International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

Human Action Recognition Using SURF and HOG Features from Video Sequences

Akila M, Rajeswari R*
Department of Computer Applications, Bharathiar University, Coimbatore, Tamilnadu, India

ABSTRACT

Human action recognition is important in a number of applications such as video indexing, video surveillance and human computer interaction. Hence, human action recognition has been greatly researched during the last decades; however, it is still regarded as a challenging task. In this paper, a human action recognition method is proposed which aims to improve action recognition using a combination of local and global features. For the local feature Speeded-Up Robust Features (SURF) are used and for global feature Histogram of Oriented Gradients (HOG) are used. Bag-of-Features representation of local features is used for representation of the features extraction and the actions are classified using Support Vector Machine. The proposed human action recognition system is tested with various action categories of videos from the KTH dataset. The experimental results show that the proposed method has better results in terms of accuracy.

Keywords: STIP, human action recognition, SURF, Bag-of-Features, HOG

I. INTRODUCTION

Humans can easily understand actions in a complex scene by using visual system. This field is closely related to other field of studies like motion analysis and action recognition. Recognizing human action is a promising area of computer vision due to its importance in large variety of applications. The main task of human action recognition is to preprocess the data, extract suitable features and classify the features to recognize the different actions. In preprocessing step, many researchers have used different approaches for the noise reduction, background subtraction and silhouettes extraction [1]. Feature extraction process is the key step of any human action recognition system. Different methods have been used for representation, and extraction of the features using silhouettes, spatiotemporal interest points, principal component analysis, motion information and independent component analysis. Classification has been done by several linear and probabilistic classifiers in the field of computer vision to classify the human activities. Although it is difficult for the action recognition system to recognize the action automatically because there are various characteristics of the incoming digital data for which the system may not be able to perform invincibly and identify the actions correctly which have been performed by the human in front of camera.

This paper introduces a novel human action recognition based on a combination of local features viz., HOG and SURF. The set of local interest point features in a video are combined using a Bag-of-Features representation that enables the comparison with other videos. The proposed human action recognition method is also evaluated using videos from KTH dataset. The rest of the paper is organized as follows: Section II reviews previous related work, Section III gives the proposed methodology for human action recognition, Section IV provides the experimental results and Section V gives the conclusion.

II. RELATED WORKS

This section presents some of the potential research works in the field of human action recognition based on Spatio Temporal Interest Points (STIP) based features.

CSEIT1726323 | Received : 11 Dec 2017 | Accepted : 31 Dec 2017 | November-December-2017 [(2)6: 1259-1264]

Local space-time features capture characteristic shape and motion information for a local region in video. They provide a relatively independent representation of events with respect to their spatiotemporal shifts and scales as well as background clutter and multiple motions in the scene. These features are usually extracted directly from video and therefore avoid possible dependencies on other tasks such as motion segmentation and human detection.

Extending the famous Harris detector to video, the Space–Time Interest Points (STIP) for action recognition was first introduced in [2]. The extended 3D Harris takes into consideration the pixel variations on space and time. The histogram of Oriented Gradients (HOG) and the Histogram of Optical Flow (HOF) features are then computed in the local neighborhood of the interest points. The promising result has given by the combination of the HOG as a spatial feature representing the local appearances and the HOF as a temporal feature describing the video motions.

Spatio-temporal corners are rare, even when interesting motion occurs, and might be too rare in certain cases, while enough characteristic motion is still present in other regions[3]. Therefore, Gabor detector was introduced, which gives denser results than the Harris3D. The Gabor detector applies a set of spatial Gaussian kernels and temporal Gabor filters. The final spatio-temporal points are detected as local maxima of the defined response function. Interest point detector which uses global information were introduced [4], i.e. the organization of pixels in a whole video sequence, by applying non-negative matrix factorization on the entire video sequence. This detector is based on the extraction of dynamic textures, which are used to synthesize motion and identify important regions in motion. The detector extracts structural information, the location of moving parts in a video, and searches for regions that have a large probability of containing the relevant motion.

The ESURF (Extended SURF) is an extension of the SURF (Speeded Up Robust Features) image descriptor to the spatio-temporal domain [5]. The ESURF divides the local neighborhood surrounding a local feature into a spatio-temporal grid, and it represents each cell of the grid by weighted sums of a vector which are uniformly sampled responses of Haar-wavelets along the three axes. The geometrical distribution of interest points are captured by 3D R transform on the interest points based on their 3D locations [6]. The 3D R transform is invariant to geometrical transformation and robust to noise. By considering the spatio-temporal semantic and structural forest for recognizing the actions a real-time solution was provided. Pyramidal Spatiotemporal Relationship Match (PSRM) technique was also introduced in [7] for capturing structural information.

Based on local features global bag-of-words histogram are combined with a bag-of-words histogram focused latent regions of interest was introduced to model a video [8]. The model parameters are learned by a correlation constrained latent SVM, in which the constraint is to enforce that the latent regions chosen across all videos of a class are coherent.

III. PROPOSED HUMAN ACTION RECOGNITION BASED ON SURF AND HOG FEATURES

This section describes the methodology of the proposed human action recognition method using SURF with HOG descriptors. Firstly, spatiotemporal features are extracted from a given video sequence. In the proposed method, local features viz., SURF and global features viz., HOG features are extracted. Then, the Bag-of-Words method is used to encode the features. Finally, a classifier is applied to determine the action class for the given video. Fig.1 summarizes the proposed action recognition method.

A. Interest Point Detection

The first step is to detect interest points in the video, which are the positions where the features are computed. These points should ideally be located at places in the video where the action is taking place. A feature detector identifies the points in the video where features are going to be extracted. These points are called as Spatio-Temporal Interest Points (STIPs). In this work, Speeded Up Robust Features (SURF) Detector is used.

SURF [9] uses integral images, which results in a notable performance boost. The integral images provide a way to calculate responses for box-type filters in constant time. SURF utilizes an approximation of the Hessian matrix for detection, which is given by:

$$\det H_{approx} = D_{XX}D_{YY} - (wD_{XY})^2 \quad (1)$$

where D_{XX}, D_{YY}, D_{XY} are approximations for Gaussian second order derivatives with the lowest scale. After building the structure, the response for any box filter of any size inside the image can be built in constant time by only four operations inside any rectangular image.

SURF detector also provides scale invariance by utilizing scale space presentation. The octaves are images with increasing Gaussian kernel size. This way, the filter is fast to calculate as the box filter is scaled instead of the image. The scale space can be constructed in parallel. Finally, the local features are selected as local maxima in 3 X 3 X 3 neighborhood in the scale-space.

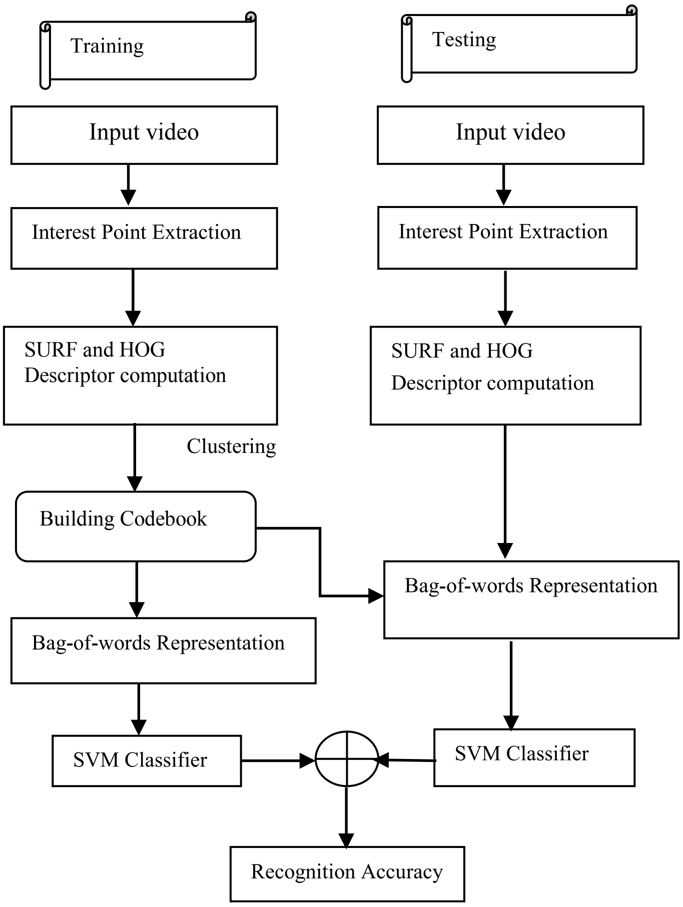

Training
Testing
Input video
Input video
Interest Point Extraction
Interest Point Extraction
SURF and HOG Descriptor computation
SURF and HOG Descriptor computation
Clustering
Building Codebook
Bag-of-words Representation
Bag-of-words Representation
SVM Classifier
SVM Classifier
Recognition Accuracy

Figure1. Flow chart of Proposed Action Recognition Method

B. Feature Description

A descriptor is a feature which is extracted to construe both shape and motion around an interest point so that it plays an important role in action recognition [10]. The descriptors are local, meaning that they are extracted in a predefined area around the detected STIPs. This area is often proportional to the scale where the STIP was detected. The descriptors are often based on gradients or optical flow, because these representations emphasize the parts in the video where changes occur.

1) Speeded Up Robust Feature (SURF) Descriptor: SURF uses a box filter as an approximation of the Gaussian derivative operator. The first step is to select an area around the keypoint, detected in scales, of size 20s. The region is split up into 4 x 4 sub-regions and for each sub-region, Haar wavelet responses are calculated for 5 x 5 blocks from the grid of sample points. The responses around the keypoint are Gaussian-weighted to increase robustness towards geometric deformations and localization errors. Each of the 4 x 4 sub-regions contains 2 x 2 smaller regions where response strengths are summed. A feature vector v calculated from these response strength sums of sub-regions is then:

$$V = \sum d_x, \sum d_y, \sum|d_x|, \sum|d_y| \qquad (2)$$

where dx, dy are the wavelet responses in horizontal and vertical directions. When the 16 vectors are combined, 16 x 4 = 64 dimensional vector is formed. By definition, SURF sums are invariant to illumination changes.

2) Histogram of Oriented Gradients (HOG) Descriptor: HOG descriptor is computed using a block consisting of a grid of cells where each cell again consists of a grid of pixels [11]. The number of pixels in a cell and number of cells in a block can be varied. HOG descriptors are used to describe global gradient information present in the detected patches. The histograms are created by accumulating space-time neighborhoods of detected interest points, where the region is given by a cuboid of the size Δx(σ)=Δy (σ)=18σ and Δt (τ)=8τ. Each cuboid's region is subdivided into a n_X X n_y X n_t grid of cells. For each cell, a 4-bin HOG histogram (4 directions) is calculated. Cell histograms are normalized to form a HOG descriptor.

SURF is a local descriptor and HOG is a global descriptor. Both these features are computed and used in the proposed method. The advantages of both these descriptors help in the improvement of human action recognition.

3) Codebook Construction and Bag-of-Features Representation: The set of local interest point features in a video has to be combined into a representation that

enables the comparison with other videos. In the first step, the bag-of-features model builds a visual vocabulary, called codebook. The codebook is generated using local features extracted from the training videos. Local features extracted from the testing videos are not used in the process of creating the codebook.

Typically, the codebook is generated using the k-means algorithm. After generating the visual vocabulary i.e. the codebook, every video can be represented by the bag-of-features model. The bag-of-features model represents a video sequence by assigning its features to the nearest elements of the created visual vocabulary, i.e. to the nearest cluster centers. Finally, the histogram representation is normalized so that the video size does not significantly change the bag-of-features magnitude.
4) Classification: Support Vector Machines (SVMs) are among the most prominent machine learning algorithms that analyze data and recognize patterns. SVMs belong to the supervised learning algorithms. It means that they use training samples, where each training sample is a pair of an input object (typically a vector) and a desired output value (class label). The classifier used is a Support Vector Machine (SVM) and the implementation used is libsvm. The SVMs analyze the training data and build an inferred function that can be used to correctly determine the class label for an unseen input object.

IV. EXPERIMENTAL RESULTS

The Experimental validation of the proposed method is performed using KTH dataset [12] which contains 6 different actions: walking, jogging, running, boxing, hand waving and hand clapping. The ground truth of this dataset is simple action annotation. 100 videos are utilized in this work, which includes all 6 actions. Out of these, 50 videos are used as training dataset and the remaining 50 videos are used as test dataset. Fig.2 shows the detected interest points for boxing and hand waving videos.

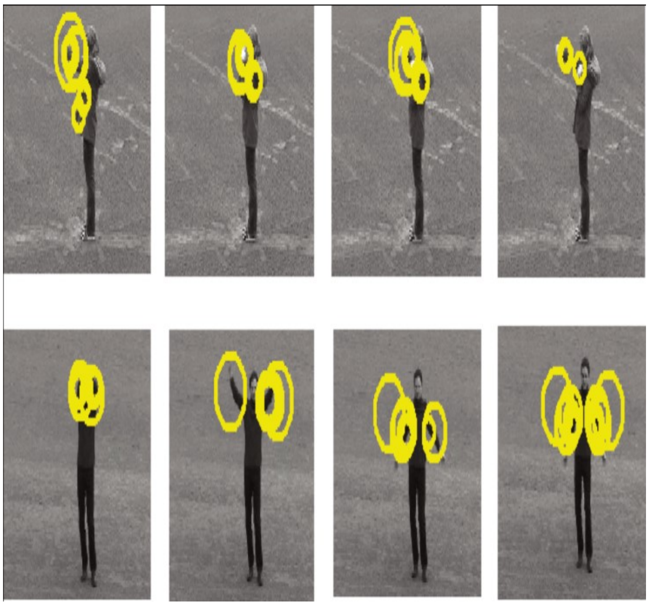


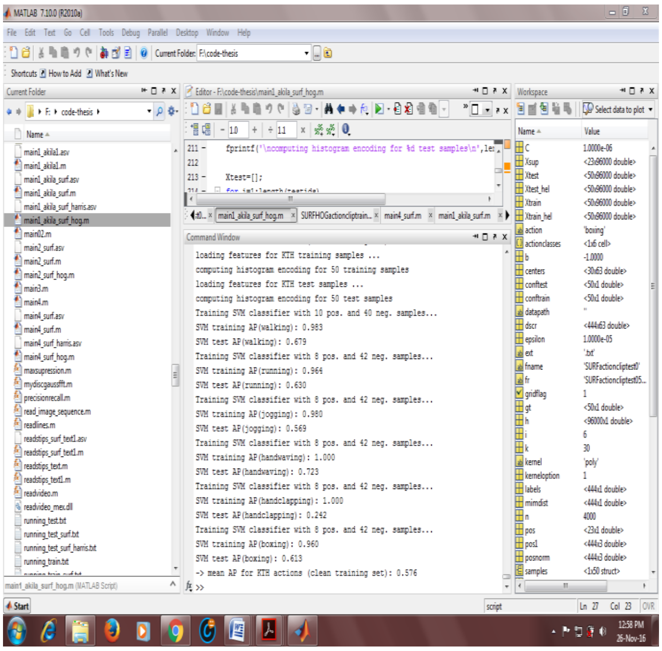
Figure 2. Detected Interest Points on Boxing and Hand waving video

For the datasets, 4000 vocabularies are built to find the best strategy for vocabulary generation. Each individual experiment utilizes the vocabulary that provides the best result for that specific experiment. The measure used for comparison is ‘‘mean average precision”. The Average precision is the average of all true positive percentages across classes. Table.I gives the obtained training and testing average precision values.

TABLE I

MEAN AVERAGE PRECISION VALUES

Features / Actions	SURF Method		Proposed Method	
	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP
Walking	0.992	0.714	0.983	0.679
Jogging	1.000	0.375	0.980	0.569
Running	0.964	0.611	0.964	0.630
Boxing	0.676	0.267	0.960	0.613
Hand waving	1.000	0.494	1.000	0.723
Hand clapping	1.000	0.658	1.000	0.242
Mean AP for KTH Dataset	0.939	0.520	0.981	0.576



```
211 -     fprintf('\ncomputing histogram encoding for %d test samples\n',le
212
213 -     Xtest=[];
```

```
loading features for KTH training samples ...
computing histogram encoding for 50 training samples
loading features for KTH test samples ...
computing histogram encoding for 50 test samples
Training SVM classifier with 10 pos. and 40 neg. samples...
SVM training AP(walking): 0.983
SVM test AP(walking): 0.679
Training SVM classifier with 8 pos. and 42 neg. samples...
SVM training AP(running): 0.964
SVM test AP(running): 0.630
Training SVM classifier with 8 pos. and 42 neg. samples...
SVM training AP(jogging): 0.980
SVM test AP(jogging): 0.569
Training SVM classifier with 8 pos. and 42 neg. samples...
SVM training AP(handwaving): 1.000
SVM test AP(handwaving): 0.723
Training SVM classifier with 8 pos. and 42 neg. samples...
SVM training AP(handclapping): 1.000
SVM test AP(handclapping): 0.242
Training SVM classifier with 8 pos. and 42 neg. samples...
SVM training AP(boxing): 0.960
SVM test AP(boxing): 0.613
-> mean AP for KTH actions (clean training set): 0.576
>>
```

Figure3. Screenshot showing the execution of BoF based action recognition

Fig.3 shows the execution screenshot of proposed BoF based proposed action recognition system. The screenshot gives the result of classified training and testing average precision values for all walking, running, jogging, hand waving, hand clapping and boxing action video sequences. And it finally shows the mean average precision value for clean training set on KTH action dataset.

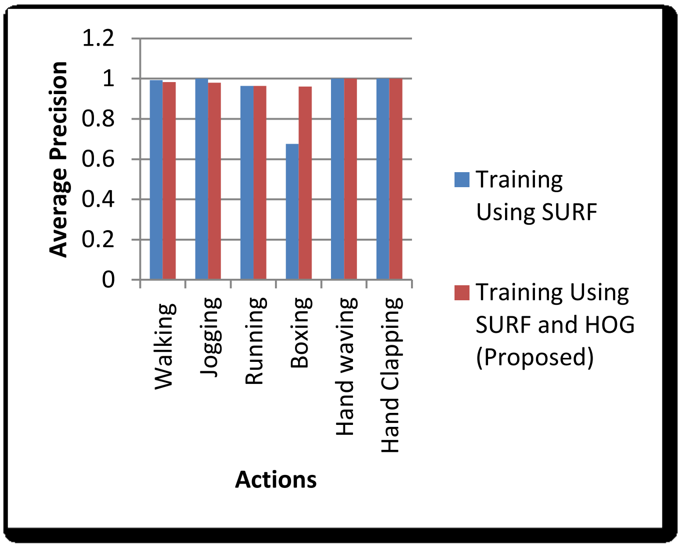


Figure4. Comparison of Average Precision Values for Training

Fig.4 shows a graph plotting the training average precision for 6 actions namely Walking, Jogging, Running, Boxing, Hand waving, Hand clapping based on SURF (existing) and SURF with HOG (proposed) features based action recognition methods. In the graph the result of mean average precision for training using SURF method and the mean average precision for training using SURF and HOG based method can be compared.

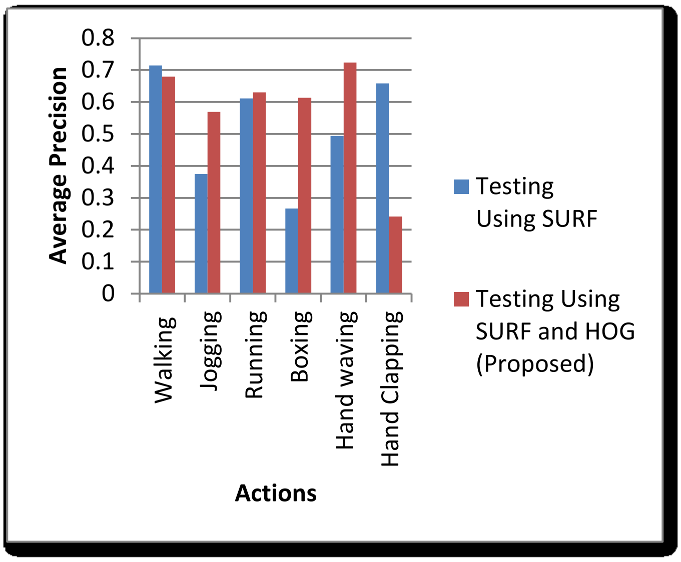


Figure5. Comparison of Average Precision Values for Testing

Similarly, the result of mean average precision for testing using SURF based method and SURF with HOG based method can be compared from fig.5. It can be seen that the proposed SURF with HOG based method gives good results in terms of mean average precision.

V. CONCLUSION AND FUTURE WORKS

This paper addressed the problem of developing a human action recognition system. STIP- based SURF feature detector is applied first and then the key points are encoded with the combined HOG and SURF feature descriptors. In the paper the advantages of local descriptor viz., SURF as well as global descriptor viz., HOG are combined to provide a set of features which will efficiently represent the detected interest points. The results indicate that the developed human action recognition system gives good performance in terms of the mean average precision. In future more number of features can be used to improve the performance of the proposed method. Hybrid or improved methods can also be incorporated in steps such as BoF representation or classification so that the proposed method can be improved.

VI. REFERENCES

[1] Wang S, Yang Y, Ma Z, Li X, Pang C, and Hauptmann A.G, "Action recognition by

exploring data distribution and feature correlation," IEEE Conference on Computer Vision and Pattern Recognition, pp.1370–1377, 2012.

[2] Ivan Laptev and Tony Lindeberg, "On Space-time Interest Points", ACM Digital Library, pp. 432–439, 2003.

[3] Dollar P, Rabaud V, Cottrell G and Belongie S, "Behavior Recognition via Sparse Spatiotemporal Features," IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp.65–72, 2005.

[4] Kwan-Yee Kenneth Wong and Roberto Cipolla, "Extracting spatiotemporal interest points using global information", 11th IEEE International Conference on Computer Vision (ICCV), pp. 1–8, 2007.

[5] Willems G,Tuytelaars T and Van Gool L, "An efficient dense and scale-invariant spatiotemporal interest point detector," European Conference on Computer Vision (ECCV), vol.5303, pp.650–663, 2008.

[6] Yuan C, Li X, Hu W, Ling H and Maybank S, "3D R transform on spatiotemporal interest points for action recognition," IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[7] Yu T.H, Kim T.K and Cipolla R, "Real-time action recognition by spatiotemporal semantic and structural forests," British Machine Vision Conference, pp. 52.1–52.12, 2010.

[8] Shapovalova N, Vahdat A, Cannons K, Lan T and Mori G, "Similarity constrained latent support vector machine: an application to weakly supervised action classification," European Conference on Computer Vision, Springer, pp.55–68, 2012.

[9] Bay H, Ess A, Tuytelaars T, and Gool L.V, "Surf: Speeded up robust features," Computer Vision and Image Understanding, pp.346-359, 2008.

[10] Yamato J, Ohya J and Ishii K, "Recognizing human action in time sequential images using hidden Markov model", Computer Vision and Pattern Recognition, pp. 379–385, 1992.

[11] Dalal N and Triggs B, "Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition, IEEE Computer Society, vol.1, pp.886–893, 2005.

[12] Christian Schuldt, Ivan Laptev and Barbara Caputo, "Recognizing Human Actions: A Local SVM Approach," 17th International Conference on Pattern Recognition, IEEE Computer Society, pp.32–36, 2004.