Indian Classical Dance Action Identification using Adaboost Multiclass Classifier on Multifeature Fusion

K.V.V.Kumar*, P.V.V.Kishore*, D.Anil Kumar*, E.Kiran Kumar*
* Biomechanics and Vision Computing Research Center, Department of ECE,
K.L. University, Green Fields, Vaddeswaram,
Guntur (DT), Andhra Pradesh, INDIA.
Kumarece405@gmail.com, pvvkishore@kluniversity.in, danilmurali@kluniversity.in, kiraneepuri@kluniversity.in

***Abstract*—Extracting and recognizing complex human movements from unconstraint online video sequence is an interesting task. In this paper a complicated problem from the class is approached using unconstraint video sequences belonging to Indian classical dance forms. A new segmentation model is developed using discrete wavelet transform and local binary pattern (LBP) features for segmentation. A 2D point cloud is created from the local human shape changes in subsequent video frames. The classifier is fed with 5 types of features calculated from Zernike moments, Hu moments, shape signature, LBP features and Haar features. We also explore multiple feature fusion models with early fusion during segmentation stage and late fusion after segmentation for improving the classification process. The extracted features input the Adaboost multi class classifier with labels from the corresponding song (tala). We test the classifier on online dance videos and on an Indian classical dance dataset prepared in our lab. The algorithms were tested for accuracy and correctness in identifying the dance postures.**

Keywords—Indian Classical Dance Identification, Adaboost Classifier, Multi Feature Fusion, Discrete Wavelet Transform (DWT), Local Binary Patterns (LBP).

I. INTRODUCTION (*HEADING 1*)

Automatic human action recognition is a complicated problem for computer vision scientists, which involves mining and categorizing spatial patterns of human poses in videos. Human action is defined as a temporal variation of human body in a video sequence, which can be any action such as dance, running, jumping or simply walking. Automation encompasses mining the video sequences with computer algorithms for identifying similarities between actions in the unknown query dataset with that of the known dataset. Last decade has seen a jump in online video creation and the need for algorithms that can search within the video sequence for a specific human pose or object of interest. The problem is to extract, identify a human pose and classify into labels based on trained human signature action models [1]. The objective of this work is to extract the signature of Indian classical dance poses from both online and offline videos given a specific dance pose sequence as input. Automatic dance motion extraction is complicated due to complex poses and actions performed at different speeds in sink to music or vocal sounds. Figure.1 shows a set of online and offline (lab captured) Indian classical dance videos for testing the proposed algorithm.

Fig.1. Online and Offline Dance datasets used in this work and the video constraints

In this work, human action recognition on Indian Classical Dance videos are performed on recordings from both offline (controlled recording) and online (Live Performances, YouTube) data. Indian classical dance forms are practised from 5000 years worldwide. However, it is difficult for a dance lover to fully hold the content of the performance as it is made up of hand poses, body poses, leg movements, hands with respect to face and torso and finally facial expressions. All these movements should synchronize in precision with both vocal song and the corresponding music for various instruments. Apart from these complications, the dancer wares complicated dresses with nice makeup and at times during performance the backgrounds are changing depending on the story which truly makes this an open-ended problem.

In this paper, we propose an multi class multi label Adaboost (MCMLA) based classification problem on multidimensional feature vector. We show that this can be used to match large unconstrained dance features which are automatically extracted from video datasets. The feature

representation of video objects depends on the efficiency of video segmentation algorithms. As illustrated in figure.2, the proposed Adaboost can effectively recover the query video frames from the dance dataset, by shape – texture observation model defied by discrete wavelet transform (DWT) and local binary patterns (LBP).

In summary, our MCMLA algorithm on online and offline Indian classical dance videos combines the representational flexibility and trivial computations. We perform experiments on two different datasets of Indian classical dance Bharatanatyam and Kuchipudi created from online downloads and offline controlled lab capture. The proposed method is compared with other GM models which are outperformed by a considerable margin in speed.

II. Proposed Methodology

The proposed algorithm framework is shown in figure.2. An Indian Classical Dance (ICD) video library is created combining online and offline videos. Dancer identification, dancer extraction, local shape feature extraction and classifier are the modules of the system. Further feature fusion concept from [2] is also explored in this wok using 5 feature types, Zernike moments, Hu moments, shape signature, LBP features and Haar features. Adaboost algorithm [3] [4] explores the relativity between the query dance sequence and known dataset.

A. Dancr Identification

The dance video sequence $V(x,y,t) \subset \mathrm{R}^+$, with $(x,y) \subset \mathrm{Z}^+$ gives pixel location and $t \subset \mathrm{Z}^+$ is the frame number. Each frame in V is having RGB planes and is of size $N \times M \times 3$. This part of the module is only for motion segmentation and object extraction; color can be discarded. RGB is converted to gray scale and contrast enhanced to improve the frame quality. The frame V^t at t is mean filtered with mask defined by $m(x,y)$ with

$$V_m^t(x,y) = V^t(x,y) \otimes m(x,y) \tag{1}$$

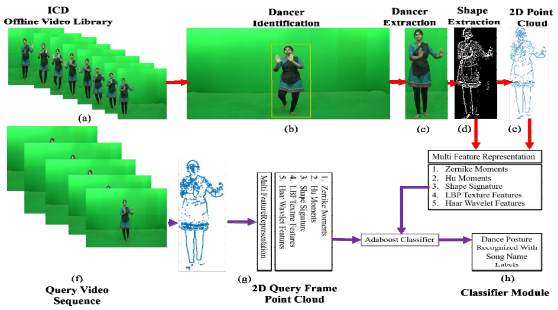

Fig.2. Flow Diagram of the proposed process for Indian Classical Dance Recognition. (a) Training Datasets, (b) Detected dance object, (c) Extracted, (d) DWT and LBP features (e) Feature points, (f) Query Dance Video, (g) Query Feature points, (h) Multi feature extraction and Classification.

The size of m is updated based on the frame size $N \times M$ for faster computations, where the object area is small compared to the background area. The $\otimes$ operator is linear convolution and the averaged frame is of same size as the input frame. The next step applies a Gaussian filter of μ mean and σ variance on the input frame V^t.

$$V_g^t(x,y) = V^t(x,y) \otimes g(\mu,\sigma) \tag{2}$$

The size of the Gaussian mask is determined by the input video frame. Euclidian distance metric $S^t(x,y)$ between V_m^t and V_g^t gives the saliency map of the moving pixels in the frame

$$S^t(x,y) = \left\| V_g^t(x,y) - V_m^t(x,y) \right\|_2 \tag{3}$$

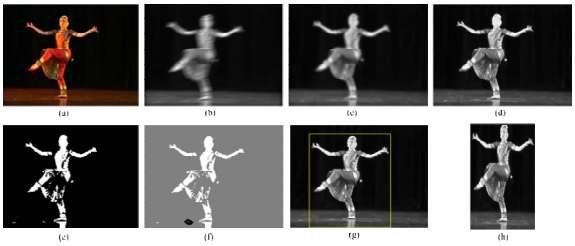

Fig.3. Dancer Extraction. (a) Original Frame, (b) Mean Filtered, (c) Gaussian Filtered, (d) Distance Saliency map, (e) silhouette Mask, (f) Connected components labelling, (g) Identified Dancer and (h) Dancer extracted

The second order normed distance map is shown in fig.3 which identifies the dancer's silhouette. However, to extract the dancer, a mask of this silhouette is used to determine the connected components in the object. Fig.3(d) shows the silhouette mask and connected component output is in fig.3(e).

The centroid of the mask is mapped on the frame to crop out the moving dancer in the frame. The method is effective in all lighting conditions putting constraints on the input video frame size in selecting the masks used for mean and Gaussian filters. The boxed and extracted dancer from the video sequence is shown in fig.3. (g) and (h) respectively. The extracted dancer is free from background variations in the video sequence. If a portion of background still appears at this stage can be nullified during the matching phase. Applying feature extraction on the extracted dancer allows for lesser computations as the background is almost eliminated and leads to good matching accuracy.

B. Feature Extraction

*1) Haar Wavelet Features -- Global Shape Descriptor:*For removing video frame noise during capture and to extract local shape information, we propose a hybrid algorithm with Discrete wavelet transform (DWT) [5] and Local Binary Patterns (LBP) [6]. The objective at this stage is to represent moving dancers shape with a set of wavelet coefficients. Here we propose to use Haar wavelet at level 1. At level 1, Haar wavelet decomposes the video frame V^t into 4 sub-bands.

Fig.4. shows the 4 sub-bands at 2 levels. At 1st level we have 4 sub-bands and at 2nd level have 8 sub-bands. In the 1st level, the three sub-bands represent the shape information at three different orientations: Vertical v, Horizontal h and Diagonal d. Combining the three sub-bands and averaging the wavelet coefficients normalizes the large values.

$$W_S^t = \frac{h+v+d}{3} \tag{4}$$

The averaged shape harr wavelet coefficients W_S^t, along with $\{h,v,d\}$ sub-band coefficients are reconstructed to spatial domain.

2) Thresholding: Apply threshold on the reconstructed ICD video frame V_r^t as

$$T^t = \sqrt{\frac{1}{NM}\sum_{j=1}^{M}\sum_{i=1}^{N}\left(V^t(j,i)\right)^2} \tag{5}$$

The binarized video frame B^t is

$$B^t = V_r^t > T^t \tag{6}$$

To extract the nodes for the graph, local pixel patterns provide exact shape representation.

3) Local Binary Patterns -- Local Shape Models: LBP compares each pixel in a pre-defined neighbourhood to summarize the local structure of the image. For an image pixel $B^t(x,y) \in \Re^+$, where (x,y) gives the pixel position in the intensity image. The neighbourhoods of a pixel can vary from 3 pixels with radius $r=1$ or a neighbourhood of 12 pixels with $r=2.5$. The value of pixels using LBP code for a centre pixel (x_c, y_c) is given by

$$L_S^t = LBP(x_c,y_c) = \sum_{j=1}^{P} B^t\left(g_p - g_c\right)2^P \tag{7}$$

$$B^t(x) = \begin{cases} 1 & \forall\, x \geq 0 \\ 0 & Otherwse \end{cases} \tag{8}$$

Where g_c is binary value of centre pixel at (x_c, y_c) and g_p is binary value around the neighbourhood of g_c. The value of P gives the number pixels in the neighbourhood of g_c. The local shape descriptor L_S^t of the human dancers pose projects maximum number of points on to graph.

C. *Multi Features -- Zernike, Hu Moments, Shape Signature, LBP, Haar*

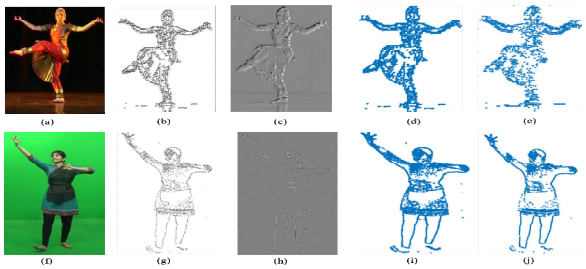

Fig.4. (a) Original dancer in Online video frame (b) LBP features from Reconstructed and Thresholded Wavelet Coefficients (c) Haar Wavelet Features, (d) Sparse Coded Wavelet reconstructed LBP features and (e) Sparse coded wavelet features. (f) – (j) same as (a) – (e) for offline lab captured video frame.

Fig.4. shows the extracted dancer represented with LBP features and Haar wavelet features. Local shape features in fig.4(a) are used to construct a graph. Given a motion frame in a ICD video sequence V^t and successfully extracted local shape features L_S^t and transformed into a binary shape matrix B_S^t of ones and zeros using eq'n 5. A sparse representation of B_S^t eliminates all zeros and retains only ones and their locations in $M_S^t(x,y,w)$, where x,y are shape point locations and w is shape feature weight vector. Fig.4(d) and 4(e) shows a sparse representation for both wavelet reconstructed LBP (WR_LBP) and only Harr wavelet features (HWF) respectively. The points on the motion object are formed by extracting the location of the pixel and its feature value determines the shape of the dance pose. From these feature point locations and values a graph is constructed in this work.

Dancer in Indian classical dance videos have a large motion vector field and Haar and LBP depend on variations in lighting, camera movement and background. To counter balance camera movements, we propose to use Zernike Moments (ZM) [7] [8] to represent dancer in each frame on the 2D point cloud extracted from WR_LBP vectors.

Hence, most of the applications use ZM magnitude as a feature vector for pattern classification. In this work, we propose to ZM magnitude on 2D shape point cloud as invariant feature representing the small variations in camera movement that occur in online video capture of the dance performance. To represent changes in the dancer dimensions which happen non-linearly, we propose to use a nonlinear function defined over geometric moments.

Hu moments in [9] are non-orthogonal centralized moments that are scale, translation and rotation invariant. Human dancers come in all shapes and sizes and the features describing them may change with dancer.

III. Results and Discussion

This session of the paper initiated to test the robustness of the proposed multi feature fusion with Adaboost classifier. Our Indian classical dance datasets consist of performances on `Bharatanatyam' and `Kuchipudi' from online YouTube videos and offline dance videos in controlled environment at K.L. University, cams department studio. We have created 4 dance videos from 5 dancers for 2 songs in two different dance styles. Similar online YouTube downloaded dance performances are also collected.

We use offline and online dataset of same and different dancer video for training and testing with early fusion and late fusion of multiple features 28 words in the dance sequence. Each mudra pose is coordinated with vocal manually by labelling each set. Variations in number of frames per label is nullified and normalized to 15 Key frames per dance pose across all video data.

The input videos from dance data set captured in the controlled environment. The dancer identification, feature extraction and graph representation for the dancer is shown in fig.5. Saliency maps from average and Gaussian distance metric creates Silhouete, which identifies the dancer in the video frame. The dimensions of the bounding box extract the dancer. Then the dancers features are extracted with segmentation as early features. Haar wavelet at level-1 is averaged in high frequency components to remove background and IDWT is performed to recover the global shapes on the dancer. Applying LBP on the resulting IDWT dance frame captures local shape information. At this stage, we perform Adaboost multi class multi label classification with the PCA based Haar and LBP feature fusion.

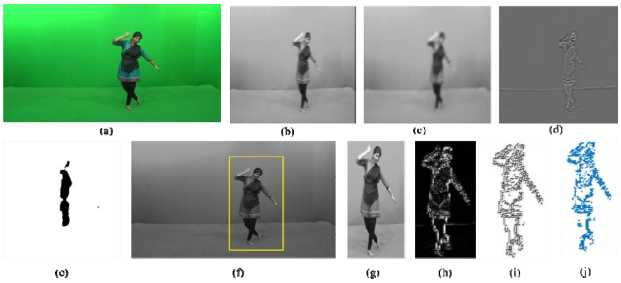

Fig.5. (a) Offline Frame shot from ICD video, (b) Gaussian smoothing, (c) Averaging, (d) Saliency map, (e) Silhouete creation, (f) Dancer Identification, (g) Extracted Dancer, (h) Wavelet reconstructed features, (i) LBP Features from Wavelet and (j) Constructed 2D Point Shape Cloud.

Early fused Haar -- LBP features of offline dance video of a dancer is used to train the Adaboost classifier. For the same dance video shot with slight variations is provided as query dance video for same set of labels. The resulting confusion matrix from early fusion of features on same training and testing set is shown in fig.6.

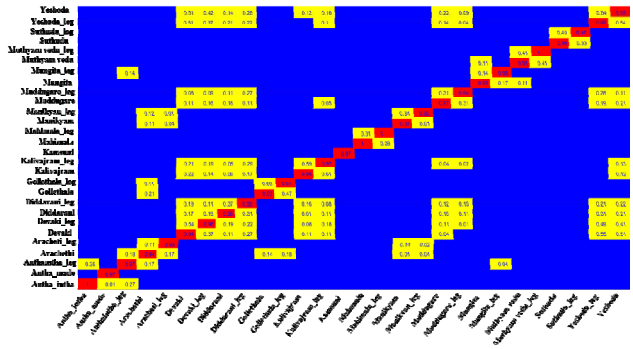

Fig.6. Early Fusion Confusion matrix with same offline dancer in training and test video

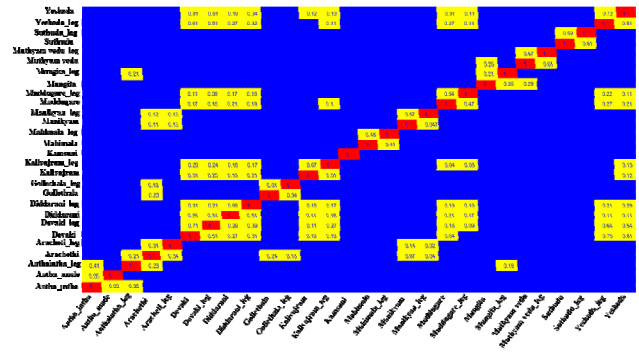

Fig.7. Late Fusion Confusion matrix with same offline dancer in training and test video

The results are repeated with all the same parameters with late fusion with Zernike moments, Hu Moments and liner shape signature along with Haar and LBP features. Late feature matrix is a 73×25 matrix per label. For our 116-label dance sequence we have a $73\times25\times116$ dimension feature vector. Fig.7 shows the results of the classifier in the form of confusion matrix.

Clearly, multiple features and late fusion has increased the ability of the classifier in recognizing dance poses correctly. False matching is very less in this experiment as the dataset used for training and testing. The results show an average of 0.99 for all dance videos in the dataset.

IV. Conclusion

Two fusion models are proposed for feature fusion. Early fusion at the segmentation stage with PCA based Haar wavelet and LBP is used and late fusion using the Zernike moments, Hu moments, Shape signatures are used with Haar and LBP are proposed. Multi class multi label Adaboost on features of early fusion and late fusion between two sets of dance video data is the classifier. Multiple experimentations on online and offline ICD video data is tested. Dance video data is labelled as per the vocal song sequence. The multi early and late features and classifiers performance tests show that the proposed late fusion features and multiclass multi label Adaboost classifier gives better classification accuracy and seed compared to AGM and SVM. More action features can be added for representing dancer more realistically by elimination backgrounds and blurring artefacts to improve the efficiency of the classifier.

References

[1] Ronald Poppe. A survey on vision-based human action recognition. Image and vision computing, 28(6):976–990, 2010.

[2] Chirag I Patel, Sanjay Garg, Tanish Zaveri, Asim Banerjee, and Ripal Patel. Human action recognition using fusion of features for unconstrained video sequences. Computers & Electrical Engineering, 2016.

[3] Kumar, K. V. V., P. V. V. Kishore, and D. Anil Kumar. "Indian Classical Dance Classification with Adaboost Multiclass Classifier on Multifeature Fusion." Mathematical Problems in Engineering 2017 (2017).

[4] KUMAR, KVV, et al. "COMPUTER VISION BASED DANCE POSTURE EXTRACTION USING SLIC." Journal of Theoretical & Applied Information Technology 95.9 (2017).

[5] PVV Kishore, ASCS Sastry, and Zia Ur Rahman. Double technique for improving ultrasound medical images. Journal of Medical Imaging and Health Informatics, 6(3):667–675, 2016.

[6] Syed Inthiyaz, BTP Madhav, and PVV Kishore. Flower segmentation with level sets evolution controlled by colour, texture and shape features. Cogent Engineering, 4(1):1323572, 2017.

[7] Dengsheng Zhang and Guojun Lu. Content-based shape retrieval using different shape descriptors: A comparative study. In null, page 289. IEEE, 2001.

[8] Manish Khare, Rajneesh Kumar Srivastava, and Ashish Khare. Object tracking using combination of daubechies complex wavelet transform and zernike moment. Multimedia Tools and Applications, 76(1):1247–1290, 2017.

[9] Ming-Kuei Hu. Visual pattern recognition by moment invariants. IRE transactions on information theory, 8(2):179–187, 1962.