Fusion of histogram based features for Human Action Recognition

Suraj Prakash Sahoo, Silambarasi R, Samit Ari
Department of Electronics and Communication Engineering
National Institute of Technology, Rourkela, India
surajprakashsahoo@gmail.com, silambarasi@gmail.com, samit@nitrkl.ac.in

***Abstract*—The human action recognition (HAR) is framed as a machine learning problem. During HAR, speed of actions, shape of action and background noise play vital role. In this work, to represent speed of action, Bag of histogram of optical flow (BoHOF) is proposed. In this technique, the optical flow is calculated over the segmented human object. Further, the features are thresholded and bagged to compute the BoHOF. Along with BoHOF, sobel edge filter is used in horizontal and vertical direction to remove shadow effect. Median filtering is applied to suppress background noise. Histogram of oriented gradients (HOG) features are extracted from 3D projected planes and combined with BoHOF to extract maximum advantages of both the features. Finally, the multi-class SVM-based classifier with radial basis kernel is applied to recognize different human actions. The experiments are conducted on the benchmark KTH dataset and the experimental findings concludes that the proposed HAR technique provides better performance compared to the state-of-the-art techniques.**

***Index Terms*—Histogram of Optical Flow (HOF), Human Action Recognition, Human Object Boundary Detection, Multi class Support Vector Machine (SVM).**

I. Introduction

Action recognition is the process of recognizing the human action by continuously observing the human with different environmental conditions. Action recognition [1] provides significant support in different applications, thus the knowledge is required to study different fields such as public security, video surveillance, human-computer interaction, robotics learning, unmanned aerial vehicles, intelligent driving, content based video retrieval, medical diagnosis and sports video analysis. Traditionally humans are used to monitor a particular area for long time. Human may lose their attention while monitoring continuously for a long time. To overcome this issue, camera based monitoring system is introduced. In video surveillance [2], video cameras are placed near the area which is to be monitored. The camera based monitoring system allows to record, store and process the monitored information continuously which is used to aid the video surveillance. By processing the monitored information, the abnormal or unusual activities [3] are identified. It helps in the field of crime analysis at sensitive areas like bank, school, airport, hospitals, restaurant, military installation.

In literature, some previous techniques are reported to support HAR. Klaser *et al*. [4] have presented histogram of oriented gradient (HOG), which is used as a local descriptor and it gives the direction of the gradient. Pers *et al*. [5] have introduced histogram of optical flow (HOF) descriptor. In contrast to local features, the global features describe the human object in the form of silhouette [6], optical flow [7] and edges [8]. For global features, holistic based representations are used, which yields unique feature descriptor even in the presence of human body part occlusion and background noise. In global feature estimation, HOG [9] and HOF [5] feature extraction techniques are commonly used. In [10], Patel *et al.* have identified that use of individual features for HAR is a critical task. To overcome this, fusion of feature concept was introduced by fusing different feature extraction techniques, different classifier and different model. Cilla *et al.* [11] presented a distribution system for action recognition based on scene view. It performed the principle component analysis (PCA) [8] and linear discriminant analysis (LDA) based dimension reduction techniques, and parametric (nearest neighbor conditional density estimator) and non parametric (k-means and naive Bayes) classification techniques. To describe the high level motion informations, region of interest is detected using the motion history image (MHI) [12], [13]. MHI is used for action representation in [13]. Samanta *et al.* [14] represented videos as 3D space time facet model and detected the STIP.

From the literature it is found that, use of local or global features is having its own advantages. Local features [5], [9] are less affected by occlusion and noise in comparison to global features. However, for better representation, local interest points should be sufficient enough which in return increases the complexity of the paradigm. When detecting motion, local feature provides information about low level motions. In order to detect high level motion, global features are suitable. In [15], HOG feature is used as global feature on extracted 3D spatio-temporal planes. However, shadow of the human, noisy background and speed of action are the problems to be handled further. In this work, sobel edge detection is applied in both horizontal and vertical direction to segment the human object from background. The detected human object is enhanced by the morphological operations to remove the sparse noise, clutter background and object shadow. Bag of histogram of optical flow (BoHOF) is proposed to represent the motion in video. Fusion of HOG and BoHOF features is used to improve the classification accuracy. HOF contributes motion and speed related features and HOG provides orientation of the motion gradients.

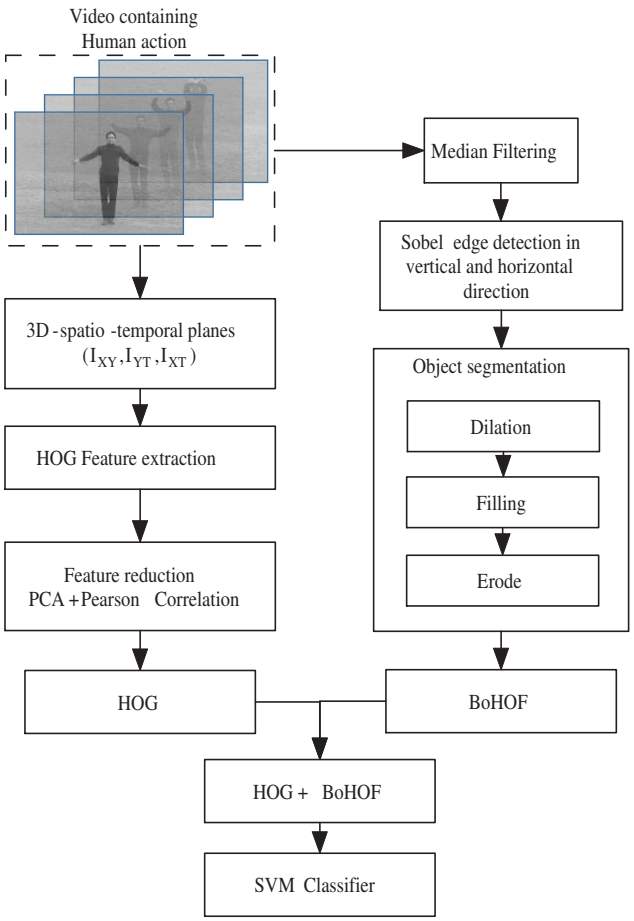

Fig. 1: Block diagram of the proposed technique for HAR.

The following sections are as follows: the proposed BoHOF feature and fusion with other features is explained in section II. Section III discusses the experimental results and comparison of the proposed work with the previously available methods in literature. Finally, the work is concluded in section IV.

II. Proposed Framework

The overview of the proposed technique is shown in Fig. 1. The proposed technique extracts features from action videos by HOG [15] and proposed BoHOF technique. As in [15], the HOG features are extracted from three MHI projected planes I_{XY}, I_{YT}, I_{XT}. PCA based feature reduction and Pearson correlation based feature selection are employed on extracted HOG features. Before BoHOF feature extraction, the human action region is segmented by using the edge detection in both horizontal and vertical direction followed by morphological operations. The BoHOF features are extracted on the boundary of the segmented action region. The two features are fused to provide a better feature vector. Finally, the actions are classified by SVM classifier with RBF kernel.

A. Preprocessing

The median filter is applied on the frames of videos by sliding the filter window. In the experiment 3×3 median filter is used for filtration. This median filter is used to enhance the image by removing the noise. Median filter is a nonlinear filter which is a popular filtering technique, because it preserves the edges while removing noise. The median filters produce better performance in the presence of salt and pepper noise. Because of preserving edges and noise removal property of median filter, the median filter is effective than the convolution operation. After median filtering, sobel edge detection is used to segment the action region. In general, the human object (foreground) in a background smoothed frame can be identified by using the edge detection methods. It may contains background noises like edges on the background scene and shadow of human object, as illustrated in Fig. 2(a,b). Boundary detection by single directional sobel edge detection contains more background noise. To remove these noises, sobel edge detection is performed in both horizontal and vertical direction. Sobel edge detector preserves the edges and eliminates the noise content. Fig. 2 shows the effect of the horizontal, vertical and bi-directional sobel edge detection on image frame. The result shows that the edge detection by bi-directional sobel filter provides boundary of an object accurately. After edge detection, various morphological opeartions are applied to remove other noise residuals. The final step is to segment the human action region as shown in Fig. 2(c).

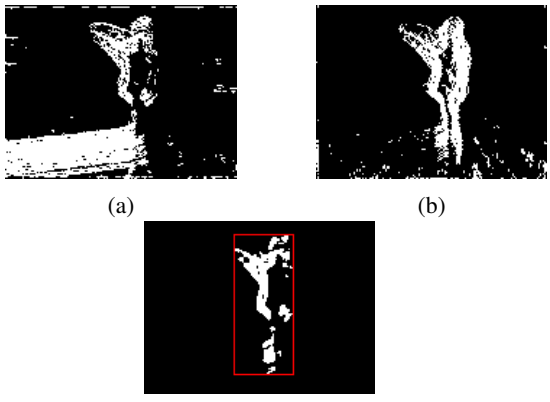

Fig. 2: Two directional sobel edge detection to segment the action region. (a) XY plane by horizontal sobel edge detection, (b) XY plane by vertical sobel edge detection, (c) XY plane by horizontal and vertical sobel edge detection.

B. Bag of HOF (BoHOF) feature extraction

The detected human object from the boundary detection method is considered for the optical flow extraction. It contains only the interest event, therefore, it is free from noise. The new optical flow image is divided into four sub region. The motion of the legs are in the third and fourth sub region and motion of the hands are in the first and second sub region. Optical flow vector has one horizontal $(u(x, y))$ and vertical $(v(x, y))$ component at each pixel. The magnitude and phase of the optical flow vectors are calculated as:

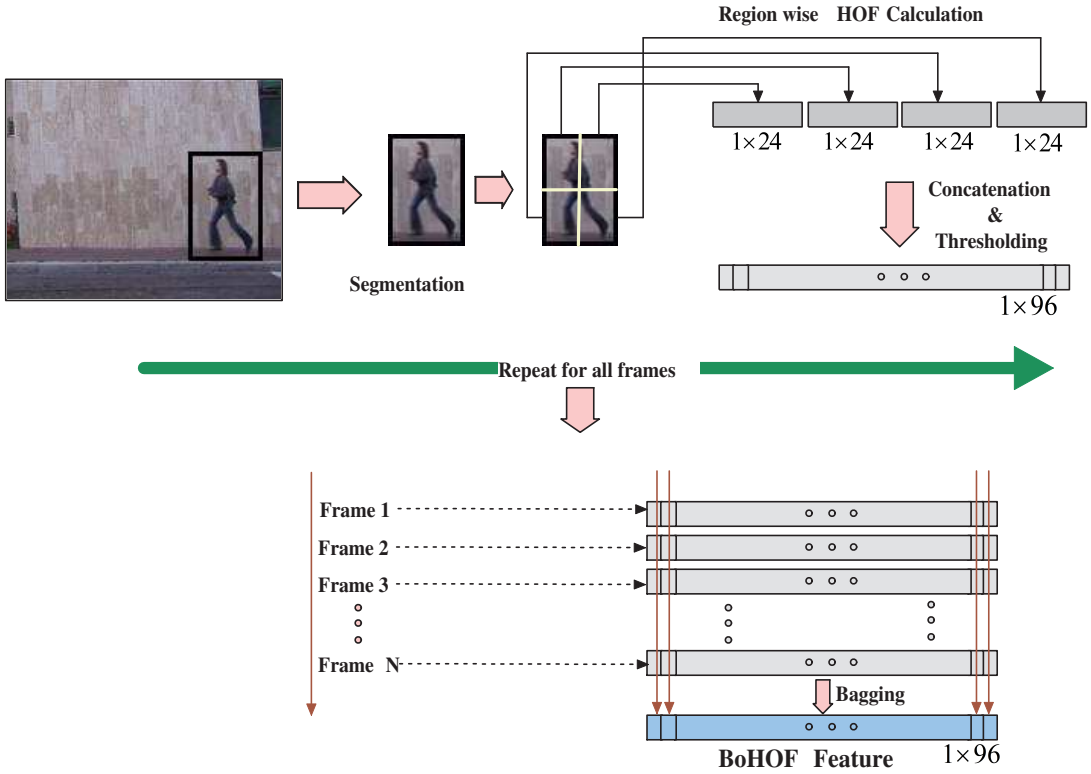

Fig. 3: Block diagram of the proposed BoHOF feature extraction technique.

$$mag = \sqrt{u(x,y)^2 + v(x,y)^2}$$
$$angle = \tan^{-1}\left(\frac{u(x,y)}{v(x,y)}\right) \quad (1)$$

In each sub region eight bins are considered. The total angle is divided into eight bins, therefore each bin has a range of 45^o. The normalized magnitude is divided into three bins. By considering these three magnitude bins and eight direction bins, histogram is computed. It gives a total of $8 \times 3 = 24$ bins in each sub region. Therefore, the histogram for a single image is of $4 \times 24 = 96$ bins. The same procedure is repeated for all frames in a video. From the histogram of optical flow only the bin with dominant motion direction and magnitude are considered based on the threshold value. The thresholded bins from all frames are bagged to produce the proposed BoHOF feature of length 96 for a video. The detailed procedure is explained in Fig. 3.

C. Fusion of HOG and BoHOF Feature

The HOF features consider the motion of human action, velocity and speed. When actions with similar shape but different speed are recognized, HOF is more applicable. When only gradient or shape is sufficient to recognize, HOG performs better. To use the advantage of both the features, HOG and BoHOF are fused in the proposed work for HAR. The feature length of the HOF is taken as 96 and the feature length of the HOG features is 200. The normalized features of HOG and HOF are fused to enhance the result. The combined feature has the length of 296. The HOG feature is extracted as in [15]. The HOG feature is extracted from the 3D projected planes of action videos. The dimension of extracted HOG feature is 12768, 7392 and 10032 from *XY*, *XT* and *YT* planes. Since the feature dimension is very high, it should be reduced. For feature reduction, PCA technique is applied and for feature selection, Pearson correlation technique is adopted. Finally the feature length of HOG is reduced to 200.

III. Results and Discussion

To evaluate the proposed fusion of feature technique for HAR, KTH dataset is used in this work. All the experimentation work is simulated using MATLAB Version: 9.0.0.341360 Release 2016a. The computer environment used is having 4GB RAM, windows 10 platform and intel core i5 processor with clock speed of 3.20 GHz. KTH dataset is one of the standard dataset for action recognition algorithm. It contains six different actions 'boxing', 'handwaving', 'hand clapping', 'running', 'jogging' and 'walking'. Each action is performed by 25 persons under four different scenarios like outdoor, indoor, different clothes and normal clothes. The dataset is captured by a static camera with scale variations. The videos are captured with 120×160 pixels resolution and 25 frames per second. The experimental setup for KTH dataset is: 16 persons' actions for training set and 9 persons' actions for testing set.

The HOG feature describes shape of an object. The problem arises when similar actions having less inter class variation

Fig. 4: Comparison of classwise accuracy of the proposed feature fusion (HOG+BoHOF) with HOG and BoHOF.

TABLE I: Confusion Matrix for KTH dataset using BoHOF feature with SVM classifier for performance evaluation

Action	Boxing	Hand clapping	Hand waving	Jogging	Running	Walking
Boxing	**0.65**	0.16	0.11	-	0.06	0.02
Hand clapping	0.12	**0.64**	0.21	-	0.02	0.01
Hand waving	0.09	0.22	**0.67**	-	0.02	-
Jogging	-	-	-	**0.82**	0.08	0.10
Running	-	-	-	0.04	**0.96**	-
Walking	-	-	-	0.04	-	**0.96**

are recognized. Class wise accuracy for KTH dataset with HOG feature is shown in Fig 4. The result shows better performance for 'boxing', 'hand waving', 'hand clapping' actions whereas, the 'running', 'jogging' and 'walking' actions are confused with each other. To handle the closely related actions where speed is a distinguishing factor, optical flow

TABLE II: Confusion Matrix for KTH dataset using the proposed HAR technique for performance evaluation

Action	Boxing	Hand clapping	Hand waving	Jogging	Running	Walking
Boxing	**0.97**	0.01	0.02	-	-	-
Hand clapping	0.01	**0.98**	0.01	-	-	-
Hand waving	0.01	0.02	**0.97**	-	-	-
Jogging	-	-	-	**0.93**	0.03	0.04
Running	-	-	-	0.04	**0.96**	-
Walking	-	-	-	0.01	-	**0.99**

TABLE III: Comparative Performance of six classes of KTH dataset using proposed technique

Action	Precision	Recall	F-score
Boxing	0.979	0.97	0.975
Hand clapping	0.971	0.98	0.975
Hand waving	0.97	0.97	0.97
Jogging	0.948	0.93	0.939
Running	0.969	0.96	0.964
Walking	0.961	0.99	0.975

TABLE IV: Comparison of action classification on KTH dataset with state-of-the-art methods using proposed technique

Method	**Accuracy** (%)
Li *et al.* [12]	93.50
Zhen *et al.* [21]	94.10
Samanta *et al.* [14]	94.90
Liu *et al.* [22]	94.30
Abdul *et al.* [23]	95.36
Megrhi *et al.* [24]	94.90
Silambarasi *et al.* [15]	95.80
Proposed HOG+BoHOF	**96.70**

based technique is used. In this work BoHOF is proposed which uses optical flows to calculate the feature vector. The effect of using BoHOF feature with SVM classifier is studied for KTH dataset. The result in the form of confusion matrix is shown in Table I. From Table I, it is clear that the BoHOF feature is more efficient to handle the actions differentiated by speed. Since, HOG is better for 'boxing', 'hand waving', 'hand clapping' actions and BoHOF is better for 'running', 'jogging' and 'walking' actions, the two features are fused to provide a better feature vector. The confusion matrix for the dataset using proposed HOG+BoHOF features is presented in Table II. The overall accuracy is found to be 96.70%. Comparative performance of six classes is provided in Table III with the performance matrices like precision, recall and F-score. Comparison of action classification using proposed technique with state-of-the-art methods is provided in Table IV. An accuracy of 96.70% is found which is better in comparison to earlier reported techniques. The increase in performance is mainly due to considering motion and shape of the actions efficiently.

IV. Conclusions

In this work, optical flow based BoHOF feature extraction technique is proposed for HAR which represents speed of action more accurately. The BoHOF feature is fused with HOG feature for better representation of actions. When HOG represents the shape of action, BoHOF represents the motion information. The human object is segmented from the background by bi-directional sobel edge detector. Median filtering is used to remove unwanted noises from background. For

classification, SVM classifier with RBF kernel is used. The proposed technique is evaluated on benchmarked KTH dataset. A performance accuracy of 96.70% is reported which is better than the earlier reported techniques.

ACKNOWLEDGMENT

The work is supported by Digital India Corporation (formerly Media Lab Asia) under Visvesvaraya Ph.D. Scheme for Electronics and IT under the department of MeitY government of India. [grant number PhD-MLA/4(13)/2015-16]

REFERENCES

[1] R. Bodor, B. Jackson, and N. Papanikolopoulos, "Vision-based human tracking and activity recognition," in *Proc. of the 11th Mediterranean Conf. on Control and Automation*, vol. 1. Citeseer, 2003.

[2] S. Arora, K. Bhatia, and V. Amit, "Storage optimization of video surveillance from cctv camera," in *Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on*. IEEE, 2016, pp. 710–713.

[3] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, "Understanding transit scenes: A survey on human behavior-recognition algorithms," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 1, pp. 206–224, 2010.

[4] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[5] J. Perš, V. Sulić, M. Kristan, M. Perše, K. Polanec, and S. Kovačič, "Histograms of optical flow for efficient representation of body motion," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1369–1376, 2010.

[6] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

[7] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. Burkitt, "Performance of optical flow techniques," in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*. IEEE, 1992, pp. 236–242.

[8] V. Vo and N. Ly, "An effective approach for human actions recognition based on optical flow and edge features," in *Control, Automation and Information Sciences (ICCAIS), 2012 International Conference on*. IEEE, 2012, pp. 24–29.

[9] T. Kobayashi, A. Hidaka, and T. Kurita, "Selection of histograms of oriented gradients features for pedestrian detection," in *International conference on neural information processing*. Springer, 2007, pp. 598–607.

[10] C. I. Patel, S. Garg, T. Zaveri, A. Banerjee, and R. Patel, "Human action recognition using fusion of features for unconstrained video sequences," *Computers & Electrical Engineering*, 2016.

[11] R. Cilla, M. A. Patricio, A. Berlanga, and J. M. Molina, "A probabilistic, discriminative and distributed system for the recognition of human actions from multiple views," *Neurocomputing*, vol. 75, no. 1, pp. 78–87, 2012.

[12] C. Li, Y. Liu, J. Wang, and H. Wang, "Combining localized oriented rectangles and motion history image for human action recognition," in *Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium on*, vol. 2. IEEE, 2014, pp. 53–56.

[13] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical filtered motion for action recognition in crowded videos," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 3, pp. 313–323, 2012.

[14] S. Samanta and B. Chanda, "Space-time facet model for human activity classification," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1525–1535, 2014.

[15] R. Silambarasi, S. P. Sahoo, and S. Ari, "3d spatial-temporal view based motion tracing in human action recognition," in *Communication and Signal Processing (ICCSP), 2017 International Conference on*. IEEE, 2017, pp. 1833–1837.

[16] T. Zhang, J. Liu, S. Liu, C. Xu, and H. Lu, "Boosted exemplar learning for action recognition and annotation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 853–866, 2011.

[17] L. Shao, L. Ji, Y. Liu, and J. Zhang, "Human action segmentation and recognition via motion and shape analysis," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 438–445, 2012.

[18] G. Goudelis, K. Karpouzis, and S. Kollias, "Exploring trace transform for robust human action recognition," *Pattern Recognition*, vol. 46, no. 12, pp. 3238–3248, 2013.

[19] S. Sadek, A. Al-Hamadi, G. Krell, and B. Michaelis, "Affine-invariant feature extraction for activity recognition," *ISRN machine vision*, vol. 2013, 2013.

[20] A. F. A. Modarres and M. Soryani, "Body posture graph: a new graph-based posture descriptor for human behaviour recognition," *IET Computer vision*, vol. 7, no. 6, pp. 488–499, 2013.

[21] X. Zhen and L. Shao, "A performance evaluation on action recognition with local features," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 4495–4500.

[22] A.-A. Liu, N. Xu, Y.-T. Su, H. Lin, T. Hao, and Z.-X. Yang, "Single/multi-view human action recognition via regularized multi-task learning," *Neurocomputing*, vol. 151, pp. 544–553, 2015.

[23] H. A. Abdul-Azim and E. E. Hemayed, "Human action recognition using trajectory-based representation," *Egyptian Informatics Journal*, vol. 16, no. 2, pp. 187–198, 2015.

[24] S. Megrhi, M. Jmal, W. Souidene, and A. Beghdadi, "Spatio-temporal action localization and detection for human action recognition in big dataset," *Journal of Visual Communication and Image Representation*, vol. 41, pp. 375–390, 2016.