 130, 2011.

[7] D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 4, pages 167–170, 2000.

[8] C. Dogbe. On the modelling of crowd dynamics by generalized kinetic models. *Journal of Mathematical Analysis and Applications*, 2011.

[9] A. Doucet, N. Gordon, and V. Krishnamurthy. Particle filters for state estimation of jump markov linear systems. *IEEE Transactions on Signal Processing*, 49(3):613–624, 2001.

[10] A. Efros, A. 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.

[11] J. Ijsselmuiden and R. Stiefelhagen. Towards high-level human activity recognition through computer vision and temporal logic. In R. Dillmann, J. Beyerer, U. Hanebeck, and T. Schultz, editors, *KI 2010: Advances in Artificial Intelligence*, volume 6359 of *Lecture Notes in Computer Science*, pages 426–435. Springer Berlin Heidelberg, 2010.

[12] P. Kelly, C. Ó Conaire, C. Kim, and N. O'Connor. Automatic camera selection for activity monitoring in a multi-camera system for tennis. In *Third ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1–8, 2009.

[13] H. W. Kuhn. The hungarian method for the assigment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

[14] 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.

[15] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial intelligence*, 1981.

[16] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quéenot. TRECVid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[17] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, 2010.

[18] M. Ryoo and J. Aggarwal. Semantic representation and recognition of continued and recursive human activities. *International journal of computer vision*, 82(1):1–24, 2009.

[19] C. Tomasi and J. Shi. Good features to track. *CVPR94*, pages 593–600, 1994.

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

[21] H. Wang, 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] P. Wilkins, A. F. Smeaton, and P. Ferguson. Properties of optimally weighted data fusion in cbmir. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 643–650, 2010.

[23] X. Yang, Y. Tian, C. Yi, and L. Cao. MediaCCNY at TRECVID 2012: Surveillance event detection. In *TRECVid 2012 - TREC Video Retrieval Evaluation Workshop*, Gaithersburg, MD, 2012.

[24] F. Zernike. *Physica*, volume 1. 1934.