Human Action Recognition in Surveillance Video of a Computer Laboratory

Abdul-Lateef Yussiff[1], Yong Suet-Peng[2] and Baharum B. Baharudin[3]
Department of Computer and Information Sciences
Universiti Teknologi PETRONAS,
Bandar Seri Iskandar,
Tronoh, Perak, Malaysia.
[1]ayussiff@gmail.com, [2]yongsuetpeng@petronas.com.my, [3]baharbh@petronas.com.my

***Abstract*—One of the driving forces of behavior recognition in video is the analysis of surveillance video. In this video, humans are monitored and their actions are classified as being normal or a deviation from the norm. Local spatio-temporal features have gained attention to be an effective descriptor for action recognition in video. The problem of using texture as local descriptor is relatively unexplored. In this paper, a work on human action recognition in video is presented by proposing a fusion of appearance, motion and texture as local descriptor for the bag-of-feature model. Rigorous experiments was conducted on the recorded UTP dataset using the proposed descriptor. The average accuracy obtained was 85.92% for the fused descriptor as compared to 75.06% for the combination of shape and motion descriptor. The result shows an improved performance for the proposed descriptor over the combination of appearance and motion as local descriptor of an interest point.**

***Index Terms*—Human Action Recognition, Video Representation, School Surveillance, Codebook Descriptor.**

I. Introduction

Due to vandalism and thefts, provided with the need to open some laboratories in school for 24 hours a day including weekends, surveillance cameras are commonly installed to monitor movements of people in and out of classrooms and laboratories. There has been a strong interest in automatic analysis of these data for behavioral analysis. However, an important first step is the ability to recognize different human actions found in these video data.

The aim of human action recognition in a video is to provide a semantic interpretation on the observed video clip as how a third person human do. It appeals to many applications i.e. video surveillance, robotics, ambient surveillance, video retrieval and indexing, human computer interaction, and among others. For example, Figure 1 shows some human behaviors, such as smoking and eating which are prohibited from the school premises. Also, there is a warning signboard indicating that the laboratory users are monitored through surveillance camera. Can the behavioral analysis on these video data be used to apprehend the violators of those stated rules?

Human action recognition, as one of the active topics in computer vision, has been extensively researched during the last decades. However, it is still regarded as a challenging task especially in realistic videos. The challenges in action recognition from video data mainly lied on large intra-class variation, background clutter, occlusions, illumination changes and noise. More so, action needs to be described in terms of spatial and temporal attributes. Therefore, this work aims to address some of the challenges by proposing a robust local feature representation for action in video.

NOTICE
CCTV
IN OPERATION
NOTICE
2014/11/10
A stern action will be taken if you are found to bring food & drinks into the laborat

Fig. 1. Notices indicating users are monitored by surveillance camera and prohibiting some behaviors (actions)such as smoking and eating at the school computer laboratory. Warning and punishments of those offense are put on the door of the laboratory too.

Central to the entire discipline of action recognition is the concept of feature representation. Gradient, flow information or combination of these two attributes have been popularly used [1], [2] for describing the shape and motion attributes of the video. One major limitation of these studies that have utilized shape and motion attributes to describe an interest point is their failure to capture the relationship among the local pixels. More so, far too little attention has been paid to the use of texture feature as local descriptor, even though there are many benefits that can be derived from it. Texture feature can be used to capture the inter-relationship among the local pixels. Feature representation would have been more robust if texture attribute is combined with the shape and motion information to describe and action in video. Therefore, one question that needs to be asked, however, is whether the addition of texture to the existing shape and motion as local descriptor of an interest point will improves the performance of an action recognition in video?

In this paper, we present the work on human action classification in video data by proposing the fusion of Haralick texture

978-1-5090-2549-7/16/$31.00 ©2016 IEEE

feature [3] and Histogram of Gradient Orientation (HOG) as appearance features, while we extract the Histogram of Optical Flow (HOF) as motion features, to form the descriptor for an interest point based on bag-of-features method of video representation. While HOG and HOF have been used as action descriptors as described in [1], to the best of authors knowledge, Haralick texture feature has not been previously considered as action descriptor.

This paper is organized as follow: Section II addresses the related work on the features used in action recognition, followed by our proposed feature descriptor in Section III. The result and discussion is addressed in Section IV, and conclusion in Section V.

II. Previous Works

In the work of action recognition from video data, shape-based action representation [4]–[7] has been the favorite among early work. This technique assumes that the actors' silhouette can be captured at any time. A binary image representing human body shape in each image frame is called a silhouette. Sequence of silhouettes constitutes a silhouette tunnel, that is, a spatial - temporal binary mask of the moving object changing its shape over time. The silhouette representation is invariant to color, texture and intensity changes of the target object. The drawback of this representation is that motion within the object itself is not captured and also the quality of the silhouettes have an effect on the final action recognition. This representation fails when detecting self-occlusion and also requires background subtraction in order to isolate action performing foreground object but good for well-controlled environment, whereby, a single actor can be assumed to be closed to view of camera and with minimal distraction.

Also, motion feature [8]–[10] provides the most discriminative characteristics and dynamic attributes of actions. Recent evidence suggests that motion based representation in general does not depend on background subtraction and this make it attractive to practical works than the shape based features [9]. Another closely related concept to motion is the trajectory which is the history of human motion in space over time. The trajectory is computed from association of consecutive frame. Based on the assumption that motion is the most informative cues for action recognition, Wang et al. [9], [10] employed dense trajectory and motion boundary descriptors for action video representation. Dense points are sampled from each frame and then tracked based on displacement information from a dense optical flow field.

Furthermore, a large and growing body of literature [1], [2], [11]–[13] has investigated the suitability of spatio temporal interest point (STIP) first proposed by Laptev [14] as a representation for space-time events. This space-time feature representation extracts the shape and motion of objects from the videos. A number of studies [1], [13] have found that STIP-based representations generally are more tolerance to noise, illumination, cluttered background and inherently robust to occlusion. Also this approach can avoid the temporal alignment problem, invariant to geometric transformation and escape from the problem of object segmentation.

Laptev et al. [14] used an extension of the Harris corner detector [15] to locate local salient pixels with significant local variations in both spatial and temporal dimensions. Gaussian and Gabor filters have also been adopted to improve STIP in detecting interest points with spatial-temporal volume [2], [12]. These interest points exhibit local maxima of an image region. Overall, interest point representation provides an advantage of no reliance on explicit body part labeling nor person detection and localization. Furthermore, interest point detectors has the advantages of providing robustness to non-homogeneity of a texture and help reduce the computational cost by selecting fewer but more characteristic points [16].

Texture is one of the primary properties for identifying objects or regions of interest in image classification. Texture refers to surface characteristics and appearance of an object in which it provides important visual cues about surface properties and scenic depth for the object of interest in images. Texture features have been used in video representation for event detection [17], [18] and keyframe detection [19], but seldom applied in action classification for video data. One notable research work on the usage of texture feature for action recognition was by Yeffet and Wolf [20]. The authors proposed the local trinary pattern (LTP) which is an extension of the textural-based local binary pattern (LBP) to the video domain for action recognition. Even though the LTP produce an impressive result for some video dataset that were collected in the controlled environment. The result was not encouraging when applied on the realistic video dataset such as Hollywood-2.

Therefore, in our work, we extended the concept of STIP's descriptor by adding Haralick texture feature to the HOG to form the appearance features while employing optical flow feature as motion features for video representation to boost up the classification accuracy.

III. Methodology

The bag-of-feature based method is one of the more practical ways of representing actions in video. The motivating idea is based on the assumption that, two instances of the same actions may be different significantly in terms of their overall appearance and motion but they will tend to have very similar intrinsic properties. Figure 2 is the pipeline which consist of the four processes; video input, interest point detector, descriptor extraction and classification. More details on each process will be explained in the following subsections.

A. Pre-processing of video data

The video inputs may have different video properties. To ensure uniformity and consistencies across all the inputs, the resolution is adjusted to 720 × 576 and frame rate is set to *30 fps*. Furthermore, a realistic video usually contains many different actions of visual characteristics, therefore needs to be segmented into a homogeneous action clip. Segmenting the video into a homogeneous action facilitates the training

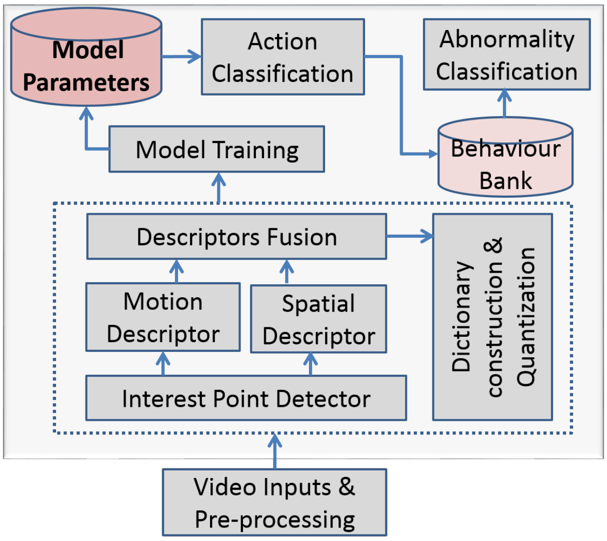


Fig. 2. The proposed computational framework for human action recognition

and hence the classification of these action classes. The segmentation task was done using ffmpeg [21]. Each video clip is having a few seconds (ranging from 2-5 seconds) of playtime. Some sample frames from the dataset reflecting the human action classes are as shown in Figure 3.

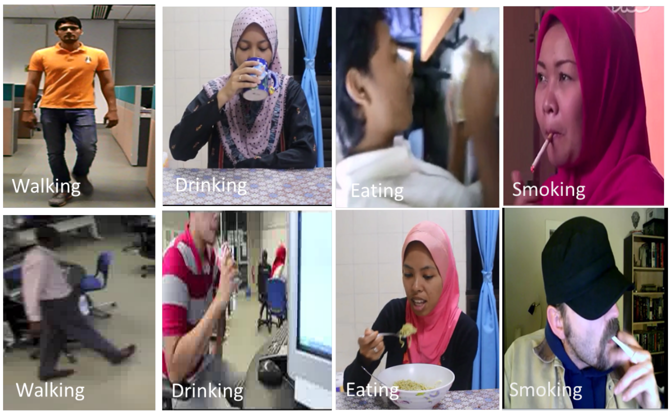


Fig. 3. A few example of representative frames from clips in each action class of human actions: Walking, Drinking, Sitting, Smoking and Eating.

B. Interest Point Detector

Actions usually contain both spatial and motion information, therefore extracting features from them has never been a trivial task. To detect the space-time characteristics from the video dataset, our work makes use of the STIP [14] detector to extract a set of interest points from the video data. Interest points are located in a region that show a high variation of image intensity in spatial and time dimension. We used Harris3D corner detector as interest point detector. The Harris3D Corner detector captured the most prominent salient points of the video and it has low computationally complexity [14]. Figure 4 shows sample frames overlay with extracted STIP features.

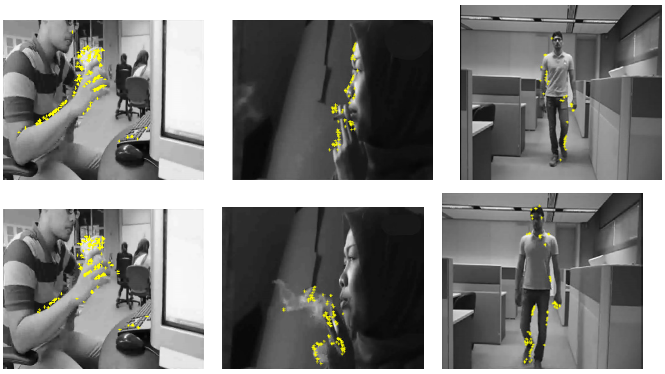


Fig. 4. Sample frames of some actions overlay with extracted spatial-temporal features as interest points, indicating that the interest points are concentrated at the regions which can distinguish the actions.

C. Descriptors Extraction

Once interest point has been detected, a description of each captured interest points was obtained from a local volume of dimension $3\times3\times2$ centered at that point. First, the local shape and texture were captured by extracting the spatial features from the 3×3 region, and secondly the motion descriptors from the motion field of the volume is captured by extracting the local optical flow of two consecutive frames of the same region. A descriptor constitutes a set of measurable properties describing appearance (shape and texture) and motion properties.

1) Spatial Descriptor: While extracting the spatial descriptor, a 4-bins Histogram of Oriented Gradient (HOG) was applied to get the shape appearance. The obtained patch descriptors were normalized into 72 dimensional HOG descriptor. Equations (1) to (4)) were employed to calculate the gradient of each patch followed by the computation of gradients' orientation and magnitude. Figure 5 shows the illustration of how the spatial (HOG) descriptors were computed and extracted from the local region describing a particular interest point.

$$G_x(x,y) = I(x+1,y) - I(x-1,y) \tag{1}$$

$$G_y(x,y) = I(x,y+1) - I(x,y-1) \tag{2}$$

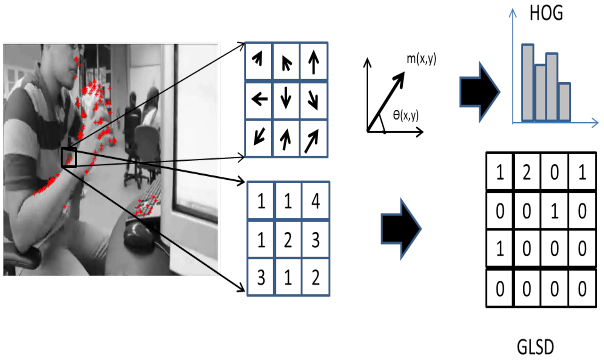
Where $G_x(x,y)$ and $G_y(x,y)$ are the x and y components of the gradient. The *I(x,y)* is the intensity value at the location *x, y*.

The orientation $\phi(x,y)$ and magnitude $m(x,y)$ were calculated as shown in Equations (3) and (4).

$$\phi(x,y) = \tan^{-1}(\frac{G_y(x,y)}{G_x(x,y)}) \tag{3}$$

$$m(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{4}$$

We further calculated the Gray Level spatial Dependencies (GLSD) of the local patch to measure the inter-dependencies



1	1	4
1	2	3
3	1	2

1	2	0	1
0	0	1	0
1	0	0	0
0	0	0	0

Fig. 5. Sample frame from the drinking action illustrating the process of computing the HOG (shape) and the GLSD (texture) descriptors from the support region of an interest point.

among the pixels in the local patch. We then selected 4-dimensional texture feature from the 13 features that can be obtained from the Haralick [3] features of the local region as texture descriptor. These features measure the Gray Level Spatial Dependency (GLSD) which is the gray scale co-occurrence among adjacent pixels. The advantage of Haralick features is that it can be computed easily and have shown to be very effective in representing images [22]. The extracted local texture features were: intensity contrast measurement between a pixel and its neighbor in the region (Eq. (5)); correlation measurement (Eq. (7)); intensity energy of the region (Eq. (6)) and the homogeneity (Eq. (8)) of the local region. The local texture feature was normalized to obtain a 4-dimensional descriptor.

$$Contrast = \sum_{I_j, I_k} |I_j - I_k|^2 \, p(I_j, I_k) \tag{5}$$

$$Energy = \sum_{I_j, I_k} p(I_j, I_k)^2 \tag{6}$$

$$Correlation = \sum_{I_j, I_k} \frac{(I_j - \mu_{I_j})(I_k - \mu_{I_k}) p(I_j, I_k)}{\sigma_{I_j} \sigma_{I_k}} \tag{7}$$

$$Homogeneity = \sum_{I_j, I_k} \frac{p(I_j, I_k)}{1 + |I_j - I_k|} \tag{8}$$

Where I_j, I_k are the intensity values and $p(I_j, I_k)$ is the probability of co-occurrence of these two intensity values in the image region. μ_I and σ_I are the mean and the standard deviation of the intensity occurrence respectively.

2) Motion Descriptor: Motion descriptor was extracted from optical flow [23] of two consecutive frames. Optical flow is use to describe flow information and was computed around the space-time interest points using the second moment matrices based on a 5-bin Histogram of the Optical Flow (HOF). The optical flow equation was derived from Equation (9).

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{9}$$

Equations (9) - (11) were the Taylor series expansion of Equation (9) followed by representation of the motion velocity in the x-direction and y-direction by u and v respectively.

$$\frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial t}{\partial t}\frac{\Delta t}{\Delta t} = 0 \tag{10}$$

$$uI_x + vI_y + I_t = 0 \tag{11}$$

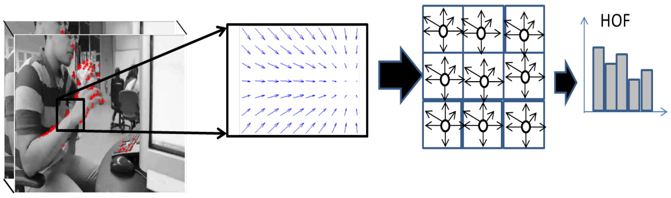


Fig. 6. Illustration of the Histogram of Optical Flow (HOF) from the local regions of two consecutive frames. The number of Bins used is 5.

The motion descriptor was then normalized to a 90-dimensional histogram vectors.

The experiments were carried out with different subsets of the local descriptor in order to evaluate which of the local descriptors has the highest positive effect on the percentage of the accuracies. Based on the empirical experimental design, three subsets were selected for evaluation. The descriptors that were selected are: the motion descriptor alone (HOF), the shape and motion descriptors (HOG + HOF) and the combination of all the three descriptors (GLSD + HOG + HOF). The selected descriptor group were then used to construct a dictionary for the bag-of-feature representation.

D. Dictionary Construction and Quantization

The set of all interest points' descriptors is used to represent a given video with Bags-of-Features method [24]. The descriptor vectors of each local volume was encoded to obtain a compact representation in a vocabulary of ”geometric words”. The process involves dictionary construction and quantization. A dictionary $D = \{V_1, V_2, ..., V_k\}$ of size k, is a set of representative vectors in the descriptor space, derived through unsupervised learning with *k-means* algorithm and Euclidean distance as the clustering metric from the pool of local descriptors. The value of k, which is the size of the dictionary was obtained through experimental design. Figure 7 depicts the codebook construction using the k-means clustering algorithm.

In the quantization step, given the dictionary $D = \{V_k\}_{k=1}^{K}$, each local descriptors is assigned to the nearest cluster and then inherit a unique cluster membership representation label that it has been assigned to based on the Euclidean distance metric. Histogram was constructed from the quantized vector for each video input in the dataset.

E. Action Classification

The normalized histogram of the k-dimensional feature descriptor were used for model training and testing. In our

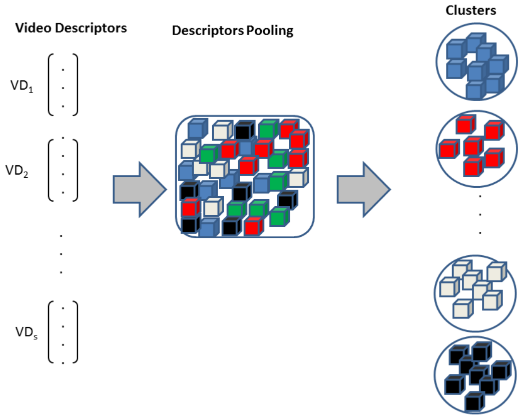


Fig. 7. Codebook formation using the k-means algorithm to group similar descriptors together in a cluster.

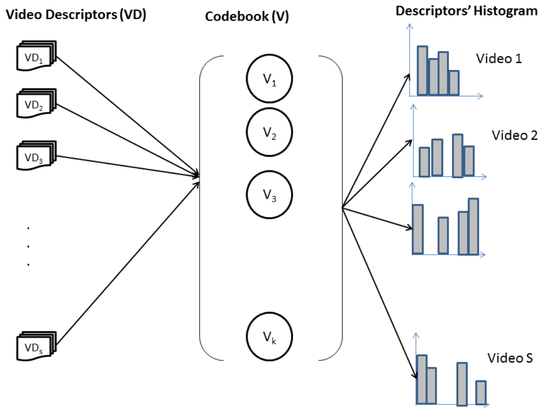


Fig. 8. This figure shows the quantization step that takes the video descriptors and Codebook as input to produce the feature histogram for each video in the dataset. The histogram feature is a compact representation of the video.

experiment, we adopted support vector machine (SVM) for action classification. The selection of the SVM was as a result of experimentation with several classifier models. The support vector machine produced the best result.

IV. Experiments and Results

In this section, experimental results are discussed. Action features are evaluated and compared with other similar descriptors in the interest point representation categories. Furthermore, The features and classifiers combination that give an optimum result was used for further experiment.

Video inputs were collected from the school computer laboratories in Universiti Teknologi PETRONAS (UTP) and also from Youtube for our experiments. The number of training dataset is 240 videos clips which is equally divided by four action ('eating', 'drinking', 'smoking' and 'walking') classes.

A. *Evaluation of our proposed technique*

This study set out with the aim of assessing the importance of texture feature as local descriptor in combination with the most commonly used spatio-temporal descriptor.

To select the most appropriate features, as previously mentioned, this work focus on local STIPs representation while exploring the discriminate power of several action descriptors such as HOG for shape, HOF for motion descriptor, and Gray Level Spatial Dependency (GLSD) for texture descriptors with their combinations. Each descriptor was extracted from the local volume as described in Section III. The detailed comparison is presented in table I. The value of *k*, which is the size of dictionary is 3000.

It has previously been shown from Laptev et al. [1] that combination of HOG and HOF produce a better performance in terms of accuracy of recognizing actions on benchmarking dataset. Therefore, the basis of this section is to investigate what has been reported and also hypothesizes that addition of texture features improves the performance.

94 videos were used as test data. Figure 9 shows the sample frame of video collected for test dataset.

Fig. 9. Sample testing dataset. These dataset were set aside from the training to prevent the same video being used for the training and testing processes.

Table I compares action recognition performance on the experimental data. As shown in table I, the proposed method reported improved accuracy than the other two groups. The result table on the average indicates that addition of texture to the existing spatio-temporal descriptor does improves the recognition performance. However, repeated measures of ANOVA showed that these results were not statistically significant at $\alpha = 0.05$. Therefore, more research on this topic needs to be undertaken before the contribution of local texture descriptor to the action recognition performance is more clearly understood.

TABLE I

Feature comparisons for motion, shape, texture and their combination. The combination of three features produce a better accuracy performance for each action class in percentage(%).

	HOF	HOG+HOF	GLSD + HOG + HOF (proposed method)
Drinking	64.8	82.18	91.67
Eating	55.5	62.2	75.4
Smoking	55.3	76.45	77.5
Walking	67.4	79.4	99.1
Average	60.75	75.06	85.92

Some misclassified actions by the classification model during testing are shown in Figure 10. While the first and the second images are the false positive images, the last image indicates the false negative action. The first two images were originally grouped under the walking action class but because in both images, two actions take place simultaneously, that is drinking while walking. So the two images were classified by the model under drinking action. In the last image of the figure, the drinking action was missed. It is due to the fact that all the drinking action examples in the training dataset involves movement of the arms but in this example, there was no movement of the arms so it was flagged negative.

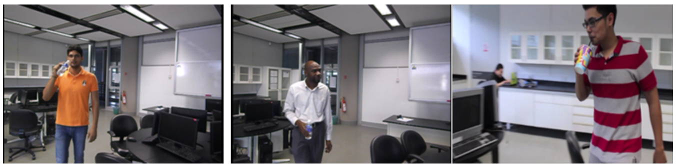


Fig. 10. Representative frames of the video dataset that were misclassified.

V. Conclusion

Accurate classification of human action in video based on good feature representation can play an important role for intelligent video surveillance applications, video searching, video indexing and even human computer interaction. This can serve as the preliminary step for behavior detection and recognition especially in application that further needs behavior to be monitored. In this paper, we complemented the popular spatio-temporal shape and motion information with textural features which describes the spatial distribution of local patterns. The bag-of-feature model was employed for video representation prior to model training and testing. Returning to the question posed at the beginning of this study, it is now possible to state that texture feature is indeed a good attributes for describing the local spatio-temporal interest point. In our experiment, our proposed feature integration provides a better performance than using those features individually. A repeated measures of ANOVA showed that these results were not statistically significant. Therefore, more research on this topic needs to be undertaken before the contribution of local texture descriptor to the action recognition performance is more clearly understood. The study has gone some way towards enhancing our understanding of local feature representation for action recognition.

A limitation of this study is that the numbers of action classes as well as the dataset size were relatively small. It is therefore recommended that further research be undertaken in the following areas: More actions to train the classifiers in order to test the robustness of the proposed features, also pattern of students' behavior needs to be studied which will eventually identifies the abnormal(forbidden) behaviors.

References

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

[2] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.

[3] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.

[4] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 3, pp. 257–267, 2001.

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 12, pp. 2247–2253, 2007.

[6] J. Zhang and S. Gong, "Action categorization by structural probabilistic latent semantic analysis," *Computer Vision and Image Understanding*, vol. 114, no. 8, pp. 857–864, 2010.

[7] J. Zhang and S. Gong, "Action categorization with modified hidden conditional random field," *Pattern Recognition*, vol. 43, no. 1, pp. 197–203, 2010.

[8] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 2, pp. 288–303, 2010.

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

[10] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[11] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 489–496.

[12] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.

[13] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2046–2053.

[14] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[15] C. Harris and M. Stephens, "A combined corner and edge detector." in *Alvey vision conference*, vol. 15. Manchester, UK, 1988, p. 50.

[16] L. Zaanen, "Bag-of-features model, application to medical image classification," 2014.

[17] J. Yongqiang, W. Jonathan, and Y. WeiWi, "Flame detection in surveillance," *Journal of Multimedia*, vol. 6, no. 1, pp. 22–32, 2011.

[18] R. Qian, N. Haering, and I. Sezan, "A computational approach to semantic event detection," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1, 1999, pp. –206 Vol. 1.

[19] S.-P. Yong, J. Deng, and M. Purvis, "Wildlife video key-frame extraction based on novelty detection in semantic context," *Multimedia Tools and Applications*, vol. 62, no. 2, pp. 359–376, 2013. [Online]. Available: http://dx.doi.org/10.1007/s11042-011-0902-2

[20] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 492–497.

[21] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[22] S.-P. Yong, J. D. Deng, and M. K. Purvis, "Modeling semantic context for key-frame extraction in wildlife video," in *Image and Vision Computing New Zealand (IVCNZ), 2010 25th International Conference of.* IEEE, 2010, pp. 1–8.

[23] B. K. Horn and B. G. Schunck, "Determining optical flow," in *1981 Technical Symposium East*. International Society for Optics and Photonics, 1981, pp. 319–331.

[24] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," in *ECCV Workshop on Statistical Learning in Computer Vision*, 2004, pp. 1–22.