Snatch Theft Detection in Unconstrained Surveillance Videos Using Action Attribute Modelling

Article *in* Pattern Recognition Letters · March 2018

DOI: 10.1016/j.patrec.2018.03.004

CITATIONS
2

READS
368

2 authors:

Debaditya Roy
Nihon University
14 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Krishna Mohan Chalavadi
Indian Institute of Technology Hyderabad
57 PUBLICATIONS **351** CITATIONS
SEE PROFILE

Snatch Theft Detection in Unconstrained Surveillance Videos Using Action Attribute Modelling

Debaditya Roy[a,*], C. Krishna Mohan[a]

[a] *Visual Learning and Intelligence Group (VIGIL)*
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad

Abstract

In a city with hundreds of cameras and thousands of interactions daily among people, manually identifying crimes like chain and purse snatching is a tedious and challenging task. Snatch thefts are complex actions containing attributes like walking, running etc. which are affected by actor and view variations. To capture the variation in these attributes in diverse scenarios, we propose to model snatch thefts using a Gaussian mixture model (GMM) with a large number of mixtures known as universal attribute model (UAM). However, the number of snatch thefts typically recorded in a surveillance videos is not sufficient enough to train the parameters of the UAM. Hence, we use the large human action datasets like UCF101 and HMDB51 to train the UAM as many of the actions in these datasets share attributes with snatch thefts. Then, a super-vector representation for each snatch theft clip is obtained using maximum aposteriori (MAP) adaptation of the universal attribute model. However, super-vectors are high-dimensional and contain many redundant attributes which do not contribute to snatch thefts. So, we propose to use factor analysis to obtain a low-dimensional representation called action-vector that contains only the relevant attributes. For evaluation, we introduce a video dataset called Snatch 1.0 created from many hours of surveillance footage obtained from different traffic cameras placed in the city of Hyderabad, India. We show that using action-vectors snatch thefts can be better identified than existing state-of-the-art feature representations.

Keywords: Gaussian mixture model, Factor Analysis, Action Recognition, Surveillance, Snatch Theft Detection

1. Introduction

Abnormal events of interest like thefts and threats in long video sequences like surveillance footage have an extremely low probability of occurrence. Manually detecting these rare events or anomalies is challenging in cities as there are hundreds of cameras which need to be monitored. Especially anomalies have localized spatio-temporal signatures where they occur over a small time window in a long sequence or a small spatial region in a wide surveillance area. The distinguishing feature of these scenarios is that outside this anomalous spatio-temporal region, regular activities are observed. Anomalous activities like chain and purse snatching are especially prevalent in many countries.

Most of the existing literature on detecting anomalous activities like snatching uses datasets collected in controlled laboratory settings with no crowd or background and with an excellent viewing angle of the activity. Even when conducted in crowded scenes as in [1], the entire pickpocket incident is staged with *apriori* knowledge of how the incident is going to take place which makes anal-

*Corresponding author: Tel.: +91-949-287-5174; fax: +91-040-2301-6032;

Email addresses: cs13p1001@iith.ac.in (Debaditya Roy), ckm@iith.ac.in (C. Krishna Mohan)

URL: https://sites.google.com/view/debadityaroy/ (Debaditya Roy)

ysis a lot easier. So, to analyze real-life thefts, we present a dataset called Snatch 1.0 [1] collected from unconstrained surveillance footage. This dataset contains surveillance footage obtained from the traffic police department of Hyderabad city of India which includes various instances of snatch thefts (details in Section 4.1). It was observed that snatching incidents in surveillance videos can occur in a variety of *scenarios* which are of diverse types and lead to different victim *reactions*. Some of the examples of snatch thefts are shown in Figure 1.

Table 1: Some cases of snatching scenarios, types and victim reactions

	Snatch Type 1 (Direct grab)	**Snatch Type 2 (Inquire and grab)**
Scenario 1 (Thief on motorbike)	Reaction 1 (Victim chases)	Reaction 1 (Victim chases)
	Reaction 2 (Victim dragged)	
	Reaction 3 (Victim falls or remains standing)	Reaction 3 (Victim remains standing)
Scenario 2 (Thief on foot)	Reaction 1 (Victim chases)	Reaction 1 (Victim chases)
	Reaction 3 (Victim falls or remains standing)	Reaction 3 (Victim falls or remains standing)

In Table 1, we list some cases of snatch thefts encountered in surveillance videos used for the present work. It is evident that snatch thefts are not only complex to model but also the definition of a snatch theft itself is non-trivial. Each interaction between individuals needs to be studied to decide whether or not it is a potential snatch theft or not. These interactions can be considered as part of the larger set of human actions [2]. Hence, we propose a framework for analyzing these interactions. At first, an unsupervised Gaussian mixture model called universal attribute model (UAM) is trained using a variety of human actions containing attributes like *punching* in HMDB51 and UCF101 datasets which is visually similar to *grabbing* in snatch thefts as shown in Figure 2. Gaussian mixture models have previously been explored to model attributes of actions [3, 4]. Using factor analysis, the essential attributes useful for describing snatch thefts are extracted and represented in the form of action-vectors. We show that action-vectors perform better than existing state-of-the-art feature descriptors while leveraging a lot of existing video data containing human actions to effectively represent snatch thefts.

The rest of the paper is arranged as follows. Section 2 discusses the related literature for anomaly detection in surveillance videos with a focus on anomalous activities. In Section 3, a detailed description of the proposed snatch theft detection framework is presented. The results and related analysis are reported in Section 4 and we conclude with directions for future work in Section 5.

2. Related Work

A majority of existing literature in the field of anomaly detection is aimed towards detection of generalized abnormal patterns [5, 6] or behaviour in case of individuals or crowds [7]. Many approaches model normal behaviour extensively and consider events which do not follow these models as anomalous activities. One approach [8] proposed a motion-influence map which localizes both global and local unusual activities. The motion-influence map considers the speed and direction of various objects to determine their relative influence on other objects for detecting unusual motion patterns. In [9], apart from magnitude and direction of motion, entropy information was further added to form a combined histogram of flow orientation, motion, and entropy (HOFME) descriptor. The usage of entropy was to determine the density of motion during normal events. In [10], a roadside surveillance scene was divided into zones like traffic lanes, stationary areas, etc. each of which was termed as a scene context and the direction and flow of persons in each of these contexts were measured. Further, the interaction between any two individuals in the close spatial vicinity was measured for gaze and motion direction information. This was termed as social context and revealed normal behaviour in different scene contexts.

Recently, deep learning methods have also been used for anomaly detection in videos. In [11], appearance and motion features were learned using two stacked denoising auto-encoders (AE) trained on patches extracted from each image in the video and corresponding optical flow map, respectively. A fusion of the two AE outputs with a one-class SVM was used for learning normal appearance and motion. For extracting both appearance and motion features simultaneously, in [12], 3D convolutional layers

[1] http://www.iith.ac.in/vilg/datasets/

Figure 1: Different snatching scenarios as captured in Snatch 1.0. Best viewed in colour.

(a) *grab* (b) *punch*

Figure 2: Visual similarity in (a) *grabbing* of Snatch 1.0 and (b) *punching* of HMDB51 dataset.

were added to the AE network above. Instead of training a one-class SVM, the reconstruction error was directly used for predicting anomalies. A variant of AE involving 2D CNN for appearance features and a long-short memory (LSTM) was introduced in [13] for motion modelling. The deep learning methods described above can register only significant deviations from normal behaviour in constrained environments and hence cannot be used for recognizing snatch thefts in unconstrained environments which vary only subtly from regular activities.

These approaches presented above assume that the testing scenarios will be same as training scenes and enough labelled data is available to learn the sequences. However, snatch thefts in real-world scenarios are rarely similar and labelled examples are scarcely available. A crime like snatching can be characterized as having the following characteristics: 1) snatching activity occurs much more infrequently than regular interactions, 2) many actions which contribute to snatching do not have significantly different characteristics from normal activities like grabbing, running, etc. Further, the infrequency of snatching events mean that the most of the completely supervised methods where both the training and testing set should contain large labelled data of abnormal patterns cannot be employed. Similarly, the closeness to normal activities makes the use of completely unsupervised methods unsuitable [14].

3. Proposed Framework

In the proposed approach we exploit the similarity of snatch thefts to normal actions to our advantage by training a large UAM which encompasses attributes across all actions. So, the training of the UAM is not dependent on the availability of labelled snatch theft examples which are difficult to obtain as they are infrequent activities. Next, we describe UAM construction in detail.

3.1. Universal Attribute Model (UAM)

Each action clip can be considered to be a sample function which realizes the random process generating the action. For estimating the sampling function, we need the

parameters of the *pdf* which describes the random process. If such a *pdf* can be estimated using a GMM, then the number of mixtures must be sufficiently large to accommodate the intra-action variances encountered in different recording conditions. We call this model the universal attribute model (UAM) which can be represented as $p(\mathbf{x}_l) = \sum_{c=1}^{C} w_c \mathcal{N}(\mathbf{x}_l|\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c)$ where the mixture weights w_c satisfy the constraint $\sum_{c=1}^{C} w_c = 1$ and $\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c$ are the mean and covariance for mixture c of the UAM, respectively. A feature $\mathbf{x}_l$ is part of a clip $\mathbf{x}$ represented as a set of feature vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L$. This feature can be histogram of optical flow (HOF) [15] or motion boundary histogram (MBH) [15] descriptor and we train the UAM using standard EM estimation. Owing to the large number of mixtures in the UAM, a small number of snatch theft examples are not enough to train the UAM. Hence, we use the clips from the training split of large human action datasets like UCF101 [16] and HMDB51 [2] containing 101 and 51 actions, respectively, to train the UAM.

As the goal is to find the *pdf* of the action that generates a clip, we need to adapt the UAM parameters using the data in the clip [17, 18]. The UAM parameters are adapted for every clip to enhance the contribution of the attributes present in it. Given L feature vectors of a clip $\mathbf{x}$, the probabilistic alignment of these feature vectors into each of the C mixture components of the UAM is calculated as a posterior $p(c|\mathbf{x}_l)$ which is computed as

$$p(c|\mathbf{x}_l) = \frac{w_c p(\mathbf{x}_l|c)}{\sum_{c=1}^{C} w_c p(\mathbf{x}_l|c)}, \tag{1}$$

where $\mathbf{x}_l$ is a $d \times 1$ feature vector and $p(\mathbf{x}_l|c)$ is the likelihood of a feature $\mathbf{x}_l$ arriving from a mixture c.

The posterior probability is used to calculate the zeroth and first order Baum-Welch statistics for a clip $\mathbf{x}$ as $n_c(\mathbf{x}) = \sum_{l=1}^{L} p(c|\mathbf{x}_l)$ and $\mathbf{F}_c(\mathbf{x}) = \left(\sum_{l=1}^{L} p(c|\mathbf{x}_l)\mathbf{x}_l\right) n_c(\mathbf{x})$, respectively.

The MAP adapted parameters of a clip-specific model can be obtained as a convex combination of the UAM and the clip-specific statistics. For every mixture c of the UAM, the adapted weights and means are calculated as

$$\hat{w}_c = \alpha n_c(\mathbf{x})/L + (1-\alpha) w_c \tag{2a}$$

and

$$\hat{\boldsymbol{\mu}}_c = \alpha \mathbf{F}_c(\mathbf{x}) + (1-\alpha)\boldsymbol{\mu}_c. \tag{2b}$$

The adapted means for each mixture are then concatenated to compute a $(Cd \times 1)$-dimensional SAV for each clip represented as $\mathbf{s}(\mathbf{x}) = [\hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_2 \cdots \hat{\boldsymbol{\mu}}_C]^t$. Obtaining a fixed-dimensional representation like the super action-vector normalizes the effect of varying length clips but results in a high-dimensional representation. This representation though contains many of the attributes that do not contribute to the clip and hence are not changed from the original UAM. Since each clip contains only a few of the total UAM mixtures (attributes), only those means are modified. Hence, the SAV is intrinsically low-dimensional, and by using a suitable decomposition, we can extract such a representation which we refer to as an action-vector.

3.2. Action-vector representation

In order to arrive at a low-dimensional representation, the super-action vector $\mathbf{s}$ is decomposed as

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{3}$$

where $\mathbf{m}$ is the supervector that is actor and viewpoint independent (can be assumed to be the un-adapted UAM supervector), $\mathbf{T}$ is a low-rank rectangular matrix known as the total variability matrix of size $Cd \times r$, and a r-dimensional action-vector $\mathbf{w}$ whose prior distribution is assumed to be a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [19]. The posterior distribution of the action-vector after observing a clip $\mathbf{x}$ as

$$P(\mathbf{w}|\mathbf{x}) \propto P(\mathbf{x}|\mathbf{w})\mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

$$\propto \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{L}(\mathbf{x}))^t \mathbf{M}(\mathbf{x})(\mathbf{w} - \mathbf{L}(\mathbf{x}))\right), \tag{5}$$

where $\boldsymbol{\Sigma}$ is a diagonal covariance matrix of dimension $Cd \times Cd$ and it models the residual variability not captured by the total variability matrix $\mathbf{T}$. The matrix $\mathbf{L}(\mathbf{x}) = \mathbf{M}^{-1}(\mathbf{x})\mathbf{T}^t\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{s}}(\mathbf{x})$ where $\tilde{\mathbf{s}}(\mathbf{x})$ is the centered supervector which appears because the posterior distribution of $\mathbf{w}$ is conditioned on the Baum-Welch statistics of the clip centered around the means of the UAM. The first order Baum-Welch statistics centered around the UAM mean can be obtained as $\tilde{\mathbf{F}}_c(\mathbf{x}) = \sum_{l=1}^{L} p(c|\mathbf{x}_l)(\mathbf{x}_l - \boldsymbol{\mu}_c)$.

We can now express $\tilde{\mathbf{s}}(\mathbf{x})$ as the concatenated first-order statistics $\tilde{\mathbf{s}}(\mathbf{x}) = [\tilde{\mathbf{F}}_1(\mathbf{x})\tilde{\mathbf{F}}_2(\mathbf{x}) \cdots \tilde{\mathbf{F}}_C(\mathbf{x})]^t$. Also, the matrix $\mathbf{M}(\mathbf{x}) = \mathbf{I} + \mathbf{T}^t\boldsymbol{\Sigma}^{-1}\mathbf{N}(\mathbf{x})\mathbf{T}$ where $\mathbf{N}(\mathbf{x})$ is a diagonal matrix

of dimension $Cd \times Cd$ whose diagonal blocks are $n_c(\mathbf{x})\mathbf{I}$, for $c = 1, ..., C$ and $\mathbf{I}$ is the identity matrix of dimension $d \times d$.

From Equation 5, the mean and covariance matrix of the posterior distribution are given by

$$E[\mathbf{w}(\mathbf{x})] = \mathbf{M}^{-1}(\mathbf{x})\mathbf{T}^t\mathbf{\Sigma}^{-1}\tilde{\mathbf{s}}(\mathbf{x}) \tag{6a}$$

and

$$\text{Cov}(\mathbf{w}(\mathbf{x}), \mathbf{w}(\mathbf{x})) = \mathbf{M}^{-1}(\mathbf{x}), \tag{6b}$$

respectively. Using EM algorithm [20], we iteratively estimate the posterior mean and covariance in the E-step and use the same to update $\mathbf{T}$ and $\mathbf{\Sigma}$ in the M-step.

In the first E-step of the estimation, $\mathbf{m}$ and $\mathbf{\Sigma}$ are initialized with the UAM mean and covariance, respectively. For the total variability matrix $\mathbf{T}$, a desired rank r is chosen, and the matrix is initialized randomly. Then $E[\mathbf{w}(\mathbf{x})]$ and $\text{Cov}(\mathbf{w}(\mathbf{x}), \mathbf{w}(\mathbf{x}))$ calculated according to Equations 6a & 6b.

In the M-step, the matrix $\mathbf{T}$ is calculated as the solution of $\sum_{\mathbf{x}} \mathbf{N}(\mathbf{x})\mathbf{T}E[\mathbf{w}(\mathbf{x})\mathbf{w}^t(\mathbf{x})] = \sum_{\mathbf{x}} \tilde{\mathbf{s}}(\mathbf{x})E[\mathbf{w}^t(\mathbf{x})]$ which results in a system of r linear equations. For each $c = 1, \cdots, C$, the residual matrix $\mathbf{\Sigma}$ is estimated mixture by mixture as $\mathbf{\Sigma}_c = \left(\sum_{\mathbf{x}} \tilde{S}_c(\mathbf{x}) - \mathbf{M}_c\right)/n_c(\mathbf{x})$ where $\mathbf{M}_c$ denotes the c^{th} diagonal block of the $Cd \times Cd$ matrix $1/2 \sum_{\mathbf{x}} \tilde{\mathbf{s}}(\mathbf{x})E[\mathbf{w}^t(\mathbf{x})]T^t + TE[\mathbf{w}(\mathbf{x})]\tilde{\mathbf{s}}^t(\mathbf{x})$ and $\tilde{\mathbf{S}}_c(\mathbf{x})$ is the second-order Baum-Welch statistics of the clip centered on the means of the UAM calculated as $\tilde{\mathbf{S}}_c(\mathbf{x}) = diag\left(\sum_{l=1}^{L} p(c|\mathbf{x}_l)(\mathbf{x}_l - \boldsymbol{\mu}_c)(\mathbf{x}_l - \boldsymbol{\mu}_c)^t\right)$.

After the final M-step i.e. estimation of $\mathbf{T}$ and $\mathbf{\Sigma}$ matrices, the action-vector for a given clip can be represented using the mean of its posterior distribution as

$$\mathbf{w}(\mathbf{x}) = (\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}\mathbf{N}(\mathbf{x})\mathbf{T})^{-1}\mathbf{T}^t\mathbf{\Sigma}^{-1}\tilde{\mathbf{s}}(\mathbf{x}). \tag{7}$$

This process of obtaining the action-vector is known as factor analysis [20]. This action-vector can now be used for training classifiers in order to detect snatch thefts.

4. Experimental protocols and Results

4.1. Snatch 1.0 : Dataset Description

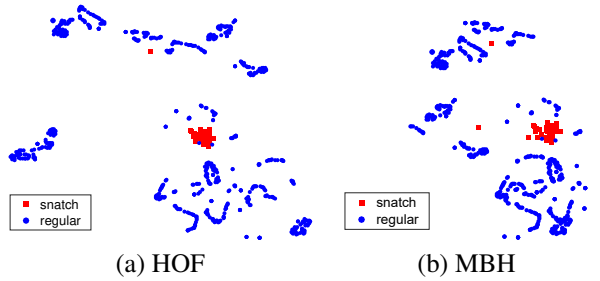
The videos obtained from the surveillance cameras are low-resolution in most surveillance setups. In cases of wide area surveillance, events of interest can occur further away from the camera because of which person detection becomes more challenging. As snatch theft incidents occur infrequently, we could only obtain a total of 35 chain snatch theft incidents after searching through the archived video footage of over six months from different surveillance cameras placed in the city of Hyderabad, India. The snatch thefts were spread within 4.5 hours of surveillance footage containing 37485 regular interactions. It was observed that snatch thefts are 4-5 seconds in duration, so we divided the entire surveillance footage into 10-second clips which resulted in a total of 816 clips.

4.2. Effect of different feature descriptors and UAM mixtures

In this work, action-vectors are formed using two state-of-the-art feature descriptors namely, HOF and MBH which are obtained as part of the improved dense trajectory set of features [15]. In Table 2, the classification performance of these action-vectors is presented for varying number of UAM mixtures. Also, different classifiers like ensemble subspace discriminant analysis (ESDA), k nearest neighbours (k-NN), and support vector machines (SVM) are used for evaluation. For each of the classifiers, 3-fold cross-validation is used. The dimension for the action-vector referred to as r in the previous section is fixed to be 200 as it is found that varying the action-vector dimension does not yield any change in classification performance. It can be observed that action-vector performs consistently across all the settings making it an effective representation. Further, even smaller UAM with less number of mixtures can help in efficien representing snatch thefts leading to proper classification. In Figure 3, the t-SNE plot of the action-vectors with HOF and MBH features using 256 UAM mixtures is shown where clear separability can be noticed between the regular interactions and snatch thefts.

Table 2: Action-vector classification performance (in %) for Snatch 1.0.

Classifier	**Number of UAM mixtures**					
	HOF			**MBH**		
	256	**512**	**1024**	**256**	**512**	**1024**
ESDA	99.2	99.3	99.3	98.3	99.4	99.4
k-NN	99.5	99.4	99.4	99.7	99.5	99.5
SVM	99.7	99.7	**99.8**	**99.8**	99.6	99.4



(a) HOF (b) MBH

Figure 3: Action-vectors for Snatch 1.0 using (a) HOF features and (b) MBH features, using 256 UAM mixtures. Best viewed in color.

4.3. Comparison with state-of-the-art feature descriptors

Action-vectors are compared against recent state-of-the-art feature descriptors for describing human actions: a) HOFM (histogram of optical flow and magnitude) [9] and b) 3D convolutional neural network (3DCNN) features [21]. For HOFM, we use the parameter settings as per [9] where a regular grid of size $30 \times 30 \times 3$ is created. The first two dimensions correspond to grid cell dimensions (width and height) in space, and the third dimension corresponds to the depth in time. For 3DCNN, a pre-trained network trained on the sports 1M dataset as per [21] was used to obtain the features. Each 3DCNN features summarizes the information in 16 frames into a single descriptor and resulting in 20 feature descriptors for every 10-second clip used in our experiments. The classification outputs of these 20 descriptors are then combined using majority voting to produce the final classification output for the clip.

From Table 3, it can be observed that the proposed framework misses only 1 snatch theft as compared 5, 20, and 24 and to the other features. In terms of both accuracy and area under the curve (AUC), action-vector representation outperforms other features as shown in Figure 4. Also, when action-vectors calculated using MBH features are compared to MBH features, a significant improvement can be observed. This shows that action-vectors can extract meaningful information from existing descriptors to produce even more discriminative representations.

We present some of the detected snatch theft scenarios in Figure 5 according to the scenarios explained in Table 1. It can be observed that action-vectors recognize a diverse set of snatch thefts which have little similarity to each other. However, a few false positive cases are also

Table 3: Comparison with state-of-the-art feature descriptors using SVM classfier.

Method	**Missed Snatches**	**Accuracy** (in %)	**AUC** (in %)
HOFM [9]	24	47.6	69.8
MBH [15]	20	59.3	72.4
3DCNN [21]	5	96.6	98.3
action-vector (HOF)	**1**	**99.8**	**99.8**
action-vector (MBH)	**1**	**99.8**	**99.9**

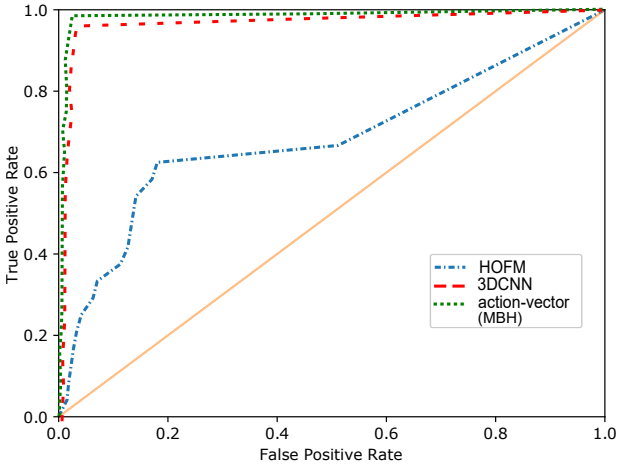


Figure 4: ROC for performance comparison with state-of-the-art. Best viewed in colour.

encountered where regular interactions are identified as snatch thefts are shown in Figure 6. Such cases are difficult to address without identifying and tracking the thief's limbs.

5. Conclusion

In this paper, we presented a framework for snatch theft detection in unconstrained videos using action attribute modelling. To learn all the action attributes in the snatch thefts, a large GMM called universal attribute model (UAM) was trained using existing video datasets of human actions. The means of the UAM were adapted to obtain a high dimensional super action-vector for each snatch theft clip. To remove redundant attributes, factor analysis was used to obtain a low-dimensional action-vector. For evaluation, we introduced a dataset called Snatch 1.0 that contains snatch thefts in surveillance

(a) **Scenario 1, Snatch Type 1 and Reaction 1**

(c) **Scenario 2, Snatch Type 2 and Reaction 1**

Figure 5: Detection results of snatch theft using the proposed framework. For visualization, a YOLO [22] detector and a kernelised correlation filter [23] based tracker was applied on each of the detected snatch theft clips. Bounding boxes are drawn (in green) with the *id* (in blue) on top of the box for different persons. Best viewed in colour.

(a) (b)

Figure 6: False positive cases where normal interactions detected as snatch thefts

videos. It was shown that action-vector provides better discriminative representation for snatch thefts than existing state-of-the-art feature descriptors. In future, we would like to work on live streams of surveillance and generate real-time alerts for a smart monitoring system which can immediately warn the security personnel at the surveillance site for possible snatch thefts in the area.

References

[1] H. Bouma, J. Baan, G. J. Burghouts, P. T. Eendebak, J. R. van Huis, J. Dijk, J. H. van Rest, Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall, in: SPIE Security+ Defence, International Society for Optics and Photonics, 2014, pp. 925–933.

[2] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in: Proceedings of the International Conference on Computer Vision (ICCV), 2011.

[3] Z. Cai, L. Wang, X. Peng, Y. Qiao, Multi-view super vector for action recognition, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 596–603.

[4] Z. Wang, Y. Wang, L. Wang, Y. Qiao, Codebook enhancement of vlad representation for visual recognition, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 1258–1262.

[5] Y. Zhang, H. Lu, L. Zhang, X. Ruan, Combining motion and appearance cues for anomaly detection, Pattern Recognition 51 (2016) 443–452.

[6] M. H. Sharif, C. Djeraba, An entropy approach for abnormal activities detection in video streams, Pattern Recognition 45 (7) (2012) 2543 – 2561.

[7] A. A. Sodemann, M. P. Ross, B. J. Borghetti, A review of anomaly detection in automated surveillance, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (6) (2012) 1257–1272.

[8] D. G. Lee, H. I. Suk, S. K. Park, S. W. Lee, Motion influence map for unusual human activity detection and localization in crowded scenes, IEEE Transactions on Circuits and Systems for Video Technology 25 (10) (2015) 1612–1623.

[9] R. M. Colque, C. Caetano, M. Toledo, W. R. Schwartz, Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos, IEEE Transactions on Circuits and Systems for Video Technology PP (99) (2016) 1–1.

[10] M. J. Leach, E. Sparks, N. M. Robertson, Contextual anomaly detection in crowded surveillance scenes, Pattern Recognition Letters 44 (2014) 71 – 79.

[11] D. Xu, Y. Yan, E. Ricci, N. Sebe, Detecting anomalous events in videos by learning deep representations of appearance and motion, Computer Vision and Image Understanding 156 (Supplement C) (2017) 117 – 127, image and Video Understanding in Big Data.

[12] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, X.-S. Hua, Spatio-temporal autoencoder for video anomaly detection, in: Proceedings of the 2017 ACM on Multimedia Conference, MM '17, ACM, New York, NY, USA, 2017, pp. 1933–1941.

[13] W. Luo, W. Liu, S. Gao, Remembering history with convolutional lstm for anomaly detection, in: 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 439–444.

[14] X. Mo, V. Monga, R. Bala, Z. Fan, Adaptive sparse representations for video anomaly detection, IEEE Transactions on Circuits and Systems for Video Technology 24 (4) (2014) 631–645.

[15] H. Wang, D. Oneata, J. Verbeek, C. Schmid, A robust and efficient video representation for action recognition, International Journal of Computer Vision 119 (3) (2015) 219–238.
URL https://hal.inria.fr/hal-01145834

[16] K. Soomro, A. R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, CoRR abs/1212.0402.
URL http://arxiv.org/abs/1212.0402

[17] F. Perronnin, Universal and adapted vocabularies for generic visual categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (7) (2008) 1243–1256.

[18] N. Inoue, K. Shinoda, A fast and accurate video semantic-indexing system using fast map adaptation and gmm supervectors, IEEE Transactions on Multimedia 14 (4) (2012) 1196–1205. doi:10.1109/TMM.2012.2191395.

[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech, and Language Processing 19 (4) (2011) 788–798.

[20] P. Kenny, G. Boulianne, P. Dumouchel, Eigenvoice modeling with sparse training data, IEEE Transactions on Speech and Audio Processing 13 (3) (2005) 345–354.

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[22] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[23] J. F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (3) (2015) 583–596.