Intelligent Human Action Recognition: A Framework of Optimal Features Selection based on Euclidean Distance and Strong Correlation

Atiqa Sharif [1], Muhammad Attique Khan [2*], Kashif Javed [3], Hafiz Gulfam Umer [4], Tassawar Iqbal [5], Tanzila Saba[6], Hashim Ali[7], Wasif Nisar[8]

[1] *Department of EE, COMSATS University Islamabad, Wah Cantt*
[2,7] *Department of Computer Science and Engineering, HITEC University, Museum Road, Taxila*
[3] *Department of Robotics, SMME NUST*
[4] *Department of Computer Science & IT, Ghazi University, D.G. Khan, Pakistan*
[5,8] *Department of Computer Science COMSATS University Islamabad, Wa Cantt, Pakistan*
[6] *Department of Information Sciences, Prince Sultan University, Saudi Arabia*
Corresponding Author: attique@ciitwah.edu.pk*

Abstract: Extracting salient and most prominent features from a given video sequence is a critical step in Human Action Recognition (HAR). The work presented within this article proposes a new method for HAR, which efficiently addresses the issue of robust feature selection. The proposed method initially fuses three different feature categories based on their highest values, and later selects most optimal features using a novel Euclidean distance (ED) and strong correlation (SC) methods. Finally, it classifies the selected features using multi-class classifier. For experimentation, four publically available datasets including Weizmann, KTH, UCF YouTube, and HMDB51 are used and results with improved classification accuracy, on average more than 94%, are obtained. Experimental results validate that the proposed approach outperforms the existing techniques.

Keywords: Intelligent surveillance; features computation; features fusion; selection

1. INTRODUCTION

In the last two decades, the HAR sought attention of many researchers in the area of computer vision (CV), due to its famous applications such as movie indexing (Weinland et al., 2011), human-computer interaction (Poppe, 2010), biometrics (Arshad et al., 2019), intelligent surveillance (Siddiqui et al., 2018) and video surveillance (Arshad et al., 2019; Aurangzeb et al., 2019; Khan et al., 2018). In video surveillance, human motion is captured from several cameras to recognize humans' activity at the time of their movement in public places such as airports, family parks, and railway stations etc. In this regard, automatic annotation and content-based video analysis (Chang, 2002) allow efficient searching of different actions, for instance, dance movements in musical videos or finding tackles in soccer matches, etc. In addition, such systems has potential to identify day to day activities of aged people and children at home, to simplify the process of their care and reduce worry of attendants (Cardinaux et al., 2011). In literature, several HAR methods are available which addresses known HAR challenges, for instance illumination and complex background. In HAR, human region extraction is one of the major challenges because several objects are moving in the given sequence, but the human only need to be treated as a region of interest. In addition, during the features extraction phase, the irrelevancy and redundancy between features degrade the accuracy of action recognition, which make it more challenging**.**

1.1. Problem Statement

The work presented in this article, deals with the several challenges including: a) low contrast video acquisition because of complex background; b) variations in the human viewpoint; c) occlusion between human and other objects; d) measurement of the body size, symmetry and support level of a human at the time of performing the action because each person has different body size and style; e) high dimensions of extracted features, and f) selection of most useful features. These listed challenges degrade the performance of HAR. In addition, the high dimensional features increase the computational time of automated framework.

1.2. Contribution

In this article, an optimized frame stretching and robust features selection method is proposed for HAR. In general, the HAR consists of series of steps including pre-processing, human detection, features extraction, and recognition. The work presented in this article, follows the same steps however considers a new method. In the new proposed method, pre-processing step considers the contrast of moving region of given video sequences, which is later utilized in the human detection phase. Thereafter, it extracts features such as texture, shape, and Gabor. The extracted features are fused based on higher value features and select the best features by utilizing Euclidean distance and strong correlation. The selected features are finally classified using different methods including M-SVM. Their performance is compared with several state-of-the-art classification methods such as KNN, Ada-boost, CT, and LDA. The major contributions are as under: a) A frame contrast stretching method is introduced which is implemented using Top-hat and Gaussian filter on the original RGB frame and their intensity values is added in one frame. The artefacts in the combined frames are removed by a 3D-Median filter in the second step, which is later

separated with their background by utilizing HSV color space in the last step. b) Extraction of Weighted HOG (WH) features from human silhouette is carried out and PCA is performed. The PCA returns score values for each vector, which is sorted into ascending order and the top 100 features based on their high values are selected. The selected high-value features are later fused with Gabor and LBP features based on parallel mode. c) The robust features are selected from a fused vector by utilizing Euclidean distance and strong correlation. The selected features are later classified by M-SVM.

2. RELATED WORK

Recently in the relevant literature, various techniques are proposed for HAR (Khan et al., 2018). These techniques can be categorized into trajectory based, feature extraction based, feature selection based and to name a few more. Yun et al. (Yi and Wang, 2017) proposed a trajectory based technique for HAR. The technique solves the motion related problems in the given video sequence, and extracts the trajectory-based covariance features for recognition. According to the reported results best performance is achieved as compared to MBH, HOG and HOF. Jonti et al. (Talukdar and Mehta, 2017) introduced a novel approach combining the good features and MLP, for HAR. The good features are iteratively combined with optical flow to capture their motion information and, later on classified by feed-forward MLP. Wing et al. (Ng et al., 2017) presented a data-driven approach for HAR in the video sequences. The proposed approach selects the number of features to minimize the error of RBFNN, and proved suitable for video datasets. Ying et al. (Zheng et al., 2018) presented the unique plans for HAR, based on two major properties including; sketch ability and objectiveness. The sketches are prepared by fast edge detection and then R-CNN is performed parallel, for human detection. After the completion of both processes, the mining is carried out. Finally, four types of sketch pooling methods are applied to get a uniform representation of human actions for given video sequences. Dinesh et al. (Vishwakarma et al., 2018) introduced a novel method for HAR which extracts the human silhouette using texture-based segmentation. Later on, it extracts shape and view based features, and adds Gabor wavelet features to improve its strength. Finally, features are fused to obtain a robust feature vector, which is classified by SVM. According to the reported results, the technique achieved the best accuracy. Wang et al. (Wang et al., 2013) used point coordinates, histograms of optical flow, HOG and MBH descriptors for HAR. The results report that, MBH outperformed optical flow, HOG, and point coordinates for each feature set. In another paper, Hussein et al. (Hussein et al., 2013) proposed a novel method for HAR which considers sequences of 3D skeleton, extracted from the depth data. For joining the skeleton locations, covariance matrix was used. The experimental results demonstrate that covariance descriptor with off-the-shelf classification method performed better as compared to the existing methods.

3. PROPOSED METHOD

For HAR, the proposed method consists of two major steps including detection of the region of interest (ROI) and action recognition. The overall architecture of the proposed system consists of four major phases: 1) data acquisition, (2) frame pre-processing, (3) ROI detection, and (4) features extraction and recognition of human actions. 1. All the phases and their sub-steps are shown in Fig. 1.

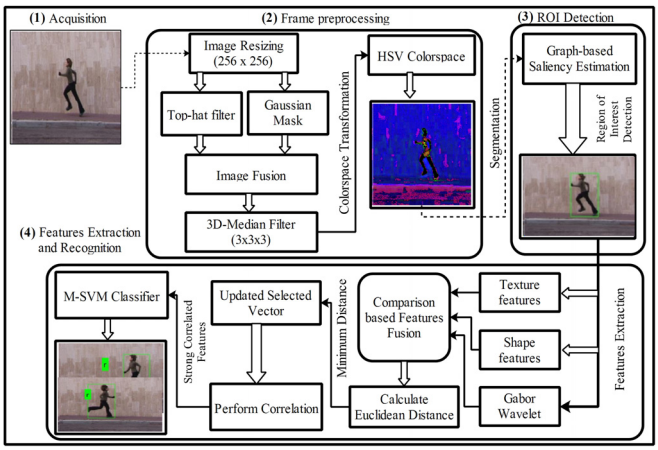

Fig. 1. System architecture of proposed method.

3.1. Detection of Region of Interest (ROI)

For ROI detection, initially pre-processing is performed on input video frames and afterwards human object is extracted from the frames using a graph-based saliency segmentation method. The process of ROI detection extracts the human region from the given frame and removes the background. There are many reasons to use ROI extraction approach such as; a) It improves the system execution time; b) It improves the recognition accuracy and reduces the error rate; c) It avoids the problem of irrelevant feature extraction. The ROI detection step further includes the pre-processing and human segmentation sub-steps as shown in Fig. 1.

3.1.1. Frame Pre-processing

The frame pre-processing is crucial to improve the visual contrast of moving regions and to reduce the noise in the input sequences. In this article, a new hybrid technique is proposed and implemented, consisting of three steps. In the first step, Top-hat and Gaussian filter is performed on the resized image. In the second step, the intensity values of a combined image are normalized and in the third step a 3D-median filter is applied on them, which removes the effects of brightness in the given frame.

The aforementioned processes are formulated as follows: Let $F(x,y)$ resizes original RGB video frame having dimensions 256×256. Let $F(x,y)$: fi $\in \mathbb{R}$ denotes the real positive pixels of each original frame and $h(x)$ denotes the structuring element. Thus the Top-hat filter is defined as:

$$F_t(x,y) = F(x,y) - h(x) \quad (1)$$

Where, structuring element $h(x)$ initialized as 13, which effects on the moving parts in the current frame. The contrast of moving object in the video frame is optimized by the addition of Gaussian function. The Gaussian function removes the brightness effects using Eq. (2):

$$G(x,y) = \frac{1}{\sigma\sqrt{2\,\pi}} e^{\frac{-1}{2}\left(\frac{F(x,y)-\mu}{\sigma}\right)^2} \quad (2)$$

Where $G(x,y)$ denotes the Gaussian function, μ denotes the mean of original RGB frame, and σ denotes the standard deviation. As mentioned earlier, the Gaussian equation removes the brightness effects and the noise, if any, in the frame and afterwards, it adds both Gaussian and top-hat frames and a new frame is obtained, which is more enhanced as compared to original top-hat and Gaussian frame. Hence, the combination of both frames is defined using Eq. (3) and Eq. (4):

$$Ad1(x,y) = \sum_{i=1}^{255}(F(x,y), F_t(x,y)) \quad (3)$$

$$Ad_F(x,y) = Ad1(x,y) - G(x,y) \quad (4)$$

Where, $Ad1(x,y)$ denotes the addition of original and top-hat filter frames, and $Ad_F(x,y)$ denotes the final addition of Gaussian frame. Afterwards, it adjusts the intensity values of a combined frame and normalizes it using gamma-correction algorithm exploiting Eq. (5) and Eq. (6):

$$\gamma C = \alpha Ad_F(x,y)^{\gamma} \quad (5)$$

$$AC(x,y) = \gamma C(Ad_F(x,y)) \quad (6)$$

Where γC denotes the gamma-correction function, α denotes the constant value, which is initialized as 2, and $AC(x,y)$ denotes the gamma-correction frame. The effects of equations (1), (2), (4), and (6) are shown in Fig. 2 (b), (c), (d), and (e) respectively. Finally, 3D-median filter on new combined frame is applied, which removes the effects of background intensities as shown in Fig. 2 (f). The final median filter frame is utilized for human extraction by using graph-cut saliency method using Eq. (7).

$$MD(x,y) = MF3(AC(x,y)) \quad (7)$$

Where, $MD(x,y)$ denotes the median filter frame and $MF3()$ denotes the 3D-median filter function (Kumar & Kumar, 2013), which is applied on gamma-correction frame.

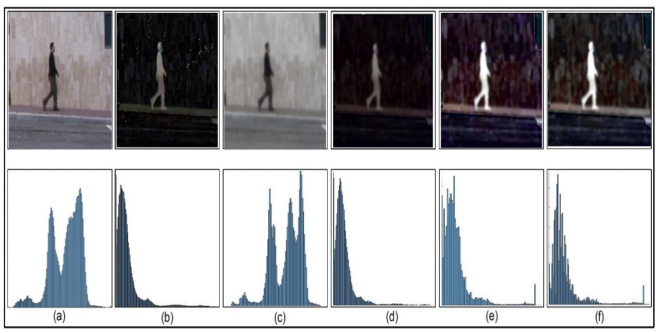

Fig. 2. Frame pre-processing effects - a) Original frame; b) top-hat frame; c) Gaussian frame; d) combined framed; e) gamma-correction frame and f) 3D-Median filter frame.

3.1.2. Human Extraction

The process of human object extraction further consists of three sub-steps including; HSV color space conversion, optical flow estimation, and graph-based saliency estimation. In the first step, HSV conversion is performed to find out the characteristics of moving objects such as their color type, movement, and brightness effect. In the second step, the Horn and Shunck (HS) optical flow method is utilized to find out the aforementioned features of moving objects in the given video frame. Finally, the extracted motion features are fed to graph-based saliency method for human extraction.

The aforementioned steps are formulated as follows: Let $MD(x,y)$ is a RGB 3D-median filtered frame. Let $F_R \in \frac{R}{MD(x,y)}$, $F_G \in \frac{G}{MD(x,y)}$, and $F_B \in \frac{B}{MD(x,y)}$ denotes the red, green and blue channel of median filtered frame. So, the HSV transformation is defined using Eq. (8) and Eq. (9):

$$Hue(x,y) = 60° \times H' \quad (8)$$

$$H'(x,y) = \begin{cases} \frac{F_G - F_R}{C} & \text{if } M = F_R \\ \frac{F_B - F_R}{C} + 2 & \text{if } M = F_G \\ \frac{F_R - F_G}{C} + 4 & \text{if } M = F_B \end{cases} \quad (9)$$

Where, $M = \max(F_R, F_G, F_B)$, $m = \min(F_R, F_G, F_B)$, and $C = M - m$. The M, m, and C denote the maximum, minimum, and Chroma components values, respectively. In addition, extraction of the Value and Saturation channels is defined using Eq. (10) and Eq. (11):

$$V(x,y) = M \quad (10)$$

$$Sit(x,y) = \begin{cases} 0 & \text{if } M = 0 \\ \frac{C}{M} & \text{Otherwise} \end{cases} \quad (11)$$

The effects of HSV transformation performed on UCF YouTube, Weizmann, and MuHAVi dataset are shown in Fig. 3. Afterwards, HS optical flow algorithm is implemented on HSV frame.

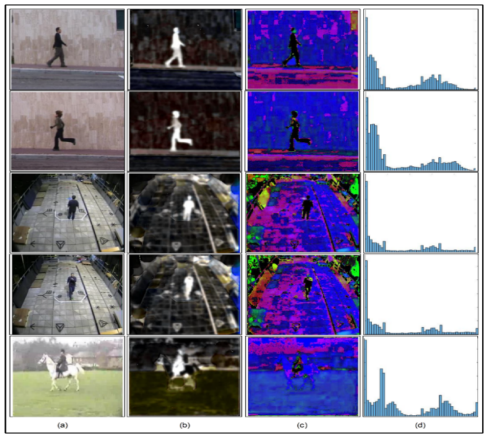

Fig. 3. HSV transformation effects - a) Original RGB frame; b) 3D-Median filter frame; c) HSV transformation, and d) histogram of HSV frame.

The HS optical flow algorithm is based on x, y, and t, where x and y denote horizontal and vertical directions and t denotes the time. The basic function of optical flow is defined using Eq (12):

$$\Delta_x U + \Delta_y V + \Delta_t = 0 \quad (12)$$

Where, U, and V denotes the horizontal and vertical optical flow of a video frame. The extracted motion features are directly fed to graph-based saliency method for extraction of moving objects in the frame. The major advantages of the extraction of motion features are: a) it neglects the background regions or static background; b) it improves the recognition accuracy and focuses, only on the human regions.

Fig. 4 shows, the moving regions that later fed to graph-based saliency method for obtaining a human silhouette frame.

Fig. 4. Motion region extraction by HS optical flow algorithm.

The visual saliency methods are organized into three different stages including, features extraction, maps activation, and normalization. In the first stage, the features are extracted at location over the panel such as human movement and human location. In the second stage, the activation maps are performed on extracted features. Finally, these extracted features are normalized and combined for concentrating mass on activation maps. In this research, an existing graph-based visual saliency (Harel et al., 2007) method is implemented to segment the moving regions. Initially, the motion features are utilized, which are extracted by HS optical flow algorithm. Then, an activation map is performed on these features to find out the dissimilarities between features. The dissimilarities between features are defined using Eq. (13) and Eq. (14):

$$\mathrm{Dis}((i,j)\|(p,q)) = \mathrm{M}(i,j) - \mathrm{M}(p,q) \tag{13}$$

After that the activation maps are normalized and combined as follows:

$$\Gamma_{\mathrm{N}}((i,j),(p,q)) \triangleq \mathrm{A}_{\mathrm{map}}(p,q).\,\mathrm{F}(i-p, j-q) \tag{14}$$

Where, $\mathrm{A}: [\mathrm{n}]^2 \rightarrow \mathbb{R}$, and n denote the pixels values which are $\{1,2,3,\ldots \mathrm{n}\}$, and $\mathrm{F}(a,b) \triangleq \exp\left(-\frac{a^2+b^2}{2\,\sigma^2}\right)$, and σ is a free parameter. The sample visual saliency effects are shown in Fig. 5.

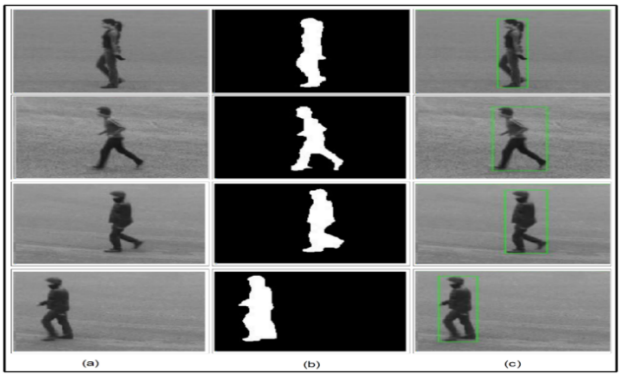

Fig. 5. Graph-based visual saliency effects- a) original frame; b) graph-based segmentation, and (c) ROI detection.

3.2. Action Representation

In a video sequence, the human actions are represented by features information which are computed using feature types-point descriptors, shape like HOG, texture such as LBP, and wavelet. The features extraction plays a key role due to well-known applications like biometrics, surveillance, agriculture etc. (Khan et al., 2018; Khan et al., 2019; Rashid et al., 2018; Sharif et al., 2018) . The proposed implemented features fusion and selection process is demonstrated in Fig. 6. Five sub-steps are performed including; 1) data acquisition in the form of binary images, b) Computing three different categories of features, 3) fusing all features into one vector by employing parallel approach along highest feature value, (4) selection of the strong features through correlation method, and (5) selection of the best features though correlation method.

3.2.1. Weighted HOG

The shape features also known as Histogram Oriented Gradient (HOG) features are originally proposed for human detection by Dalal et al. (Zhu et al., 2006) in 2005. The HOG features performed significantly well for many research domains such as object detection and in the field of medicine where each object and the infected lesion has a different shape. However, the HOG features rarely could perform well for human action recognition due to high number of sample frames. Consequently, in this article, a Weighted-HOG (W-HOG) scheme is proposed, which performed efficiently for a large number of frames. The main benefit of assigning weights is to reduce the error rate and to increase the recognition accuracy. The W-HOG features are computed using Eq. (15):

$$\varphi^{\omega}(x,y) = \mathbb{G}_{\vec{d}\times\vec{d},\sigma} \times \Gamma(x,y) \tag{15}$$

Where, $\varphi^{\omega}(x,y)$ represents weighted function, $\mathbb{G}_{\vec{d}\times\vec{d},\sigma}$ is a Gaussian matrix with dimensions$d \times d$, σ is a normalizing parameter, and $\Gamma(x,y)$ denotes the segmented frame, which is obtained from the equations (13) and (14). The final feature vector is computed using Eq. (16):

$$\varphi^{\mathrm{F}}(x,y) = \sum_{\forall(x,y)\in\epsilon} \begin{cases} \varphi^{\mathrm{F}} & \text{When } \left(\mathrm{Dir}(x,y) \times \frac{\mathbb{L}}{\pi}\right) \bmod \mathbb{L} = \epsilon \\ 0 & \text{Otherwise} \end{cases} \tag{16}$$

Where, $\left(\mathrm{Dir}(x,y) \times \frac{\mathbb{L}}{\pi}\right) \bmod \mathbb{L} = \epsilon$ is a condition of true weighted function for same pixels values, $\mathbb{L}$ denotes the number of extracted HOG features in the video frame, and π is a gradient direction parameter. Hence, the dimension of final W-HOG feature vector is 1×3780. Later on, the dimension of W-HOG vector is reduced using PCA. The PCA returns the principle score against each vector, which is sorted into ascending order in this research. Finally, the top 100, highest feature values are selected, which later on fused with texture and Gabor feature.

3.2.2. LBP Features

The Local Binary Pattern (LBP) features commonly known as texture feature, proposed by (Zhang et al., 2007) for face recognition are considered. The texture feature, a simple and effective grey scale and rotation invariant texture operator are used in several applications (Chen et al., 2017). The primary objective of using LBP features is to address the problem of fixed view, and it also provides the view-independent

analysis and maintains good identification knowledge of human activities (Kushwaha et al., 2017). In the proposed work, the LBP feature is extracted from a segmented frame, which provides a feature vector of dimensions 1 x 59 against each frame. The LBP features are calculated using Eq. (17) and Eq. (18):

$$\mathrm{LBP}(x,y) = \sum_{p=0}^{p-1} s(g_p - g_c)2^p \tag{17}$$

$$s(v) = \begin{cases} 1 & \text{if } v \geq 0 \\ 0 & \text{if } v < 0 \end{cases} \tag{18}$$

Where, the position of a central pixel is denoted by (x,y), g_c represents the intensity value of a central pixel, and g_p denotes the intensity value of neighbourhood pixel. Finally, it extracts the Gabor Wavelet features (C. Liu and Wechsler, 2002) from the segmented image and obtains a feature vector of dimensions 1×30, which are optimized by the fusion of shape and texture features discussed in the following sub-sections.

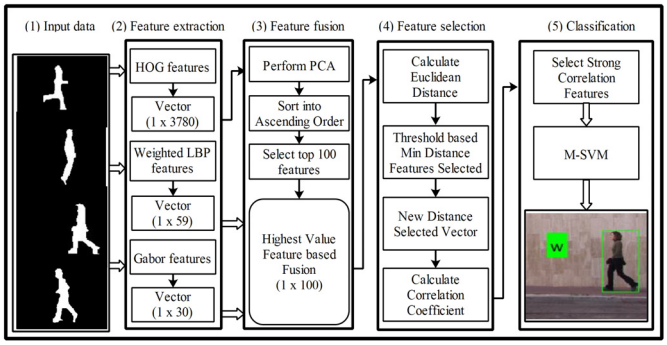

Fig. 6. Description of extracted set of features and selection.

3.2.3. *Features Fusion*

The main objective of feature fusion is to combine the characteristics of different features into one vector, which performs better for recognition as compared to individual feature type (Sun et al., 2005). In addition, it removes the redundant information between features. In this research, a simple and efficient technique is proposed which considers highest value feature based parallel fusion. Exploiting this technique, each index feature is compared to other features and the highest value feature is stored in a fused vector. Hence, the size of a fused vector depends on the size of the high feature vector. The brief description of features fusion is given in Algorithm 1.

3.2.4. *Features Selection*

In contrast to the other features reduction techniques such as PCA, the features selection methods selects the subset of features instead of original description of features (Arshad et al., 2019; Saeys et al., 2007), and for the same reason, it is considered for several applications such as video surveillance, bioinformatics, and transportation (Khan et al., 2017). In this research, the FS method is considered for supervised learning, because the irrelevant, redundant, and high dimensional features degrade the classification accuracy. The implemented FS method consists of two pipelined steps. In the first step, it calculates the Euclidean distance between a fused feature vectors and sort them into descending order. In the second step, it defines a threshold function for selecting the features with minimum distance. Euclidean distance and threshold function is calculated using Eq. (19):

$$D(fv) = \sqrt{\sum_{i=1}^{Q1} (fv_i - fv_{i+1})^2} \tag{19}$$

Where, $Q1$ denotes the length of fused vectors, which is 100. Then, the threshold function is implemented on distance vector $D(fv)$ using Eq. (20):

$$T_D = \begin{cases} \mathrm{Min} & \text{if } D_i \leq 0.5 \\ \mathrm{Max} & \text{if } D_i > 0.5 \end{cases} \tag{20}$$

The aforementioned function shows that, if the distance between two features is less than or equal to 0.5, then it is selected as a minimum distance feature, otherwise as maximum distance feature. Finally, the correlation is calculated between minimum distance features, and the strong correlation-based features for the classification. The strong correlation denotes those features having correlation value near to 1. The features having strong correlation, are later on fed to One-against-All multi SVM for final classification (Y. Liu and Zheng, 2005).

Algorithm 1: Features Fusion

Input: $\varphi^F(x,y), \mathrm{LBP}(x,y), \mathrm{GB}(x,y)$

Output: Fused Vector ← Fused(fv)

$M_L(\mathrm{vec}) \leftarrow \max(\varphi^F, \mathrm{LBP}, \mathrm{GB})$

$\varphi^F\left(\mathrm{length}(\varphi^F) + 1 : M_L(\mathrm{vec})\right) = 0$

$\mathrm{LBP}\left(\mathrm{length}(\mathrm{LBP}) + 1 : M_L(\mathrm{vec})\right) = 0$

$\mathrm{GB}\left(\mathrm{length}(\mathrm{GB}) + 1 : M_L(\mathrm{vec})\right) = 0$

$\varphi^F(x,y) \leftarrow f_1, f_2, \ldots f_n$

$\mathrm{LBP}(x,y) \leftarrow f_1, f_2, \ldots f_m$

$\mathrm{GB}(x,y) \leftarrow f_1, f_2, \ldots f_k$

$Q = m + n + k$

For $i \leftarrow 1$ to length(Q)

{

$\mathrm{FV}(i) \leftarrow \max(\varphi(i), \mathrm{LBP}(i), \mathrm{GB}(i))$

$i \leftarrow i + 1$

4. RESULTS

The experimental results are evaluated using five publically available datasets including Weizmann, KTH, MuHAVi, MSR action, and UCF YouTube. In addition, for the system evaluation, the Multi-SVM is used in this method and compared with linear SVM (L-SVM), complex tree (CT), quadratic SVM (Q-SVM), linear discriminant analysis (LDA), F-KNN, weighted KNN, and ensemble boosted tree (EBT), for performance. In this regard, eight performance parameters are considered including sensitivity, precision, accuracy rate, AUC, execution time, and name a few more. The execution time of the system is calculated for only selected testing sequences. Therefore, this method is compared with others aforementioned methods directly using accuracy rate, execution time, sensitivity rate and precision. All experiments are done on MATLAB 2017a, with a machine having Core I7 processor, 16 GB of RAM and 8GB of a graphics card.

4.1. *KTH Dataset*

The KTH human action dataset consists of six human action classes including clapping (C), walking (W), boxing (B), running (R), jogging (J), and hand waving (H). The dataset

contains total of 600 video sequences performed by 25 actors whereas all video sequences are performed in four different scenarios.

For experiments, 60:40 approach is utilized for training and testing. The 10 fold cross validation (10fV) is performed and achieved maximum recognition accuracy of 99.5% on M-SVM with FP rate of 0.00, sensitivity of 99.33%, and AUC 1.00 as given in Table 1.

Another method named LDA also outperforms for KTH dataset and achieved recognition accuracy of 99.2% with FP rate 0.0016, and AUC 1.00. The results show that the testing compile time of M-SVM is 109.61 seconds and of LDA is112.20 seconds. Thus the proposed method performed better on M-SVM as presented in Table 1, which is further verified using confusion matrix (CM) as described in Table 2. In addition, a comparison is conducted with few existing techniques, and promising results are found in terms of accuracy for KTH dataset, as presented in Table 3.

Table 1. Recognition performance on KTH dataset- The symbol Sen denotes the sensitivity, Prec denotes the precision rate, FN denotes the false negative rate, and Acc is accuracy, respectively

M	Sen (%)	Prec (%)	FN (%)	AUC	Acc (%)	Time (S)
L-SVM	97.5	97.6	2.5	0.99	97.5	130.4
CT	87.5	87.6	12.6	0.94	87.4	177.2
QSVM	99.16	98.83	1.00	**1.00**	99.0	137.5
LDA	99.16	99.16	0.8	**1.00**	99.2	112.2
MSVM	**99.33**	**99.26**	**0.5**	**1.00**	**99.5**	**109.6**
F-KNN	99.16	99.00	0.9	0.99	99.1	181.3
WKNN	98.16	98.5	1.8	**1.00**	98.2	188.7

Table 2. Confusion matrix of KTH dataset

Action Class	B	Cl	HW	J	R	W
Boxing	**100%**					
Clapping		**100%**				
Hand W			**100%**			
Jumping				**99%**	0.5%	0.5%
Running				1%	**98%**	1%
Walking				1%		**99%**

Table 3. Comparison for KTH dataset with existing techniques

Reference	Year	Accuracy
(Vu et al., 2015)	2015	96.20%
(Shao et al., 2014)	2014	95.00%
(Shi et al., 2013)	2013	93.00%
(Sharif et al., 2019)	2019	98.12%
Proposed	**2019**	**99.50%**

4.2. Weizmann Dataset

The Weizmann dataset was originally developed in 2005. It comprises of 90 video sequences of 10 human action classes including running, walking, waving, and few more. All videos are performed by 9 actors within a static environment. For experiment 60:40 ratio is considered and perform 10fV. The 40% videos are used for testing the proposed HAR system, whereas remaining for training. The maximum recognition accuracy of 98.2%, in case of Weizmann dataset is recorded, which is obtained on M-SVM with FNR 1.8, the precision rate of 98.1, and sensitivity of 98.2% is achieved. Moreover, the LDA and F-KNN also performed better and achieved recognition accuracy of 98.0%. The recognition results are presented in Table 4, and further verified by Table 5. Finally, in the Table 6, a comparison is conducted for Weizmann dataset with the existing techniques. The results shows that the proposed method outperforms in terms of accuracy measure.

Table 4. Recognition results on WEIZMANN dataset. The M denotes the method or classifier and T represents the time

M	Sen (%)	Pre (%)	FNR (%)	AUC	Acc (%)	T (sec)
L-SVM	94.6	94.5	5.3	0.99	94.7	142
CT	79.8	79.9	20.2	0.90	79.8	197
QSVM	97.7	97.7	2.1	0.99	97.9	145
LDA	98.0	98.0	2.0	1.00	98.0	165
MSVM	**98.2**	**98.1**	**1.8**	**1.00**	**98.2**	**92.3**
F-KNN	97.8	98.0	2.0	0.99	98.0	101
WKNN	96.0	96.0	4.0	0.99	96.0	166
EBT	94.9	95.2	4.9	0.99	95.1	196

Table 5. CM for Weizmann dataset. The all values are in the form of percentage (%)

C	B	K	J	P	R	D	S	W	W1	W2
B	**99**		1					1		
K		**99**				1				
J			**100**							
P				**100**						
R					**92**		6	2		
D						**99**		1		
S					5		**95**			
W					2			**98**		
W1					1				**99**	
w2									1	**99**

Table 6. Comparison with existing methods for Weizmann dataset

Reference	Year	Accuracy
(Moussa et al., 2015)	2015	96.66%
(Abdul-Azim and Hemayed, 2015)	2015	97.77%
(Vishwakarmaet al., 2015)	2015	96.64%
(Sharif et al., 2019)	2019	98.12%
Proposed	**2019**	**98.20%**

4.3. UCF YouTube Action Dataset

The UCF YouTube dataset consists of total of 11 human action classes including driving, horse riding, golf swing, basketball shooting, and few more. The dataset contains many challenges due to change in camera motion, pose and appearance problem, complex background, brightness issues and object scale. In the dataset, each action class consists of 25 groups, where each group has more than 4 action sequences. In this research, 9 action classes are selected including basketball, horse riding, tennis swing, walk with Dog, boxing, golfs swing, soccer juggling, swing, and volleyball spiking. For the experiments 60:40 ratio is used for training and testing. The results are calculated using 10-fold cross-validations. The results show, recognition accuracy of 100% on M-SVM with FPR 0.0%, and the precision rate of 100%. The testing execution time of M-SVM is recorded 262.39 seconds, which is better than other classification methods as presented in Table 7. Finally, a comparison is carried out with the existing HAR methods on the YouTube dataset, as given in Table 8. The results proof that the proposed method outperforms the other methods.

Table 7. Recognition results on UCF YouTube dataset

Method	Sen	Pre	FN	AUC	Acc	Time
L-SVM	87.5	87.6	12.6	0.94	87.4	500.5
CT	96.33	96.22	3.8	0.98	96.2	604.6
Q-SVM	97.70	97.7	2.1	0.99	97.9	392.3
LDA	97.50	97.6	2.5	0.99	97.5	307.8
MSVM	**100**	**100**	**0.0**	**1.000**	**100**	**262.3**
F-KNN	99.88	99.88	0.2	1.000	99.8	514.4
W-KNN	99.64	99.60	0.4	99.99	99.6	510.6

Table 8. Comparison with existing methods for UCF YOUTUBE dataset

Reference	Year	Accuracy
(Moussa et al., 2015)	2015	96.66%
(Abdul-Azim and Hemayed, 2015)	2015	97.77%
(Vishwakarma et al., 2015)	2015	96.64%
Proposed	**2019**	**98.20%**

4.4. HMDB 51 Dataset

The hmdb51 action recognition dataset includes total of 51 human actions like push-up, jump, stand-up, wave, and few more. In this research, we considered 12 human action categories including Ride Bike, Run, Jump, kick, punch, push up, sit-up, stand, throw, turn, walk, and wave. All the experiments are performed using 10fV with the ratio of 60:40 and achieve the recognition accuracy of 92.6% as presented in Table 9. The M-SVM outperformed and achieved this accuracy in 47.235 seconds. Moreover, the weighted KNN also performed better and achieved recognition accuracy of 90.9% with sensitivity rate of 89.67%, which is higher after M-SVM. The recognition accuracy of M-SVM is presented in Table 9. Finally, the recognition results of the proposed method are compared with the existing methods and presented in Table 10, which shows that the proposed method outperforms the existing method.

Table 9. Recognition results on HMDB51 dataset

Method	Sen (%)	Pre (%)	FNR (%)	FPR	AUC	Acc (%)
L-SVM	62.99	64.73	35.6	0.031	0.91	64.4
CT	62.96	64.21	35.4	0.032	0.91	64.6
Q-SVM	42.67	44.33	55.2	0.058	0.78	44.8
LDA	50.99	52.97	43.2	0.044	0.87	53.2
M-SVM	**92.17**	**92.33**	**7.4**	**0.007**	**0.97**	**92.6**
F-KNN	58.75	61.33	38.5	0.037	0.89	61.5
W-KNN	89.67	89.67	9.1	0.013	0.96	90.9

Table 10. Comparison of proposed algorithm for HMDB51 dataset

Method	Year	Accuracy (%)
(Girdhar and Ramanan, 2017)	2017	52.2
(Duta et al., 2017)	2017	61.0
(Bilen et al., 2016)	2016	65.2
(Wang et al., 2017)	2017	71.0
Proposed	***2019***	***92.6***

4.5. Discussion

In this article, a novel HAR method is proposed, which consists of two major steps including ROI detection and recognition. The ROI detection step further comprises of two sub-steps including frame contrast stretching and human extraction. Similarly, the classification step consists of three sub-steps including feature extraction, best feature selection, and recognition as shown in Fig. 1. The frame stretching effects are given in Fig. 2 and Fig. 3, which are later utilized for motion estimation and human extraction as given in Fig. 4 and Fig. 5. In action recognition step, W-HOG, LBP, and Gabor feature are extracted and fused based on parallel method. The fused features are later selected by proposed ED and SC based method as shown in Fig. 6. The proposed method is tested on four publically available datasets including Weizmann, KTH, Muhavi, UCF YouTube, and HMDB51. The maximum recognition results are achieved on M-SVM with 92.3%, 99.5%, 92.6%, and 100% for Weizmann, KTH, UCF YouTube, and HMDB51, respectively. The recognition results are given in Table 1, 4, 7, and 9. For further validation of results confusion matrixes are given in Table 2 and 5.

Moreover, a comprehensive comparison of proposed method is performed for each selected dataset as given in Table 3, 6, 8, and 10, which clearly show that the proposed method is

significantly performed on M-SVM as compared to existing methods.

The more recent, Sharif et al. (Sharif et al., 2019) presented a HAR method by employing weighted segmentation and feature reduction. The presented method works well against complex background data due to the pre-processing step. The pre-processing step followed the segmentation step in which harmonic mean is computed at the initial phase of the processed frame and updates the background through mean and SD value which followed the final weight function. Later, ternary features are extracted and serially fused. They also introduced a strong correlation based feature reduction method which outperforms on selected datasets (Weizmann, KTH, WVU, Muhavi, UCF sports, and MSR action of to achieve a recognition rate of 98.12, 99.4, 99.1, 99.5, 99.40, and 90.20%, respectively) as compared to individual feature sets. Sharif et al. (Sharif et al., 2017) presented a hybrid framework for human detection and action recognition. The human in the video sequences is detected through a fusion of EM algorithms and uniform method. Later, LBP, HOG, and texture information is computed but the fusion process involves some noisy data which effects on the system performance. This kind of issue is resolved by authors through best data selection name distance calculation and PCA along with Entropy method. Four datasets are utilized and achieve above 95% accuracy with the lowest error rate.

The above results show the improvement of proposed results as compare to existing techniques.

5. CONCLUSION

The frame stretching shows a major role for improving the visual contrast of a given frame which later plays a key role for ROI detection and robust feature extraction. In addition, the robust and prominent feature extractions are important for improving the system classification accuracy because high dimensional and irrelevant features may affect overall system performance in the term of accuracy and execution time. The proposed method is evaluated on five publically available datasets and the accuracy of more than 90% for all datasets is achieved. Moreover, the comparison with existing methods proved the system authentication. In future, deep learning method needs to be implemented on the same datasets, and evaluation should be carried out using all 51 action classes.

ACKNOWLEDGEMENT

Prince Sultan University Riyadh KSA, Grant/Award Number: 11-02-2019

REFERENCES

Abdul-Azim, H. A., & Hemayed, E. E. (2015). Human action recognition using trajectory-based representation. *Egyptian Informatics Journal, 16*(2), 187-198.

Arshad, H., Khan, M. A., Sharif, M., Yasmin, M., & Javed, M. Y. (2019). Multi-level features fusion and selection for human gait recognition: an optimized framework of Bayesian model and binomial distribution. *International Journal of Machine Learning and Cybernetics*, 1-18.

Aurangzeb, K., Haider, I., Khan, M. A., Saba, T., Javed, K., Iqbal, T., . . . Sarfraz, M. S. (2019). Human Behavior Analysis Based on Multi-Types Features Fusion and Von Nauman Entropy Based Features Reduction. *Journal of Medical Imaging and Health Informatics, 9*(4), 662-669.

Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., & Gould, S. (2016). *Dynamic image networks for action recognition.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Cardinaux, F., Bhowmik, D., Abhayaratne, C., & Hawley, M. S. (2011). Video based technology for ambient assisted living: A review of the literature. *Journal of Ambient Intelligence and Smart Environments, 3*(3), 253-269.

Chang, S.-F. (2002). The holy grail of content-based media analysis. *IEEE MultiMedia, 9*(2), 6-10.

Chen, C., Liu, M., Liu, H., Zhang, B., Han, J., & Kehtarnavaz, N. (2017). Multi-Temporal Depth Motion Maps-Based Local Binary Patterns for 3-D Human Action Recognition. *IEEE Access, 5*, 22590-22604.

Duta, I. C., Uijlings, J. R., Ionescu, B., Aizawa, K., Hauptmann, A. G., & Sebe, N. (2017). Efficient human action recognition using histograms of motion gradients and VLAD with descriptor shape information. *Multimedia Tools and Applications*, 1-28.

Girdhar, R., & Ramanan, D. (2017). *Attentional pooling for action recognition.* Paper presented at the Advances in Neural Information Processing Systems.

Harel, J., Koch, C., & Perona, P. (2007). *Graph-based visual saliency.* Paper presented at the Advances in neural information processing systems.

Hussein, M. E., Torki, M., Gowayyed, M. A., & El-Saban, M. (2013). *Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations.* Paper presented at the IJCAI.

Khan, M. A., Akram, T., Sharif, M., Awais, M., Javed, K., Ali, H., & Saba, T. (2018). CCDF: Automatic system for segmentation and recognition of fruit crops diseases based on correlation coefficient and deep CNN features. *Computers and electronics in agriculture, 155*, 220-236.

Khan, M. A., Akram, T., Sharif, M., Javed, M. Y., Muhammad, N., & Yasmin, M. (2018). An implementation of optimized framework for action classification using multilayers neural network on selected fused features. *Pattern Analysis and Applications*, 1-21.

Khan, M. A., Lali, M. I., Sharif, M., Javed, K., Aurangzeb, K., Haider, S. I., . . . Akram, T. (2019). An Optimized Method for Segmentation and Classification of Apple Diseases based on Strong Correlation and Genetic Algorithm based Feature Selection. *IEEE Access*.

Khan, M. A., Sharif, M., Javed, M. Y., Akram, T., Yasmin, M., & Saba, T. (2017). License number plate recognition system using entropy-based features selection approach with SVM. *IET Image Processing*.

Kumar, K. G., & Kumar, K. K. (2013). 3D Median Filter Design for Iris Recognition. *International Journal of Modern Engineering Research, 3*(5), 3008-3011.

Kushwaha, A. K. S., Srivastava, S., & Srivastava, R. (2017). Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns. *Multimedia Systems, 23*(4), 451-467.

Liu, C., & Wechsler, H. (2002). Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing, 11*(4), 467-476.

Liu, Y., & Zheng, Y. F. (2005). *One-against-all multi-class SVM classification using reliability measures.* Paper presented at the Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on.

Moussa, M. M., Hamayed, E., Fayek, M. B., & El Nemr, H. A. (2015). An enhanced method for human action recognition. *Journal of advanced research, 6*(2), 163-169.

Ng, W. W., Li, J., Zhang, J., Wu, Q., & Li, J. (2017). *Visual words selection for human action recognition using rbfnn via the minimization of L-GEM.* Paper presented at the Wavelet Analysis and Pattern Recognition (ICWAPR), 2017 International Conference on.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing, 28*(6), 976-990.

Rashid, M., Khan, M. A., Sharif, M., Raza, M., Sarfraz, M. M., & Afza, F. (2018). Object detection and classification: a joint selection and fusion strategy of deep convolutional neural network and SIFT point features. *Multimedia Tools and Applications*, 1-27.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics, 23*(19), 2507-2517.

Shao, L., Zhen, X., Tao, D., & Li, X. (2014). Spatio-temporal Laplacian pyramid coding for action recognition. *IEEE Transactions on Cybernetics, 44*(6), 817-827.

Sharif, M., Khan, M. A., Akram, T., Javed, M. Y., Saba, T., & Rehman, A. (2017). A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection. *EURASIP Journal on Image and Video Processing, 2017*(1), 89.

Sharif, M., Khan, M. A., Faisal, M., Yasmin, M., & Fernandes, S. L. (2018). A framework for offline signature verification system: Best features selection approach. *Pattern Recognition Letters*.

Sharif, M., Khan, M. A., Zahid, F., Shah, J. H., & Akram, T. (2019). Human action recognition: a framework of statistical weighted segmentation and rank correlation-based selection. *Pattern Analysis and Applications*, 1-14.

Shi, F., Petriu, E., & Laganiere, R. (2013). *Sampling strategies for real-time action recognition.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Siddiqui, S., Khan, M. A., Bashir, K., Sharif, M., Azam, F., & Javed, M. Y. (2018). Human action recognition: a construction of codebook by discriminative features selection approach. *International Journal of Applied Pattern Recognition, 5*(3), 206-228.

Sun, Q.-S., Zeng, S.-G., Liu, Y., Heng, P.-A., & Xia, D.-S. (2005). A new method of feature fusion and its application in image recognition. *Pattern Recognition, 38*(12), 2437-2448.

Talukdar, J., & Mehta, B. (2017). Human Action Recognition System using Good Features and Multilayer Perceptron Network. *arXiv preprint arXiv:1708.06794*.

Vishwakarma, D., Dhiman, A., Maheshwari, R., & Kapoor, R. (2015). Human motion analysis by fusion of silhouette orientation and shape features. *Procedia Computer Science, 57*, 438-447.

Vishwakarma, D. K., Gautam, J., & Singh, K. (2018). A Robust Framework for the Recognition of Human Action and Activity Using Spatial Distribution Gradients and Gabor Wavelet *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications* (pp. 103-113): Springer.

Vu, T. L., Do, T. D., Jin, C.-B., Li, S., Nguyen, V. H., Kim, H., & Lee, C. (2015). Improvement of Accuracy for Human Action Recognition by Histogram of Changing Points and Average Speed Descriptors. *Journal of Computing Science and Engineering, 9*(1), 29-38.

Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision, 103*(1), 60-79.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2017). Temporal Segment Networks for Action Recognition in Videos. *arXiv preprint arXiv:1705.02953*.

Weinland, D., Ronfard, R., & Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding, 115*(2), 224-241.

Yi, Y., & Wang, H. (2017). Motion keypoint trajectory and covariance descriptor for human action recognition. *The Visual Computer*, 1-13.

Zhang, L., Chu, R., Xiang, S., Liao, S., & Li, S. Z. (2007). *Face detection based on multi-block lbp representation.* Paper presented at the International Conference on Biometrics.

Zheng, Y., Yao, H., Sun, X., Zhao, S., & Porikli, F. (2018). Distinctive action sketch for human action recognition. *Signal Processing, 144*, 323-332.

Zhu, Q., Yeh, M.-C., Cheng, K.-T., & Avidan, S. (2006). *Fast human detection using a cascade of histograms of oriented gradients.* Paper presented at the Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on.