Grammar of Dance Gesture from Bali Traditional Dance

Yaya Heryadi[1], Mohamad Ivan Fanany[2] and Aniati Murni Arymurthy[3]

**[1] School of Computer Science, Bina Nusantara University
Jakarta 10270, Indonesia**

**[2] Fakultas Ilmu Komputer, Universitas Indonesia
Depok 16424, Indonesia**

**[3] Fakultas Ilmu Komputer, Universitas Indonesia
Depok 16424, Indonesia**

Abstract

Automatic recognition of dance gesture is one important research area in computer vision with many potential applications. Bali traditional dance comprises of many dance gestures that relatively unchanged over the years. Although previous studies have reported various methods for recognizing gesture, to the best of our knowledge, a method to model and classify dance gesture of Bali traditional dance is still unfound in literature. The aim of this paper is to build a robust recognizer based on linguistic motivated method to recognize dance gesture of Bali traditional dance choreography. The empiric results showed that probabilistic grammar-based classifiers that were induced using the Alergia algorithm with Symbolic Aggregate Approximation (SAX) discretization method achieved 92% of average precision in recognizing a predefined set of dance gestures. The study also showed that the most discriminative features to represent Bali traditional dance gestures are skeleton joint features of: left/right foot and left/right elbow.

Keywords: *Bali traditional dance, dance gesture recognition.*

1. Introduction

Automatic recognition of dance gesture is one area of computer vision with many potential applications such as dance self-assessment and immersion computer game development. Given arbitrary human rhythmic motions, a dance gesture recognition system aims to detect and classify the dance gesture that was performed.

Bali traditional dance choreography is a traditional dance that has its origin in ritual ceremony performed in a *pura* or Hindu temple. In contrast to a contemporary dance in which underlying dance movements are free and only constrained by the limits of the body [1], the dance gesture of a Bali traditional dance have been kept unchanged and passed from generation to generation. Therefore, it is hypothesized that the dance gesture can be explained using a set of grammatical rules that captures body-part motions of the dance performer.

Based on its function within a ceremony, Bali traditional dances can be categorized into: (1) *wali* dance which have its origin in ritual and sacred dances; (2) *bebali* or semi-sacred dance are the dances which are not necessarily connected with ritual (e.g. *Legong* dance); and (3) *balih-balihan* dance which are public dances for entertainment purposes [2]. Although having some differences, many Bali traditional dances also share some common dance gestures.

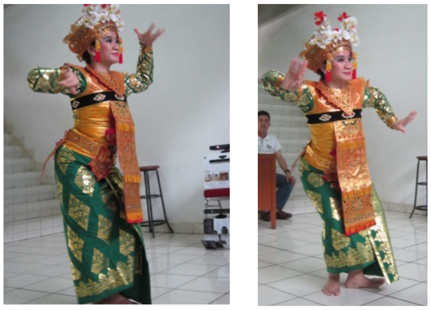

Figure 1. Two Poses of the *Legong* Dance Performance

The dance gesture in Bali traditional dance portrayes a legend character by means of articulated body-part motions including head, neck, eye, hand and feet; and face expression. With such articulated body-part motions, recognition and evaluation of dance gesture from Bali traditional dance are very challenging.

This study presents a method to recognize dance gestures from Bali traditional dance by means of probabilistic grammar. The dance gesture in this study is limited to those that represent common gestures of Bali traditional dances.

The remaining of this paper is organized as follows. Chapter 2 will describe some related works. Chapter 3 will explain the research methodology. Chapter 4 will show the results of the research followed by conclusion in chapter 5.

2. Related Works

Dance gesture recognition has gain wide attention from various research communities resulted in many published works. Many of those studies explored various visual features to represent dance gesture such as texture properties and Scale-Invariant Feature Transform (SIFT) extracted from 2D image [3], hand and feet movement trajectories [4], body-part movement trajectories [5,6], spherical coordinates of skeleton joints [7], and mixed of audiovisual and skeleton joint features [8,9]. Despite some methods have been proposed for dance performance evaluation, none of these method was based on Bali traditional dance.

2.1 Skeleton Descriptor

Recently skeleton feature has gain wide attention from computer vision researchers to represent human body-part motion thanks to the availability of the depth sensor camera. By using the depth sensor camera, the stream of skeleton feature descriptor can be estimated from data stream produced by the sensor camera.

The study by [7], for example, divides human skeleton joints into torso frame, first-degree joints, and second-degree joints. The first and second principal components of the torso frame are computed as the basis of coordinates of the other skeleton joints. Relative spherical coordinate (θ, φ) using torso PCA frame as the reference, θ denoted inclination and φ denoted azimuth, is computed from each joint to its neighboring joint to which it is directly connected (see Figure 2). The R component of the spherical coordinate is normalized so it can be ignored.

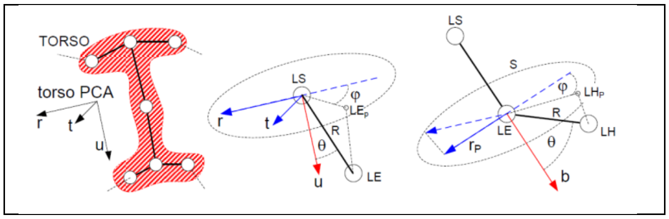

Figure 2. Skeleton Descriptor (Source: [7])

Finally, the skeleton joint descriptor is represented as a two dimensional time series of: (1) spherical coordinate (θ, φ) of the first-degree and the second-degree joints; and (2) rotation matrix parameters of the torso frame to the camera coordinate system.

2.1 Time Series Representation

A gesture element is typically represented by a long multiple time series. In order to speed-up computation, simplified time series is often preferable. Several algorithms have been proposed to simplify time series data such as Discrete Fourier Transformation [10], Discrete Wavelet Transformation [11], Piecewise Aggregate Approximation (PAA) [12], and Symbolic Aggregate Approximation (SAX) [13, 14]. In this study, the last two algorithms are combined to represent and simplify the time series that represents body motion over time. The PAA algorithm is used to discretize the time series using the mean value of each time series segment. The SAX algorithm is used to convert the value of time series in each segment into a string of symbolic labels based on boundary value of normalized Gaussian distribution. With these algorithms, the numeric time series is converted into a string of symbolic labels.

2.2 Dance Gesture Classification

Automatic recognition, description, and classification of dance gesture are important problem in many disciplines. Dance gesture classification aims to map each dance gesture into one of the predefined classes. Currently a plethora of gesture recognition methods can be divided broadly into (1) statistical pattern recognition, and (2) syntactical pattern recognition based methods. Each of those methods has comparable level of expressiveness and flexibility.

Probabilistic grammar induction is a syntactical pattern recognition method to model data sequence represented as a strings of symbols using a set of production rules. Given a set of examples, the task is to estimate the probabilistic grammar that produced the given examples. The prominent algorithm to induce probabilistic grammar from examples are Alergia [15]. By using this method, the grammar is chosen from a grammar class that produces regular languages. Given a string of symbols as examples, the grammar induction algorithm induces regular languages, which in turn will induce probabilistic finite state automata (DFA) corresponds to the probabilistic grammar that accepts the given examples.

To address weak classifier, the study by [16] has shown that a strong classifier can be built from relatively weak classifiers by building a cascaded classifier that forms a tree. By using the cascaded classifier, each successive classifier is evaluated only on the testing data which pass through the preceding classifiers. Hence, no further processing is performed if at any stage in the cascade a classifier rejects the tested data.

Another schematic to built a classifier proposed by Bourdev [17] that uses voting schema to make final decision that involves weak recognizers.

3. Research Methodology

3.1 Pattern Classes

The pattern classes in this study are a set of dance gesture from Bali traditional dance. Eleven dance gestures, as shown in the following table, are selected as the pattern classes for dance gesture recognition.

Table 1. The Dance Gesture Classes of Bali Traditional Dance

No	Dance Gesture Name	No	Dance Gesture Name
1	*Agem Kanan*	7	*Ngegol*
2	*Agem Kiri*	8	*Ulap-ulap Kanan*
3	*Piles*	9	*Ulap-ulap Kiri*
4	*Ngeseh*	10	*Mungkahlawang*
5	*Luk Nerudut*	11	*Nayog*
6	*Malpal*		

These gesture classes are selected purposively from a set of common dance gestures of Bali traditional dances.

3.2 Data Collection

The dance gestures performed by the Bali traditional dancer are captured using a static mounted depth sensor that produces skeleton coordinates at the rate of 30 fps and recorded in an ONI file format (see Figure 3). The resolution of the depth map is 320×240 and resolution of the RGB image is 640×480. In this study, the total of 63 samples over 11 basic dances was recorded. Data recording starts with initial step to calibrate the sensor (the dancer in the position of body stand up straight and both hands up). The length of the dance gestures range from 100 to 200 frames. For simplicity reason, each gesture is recorded without Bali traditional music.

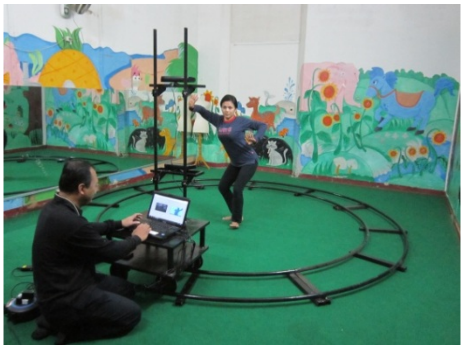

Figure 3. Equipment Setting for Data Acquisition

3.3 Feature Extraction

The extracted features in this study are similar to the features proposed by [7]. Initially, as many as 15 skeleton joint coordinates are extracted from the output data stream of the kinect. For simplicity reason, head skeleton joint is excluded from analysis. The next six skeleton joints form a frame called torso-frame from which principal components are extracted. The principal components (PCA) of the torso frame are used to span a 3-dimensional space into which the coordinate of the remaining skeleton joint (left/right elbow, left/right hand, left/right knee, and left/right foot joints) are mapped in order to achieve joint coordinate independency from the kinect position from the targeted object.

The skeleton feature descriptors are then represented as a parameter time series of spherical coordinates and the torso frame rotation matrix as follows:

$$\{(\theta_t^i, \varphi_t^i) | 1 \le i \le 8\} \cup \{(\psi_t, \alpha_t, \pi_t)\} \qquad (1)$$

where: θ_t^i denotes inclination and φ_t^i denoted azimuth of joint spherical coordinates; and $(\psi_t, \alpha_t, \pi_t)$ denotes rotation matrix of torso-frame principal component to the camera coordinate system that was parameterized by Tait-Bryan angles.

In contrast to the study by [7], that uses the whole set of extracted skeleton feature descriptors, this study only uses a subset of these skeleton feature descriptor that are highly discriminative. The skeleton features to represent the dance gesture are selected by means of clustering analysis. Given a skeleton feature, first, the dance gesture examples are clustered using hierarchical clustering method. Next, clustering performance is measured using Cophenet Correlation Coefficient (CCC) [21]. The CCC measures performance of a hierarchical cluster tree (dendrogram) in preserving the pairwise distances between the original data points. High CCC value for a given fature can be interpreted as a measures of the given feature to divide the dance gesture examples into clusters such that similarity between examples in a cluster is higher that those from two different clusters. Finally, the skeleton feature descriptors are selected from the top list that give the high CCC values.

3.4 Pattern Representation

The skeleton feature sequence data is extracted from kinect output data stream and converted into a string of symbols using SAX discretization method [15] as follows. Each time series is divided into a number of segment of 10 lengths. The SAX algorithm then examines the normalized Gaussian distribution of values and divides the distribution into 8 equal parts, each part is represented by a label set of $\{a, b, c, d, e, f, g, h\}$. Given time series x_t, the time series label L_{x_t} is computed from each segment based on the average value of the segment using the following formula:

$$L_{x_t} = \begin{cases} a, & \bar{x}_t < -1.15 \\ b, & -1.15 \le \bar{x}_t < -0.67 \\ c, & -0.67 \le \bar{x}_t < -0.32 \\ d, & -0.32 \le \bar{x}_t < 0 \\ e, & 0 \le \bar{x}_t < 0.32 \\ f, & 0.32 \le \bar{x}_t < 0.67 \\ g, & 0.67 \le \bar{x}_t < 1.15 \\ h, & 1.15 \le \bar{x}_t \end{cases} \qquad (2)$$

3.5 Feature Selection

The feature selection aims to select the most discriminative feature to represent the dance gestures. In this study, the feature selection is implemented by means of data clustering. The selected features are among those that give the highest Cophenetic Correlation Coefficient (CCC) of the data clusters among the tested features. The CCC, c, is defined by [18] as follows.

$$c = \frac{\sum_{i<j}(Y_{ij}-\bar{y})(Z_{ij}-\bar{z})}{\sqrt{\sum_{i<j}\left((Y_{ij}-\bar{y})^2(Z_{ij}-\bar{z})^2\right)}} \qquad (3)$$

where: Yij denotes a similarity distance between time series-i and time series-j; $\bar{y}$ and $\bar{z}$ denote mean of Yij and Zij respectively; Zij denotes hierarchical binary tree matrix that contains information about leaf node-i, leaf node-j, and the distance between the node-i and the node-j in a hierarchical binary tree. The CCC measures performance of a dendrogram in preserving the pairwise distances between the original data points.

3.5 Inducing Dance Gesture Grammar

Prior to training dance gesture classifier, training dataset synthesis is implemented using the observable examples as the seed. The aim of this step is to capture as much as possible unforeseen variation of dance gestures base on the gesture example at hands. It is assumed that two dance gestures from the same class can be represented by time series with different length. In order to have time series which vary in time dimension, dynamic time warping algorithm is adopted. From a pair of time series, the dynamic time warping algorithm can return another pair of time series with vary in time dimension.

Given a set of strings that represent dance gesture examples of a gesture element feature, the Alergia method is used to learn underlying grammar that explains the testing dataset. The final decision of dance gesture recognition is implemented using voting technique.

3.6 Cross Validation

The dance gesture recognizer is evaluated using hold-out cross-validation. Each of the dance gesture recognizers is trained using examples that are divided into 2 parts: 80% of the whole examples as training dataset, and the last 20% of the examples as testing dataset.

4. Experimental Results

4.1 Feature Selection

Hierarchical clustering of the dance gesture examples that used all skeleton joint features gave the following result.

Table 2. Cophenet Correlation Coefficient (CCC) of Hierarchical Clustering

Feature	CCC	Feature	CCC
L-Foot θ(LFT)	**0.60**	R-Knee θ (RKT)	0.53
L-Foot φ (LFP)	**0.57**	R-Knee φ (RKP)	0.53
R-Foot θ (RFT)	**0.56**	R-Hand φ (RHP)	0.52
R-Foot φ(RFP)	**0.55**	L-Hand φ (LHP)	0.50
L-Elbow φ (LEP)	**0.55**	L-Knee φ (LKP)	0.50
R-Elbow φ (REP)	**0.55**	L-Hand θ (LHT)	0.50
L-Elbow θ (LET)	**0.54**	R-Elbow θ (RET)	0.49
L-Knee θ (LKT)	0.53	R-Hand θ (RHT)	0.48

Note: L- (Left-), R-(Right).

The above results showed that the most discriminative skeleton joint feature for data clustering descriptors were: (1) Left Foot: LFT, LFP; (2) Right Foot: RFT, RFP; (3) Left Elbow: LEP, LET; and (4) Right Elbow: REP. Interestingly, hand and elbow skeleton features were less discriminative in compare to the other tested features.

4.2 Inducing Dance Gesture Grammars

From the experiment, it was discovered that an induced probabilistic grammar from single dance gesture examples failed to recognize an unseen dance gesture from the respected dance gesture. To address this problem, a probabilistic grammar is induced using examples from a pair of dance gestures. Given dance gestures K_i and $K_j(1 \le i \le 11,\ i+1 \le j \le 11)$, a probabilistic grammar G_m were induced using examples training examples that can be categorized into:

1) Original examples comprises of time series $t_i \in K_i$ and $t_j \in K_j$

2) Synthetic examples as the result of time warping between time series: $t_i \in K_i$ and $t_j \in K_j$

Final decision for dance gesture recognition was implemented using voting technique. Performance of the dance gesture recognizer was evaluated using Precision metric [19] such that:

$$Precision = \frac{The\ number\ of\ correct\ estimated\ class}{The\ total\ number\ of\ class\ estimation} \quad (4)$$

The result of cross-validation, as can be seen in the following table, showed that the skeleton features extracted from performer's elbow and feet have high discriminative power for recognizing dance gesture from Bali traditional dance.

The dance gesture recognizer showed low performance for recognizing the *Agem-kanan, Agem-kiri and Piles* dance gestures.

Table 3. Average Precision of the Dance Gesture Recognizer by Feature

	Feature						
	LET	LEP	REP	LFT	LFP	RFT	RFP
Mean of AP	**0.91**	**0.91**	**0.91**	**0.95**	0.82	**0.91**	**1.00**

Table 4. Average Precision of the Dance Gesture Recognizer by Dance Gesture Class

DanceGesture Class	AP
Agem-kanan	0.79
Agem-kiri	0.79
Piles	0.79
Ngeseh	**0.93**
Luk nerudut	0.86
Malpal	**1.00**
Ngegol	**1.00**
Ulap-ulap kanan	**0.93**
Ulap-ulap kiri	**1.00**
Mungkah-lawang	**1.00**
Nayog	**1.00**
Mean of AP	0.92

On the other hand, the dance gesture recognizer showed high performance to recognize the following dance gesture classes: *malpal, ngegol, ulap-ulap kiri, mungkahlawang* and *nayog*.

The low recognition result of the these dance gestures perhaps was due to its high articulated body-part motions; on the other hand, variations that were captured in the dance gesture examples are reather limited.

5. Conclusions

This paper has shown some empiric results that dance gesture from Bali traditional dance can be modeled using a probabilistic grammar. As a dance gesture recognizer, the probabilistic grammar induced from the dance gesture examples achieved high performance for recognizing the predefined set of dance gestures from Bali traditional dance. The experimental results also showed that the skeleton features extracted from performer's elbow and feet are features with the most discriminative power to represent the dance gesture from Bali traditional dance.

The next step of this research is to validate the result of this research by using more samples and variety of dance gesture of Bali traditional dance.

Acknowledgments

This work is funded by the Ministry of Research and Technology, Republic of Indonesia, under the National Synergy Research Grant SINas, contract No. RT-2012-1117.

References

[1] Herbison-Evans D. "Dance and the computer: a potential for graphic synergy," Technical Report 422, Basser Department of Computer Science, University of Sydney, Australia, 1991.

[2] Hobart, M. "Rethinking Balinese Dance," in Indonesia and the Malay World, Vol. 35, No. 101, pp.107-128, 2007.

[3] Hassan, E., S. Chaudhury, and M. Gopal. "Annotating Dance Posture Images using Multi Kernel Feature Combination," in Proceeding of 2011 Third National Conference on Computer Vision, Pattermn Recognition, Image Processing and Graphics, pp. 41-45, 2011.

[4] Min, J. and R. Kasturi."Activity Recognition Based on Multiple Motion Trajectories," In Proceeding of the Pattern Recognition, 17th International Conference on (ICPR'04), Vol.4, pp. 199- 202, 2004..

[5] Boukir, S. and F. Cheneviere. "Compression and recognition of dance gestures using a deformable model," Pattern Anal Applic (2004) 7: 308–316 DOI 10.1007/s10044-004-0228-z.

[6] Ciglar, M. "A Full-body Gesture Recognition System and its Integration in the Composition '3rd. Pole'", in the Proceeding of ICMC 2008.

[7] Raptis, M., D. Kirovski, H. Hoppe. "Real-Time Classification of Dance Gestures from Skeleton Animation," Symposium on Computer Animation, pp. 147-156, 2011.

[8] Gowing, M., P. Kell, N. E. O'Connor, C. Concolato, S. Essid, J. Lefeuvre, R. Tournemenne, E. Izquierdo, V. Kitanovski, X. Lin, and Q. Zhang. "Enhanced Visualisation of Dance Performance from Automatically Synchronised Multimodal Recordings," In Proceedings of the 19th ACM international conference on Multimedia (MM '11), pp. 667-670, 2011.

[9] Essid, S., D. Alexiadisy, R. Tournemenne, M. Gowingz, P. Kellyz, D. Monaghanz, P. Darasy, A. Dremeau, and N. E. O'Connor. "An Advance Virtual Dance Performance Evaluator," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2012.

[10] Agrawal, R., C. Faloutsos, and A. N. Swami, "Efficient similarity search in sequence databases," in Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms, ser. FODO '93. London, UK: Springer-Verlag, 1993, pp. 69–84.

[11] Chan, K. and A. Fu, "Efficient time series matching by wavelets," in Proceedings of the 15th International Conference on Data

Engineering, ser. ICDE ’99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 126–.

[12] Keogh, E. “A decade of progress in indexing and mining large time series databases,” in Proceedings of the 32nd international conference on Very large data bases, ser. VLDB ’06, 2006, pp. 1268–1268.

[13] Lin, J., E. Keogh, L. Wei, and S. Lonardi, “Experiencing sax: a novel symbolic representation of time series,” Data Min. Knowl. Discov., Vol. 15, pp. 107–144, October 2007.

[14] Ordonez, P., T. Armstrong, T. Oates, and J. Fackler. “Classification of Patients using Novel Multivariate Time Series Representations of Physiological Data,” in Proceeding 2011 10th International Conference on Machine Learning and Applications,” pp. 172-179, 2011.

[15] Carrasco, R.C.. and J. Oncina. “Learning Stochastic Regular Grammar by Means of a State Merging Method,” in Proceeding ICGI-94, Springer, pp. 139-150, 1994.

[16] Viola, P. and M. Jones. “Robust Real-time Face Detection,” in International Journal of Computer Vision, Vol. 57, No. 2, pp. 137–154, 2004.

[17] Bourdev, L. “Poselets and Their Applications in High-Level Computer Vision,” PhD Thesis, University of California, Berkeley, USA,2011.

[18] Sokal, R. R. and F. J. Rohlf. “The comparison of dendrograms by objective methods. Taxon, Vol. 11, No.2, pp. 33-40, 1962.

[19] Baeza-Yates, R. and B. Ribeiro-Neto. “Modern Information Retrieval,” Addison-Wesley, Harlow, UK, 1999.

Yaya Heryadi is a researcher and lecturer at School of Computer Science, Bina Nusantara University. He got a Sarjana in Statistics and Computation from Bogor Agricultural University (IPB) in 1984, Bogor; Master of Science degree in Computer Science from Indiana University at Bloomington (IUB), Indiana, USA, in 1989. His research interests include computer vision, and image processing.

Dr. Mohamad Ivan Fanany is a researcher and lecturer at Faculty of Computer Science, University of Indonesia. His research interests include imaging science and engineering, 3D perception, reconstruction, recognition, and data mining. He got a degree in Physics at the Faculty of Mathematics and Natural Science in 1995. Before joining the faculty, he worked at Future Project Div. Toyota Motor Corp, Japan, as a member of middleware development and recognition team; NHK Engineering Services Inc., as a researcher of IT21 Millennium Project on Advanced High Resolution and Highly Sensible Presence 3D Content Creation funded by NICT Japan; and a JSPS Fellow and Research Assistant at Imaging Science and Engineering, Graduate School of Information Science and Engineering, Tokyo Institute of Technology (Titech). He served as the Chairman of Titech IEEE student branch 2002-2003 and member of IAPR, IEEE, and ACM SIGGRAPH.

Dr. Aniati Murni Arymurthy, Sarjana degree in Electrical Engineering from University of Indonesia; Master of Science degree in Computer Science from Ohio State University; and Doctor of Computer Science from University of Indonesia in 1997; and Professor at Faculty of Computer Science University of Indonesia.