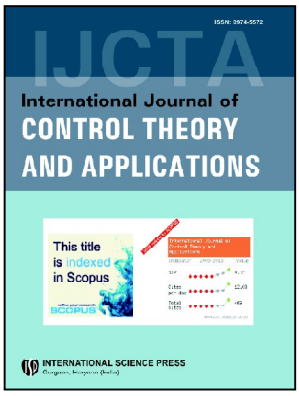


International Journal of Control Theory and Applications

ISSN : 0974-5572



Volume 10 • Number 6 • 2017

# Magnitude and Angle based Histograms for Human Action Recognition

**Mandlem Gangadharappa[1] and Rajiv Kapoor[2]**

[1] *Department of Electronics and Communication Engineering, Ambedkar Institute of Advanced Communication Technologies and Research (AIACTR), India, E-mail: iitkgangadhar@gmail.com*
[2] *Department of Electronics and Communication Engineering, Delhi Technological University (DTU), India, E-mail: rajivkapoor@dce.ac.in*

***Abstract:*** Human Action Recognition is a major area of research in computer vision as it is finding wide applications in systems such as content based video retrieval, intelligent video surveillance, gaming and animation, human computer interaction, anomalous behavior analysis etc. In this paper, Human action recognition problem is addressed based on the magnitude and angle histograms obtained from optical flow. In this method, the new bounding box sequence is extracted from the original sequence of images. Then the frames are cropped with respect to the bounding box coordinates in each of the frames of the video such that the centroid of the subject coincides with the center of the cropped frame. The optical flow field is computed over these newly cropped consecutive frames of the video. The resultant cropped frames of the subject are divided into eight regions and subsequently each frame is represented with eight histograms based on the magnitude and angle values of the optical flow vectors. For recognizing actions, training of the dataset is performed through k-means clustering. The testing of the system is done through the K nearest neighbor classifier. The algorithm found more robust in distinctively classifying different human actions over Weizmann dataset. The software MATLAB 2013a used for evaluating the results.

***Keywords:*** Optical Flow, Human Action Recognition, Motion Descriptors, K-means Clustering

## 1. INTRODUCTION

The objective of Human Action Recognition is to identify different actions of a human in a given video, taken under different scenarios automatically. In general HAR comprises of three basic steps, namely human motion detection, tracking and recognizing the relevant action [1-3]. The first step human motion detection involves motion segmentation and object classification. The motion segmentation is a difficult task to obtain in a given sequence, which identifies the movable region from the rest of the image based on several algorithms depending on the context and application. Object classification ensures motion of the human can be differentiated from other moving objects such as vehicles, flying birds, clouds etc. in the successive frames of a video.

The second step object tracking involves the process of locating the objects of interest in the sequential frames of a video. The object tracking methods are subdivided into six major categories: region based tracking,

contour based tracking, feature based tracking, model based tracking, hybrid tracking and optical flow based tracking. In region based tracking method, the objects are tracked by a particular group of varying pixels that constitutes different regions in the image corresponding to the moving objects in a sequence of images. In contour based tracking method, only the contour that is boundary of the object is tracked instead of the complete set of pixels that representing an object. This hugely reduces the computational cost. In feature based tracking method, the static features of the image such as geometry, texture and color, those are extracted from the object used for tracking. The dynamic features are also used in the literature for object tracking using Kalman filter, particle filter etc. In a model based tracking method, the target object is tracked with respect to an already defined model based on computer vision techniques. In hybrid tracking method, the combination of different number of approaches of object tracking, which are mentioned above are used instead of the particular method. The optical flow based tracking method uses the optical flow, a vector field, for computing intensity changes in the successive images with time [4]. Efros *et al.* [5] introduced motion descriptor based on Optical flow and motion similarity. Local motion descriptors are calculated from the optical flow orientation histograms used for human action recognition [6, 7]. A set of kinematic features based on optical flow vectors is used for human action recognition was proposed by Ali and Shah [8].

The third step, action recognition is based on nonhierarchical and hierarchical approaches. The nonhierarchical methods make use of space, time and sequential approaches. The nonhierarchical approach considers simple and short activities for recognition for eg. running, jumping, waving etc. It runs the matching algorithm on the testing image sequences with reference to the training set of predefined activities to recognize actions in the unknown image sequences. In the space time approach, activities are represented in volume, trajectories and set of features. In the sequential approach, human activities are recognized by analyzing the sequences of features extracted from the input video frames. For a particular class of activity in videos, the sequences of features are called observation sequences. Then the similarity measures are identified for the extracted sequences of observation in an input video. Hierarchical approaches define the recognition methodologies for complex human activities such as human object interactions and group based activities. One such hierarchical approach based on statistics is Hidden Markov Model [9] where, the multiple layers of statistical state based models are used for recognizing the actions. Some of the recent work [10], [11] in the area of human action recognition considers silhouette based segmentation analysis to improve the recognition accuracy.

The motivation behind the proposed work is to make the algorithm more robust to recognize human actions by constructing efficient motion descriptor. The proposed approach to the problem of Human activity recognition is to compute the motion descriptors based on the magnitude and angle histograms of optical flow vectors. The merits of motion based features are, invariant to subject appearance and can be effectively obtained irrespective of the video having lower resolution. The local motion descriptors can be obtained by subdividing the subject frame into various regions and corresponding histogram characteristics. The time sequence information can be obtained from the video can be used to discriminate human actions with the global motion histogram characteristics.

The rest of the paper is organized as follows. Proposed algorithm and the methodology are explained in section 2. The results along with corresponding discussions to appreciate the proposed method are presented in section 3. Concluding remarks are given in section 4.

## 2. PROPOSED ALGORITHM

The basic steps involved in the formulation of the proposed methodology are given as follows. Initially the bounding boxes are extracted for each frame of the video. A new sequence of frames is obtained by scaling and centrally aligning the bounding boxes. The optical flow vectors are computed for each frame of the video sequence. Further, each frame is divided into eight regions and corresponding magnitude and angle histograms of optical

flow vectors are obtained. The concatenation of the histograms of the eight regions is formed get the final motion descriptor for each frame. The clustering is done based on similar action frames and label them with the same cluster ID and the actions are classified using K-Nearest Neighbor classifier. The Proposed methodology is explained with the help of the block diagram as shown in Figure 1.

## 2.1. Bounding box Extraction and new sequence formation

First of all, the initial human action video frames are converted to grayscale images and background subtraction is performed. The subsequent step is to extract bounding box in all the frames of the input video. The centroid and bounding box coordinates are found to create a new sequence of frames. This sequence of images (i.e. comprising the subject) is cropped by taking corresponding bounding box coordinates into consideration, such that the centroid and the center of the cropped sequence of frames coincide. Thus, the new centrally aligned foreground (grayscale) depicting a sequence of frames is obtained. These frames are then used for further processing.

## 2.2. Optical Flow Computation

The next step is to compute the magnitude as well as angle values of optical flow vectors for a new sequence of frames. The method used for computing optical flow is Lucas Kanade method [4]. Optical flow, calculates the velocity for points within the images, and provides estimation that, where points could be in the next image sequence. Before applying this method the frames are subject to a low pass filter to reduce the noise in the images. Then for each two consecutive frames, the angle and magnitude values of the optical flow vectors are evaluated using equation (1) and (2) given as,

$$mag = \sqrt{u^2 + v^2} \tag{1}$$

$$theta = \tan^{-1}\frac{v}{u} \tag{2}$$

Where $u$ = horizontal component of optical flow vector

$v$ = vertical component of optical flow vector

The optical flow equation (3) has a constraint that it has two unknown's $u$ and $v$ in terms of intensity values is given as,

$$\mathbf{I_x}\,u + \mathbf{I_y}\,v + \mathbf{I_t} = 0 \tag{3}$$

**W**here $\mathbf{I_x}$, $\mathbf{I_y}$ and $\mathbf{I_t}$ are the partial derivatives of intensity with respect to x, y coordinates and time't' respectively.

From the above equation, only the normal component of the flow can be obtained, the Lucas Kanade method is one such method which helps in removing this constraint. Lucas Kanade method treats the following two assumptions:

(i) In a sequence of images, any two consecutive images are separated by a small time constraint, that is, the object in the two images is not displaced much.

(ii) There are no illumination changes.

It solves the optical flow equations for all the pixels in the neighborhood of a given pixel since it assumes that that optical flow is constant within the neighborhood of a given pixel. It is done by Least Squares criterion.

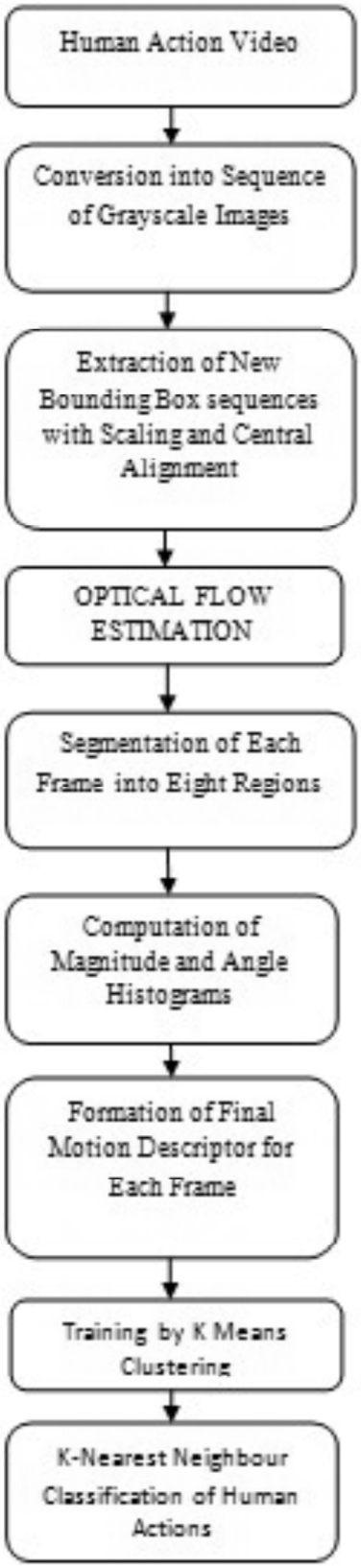


**Figure 1: Flow Diagram of Proposed Methodology**

### 2.3. Segmentation of each frame into eight regions

The segmentation process is used to divide the frames of the new sequence into eight regions. A vertical and a middle horizontal line passing through the center of the frame are created. Then a top horizontal line and bottom horizontal line are created on the frames. The top horizontal line is made through the middle of the top subject region and the middle horizontal line. Likewise, the bottom horizontal line is made through the middle of the middle horizontal line and the bottom subject region.

### 2.4. Computation of normalized Magnitude and Angle Histograms

The next step is to construct the normalized histograms of magnitude and angle in each region of each frame, containing optical flow vectors. The magnitude (mag) values of equation (1) are normalized for each frame such that they lie in 0 to 1 range. The theta values of equation (2) are converted from radian to degrees into a 0 to 360 degrees scale. The theta values are then quantized into a set of $\{0^0, 45^0, 90^0, 135^0, 180^0, 225^0, 270^0, 315^0\}$. Two cases are considered with respect to magnitude. One in which the magnitude is quantized to two levels and the other in which the magnitude is quantized into four levels. In case of two levels, the magnitude values are quantized into {0.5, 1}. In case of four levels, the magnitude values are quantized into {0.25, 0.5, 0.75, 1}. Each bin is having an angle of 45 degrees spacing and magnitude of 0.25 spacing. Here, the concept of magnitude quantization is fused with the idea of angles quantization as given by Lertniphonphan et al. [7] and Lucena et al. [6]. This proposed methodology makes use of quantized values of both magnitude and theta to define the motion descriptor. The vertical axis in the histogram plot denotes the number of pixels in each of these bins. The horizontal axis in the histogram plot denote the different bins used. Then the histograms in each of the regions are normalized i.e. the number of pixels in each bin is divided by the sum total of all the pixels present in all the bins.

### 2.5. Formation of final motion features for each frame

For each frame, a final motion feature vector is obtained by concatenating the histograms of all the eight regions. In case of magnitude of two levels, each region consists of a histogram of 16 bins. By horizontally concatenating the eight region histograms, a final motion feature vector of 16*8=128 features or values are obtained. In case of magnitude of 4 levels, each region consists of a histogram of 32 bins. By horizontally concatenating the eight region histograms, a final motion feature vector of 32*8=256 features or values are obtained. These feature values are used to define the final motion descriptor.

The training based on obtained motion descriptors, is done using K-means clustering and testing of any given sequence of human action is classified into particular action using K nearest neighbor classifier.

## 3. RESULTS AND DISCUSSIONS

The results are obtained by training and testing the proposed algorithm on Weizmann video dataset of human actions, for illustration one frame of each action is represented in Figure 2.

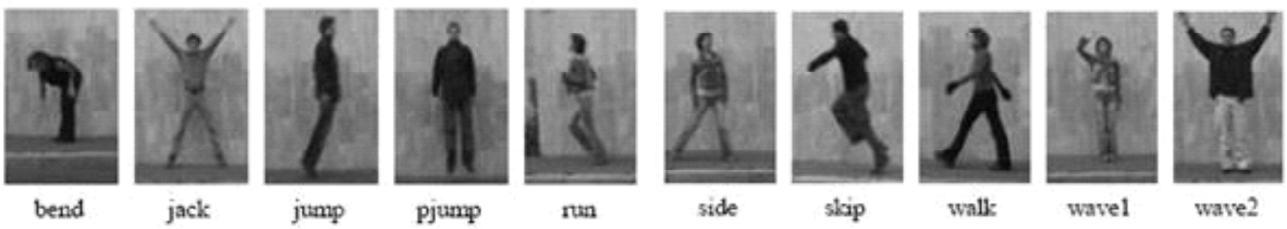


**Figure 2: Weizmann Human action dataset, single frame depiction for all 10 actions**

The screenshots at various stages of the proposed methodology is presented in Figure.3. The proposed methodology implemented using the software MATLAB 2013a. In Figure 3, the first Image (a) represents one of the original frame of human action taken from standard Weizmann dataset, the Image (b) obtained after background subtraction to consider only motion based pixels. Figure.3 (c), represents the frames depicting centroid with bounding box and (d) is obtained frame after alignment and cropping. Optical Flow vectors are shown over the interested human action image in figure 3 (e) to evaluate the magnitude and angle parameters. The Image is segmented into eight regions as shown in Figure 3 (f) and the dataset of magnitude and angle based histograms are extracted to define motion descriptor for each frame

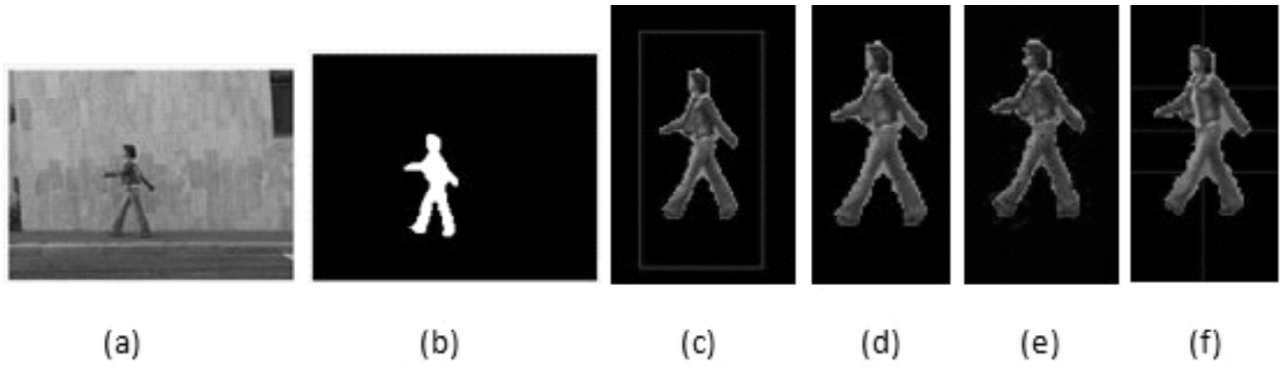


**Figure 3**

(a) Input video frame
(b) Background subtracted frame
(c) Frame depicting centroid and bounding box around the subject.
(d) Aligned and cropped frame
(e) Frame of new sequence depicting the optical flow vectors
(f) Frame of new sequence showing eight regions.

K-means clustering is used as a clustering technique to partition the different actions of Weizmann dataset based on the motion descriptor features obtained by magnitude and angle based histograms. The training in the system is done through K-means clustering. The classification model is built at this stage only, by storing the values of the final centroids, output by the system. These centroids will act as the representative of a whole cluster at the classification stage.

The testing of an unknown and random human video action is done using the K-Nearest Neighbors (KNN) classifier. KNN is a nonparametric method of classification. It is a simple method that stores all the representative cluster centroids of different clusters and based on them assign a class to the unknown observation with the help of the similarity distance measure. We have used the squared Euclidean distance as the distance measure. Initially the parameter K defines the number of nearest neighbors that needs to be taken into consideration for the final output. The distance between the query instance and the training samples are calculated. The distance measures are sorted and the nearest neighbors are determined based on the K'th minimum distance. The corresponding categories of the nearest neighbors are determined. The simple majority of the category of nearest neighbors is used as the prediction value of the query instance.

Table 1 Represents "Pjump" human action classified and tested over different persons performing the same action from Weizmann Dataset. Out of 212 frames tested only 12 frames are misclassified and 200 frames are classified as correct cluster ID.

**Table 1**
**Results of Human action (Pjump) by the proposed algorithm**

| *Human Action (Pjump)* | *Total frames tested* | *Misclassified frames* |
|---|---|---|
| Denis_pjump | 26 | 3 |
| Ido_pjump | 23 | 0 |
| Ira_pjump | 63 | 1 |
| Lena_pjump | 24 | 2 |
| Lyova_pjump | 27 | 1 |
| Moshe_pjump | 22 | 1 |
| Shahar_pjump | 27 | 4 |

Similarly, Table 2 Represents 'Walk" human action classified and tested over different persons performing the same action from Weizmann Dataset. Out of 189 frames tested only one frame is misclassified and 188 frames are classified with the same action class.

**Table 2**
**Results of Human action (Walk) by the proposed algorithm**

| *Human Action (Walk)* | *Total frames tested* | *Misclassified frames* |
|---|---|---|
| Eli_walk | 35 | 0 |
| Lyova_walk | 24 | 0 |
| Lena_walk | 36 | 0 |
| Denis_walk | 35 | 1 |
| Moshe_walk | 38 | 0 |
| Ido_walk | 21 | 0 |

Similarly, all ten Human actions (nine persons) of Weizmann dataset were tested with the proposed algorithm and gives the phenomenal Average recognition accuracy of 98.8%, i.e. only 1.2 % frames are classified erroneously.

## 4. CONCLUSION

The proposed methodology is applied to the standard Weizmann dataset. The Weizmann dataset contains ten actions performed by nine persons. The proposed algorithm can able to distinguish the different human actions with very few frames are misread in wrong action class. Overall, the proposed algorithm was found more robust to classify all 10 human actions efficiently and gives a remarkable recognition accuracy of 98.8%. The proposed method is also finding computationally efficient to enable real-time applications.

## REFERENCES

[1] Ronald Poppe, "A Survey on vision-based Human Action Recognition" Image and Vision Computing, Elsevier, vol. 28, 2010, pp.976-990.

[2] D.Weinland, R. Ronfard, E.Boyer, "A Survey on vision-based methods for action Representation, segmentation and Recognition". Computer Vision and Image Understanding, Elsevier, vol. 115, 2011, pp. 224-241.

[3] Sarvesh Vishwakarma, Anupam Agrawal, "A Survey on activity recognition and behavior understanding in video surveillance," Vis. Comput., Springer, vol. 29, September 2012, pp. 983-1009.

[4] Lucas B.D, Kanade T. "An Iterative Image Registration Technique With An Application To Stereovision" Proceedings Imaging Understanding Workshop, 1981, pp.121– 130.

[5] Efros A.A, Berg A.C, Mori G, Malik J, "Recognizing Action At A Distance" Proceedings of International Conference on Computer Vision, France, 2003.

[6] Lucena M, Perez de la Blanca N, Fuertes J M, "Human Action Recognition Based on Aggregated Local Motion Estimates". Machine Vision and Applications, Springer, 2012, pp.135-150

[7] Lertniphonphan k, Aramvith S, Chalidabhongse T H, "Human Action Recognition using Direction Histograms of Optical Flow". 11th International symposium on Communications & Information Technologies, 2011.

[8] Saad Ali, Mubarak Shah: 'Human action recognition in videos using kinematic features and multiple instance learning', IEEE Trans. Pattern Analysis And Machine Intelligence, February 2010, vol.32, no.2, pp. 288-303

[9] Zia Moghaddam, Massimo Piccardi: 'Training initialization of Hidden Markov Models in Human Action Recognition', IEEE Trans. Automation Science and Engineering, April 2014, vol.11, no.2, pp. 394-408

[10] Vishwakarma, D.K., Rajiv Kapoor: 'Hybrid classifier based human activity recognition using the silhouette and cells', Expert Systems with Applications, 2015, vol.42, pp. 6957-6965

[11] Jian Cheng, Haijun Liu, Feng Wang et al.: 'Silhouette analysis for human action recognition based on supervised temporal t-SNE and incremental learning', IEEE Trans. Image Proccessing, October 2015, vol.24, no.10, pp. 3203-3217.