International Journal of Engineering & Technology, 7 (2.20) (2018) 207-213

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A novel feature fusion based Human Action Recognition in 2D Videos

K. Rajendra Prasad[1]*, P. Srinivasa Rao[2]

[1]Research Scholar, Department of Computer Science and Systems Engineering
[2]Professor, Department of Computer Science and Systems Engineering
College of Engineering, Andhra University
**Corresponding author E-mail: kalapalarajendraprsd11@gmail.com*

Abstract

Human action recognition from 2D videos is a demanding area due to its broad applications. Many methods have been proposed by the researchers for recognizing human actions. The improved accuracy in identifying human actions is desirable. This paper presents an improved method of human action recognition using support vector machine (SVM) classifier. This paper proposes a novel feature descriptor constructed by fusing the various investigated features. The handcrafted features such as scale invariant feature transform (SIFT) features, speed up robust features (SURF), histogram of oriented gradient (HOG) features and local binary pattern (LBP) features are obtained on online 2D action videos. The proposed method is tested on different action datasets having both static and dynamically varying backgrounds. The proposed method achieves shows best recognition rates on both static and dynamically varying backgrounds. The datasets considered for the experimentation are KTH, Weizmann, UCF101, UCF sports actions, MSR action and HMDB51.The performance of the proposed feature fusion model with SVM classifier is compared with the individual features with SVM. The fusion method showed best results. The efficiency of the classifier is also tested by comparing with the other state of the art classifiers such as k-nearest neighbors (KNN), artificial neural network (ANN) and Adaboost classifier. The method achieved an average of 94.41% recognition rate.

Keywords: *Human action recognition (HAR); Feature Fusion, Support vector machine; Scale invariant feature transform; Speed up robust features; Histogram of oriented gradient; Local binary patterns.*

1. Introduction

Human action recognition from 2D videos is a challenging task due to the human appearance and posture variations with in the same category of action. Human action recognition has the demanding application areas, such as human computer interaction, intelligent surveillance video monitoring. The goal of human action recognition is to recognize ongoing actions from unknown video consisting of a sequence of images. Occlusions, camera movements, various cluttered background and change in the illumination makes the recognition task more complex. Extraction of robust features which describes the human action and selection of the suitable classifier is important for reliable human action recognition. Human action representation and recognition methods are broadly classified into two, appearance-based approaches [1] and model-based approaches [2]. These methods are further sub categorized into interest point based approaches [3], tracking based approaches [4], spatio-temporal shape template based approaches [1][5] and flow based approaches [6].

Optical flow features are extracted to describe the motion in the flow based approaches. As the optical flow descriptor is noise sensitive, the false motions may reduce the recognition accuracy. In spatio-temporal shape based approach the huge number of features are extracted from the 3D volume, which increases the computational cost for real time applications. Tracking based approaches are also suffers from same problem. In real time, processing of huge number of features is not preferred and a short feature vectors are encouraged. Interest point based approaches incorporates the advantage of short feature descriptors. Treating the videos as documents and visual features as words [7], Bag of video words (BoVW) [8] is the most widely used action recognition technique that eliminates the problem of location change and noise.

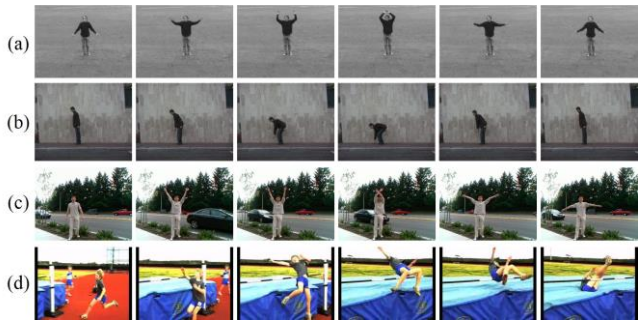

Fig. 1: Human Action Datasets. a) KTH dataset – Hand Waving, b) Weizmann dataset – Bending, c) MSR action dataset – Hand Waving, d) UCF101 dataset – High Jump.

In this paper we proposed to use fusion of various feature descriptors to improve the accuracy of recognizing human actions in a 2D videos. Scale invariant feature transform, speeded up robust features, histogram of oriented gradients features, local binary pattern features are extracted for each action in the action datasets. The sample actions from various online available datasets are shown in figure 1. All these features are fused together to improve

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

the recognition rate. The SVM classifier is used to classify the actions. The other state of the art classifiers such as ANN and Adaboost are also implemented with these fusion of features as input and compared the performance with SVM classifier.
The rest of the paper is organized as: Section 2 gives the related work in this field, and the datasets used in this work are briefly explained in section 4. The proposed methodology is described in section 4. The. In section 5 experimental results are presented and finally the conclusion.

2. Related Work

For any action recognition activity, the foremost important task is the representation of 2D video. There are many video representation models are proposed using the appearance and motion descriptors. This section brings out the brief literature of the previously presented works in human action recognition. The popular descriptors that represents the human action videos are Scale invariant feature transform [9,10], Speed up robust features [15], histogram of oriented gradients [12], histogram of oriented optical flow, 3D histogram of oriented gradients [13], BRIEF [14] and local binary patterns [16]. Many of these descriptors are based on obtaining the interest points using Harris condition [11]. The SIFT provides the local appearance and motion descriptors which plays vital role in action recognition.
In [17], Zhang et al. extracted key points and the tracking information from SIFT flow for human activity recognition. Hassaan ali et al. used SVM classifier to classify actions from extracted SIFT and HOG features in [18]. The MoSIFT features are proposed by Chen and Hauptmann in [19] which detects interest points. The encoding process is adopted to describe the local appearance and local motion in an action video frame.
Thi and Zhang et al. performed the human action recognition by evaluating the silhouette based features [20] on multi views. The multi views are obtained by computing the R transform on the silhouette surface. In [21], the pose of the human in the videos are represented using the contour points of the silhouette. In [16], Nowozin and Bakir extracted the local binary patterns (LBP) and contour pose features to make the human action recognition system as a view invariant. The silhouette based features and an optical flow features are combinedly used in [22]. They adopted principle component analysis to reduce the dimensionality of the feature descriptor to speed up the recognition activity. The method proposed in [23] uses local partitioned histogram of oriented gradients (HOG) features to accurately classify the human actions.
The geometrical and hu-moments are extracted and fed to the multiclass SVM to classify the human action in [24]. Wang et al. in [25] combined different feature detection methods to extract the feature descriptors and the performance is tested on static and dynamic background datasets.
In this work we proposed the fusion of various shape and motion features to describe a human action from a 2D videos. The features are classified using the state of the art SVM classifier to produce the best performance. Datasets with static and dynamic backgrounds are considered for the experimentation. The datasets are KTH, Weizmann, UCF101, UCF Sports actions, HMDB51 and MSR action. The performance of the proposed method is compared with the other state of the art classifiers.

3. Datasets

This section gives brief explanation about different datasets used in this work for evaluating the proposed method. Six online available datasets namely, KTH dataset, Weizmann dataset, UCF sports dataset, UCF101 dataset, MSR action dataset and HMDB51 action dataset were considered for evaluating the current proposed method. The datasets are briefly explained in the following section.

3.1. KTH Dataset

The current database[26] explores six actions - walking, jogging, running, boxing, hand waving and hand clapping performed by 25 subjects repeatedly in four different scenarios outdoors, outdoors with scale variation, outdoors with different clothes and indoors. A total of 2391 sequences were recorded with a static camera with 25fps frame rate.

3.2. Weizmann Dataset

The database covers 10 natural actions - running, walking, skipping, jumping-jack, jumping-forward-on-two-legs, jumping-in-place-on-two-legs, galloping sideways, waving-two-hands, waving one- hand and bending performed by nine subjects [27]. It contains a total of 93 sequences. All sequences are taken with a static camera with 25fps frame rate, down sampled to the spatial resolution of 180x144 pixels. The dataset also has ten additional sequences of walking captured from a different viewpoint varying between 0 and 81 relative to the image plane. The extracted masks after background subtraction and background sequences are provided.

3.3. UCF Sports Dataset

This dataset [28] consists of several actions collected in 2008 from various sporting events which are typically featured on broadcast television channels such as the BBC and ESPN. The video sequences were obtained from a wide range of stock footage web sites including BBC Motion gallery, and GettyImages. The dataset comprises a natural pool of actions featured in a wide range of scenes and viewpoints.

3.4. UCF 101 Dataset

The UCF-101 [29] is composed of realistic web videos, which are typically captured with camera motions and under various illuminations, and contain partial occlusion. It has 101 categories of human actions, ranging from daily life to unusual sports (such as “Yo Yo”). UCF-101 has more than 13K videos with an average length of 180 frames per video. It has 3 split settings to separate the dataset into training and testing videos.

3.5. MSR Action Dataset

The MSR Action Dataset [30], created in 2009 to study the behavior of recognition algorithms in presence of clutter and dynamic backgrounds and other types of action variations. The dataset contains 16 video sequences and includes 3 types of actions: hand clapping (14 examples), hand waving (24 examples), and boxing (25 examples), performed by 10 people. Each sequence contains multiple types of actions. Some sequences contain actions performed by different people. There are both indoor and outdoor scenes. All the video sequences are captured with clutter and moving backgrounds with lengths ranging from 32 to 76 seconds.

3.6. HMDB51 Dataset

The Serre lab at Brown University, USA, introduces the HMDB dataset [31] collected in 2011 from various sources which are mostly from movies and, a small proportion, from public databases, such as the Prelinger archive, YouTube and Google videos. The dataset contains 6849 clips divided into 51 action categories, each containing a minimum of 101 clips.

4. Proposed Method

Our proposed approach of human action recognition based on various feature fusion technique improves the recognition rate. Let us consider an action video sequence $A(x,y,t) \in \Re^{3\times3}$ having N frames at frame rate of 30 frames per second. Giving A as input to a classifier, we aimed to obtain a corresponding label L. We started the human action recognition activity by extracting the key frames from the action sequence. This process eliminates the action less frames from the entire action sequence. The key frames contain action only frames. Several discriminative features are extracted for the key frames and the feature vectors are built to input the classifier. The proposed method of human action recognition in 2D videos is shown in figure 2.

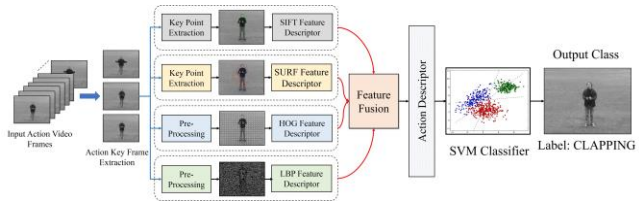

Fig. 2: The block diagram of proposed method of human action recognition in 2D videos.

4.1. Action Key Frame Extraction

The action key frame extraction is achieved based on spatial and temporal saliency attention values. The spatial and temporal saliency maps are generated in the range of [0,1]. The spatial attention value $S_A(t)$ is calculated as

$$S_A(t) = mean\left(\frac{S_S\{A(x,y,t)\}}{\max\left(S_S\{A(x,y,t)\}\right)}\right) \quad (1)$$

Where, $S_A(t) \in [0,1]$ and $S_S\{A(x,y,t)\}$ is the spatial saliency map. Non-action frames get minimum attention value, while the action frames get the attention value ≈ 1.

The temporal change in the consecutive action frames in a sequence is achieved using optical flow [32] tracking algorithm. The temporal saliency maps $S_T\{A(x,y,t)\}$ are derived and the temporal attention value $T_A(t)$ is calculated as

$$T_A(t) = mean\left(S_T\{A(x,y,t)\}\right) \quad (2)$$

The maximum value of $S_T(t)$ denotes the action frames in the sequence and the low value denotes the action non-action frame. Both the spatial and temporal saliency attention values are fused [33] to obtain key frames from the entire action video sequence $A(x,y,t)$. The process of action key frame extraction is shown in figure 3.

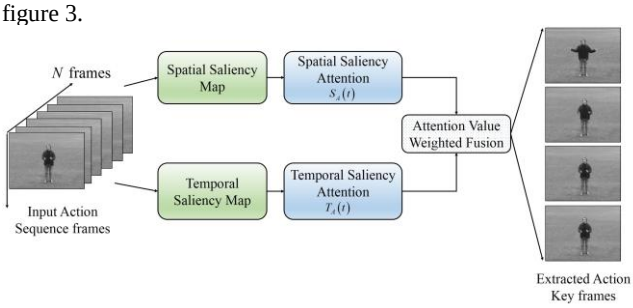

Fig. 3: Process of action key frame extraction.

This paper proposed to extract various handcrafted features on action key frames such as SIFT, SURF, HOG and LBP, fused together to make human action recognition more reliable. The features are extracted on key frames of different human action dataset videos.

4.2. Scale Invariant Feature Transform (SIFT) Features

Lowe et al. in [34] extracted the invariant features based on invariant descriptors. The SIFT features are invariant to translation, scaling, intensity variations, rotation and noises. The SIFT feature descriptor is constructed in two stages. In stage one, the direction parameters of feature points are determined, and a 128-dimensional feature vector is constructed using the graphic information about feature interest points in stage two.

To make sure SIFT features to be rotation invariant, the feature descriptor is created in the main direction of feature interest points. The gradient is calculated on the detected feature interest points and the difference is figured out. The gradient module of a pixel at point (x,y) is defined as

$$\nabla(x,y) = \sqrt{l_1^2 + l_2^2} \quad (3)$$

where $l_1 = l(x+1,y) - l(x-1,y)$ and $l_2 = l(x,y+1) - l(x,y-1)$.

$l(x,y)$ is pyramidal image grayscale of an interest point at (x,y).

The gradient angle θ at a feature interest point (x,y) is given as

$$\theta(x,y) = \arctan\left(\frac{l(x,y+1) - l(x,y-1)}{l(x+1,y) - l(x-1,y)}\right) \quad (4)$$

The local region is rotated around the feature point by the gradient magnitude θ. The rotation invariance feature of the descriptor is attained by considering it in the main direction. The entire rotated region is equally divided into 16×16 rectangular windows with the feature point as center into 4×4 sub-regions. In each sub-region, the gradient histogram is calculated in eight directions. In this process a 128-dimensional feature vector is generated by each feature interest point.

4.3. Speed Up Robust Features (SURF)

Speed Up Robust Features algorithm is rooted from the multi-scale space theory. The SURF features are resilient to rotation, scale and illumination variations. The algorithm is implemented in four stages [35,36]. In stage one the integral image is generated followed by Fast-Hessian detection in stage two. Descriptor orientation assignment is done in stage three and finally the descriptor is generated in stage four.

The integral image is obtained as

$$I_\Sigma(x,y) = \sum_{i=0}^{x}\sum_{j=o}^{y} I(i,j) \quad (5)$$

The surface integral of any size from the image given is computed by reading only four-pixel values.

The key points / interest points are detected using the fast Hessian detector. The Hessian matrix is defined as

$$H(x,y) = \begin{vmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x\,\partial y} \\ \frac{\partial^2 f}{\partial x\,\partial y} & \frac{\partial^2 f}{\partial y^2} \end{vmatrix} \quad (6)$$

4.4. Histograms of Oriented Gradients (HOG) Features

Histograms of Oriented Gradients [37] is a popular 2D descriptor originally developed for person detection. The important components of the detector are shown in figure 4. A HOG descriptor is computed using a block consisting of a grid of cells where each

cell again consists of a grid of pixels. The number of pixels in a cell and number of cells in a block can be varied. The structure performing best according to the original paper is 3 × 3 cells with 6 × 6 pixels.

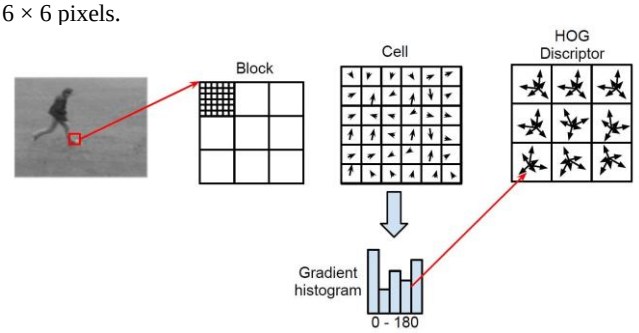

Fig. 4: Block diagram of HOG method.

For each cell in the block, a histogram of the gradients in the pixels is computed. The histogram has 9 bins and a range of either 0-180 or 0-360, where the former is known as unsigned and the latter as signed. Each gradient votes for the bin corresponding to the gradient direction, with a vote size corresponding to the gradient magnitude. Finally, each block is concatenated into a vector v and normalized by its L_2 norm

$$v_{norm} = \frac{v}{\sqrt{\|v\|_2^2 + \varepsilon^2}} \tag{7}$$

where ε is a small constant to prevent division by zero. The HOG descriptor is very similar to the descriptor used in SIFT [38]. The difference is that that the SIFT descriptor is rotated according to the orientation of the interest point.

4.5. Local Binary Patterns (LBP) Features

The LBP operation [39] is a plain sailing method yet constructive gray scale and rotational invariant texture operation, being used in various potential applications. It labels the image pixels with decimal numbers which encodes the local texture information. Given a pixel (scalar value) g_c in an image, its neighbour set contains pixels that are equally spaced on a circle of radius r (r > 0) with the center at g_c. If the coordinates of g_c are (0,0) and m neighbours $\{g_i\}_{i=0}^{m-1}$ are considered, the coordinates of g_i are $\left(-r\sin(2\pi i/m), r\cos(2\pi i/m)\right)$. The gray values of circular neighbours that do not fall in the image grids are estimated by bilinear interpolation [39].

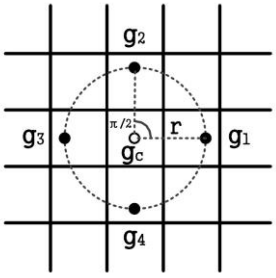

Fig. 5: Center pixel g_c and its 4 circular neighbours $\{g_i\}_{i=0}^{3}$ with radius r for the LBP operator.

Fig. 5 illustrates an example of a neighbour set for $(m=4, r=1)$ (the values for m and r may change in practice). The LBP is created by thresholding the neighbours $\{g_i\}_{i=0}^{m-1}$ with the center pixel g_c to generate a m-bit binary number. The resulting LBP for g_c can be expressed in decimal form as follows:

$$LBP_{m,r}(g_c) = \sum_{i=0}^{m-1} U(g_i - g_c)2^i \tag{8}$$

Where $U(g_i - g_c) = 1$ if $g_i \geq g_c$ and $U(g_i - g_c) = 0$ if $g_i \leq g_c$.

Although the LBP operator in Eq. (4) produces 2^m different binary patterns, a subset of these patterns, named uniform patterns, is thus able to describe image texture [39]. After obtaining the LBP codes for pixels in an image, an occurrence histogram is computed over an image or a region to represent the texture information.

4.6. Support Vector Machine (SVM) Classifier

Support vector machine is a powerful classifier introduced by Vapnik [40] and Cortes [41]. It has been widely used with outstanding results in many pattern and action recognition applications [42]. It is the state of the art algorithm which solves many linear and non-linear classification tasks [40]. The SVM has a very good prediction capability and flexibility. The implementation of SVM considers the minimization of structural risk, rather than empirical risk. The minimization of empirical risk is traditionally used in the artificial neural networks [40].

Basically, SVM determines the hyper-plane which separates the different classes. A decision surface is constructed which maps the sample points turns into a feature space of high dimensionality. The feature space is categorized using a nonlinear transformation Φ.

$$f(x) = W^T \Phi(x) + b \tag{9}$$

Where $W \in R^n$, $b \in R$ and $\Phi(x)$ is a feature map.

The optimal hyper-plane is obtained by solving a quadratic programming problem which is reliant on regularization parameters. This transformation was carried out by kernel functions like linear, radial basis function, sigmoid and polynomial kernel types:

The linear kernel: $K(x,y) = x \times y$ (10)

The polynomial kernel: $K(x,y) = \left[(x \times y) + 1\right]^d$ (11)

The Sigmoid kernel: $K(x,y) = \tanh\left(\beta_0 xy + \beta_1\right)$ (12)

RBF kernel (Radial Basis Function): $K(x,y) = \exp\left(-\gamma \|x - y\|^2\right)$ (13)

With d, β_0, β_1 and γ are parameters that will be determinate empirically.

In this work, we adopted a transformation by mapping the input data (x_i, y_i) into a feature space of high dimensionality with the help of a non-linear operator $\Phi(x)$. Hence, the optimal hyper-plane can be written as:

$$f(x) = \text{sgn}\left(\sum y_i a_i K(x_i, x) + b\right) \tag{14}$$

Where $K(x_i, x) = \exp\left(-\gamma \|x_i - x\|^2\right)$ is the kernel function founded on a radial basis function (RBF), and $\text{sgn}(.)$ is the sign function. This classifier model called RBF kernel SVM.

5. Experimental Results and Discussion

This section explains various human action recognition experiments performed. The robustness of the proposed feature fusion

model along with SVM classification is tested under various experiments. The experimentation considers static background datasets and dynamically varying datasets. The statics background datasets are KTH and Weizmann datasets which are having a simple constant background and makes the recognition process easier. But in real time the background varies dynamically. The dynamic background datasets such as UCF sports, UCF101, HMDB51 and MSR action datasets are considered for this work. The local features such as SIFT, SURF, HOG and LBP are extracted on the sequence of different human action video frames and the respective feature descriptors are constructed. As a first step the individual feature descriptor are inputted to SVM classifier to classify the action, later the fused feature descriptor is applied to the classifier. The recognition rates obtained in both the cases for different action datasets are tabulated in table 1.

Table 1: SVM classifier performance (Recognition rates) with various feature descriptors.

Dataset	Recognition Rates (%)				
	Exp-1	Exp-2	Exp-3	Exp-4	Exp-5
	SIFT	SURF	HOG	LBP	SIFT+SURF+ HOG+LBP
KTH	86.52	87.43	89.41	83.10	95.98
WEIZMANN	85.26	86.99	88.79	82.49	96.02
MSR Action	83.19	85.76	86.32	79.49	93.28
UCF101	81.56	84.68	87.49	78.43	94.74
UCF Sports	82.40	83.29	85.66	79.15	94.68
HMDB51	79.53	82.46	83.19	76.09	91.75

In experiment-1, only SIFT features are used to classify the action data using SVM classifier and found moderate recognition rates. Later in experiment-2, the SURF features alone used to classify the action and ended with unreliable testing results. In experiment-3, we tried implementing the task using the HOG tracking features alone and found some reliable results on clean background data. This method showed misclassification for dynamically varying complex backgrounds. Experiment-4, moves to LBP features and found that LBP features alone are not enough to classify the human action in videos. The proposed method is implemented as experiment-5 to improve the accuracy. This method uses the feature fusion technique and an SVM classifier to classify. In this experiment we found reliable recognition rates. From the table 1, it is observed that the recognition rates are quite more in the case of feature fusion when compared to the classification cases with individual feature descriptors. The recognition rates are good for both static and dynamic background datasets. The average recognition rate achieved by SVM classifier using SIFT features alone is observed as 83.08% and using SURF alone gives an average of 85.10%. In case of HOG and LBP features 86.81% and 79.79% observed. A good amount of average recognition rate i.e. 94.40% is obtained when all these features were combined.

Further the performance of the proposed method of feature fusion to the SVM classifier is tested on individual datasets and the confusion matrices were plotted as shown in figure 6. The figure shows the performance of the proposed on various datasets is reliable. The experimental evaluation on static background datasets KTH and Weizmann datasets are shown in fig. 6(a) and fig. 6(b) respectively. As the background is static and simple the recognition rates obtained on these two datasets found to be higher, when compared to the recognition rate obtained on the dynamic background datasets. The average recognition rate obtained on static background datasets is around 95–97% and for the dynamic background datasets is observed in between 92 to 94 %.

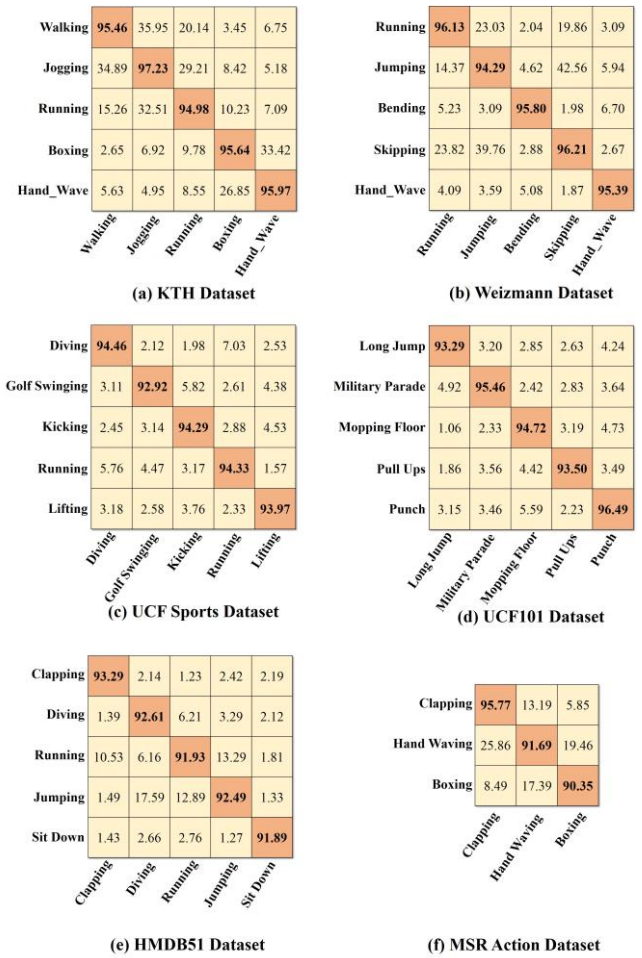

Fig. 6: Confusion Matrix obtained using the proposed method on a) KTH dataset, b) Weizmann dataset, c) UCF sports action dataset d) UCF101 action dataset, e) HMDB51 dataset and f) MSR action dataset.

To further know the robustness of the SVM classifier with multi feature fusion, it is compared with other state of the are classifiers such as KNN, ANN and Adaboost. The same multi feature fusion data is inputted to these classifiers and the recognition rates obtained on different datasets are tabulated in table 2. From the table 2, it can be observed that the multi feature fusion SVM classifier is outperformed on other classifiers. The ANN is also giving somewhat positively reliable recognition rates which are nearer to the SVM classifier recognition rates. However, the ANN requires more training to achieve these recognition rates. Adaboost classier is fast in execution but the average recognition rates achieved is around 88.14%. The classifier KNN gives an average of 85.61% recognition rate. From the table 2, it is clearer that the SVM classifier performs well with multi feature fusion for human action recognition tasks.

Table 2: Performance comparison of SVM classifier with other state of the art classifiers.

Dataset	Recognition Rates using different Classifiers			
	KNN	ANN	SVM	Adaboost
KTH	87.35	92.02	95.98	89.54
WEIZMANN	87.39	92.06	96.02	90.58
MSR Action	84.65	89.32	93.28	86.84
UCF101	86.11	90.78	94.74	88.30
UCF Sports	85.05	90.72	94.68	88.24
HMDB51	83.12	87.79	91.75	85.31
Average Recognition Rate	**85.61**	**90.45**	**94.41**	**88.14**

The below figure 7 shows the performance of SVM classifier with other state of the art classifiers on different actions taken randomly

form various action datasets. The plot shows the SVM classifier can perform well with multi feature fusion concept.

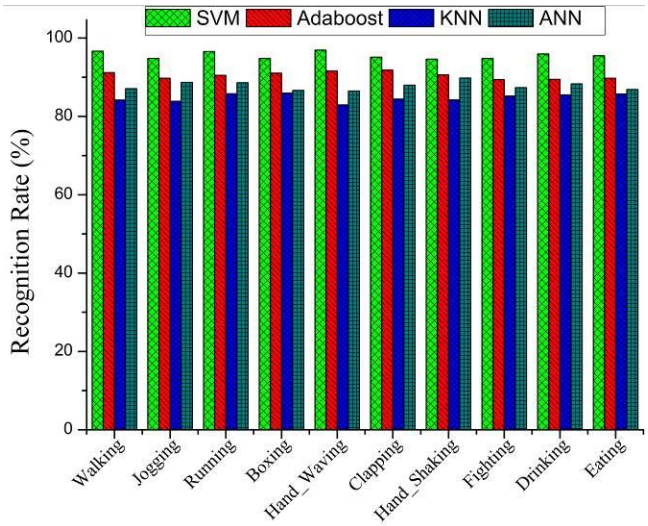

Fig. 7: Performance comparison of SVM with multi feature fusion over state of the art classifiers with multi feature fusion.

Further we compared the results obtained in our method with various previously proposed methods in table 3. It is observed that the proposed method outperforms over the other methods for human action recognition and classification.

Table 3: Comparison of the proposed method with existing methods.

Method	Recognition rate
Liu and Shah [43]	93.2
Niebles et al. [44]	83.3
Schuldt et al. [45]	71.72
Dollar [46]	81.17
Zhang [47]	91.33
Lin [48]	93.43
Bregonzio et al. [49]	93.33
Proposed Method	**94.41**

The proposed method is showing good performance with reliable recognition rates. The method is compared with the previously proposed and found that the proposed method is showing its best in correctly recognizing the human actions in 2D videos.

6. Conclusion

A novel feature fusion technique is proposed in this paper for human action recognition in 2D videos. Various handcrafted features such as SIFT, SURF, HOG and LBP were considered, and a feature descriptor is generated by fusing these four features. The SVM classifier is used in this work to classify the actions. The SVM with individual feature descriptor is also tested to know the capability of our feature fusion model. An average of 94.41% recognition rate is achieved using the proposed method. The method is tested on various datasets, which are with static and dynamically varying backgrounds. The proposed methods work well on all category datasets. The drawbacks in each feature alone is rectified by the other feature fused to it. The robustness of the classifier is tested by comparing it with other state of the art classifiers such as KNN, ANN and Adaboost. Further, the recognition rate can be improved by adding more robust and view invariant features for classification.

References

[1] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In ICCV, 2005.
[2] C. Fanti, L. Zelnik-manor, and P. Perona. Hybrid models for human motion recognition. In ICCV, 2005.
[3] Chen MY, Hauptmann AG. MoSIFT: recognizing human actions in surveillance videos. Technological report, CMU-CS-09-161, Carnegie Mellon University; 2009. pp. 9–161.
[4] Sheikh Y, Sheikh M, Shah M. Exploring the space of a human action. Int Conf Comput Vision, ICCV IEEE 2005:144–9.
[5] Ke Y, Sukthanka R, Hebert M. Efficient visual event detection using volumetric features. Int Conf Comput Vision, ICCV IEEE 2005;1:166–73.
[6] Fathi A, Mori G. Action recognition by learning mid-level motion features. Computer Vision Pattern Recognition, CVPR IEEE 2008:1–8.
[7] Gemert J, Geusebroe J, Veenman C, Smeulders A. Kernel codebooks for scene categorization. Proc Euro Conf Comput Vision, ECCV 2008:696–709.
[8] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach. Int Conf Pattern Recogn, ICPR IEEE 2004;3:32–6.
[9] D. G. Lowe, Object Recognition from Local Scale-Invariant Features, International Conference on Computer Vision, 1999.
[10] D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 2004.
[11] C Harris, M. Stephens. A combined corner and edge detector, In M.M. Mathews, editor, Proc of Alvey vision conference, pages 147-151, University of Manchester, England. September, 1988.
[12] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2005.
[13] N. Buch, J. Orwell, S. A. Velastin. 3D Extended Histogram of Oriented Gradients (3DHOG) for Classification of Road Users in Urban Scenes, British Machine Vision Conference, 2009.
[14] M. Calonder, V. Lepetit, C. Strecha, P. Fua. Brief: Binary robust independent elementary features. European Conference on Computer Vision, 2010.
[15] H. Bay, T. Tuytelaars, L. Van Gool. Surf: Speeded up robust features. European Conference on Computer Vision, May 2006.
[16] Kushwaha, A.K.S.; Srivastava, S.; Srivastava, R. Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns. Multimedia Syst. 2016.
[17] Zhang, Jia-Tao, Ah-Chung Tsoi, and Sio-Long Lo,"Scale invariant feature transform flow trajectory approach with applications to human action recognition," Neural Networks (IJCNN), 2014 International Joint Conference on. IEEE, 2014.
[18] Qazi, Hassaan Ali, Umar Jahangir, Bilal M. Yousuf, and Aqib Noor. "Human action recognition using SIFT and HOG method." In Information and Communication Technologies (ICICT), 2017 International Conference on, pp. 6-10. IEEE, 2017.
[19] Chen MY, Hauptmann AG. MoSIFT: recognizing human actions in surveillance videos. Technological report, CMU-CS-09-161, Carnegie Mellon University; 2009. p. 9–161.
[20] Souvenir, R.; Babbs, J. Learning the viewpoint manifold for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2008 (CVPR 2008), Anchorage, AK, USA, 23–28 June 2008; pp. 1–7.
[21] Chaaraoui, A.A.; Climent-Pérez, P.; Flórez-Revuelta, F. Silhouette-based human action recognition using sequences of key poses. Pattern Recognit. Lett. 2013, 34, 1799–1807.
[22] Ahmad, M.; Lee, S.-W. HMM-based human action recognition using multiview image sequences. In Proceedings of the 18th International Conference on Pattern Recognition 2006 (ICPR 2006), Hong Kong, China, 20–24 August 2006; pp. 263–266.
[23] Weinland, D.; Özuysal, M.; Fua, P. Making Action Recognition Robust to Occlusions and Viewpoint Changes. In Computer Vision–ECCV 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 635–648.
[24] Sargano, Allah Bux, Plamen Angelov, and Zulfiqar Habib. "Human action recognition from multiple views based on view-invariant feature descriptor using support vector machines." Applied Sciences 6, no. 10 (2016): 309.

[25] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In BMVC, 2010.
[26] Schuldt, Christian, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach." In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 3, pp. 32-36. IEEE, 2004.
[27] Gorelick, Lena, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. "Actions as space-time shapes." IEEE transactions on pattern analysis and machine intelligence 29, no. 12 (2007): 2247-2253.
[28] University of Central Florida, UCF sports action dataset, February 2012. <http://vision.eecs.ucf.edu/datasetsActions.html>.
[29] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR, abs/1212.0402, 2012.
[30] J. Yuan, Z. Liu, and Y. Wu. Discriminative video pattern search for efficient action detection. http://users.eecs.northwestern.edu/~jyu410/index_files/actiondetection.html, January 2012.
[31] Serre lab. Hmdb: A large video database for human motion recognition. http://serre-lab.clps.brown.edu/resources/HMDB/index.htm, November 2011.
[32] Tao, Michael, Jiamin Bai, Pushmeet Kohli, and Sylvain Paris. "SimpleFlow: A Non-iterative, Sublinear Optical Flow Algorithm." In Computer Graphics Forum, vol. 31, no. 2pt1, pp. 345-353. Blackwell Publishing Ltd, 2012.
[33] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. IEEE transactions on pattern analysis and machine intelligence, 35(1):185–207, 2013.
[34] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60, no. 2 (2004): 91-110.
[35] Svab, Jan, Tomas Krajnik, Jan Faigl, and Libor Preucil. "FPGA based speeded up robust features." In Technologies for Practical Robot Applications, 2009. TePRA 2009. IEEE International Conference on, pp. 35-41. IEEE, 2009.
[36] Bay, Herbert, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. "Speeded-up robust features (SURF)." Computer vision and image understanding 110, no. 3 (2008): 346-359.
[37] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886–893. IEEE, 2005.
[38] D.G. Lowe. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2):91–110, 2004.
[39] T. Ojala, M. Pietikäinen, and T. Mäenpää, ``Multiresolution gray-scale and rotation invariant texture classication with local binary patterns," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971987, Jul. 2002.
[40] V. Vapnik, “Statistical Learn Theory,” John Wiley, New York, 1998.
[41] C. Cortes, V. Vapnik, “Support vector networks,” Machine Learning, vol. 20, pp. 273-297, 1995.
[42] C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining Knowledge Discovery, vol. 2(2), pp. 121-167, 1998.
[43] Liu J, Shah M. Learning human actions via information maximization. Comput Vision Pattern Recogn, CVPR IEEE 2008:1–8.
[44] Niebles J, Wang H, Fei-Fei L. Unsupervised learning of human action categories using spatial-temporal words. Int J Computer Vision 2008;79(3):299–318.
[45] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach. Int Conf Pattern Recogn, ICPR IEEE 2004;3:32–6.
[46] Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features. IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance 2005:65–72.
[47] Zhang Z, Hu Y, Chan S, Chia LT. Motion context: a new representation for human action recognition. Proceedings of the European conference on computer vision, ECCV Springer 2008; 5305:817–29.
[48] Lin Z, Jiang Z, Davis LS. Recognizing actions by shape motion prototype trees. Int Conf Comput Vision, ICCV IEEE. p. 1–8.
[49] Bregonzio M, Xiang T, Gong S. Fusing appearance and distribution information of interest points for action recognition. Pattern Recognition 2012;45(3):1220–34.