An Automatic Extraction Tool for Ethnic Vietnamese Thai Dances Concepts

Truong-Thanh Ma[2], Salem Benferhat[1], Zied Bouraoui[1], Karim Tabia[1], Thanh-Nghi Do[2], Nguyen-Khang Pham[2]
[1]CRIL CNRS & Artois University, Lens, France
Email: {benferhat, bouraoui, tabia}@cril.fr
[2]CICT, Can Tho University, Can Tho city, Vietnam
Email: truongthanh1511@gmail.com, {dtnghi, pnkhang}@cit.ctu.edu.vn

***Abstract*—In recent year, preservation and promotion of the ICHs are one of the problems of interest. In this paper, we focus on modelling the traditional dance domain, particularly modelling traditional Vietnamese dances. To conserve significant characteristics of dances, we proposed an ontology to represent the significant movements features of Ethnic Vietnamese Thai Dances (EVTDs). Particularly, a detailed description of the movement schemas of EVTDs is presented in this paper. Additionally, we present how to build an automatic extraction tool to collect the fundamental movements data of EVTDs using machine learning. Finally, we represented explicitly how to store those extracted features from raw dance videos into prioritized Ontology-based proposed.**

Keywords: *Vietnamese Traditional Dance,machine learning, Ethnic Vietnamese Thai Dance*.

I. Introduction

Vietnam is a multi-ethnic country existing many different cultures [4] with fifty-four-ethnic groups living in a territory. The traditional dances had became "spiritual foods" of each Vietnamese people, it influences directly the real life from urban to rural. Most of the traditional Vietnamese dance (TVD) is transfered by "word of mouth" [1], the present generation would instruct fundamental movements to the adjacent generation.

Almost Vietnamese traditional dances built up from the ethnic groups culture, life environments and regions, it contains the large number of the significant characteristics of region-zone. Particularly, the dance movements of the ethnic groups originated from the life activities, each posture is depicted an characteristic action of their life. Therefore, the fundamental movements would be one of the stable foundation as well as being the essential features to determine the different dances. In this paper, we selected a remarkable dance of Thai community in Vietnam to illustrate how to represent the important features. The movements of EVTDs are the combination of fundamental postures presented in [3].

One of the main contributions of this paper is to build an automatic extraction tool using machine learning in order to store the significant features of ethnic Vietnamese Thai dance movements (EVTDMs). These characteristics are put into a lightweight prioritized ontology-based (LPO) attached to recognized probabilistic of each body part in the same frame. Based on the probabilistic of each posture to decide whether the tool put those features into LPO or not.

In the research process, we decomposed our approach into two main stages: the first aspect is to reconstruct a schema of panorama overview of traditional Vietnamese dances; the second aspect concerns a character with the fundamental movements of each EVTD. In this paper, we concentrate primarily on the second stages with respect to detecting, extracting and storing automatically the fundamental movements. Our primary challenge is to determine the principle concepts from EVTD's movements combined with a set of desirable properties to build a useful dance search engine, moreover, because the movements of the performers is quite uncertain and inconsistent regarding the different cases (amateur, speed of music, etc.) as well as the sequential frame of videos is not explicit to recognize those extract movements, our remarkable aim is how to handle the uncertainties and inconsistencies in processing.

The remainder of this paper is structured as follows. In the next section (section 2) we give recent related works. Section 3 provides a description of EVTD's features and how to encode them into Ontology. How to detect automatically concepts and a methodology to build a LPO are discussed in section 4. Finally, section 5 concludes the paper.

II. Ethnic Vietnamese Traditional Dances

A. Ethnic Vietnamese Thai Dances

Thai community in Vietnam is one of the ethnic groups existing the large number of the traditional dances. There are many significant festivals of Thai ethnic group to be held in villages as well as regions during the whole year. In order to understand explicitly with respect to EVTDs, in this section, we present several fundamental movement features to identify the EVTDs in the Vietnam's territory.

In each EVTD, the remarkable characteristics to determine EVTDs is the fundamental movements, in which is the foundation of the creative combination in each motion in order to create the specific dances of Vietnamese Thai people. Correspondingly, the detection of the basis movements is one of the important steps collected automatically the dance dataset for LPO. The following we would present a basis movements schema of EVTDs as well as represent EVTD's LPO based

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00252

on the schema proposed. Basis movements of EVTDs are divided in five characteristics [3]: Orientation, Arm Posture, Leg Posture, Sitting Posture, Standing Posture.

1) Orientation: Regarding orientation features, it is one of the most significant characteristics because the motions, postures and gestures of Vietnamese traditional dances are always described explicitly through the orientations in almost all documents. They are split in eight orientations as Figure 1, including from orientation 1 to orientation 8. In [3], orientation 1 is the direction of the dancer opposite to spectator (in front of audience), it is also utilized for the first preparation step of performing

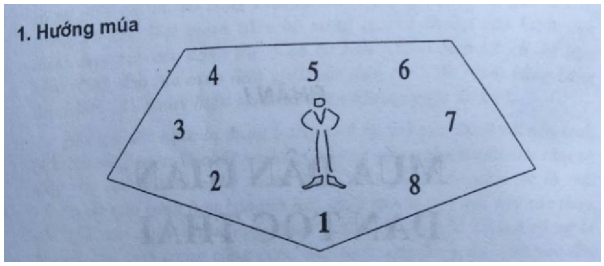

Figure 1: Orientations of EVTDs

2) Arm Posture: Most of arm postures is concentrated on depicting life activities in Thai community, therefore the basis postures is simple and habitual. It is divided in five primary postures: VN-Thai-Thế-[i]-Arm (i=1..5). They are grouped into two distinct clusters: open-arm posture and close-arm posture.

3) Leg Posture: There are five significant leg postures to represent for EVTD movements consisting of VN-Thai-Thế-[j]-Leg (j=1..5).

4) Sitting and standing Posture: Sitting posture is divided in two postures, it consists of VN-Thai-Thế-1-Sitting, VN-Thai-Thế-2-Sitting. There are three standing posture in EVTD VN-Thai-Thế-5-Standing, VN-Thai-Thế-2-Standing, VN-Thai-Thế-4-Standing.

III. On the Automatic Detection of Concepts

A. Human Pose Estimation

We utilize TF-Openpose (written in python using Tensorflow library instead of Caffe library) for estimating the positions of human joints and articulated pose estimation in order to support for depicting each movements in EVTD. Moreover, we ameliorated and improved TF-Openpose through algorithms of input image processing and modified several essential arguments of CNNs.

The primary purpose of using HPE for EVTD movements is to determine concretely parts of body in raw dance videos to aim at extracted and represent the motions in each dance movement. The architecture simultaneously predicts detection confidence maps and affinity fields that encode part-to-part association as in Figure2. The network is split into two branches: Branch 1 is responsible for predicting confidence maps, and Branch 2 is to predict the affinity fields. TF-Openpose takes a 2D color image as input and produces the 2D location of anatomical key-points for each person. The (x,y) coordinates of the final pose data array could be normalized to the range depending on the key-point scale. It can be estimated 18 key-point body pose from COCO 2016 dataset.

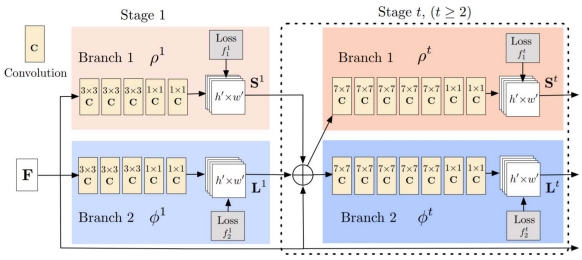

Figure 2: The architecture using CNN in Openpose

Realizing the requirements of a high configuration regarding GPU for Openpose handled, we proposed to utilize TF-Openpose [1] instead of orginal Openpose version. It is a human pose estimation library developed based upon the foundation of the Openpose library using Tensorflow and OpenCV. It also provides several variants that have made the changes to the network structure for real-time processing on the CPU or low-power embedded devices. We concentrated on two variations of models to find optimized network architecture: CMU [6] and Mobile-Net [9]. (1) With respect to CMU, it is the model based VGG pre-trained network which described in the Openpose's original paper using COCO dataset for training, it is converted from Caffe format to utilize in Tensorflow; (2) Based on the Mobile-Net paper [9], with 12 convolutional layers are used as feature-extraction layers. The experimental result as in figure 3.

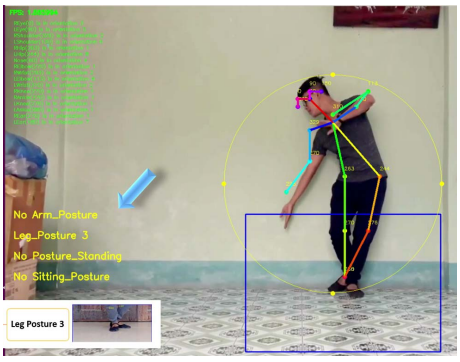

Figure 3: Using HPE and DCNN to detect the basis features

B. Using Deep Convolutional Neural Networks

In order to detect and recognize the significant postures (particularly in Leg, Standing and Sitting Posture), we proposed

[1] https://github.com/ildoonet/tf-pose-estimation

to utilize deep convolutional neural networks (DCNNs) which have been applied to visual tasks since the late 1980s.

As we had known, there are three main types of layers used to build Deep CNN architectures: convolutional layer, pooling layer and fully connected layer. Most of the CNN architectures is obtained by stacking the number of these layers. Deep convolutional neural networks, trained on large datasets, achieve convincing results and are currently the state-of-the-art approach for this task illustrated in figure4.

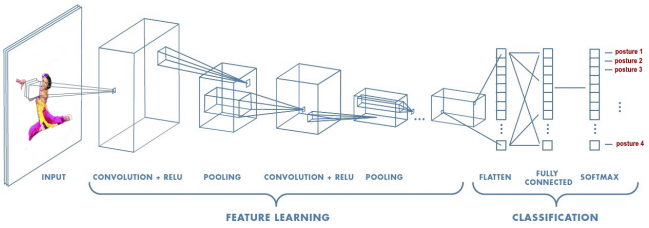

Figure 4: Neural network with many convolutional layers

Because the limited number of the postures features to represent EVTDs, we selected DCNNs to detect automatically those postures on the image frames of video. We utilized an open source neural network library written in Python called Keras [2] in which integrated many architectures being compatible with all the backends (TensorFlow, Theano, and CNTK).

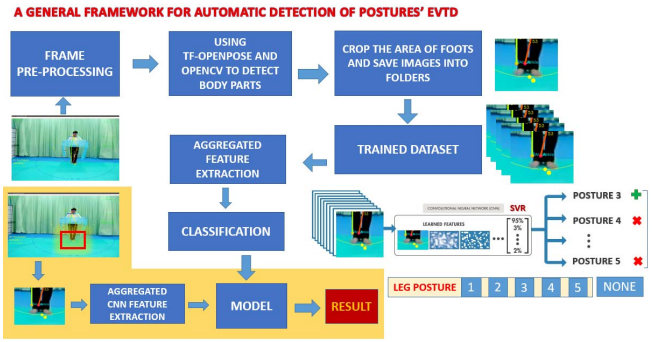

Figure 5: Framework to detect automatically dance postures

In this research, we implemented a general framework for automatic detection and classification of EVTD's fundamental postures as in figure 5. We introduce a classification model aggregating consequently the deep CNN architectures to extract the features (including Xceptions model, Inceptionv3 model and MobileNet model as in Figure 6). Realizing that each CNN architecture deals with the different cases of the particular dataset as well as combining the features to represent a vector is quite expected because it will be boosted strongly the discrimination between the classifications. In addition, the

[2]https://keras.io/models/model/

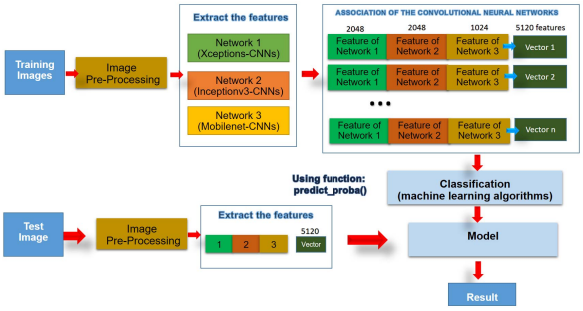

Figure 6: The aggregation of three DCNNs architectures to extract features

dataset of ETVDs focus principally on the high resolution image frames of dance videos, specially, the resolution and size of images will be grown rapidly in the future. For these reasons, we selected the methodology using deep CNNs to extract the significant features and ML algorithms to classify. After having sets of the collected images, we used several algorithms to advance the quality of images (image pre-processing). In each image frame, we extracted the features from three CNN architectures, particularlly, including Xceptions [7] (2048 features), InceptionV3 [8] (2048 features), Mobilenet [9] (1024 features). The next step, we aggregated consequently (respectively) the extracted features to have a feature vector with 5120 dimensions. After that, the comparison among several ML algorithms to have a best selection for classification is fully essential presented in section V.

The main idea of this model is a classification tool to update flexibly the novel architectures which will be published in the future aimed at advancing the accuracy in classification as soon as the dataset accelerated. In addition, we are also able to extract the features in parallel with each CNN architecture, nevertheless, we do not experiment the parallel models in this paper.

C. An automatic detection tool

In order to store and to detect the fundamental features into LPO, we implemented an automatic extraction tool of EVTD's movements (poses) by python language illustrated in figure **??**. We utilized DCNN's architecture models proposed to classify the postures (leg, sitting and standing postures). Additionally, we also used HPE to describe each body parts of performer/dancer. Furthermore, we selected Owlready2 library (using python language) with HermiT reasoner to build our lightweight prioritized ontology. Our tool allowing users is able to extract automatically or extract manually (in each frame) features to put into ontology. On the other hand, because the contrast between light and shade is different as well as the resolution of each video is also distinct, our tool implemented a simple adjustable set of image processing manually to support detection and classification.

In order to be compatible for extracting the features of EVTDs where existing many fundamental movements sepa-

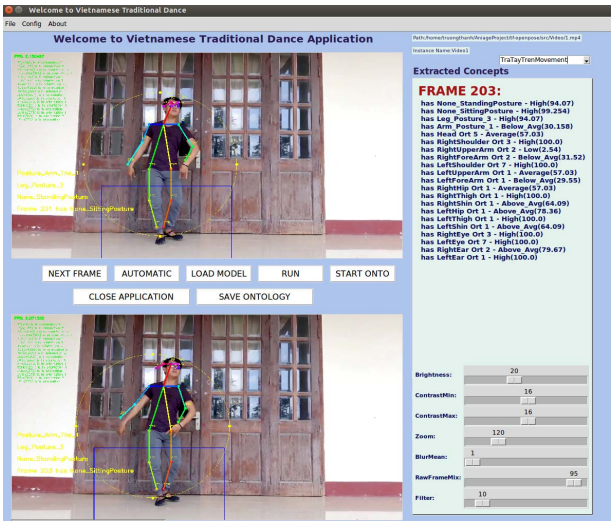

Figure 7: An automatic extraction tool of EVTDs

Table I: Datasets of the fundamental posture of EVTDs

	Postures (Pos)						Total
	P1	P2	P3	P4	P5	None Pos	
Leg Postures	1260	685	1185	494	673	1407	5704
Sitting Posture	1252	798	none	none	none	2163	4213
Standing Posture	none	685	none	494	673	2084	3936

rately. We also designed a simple function to change flexibly the classes/concepts for storing easily the different movements.

IV. Experimental Result

We implemented the propositional tool on the computer supporting graphical card (NVIDIA GeForce GTX 950M with total memory is 8107 MB) to run TensorFlow on multiple GPUs. We used python programming language and Keras libary to implement the EVTD postures automatic detection framework on each frame. We randomly split the dataset into two different training(2/3) and test (1/3) sets *(with the posture image datasets collected as in Table I, collected from 15 videos)*. In addition, we also implemented scikit-learn library [3] to use several ML algorithms including logic regression, Support vector machine (SVM - C=10000, Gama=0.002), Random Forest (200 decision trees), Stochastic Gradient Descent classifier (SGD), K-Nearest Neighbors (KNN - K=5), Naïve Bayes classifier. The experimental results of the propositional CNN model are in Table II for Rank-1 accuracy). The result achieved the high accuracy (F1) using Logic Regression as follows: 98.88% of Leg postures (for 06 classes), 99.84% of Sitting Posture (for 03 classes), 99.37% of Standing Posture (for 04 classes). In general, most of the classification results achieved the high accuracy around above 90%.

[3]http://scikit-learn.org/stable/

Table II: Comparisons of Rank-1 accuracy of algorithms

Algorithm	Postures		
	Leg Posture	Sitting Posture	Standing Posture
Logic Regression	98,88	99.84	99,37
SVM (Linear)	98,84	99.80	99,27
SGD Classifier	92,18	94,23	93,78
Random Forest	96,47	97,63	98,02
K-Nearest Neighbors	94,11	96,15	95,07
Naïve Bayes	96,30	97,06	97,34

V. Conclusion and Future Works

With the aim of the preservation and promotion in the intangible cultural heritage in general as well as developing an application to store the Vietnamese traditional dances. In particular, we presented a methodology to identify automatically the significant concepts of EVTDs to build an intelligent repository. Using the machine learning algorithms combines with the CNN architectures and HPE to detect the important features are discussed in this paper. On the basis of the propositional tool, we collected and managed the heterogeneous dance data of EVTDs (from raw videos with low revolution).

Acknowledgements

This work has received support from the European Project H2020 Marie Sklodowska-Curie Actions (MSCA), Research and Innovation Staff Exchange (RISE): Aniage project (High Dimensional Heterogeneous Data based Animation Techniques for Southeast Asian Intangible Cultural Heritage Digital Content), project number 691215.

REFERENCES

[1] L.T.Loc, "Mua dan gian cac dan toc Viet Nam", in *Thoi-dai Publishing house*, 1994.
[2] V.Hoc, "Nghe thuat mua Viet Nam, thoang cam nhan", in *Nation Publishing House*, 2001.
[3] T.V. Son, Đ. T. Hoàn, N. T. M. Hương, "Mua dan gian mot so dan toc vung Tay Bac", in *Culture and Nation Publishing House*, 2003.
[4] L.N.Canh. "Đại cương nghệ thuật múa", in *Culture and information publishing house*, 2003.
[5] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications", arXiv preprint, arXiv:1704.04861, 2017.
[6] Z. Cao, T. Simon, S. Wei, Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", arXiv preprint, arXiv:1611.08050, 2016.
[7] F. Chollet. Xception: Deep learning with depthwise separable convolutions. arXiv preprint arXiv:1610.02357, 2016.
[8] X. Xia, C.Xu, Inception-v3 Flower Classification, 2017 2nd International Conference on Image, Vision and Computing, 978-1-5090-6238-6/17/ 2017 IEEE
[9] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications", arXiv preprint, arXiv:1704.04861, 2017.