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 [4], Tassawar Iqbal [5]

1 Department of EE, COMSATS University Islamabad, Wa Cantt
2 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 Department of Computer Science COMSATS University Islamabad, Wa Cantt, Pakistan
***Corresponding Author*:** attique@ciitwah.edu.pk*

Abstract: Extracting salient and most prominent features from a given video sequence is one of the most critical steps in human action recognition. In this article, proposed a novel method for human action recognition, which efficiently addresses the problem of selection of most prominent and robust features in the feature selection step. Three types of features are fused based on their highest values and later most optimal features are selected with an implementation of a novel Euclidean distance (ED) and strong correlation (SC) method. In the final phase, selected features are classified using multi-class support vector machine (M-SVM). Four publically available datasets are utilized including Weizmann, KTH, UCF YouTube, and HMDB51 with an improved classification accuracy of average more than 94%. Experimental results authenticate our claim that the proposed method outperforms compared to several existing methods.

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

1. INTRODUCTION

In the field of computer vision, Human Action Recognition (HAR) sought attention of many researchers since the last two decades, due to its applications in the fields of video indexing (Weinland, Ronfard, & Boyer, 2011), human-computer interaction (Poppe, 2010), intelligent surveillance and video surveillance (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 areas such as airports, 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, monitoring systems that can identify activities of daily living used in the home for the care of old people and children's, reduced burden of care and anxiety (Cardinaux, Bhowmik, Abhayaratne, & Hawley, 2011). In daily life, several actions are performed by human for example; waving, jumping, punching, walking, running, sliding and to name a few more. In literature, Several HAR methods are available, which were focusing on different challenges such as, variations in illumination, complex background, and occlusion. In HAR, the major problem is the extraction of the human region because several objects are moved in the given sequence but the human is treated as a region of interest. Secondly, in the features extraction phase, the irrelevant and redundancy between features degrade the accuracy of action recognition.

1.1. Problem Statement

In this article, dealt with several types of challenges such as: a) low contrast video acquisition because of complex background; b) change in the human viewpoint and variation; c) occlusion between human and other objects; d) measuring the body size, symmetry and support level of a human at the time of performing the action because each person has their own body size and style; e) extraction of most useful features because the irrelevant features degrade the recognition accuracy, and f) high dimension of extracted features. These listed challenges have degraded the accuracy of HAR. Therefore, it is important to resolves above challenges for improving the accuracy of HAR and also improves the computation time, which is an important factor in latest research in this area.

1.2. Contribution

Therefore, 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 such as preprocessing, human detection, features extraction, and recognition. In this article, we deal with these steps and implement a new method in which preprocessing step incorporates the contrast of moving region of given video sequences, which is later utilized in the human detection phase. Thereafter, extract features such as texture, shape, and Gabor. The extracted features are fused based of higher value feature and select the best features by utilizing Euclidean distance and strong correlation. The selected features are finally classified by utilizing several methods such as M-SVM and compare their performance with several state-of-

the-art classification methods such as KNN, Ada-boost, CT, and LDA. The major contributions are: a) A frame contrast stretching method is introduced based on three series of sub-steps. In the first step, Top-hat and Gaussian filter is performed on the original RGB frame and add their intensity values in one frame. The artifacts in combined frames are removed by a 3D-Median filter in the second step, which is later separate with their background by utilizing HSV color space in the last step. b) Extract Weighted HOG (WH) features from human silhouette and performed PCA. The PCA return score values for each vector, which is sort into ascending order and select the top 100 features based on their high values. 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.

1.3. Paper Organization

The rest of this article is organized as follows: Section 2 present the related work of recent works. Proposed work is discussed in Section 3. Section 4 demonstrates the experimental results, and Section 5 followed the conclusion.

2. RELATED WORK

Many techniques have been recently 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 & 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. It achieved best performance as compared to MBH, HOG and HOF. Jonti et al. (Talukdar & Mehta, 2017) introduced a novel approach for HAR by the combination of good features and MLP. 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, Li, Zhang, Wu, & Li, 2017) extended 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 it performs well for video datasets. Ying et al. (Zheng, Yao, Sun, Zhao, & Porikli, 2018) presented a 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 completing both processes, the mining is carried out. Finally, four types of sketch pooling methods are performed to get a uniform representation of human actions for given video sequences. Dinesh et al. (D. K. Vishwakarma, Gautam, & Singh, 2018) introduced a novel method for HAR based on spatial distribution gradients and Gabor wavelet. Initially, the human silhouette is extracted using texture-based segmentation. After that, shape and view based features are extracted, and Gabor wavelet features are added to improving its strength. Finally, features are fused to obtain a robust feature vector, which is classified by SVM. According to results, the technique achieved the best accuracy. Wang et al. (H. Wang, Kläser, Schmid, & Liu, 2013) used point coordinates, histograms of optical flow, HOG and MBH descriptors for HAR. Performance evaluation is evaluated on each feature set and MBH outperforms as compare to optical flow, HOG, and point coordinates. Hussein et al. (Hussein, Torki, Gowayyed, & El-Saban, 2013) have presented a novel method for HAR based on the 3D skeleton sequences, extracted from the depth data. Covariance matrix was used for joining the skeleton locations. The experimental results demonstrate that covariance descriptor with off-the-shelf classification method performs better as compared to the existing methods.

3. PROPOSED METHOD

The proposed method consists of two major steps including; detection of the region of interest (ROI) and action recognition. In ROI detection step, initially preprocessing is performed on input video frames and then human is extracted from the frames by utilizing a graph-based saliency segmentation method. In the second step, the best features out of all fused multiple features are selected for action recognition using M-SVM. The system architecture of proposed method is shown in Fig. 1.

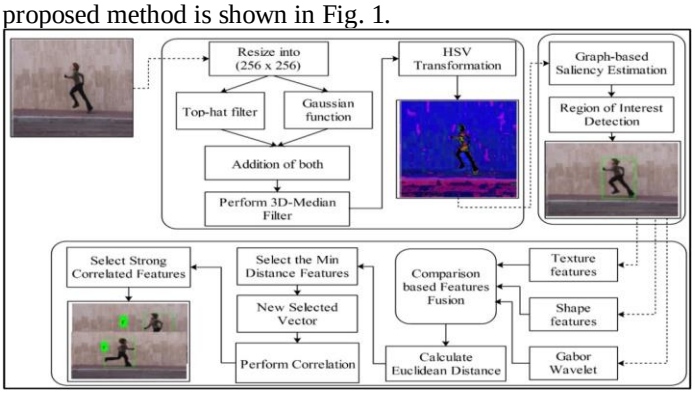

Fig. 1: A system architecture of proposed method.

3.1. Detection of Region of Interest (ROI)

The principle 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 consists of series of sub-steps including preprocessing and human segmentation as shown in Fig. 1. The preprocessing sub-step, initially resizes the video frame into 256 × 256 and performs Top-hat and Gaussian filter separately. After that, it combines both Gaussian and Top-hat frame into one frame using addition, and it adjusts their intensity values using normalization function. Finally, a 3D-median filter is applied on the normalized frame to remove hidden noise such as brightness effects. The human segmentation sub-step performs HSV color space on a median filtered frame and implements a graph-based saliency method on them. The graph-based saliency method extracts the human from them, which is later used for feature extraction. The detail of each sub-step is described in the following sub-sections.

3.1.1. Frame Preprocessing

The purpose of frame preprocessing step is to improve the visual contrast of moving regions and to reduce the noise in the input sequences. In this research, a new hybrid technique is 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 a 3D-median filtered is applied on them, which removes the effects of brightness in the given frame. The above processes are defined as follows:

Let $F(x,y)$ resizes original RGB video frame having dimensions256×256. Let $F(x,y)$: fi $\in \mathbb{R}$ denotes the real positive pixels of each original frame and $h(x)$ denotes the structuring element. Then 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 as follows:

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

Where $G(x,y)$ denotes the gaussian function, μ denotes the mean of original RGB frame, and σ denotes the standard deviation. The Gaussian equation removes the brightness effects and the noise, if any, in the frame. Later on, it adds both Gaussian and top-hat frames and obtained a new frame, which is more enhanced as compared to original top-hat and Gaussian frame. The combination of both frames is defined in the following equations:

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

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

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 as follows:

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

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

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 equation (1), (2), (4), and (6) are shown in Fig. 2 (b)(c)(d)(e). Finally, 3D-median filter on new combined frame, which removes the effects of background intensities as shown in Fig. 2 (f) are performed. The final median filter frame is utilized for human extraction by using graph-cut saliency method.

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

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.

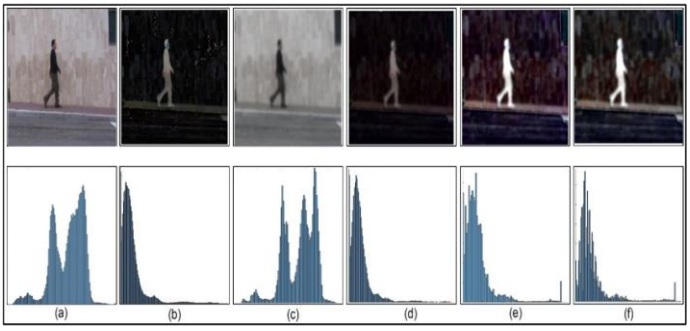

Fig. 2: Frame preprocessing effects. a) Original frame; b) top-hat frame; c) Gaussian frame; d) combined framed; e) gamma-correction frame and f) 3D-Median filter frame.

3.2. Human Extraction

Extraction of human is performed using series of three 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 motion 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 detail description of each step is given below.

Let $MD(x,y)$ is a RGB 3-dimensional 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. Then, the HSV transformation is defined as follows:

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

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

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 denotes the maximum, minimum, and chroma components values, respectively. Moreover, extract the Value and Saturation channels which are defined as follows:

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

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

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

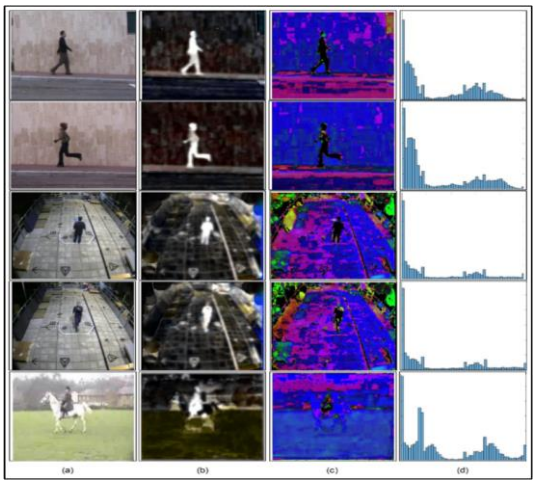

Fig. 3: HSV transformation effects. a) Original RGB frame; b) 3D-Median filter frame; c) HSV transformation, and d) histogram of HSV frame.

The HS optical flow algorithm 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:

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

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 video frames of KTH dataset are utilized and they obtained motion features of moving regions. Later on, these motion frames are fed to graph-based saliency method and obtain a human silhouette frame.

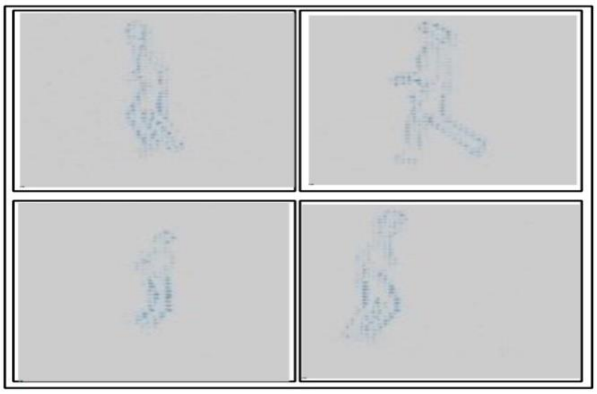

Fig. 4: Motion region extraction by HS optical flow algorithm.

The visual saliency methods are organized into three different stages including, features extraction, maps activation, and normalization. In the first step, the features are extracted at location over the panel such as human movement and human location. In the second step, 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, Koch, & Perona, 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 as follows:

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

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

$$\Gamma_N((i,j),(p,q)) \triangleq A_{map}(p,q).F(i-p,j-q) \quad (14)$$

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

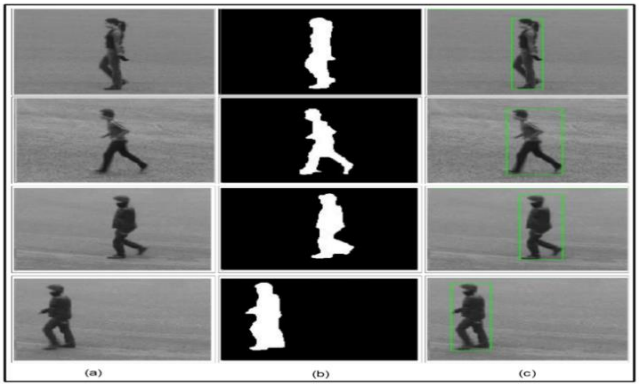

Fig. 5: Graph-based visual saliency effects: a) original frame; b) graph-based segmentation, and (c) ROI detection

3.3. Action Representation

In a video sequences, the human actions are represented by extraction of several types of features such as shape, texture, color, and point. Features extraction plays a key role in machine learning due to their several applications including video surveillance, robotics, and biometrics (Sharif, Khan, Faisal, Yasmin, & Fernandes, 2018). In this research, three types of features are extracted such as texture, shape, and Gabor wavelet. The extracted features are fused into one vector based on their highest feature values. The fusion of extracted features is performed to reduce the recognition error rate and to improve the overall recognition accuracy. After that, the best features from the fused vector based on their minimum Euclidean distance and strong correlation are selected. The aim of features selection is to pick the most prominent features and to remove the redundancy between them, because the redundancy between features degrades the recognition accuracy. The major advantage of features selection is to improve the system execution time. The detailed description of features extraction and best features selection is shown in Fig. 6.

3.3.1. Weighted HOG

Histogram Oriented Gradient (HOG) features also known as shape features are originally proposed by Dalal et al. (Zhu, Yeh, Cheng, & Avidan, 2006) in 2005 for human detection. 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. But, the HOG features have not performed well for human action recognition due to a high number of sample frames. Consequently, in this article, a Weighted-HOG (W-HOG) scheme is proposed, which performs efficiently for a large number of frames. The main benefit of assigning weights is to reduce the error rate and increase the recognition accuracy. The W-HOG features are computed as follows:

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

where $\varphi^{\omega}(x,y)$ represented weighted function, $\mathbb{G}_{\vec{d}\times\vec{d},\sigma}$ is a Gaussian matrix with dimension $d \times d$, σ is a normalizing parameter, and $\Gamma(x,y)$ denotes the segmented frame, which is obtained from equation (13) and (14). The final feature vector is computed as follows:

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

Where, $\left(\mathrm{Dir}(x,y) \times \frac{\mathbb{L}}{\pi}\right) \bmod \mathbb{L} = \epsilon$ is 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. After that, the dimension of W-HOG vector is reduced using PCA. The PCA returns the principle score against each vector, which we have sorted into ascending order. Finally, it selects the top 100, highest feature values, later on these are fused with texture and Gabor feature.

3.3.2. *LBP Features*

The Local Binary Pattern (LBP) features, originally proposed by (Zhang, Chu, Xiang, Liao, & Li, 2007) for face recognition are also known as texture feature. The texture feature a simple and effective grayscale and rotation invariant texture operator used in several applications (Chen et al., 2017). The major aim of using LBP features is to address the problem of fixed view, it provides the view-independent analysis and maintains good identification knowledge of human activities (Kushwaha, Srivastava, & Srivastava, 2017). In this research, 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 as follows:

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

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

Where, (x,y) denotes the position of a central pixel, g_c denotes the intensity value of a central pixel, and g_p denotes the intensity value of neighborhood pixel. Finally, it extracts the Gabor Wavelet features (C. Liu & Wechsler, 2002) from segmented image and a feature vector of dimensions 1×30 is obtained, which are optimized by fusion of shape and texture features in below section.

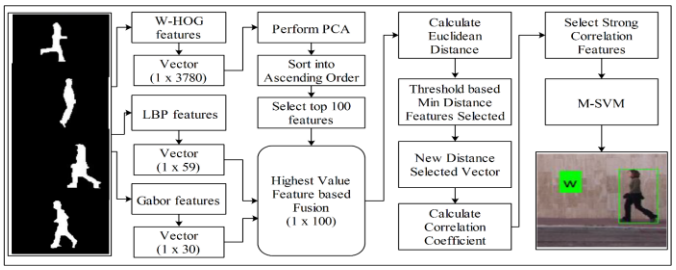

Fig. 6: Description of extracted set of features and selection.

3.3.3. *Features Fusion*

The features fusion is an active research topic in the fled of pattern recognition. (Sun, Zeng, Liu, Heng, & Xia, 2005). The major use of feature fusion is to combine the characteristics of different features into one vector, which performs well for recognition as compared to individual feature type. Moreover, it removes the redundant information between features. In this research, a simple and efficient technique is realized based on 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.

Therefore, 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.

3.3.4. *Features Selection*

In present times, the Features Selection (FS) method has become a manifest need in several applications such as video surveillance, bioinformatics, and transportation (Khan et al., 2017). In contrast to other features reduction techniques such as PCA, the features selection methods selects the subset of features instead of original description of features (Saeys, Inza, & Larrañaga, 2007). 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 is consisting of two pipelined steps. In the first step, it calculates the Euclidean distance between a fused feature vector and sort them into descending order. In the second step, it defines a threshold function, to select the minimum distance features. The calculation of Euclidean distance and threshold function is defined as follows:

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

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

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

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

Algorithm 1: Features Fusion

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

Output: Fused Vector $\leftarrow$ Fused(fv)

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

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

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

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

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

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

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

$Q = m + n + k$

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

{

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

$i \leftarrow i + 1$

4. RESULTS

The experimental results are evaluated on five publically available datasets including Weizmann, KTH, Muhavi, MSR action, and UCF Youtube. For the system evaluation, the Multi-SVM is used for the proposed method and compared with linear SVM (L-SVM), complex tree (CT), quadratic SVM (Q-SVM), linear discriminant analysis (LDA), fine KNN (F-KNN), weighted KNN (W-KNN), and ensemble boosted tree (EBT), for performance. In this regard, 8 performance parameters including sensitivity, precision, false discovery rate (FDR), false negative rate (FNR), false positive rate (FPR), the area under the curve (AUC), accuracy rate, and execution time are considered. The execution time of the system is calculated for only selected testing sequences. Therefore, comparison of the proposed method with others mentioned earlier, is directly performed 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 6 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, which are performed by 25 actors in four different scenarios including outdoors, indoor, outdoors with scale variation, and outdoors with various clothes.

For experiments, 60:40 is used for training and testing. The 60% video sequences are utilized for training the system and remaining sequences are utilized for testing. All experimental results are obtained using 10-fold cross-validation. The proposed method achieved maximum recognition accuracy of 99.5% on M-SVM with FP rate 0.00, sensitivity of 99.33%, and AUC 1.00 as given in Table 1.

Similarly, the LDA also performed well and achieved recognition accuracy of 99.2% with FP rate 0.0016, and AUC 1.00. The results show that the test execution 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 confirmed by their confusion matrix in Table 2. Moreover, comparison of the proposed technique with the existing methods is presented in Table 3, which shows that the proposed method performed better as compared to latest action recognition.

Table 1. Recognition performance on KTH dataset

Meth	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
Q-SVM	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
M-SVM	**99.33**	**99.26**	**0.5**	**1.00**	**99.5**	**109.6**
F-KNN	99.16	99.00	0.9	0.99	99.1	181.34
W-KNN	98.16	98.5	1.8	**1.00**	98.2	188.73

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

Reference	Year	Accuracy
(Vu et al., 2015)	2015	96.20%
(Shao, Zhen, Tao, & Li, 2014)	2014	95.00%
(Shi, Petriu, & Laganiere, 2013)	2013	93.00%
Proposed	**2017**	**99.50%**

4.2. Weizmann Dataset

The Weizmann dataset was originally developed in 2005. It consists 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: 60% videos from each class are utilized for training the classifier and remaining 40% videos are used for testing. All experimental results are evaluated using 10-fold cross-validation. The maximum recognition accuracy, in case of Weizmann dataset is recorded 98.2%, which is obtained on M-SVM with FNR 1.8, the precision rate 98.1, and sensitivity of 98.2%. Moreover, the LDA and F-KNN also perform better and

achieved recognition accuracy of 98.0%. The recognition results are presented in Table 4, and further confirmed by their confusion matrix as shown in Table 5. Finally, proposed results are compared with existing methods as given in Table 6, which shows the system authenticity.

Table 4. Recognition results on WEIZMANN dataset.

Method	Sen	Pre	FNR	AUC	Acc	Time (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
Q-SVM	97.7	97.7	2.1	0.99	97.9	145
LDA	98.0	98.0	2.0	1.00	98.0	165
M-SVM	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
W-KNN	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. Confusion matrix for Weizmann dataset

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, Hamayed, Fayek, & El Nemr, 2015)	2015	96.66%
(Abdul-Azim & Hemayed, 2015)	2015	97.77%
(D. Vishwakarma, Dhiman, Maheshwari, & Kapoor, 2015)	2015	96.64%
Proposed	**2017**	**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 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 such as 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. All 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 of100%. 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 general comparison is done with the existing methods in Table 8, which shows that the proposed method is performed significantly better than the compared 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
M-VM	**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 & Hemayed, 2015)	2015	97.77%
(D. Vishwakarma et al., 2015)	2015	96.64%
Proposed	**2017**	**98.20%**

4.4. HMDB 51 Dataset

The hmdb51 dataset consists of total 51 human action classes including dive, jump, push up, run, sit, stand up, walk, wave, and few more. It contains total 6849 video sequences of all action classes. This dataset represents the everyday human actions, which are collected from several internet sources. This dataset has many challenges such as the high difference between video sequences, blur, and a problem of lower quality. In this article, we select the 12 action classes for

testing including Ride Bike, Run, Jump, kick, punch, push up, sit up, stand, throw, turn, walk, and wave. The 60:40 strategies are adopted for training and testing of the system. All experiments are done using 10 fold cross validation. The maximum testing recognition accuracy is 92.6%, which is achieved on M-SVM. The M-SVM achieved this accuracy in 47.235 seconds. Moreover, the weighted KNN also perform well and achieved recognition accuracy is 90.9% and sensitivity rate is 89.67%, which is higher after M-SVM. The recognition accuracy of M-SVM is presented in Table 9. Finally, the comparison of proposed recognition results with existing method is presented in Table 10, which shows that the proposed method perform significantly well as compare to 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 & Ramanan, 2017)	2017	52.2
(Duta et al., 2017)	2017	61.0
(Bilen, Fernando, Gavves, Vedaldi, & Gould, 2016)	2016	65.2
(L. Wang et al., 2017)	2017	71.0
Proposed	***2017***	***92.6***

4.5. Discussion

In this section, we interpret our proposed HAR method, which is based on two major steps including ROI detection and recognition. The ROI detection step consists 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 as 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. Also the confusion matrix 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.

5. CONCLUSION

From all above discussion, we conclude that the frame stretching show a major role for improving the visual contrast of given frame, which later play a key role for ROI detection and robust feature extraction. The robust and prominent feature extractions are much important for improving the system classification accuracy because high dimensional and irrelevant features degrade the system accuracy and also effect on the system execution time. The proposed method is evaluated on five publically available datasets including KTH, Weizmann, UCF-YouTube, MSR-Action, and HMDB51 achieved recognition accuracy 99.5%, 98.2%, 100%, 90.2%, and 92.6%, respectively. Moreover, the comparison with existing methods shows the system authentication.

6. REFERENCES

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

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., 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., 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.

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., Faisal, M., Yasmin, M., & Fernandes, S. L. (2018). A Framework for Offline Signature Verification System: Best Features Selection Approach. *Pattern Recognition Letters*.

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.

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.