I J C T A, 9(3), 2016, pp. 1535-1543

CMI-HoG for Human Emotion Recognition in Video using Tree based Classifiers

S. Gowsalya*, M. Kalaiselvi Geetha and J. Arun Nehru*****

Abstract: Human emotion recognition is an extensive research area in human computer vision community. Human emotions are identified by gesture of body movements. In this paper, an emotion recognition approach based on body gesture is proposed. Cumulative Motion Image based Histogram Oriented Gradient (CMI-HoG) is extracted as features. The experiments were carried out using emotion dataset considering nine actions (Happy-walking, Happy-sitting, Happy-jumping, Angry-walking, Angry-sitting, Angry-jumping, Fearful-walking, Fearful-sitting and Fearful-jumping) and the various tree based classifier like Decision Tree (J48), Random Tree, and Random Forest are utilized. In the experimental results, random forest classifier showed the best performance with an overall accuracy rate of 82.42% which outperformed other algorithms.

Keywords: Video surveillance. Action recognition. Frame difference. Cumulative motion image. Decision tree. Random tree. Random forest

1. INTRODUCTION

Human computer interaction based system is to give the ability to computers to evaluate the human emotion robustly. The large application of human emotion and action recognition is found in interactive educational systems, human recognition, web movies and surveillance, just to name a few all give rises to the need of robust human emotion recognition. A video surveillance system observes action in open public areas such as airport, railway station, bus stand, banks, gas stations, ATM, and commercial buildings for real-time or later analysis, and it is deterrent to crime. Three major principal emotions are: anger, happy and fearful. The aim of emotion recognition is an automatic interpretation of current events and their context from video sequences when a human performs an action.

Human emotion recognition aims to recognize and classify the action of a person into certain categories, i.e., walking, sitting, jumping, bending, skipping, and hand waving. It is a complex process to recognize human activity due to many factors such as occlusion, illumination changes, postures, speed, shadows, and clothing. Gesture recognition in identifying the action and emotion of human behaviors helps to find out the minute gesture dynamics of video sequences. Gesture can initiate from any bodily motion or formal but generally create from the face or hand. Current focus in the field includes emotion recognition from human body gestures. The recognition of whole-body expressions is significantly durable, because the form of the human body has more degrees of freedom than the face on your own, and its overall shape varies strongly during expressed motion. In this paper, propose human emotion recognition of whole body expression with image sequence features.

1.1. Related Work

Recent surveys in the area of human action analysis in [1] and [2] focus on the feature descriptor, representation, and classification model in video sequences. Survey by Turaga et al.[3] centers around

* Dept. of Computer Science Engg., Annamalai Universtiy, Annamalainagar, India, *Email: gowsalyasub@gmail.com*

** Dept. of Computer Science Engg., Annamalai Universtiy, Annamalainagar, India, *Email: geesiv@gmail.com*

*** Dept. of Computer Science Engg., Annamalai Universtiy, Annamalainagar, India, *Email: arunnehru.aucse@gmail.com*

recognition of human activity. Wang et al. [4] propose a method which relies on optical flow and edge features, where these two discriminative features were combined to extract the motion and shape descriptors to distinguish one action from another. Reddy et al. [5] present a method to recognize action-based on sphere/rectangle tree structure that is built with spatio temporal interest point features. In [6], vocabulary forest is constructed with local static and optical flow features and uses trees and forests for action recognition classification. Zhu et al. [7] address a multi-view action recognition algorithm based on local similarity random forests and sensor fusion, and normalized silhouettes are used as pose features and effective for multiple view human action recognition. Haiyong et al. [8] present a novel classification method to recognize the actions from videos based on centroid-radii model descriptor and to train and classify video sequences by nonlinear SVM decision tree (NSVMDT). In [9], a method was proposed based on extended motion template from human silhouettes. The holistic structural features were extracted from motion templates to discriminate the human action, which represents local and global information. Zhang et al. [10] have used the spatial–temporal with optical flow features. The temporal consistence of motion is improved with an enhanced DTW method to recognize the human actions. Wang et al. [11] propose a method which relies on optical flow and edge features, where these two discriminative features were combined to extract the motion and shape descriptors to distinguish one action from another.

1.2. Outline of the Work

This paper deals with human emotion recognition that aims to understand human actions from video sequences and then to identify their emotions while doing these action. The proposed method is evaluated using University of York [12] emotion dataset with the person showing actions such as walking, sitting and jumping with emotion like happy, angry and fearful. Difference image is obtained by subtracting the consecutive frames. Cumulative motion image (CMI) is calculated by combining the five frame difference images and then HoG features are extracted from CMI. The extracted feature is fed to the tree-based classifier such as decision tree (J48), random tree and random forest.

The rest of the paper is organized as follows. Section 2 explains the feature extraction method. Section 3 explains the workflow of the proposed approach. Section 4 presents the Experimental results. Finally, Section 5 concludes the paper.

2. FEATURE EXTRACTIONS

Feature is a descriptive characteristic extracted from an image or video sequences, which represent the meaningful data that are vital for further analysis. The following subsections present the description of the feature used in this work. The block diagram of the proposed approach is shown in Fig. 1.

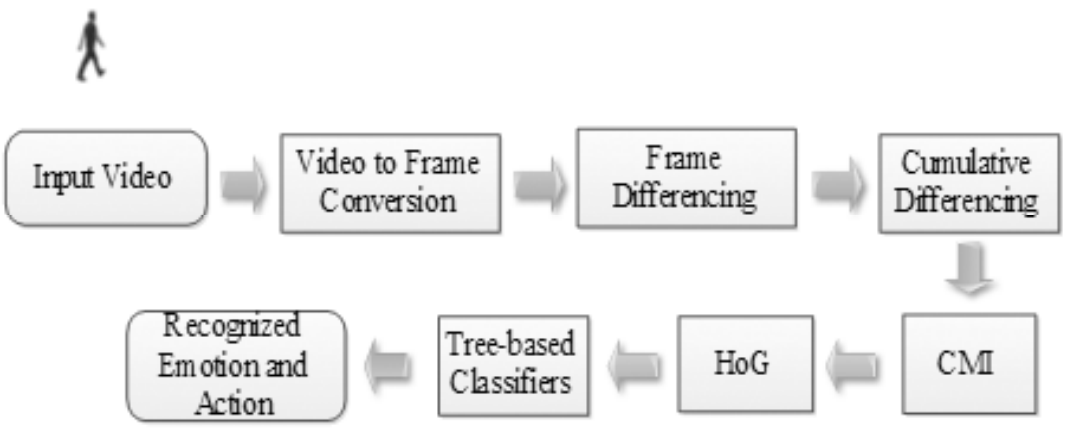

Figure 1: Block diagram of the proposed approach.

2.1. Frame Differencing

Frame differencing is defined by the difference between two consecutive frames, as an replacement of subtracting a predefined background while in motion, the subtraction of frames model considers every pair of consecutive frames of time t and $t + 1$ to extract any motion details in it. In order to find the regions of interest, by previous frame simply subtracting the current frame on a pixel by pixel model, Fig. 2(a), Fig. 2(b) shows the consecutive frames of the emotion dataset. The frame difference image of the emotion and action is shown in Fig. 2(c). Then the value of the difference image is related with a determine threshold value. The image at time t is given by:

$$D_t(x, y) = \left| I_t(i, j) - I_{t+1}(\mathrm{i}, \mathrm{j}) \right| \quad 1 < i < w, 1 < j < h \tag{1}$$

$I_t(i, j)$ is the amount of the pixel (i, j) in the emotion dataset frame, w and h are the width and height of the image respectively. The proposed method uses an image size of 960 × 540.

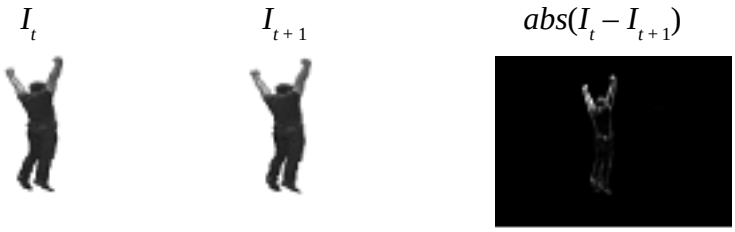

Figure 2: (a), (b) Two consecutive frames. (c) Consecutive frames of (a) and (b) from emotion dataset.

2.2. Cumulative differencing

n-frame cumulative differencing is applied to identifying the region as seen in Fig. 3.

$$D_k(\mathrm{x}, \mathrm{y}) = I_t(x, y) - I_{t+1}(x, y) \quad 1 \le x \le w,\ 1 \le \mathrm{y} \le \mathrm{h} \tag{2}$$

The next step is to calculate then D_k is the difference image found by subtracting by two consecutive frames I_t and I_{t+1}. $I(x, y)$ is the pixel intensity values of (x, y), w and h are width and height of the image respectively. Consecutive difference images are calculated as follows:

$$\begin{aligned} D_n(x, y) &= I_n(x, \mathrm{y}) - I_{p+1}(x, y) \\ D_{n+1}(\mathrm{x}, \mathrm{y}) &= I_{p+1}(\mathrm{x}, \mathrm{y}) - \mathrm{I}_{\mathrm{p}+2}(\mathrm{x}, \mathrm{y}) \\ D_{n+2}(x, y) &= I_{p+2}(x, y) - I_{p+3}(x, y) \\ D_{n+k}(x, y) &= I_{p+k}(x, y) - I_{p+k+1}(x, y) \end{aligned} \tag{3}$$

2.3. Cumulative Motion Images (CMI)

Motion information in a video sequence is extracted by pixel-wise differencing of consecutive frames. CMI creation processes for walking action is shown in Fig 4, the moving image is gradually received by the video. CMI is calculated using:

$$H(x, y, t) = \begin{cases} \tau & \text{if B(x,y,t)=1} \\ \max(0, \mathrm{H}(\mathrm{x},\mathrm{y},\mathrm{t}\text{-}1) - \delta) & \text{otherwise} \end{cases} \tag{4}$$

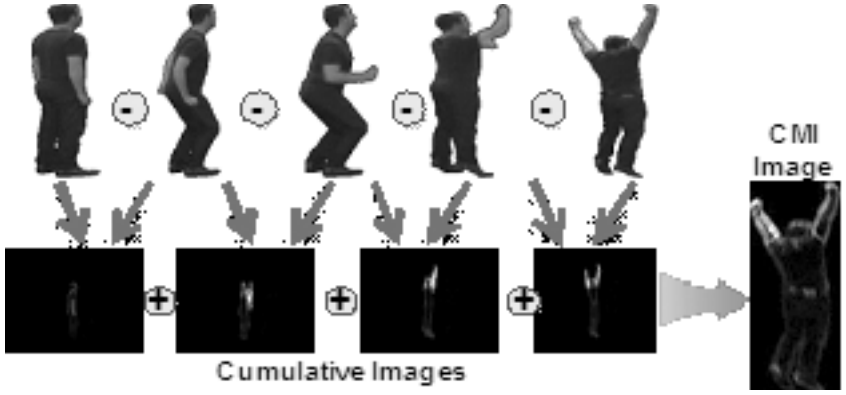

Figure 3: Cumulative Motion Image (CMI).

Where $H(x, y, t)$ is the present CMI, $H(x, y, t\text{-}1)$ is the earlier CMI, $B(x, y, t)$ is the current binary image, τ is the maximum value of importance degree, and δ is the reducing value of the importance degree. If the pixel value of the current incoming binary image $B(x, y, t)$ is one, the pixel value of CMI is the maximum value. CMI subtracts the reduction coefficient from the pixel value of the previous CMI; a higher value is then selected after comparing the subtraction value and minimum value (zero). Accordingly, the pixel value where the action is not shown is zero. After the CMI creation, features are executed using HoG.

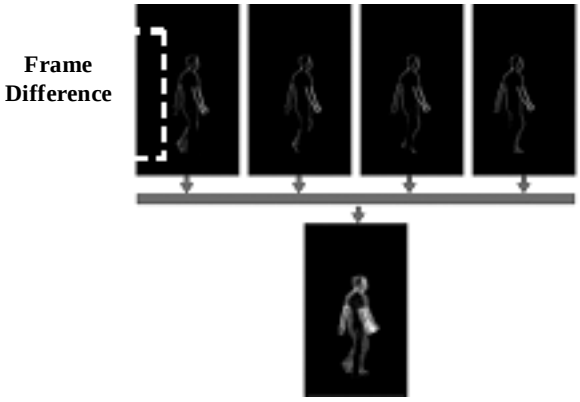

Figure 4: CMI obtained for walking action.

2.4. Histogram of oriented Gradients (HoG)

Histograms are collected totals of data prepared into a set of predefined bins. Imagine that a matrix contains information of an image intensity value in the range 1 to 256. The histogram of an image is a design of the gray-level values vs. the number of pixels at that value. The feature will be used then for classification of emotion and action recognition. The HoG feature process first calculates the gradient of the input image. A typical method is to apply a one-dimensional discrete differential mask as the horizontal orientation (D_k = [–1 0 1]) and vertical orientation

$$(D_Y = [-101]^T).$$

Convolution mask of horizontal orientation:

$$I_X = H(x, y, t) * D_X \tag{5}$$

Convolution mask of vertical orientation:

$$I_Y = H(\mathrm{x}, \mathrm{y}, \mathrm{t}) * \mathrm{D}_\mathrm{Y} \tag{6}$$

Size of gradient: $|G| = \sqrt{I_X{}^2 + I_Y{}^2}$ (7)

Orientation of gradient: $\theta = \arctan \frac{I_Y}{I_X}$ (8)

Signed gradient: $\alpha_{Signed} = \begin{cases} \alpha & \alpha \geq 0 \\ \alpha + 360 & \alpha < 0 \end{cases}$ (9)

Unsigned gradient: $\alpha_{Signed} = \begin{cases} \alpha & \alpha \geq 0 \\ \alpha + 180 & \alpha < 0 \end{cases}$ (10)

When the image is created, convolution masks of the horizontal and vertical orientations (Equations (5) and (6), respectively) are applied to the image, and the orientation and slope size are calculated. Second, calculate histograms of the cells are divided. The value of each pixel in the cell is calculated as the gradient of the orientation through an advanced gradient calculation. These orientation histogram bands are spread on values, which are set as the number of bins. Rectangular shapes in the image of comprised are cells. As an expression of the gradient, the histogram bands are evenly distributed from 0 to 360 degrees (Equation (9)) or from 0 to 180 degrees (Equation (10)). HoG features to extract the CMI images are obtained. Then image is divided into 3 × 3 cells, where each cell consists of 9 bin histogram [13], in order to obtain the HoG features. The compute extraction process of CMI-HoG is shown in the Fig. 5.

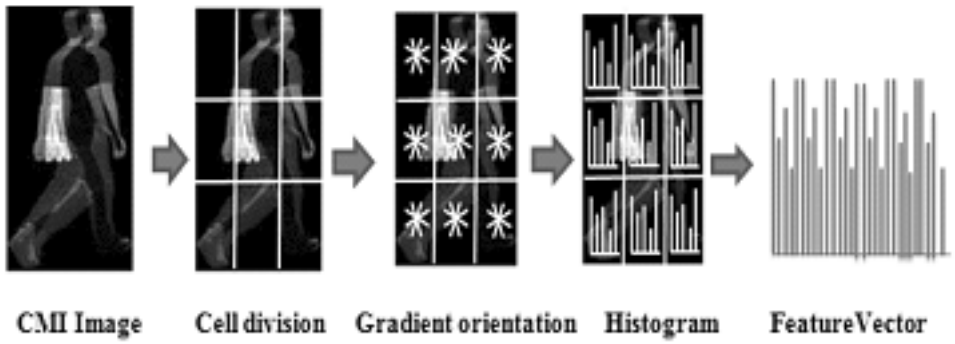

Figure 5: Extraction process of CMI-HoG feature vector.

3. WORKFLOW OF THE PROPOSED CMI ACTION RECOGNITION APPROACH

Emotion datasets are used for experimental purpose. The video is processed at 25 frames per second. It is essential to preprocess all video sequences to remove noise for fine features extraction and classification. CMI-HoG features are extracted as discussed in Section 2. The obtained features are fed to the tree-based classifier for emotion recognition. In this work, different tree-based classifiers such as decision tree (J48) [14], random tree and random forest [15] are used in order to evaluate the effectiveness of these classifier on emotion dataset.

4. EXPERIMENTAL RESULTS

In this section, the proposed method is evaluated using emotion dataset. The experiments are carried out in MATLAB 2013a in Windows 7 Operating System on a computer with Intel Xeon Processor 2.40 GHz with 4 GB RAM. The obtained CMI-HoG features are fed to tree-based classifiers such as decision tree (J48), random tree and random forest using open source machine learning tool WEKA [16] to develop the model for each activity, and these models are used to test the performance for each tree-based classifier.

4.1. Emotion Dataset

The proposed approach is evaluated on theEmotion dataset (University of York) is a publicly available dataset, containing four different emotions (happy, angry, fear and sad) performed by 25 actors.The sequences were taken over static (black) background with the frame size of 1920 × 1080 pixels at a rate of 25 fps. For each emotion, actors are performed five different actions: walking, jumping, box picking, box dropping and sitting having an approximate length of 15 seconds of video. In this work, three emotions (happy, angry and fear) and three actions (walking, jumping and sitting) of 10 persons (male and female) are considered for experimental purpose. The sample frames of the emotion dataset are shown in Fig. 6.

Figure 6: Sample frames from the emotion dataset.

4.2. Quantitative Evaluation

As explained in Section II, D, the 81-dimensional feature vectors are extracted. The performance is evaluated using tenfold cross-validation approach. The obtained features are fed to the tree-based classifier. To evaluate the performance of the proposed approach, precision (P) = TP/TP + FP, recall (R) = TP/TP + FN, specificity (S) = 100 × *TN/TN* + *FP*, F-measure (F1) = 2P.R/P + R are used, where TP and FP are the numbers of true-positive and false positive predictions for the particular class, and TN and FN are the number of true negative predictions and false-negative predictions for the particular class, and build time of all tree-based classifiers is also measured.

4.3. Experimental Results on Tree-based Classifiers

The experiment was done using an open source machine learning tool WEKA. The performances of the classifiers were measured using tenfold cross-validation model. Table 1 shows the confusion matrix of the decision tree classifier for the emotion dataset. Table 2 shows the confusion matrix of the random tree classifier for the emotion dataset. Table 3 shows the confusion matrix of the random forest classifier. Nine emotion based actions are considered viz (Happy-walking, Happy-sitting, Happy-jumping, Angry-walking, Angry-sitting, Angry-jumping, Fearful-walking, Fearful-sitting and Fearful-jumping) dataset, where correct response defines the main diagonal, and majority of actions are correctly classified. Table 4 shows the accuracy results obtained for decision tree (J48), random tree and random forest on Emotion dataset. When represents the precision, recall, specificity, F-measure, and time (s) values for all three tree based classifiers and shows that random tree takes less time and gives better accuracy rate of 74.53%, the accuracy of J48 is 72.10%, Decision tree consumes minimal time to build the model but also shows the accuracy rate of 72.10 % having rank two. Random forest classifier consumes nominal time to build the model but also shows the

Table 1
Confusion Matrix of the Decision Tree Classifier for the Emotion Dataset

Class	Happy (walk)	Happy (sit)	Happy (jump)	Angry (walk)	Angry (sit)	Angry (jump)	Fearful (walk)	Fearful (sit)	Fearful (jump)
Happy (walk)	66.48	0.00	3.30	12.64	0.00	3.30	10.99	0.00	3.30
Happy (sit)	0.00	79.82	1.20	0.00	10.84	0.60	0.00	6.02	1.51
Happy (jump)	5.22	3.21	71.89	2.41	0.00	6.43	2.41	2.01	6.43
Angry(walk)	14.56	0.63	3.16	69.62	0.63	1.27	6.96	0.63	2.53
Angry(sit)	0.37	14.29	0.37	1.10	75.09	1.10	0.00	6.23	1.47
Angry(jump)	2.11	2.11	10.53	1.58	2.63	69.47	1.58	0.00	10.00
Fearful (walk)	10.92	0.57	6.90	8.05	0.00	2.30	66.09	1.15	4.02
Fearful (sit)	0.41	8.71	2.07	0.41	10.37	1.24	0.41	74.69	1.66
Fearful (jump)	0.93	1.86	10.70	1.40	3.72	9.30	3.26	1.40	67.44

Table 2
Confusion Matrix of the Random Tree Classifier for the Emotion Dataset

Class	Happy (walk)	Happy (sit)	Happy (jump)	Angry (walk)	Angry (sit)	Angry (jump)	Fearful (walk)	Fearful (sit)	Fearful (jump)
Happy (walk)	70.33	0.00	4.40	10.99	1.10	0.55	9.89	1.10	1.65
Happy (sit)	0.60	81.02	0.90	0.00	7.53	1.20	0.30	7.53	0.90
Happy (jump)	2.01	2.81	76.31	0.40	1.61	6.02	2.41	2.81	5.62
Angry(walk)	15.82	0.00	0.63	68.35	1.90	1.90	8.23	1.90	1.27
Angry(sit)	1.10	8.06	1.83	0.73	79.12	0.73	0.00	6.23	2.20
Angry(jump)	2.11	2.11	7.37	3.68	2.63	70.53	3.16	1.58	6.84
Fearful (walk)	5.17	0.57	5.17	10.92	1.15	2.87	70.11	0.57	3.45
Fearful (sit)	1.66	10.79	1.66	0.41	5.81	2.07	1.66	74.69	1.24
Fearful (jump)	3.72	0.93	7.91	1.86	1.86	5.58	4.19	2.33	71.63

Table 3
Confusion Matrix of the Random Forest Classifier for the Emotion Dataset

Class	Happy (walk)	Happy (sit)	Happy (jump)	Angry (walk)	Angry (sit)	Angry (jump)	Fearful (walk)	Fearful (sit)	Fearful (jump)
Happy (walk)	76.37	0.00	3.30	8.79	0.00	3.30	7.69	0.00	0.55
Happy (sit)	0.00	87.95	0.60	0.00	3.31	0.90	0.00	7.23	0.00
Happy (jump)	1.61	1.20	86.75	0.00	0.40	4.82	0.00	0.80	4.42
Angry(walk)	13.92	0.00	1.27	75.95	0.00	0.63	7.59	0.63	0.00
Angry(sit)	0.00	7.69	0.73	0.00	83.88	0.37	0.00	6.96	0.37
Angry(jump)	1.05	0.53	12.11	0.00	2.11	77.37	0.00	0.53	6.32
Fearful (walk)	13.79	0.57	0.00	4.02	0.00	1.15	78.74	0.00	1.72
Fearful (sit)	0.41	6.22	1.24	0.00	5.81	0.83	0.00	85.48	0.00
Fearful (jump)	0.93	0.93	9.77	0.00	2.33	3.72	1.40	0.00	80.93

accuracy rate of 82.42 %, and. Finally, the random forest classifier shows highest accuracy percentage when compared to other classifier algorithms.

From this, Happy-walk, Angry-walk and Angry-jump emotions are confused. The overall performance

Table 4
Performance Measure of the Tree-based Classification Algorithms

Classifier	Precision (%)	Recall (%)	Specificity (%)	F-measure (%)	Time (Sec)
Decision Tree	71.41	71.18	96.50	71.27	1.38
Random Tree	73.56	73.57	96.81	73.55	0.09
Random Forest	82.10	81.49	97.79	81.71	5.48

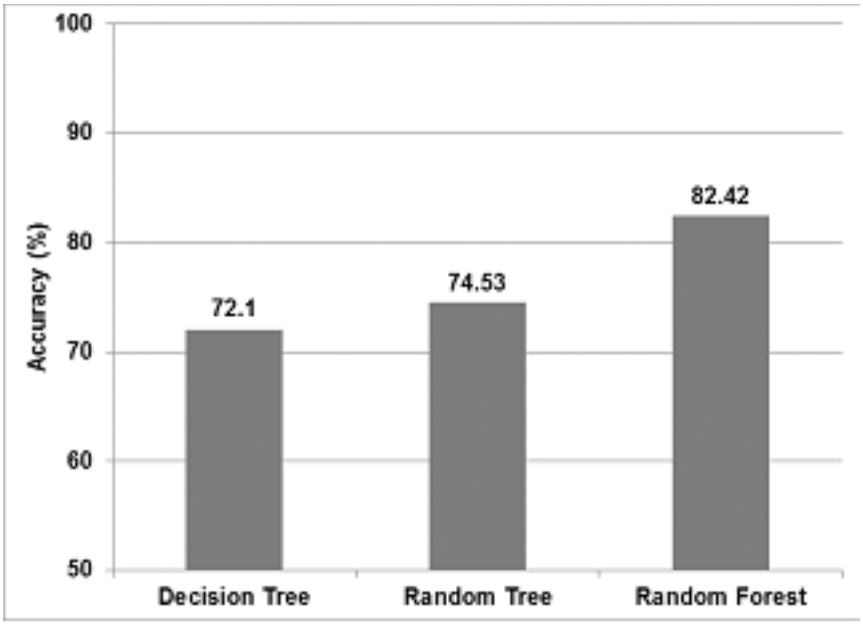

Figure 7: Performance measure of the different classifiers

of the proposed CMI-HoG with different tree-based classifier on Emotion dataset is shown in Fig. 7. As expected, random forest excelled in accuracy among other approaches.

5. CONCLUSIONS AND FUTURE WORK

This paper presented a method for emotion recognition using CMI-HoG as a feature. Experiments are conducted on Emotion dataset considering different actions viz (Happy-walking, Happy-sitting, Happy-jumping, Angry-walking, Angry-sitting, Angry-jumping, Fearful-walking, Fearful-sitting and Fearful-jumping). This approach evaluates the performance of CMI-HoG feature in video sequence using tree-based classification algorithms such as decision tree, random tree, and random forest. The recognition results are obtained for the emotion dataset. It shows that that the average recognition accuracy was 82.42% with random forest and the random forest performs best, when compare with Random Tree, and Decision Tree classifiers, were 72.10 and 74.53 respectively.

References

[1] R. Poppe, Vision-based human motion analysis: an overview. Comput. Vis. Image Underst. (CVIU) 108(1–2), 4–18 (2007).

[2] R. Poppe, A survey on vision-based human action recognition. IVC 28, 976–990 (2010).

[3] P. Turaga, R. Chellappa, S. Venkatramana Subrahmanian, O. Octavia Udrea, Machine recognition of human activities: a survey. IEEE Trans. Circ. Syst. Video Technol. 18(11), 1473–1488 (2008).

[4] L. Wang, Y. Wang, T. Jiang, D. Zhao, W. Gao, Learning discriminative features for fast frame-based action recognition. Pattern Recogn. 46(7), 1832–1840 (2013).

[5] K. Reddy, J. Liu, M. Shah, Incremental action recognition using feature-tree, in International Conference on Computer Vision (2009).

[6] Z. Lin, Z. Jian, L. Davis, Recognizing actions by shape motion prototype trees, in International Conference on Computer Vision (2009).

[7] F. Zhu, Ling Shao, Mingxiu Lin.: Multi-view action recognition using local similarity random forests and sensor fusion. Pattern Recogn.Lett. 34(1), 20–24 (2013).

[8] H. Zhao, Z. Liu, Human action recognition based on non-linear SVM decision tree. J. Computat. Infor. Syst. 7, 2461–2468 (2011).

[9] D. Wu, L. Shao, Silhouette analysis-based action recognition via exploiting human poses. IEEE Trans. Circuits Syst. Video Techn. 23(2), 236–243 (2013).

[10] W. Zhang, Y. Zhang, C. Gao, J. Zhou, Action recognition by joint spatial-temporal motion feature. J. Appl. Math. (2013).

[11] L. Wang, Y. Wang, T. Jiang, D. Zhao, W. Gao, Learning discriminative features for fast frame based action recognition. Pattern Recogn. 46(7), 1832–1840 (2013).

[12] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in Proceedings of IEEE International Conferences on Computer Vision (2005), pp. 1395–1402.

[13] Eum, Hyukmin, et al. “Continuous Human Action Recognition Using Depth-MHI-HOG and a Spotter Model.” Sensors 15.3 (2015): 5197-5227.

[14] L. Breiman, Random forest. Mach. Learn. 45(1), 5–32 (2001).

[15] P. Langley, W. Iba, K. Thompson, An analysis of bayesian classifiers, in Proceedings of the Tenth National Conference on Artificial Intelligence (1992), pp. 223–228.

[16] J.R. Quinlan, C4.5: Programs for Machine Learning (Morgan Kaufmann Publishers, Burlington, 1993).

[17] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (Morgan Kaufmann Publishers, Burlington, 1999).