AN AUTOMATIC DETECTION OF FUNDAMENTAL POSTURES IN VIETNAMESE TRADITIONAL DANCES

Ngan-Khanh Chau[1] and Truong-Thanh Ma[2]

[1]An Giang University, Vietnam National University Ho Chi Minh City, Vietnam
[2]Soc Trang Community College, Vietnam

ABSTRACT

Preserving and promoting the intangible cultural heritage is one of the essential problems of interest. In addition, the cultural heritage of the world has been accumulated and early respected during the development of human society. For preservation of traditional dances, this paper is one of the significant processed steps in our research sequence to build an intelligent storage repository that would help to manage the large-scale heterogeneous digital contents efficiently, particularly in dance domain. We concentrated on classifying the fundamental movements of Vietnamese Traditional Dances (VTDs), which are the foundations of automatically detecting the motions of the dancer's body parts. Moreover, we also propose a framework to classify basic movements through coupling a sequential aggregation of the Deep-CNN architectures (to extract the features) and Support Vector Machine (to classify the movements). In this study, we detect and extract automatically the primary movements of VTDs, we then store the extracted concepts into an ontology that serves for reasoning, query-answering, and searching dance videos.

KEYWORDS

Vietnamese Traditional Dance, Deep learning, Support vector machine.

1. INTRODUCTION

The cultural heritage, which has been accumulated throughout the development of humanity, is early respected. Vietnam possesses an array of intangible cultural heritages that have been recognized in the world. Indeed, the intangible cultural heritage (ICH) in Vietnam shows the diversity of styles and expressions. Dances are widely displayed in community cultural activities. Each nation is proud of its own dances and preserves it from generation to generation. Dancing is also a quintessential form of human motion and expression. Most of ICH in the ethnic communities has been collected and studied through the Vietnamese traditional dances (VTDs). These colourful dance paintings naturally show the cultures of ethnic groups as well as regions through key postures. Identifying key moves keeps an important role in spreading traditional dances as well as native culture. Fortunately, most of the moving phrases of the distinguishable dances originate from the fundamental movements, the detection of the basic movements to identify the distinct dances is completely expected. In this paper, we would focus on detecting automatically the essential movements and storing the primary concepts into VTD's movement ontology-base. With this study, we will give our part to the preservation and development of national cultural arts.

David C. Wyld et al. (Eds): ITCSE, NLCA, ICAIT, CAIML, ICDIPV, CRYPIS, WiMo - 2020
pp. 263-277, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.100524

From an academic point of view, there is much research in dance analysis such as developing algorithms that may recognize how well a user can imitate certain motion patterns presented by an avatar [34, 35]. Several dance topics have recently been studied, including ballet dance [6], Tsamiko dance [7, 8], and Salsa dance [9, 15].

Automatic recognizing performer's movements is a complicated problem for artificial intelligence (AI) scientists, which involves in mining and categorizing spatial patterns of human postures from videos. Dance movements are well-defined as a temporal variation of human body. The problem is to extract and detect the human posture and classify it into a label based on the trained CNN features. The goal of this work is to extract the features from multiple CNN-architectures of different VTDs postures in dance videos. In this paper, we concentrate on classifying the fundamental movements of ethnic Vietnamese Thai Community (EVTC) composed of “'Standing movements, Leg movements, Hand movements, Sitting movements”.
One of the main contributions of this paper is to present a general framework using the aggregated CNN-architectures to automatically detect the key postures of ethnic Vietnamese Thai dances (EVTD). Our classification model focuses on extracting the features and aggregating the features to support the classification of the postures. Especially, we used three of the effective existing CNNs architectures to extract the significant features and using algorithms of machine leaning to classify. Most of our datasets are collected from the raw dance videos on the social networks (almost all from YouTube).

In research process, we decomposed our approach into three main stages: the first is reconstructing a schema for panorama view of VTDs; the second is region-zone of VTDs, considering this aspect as a large branch because most of the VTDs originate from distinguishable ethnic groups living in distinct regions; the last is the fundamental postures to identify the name of dances. In this article, we mainly concentrate on the third stages involved in automatic detection of basic motions. Our primary challenge is to determine the principal concepts from EVTD's postures combined with a set of desirable properties to build a useful dance search engine.

This paper is structured as follows section 2 gives an overview and review of recent related works. Section 3 provides a description of EVTD's basic postures, we then present the proposed model to classify and detect the specific postures automatically in section 4. In section 5, we discuss the experimental results. The last section is the conclusion (in section 6).

2. OVERVIEW AND RELATED WORKS

2.1. Vietnamese Traditional Dance Overview

Vietnam is a multi-ethnic country with many different cultures [25] due to the combination of fifty-four-ethnic groups living in one territory. The traditional dances had become the spiritual foods of Vietnamese people, it explicitly influences the life from urban to rural. Most of the VTDs are taught from previous generations using "word of mouth", the present generation would instruct fundamental movements to the adjacent one. Additionally, VTD is a steady bridge in education of human dignity, morality, and even historical knowledge. Instead of learning the historical lessons in regular classes as well as participating in the training courses for life skills, dance has become a digital channel to effectively educate personality, knowledge, and even ethnicity for generations.

Generally, the VTDs [1] concentrate on ethnicity, aestheticism and bringing many significations. The worship of gods as well as ancestors plays an important role in the life of people in Vietnam, therefore several dances are performed in festivals and celebrations with desiring to be blessed. In addition, combining particular props and traditional costumes would be the remarkable characteristic in VTDs.

The Vietnamese culture brings many multiform traditional cultural properties, which plays an important role in Vietnamese community. Most VTDs are built up from the ethnic groups in different life environments and regions, they contain a large number of the significant characteristics of specific regions. The dance movements of the ethnic groups stem from the life activities, each posture in the dance is depicted an action of their life. Therefore, the basic postures will be one of the stable platforms as well as the being the essential features for identifying different dances.

2.2. Related Work

During the last two decades, people have tried to develop different algorithms for human activity analysis [20, 24] for wide applications in the field of surveillance, patient monitoring and more. Most of the studies have been reported on classifying human activity from videos. Recently, researchers are trying to classify an activity from a single image [22, 23, 27]. In the video based activity recognition, people have tried with different human activities like walking, jogging, running, boxing, hand waving, hand clapping, pointing, digging and carrying for a single actor [4, 24].

There are also a few works on group activities such as [12, 16]. With the best of our knowledge, no one has addressed the dance classification problem so far, at least in computer vision domain. Due to the increase in multimedia data access through the internet, multimedia data, particularly video data indexing becomes more and more important. Not only in the retrieval but also for digitization of cultural heritage, this could be an interesting problem. It can be used to analyse a particular dance language.

Some researchers used space time features to classify the human action. Blank et al. represented the human action as three-dimensional shapes included by the silhouettes in the space-time volume [13]. They used space-time features such as local space-time salience, action motivations, shape structure, and orientation to classify the actions. , they recognised human action based on space-time locally adaptive regression kernels and the matrix cosine similarity measure [2]. Klaser et al. localized the action in each frame by obtaining generic spatio-temporal human tracks [11]. They used a sliding window classifier to detect specific human actions.

There are several attempts to recognize the postures from multiple videos sources [2, 24]. Aggarwal et al. has categorized the recognition of human activities in two classes as single-layered approach and hierarchical approach [24]. In single-layered approach, activities are recognized directly from videos, while in hierarchical approach, an action is divided into sub-actions [10]. The action is represented by classifying it into sub-actions. Wang et al. have used topic modelling to model the human activity [18]. They proposed a video by Bag-of-Wards representation. Later, they have used a model which is popular in object recognition community, called Hidden Conditional Random Field (HCRF) [19]. They modelled human action by flexible constellation of parts conditioned on image observations and learn the model parameters in max-margin framework and named it max-margin hidden conditional random field. Md Faridee et al. has built a tool to recognize dance activities automatically [32]. They applied the Convolution Neural Network based on body sensor network to assess steps of dances. Mallick et al. has captured and extracted postures of the Indian classical Dances using the Hidden Markov Model and Kinect [33].

3. EVTD POSTURES DESCRIPTIONS

3.1. Ethnic Vietnamese Thai Dances

Thai community in Vietnam is one of the ethnic groups which possesses a 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 the explicitly with respect to ethnic Vietnamese Thai dance (EVTD), in this paper, we gave more attention on analysing and determining the postures' features which keep a vital role to classify the Thai dances in the Vietnamese territory.

In general, the remarkable dances are always performed in most of the festivals or private celebration days in community. Specially, when it comes to the Thai ethnic group, certainly, "Xoe dances" [5] is one the most important and popular dances in this community. It is not only the worth cultural heritage as well as being the Thai community symbol but also one of the indispensable intellectual nourishment of Thai people in the festivals and ceremony. There are six kinds of "Xoe" dances including "Kham khan moi lau" (representing the culture of communicating), "Pha xi" (the unity of communities), "Nhom khan" (representing happiness of ethnic group), Don hon (steels and sweet hearts), "Kham khen" (work together as communities), "Om lop top mu" (recall with nostalgia when saying goodbye). Several dance evidences in "Xoe dances" including Handkerchief Dance, Tinh Tau Dance, Xe-Ma-Hinh Dance, Vi-Khan Dance performed in the important festivals, for example as Kin pan then, is a significant festival taken place on March 10th (according to lunar calendar) in Vietnam's North Southern region. In each EVTD, apart from the inspiration of choreographers and the traditional costumes, the remarkable characteristics to determine EVTDs are the foundation postures which are the base of the creative combination in each motion in order to create the particular dances of Vietnamese Thai people. Correspondingly, the detection of the basic postures is one of the important steps to collect automatically the dance dataset for an ontology-based of EVTDs.

3.2. Fundamental Postures of EVTDs

Representing the details of each motion in EVTDs is necessary when considering basic motion characteristics. It is divided in five primary characteristics in overview [5]: Orientation, Arm Posture, Leg Posture, Sitting Posture, Standing Posture. The orientation of EVTDs is split into eight orientations (from orientation 1 to orientation 8). In this article, we focused on the postures of the lower-body area including leg postures, sitting postures, standing postures to collect training dataset and classification. We particularly describe the remaining postures as follows:

3.2.1. Arm Posture

Most arm postures are concentrated on depicting life activities in Thai community, therefore the basic postures are simple and habitual. They are divided in five primary postures: VN-Thai-The-1-Arm, VN-Thai-The-2-Arm, VN-Thai-The-3-Arm, VN-Thai-The-4-Arm, VN-Thai-The-5-Arm. They are grouped into two distinct clusters: open-arm posture and close-arm posture.

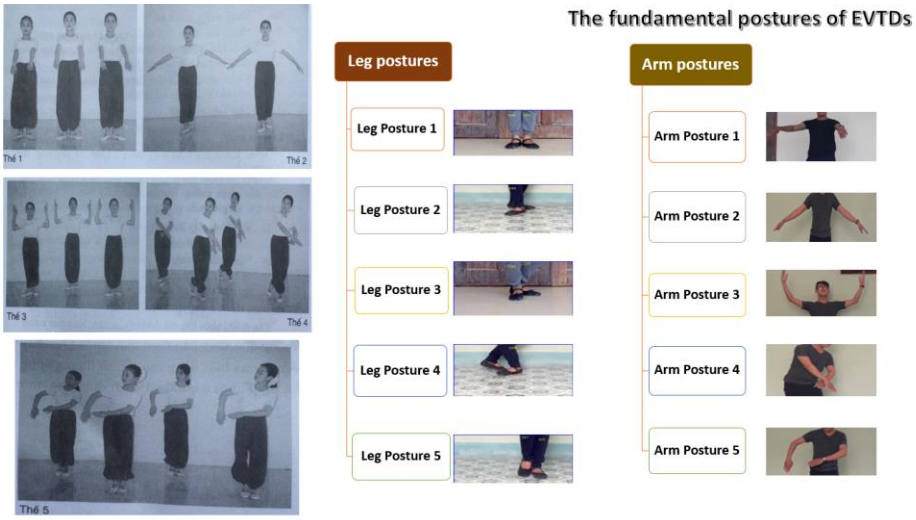


Figure 1. Arm postures and Leg postures of EVTDs

3.2.2. Leg Posture

There are five significant leg postures to represent for EVTD postures. It includes VN-Thai-The-1-Leg, VN-Thai-The-2-Leg, VN-Thai-The-3-Leg, VN-Thai-The-4-Leg, VN-Thai-The-5-Leg.

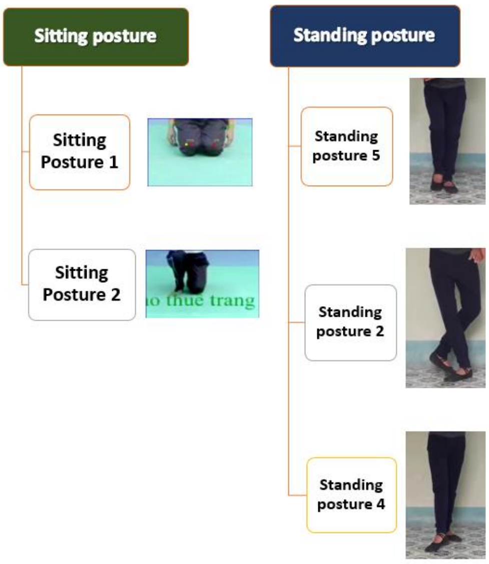


Figure 2. Sitting postures and standing postures of EVTDs

3.2.3. Sitting and standing Posture

Sitting posture is divided into two types, it consists of VN-Thai-The-1-Sitting, VN-Thai-The-2-Sitting. There are three standing postures in EVTD: VN-Thai-The-5-Standing, VN-Thai-The-2-Standing, VN-Thai-The-4-Standing.

4. Proposed Methodology

4.1. Human Pose Estimation

Human pose estimation (HPE) is one of the most challenging problems in computer vision and plays an essential role in human body modelling. Regarding HPE, existing two main approaches are bottom-up and top-down estimation. The state-of-the-art solutions model for the two key issues in bottom-up approach are joint detection and spatial configuration refinement, together using convolutional neural networks (CNN) for training and classifying. A real-time method to estimate multi-person pose efficiently is the so-called Openpose [29] (written in C++ using OpenCV and Caffe), developed by Carnegie Mellon University. In this paper, 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 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 postures is to determine concretely parts of body in raw dance videos aiming at describing most motions. The architecture also predicts detection reliable maps and affinity fields that are encrypted part-to-part association simultaneously as shown in Figure 3. The network is split into two branches: Branch 1 is responsible for predicting reliable maps, and Branch 2 predicts the affinity fields. TF-Openpose takes a 2D colour image as an 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 a range depended on the key-point scale. It can be estimated 18 key-points body pose from COCO 2016 dataset.

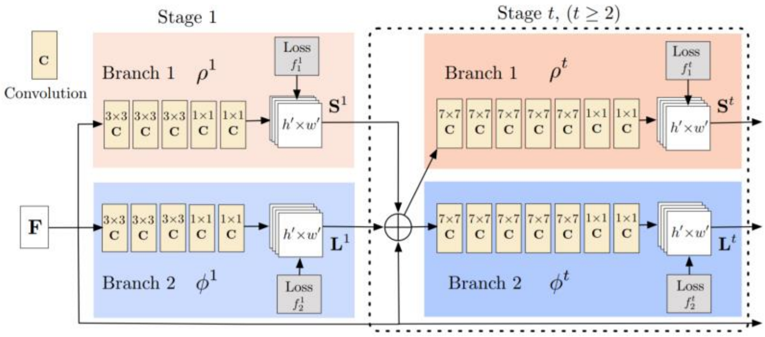


Figure 3. The architecture of the two-branches using CNN in Openpose

Each branch is an iterative prediction architecture which refines the predictions over successive stages, $t \in \{1,2,...,T\}$, with the intermediate supervision at each stage. The frames from raw video are analysed by CNNs, generating a set of feature maps F that is the input to the first stage of each branch. At the first stage, the network produces a set of detection reliable maps $S^1 = \rho^1(F)$ and a set of partial affinity fields $L^1 = \phi^1(F)$, where ρ^1 and ϕ^1 are the CNNs for inference at Stage 1. In each subsequent stage, the predictions from both branches in the previous stage, along with the original frame features F, are concatenated and used to produce refined predictions,

$$S^t = \rho^t(F, S^{t-1}, L^{t-1}), \forall t \geq 2,$$
$$L^t = \phi^t(F, S^{t-1}, L^{t-1}), \forall t \geq 2,$$

where ρ^t and ϕ^t are the CNNs for inference at Stage t.

Realizing the requirements of a high configuration regarding GPU for Openpose handled, we proposed to apply 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 made 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 [29] and Mobile-Net [28]. (1) With regard 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 for use in Tensorflow; (2) Based on the Mobile-Net paper [28], with 12 convolutional layers are used as feature-extraction layers.

4.2. Deep Convolutional Neural Networks – DCNNs

Convolutional neural networks (CNNs) have been applied to visual tasks since the late 1980s. With a few distributed applications, they were dormant until the mid-2000s when developments in computing power and the advent of a large amount of labelled data, supplemented by improved algorithms, contributed to their advancement and brought them to the forefront of a neural network renaissance that has seen rapid progression since 2012. CNNs are feed-forward networks in which information flow only takes place in one direction, from their inputs to their outputs.

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 are 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 as illustrated in Figure 4.

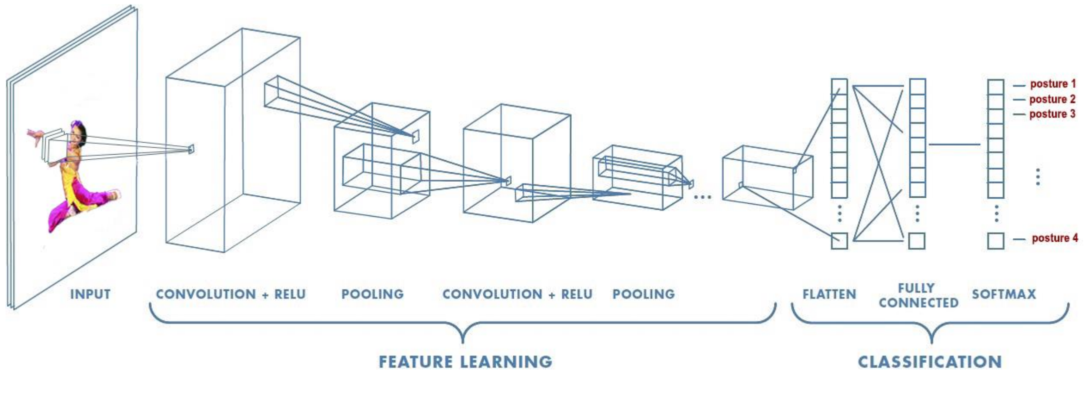


Figure 4. Neural network with many convolutional layers

In addition, the DCNNs are also developing rapidly as a result many DCNNs novel architectures are proposed. Each DCNN architecture undertakes a distinguishable role to train the different datasets. In this study, we used the existing successful DCNN architectures to extract features, our proposed framework is associating the features of the different extracted DCNN architectures. It can be seen that the dataset of EVTDs would accelerate considerably in the following years as well as the angles of the camera are flexible and multiform. Furthermore,

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

combining the features to increase the explicit discrimination between the featured vectors is quite expected. We therefore proposed a flexible framework to easily select the new architectures for extracting the features. Our contributed framework concentrated on the sequential aggregation of the features of many effective DCNN architectures to support detecting the postures of EVTDs automatically.

At present, there are a large number of the published CNN architectures. Fortunately, 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). In this research, we selected three of the efficient architectures published recently that will be discussed in subsection 4.3.

4.3. A general framework

In this subsection, we present a general framework to serve in training, detecting and extracting automatically the fundamental postures of EVTDs aiming at storing the primary concepts into a lightweight ontology-based in order to support for classification, query-answering, reasoning, and searching dance videos. The below model is a sequential process to train the dataset issued from the frames of the dance videos. We used Openpose and Tensorflow library (called by TF-Openpose) to detect body parts, the reason we used these libraries is we need a useful tool to detect human skeleton aiming at describing the actions and motions in each frame of EVTD videos. However, regarding the leg part, the postures of this part could not be covered using TF-Openpose, i.e. foots, we therefore addressed on classifying *legs-postures* by machine learning. In detail, we crop the parts of legs and save them into the distinct folders for collecting trained dataset. The next steps are the extraction of the features *(based upon the deep CNN models proposed)* and the classification with machine learning algorithms.

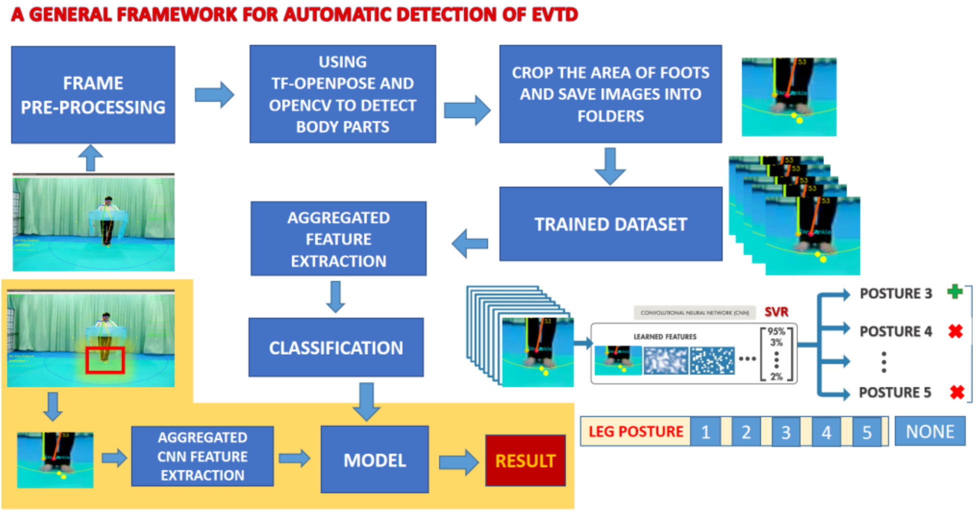

A GENERAL FRAMEWORK FOR AUTOMATIC DETECTION OF EVTD
FRAME PRE-PROCESSING
USING TF-OPENPOSE AND OPENCV TO DETECT BODY PARTS
CROP THE AREA OF FOOTS AND SAVE IMAGES INTO FOLDERS
TRAINED DATASET
AGGREGATED FEATURE EXTRACTION
CLASSIFICATION
CONVOLUTIONAL NEURAL NETWORK (CNN)
SVR
LEARNED FEATURES
95%
3%
2%
POSTURE 3
POSTURE 4
POSTURE 5
LEG POSTURE
1
2
3
4
5
NONE
AGGREGATED CNN FEATURE EXTRACTION
MODEL
RESULT

Figure 5. A general framework for automatic detection of EVTD

Additionally, we classified on three sets of the basic postures of EVTDs including sitting postures, standing postures, and legs postures. With each kind of the different postures, we create an extra folder in which contains the common motion images called non-posture, the particular examples consist of non-leg-posture, non-standing-posture, non-sitting-posture. As soon as the

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

postures are undetermined precisely, non-posture classification is expected to solve the problems in the online detection.

4.4. Extraction of aggregated features

In this subsection, we introduce a classification model aggregating consequently the deep CNN architectures to extract the features. As can be seen that each CNN architecture deals with the different cases of the particular datasets as well as combining the features to represent a vector is expected because it will be boosted strongly the discrimination between the classifications. In addition, the datasets of ETVDs mainly focus 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 method 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 pro-processing). In each image frame, we extracted the features from three CNN architectures, including Xceptions [30] (2048 features), InceptionV3 [31] (2048 features), Mobilenet [28] (1024 features). The next step, we aggregated the extracted features to have a feature vector with 5120 dimensions. In order to have a best candidate for classification, the comparison of ML algorithms is essential as presented in the next section.

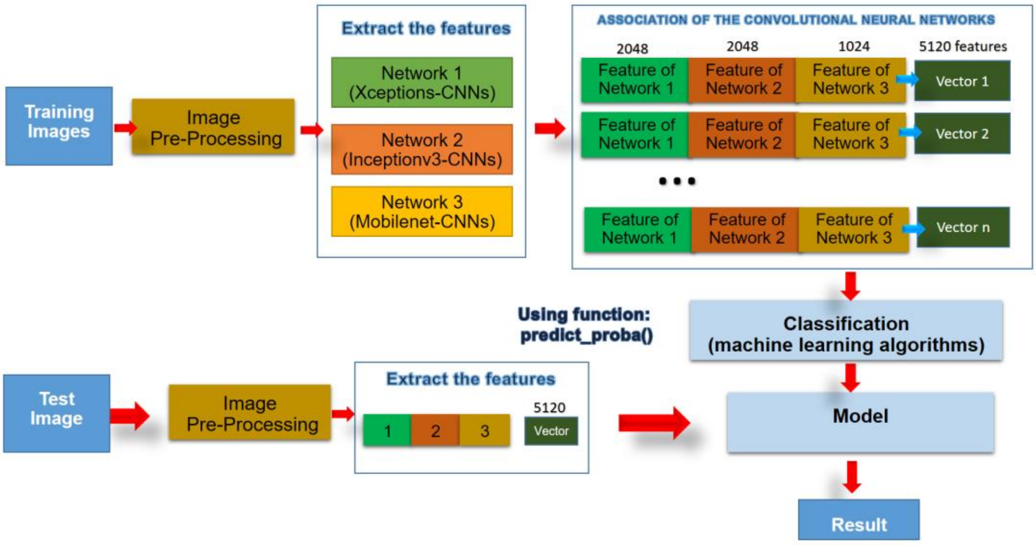


Figure 6. A classification model aggregating the deep CNN architectures

The main idea is to have a classification tool that allows flexible updating of novel architectures that will be published in order to improve classification accuracy as soon as the dataset is accelerated. In addition, we can also extract the features using CNN architectures in parallel, nevertheless, we do not experiment the parallel models in this paper.

5. EXPERIMENTAL RESULT

We implemented the propositional framework on the computer supporting graphical card (NVIDIA GeForce GTX 950M with total memory is 8107 MB) to run TensorFlow on multiple GPUs. To evaluate the proposed model, we had experimented on the Flower17 dataset of University of Oxford including 17 categories of the common flowers in the UK with 80 images for each class. Particularly, the images have large scale, pose and light variations and there are also classes with large variations of images within the class and quite similar to other classes. We randomly split the dataset into two different train (2/3) and test (1/3) sets. 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 Neighbours (KNN - K=5), Naïve Bayes classifier. The experimental results of the propositional CNN model are in Figure 7 and Table 1 for Rank-1 accuracy). Additionally, we also test our model with Rank-3 accuracy presented the results in Figure 8 and Table 2.

After comparing of the experimental results with Rank-1 and Rank-3 accuracy, the proposed model obtained the highest accuracy in the ML algorithms. The noticeable results are Logic regression and SVM algorithms achieved above 98% of Rank-1 accuracy and 100% of Rank-3 accuracy. The result on Flower17 dataset also showed that our model is more accuracy than the single architecture models.

Table 1. Accuracy of Flower17 dataset (Rank-1)

Algorithm	Deep CNN Architectures (Rank -1)			
	Xception	InceptionV3	Mobilenet	Proposed Method
Logic Regression	92,26	92,29	97,14	98,24
SVM (Linear)	94,49	91,85	96,92	98,02
SGD Classifier	80,84	77,31	86,12	88,99
Random Forest	86,56	85,68	91,85	91,85
K-Nearest Neighbors	83,48	84,58	92,95	94,05
Naïve Bayes	86,12	84,12	89,87	92,07

Table 2. Accuracy of Flower17 dataset (Rank-3)

Algorithm	Deep CNN Architectures (Rank -1)			
	Xception	InceptionV3	Mobilenet	Proposed Method
Logic Regression	99,12	98,68	99,34	100,00
SVM (Linear)	98,90	98,46	99,56	100,00
SGD Classifier	84,36	82,82	88,33	89,87
Random Forest	96,48	96,48	96,70	97,58
K-Nearest Neighbors	95,15	95,15	98,02	99,12
Naïve Bayes	96,70	96,04	96,04	97,80

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

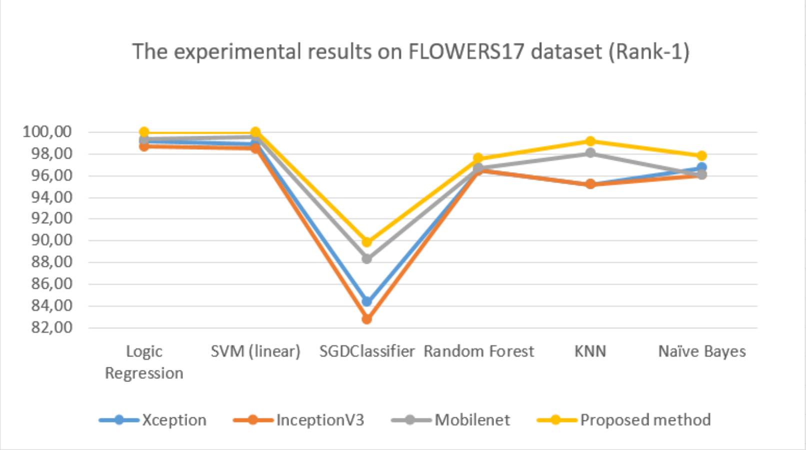


Figure 7. Comparison of Rank-1 accuracy

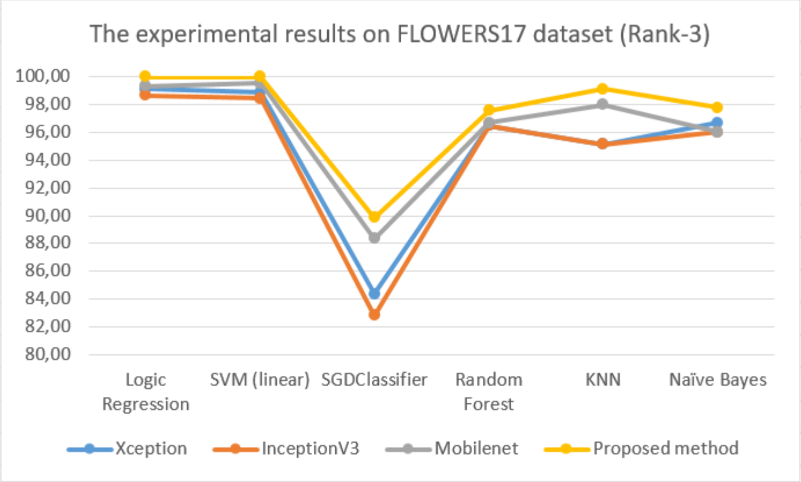


Figure 8. Comparison of Rank-3 accuracy

Furthermore, our main task is building a framework to detect automatically the fundamental postures of EVTDs. Based on the framework presented in Figure 5, we had collected a dataset including six images folders of leg postures, three images folders of sitting postures and four images folders of standing postures to support for training and evaluating. The particular number of each posture showed in Table 3. After collecting the dataset, we also divided the dataset into 2 distinct training (2/3) and test (1/3) sets, our model and framework gained the experimental results is fully expected. They achieved the high accuracy as follows: 98.88% of Leg postures, 99.84% of Sitting Postures, 99.37% of Standing postures. Comparisons of rank-1 accuracy showed in Table 4 and Figure 9. In general, most of the classification results achieved the high accuracy all above 90%.

After evaluating the model with the collected dataset, the implemented result about the propositional framework is demonstrated in the Figure 10. In each frame, we detected and extracted the essential postures based upon our proposal framework, this is one of the important foundation to provide automatically for the fundamental concepts into lightweight ontology-based served in preserving and promoting the intangible cultural heritage of traditional dances in

Vietnam. Moreover, Logic Regression algorithm and Linear SVM algorithm give the highest and quite similar results as they reach over 98% for all postures (Table 4, Figure 9).

Table 3. Datasets of the fundamental of EVTDs

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

Table 4. Comparisons of Rank-1 accuracy of algorithms

Algorithm	Postures (Pos)		
	Leg Posture	Sitting Posture	Standing Posture
Logic Regression	98,84	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

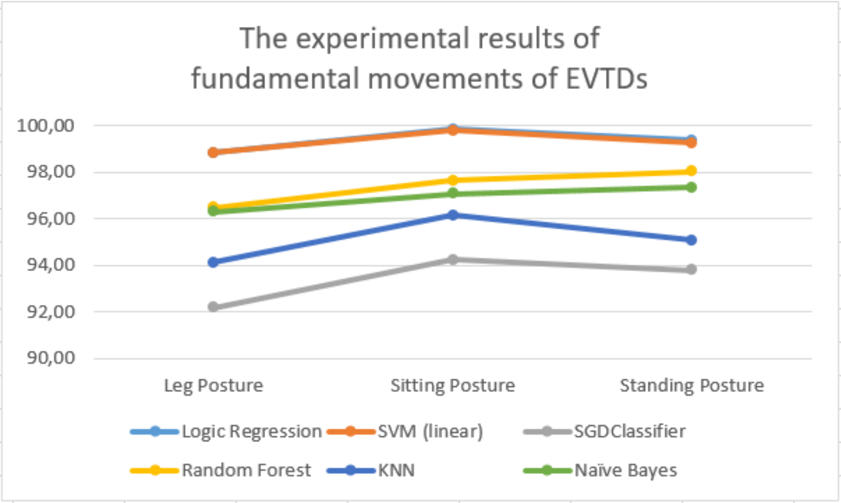
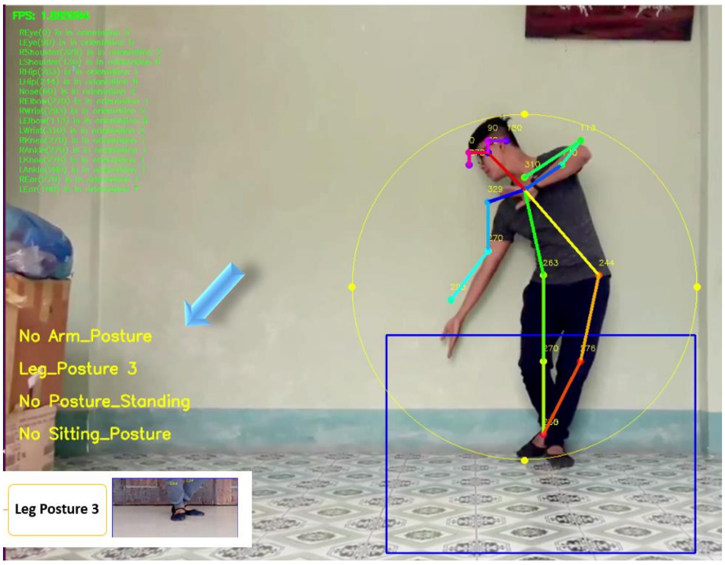


Figure 9. Comparisons of Rank-1 accuracy in EVTDs

Figure 10 illustrates the online identification of a performer's gestures. As can be seen, when the performer makes a move, our system will predict and display an illustration corresponding to the performed movement (*in this case, the results are "Leg_Posture 3" message and "Leg Posture 3" image*).



No Arm_Posture

Leg_Posture 3

No Posture_Standing

No Sitting_Posture

Leg Posture 3

Figure 10. The experimental result of the propositional framework

The result of this paper will be one of the preliminary of collected dataset and classification. In implemented process, we realized that this collected dataset is not able to represent and reflect most of the different angle in dance. There are some difficulties in classification including the differences between the postures and gestures of a professional dancer and an amateur person as well as the distinction from different directions to look. Therefore, it is necessary to collect a huge dataset in the future. We would also update the architectures into our framework to support for classification aiming to build the intelligent repository and a query-answering and reasoning tool[4]. Although our proposed system has good results in detecting and classifying dances which are performed by a single person (in dance courses or practice sessions), it is difficult to identify dances performed by many dancers in performances or festivals due to gestures may be obscured by other dancers or performance costumes.

6. Conclusion and Future Works

With the aim of the preserving and promoting 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 serve in building an intelligent repository. Using the machine learning algorithms combined with the CNN architectures to classify dataset are discussed in this paper. On the basis of the propositional framework, we proposed a model aggregating consequently the CNN architectures to extract the features. After an experimented process is done, we gained the results which are absolutely expected.

[4] Our tool to detect EVTD's postures and to store into EVTD's ontology published in the Github at: https://github.com/truongthanhmastcc/VietnameseDance

The presented work in this paper is one of the important first steps on preservation and promotion of EVTDs based on the background of artificial intelligent. These initial steps would be the foundation for creating universal traditional dance repository aiming to support for advanced heterogeneous digital storage, indexing, classification, reasoning and searching dance videos. Based on the concepts extracted automatically, the next step will put all concepts into ontology-based. In the future, we will expand the collection of the dance dataset as well as improve the model through the novel architectures which will be more compatible in dance domain. We will also build a lightweight ontology-based for this expanded dataset.

REFERENCES

[1] L.T.Loc, ”Mua dan gian cac dan toc Viet Nam”, in Thoi-dai Publishing house, 1994.

[2] H. Seo and P. Milanfar. Action recognition from one exam-ple.IEEE Trans. on Pattern Analysis and Machine Intelli-gence, 33(5):867–882, 2011.

[3] V.Hoc, ”Nghe thuat mua Viet Nam, thoang cam nhan”, in Nation Publishing House, 2001.

[4] I. Laptev and T. Lindeberg. Velocity adaptation of space-time interest points. InICPR, pages 52–56, September 2004

[5] T.V. Son, D. T. Hoan, N. T. M. Huong, ”Mua dan gian mot so dan toc vung Tay Bac”, in Culture and Nation Publishing House, 2003.

[6] Kyan, G. Sun, H. Li, L. Zhong, P. Muneesawang, N. Dong, B. Elder, and L. Guan, “An approach to ballet dance training through ms kinect and visualization in a cave virtual reality environment,” In ACM Transactions on Intelligent Systems and Technology, vol. 6, no. 2, p. 23, 2015.

[7] G. Chantas, A. Kitsikidis, S. Nikolopoulos, K. Dimitropoulos, S. Douka, I. Kompatsiaris, and N. Grammalidis, ”Multi-entity bayesian networks for knowledge-driven analysis of ich content”, in Proc. 1st International Workshop on Computer vision and Ontology Applied Cross-disciplinary Technologies in conj. with ECCV, pp. 355–369, Springer, 2014.

[8] A. Kitsikidis, K. Dimitropoulos, E. Yilmaz, S. Douka, and N. Grammalidis, ”Multi-sensor technology and fuzzy logic for dancer motion analysis and performance evaluation within a 3d virtual environment”, in Universal Access in Human-Computer Interaction. Design and Development Methods for Universal Access, pp. 379–390, Springer, 2014.

[9] A. Masurelle, S. Essid, and G. Richard, ”Multimodal classification of dance movements using body joint trajectories and step sounds”, in Proc. 14th International Workshop on Image Analysis for Multimedia Interactive Services, pp. 1–4, IEEE, 2013.

[10] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understand-ing videos, constructing plots: Learning a visually groundedstoryline model from annotated videos. InCVPR, 2009.

[11] A. Klaser, M. Marszałek, C. Schmid, and A. Zisserman. Hu-man focused action localization in video. InInternationalWorkshop on Sign, Gesture, Activity, 2010

[12] M. S. Ryoo and J. K. Aggarwal. Recognition of compositehuman activities through context-free grammar based repre-sentation. InCVPR, pages 1709 – 1718, October 2006.

[13] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri.Actions as space-time shapes. InICCV, volume 2, pages1395–1402, October 2005.

[14] L.N. Canh, Nghe Thuat mua Ha Noi Truyen thong va Hien dai”, in Ha Noi Publishing House, 2011.

[15] Karavarsamis S., Ververidis D., Chantas G., Nikolopoulos S., Kompatsiaris Y. Classifying salsa dance steps from skeletal poses; Proceedings of the 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI); Bucharest, Romania. 15–17 June 2016; pp. 1–6.

[16] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions:Discriminative models for contextual group activities. InNIPS, 2010.

[17] L.N. Canh, “Nghe thuat mua truyen thong Khmer Nam Bo”, in Vietnamese Dance College, Cultural and Nation publishing house, 2013.

[18] Y. Wang and G. Mori. Human action recognition by semi-latent topic models.IEEE Trans. on Pattern Analysis and Machine Intelligence Special Issue on Probabilistic Graphical Models in Computer Vision, 31(10):1762–1774, 2009.

[19] Y. Wang and G. Mori. Hidden part models for human ac-tion recognition: Probabilistic vs. max-margin.IEEE Trans.on Pattern Analysis and Machine Intelligence, 33(7):1310–1323, 2011

[20] N. Nayak, R. Sethi, B. Song, and A. Roy-Chowdhury. Mo-tion pattern analysis for modeling and recognition of com-plex human activities.Visual Analysis of Humans: Lookingat People, Springer, 2011.

[21] L.N.Canh. ”Mua tin nguong dan gian Viet Nam”, in Social Science Publishing house, 1998.

[22] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In CVPR, pages 2030–2037, June 2010.

[23] B. Yao, A. Khosla, and L. Fei-Fei. Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. InICML, Bellevue, USA, June 2011.

[24] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: Areview.ACM Computing Surveys(To appear), 2011

[25] L.N.Canh. ”Dai cuong nghe thuat mua”, in Culture and information publishing house, 2003.

[26] D.Lin and P.Paten. Induction of semantic classes from natural language text. In proceedings of SIGKDD-01. pp.317-322. 2001.

[27] S. Maji, L. Bourdev, and J. Malik. Action recognition from adistributed representation of pose and appearance. InCVPR,pages 3177–3184, June 2011.

[28] 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.

[29] Z. Cao, T. Simon, S. Wei, Y. Sheikh, ”Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”, arXiv preprint, arXiv:1611.08050, 2016.

[30] F. Chollet. Xception: Deep learning with depthwise separable convolutions. arXiv preprint arXiv:1610.02357, 2016.

[31] 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.

[32] Md Faridee, Abu Zaher, Sreenivasan Ramasamy Ramamurthy, and Nirmalya Roy. “”HappyFeet: Challenges in building an automated dance recognition and assessment tool”. *GetMobile: Mobile Computing and Communications* 22.3 (2019): 10-16.

[33] Mallick, Tanwi, Partha Pratim Das, and Arun Kumar Majumdar. "Posture and sequence recognition for Bharatanatyam dance performances using machine learning approach." *arXiv preprint arXiv:1909.11023*, 2019.

[34] Gibbs B, Quennerstedt M, Larsson H. Teaching dance in physical education using exergames. *European Physical Education Review*. 2017 May;23(2):237-56.

[35] Kico, Iris, et al. "Visualization of Folk-Dances in Virtual Reality Environments." *Strategic Innovative Marketing and Tourism*. Springer, Cham, 2020. 51-59.