Performance Evaluation of Descriptors Extracted by MSER Detector for Human Action Recognition

R. Rajeswari
Department of Computer Applications
Bharathiar University
Coimbatore, Tamilnadu
rajeswari@buc.edu.in

P. Ramya
Department of Computer Applications
Bharathiar University
Coimbatore, Tamilnadu
sramya1819@gmail.com

Abstract **– Human action recognition helps in automatically analyzing various events in video data. It has been used for recognizing human actions in many applications including surveillance, healthcare and human-computer interface. In order to recognize human actions in videos various feature descriptors and detectors have been proposed in the literature. The feature detectors help in extracting feature descriptors which provide vital information related to the human actions in video frames. One such feature detector is maximally stable extremal regions (MSER) which is widely used for detecting blobs in video frames. In this paper, the performance of various feature descriptors such as Binary Robust Invariant Scalable Keypoints (BRISK), Histogram of Gradients (HOG) and Speeded Up Robust Features (SURF) extracted by MSER for human action recognition is investigated. Experiments are performed on KTH action dataset.**

Keywords – action recognition, descriptors, detectors, MSER

1. INTRODUCTION

Human action recognition is an active research topic in the field of computer vision. One of the main purposes of computer vision is to make machines analyze and recognize human actions using motion information. Due to the rapidly increasing amount of video records and the large number of potential applications such as visual surveillance (Hu, 2004), human-machine interfaces (Pretlove, 2010), sports video analysis, and video retrieval (Wang, 2012). Among these applications, one of the most interesting is video analysis especially high-level human action recognition plays an important role.

The main task of human action recognition is to pre-process the data, extract suitable features and classify the features to recognize the different actions. In the pre-processing step, many researchers have used different approaches for the noise reduction, background subtraction, and silhouettes extraction (Kim, 2007). Feature extraction process is the key block of any human action recognition system. Different methods have been used for representation, silhouettes, spatiotemporal interest points and extraction of the features using R transform, principal component analysis, motion information and independent component analysis (jalal Ahmad, 2015). Although a number of approaches have been proposed, it is still challenging to recognize a specific object from a dataset of images due to viewpoint change, illumination, partial occlusions, and intra-class difference. The existing methods still need improvement, especially for realistic movies which have wide variations in people's posture and clothes, dynamic background, and partial occlusions.

This paper evaluates the performance of BRISK, HOG and SUEF features for human action recognition. The main objectives of this work are 1) to recognize human actions in videos using feature descriptors and detectors that have been proposed in the literature, particularly for maximally stable extremal regions (MSER) detector 2) Implement human action recognition using Bag-of-Words Representation which use various feature descriptors such as Binary Robust Invariant Scalable Keypoints (BRISK), Histogram of Gradients (HOG) and Speeded Up Robust Features (SURF) extracted by MSER detector 4) Evaluate the performance of MSER detector with the above mentioned descriptors on KTH action dataset.

The rest of the paper is organized as follows. Section II gives a review of the literature for human action recognition methods. Section III presents the evaluation of the proposed human action recognition system based on a combination of MSER/SURF, MSER/SURF, and MSER/BRISK features. Section IV presents the experimental results of the evaluated human action recognition methods. Section V gives the conclusion.

2. REVIEW OF LITERATURE

Action recognition based on local features is one of the most active research topics. Local image video features have been successfully used in many action recognition applications such as object recognition, scene recognition and activity recognition. Feature detectors usually select characteristic spatio-temporal locations and scales in videos by maximizing specific saliency functions. These features are usually extracted directly from video and therefore avoid possible dependencies on other tasks such as motion segmentation and human detection. Some of the action recognition methods are includes:

Wong et al., (Cipolla, 2007)have proposed an interest point detector which uses global information, i.e. the organisation of pixels in a whole video sequence, by applying non-negative matrix factorization on the entire video sequence. The proposed 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.

Willems et al., (Willems G, 2008) proposed a spatio-temporal image descriptor SURF (Speeded Up Robust Features) and the ESURF (Extended SURF). The ESURF divides the local neighborhood surrounding a local feature into a spatio-temporal grid, and it represents each cell of the grid by a vector of weighted sums of uniformly sampled responses of Haar-wavelets along the three axes.

Laptev et al., (Lindeberg, Local Descriptiors for spatio-temporal recognition, 2006) proposed a Histogram of Oriented Gradients (HOG) as cuboid descriptor to represent local object appearance and shape and can be characterized by the distribution of local intensity gradients. HOG descriptor is implemented by dividing the cuboid into small space-time regions and accumulating a local one dimensional histogram of gradient directions over the pixels of each sub-region. The combined histogram entries form the representation. Another approach used by Laptev is Histogram of Optic Flow (HOF). The idea is the same as the previous descriptor HOG, with the only difference that the histogram of optic flow is computed for each sub-region. Laptev proved the combination of HOG and HOF, named HOG-HOF, to perform better than each separate method.

Leutenegger et al., (S.Leutenegger, 2011) proposed a semi binary-based feature detector-descriptor based on the BRISK detector, which can detect and represent videos with significantly reduced computational requirements, while achieving comparable performance to the state-of-the-art spatio-temporal feature descriptors. This proposed feature detector/descriptor can be used not only in action recognition but also in different video-based applications such as motion analysis, anomalous event analysis and video retrieval.

Lu et al., (J, 2006) have proposed novel spectral methods to learn latent semantics from abundant mid-level features by spectral embedding with nonparametric graphs and hyper graphs. A new semantics-aware representation for example, histogram of high-level features, is derived for each video from the original Bag-of-Words (BOW) representation, and actions are classified by a SVM with a histogram intersection kernel based on the new representation. These spectral methods for semantic learning can discover the manifold structure hidden among mid-level features, which results in compact but discriminative high-level features.

3. EVALUATION OF MSER DETECTOR FOR VARIOUS DESCRIPTORS

In this framework, first, the MSER interest point feature detector is applied on a frame-by-frame basis to detect interest points. Amongst the detected points, only the points with significant motion are retained. Then the retained key points are encoded with the combined SURF, HOG and BRISK feature descriptor in the spatial domain. The SURF descriptor is a local descriptor, whereas HOG descriptor is a global descriptor and BRISK is a binary descriptor. Hence, in the present work a local descriptor, a global descriptor and a binary descriptor for MSER detector are evaluated (J.Matas, 2004).

The set of interest point features in a video are combined using a Bag-of-Features representation that enables the comparison with other videos. This method is described in figure 1. 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 centres. The experimental results were conducted on KTH dataset using a linear Support Vector Machine (SVM) for classification. Three human action recognition methods based on the combinations of MSER/SURF, MSER/HOG, and MSER/BRISK feature descriptors are evaluated. The experimental results prove that the MSER/HOG method performs well in terms of the Mean average precision measures.

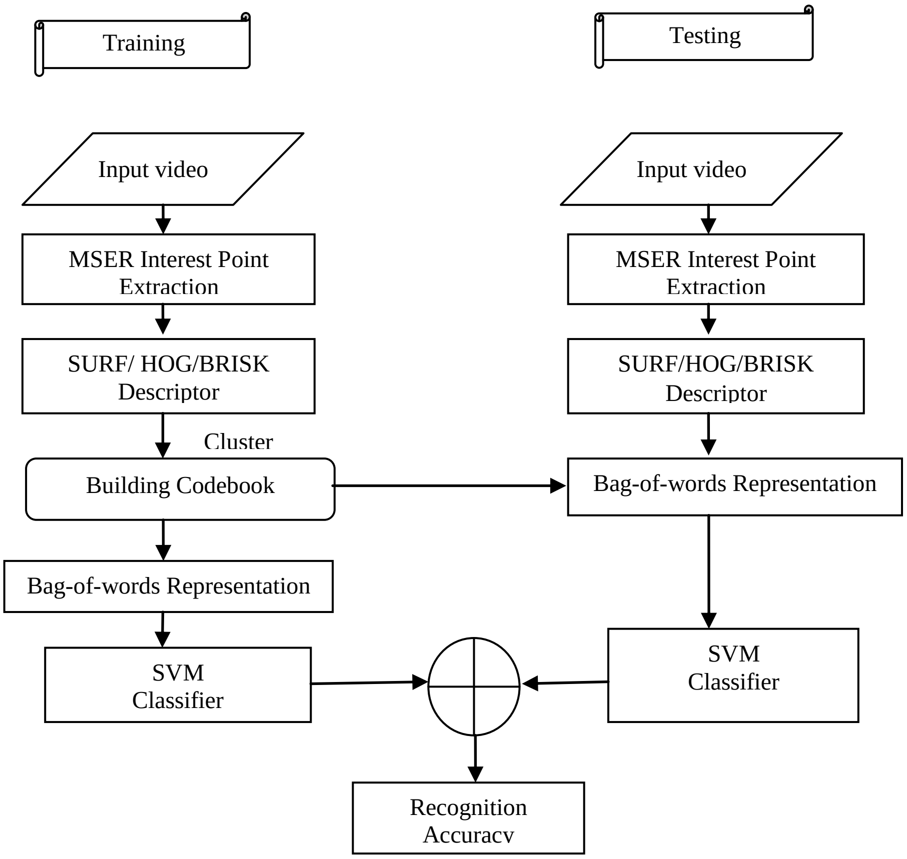

Training
Testing
Input video
Input video
MSER Interest Point Extraction
MSER Interest Point Extraction
SURF/ HOG/BRISK Descriptor
SURF/HOG/BRISK Descriptor
Cluster
Building Codebook
Bag-of-words Representation
Bag-of-words Representation
SVM Classifier
SVM Classifier
Recognition Accuracy

Figure 1. Flow chart of Human Action Recognition algorithm used for MSER evaluation.

3.1. Maximally Stable Extremal Regions (MSER)

MSER is an interest region detector as well as shape descriptor because its regions can be much larger than other interest point methods such as Harris or Features from Accelerated Segment Test (FAST) (Kimmel, 2011). The MSER detector was developed for solving disparity correspondence in a wide baseline stereo system. Stereo systems create a warped, complex geometric depth field, depending on the baseline between cameras and the distance of the subject to the camera. In a wide baseline stereo system, features nearer the camera are more distorted under affine transforms, making it harder to find exact matches between the left/right image pair. The MSER approach attempts to overcome this problem by matching on blob-like features. Its regions are similar to morphological blobs and are fairly robust to skewing and lighting. This method involves sorting pixels into a set of regions based on binary intensity thresholding. Regions with similar pixel value over a range of threshold values in a connected component pattern are considered maximally stable.

The computation of MSER involves the following steps. Pixels are sorted in a binary intensity thresholding loop, which sweeps the intensity value from min to max. The binary threshold is set to a low value such as zero on a single image channel— luminance, for example. Pixels < the threshold value are black, pixels >=are white. At each threshold level, a list of connected components or pixels is kept. The intensity threshold value is incremented from 0 to the max pixel value. Regions that do not grow or shrink or change as the intensity varies are considered maximally stable, and the MSER descriptor records the position of the maximal regions and the corresponding thresholds. In stereo applications, smaller MSER regions are preferred and correlation is used for the final correspondence, and similarity is measured inside a set of circular MSER regions at chosen rotation intervals.

Some interesting advantages of the MSER include: 1) Multi-scale features and multi-scale detection. Since the MSER features do not require any image smoothing or scale space, both coarse features and fine-edge features can be detected. 2) Variable-size features computed globally across an entire region, not limited to patch size or search window size. 3) Affine transform invariance, which is a specific goal. 4) General invariance to shape change, and stability of detection, since the extremal regions tend to be detected across a wide range of image transformations. 5) The MSER can also be considered as the basis for a shape descriptor, and as an alternative to morphological methods of segmentation.

3.2. Scale Invariant Local Feature (SURF)

The SURF descriptor is a rotation and scale invariant local feature descriptor proposed by (Bay H, 2008). The rotation invariance is achieved by finding reproducible orientation for the local neighborhood of a keypoint. When the scale of a detected keypoint is s, the responses of Haar wavelet in both x and y directions are calculated in the circular neighbordhood of size 6s. After calculating the filter responses, the local neighborhood is weighted with a Gaussian with σ= 2 to make the nearest most significant intensity changes. In practice, the calculated wavelet responses are handled as points in 2-D space, X and Y axes represent responses in horizontal and vertical directions, respectively. A sliding “orientation window” (a sector) of size $\frac{\pi}{3}$ is used around the keypoint surroundings to calculate the sum of horizontal and vertical responses. Sums of responses are then used to calculate a local orientation vector for each direction. The longest such vector is finally selected to represent the dominant orientation of a descriptor. The SURF Descriptor is demonstrated in figure 2.

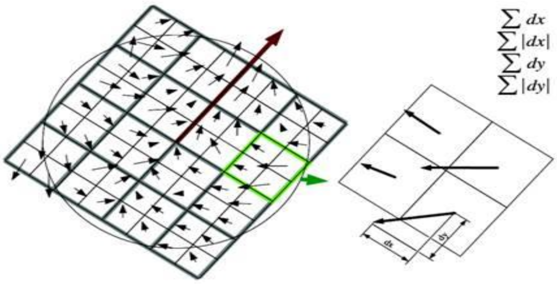


Figure 2. SURF Descriptor

3.3 Histograms of Oriented Gradients (HOG)

HOG (B, 2005) is a popular 2D descriptor originally developed for person detection. The important components of the detector are shown in figure 2. 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.

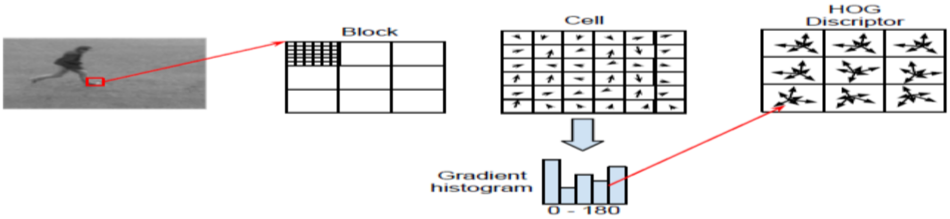


Figure 3. Block diagram of HOG Method

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 unsigned and the latter is 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 L2 norm

$$v_{norm} = \frac{v}{\sqrt{\|v\|_2^2+\epsilon^2}} \tag{1}$$

where ϵ is a small constant to prevent division by zero. The HOG descriptor is very similar to the descriptor used in SIFT (D.G, 1999). The difference is that the SIFT descriptor is rotated according to the orientation of the interest point.

3.4 Binary Robust Invariant Scalable Key points (BRISK)

Brisk (Heinly, 2012) is a unique method, using a novel FAST detector adapted to use scale space, reportedly achieving an order of magnitude performance increase over SURF with comparable accuracy. It is a local binary method using a circular-symmetric pattern region shape and a total of 60 point-pairs as line segments arranged in four concentric rings. This method uses point-pairs of both short segments and long segments, and provides a measure of scale invariance. Subsequently short segments map better for fine resolution and long segments map better at coarse resolution.

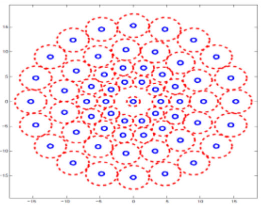


Figure 3. Computation of BRISK sampling pattern

A BRISK sampling pattern is shown in figure 3. The main computational steps in the BRISK computation algorithm (S.Leutenegger, 2011) are 1) Detect key points using FAST or AGHAST based selection in scale space. 2) Perform Gaussian smoothing at each pixel sample point to get the point value. 3) Make three sets of pairs: long pairs, short pairs, and unused pairs (the unused pairs are not in the long pair or the short pair set; 4) Compute gradient between long pairs, sums gradients to determine orientation. 5) Use gradient orientation to adjust and rotate short pairs. 6) Create binary descriptor from short pair point-wise comparisons. Examples for BRISK descriptors for three sets of pairs are shown in figure 4.

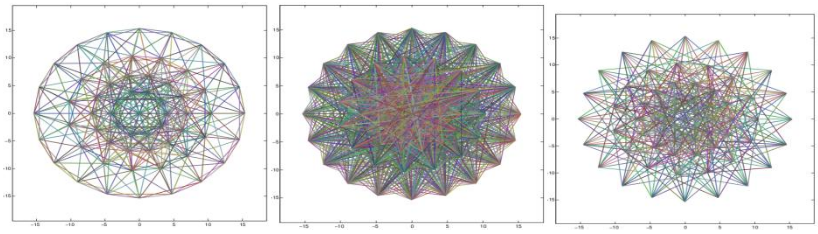


Figure 4. BRISK descriptor: Short-distance pairs (512), Long-distance pairs (870), unused pairs (388)

4. EXPERIMENTAL RESULTS

This section presents the results obtained for human action recognition using MSER detector in combination with descriptors such as SURF, HOG and BRISK. The general framework for human action recognition described in section III is used for evaluating the MSER detector. K-means clustering algorithm is used to cluster all the extracted descriptors. Bag-of-Feature representation is used as the location features encoding technique. The videos from KTH dataset are used for the experiments. The KTH Action dataset shown in figure 5, has been introduced by (Schuldt, Ivan, & Barbara, 2004). The dataset is available in http://www.nada.kth.se/cvap/actions/ website. It contains videos of 6 types of human actions. The six types of human actions included in this dataset are 1) boxing 2) jogging 3) hand clapping 4) hand waving 5) running and 6) walking. Each action is performed several times by 25 different subjects. Each subject performs actions in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors.

All videos are recorded over homogeneous backgrounds and are down-sampled by the authors to the spatial resolution of 160 × 120 pixels. The sequences are recorded using a static camera with 25 frames per second frame rate, and have a length of four seconds on average. In total, the dataset contains 600 video files. The ground truth of this dataset is simple action annotation.

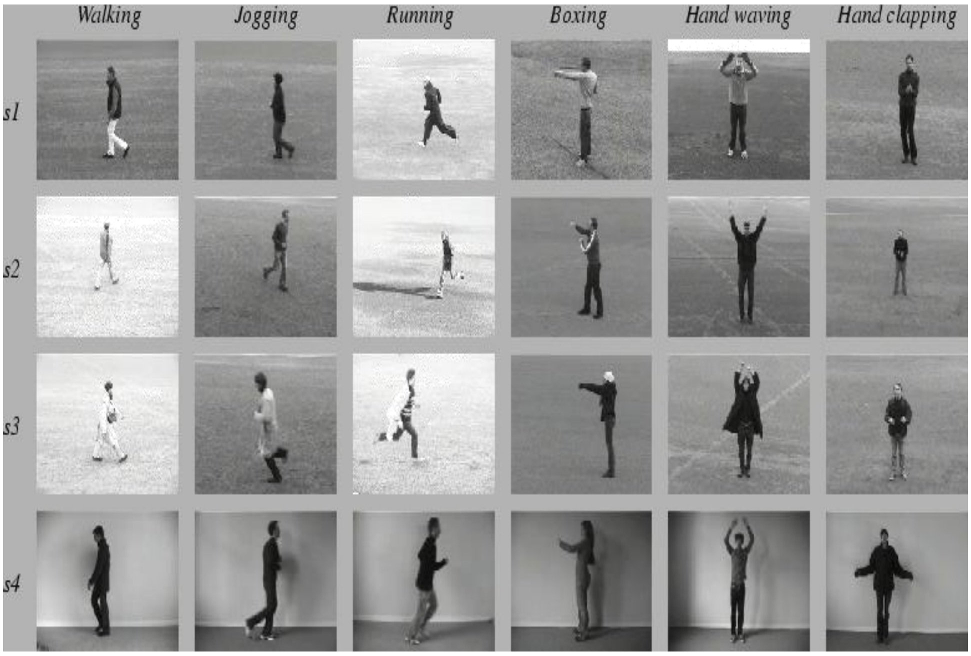
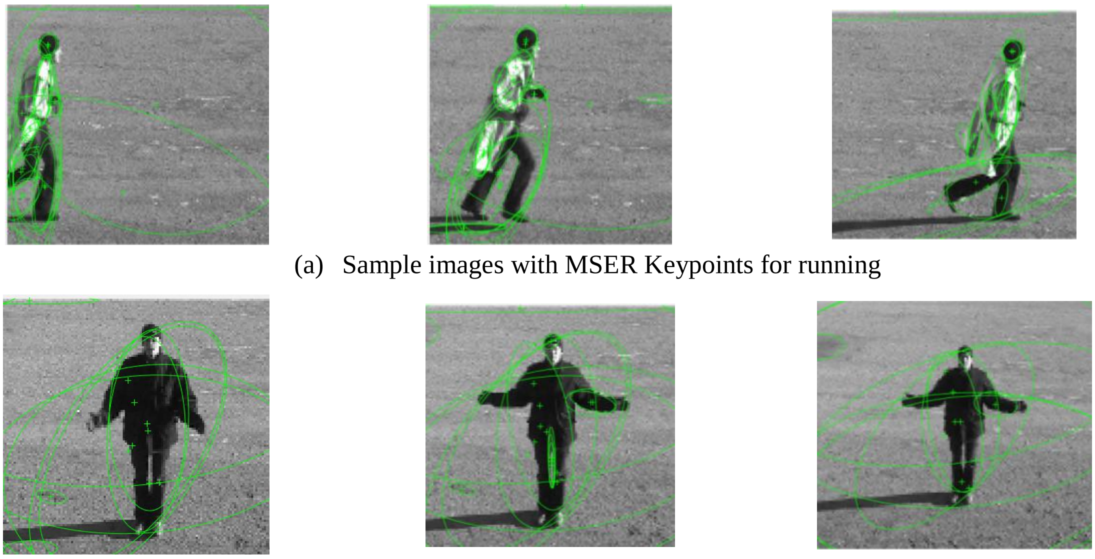


Figure 5. Example images from KTH dataset



(a) Sample images with MSER Keypoints for running

(b) Sample images with MSER Keypoints for hand waving

Figure 6. sample images representing MSER keypoints

In this work, 100 videos are utilized which includes all 6 actions. Out of these videos, 50 videos are used as training dataset and the remaining 50 videos are used as test dataset. Figure 6 (a) shows the detected interest points for running video and figure 6 (b) shows the detected interest points for hand waving video. The MSER detectors help in identifying blobs in images which are connected gray level regions in the images. The centroids of these regions are identified and used for extracting the features. In order to extract features in these centroid points, feature descriptors such as SURF, HOG and BRISK are used. The number of features extracted using the feature descriptors, SURF, HOG and BRISK is 64, 36 and 64 respectively.

The performance of the descriptors viz., SURF, HOG and BRISK evaluated based on how they help in recognizing the actions from the video sets. For this purpose, the cluster size of the code books is varied and the average precision values are obtained for various actions. The cluster sizes taken into consideration are 5, 70 and 500. The mean average precision helps in evaluating the performance of multi-class classification. The formula for precision is given by the equation 2.

$$precision = \frac{True\ Positive}{Total\ Predicted\ Positive} \tag{2}$$

The precision value helps in determining how many of the predicted positive values from the proposed method are actual positive. The mean average precision is given by the equation 3.

$$Mean\ Average\ Precision\ (MAP) = \frac{\sum_{q=1}^{Q} Average(p)}{Q} \tag{3}$$

where p is the precision for every action, Q is the total number of actions ie. 6 representing the actions boxing, jogging, hand clapping, hand waving, running and walking.

Table 1. Average Precision Values when cluster size is 5

Features Actions	MSER &SURF		MSER &HOG		MSER &BRISK	
	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP
Walking	0.227	0.200	0.853	**0.736**	0.703	0.344
Jogging	0.171	0.262	0.862	0.482	0.794	**0.591**
Running	0.222	0.335	0.841	**0.565**	0.836	0.416
Boxing	0.404	0.174	0.701	**0.255**	0.648	0.199
hand waving	0.423	0.167	0.659	0.320	0.723	**0.322**
hand clapping	0.174	0.185	0.745	**0.420**	0.554	0.214
Mean AP for KTH Dataset	0.270	0.220	0.776	**0.463**	0.709	0.347

Table 2. Average Precision Values when cluster size is 20

Features Actions	SURF		HOG		BRISK	
	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP
Walking	0.260	0.214	0.933	**0.639**	0.910	0.275
Running	0.270	0.207	0.950	**0.557**	0.970	0.540
Jogging	0.160	0.207	1.000	**0.679**	1.000	0.399
hand waving	0.372	0.190	0.746	**0.226**	0.960	0.208
hand clapping	0.430	0.176	0.870	**0.352**	0.970	0.286
Boxing	0.163	**0.415**	0.941	0.236	0.926	0.275
Mean AP for KTH Dataset	0.275	0.234	0.906	**0.448**	0.957	0.330

Table 1 gives the obtained training and testing average precision values when the cluster size of the codebook used in Bag-of-Words representation is 5 (k=5). It can be seen from the table that for most of the actions the average precision obtained using MSER & HOG in the test set is good compared to MSER & SURF and MSER & BRISK. Even the mean average precision for MSER & HOG is good.

Table 2 gives the obtained training and testing average precision values when the cluster size of the codebook used in Bag-of-Words representation is 20 (k=20). It can be seen from the table that for most of the actions, except boxing, the average precision obtained using MSER & HOG in the test set is good compared to MSER & SURF and MSER & BRISK. Even the mean average precision for MSER & HOG is good.

Table 3 gives the obtained training and testing average precision values when the cluster size of the codebook used in Bag-of-Words representation is 70 (k=70). It can be seen from the table that for three of the actions the average precision obtained using MSER & HOG is good and for other three actions average precision obtained from MSER & BRISK is good. The mean average precision for MSER & HOG is good.

Table: 3 Average Precision Values when cluster size is 70

Features Actions	MSER &SURF		MSER &HOG		MSER &BRISK	
	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP
Walking	0.255	0.406	1.000	**0.638**	1.000	0.320
Jogging	0.200	0.165	1.000	0.530	0.970	**0.646**
Running	0.160	0.160	1.000	**0.542**	1.000	0.515
Boxing	0.409	0.274	0.970	**0.360**	1.000	0.224
hand waving	0.346	0.361	0.964	0.308	1.000	**0.364**
hand clapping	0.190	0.400	1.000	0.262	1.000	**0.415**
Mean AP for KTH Dataset	0.26	0.294	0.989	**0.440**	0.995	0.414

Table 4 gives the obtained training and testing average precision values when the cluster size of the codebook used in Bag-of-Words representation is 100 (k=100). It can be seen from the table that for most of the actions except hand clapping the average precision obtained using MSER & HOG is good. The mean average precision for MSER & HOG is good.

Table: 4 Average Precision Values when cluster size is 100

Features Actions	MSER &SURF		MSER &HOG		MSER &BRISK	
	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP
Walking	0.293	0.449	0.992	**0.686**	1.000	0.443
Jogging	0.209	0.197	1.000	**0.782**	1.000	0.578
Running	0.160	0.160	1.000	**0.628**	1.000	0.385
Boxing	0.388	0.239	1.000	**0.364**	1.000	0.212
hand waving	0.323	0.359	1.000	**0.398**	1.000	0.235
hand clapping	0.233	**0.394**	1.000	0.213	1.000	0.318
Mean AP for KTH Dataset	0.267	0.299	0.998	**0.511**	1.000	0.361

Table 5. Average Precision Values when cluster size is 500

Features Actions	SURF		HOG		BRISK	
	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP	Training Mean AP	Testing Mean AP
Walking	0.248	0.236	1.000	**0.746**	1.000	0.317
Running	0.170	0.167	1.000	**0.683**	1.000	0.571
Jogging	0.160	0.160	1.000	**0.640**	1.000	0.366
Hand waving	0.362	0.496	1.000	**0.626**	1.000	0.429
Hand clapping	0.295	0.272	1.000	**0.652**	1.000	0.614
Boxing	0.315	0.293	1.000	0.407	1.000	**0.435**
Mean AP for KTH Dataset	0.258	0.270	1.000	**0.625**	1.000	0.455

Table 5 gives the obtained training and testing average precision values when the cluster size of the codebook used in Bag-of-Words representation is 500 (k=500). It can be seen from the table that for most of the actions except boxing the average precision obtained using MSER & HOG is good. The mean average precision for MSER & HOG is good. The results show that for identifying human actions using MSER, the descriptor HOG is more accurate than the other descriptors such as SURF and BRISK.

5. CONCLUSION

This paper presented a brief introduction about human action recognition, the MSER detector and SURF, HOG and BRISK descriptors is given. It also gives the performance evaluation of several local descriptors for the region detector MSER. Three feature descriptors viz., SURF, HOG and BRISK are tested with different cluster sizes 5, 70 and 500. The experimental results proved that MSER and HOG combination outperformed other descriptors for various cluster sizes.

REFERENCES

Hu, W. T. (2004). A survey on visual surveillance of object motion and behaviors . IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) , 334-352.

Pretlove, J. C. (2010). Method to generate a human machine interface. U.S.Patent 7,787,992.

Wang, J. Z. (2012). Mining actionlet ensemble for action recognition with depth cameras. In Computer Vision and Pattern Recognition (CVPR)2012 IEEE Conference on, 1290-1297.

Kim, H. R. (2007). Robust Silhouette Extraction Technique Using Background Subtraction. In 10th Meeting on Image Recognition and Unders

jalal Ahmad, S. K. (2015). A Spatiotemporal motion variation features extraction approach for human tracking and pose-based action recognition . Informatics, Electronics and Vision (ICIEV), International Conference, 1-6.

Cipolla, K.-Y. K. (2007). Extracting spatiotemporal interest points using global information. 11th IEEE International Conference on Computer Vision(ICCV), 1-8.

Willems G, T. T. (2008). An efficient dense and scale-invariant spatiotemporal interest point detector. European Conference on Computer Vision(ECCV), 650-663.

Lindeberg, I. L. (91-103). Local descriptors for spatio-temporal recognition. springer Lecture Notes in Computer Science, 2006.

S.Leutenegger, M. a. (2011). Brisk:Binary robust invariant scalable keypoints. Computer Vision(ICCV) , 2011 IEEE International Conference , 2548-2555.

J, L. W. (2006). Simultaneous tracking and action recognition using the pca-hog descriptor . Computer and Robot Vision, 6-6.

J.Matas, O. M. (2004). Robust Wide-Baseline Stereo from Maximally Stable Extermal Regions. In British Machine vision Conference (BMVC), 384-396.

Kimmel, R. Z. (2011). Are MSER features really interesting? IEEE Transactions on Pattern Analysis and Machine Intelligence, 2316-2320.

Bay H, E. A. (2008). Surf:Speeded up robust features. Computer Vision and Image Understanding , 346-359.

B, D. N. (2005). Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition, IEEE Computer Society, 886-893.

D.G, L. (1999). Object recognition from local scale-invariant features. International Conference on Computer Vision(ICCV), 1150-1157.

Heinly, J. D.-M. (2012). Comparative evaluation of binary features. In Computer Vision-ECCV, Springer,Berlin,Heidelberg, 759-773.

Schuldt, C., Ivan, L., & Barbara, C. (2004). Recognizing Human Actions: A Local SVM Approach. 17th International Conference on Pattern Recognition , IEEE, 32-36.