EVENT DETECTION USING LOCAL PART MODEL AND RANDOM SAMPLING STRATEGY FOR VISUAL SURVEILLANCE

Si Wu, Feng Shi, Mathieu Jobin, Emilienne Pugin, Robert Laganière, and Emil Petriu

VIVA Research Lab
School of Electrical Engineering and Computer Science
University of Ottawa, ON, Canada

ABSTRACT

We proposed an approach based on a local part model (LPM) and a random sampling strategy for detecting seven types of events defined in TRECVID Surveillance Event Detection task, such as CellToEar, Embrace, ObjectPut, PeopleMeet, PersonRun, and Pointing. To extract spatio-temporal points, a very dense spatio-temporal grid is simply used for random sampling. At the locations of the points, LPM is employed to compute features that capture both the global and local motion information in the spatio-temporal cuboids. In a bag-of-features representation of human activities, a set of visual words are produced from the obtained LPM-based features, and the events contained in videos are represented as histograms of visual word occurrences. For each type of events, a discriminative support vector machine classifier using RBF-χ^2 kernel is trained. For purpose of event detection in long surveillance videos, spatio-temporal volumes are slid in temporal direction to find candidates containing likely events, and the final decision is made by SVM response thresholding and local-maximum filtering. In addition, an event viewer tool has been developed for manually and efficiently remove false positives in the results.

Index Terms— Event detection, local part model, random sampling, bag-of-features, event viewer.

1. INTRODUCTION

As a result of low cost of cameras and communication devices and growing demand for security, computer vision techniques have started to play an important role in visual surveillance. It is now becoming possible to develop intelligent systems to automatically identify specific events in massive surveillance videos. Many works with great progress have been done in this field [1] [2] [3]. However, human behavior recognition is still a challenging problem. It is far beyond for machine to reach human capabilities in analyzing video content [4].

In the TRECVID surveillance event detection (SED) task [5], there are seven types of human activities including CellToEar, Embrace, ObjectPut, PeopleMeet, PeopleSplitUp, PersonRuns, and Pointing. The training and test videos were taken by five surveillance cameras installed in London Gatwick Airport. As shown in Fig.1, this task is challenging because of viewpoint variations, clutter, high density crowd, occlusion, and so on. To analyze video content, spatio-temporal methods achieved good results. These methods can be divided into two categories: global and local representation-based methods extracting spatio-temporal features from 3D volumes in spatio-temporal dimension. The former represents the entire spatio-temporal volume or the volume of interest by a single descriptor, and applies it to simple activity recognition [6] [7]. By incorporating bag-of-features representation, the latter achieved better results especially for more complex human activities [8] [9] [10] [11] [12]. Specifically, local representation-based methods usually first detect spatio-temporal interest points, and then extract local features. For example, the MoSIFT feature [13] is a 3D extention of SIFT [14], which captures substantial motion in local spatio-temporal volume. In [15], the MoFREAK detector is proposed that represents events using fast-to-compute binary spatio-temporal descriptors. To im-

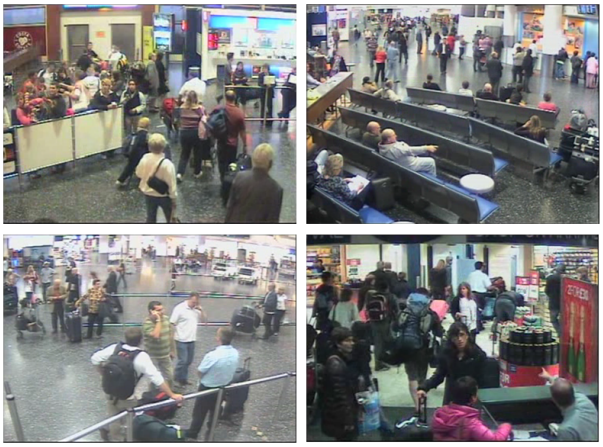

Fig. 1. Human activities captured by four different surveillance cameras.

prove computation efficiency, Willems et al. [16] developed a Hessain 3D detector and an extended SURF descriptor [17]. Furthermore, sampling strategies have been successfully applied to bag-of-features based image classification [18], and the recent works [19] [3] [11] have reported that direct dense sampling of spatio-temporal points produces good results for action recognition. A bag-of-words approach is used [20] to match local features to a set of visual words such that the video can be represented as a histogram of visual word occurrences. The spatio-temporal relationships among the local features are however occluded by the histogramming performed in this kind of approaches.

In this work, we employ a local part model (LPM) [21] to capture both global and local motion information. Specifically, we sample spatio-temporal points on a very dense grid instead of extracting spatio-temporal interest points, which is of benefit to computation efficiency. At each point location, the LPM-based feature is computed from the corresponding local spatio-temporal cuboid. A LPM-based feature is composed of a root descriptor and eight part descriptors to capture global and local motion information in the cuboid respectively. Finally, the bag-of-feature representation is fed to a pre-trained support vector machines (SVM) [22] for classification. The flowchart of the proposed approach is shown in Fig.2. Furthermore, we developed an event viewer tool which allows users to manually flag the result of our event detection approach.

The remainder of this paper proceeds as follows. In Section II, we present the details of our approach for event detection, especially the LPM-based feature and the random sampling strategy for bag-of-feature representation. In Section III, we introduce an event viewer tool for manually removing false positives in the results returned by our event detection approach. The result of our approach performed on the TRECVID SED task is given in Section IV, and we conclude this paper in Section V.

2. PROPOSED APPROACH FOR EVENT DETECTION

In this section, we introduce the proposed approach for event detection, especially the local part model for feature extraction and the random sampling strategy for bag-of-features representation.

2.1. Motion Boundary Histogram

To use motion for event detection, we employ motion boundary histogram (MBH) [23], a flow based-feature for coding motion boundaries or displacements of moving objects, which has been applied to action recognition with good performance [3]. Similar to the histogram of oriented gradient [24], the optical flow field is separated into its horizontal and vertical components. Spatial derivatives are calculated for each component. The gradient magnitudes and orientations are used as weighted votes into orientation histogram. As a result, each component is associated with a histogram. The two histograms are combined as the final descriptor. The MBH descriptor represents the gradient of optical flow, and constant motion information is ignored, which is of benefit to suppressing noise resulting from background motion.

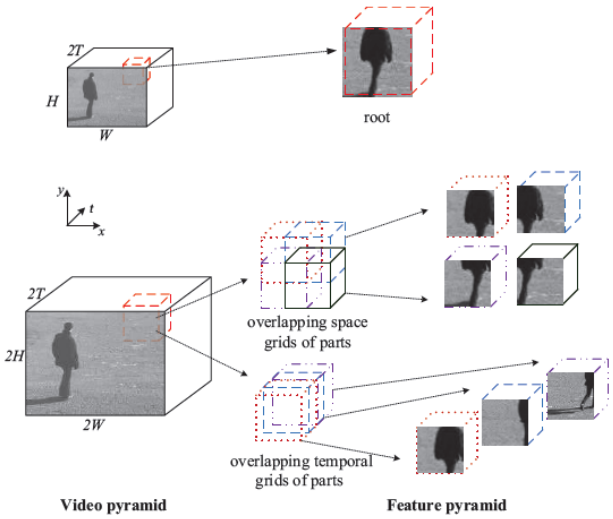

Fig. 3. An example illustrating the local part model.

2.2. Local Part Model

Inspired by the deformable part model [25] for object detection, a local part model [21] has been developed to capture the motion in a local spatio-temporal cuboid. The LPM-based feature is computed by a root filter that covers the entire cuboid and a set of higher resolution part filters that cover smaller sub-cuboids. The sub-cuboids with fixed size and location are defined on a grid that has twice the spatial resolution of the root. The part filters thus capture the fine-grained information. The MBH descriptor is used as both the root and part filters to capture global and local motion information respectively. Fig.3 illustrates the local part model. To implement fast computation of LPM-based feature, two integral videos are computed, one for the root filter and the other for the part filters at twice spatial resolution. In our experiments, one root filter and eight parts filter are used in the local part model, and the corresponding descriptors are concatenated into one vector, which is then nine times of the original MBH descriptor.

2.3. Random Sampling and Bag-of-Features

Random sampling has been applied to real-time action recognition and leads to excellent performances [21]. Given a video, a very dense spatio-temporal grid is used for feature

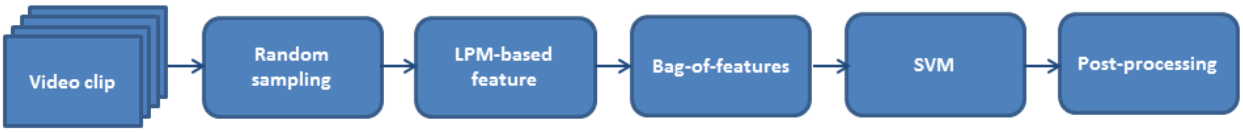

Fig. 2. The flowchart of the proposed approach for event detection.

pooling. Each spatio-temporal point is associated with a local 3D cuboid. In our experiments, multiple scale cuboid are used and the consecutive scales are generated by scale factor of 2. To obtain high density sampling, the intersection between two neighboring cuboids is set at 80%. For example, there are total of 87,249 spatio-temporal points generated for a video with resolution of 182x120 and 95 frames. For the selected 3D points, the LPM-based features are computed for the corresponding cuboids.

A standard bag-of-feature approach [20] is used to represent events contained in videos. We use the k-means clustering algorithm [26] to group a subset of randomly selected LPM-based features, and construct a dictionary of visual words. We fix the number of visual words in the experiment. For a video clip, the obtained LPM-based features are assigned to their nearest visual words using Euclidean distance, and the obtained histogram of visual word occurrences is used for event recognition. High dimensional feature generally lead to higher accuracy. To achieve a trade-off between accuracy and matching efficiency, principal component analysis (PCA) [27] is used for dimension reduction. In our case, the root filter and part filters highly overlap, and PCA is helpful for removing redundant information.

2.4. Detection

For the purpose of event detection, we train non-linear SVMs with RBF-χ^2 kernel [28] given as follows

$$K(\mathbf{x}, \mathbf{y}) = \exp(-D(\mathbf{x}, \mathbf{y})),$$
$$D(\mathbf{x}, \mathbf{y}) = 1 - \sum_{n=1}^{N} \frac{(x_n - y_n)^2}{\frac{1}{2}(x_n + y_n)}. \quad (1)$$

Since our object is to detect seven types of events in surveillance videos, we train seven models using LIBSVM [29] by using one-versus-rest approach. The videos in the Gatwick dataset are used for training model. For each query video clip, a bag-of-features representation is build according to the method described in the above sub-sections, and fed into the pre-trained SVMs. Each of these SVM responses indicates how likely the query contains a specific event. To make decision, we first use an empirical threshold to remove most candidates. For each candidate passing the first filtering stage, if there is a peak maximum response centered in a temporal window, we consider that a specific type of event has occurred.

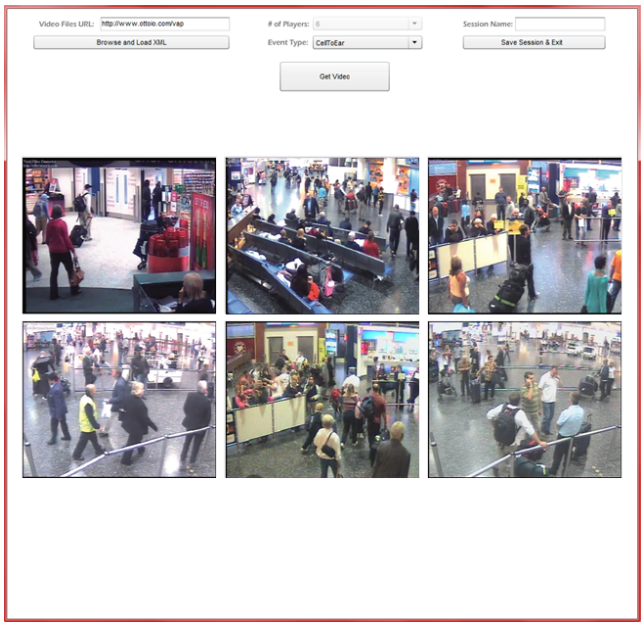

Fig. 4. The interface of the event viewer tool.

2.5. Summary

The proposed approach for event detection is summarized as follows.

1: Specify event E.
2: Input query video V.
3: Split V into clips $\{V_1, V_2, \ldots, V_K\}$.
4: **for** $k = 1$ to K **do**
5: Randomly select a set of spatio-temporal points in V_k .
6: Compute LPM-based features from cuboids associated with obtained 3D points.
7: Generate bag-of-features representation.
8: Calculate response of SVM associated with E.
9: **end for**
10: Generate candidates by thresholding SVM responses.
11: Filter candidates by local maximum filtering.

3. FALSE POSITIVE REMOVAL USING EVENT VIEWER TOOL

To further remove false positives in the results returned by the proposed event detection approach described in the previous

section, an event viewer tool as shown in Fig.4 is developed for visual and manual filtering. The event viewer tool is designed to allow a user to quickly and accurately flag or reject video clips for the occurrence of specific types of events. After specifying the number of simultaneous players and the event type, the user can mouse over each clip playing on a loop and select REJECT if the specified event is not present, or FLAG if the event occurs. Once a REJECT/FLAG decision has been made on each of currently displayed video clips, more sets of video clips load and play on a loop until all clips have been inspected.

Having more videos open at once may allow the user to observe human activities in several clips at the same time, which may be of benefit to greater efficiency. On the other hand, as more players play videos simultaneously, the view area for each video diminishes, which may lead to loss of details. Therefore, it makes sense to find an appropriate number of players in terms of both viewing speed and flagging accuracy. Students participated in testing the event viewer tool with different number of players for flagging the seven types of events defined in the TRECVID SED task.

4. EXPERIMENTAL RESULTS

The proposed event detection approach is trained on the Gatwick dataset. As described above, LPM-based features are used to train seven SVMs with RBF-χ^2 kernel for the seven specific types of events using one-versus-rest strategy respectively. We use sliding spatio-temporal volumes to detect likely events in the evaluation videos. The final results are generated from SVM response thresholding and local maximum filtering.

4.1. Dataset

The Gatwick dataset contains videos captured by five surveillance cameras installed at the London Gatwick airport at a resolution of 720×480 at 25 frames per second. Specifically, the first camera observes passengers entering and exiting a door, the second and third cameras observe two waiting areas, the fourth camera observes two elevator doors, and the fifth camera observes multiple passenger channels. These scenes involve high density crowd and contain significant occlusion amongst people. There are several hours of video taken by each camera. The videos are annotated for seven specific types of events, such as CellToEar, Embrace, ObjectPut, PeopleMeet, PeopleSplitUp, PersonRuns, and Pointing. The dataset is composed of two subsets for development and evaluation. The proposed event detection approach is trained on selected portions of the development videos. For each type of event, the annotated video clips are used as positive samples, both the video clips annotated for other types of event and 1500 video clips without containing any specific type of event are used as negative samples, and a non-linear SVM is trained on the obtained training set.

4.2. Event Detection Results

There are a few parameters in the proposed approach, and we determine them by analyzing training videos. Specifically, we resize the spatial resolution of training and test videos to 0.25 times of the original resolution to make the computation manageable while maintain an adequate sampling density. For the purpose of event detection in long videos, we slide a spatio-temporal volume inside the original resolution of the video. The volume has a temporal size of 50 frames with intervals of 10 frames. The minimal spatial and temporal sizes of 3D cuboid are set to 16×16 pixels and 10 frames respectively. There are total 8 spatial scales and 2 temporal scales. To capture motion information in each cuboid, we set one root filter and eight $(2 \times 2 \times 2)$ partially overlapping part filters in both spatio and temporal directions. For bag-of-features representation, we randomly selected $120,000$ LPM-based features extracted from the training set, and used k-means to cluster these features to produce $6,000$ visual words. The original LPM-based feature has 1152 dimensions which is 9 times of the MBH descriptor. To improve matching efficiency, we use PCA to project LPM-based feature to a sub-space of 96 dimensions.

The final evaluation results of the proposed event detection approach we got from TRECVID are shown in Table I and Fig.5. The rates of false alarms (RFA) for all the events are high, which may result from the fact that there is no bootstrapping stage included in model training. For the events of Embrace, PeopleMeet, PeopleSplitUp and PersonRuns, the percentages of missed detections (PMiss) are much smaller than that of CellToEar, ObjectPut and Pointing while the values of RFA are close. A possible reason for such a situation may be that the random sampling strategy used for bag-of-features is more suitable for capturing motion information in the case of specific human activities occupying more spatio-temporal region in the detection window. Therefore, a possible solution would be to use a detection window with a lower spatial resolution sliding in both spatial and temporal directions.

4.3. Test Result of Event Viewer Tool

In this experiment, we compare the results of using 1, 2, 4, 6 and 9 video players in the event viewer tool for flagging the video clips. The seven types of events defined in the TRECVID SED task were all tested. All of the video clips used were cut from the Gatwick dataset. The number of players and the event type were randomly chosen for each participant. Before the test, 5 sample videos containing the event were shown to each participant in order to show typical manifestations of the event, and then 25 randomly select videos were shown for training. In the test phase, 200 video clips

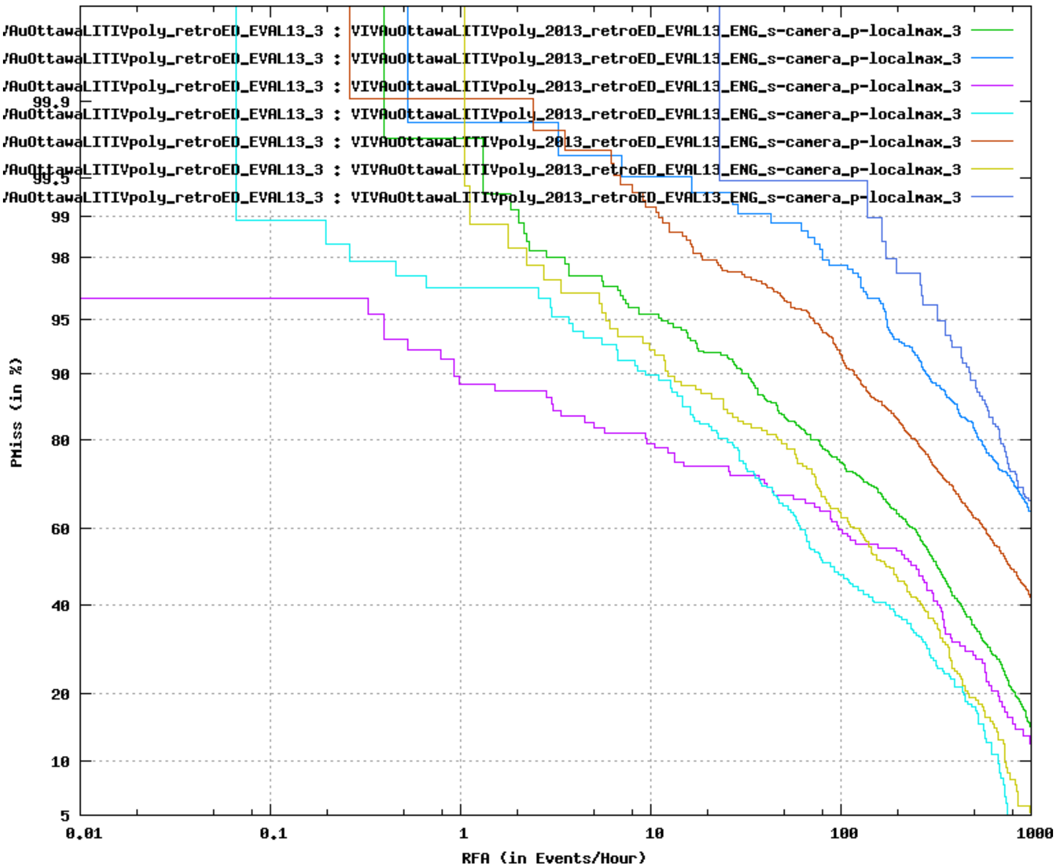

Fig. 5. DET curve of the proposed approach for the seven types events in the TRECVID SED task.

Table 1. Evaluation results of the proposed approach for the seven types of events in the TRECVID SED task.

Event	#Targ	#NTarg	#Sys	#CorDet	#FA	#Miss	RFA	PMiss	DCR
CellToEar	194	27041	27173	62	14135	132	927.05	0.680	5.3157
Embrace	175	27001	27173	166	15820	9	1037.57	0.051	5.2393
ObjectPut	621	26736	27173	262	17773	359	1165.65	0.578	6.4064
PeopleMeet	449	26752	27173	420	24620	29	1614.72	0.065	8.1382
PeopleSplitUp	187	26987	27173	185	18149	2	1190.32	0.011	5.9623
PersonRuns	107	27067	27173	98	17467	9	1145.59	0.084	5.8120
Pointing	1063	26330	27173	789	23825	274	1562.58	0.258	8.0707

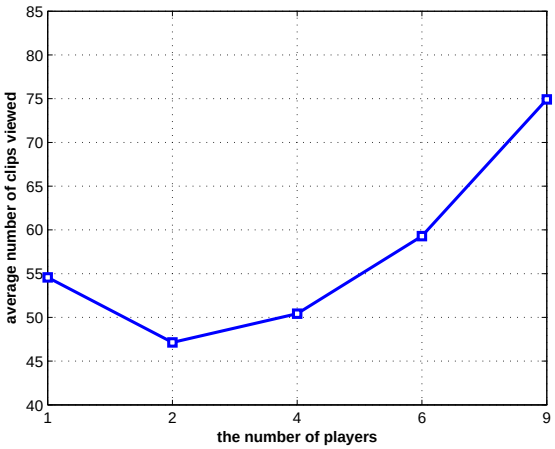

Fig. 6. The average number of clips viewed.

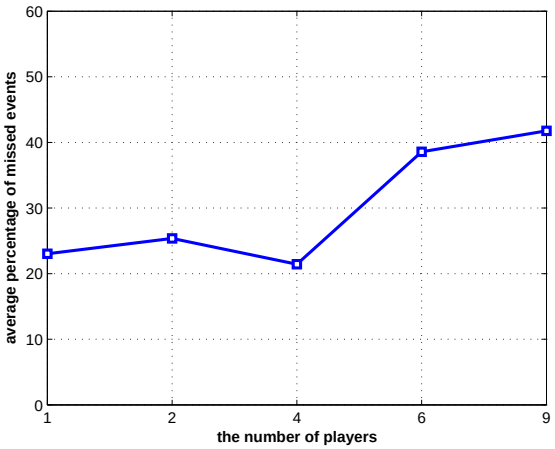

Fig. 7. the mean rate of missed events.

(5% of these contained the specific events) were randomly chosen, and the participant was required to flag the clips as much as possible within 10 minutes. Figs.6 and 7 show the average number of clips viewed and the mean percentage of missed events (the percentage was calculated by taking the number of events that were failed to be flagged, and dividing by the total number of clips that were viewed and contained an event). When 9 players were used, the average number of viewed videos is the highest, however the corresponding miss rate is also the highest. Given these results, we use 6 players in the event viewer tool which allow the user to have access to more simultaneously playing clips without losing too much detail.

5. CONCLUSION

This paper presents an approach for detecting seven types of events such as CellToEar, Embrace, objectPut, PeopleMeet, PeopleSplitUp, PersonRun, and Pointing defined by TRECVID SED task. To represent events, a local part model is employed to extract features from local spatio-temporal cuboids and random sampling strategy is used to produce bag-of-features representation. Seven pre-trained SVMs with RBF-χ^2 kernel are used for classification. On the other hand, a event viewer tool is developed for further filtering the results by manually and efficiently removing false positives. The evaluation results demonstrate that the performance of the proposed approach in detecting Embrace, PeopleMeet, PeopleSplitUp and PersonRuns is better than that in detecting CellToEar, ObjectPut and Pointing. In addition, a bootstrapping stage should be included in model training, which would be of benefit to reducing false positive rate.

6. REFERENCES

[1] Sreemanananth Sadanand and Jason J Corso, "Action bank: A high-level representation of activity in video," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1234–1241.

[2] Michael Sapienza, Fabio Cuzzolin, and Philip Torr, "Learning discriminative space-time actions from weakly labelled videos," in *Proceedings of European Conference on Computer Vision*, 2012, pp. 1–12.

[3] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajecto-

ries," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3169–3176.

[4] Hasan Mahmudul, Zhu Yingying, Sunderrajan Santhoshkumar, Pourian Niloufar, Manjunath B.S., and Chowdhury A.R., "Activity analysis in unconstrained surveillance videos," in *TRECVID Technical Report*, 2012.

[5] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, and Georges Quenot, "Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2013*. NIST, USA, 2013.

[6] Alexei A Efros, Alexander C Berg, Greg Mori, and Jitendra Malik, "Recognizing action at a distance," in *Proceedings of IEEE International Conference on Computer Vision*, 2003, pp. 726–733.

[7] Wei-Lwun Lu and James J Little, "Simultaneous tracking and action recognition using the pca-hog descriptor," in *Proceedings of Canadian Conference on Computer and Robot Vision*, 2006, pp. 1–6.

[8] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005, pp. 65–72.

[9] Frederic Jurie and Bill Triggs, "Creating efficient codebooks for visual recognition," in *Proceedings of IEEE International Conference on Computer Vision*, 2005, vol. 1, pp. 604–610.

[10] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of International Conference on Pattern Recognition*, 2004, vol. 3, pp. 32–36.

[11] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, Cordelia Schmid, et al., "Evaluation of local spatio-temporal features for action recognition," in *Proceedings of British Machine Vision Conference*, 2009, pp. 1–11.

[12] Juan Carlos Niebles and Li Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[13] Ming-yu Chen and Alexander Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos," in *Technical Report CMU-CS-09-161*. Carnegie Mellon University, 2009.

[14] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[15] Chris Whiten, Robert Laganiere, and Guillaume-Alexandre Bilodeau, "Efficient action recognition with MoFREAK," in *Proceedings of International Conference on Computer and Robot Vision*, 2013, pp. 319–325.

[16] Geert Willems, Tinne Tuytelaars, and Luc Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proceedings of European Conference on Computer Vision*, 2008, pp. 650–663.

[17] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *Proceedings of European Conference on Computer Vision*, 2006, pp. 404–417.

[18] Eric Nowak, Frédéric Jurie, and Bill Triggs, "Sampling strategies for bag-of-features image classification," in *Proceedings of European Conference on Computer Vision*, 2006, pp. 490–503.

[19] Feng Shi, Emil M Petriu, and Albino Cordeiro, "Human action recognition from local part model," in *Proceedings of IEEE International Workshop on Haptic Audio Visual Environments and Games*, 2011, pp. 35–38.

[20] Li Fei-Fei and Pietro Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 2, pp. 524–531.

[21] Feng Shi, Emil Petriu, and Robert Laganiere, "Sampling strategies for real-time action recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2595–2602.

[22] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[23] Navneet Dalal, Bill Triggs, and Cordelia Schmid, "Human detection using oriented histograms of flow and appearance," in *Proceedings of European Conference on Computer Vision*, 2006, pp. 428–441.

[24] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.

[25] Pedro Felzenszwalb, David McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[26] James MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Proceedings of Berkeley Symposium on Mathematical Statistics and Probability*. California, USA, 1967, vol. 1, p. 14.

[27] Ian Jolliffe, *Principal component analysis*, Wiley Online Library, 2005.

[28] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[29] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27, 2011.