Stochastic Regular Grammar-based Learning for Basic Dance Motion Recognition

Yaya Heryadi[1]
[1]*School of Computer Science*
Binus International – Binus University
Jakarta, Indonesia
Email: [1]yayaheryadi@binus.edu

Mohamad Ivan Fanany[2], Aniati Murni Arymurthy[3]
[2,3]*Fakultas Ilmu Komputer*
Universitas Indonesia
Depok, Indonesia
Email: [2]ivan@cs.ui.ac.id, [3]aniati@cs.ui.ac.id

Abstract**—this paper presents a simple and computationally efficient framework for 3D dance basic motion recognition based on syntactic pattern recognition. In this research, a class of basic dance motions is modeled by a stochastic regular grammar (SRG), inferred from training dataset, in which key body poses that are learned from training dataset are selected as gesture primitives. To represent a dance motion, body pose of a dancer is represented by angular coordinate of 15 skeleton joints. This feature is later compacted into one-dimensional string of labels for grammar inference which makes the recognition process is considerably fast compared to statistical pattern classifier such as k-nearest neighbor (kNN). A single test using the learned grammar in average takes only about 5 ms compared to around 20s using kNN whilst the overhead time to build all grammars takes only about 3.4s. This compacting process, however, leads to information loss which is observed in slightly degraded recognition performance for low articulated motions but quite large degradation for high articulated dance motions. To overcome this, we investigate several reliable feature selection methods such as Sequential Feature Selection (SFS), Principal Component Analysis (PCA), and Heuristic Sequential Feature Selection (HSFS) compared to the use of whole features. Based on our experiment, the HSFS is the most suitable feature selection to overcome this problem.**

Keywords—dance motion recognition, syntactic model

I. Introduction

Dance is a highly articulated motion and considered as an important subset of non-rigid motion where the object of interest is composed of several rigid components connected to each other by ball and hinge joints. Dance motion recognition is a challenging problem that has received considerable attention from the computer vision community in recent years. Its applications are diverse, spanning from its use in game and interactive animation to digital archiving of cultural heritage.

The challenges in solving this problem, however, are multifold due to complexity of human motions thanks to the flexibility of human body, the spatial and temporal variations exhibited due to differences in duration of different actions performed, and the changing spatial characteristics of the human form in performing each action. One key difference between human regular motions and dance motions is the existence of structure in which a motion pattern composes of sub-pattern (basic motion) in a hierarchical structure.

Related works. Dance motion is human expression using aesthetic and rhythmic body part motion. The feature that dominates a dance motion is the repetitive pattern of skeleton joint coordinates. Most of previous studies to motion recognition concentrated on the use of statistical pattern recognition [1,2,3]. The study by [4] demonstrated that syntactic pattern recognition has become a promising alternative approach to motion recognition. In this approach, a motion is hypothesized contains as a structure of sub-patterns that occur repeatedly according to particular rules. In their report, [4] further proposed a method to recognize daily human motion activities using syntactic approach based on upper-body skeleton features. The inputs for motion recognition are images from video data sequence. The used motion features are 3D coordinate of 8 nodes: 2 head nodes and 6 arm nodes that are tracked using adds on devices attached to corresponding part of the body.

Contribution. The aim of this research is to build an efficient method to recognize dance basic motion using syntactic pattern recognition approach. In contrast to the previous study with similar approach [4], the inputs of this research are depth image captured using a single Microsoft kinect depth sensor camera positioned in front of a dance performer. In [4], the transition of the grammar is predefined a priori and transition probability of the grammar is computed directly from examples. Primitives of the grammar are direction of body part movements which are determined a priori namely: left, right, up, and down. This study is also different from other dance motion recognition reported by [3], which uses basically statistical approach that is based on

Manuscript received July 18, 2013. 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 and No. KP-2013-1679

separate joints that are combined later. In addition, the relation between two motions is measured using cross-correlation coefficient of skeleton features extracted from 7 skeleton joints.

The basic idea of our proposed model is to divide each dance basic motion into a number of key pose units as primitive gestures so that each dance basic motion can be represented using syntactic model according to a particular grammar (see Figure 1). Two types of syntactic-based classifiers are built from the training motion sets. The main classifier is probabilistic deterministic finite automata which is designed to take advantage of structure contains in a motion in measuring similarity between training motions and unseen motion. The syntactic-based classifier is trained using Alergia algorithm. For comparison, a statistical-based classifier is also designed to measure similarity between training and unseen motion based on a particular distance function. Although both classifiers can recognize a class of unseen motions, the probabilistic deterministic finite automata classifier appears to be more appealing than the later as it is capable not only to recognize but also to generate a correct sequence of basic key body poses in a dance.

This paper is organized as follows. Chapter 2 describes stochastic regular grammar that is used for motion modeling, representation and recognition. Chapter 3 explains experimental results and discussion. Finally, Chapter 4 draws up conclusions including future study and development.

II. Motion Modeling using SRG

Syntactic pattern recognition comprises of two main steps [5]. First, it determines a set of primitives. For motion recognition, it is natural to view human body key pose as primitive of a motion for two reasons: (1) a motion can be decomposed into a set of body poses in a structure formed by concatenation relation; and (2) the key poses can be easily constructed using clustering algorithm to all body poses from training motions so that the set of key poses comprises of a number of distinguishable poses. Second, it infers grammar to describe the motion patterns. For this purpose, Alergia is a prominent algorithm [6] to infer probabilistic grammar for a motion class. This algorithm infers probabilistic grammar from positive examples using state-merging technique on probabilistic criteria. This grammar inference algorithm has been successfully applied in music genre recognition [7], normal or abnormal chromosome recognition [8], digit and shape recognition [9], and text and speech recognition [10,11].

Syntactic pattern recognition works with string of symbols. In order to adopt Alergia algorithm; therefore, each key pose should be given a label so that a motion can be represented as a string of label. Alergia algorithm is a grammar inference algorithm that takes only positive examples. The grammar inference comprises of three main steps [6]. First, constructing a prefix tree automaton (PTA) to recognize all prefixes found in the training examples. Each transition of the PTA is set to a probability corresponded to the number of times it is traversed during the PTA construction. Second, merging pair of nodes in a lexicographical order. Two nodes are merged if the resulting automaton is considered equivalent to the PTA. The state merging is computed iteratively until no further merging is possible. Finally, generating a weighted deterministic finite-state automaton (DPFA). The DPFA was then applied to capture all the strings found in the training examples as well as strings that represent gesture that were not part of the training examples. Alergia infers deterministic probabilistic finite state automaton that corresponds to stochastic regular grammar.

A deterministic probabilistic finite-state automaton (DPFA) is defined as 6-tuple $A=(Q, \Sigma, \delta, I, F, P)$ where Q is a finite set of states, Σ is the alphabet, δ is transition relation such that $\delta \subseteq Q\times\Sigma\times Q$, I is the initial state probabilities where $\exists\ q_0 \in Q$ such that $I(q_0) = 1$ where q_0 is the initial state, $\forall q \in Q$, $\forall\ a \in \Sigma$, $|\{q': (q,a,q') \in \delta\}| \leq 1$. F is the final-state probabilities such that $F: Q \rightarrow R+$, and P is the transition probabilities such that $P: \delta \rightarrow R+$ [12].

A stochastic regular grammar (SRG) is defined as a 4-tuple $G=(V_N, V_T, P, S, F)$ where V_N and V_T are finite sets of nonterminal and terminals, $S \in V_N$ is the start symbol, and P is a finite set of production rule each of which is in the form: $P \subseteq V_N \times (V_T V_N \cup V_T)$ and $F: P \rightarrow [0:1]$ which assigns probability to production rules [13].

Deterministic probabilistic finite-state automata and stochastic regular grammar are equivalent. As describe in detail in [5], for a given deterministic probabilistic finite-state automata, it is possible to construct an equivalent stochastic regular grammar.

Theoretical strength and weakness of Alergia algorithm for grammar inference is as follows. Alergia algorithm allows SRG inference based on positive examples only with reasonably fast training time; however, learning grammar from positive examples only brings about several shortcomings [14] such as: the risk of over-generalization in which learned languages are larger than the target language, and regular languages are theoretically un-learnable from positive information alone.

A. Data Collection and Definition of Pattern Classes

Observation data of dance motions are captured using a static mounted depth sensor camera positioned in front of the performer that produces skeleton coordinates at the rate of 30 fps. Data recording are started with depth sensor calibration to track initial position of the dancer skeleton joints. The pattern classes used in this study are some basic gestures from Balinesse Pendet traditional dance. The data consist of 75,197 poses which is segmented into 705 dance motion segments that are annotated into 9 classes of dance motions.

B. Feature Extraction and Pre-processing

Following [3] skeleton joint features are extracted from depth images that are captured using a depth sensor camera. The eight joints are namely: left/right elbow, left/right hand, left/right knee, and left/right foot joints. Each feature is then represented by parameter of spherical coordinate as follows $\{(\theta_t^i, \varphi_t^i) \mid 1 \leq i \leq 8\}$ where: $\{(\theta_t^i, \varphi_t^i)$ for i=1,2,…,8 denotes spherical coordinates of the 8 joints such that θ denoted inclination and φ denoted azimuth which uses neck as coordinate reference.

In attempt to find optimum feature for dance motion representation, some feature selection algorithms are adopted namely: Principal Component Analysis (PCA) [15], Sequential Feature Selection (SFS) [16] and Heuristic Sequential Feature Selection (HSFS). In order to improve recognition accuracy of the motion model, two dimensional reduction methods are adopted. In SFS method a subset of joints are selected from set of joints that best predict the data in training dataset. The SFS method works by sequentially selecting joint until there is no improvement in prediction. For simplicity, we use forward selection approach.

On the other hand, in the HSFS method, joint skeleton of human body can be viewed as a structure such that each joint have a parent, except the root parent. In this structure, position of each child joint is influenced by its parent joint position. In HSFS, we consider that end point of joint skeleton are very important since they are highly influenced by the parent-of-parent joint movement. The term heuristic in HSFS refers to the approach in feature selection such that the end point joints are kept and then do sequential forward feature selection.

C. Building Key Pose Vocabulary

Kmeans algorithm is used to cluster data observation into a group of key poses. Having the key poses been estimated, an instance of skeleton joint features are given label by the closest key poses based on Euclidean distance (see Figure 2). Therefore, dance motion is then represented as a string of cluster labels. String representation of motion is then used to build the motion model both using the syntactic (Grammar inference) and statistical approach (kNN).

D. Classifier Training and Validation

In this study, model of the dance motion are built using probabilistic grammar whereas 1NN and 3NN classifier are used for comparison. The former approach is adopted to solve classification problem under hypothesis that the observation data contains a structure. The kNN algorithm is adopted to map observation data into a predefined set of classes using statistical pattern recognition approach. Various methods are proposed to improve the computational cost of finding the nearest neighbour such as employing K-D tree or using condensed NN [17], but currently we compare our method only to the conventional kNN using Dynamic Time Warping (DTW) as distance function. DTW has also been used for dance similarity matching in [3].

The model cross validation is implemented using 5-fold technique and aimed to gain preliminary information to lead further steps in efforts to find the best model formulation. The performance of a gesture classifier is measured using accuracy metric defined as the number of true positive/the total number detected gestures.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. Impact of the Number of Clusters

The impact of the number of clusters (c) to dance motion recognition performance measured by average accuracy is shown in the Figure 3 and 4. The table below showed classification performance for c ∈ {5,10,15,20,25,30,35} gained by two models as follows where c is the number of clusters.

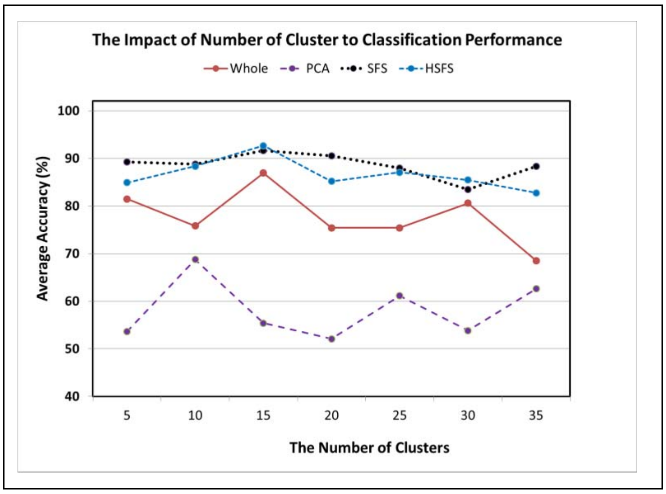


Figure 3. The impact of number of clusters to classification performance using Alergia

The figure above shows that SFS and HSFS can select the most discriminative features for motion classification using Alergia algorithm. In contrast, the whole features and selected features using PCA achieved lower recognition accuracy than previous selected features.

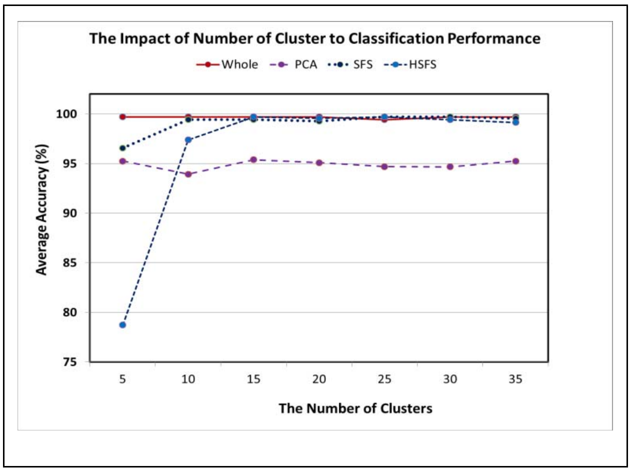


Figure 4. The impact of number of clusters to classification performance using 3NN

It can also be seen in Figure 4, the 3NN algorithm achieved higher average classification performance despite any feature algorithm adopted. The only factor that causes conventional kNN less appealing is the speed of testing phase. While the computational time complexity for testing phase of probabilistic regular grammar is linear to the length of testing string, i.e.,O(n), the complexity for testing phase of conventional kNN is O(n^2). The slower model testing of kNN is due to high number of pairwise distance comparison involving all examples; while, the model testing of Alergia is only involve examples of the same class. Interestingly, both approach gain maximum performance for the number of cluster c = 15.

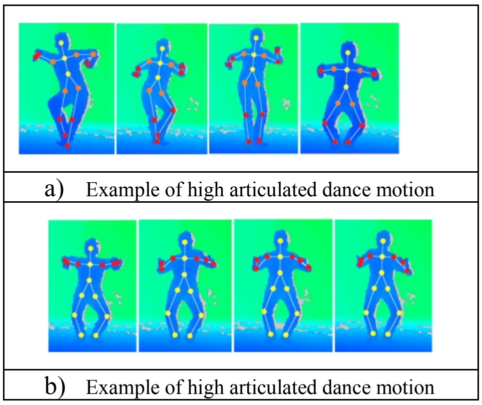
B. *Impact of the Classification Method*

Further analysis to recognition performance of each class of gesture is as follows. The motion recognition accuracy by motion class using stochastic regular grammar and kNN algorithm is shown in Table 1 below.

TABLE 1.
COMPARISON OF MOTION RECOGNITION PERFORMANCE USING ALERGIA (C=15)

Dance Motion	Recognition Accuracy (%)			
	whole	PCA	SFS*	HSFS**
Agem Kanan	100.0	92.2	100.0	100.0
Agem Kiri	100.0	59.3	100.0	100.0
Piles	100.0	81.1	100.0	96.8
Ngeseh	70.5	7.5	83.5	93.5
Luk Nerudut	100.0	91.7	97.3	97.3
Malpal	59.8	13.5	73.3	90.7
Ngegol	100.0	47.6	88.8	88.8
Mungkahlawang	59.2	34.8	91.3	74.1
Nayog	86.7	73.3	93.3	93.3
Average	86.9	55.4	91.7	92.7

Note: (*) The selected features are skeleton joint of: hand (left), elbow (right), shoulder (left), foot (right), knee (left). (**) The selected features are skeleton joint of: hands (left, right), feet (left, right), and knees (left, right).



a) Example of high articulated dance motion

b) Example of high articulated dance motion

Figure 5. Comparison of skeleton feature between high and low articulated motion

Ngeseh, Malpal, Mungkahlawang, and *Nayog* are some basic dance motions in Balinese Pendet traditional dance with low body pose variations. Figure 3 shows comparison of skeleton feature between high and low articulated motion where the red joints denote very high articulated joints, orange joints denote high articulated joints, and yellow joints denote low articulated joints. Representing an instance of skeleton joint features that are originally composed of 14 angles with the closest cluster centroid causes false representation. For example, a motion is represented by one/two long segments of similar symbols which causes pose variations in the original motion undetected. False representation of low articulated motions has made the trained classifier quite sensitive to the difference of motion representation as can be seen in Figure 5 below.

TABLE 2.
COMPARISON OF MOTION RECOGNITION PERFORMANCE USING 3NN (C=15)

Dance Motion	Recognition Accuracy (%)			
	whole	PCA	SFS*	HSFS**
Agem Kanan	100.0	100.0	100.0	100.0
Agem Kiri	100.0	96.0	100.0	100.0
Piles	97.9	95.8	98.0	100.0
Ngeseh	100.0	91.2	100.0	100.0
Luk Nerudut	100.0	98.1	100.0	100.0
Malpal	100.0	94.7	97.3	97.1
Ngegol	100.0	95.0	100.0	100.0
Mungkahlawang	100.0	89.7	100.0	100.0
Nayog	100.0	100.0	100.0	100.0
Average	99.7	95.4	99.4	99.7

Note: (*) the selected features are skeleton joint of: hands (left), elbow (right), shoulder (left), feet (right), and knee (left). (**) the selected features are skeleton joint of: hands (left, right), feet (left, right), and knees (left, right).

In general, average recognition accuracy of probability grammar is improved by selecting feature using SFS and HSFS. Probability grammar using PCA achieved the lowest recognition accuracy as PCA only reduces dimension but does not maintain the original class. The Table 2 shows that kNN achieves high recognition accuracy despite adopted feature selection algorithm.

IV. CONCLUSION

In this paper we pursue an efficient framework for recognizing basic dance motion using stochastic regular grammar. The current result is promising which show that for low articulated dance motion, the recognition performance is comparable to statistical recognition. The most appealing factor is its fast testing with only small overhead of building all grammars. Degradation in performance caused by loss of information using the stochastic regular grammar are tried to be alleviated by automatically selecting the best skeleton joints as features. A Heuristic Sequential Feature Selection (HSFS),which is based on the notion that end point of joint skeleton are very important since they are highly influenced by the parent-of-parent joint movement, can give significant recognition performance. For future study, we plan to move forward from the current dance basic motion recognition into dance motion analysis. By combining with automatic dance video segmentation method, the proposed method will be integrated to extract relation among each individual dance so that we are able automatically choreography and recognize higher level abstraction of dance such as dance modifications and compound dance analysis. In addition, further elaboration on dance sub-pattern should also be investigated.

REFERENCES

[1] A. Nakazawa, S. Nakaoka, K. Ikeuchi, and K. Yokoi. Imitating Human Dance Motions through Motion Structure Analysis, 2002. Proceeding of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, EPFL, Lausanne, Switzerland.

[2] T. Shiratori, A. Nakazawa, and K. Ikeuchi. Detecting Dance Motion Structure Using Motion Capture and Musical Information, 2004. In Proc. International Conference on Virtual Systems and Multimedia (VSMM) (Vol. 3).

[3] M. Raptis, D. Kirovski, and H. Hoppe. Real-Time Classification of Dance Gestures from Skeleton Animation, 2011. Eurographics/ ACM SIGGRAPH Symposium on Computer Animation.

[4] K. Cho, H. Cho, and Kyhyun Um. Human Action Recognition by Inference of Stochastic Regular Grammars, 2004. Structural, Syntactic, and Statistical Pattern Recognition, Lecture Notes in Computer Science 3138, Springer Science + Business Media, Inc.

[5] K. S. Fu. Statistic Pattern Recognition and Application, 1982. Prentice-Hall, Inc.

[6] R. C. Carrasco, and J. Oncina. Learning deterministic regular grammars from stochastic samples in polynomial time, 1999. Theoretical Informatics and Applications, Vol. 33(1), pp. 1-20.

[7] Cruz-Alcazar, P.P. and Vidal-Ruiz, E.: 'Modeling musical style using grammatical inference techniques: a tool for classifying and generating melodies', Third International Conference on Web Delivering of Music, (2003) 77-84.

[8] Fu, K.S.: Syntactic Methods in Pattern Recognition, Academic Press (1974) 54-55, 124-229

[9] Vidal, E., 'Application of the error-correcting grammatical inference algorithm (ECGI) to planar shape recognition', IEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives, (1993) 24/1 - 24/10.

[10] Atwell, E., et al, 'Multi-level disambiguation grammar inferred from English corpus, treebank,and dictionary', IEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives, (1993) 9/1 - 9/7

[11] Galiano, I. and Segarra, E., 'The application of k-testable languages in the strict sense to phone recognition in automatic speech recognition', IEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives, (1993) 22/1 - 22/7

[12] E. Vidal, F, Thollard, C. De La Higuera, F. Casacuberta, and R.C. Carrasco. Probabilistic finite-state machines-part I, 2005. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27(7), pp. 1013-1025.

[13] M. Schwem and A. Ost. Inference of Stochastic Regular Grammars by Massively Parallel Genetic Algorithms, 1995. In Proceedings of the Sixth International Conference on Genetic Algorithms. Morgan Kaufmann Publishers, Inc., San Francisco, CA.

[14] A. Stevenson and J. R. Cordy. Grammatical Inference in Software Engineering: An Overview of the State of the Art, 2013. In Software Language Engineering (pp. 204-223). Springer Berlin Heidelberg.

[15] K. Pearson. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space, 1901. Philosophical Magazine 2 (11): 559–572

[16] J. Doak. An evaluation of feature selection methods and their application to computer security (Technical Report CSE-92-18), 1992. Department of Computer Science, University of California, Davis.

[17] S. Gregory, D. Trevor, and I. Piotr. Nearest Neighbor Methods in Learning and Vision: Theory and Practice, 2006. Neural Information Processing, ISBN 026219547X, The MIT Press.

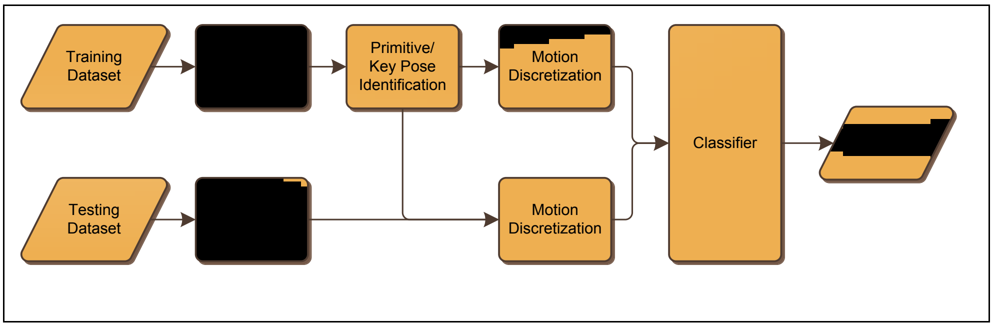


Figure 1. Process Flow Diagram of the Proposed Method

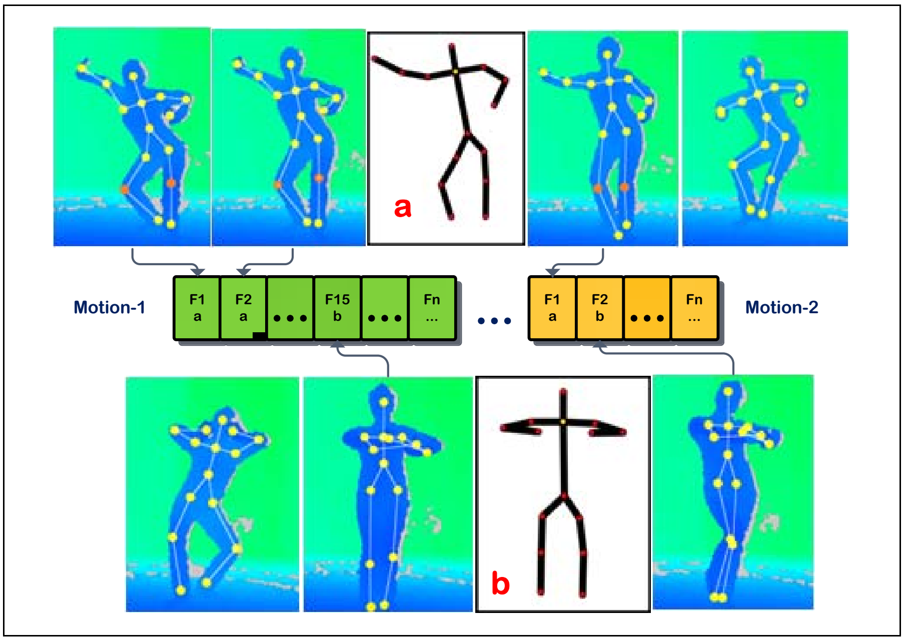


Figure 2. Cluster's Centroid as Dance Gesture Primitive