# Automatic annotation of traditional dance data using motion features

Huma Chaudhry
Faculty of Computing
University Technology Malaysia
Skudai, Johor
chuma2@live.utm.my

Karim Tabia
CRIL UMR CNRS 8188
Artois University,
Lens, France
tabia@cril.fr

Shafry Abdul Rahim
Faculty of Computing
University Technology Malaysia
Skudai, Johor,
shafry@utm.my

Salem BenFerhat
CRIL UMR CNRS 8188
Artois University,
Lens, France
benferhat@cril.fr

***Abstract*— This study investigates the accuracy of human dance motion capture and classification from selected Malaysian dances for Malaysian Dance Annotation (MDA). Dance motion classification is a new scope to motion classification. Recent studies focus on basic movements classification where motion is not very complex, such as walking or waving hand. In this paper, an attempt has been made to classify complex motion such as a dance. This work is motivated by the need to develop an automated tool to annotate dance videos to build traditional dance video management and retrieval system. We evaluate our system on a dataset of different types of Malaysian dances, collected from Youtube. Despite the complex movements in dance, the proposed solution requires less human input effort and gets a suitable accuracy for complex dance motion annotation.**

***Keywords*— Motion; Dance; Malaysian; Annotation; Classification***

## I. Introduction

Motion in video carries important information, which is of a multi-fold nature. Motion sometimes needs to be interpreted for understanding or anticipating any immediate reaction to the motion. Hence, the significance of motion perception in a video is encountered in several of the present systems and it remains to be an active research area. Movements in dances encompass different types of information; sacred rituals, social dialogue or cultural expressions. This calls for the need to completely and precisely capture motion details and make it digitally available for further processing. Several applications can result in dance video automatizing, such as dance video retrieval, classification of video databases or animation of dances using modeling of stored information, to name a few.

This work is done in the framework of an EU research project dealing mainly with traditional Asian dances data. This study focuses on investigating the complex human movement through analyzing the different steps and stages in a dance video. The target videos for this research are real traditional Malaysian dance videos acquired from various online resources, such as Youtube. The methodology of this research includes motion features extraction that can best represent a complex human motion and then annotate each motion of target based on sequence of motion observed over a period of time. The motivation behind this work is the need to develop an automated tool to annotate dance videos to build a dance video management and retrieval system.

This paper is divided into five sections. The first section mainly introduces the subject of human dance motion annotation and classification, the significance and motivation of this research problem. The second section provides a general literature overview of human motion and annotation systems as well as their aim and if they can be extended to a complex movement such as in dances. Section 3 discusses the methodology. Finally, in Section 4, the results are discussed and analyzed, followed by Conclusion and future works in Section 5.

## II. Literature Review

Human pose tracking significantly differs from the traditional object tracking as it combines the information of human structure and local parts. This results in the need for tracking the local parts, as well as maintaining the global structure in terms of connectivity. An early work in this scope goes back to 1996 [1], but due to the complexity of the problem, human pose estimation and tracking remains an insufficiently explored field. Using field of view or viewpoint restrictions, some works have successfully detected the body parts, especially when the motion is not a complex one, such a walking, or running [2-5]. However, still image pose estimation can be extended and to study inter-frame dependencies to perform tracking, such as by incorporating optical flow information [6], [7]. Ramakroshna et al. [7] modeled human body as a combination of singleton (e.g., head, neck) and symmetric pair of parts (e.g, arms and legs) such that pose tracking becomes a multi-target (parts) tracking problem where targets are related as a global structure. This approach incurs high computational cost. Video and shot annotation methods have also been explored using non-visual information such as movie scripts [8]. To obtain a sufficient number of action samples from movies for visual training manually is a hard task. Laptev et al. [8] avoided this difficulty of manual annotation by using movie scripts and extracting text description of the movie content in terms of scenes for annotation.

Activity classification of a video can be categorized from various points of view, the two most important being sparse/dense methods and supervised/unsupervised methods. Sparse activity analysis techniques focus on human body part motions. The frames are analyzed based on patches of interest using the spatial and temporal information; spatio-temporal volume, hence focusing on regions of interest. These methods can be categorized as local approaches for activity recognition. The state of the art detectors for space-time interest points are sparse spatio-temporal features [9] and STIP features [2]. On the other hand, dense methods take the data as a whole and attempt designing classification methods for a large number of features. Global approaches use global features using optical flow method to represent the state of motion in a video at a time instant. Ground truth annotations are required to train data using supervised techniques. Unsupervised methods focus on action discovery without any directed input.

For the application in hand, as the types of the different dance segments and their labels are known, therefore to annotate a dance video into different dance segments, supervised machine learning techniques are used.

## III. METHODOLOGY

In this section, we introduce the proposed framework for human dance classification. Fig. 1 presents the overall picture and building blocks of the proposed framework. Dance videos have been acquired from Internet repositories, such as Youtube. The dance category is native Malaysian dances.

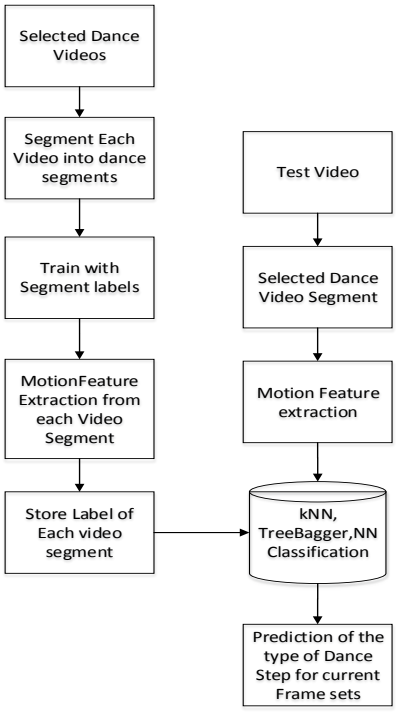



Fig. 1. Framework for the proposed dance classification system

### A. System Overview

Figure 1 shows the system overview. The system acquires videos in the database. Each video is manually divided into segments and each segment is labelled. Then, the Space Time Interest Point (STIP) and Histogram of Optical Flow (HOOF) based features are employed to extract motion information from each dance segment. For classification and annotation, we use a simpke *k*-Nearest Neighbour classifier to get the classification results. Also, Neural Network based classification and TreeBagger classifier have been used to compare better annotation performance.

### B. Motion Feature Extraction

For concise and accurate information extraction from a complex motion of a dance, we extract features at local and global levels. Since the assumption that motion is only introduced by human dance performer holds, therefore, a better analysis from both levels can be made.

*STIP (Space Time Interest Points):*

Local space-time features capture local events in video and can be adapted to the size, the frequency and the velocity of moving patterns. These features are local level spatiotemporal features and can be used for recognizing complex motion patterns. Hence, these features are chosen to represent the information in a complex motion, such as a human dance.

To detect spatio-temporal events, [2] used the idea of the Harris and Forstner interest point operators. They identify local structures in space-time where the image values have significant local variations in space and time domain. Using 3D Gaussian kernels convolution for maximizing a normalized spatio-temporal Laplacian operator over spatial and temporal scales, they estimate the spatio-temporal extents of the detected events. The spatio-temporal separable Gaussian kernel is defined as

$$g(x, y, t; \sigma_l^2, \tau_l^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_l^4 \tau_l^2}} \times \exp(-(x^2 + y^2)/(2\sigma^2) - (t^2/2\tau_l^2)) \quad (1)$$

This is essentially different from spatial Gaussian in keeping a separate scale parameter for the temporal domain as the spatial and the temporal extents of events are in general independent. The spatio temporal second moment matrix illustrates how the spatial and temporal derivatives are averaged using the spatio temporal Gaussian Kernel

$$\mu = g(:,:;\sigma_l^2, \tau_l^2) \times \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix} \quad (2)$$

where $L_x$ is the differential in x axis.

*Optical Flow:*

In our work, we assume that the magnitude of optical flow is high around the moving person. Thus, optical flow can prove to be a useful method for detecting a global level motion information using motion vectors. An optical flow in a pixel is represented by vector $\begin{bmatrix} u \\ v \end{bmatrix}$, where *u* and *v* are the horizontal and vertical flows or displacements, where optical flow, for a window of 3x3 pixels is measured as

$$\begin{pmatrix} f_{x1} & f_{y1} \\ \vdots & \vdots \\ f_{x9} & f_{y9} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = - \begin{pmatrix} f_{t1} \\ f_{t9} \end{pmatrix} \qquad (3)$$

Which represents the equation of optical flow of each pixel.

To enable invariance to scale and direction of motion, [10] proposed Histogram of Optical Flow features (HOOF). Normalization of Optical Flow at each frame makes the histogram representation scale-invariant. HOOF features $h_t = [h_{t;1}, h_{t;2}, \dots, h_{t;B}]$ are defined at each frame of the video. Since each dance segment is labelled individually, therefore, for classification stage, it is necessary to compare testing and training HOOF features.

*Classification:*

*k*NN classification combined with motion descriptors in terms of local features (STIP) and global feature histograms (HOOF) analyses a field of less explored techniques for motion recognition and annotation. K Nearest Neighbour classifiers (*k*-NNs) are state-of-the-art nonlinear multi-class classifiers which are commonly used for multiple classes problems. For more details and similar works, we refer the reader to [11-13]. Let $X = \{ x_1, x_2, x_3, \dots, x_s \}$ be the feature vectors corresponding to the dance segment S, extracted from a given sequence of a dance video. To classify this sequence, we project each of these vectors $x_i$ to a d-dimensional vector in Eigengait space, and determine its class. k-Means clustering with k=5 was applied to the sequence for populating X into clusters $C_{sn}$ based on the k-nearest neighbour rule, each denoted $c_i$ and sn is the number of different dance segments, for this case sn=20. The k-nearest neighbour classifier can be viewed as assigning a weight $1/k$ for k nearest neighbours and all others a weight of 0. This can be generalised to weight nearest neighbour classifiers.

TreeBagger {Breiman, 1996 #421}, a tree based classifier, is also used for classification. TreeBager algorithm is known to have a better stabile and accurate decision tree. The bagged decision trees or TreeBagger technique creates an ensemble of decision trees where each tree (learner) is grown at an independent bootstrap replica of randomly selected N observations of input data. Also, every tree in the ensemble can randomly select predictors (data values) for decision splits to increase the accuracy of bagged trees. 150 trees are aggregated to create the classifier.

Neural Networks (NN) have been used for classification in various applications. NN are less sensitive to noise and can handle multi-class classification problems. A basic NN has *n-1* inputs with 1 bias, the feed, and 1 output node. The total input stimuli to a neuron in the output layer is

$$z_{in} = \sum_{i=0}^{n} x_i w_i = x_0 w_0 + x_1 w_1 + .. + x_n w_n$$

$Z_{in}$ is the output to the activation function $f(Zin) = 1/(1 + e^{-\sigma x})$. The difference between this value and the correct value is the error function $E = \frac{1}{2}\sum_{k=1}^{m}(t_k - y_k)^2$. In this work, NN has 30 input and 20 output nodes, connected by 10 hidden layers. The input and output nodes' number represent the length of the features and the binary vector length of classes each feature can belong to.

## IV. Results and Discussion

This dataset contains two types of traditional Malaysian dances. These videos contain 20 short videos of different dance segments in each performance. The dances are performed by different persons and the number of subjects in each video is 1. The classes of dance segments include labels such as InangSideBend, InangForward, HandBirdCen, etc. These labels are provided manually. STIP features are extracted from the set of frames in dance segment. Optical flow is computed on each of the dance segments and HOOF features are extracted from each frame and further processed to represent each dance segment of each video sequence. For testing, the periodic segments for each dance segment are clipped to be used as testing video. For a more robust evaluation, a 10-fold cross-validation was used for all experiments. The data has been divided for 70% training and 15% testing data, while 15% is used for validation test.

### *A. Results*

We compare results of combining two different representations of dance motions with *k*-NN classifiers. The representations are local features described by spatio-temporal interest points, STIP, and 30-bin histograms of global optical flow based features, HOOF. Some of the dance segments in two different dances can share similar motion patterns. An example of such segments is tabulated in Table I. Hence for such dance segments, a lower recall value is expected.

TABLE I. Similar Motion Segments in Dances

| **Dance Name** | **Dance Segments with Similar Motion** | | |
|---|---|---|---|
| | ***Segment Name*** | ***Dance Name*** | ***Similar Segment*** |
| Cengung | Class13 | Inang | Class8 |
| | Class14 | Inang | Class2 |
| Inang | Class1 | Cengung | Class16 |
| | Class5 | Cengung | Class17 |

The accuracy with k-NN classifier and NN classifier is ilustrated in Table II. While kNN and TreeBagger show a good result, NN technique for classification does not produce

good results. A larger dataset can improve the overall accuracy in this case.

TABLE II. RECALL RATE FOR DACE SEGMENTS USING COMBINED FEATURES

| Classifier | Parameters | Accuracy (%) |
|---|---|---|
| *kNN classifier* | K=3 | 70.3 |
| NN classifier | N-fold Test, n=5 | 42.1 |
| TreeBagger | N=150 trees | 92.6 |

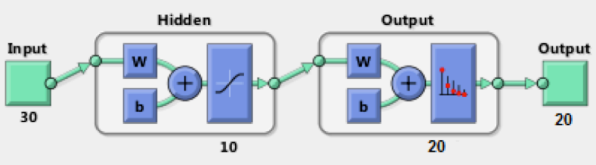
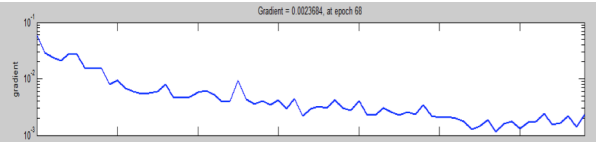
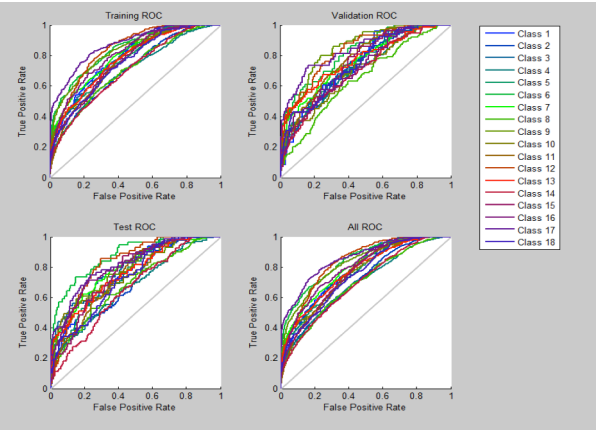



Fig. 2. Neural Network setup for dance segment classification.

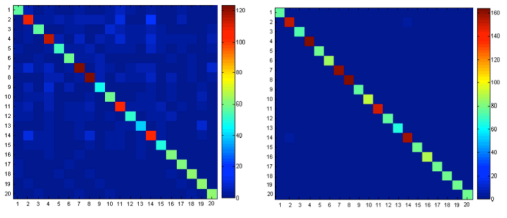



(a)




(b)

Fig. 3. (a) Illustration of Gradient decent during the training State of dance segment with NN classfier (b) Training for 20 of the dance segments and their ROC



(a) (b)

Fig. 4. Confusion Matrix for Dance Segment classification using (a) *k*-NN classifier (b) TreeBagger classifier

Fig 2 shows the network setup for dance segment annotation. Input value of 30 shows the length of each feature vector and the output layer shows the number classes of dance segments that the result can belong to. The training state convergence can be seen in Fig 3a with the gradient descent. Another illustration of how well the neural network has fit data is the receiver operating characteristic plot shown in Fig 3b.

## *B. Discussion*

The ROC curve of class 7 in Fig 3 shows the best characteristics. This is because the dance segment was repeated several times during the dance video and therefore, several training clips could be extracted from the main dance video. Confusion matrices in Fig. 4 and overall results demonstrate that tree based classifier outperforms the other classifiers as it tends to achieve high true positive while minimizing the false positive. Class 17, for instance, has low accuracy because this dance segment had complex hand movement white sitting and turning in circle. This increased occlusion and has not been classified well. This result can be improved using better motion tracking features. The detection however has a better accuracy with TreeBagger classifier. Also, it can be seen in Fig. 4 that TreeBagger can reduce classification noise. Confusion for matches found for similar actions (Class2, Class14) has lower recall since the moments share mostly similar motions. The TreeBagger classifier proves to be the most robust recognition algorithm, producing the best performance with an output annotation rate remaining very close to 90%, for different combination of features.

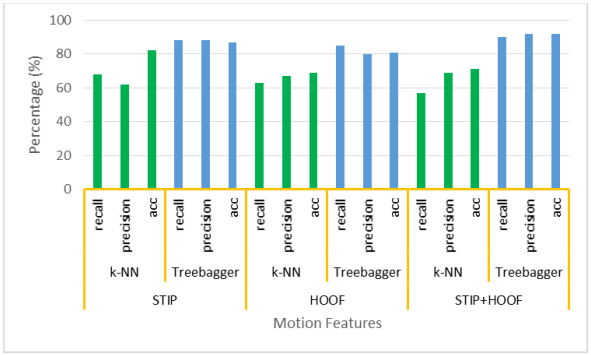



Fig. 5. Comparison of different features and classifier combinations for dance segment annotation

Fig 5 shows the result of different features and their combination with different classifier. Due to the local nature of features, STIP can better distinguish test video features based on local information. Taking the spatial and the temporal consistency of local and global features, STIP and HOOF, show an improvement in precision of dance segment annotation.

Global motion of subjects in the database is a strong cue for discriminating between, for instance, the leg and the arm actions when using histograms of optical flow. This information, however, is intermingled when representing the

actions of dance segments that cover similar spatial region. Similarly, spatio-temporal features can be further combined with skeletal trajectories of relative position across the frames to give better recognition and tracking. In this work, when classifying segments, the information of relation among the dance patterns is lost since we assume that dance segments are independent. This is not generally true for a dance video.

Moreover, for annotating a long dance video which is composed of different dance steps, temporal segmentation of dance videos can assist in automating a full system for video annotation. This information can be extracted from temporal pauses or dance segments returning to basic position of dance, which is applicable to Malaysian dances.

## V. Conclusion

Computer recognition of human activities is an important area of research in computer vision with applications in many diverse fields. Dance annotation has applications in many fields such as dance-based video retrieval system, video management, games and graphics, as well as automatic annotation. In this paper, we proposed an algorithm using basic video processing techniques to annotate different dance video segments. Using different classifiers, kNN and TreeBagger have shown more promising results. The obtained preliminary results can be improved with a larger dataset and more dynamic methods such as tracking.

## Acknowledgment

This research is funded by European Union under project "Marie Skłodowska-Curie Actions RISE (Research and Innovation Staff Exchange)" project AniAge, H2020-MSCA-RISE-2015.

## References

[1] Ju, S.X., M.J. Black, and Y. Yacoob. *Cardboard people: A parameterized model of articulated image motion*. in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*. 1996. IEEE.

[2] Laptev, I., *On space-time interest points.* International Journal of Computer Vision, 2005. **64**(2-3): p. 107-123.

[3] Cherian, A., et al. *Mixing body-part sequences for human pose estimation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

[4] Kiefel, M. and P.V. Gehler. *Human pose estimation with fields of parts*. in *European Conference on Computer Vision*. 2014. Springer.

[5] Schuldt, C., I. Laptev, and B. Caputo. *Recognizing human actions: a local SVM approach*. in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. 2004. IEEE.

[6] Yang, Y. and D. Ramanan, *Articulated human detection with flexible mixtures of parts.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013. **35**(12): p. 2878-2890.

[7] Ramakrishna, V., T. Kanade, and Y. Sheikh. *Tracking human pose by tracking symmetric parts*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013.

[8] Laptev, I., et al. *Learning realistic human actions from movies*. in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. 2008. IEEE.

[9] Dollár, P., et al. *Behavior recognition via sparse spatio-temporal features*. in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. 2005. IEEE.

[10] Chaudhry, R., et al. *Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions*. in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. 2009. IEEE.

[11] BenAbdelkader, C., R. Cutler, and L. Davis. *Motion-based recognition of people in eigengait space*. in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. 2002. IEEE.

[12] Qi, Y., A. Hauptmann, and T. Liu. *Supervised classification for video shot segmentation*. in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*. 2003. IEEE.

[13] Altman, N.S., *An introduction to kernel and nearest-neighbor nonparametric regression.* The American Statistician, 1992. **46**(3): p. 175-185.