Dance Modelling, Learning and Recognition System of Aceh Traditional Dance based on Hidden Markov Model

N. Anbarsanti, Ary S. Prihatmanto
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia

Outline

1. Background
   * Problem
   * The Solution
2. Hidden Markov Model
3. Modelling the Whole Dance
4. Learning and Recognition System of 6 gesture classes
5. Data Collection and Segmentation
6. Skeleton Representation
   * Human Pose Representation
   * Pose Vector Quantization based on Range of Movement
7. Learning using Baum Welch Algorithm
8. Classification using Forward Algorithm
9. Effect of using Skeleton Representation and Max-likelihood Thresholding
10. Dance Phrase Recognition
11. Realtime Implementation
12. Conclusion
    * Suggestion

Problems

* Culture is the identity of a civilization
* Our duty to maintaining the Indonesian culture
* Globalization hits today's youth,
  * Lose their identity
  * Lose the good role models
  * Lose the value of good wisdom Indonesian culture.
* If the Indonesian culture is not well maintained
  * It will disappear in the next few decades.

How to present Indonesian Culture interactively?

The Solution

* Integrating the traditional dance with technologies

* Unique and Attractive dance :

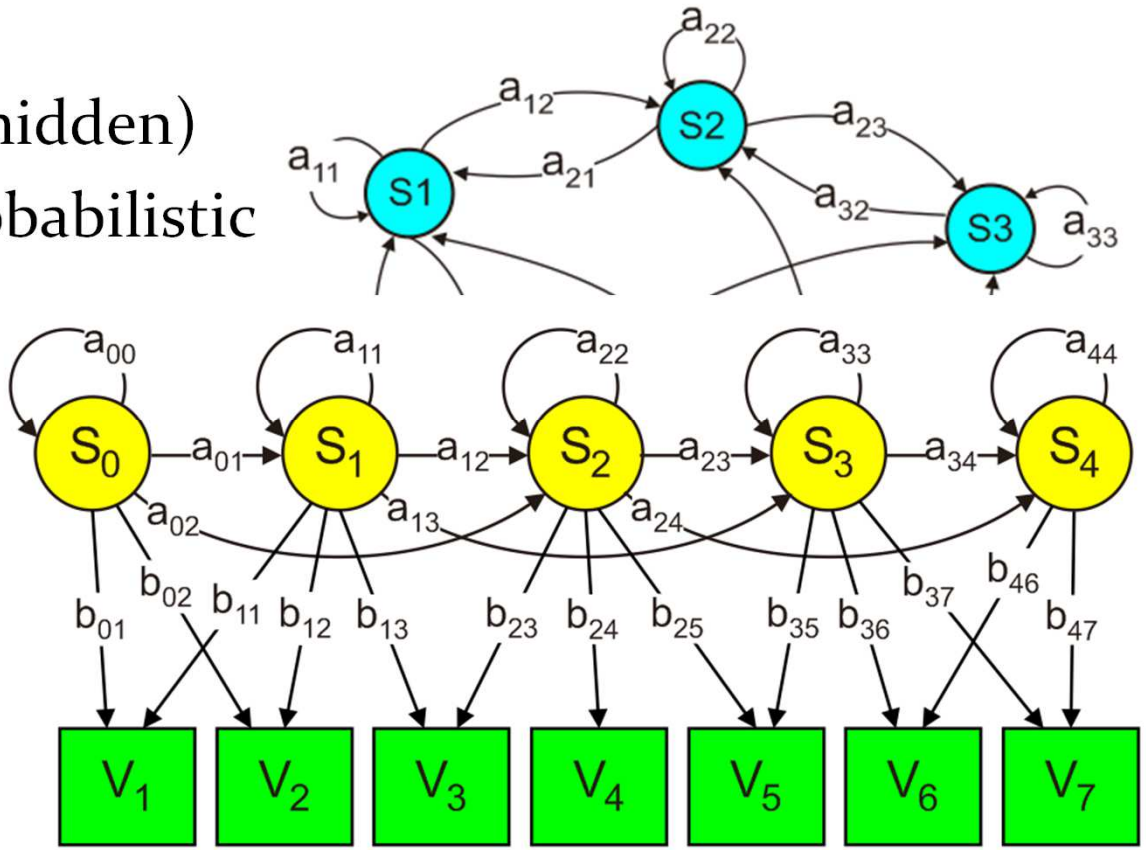
  Likok Pulo dance from Aceh

  * Strong and decent identity.
  * The choreography is mostly keep unchanged
  * Requires synchronous motion among the group
  * Requires precision timing of gestures with the rhythm
  * The rhythm linearly change m
  * Several gestures performe
  * Well accepted in the inter environment

How to model the dance itself?

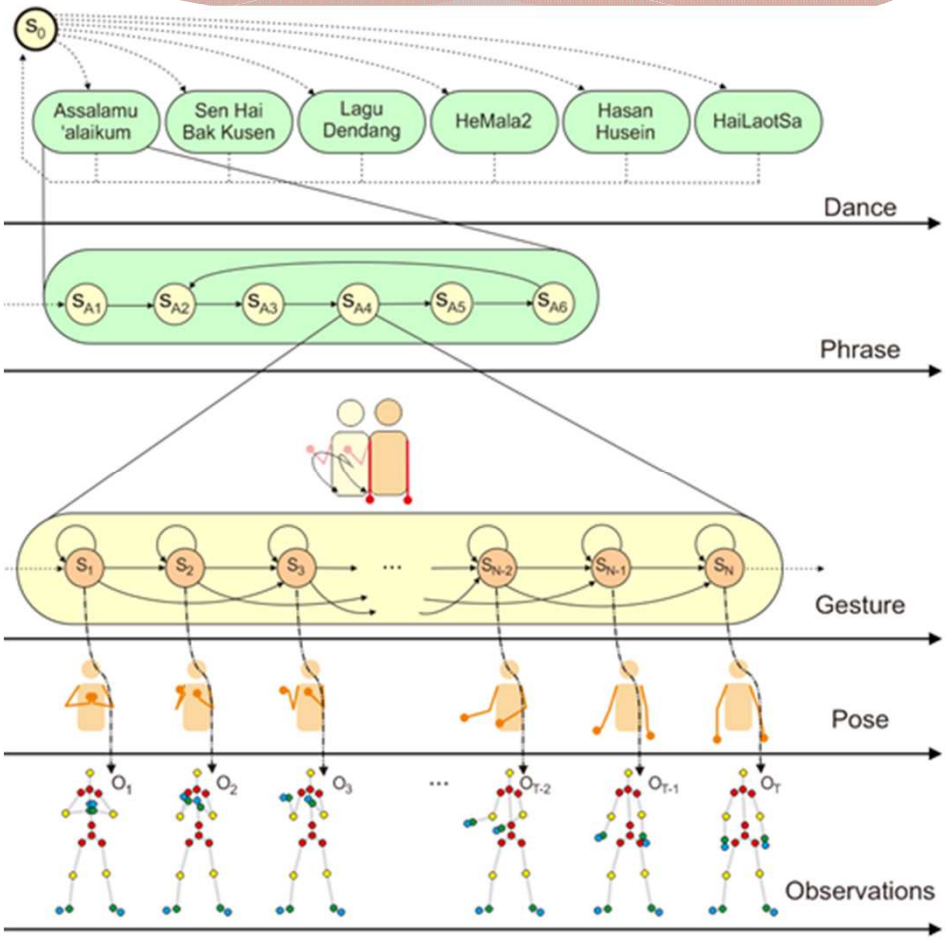
Hidden Markov Model

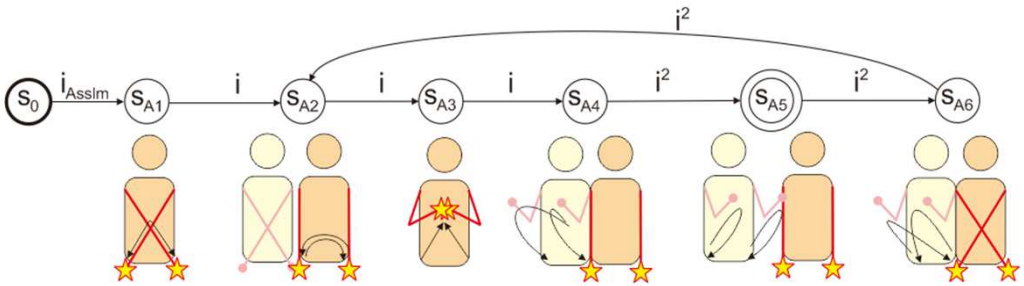
* Markov model
* Not observable state (hidden)
* The observation as probabilistic function of the state



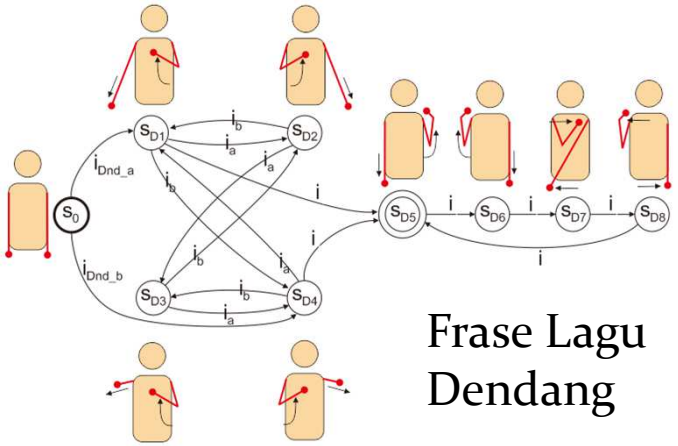
Modelling the Whole Dance

* Hierarchy of the dance
* $\mathcal{L} = (S, I, P, O, f, e, s_0, S_t)$
  * S, the finite nonempty set of states.
    Correspond to gestures.
  * I, the finite nonempty set of input.
  * P, the vocabulary of all possible discrete pose of dance.
  * O, the finite nonempty set of output.
    Where $O = \{o_1, o_2, \dots, o_T\}$, $o_i \in P^*$, $i \in \{1,2, \dots, T\}$.
  * f, state transition function $f : S \times I \to S$.
    Corresponds to gesture transitions.
  * e, the output map $e : S \times I \to O$.
  * s_0, initial state, $s_0 \in S$.
    Corresponds to initial pose/gesture of all phrases of Likok Pulo.
  * S_t – set of final (or accepting) states, $S_t \subseteq S$. Final states correspond to the end of the phrase.

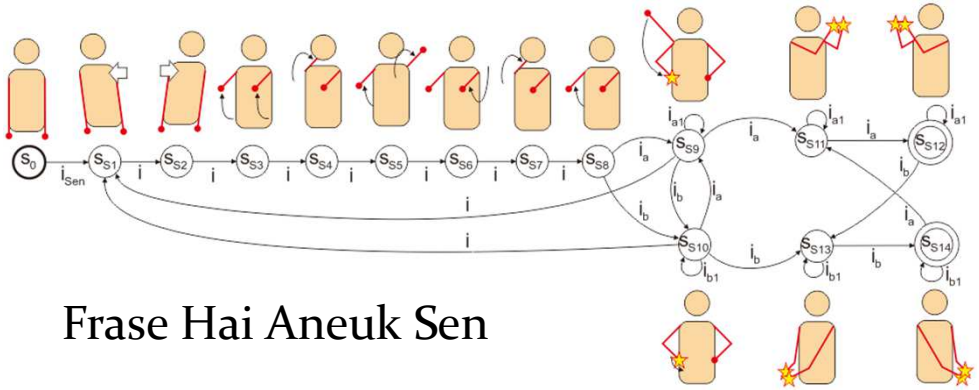




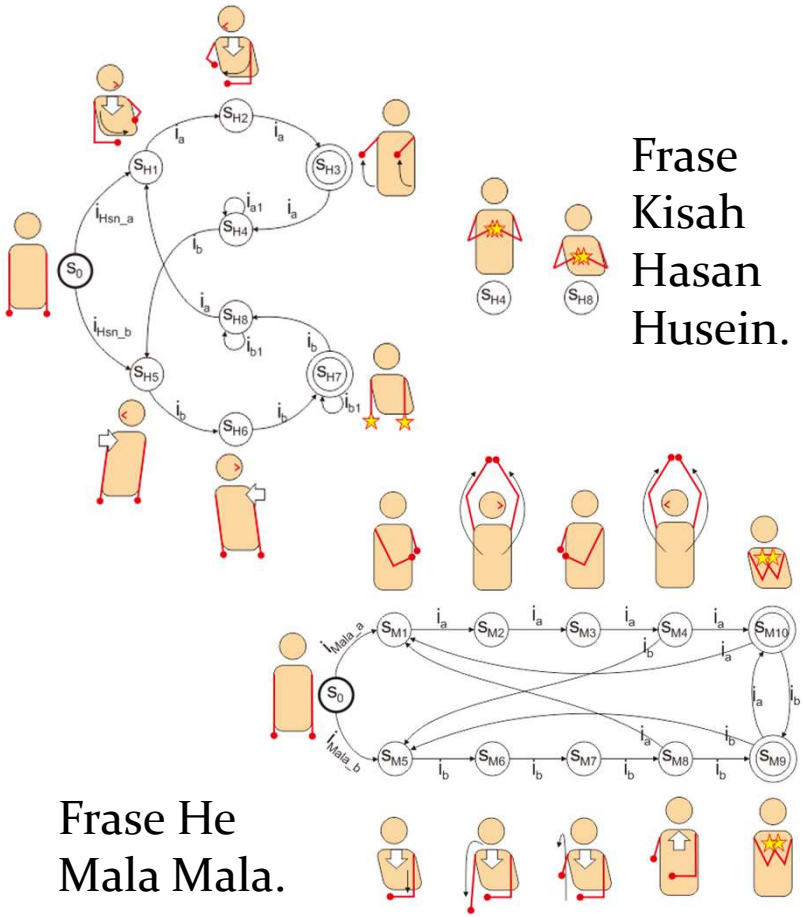
Frase Assalamualaikum



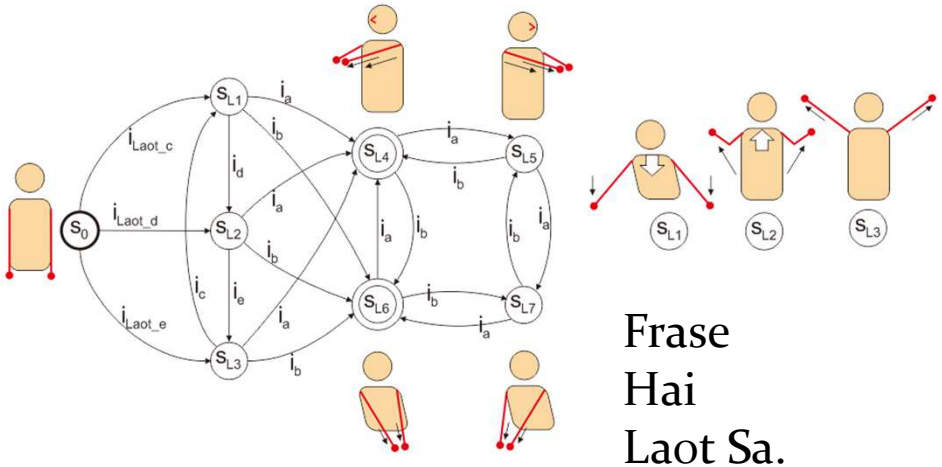
Frase Lagu Dendang



Frase Hai Aneuk Sen



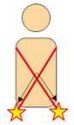
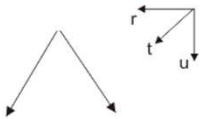
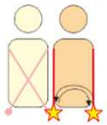
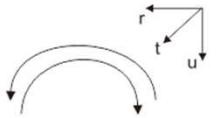
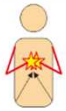
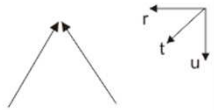
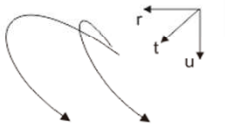
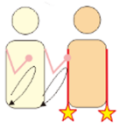
Frase Kisah Hasan Husein.

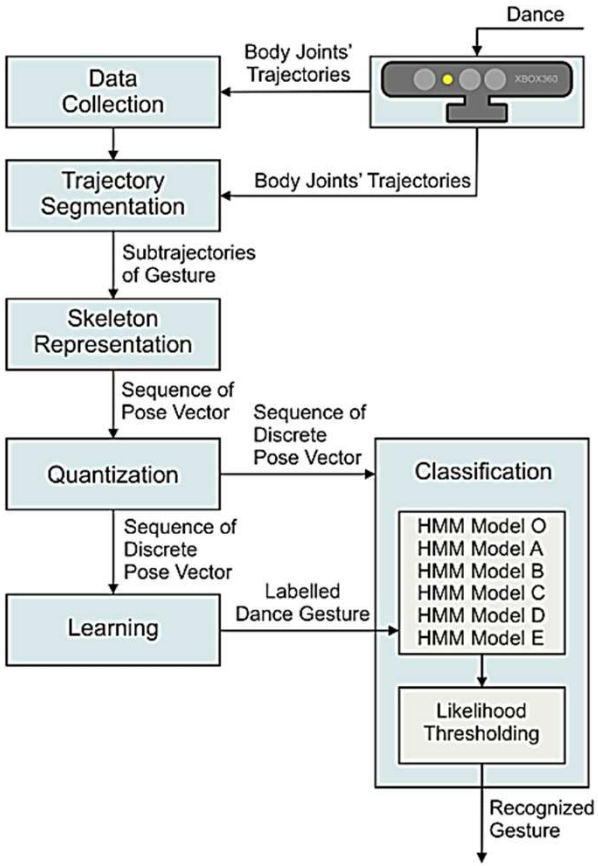


Frase Hai Laot Sa.

Frase He Mala Mala.

Learning and Recognition System for 6 Gesture Classes

Gesture	Wrist Trajectories	Class	Description
		O	"Clapping the hand in front of the chest" to "crossing the hand over the thigh".
		A	"Crossing hand over the thigh" to "straightening the hand over the thigh".
		B	"Straightening the hand over the thigh" to "clapping the hand in front of the chest".
		C	"Clapping the hand in front of the chest" to "swinging the hand to the rightside" to "straightening the hand over the thigh".
		D	"Straightening the hand over the thigh" to "swinging the hand to the leftside" to "straightening the hand over the thigh".
		E	"Straightening the hand over the thigh" to "swinging the hand to the rightside" to "crossing the hand over the thigh".



Data Collection and Segmentation

* Has been collected 2169 isolated gestures data
* Classified to 6 sets of data for each of dance gesture classes
* Partitioned into 80% training data and 20% test data.
* Segmentation between gestures
  * Detecting hand's clap
  * Using a time window

Gesture	Only Kinect	W/Clap Sensors	Total Data	Training Data	Test Data
O	387	0	387	310	77
A	394	71	465	372	93
B	415	92	507	406	101
C	488	127	615	492	123
D	415	131	546	437	109
E	386	113	499	400	99

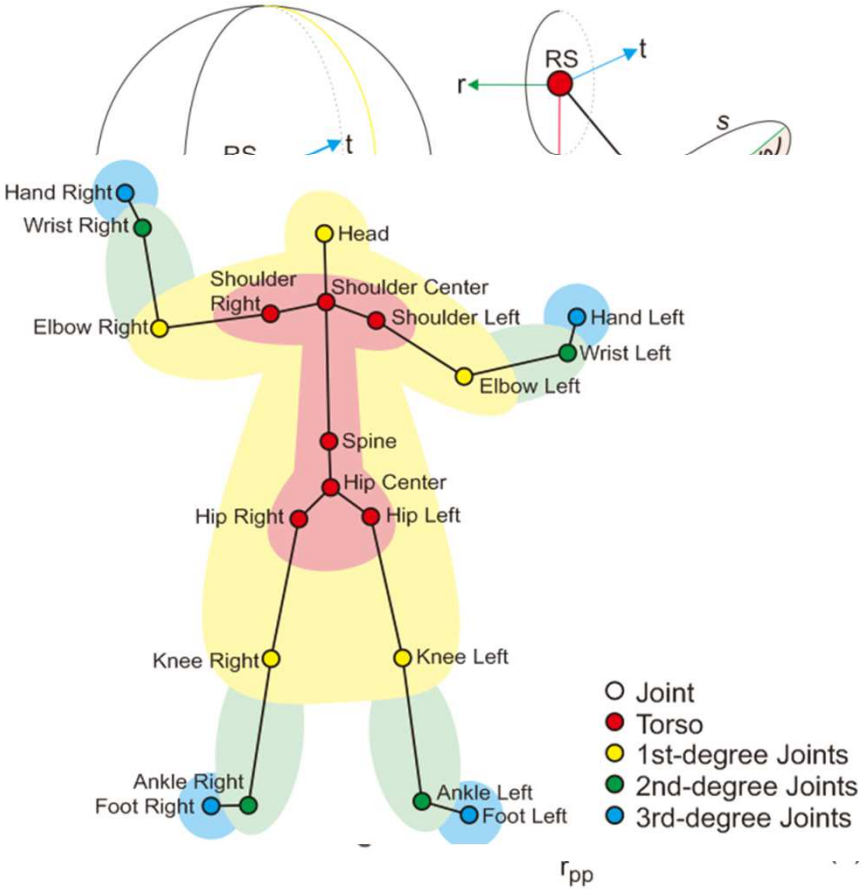
Skeleton Representation

Objectives :

* Robust coordinate system based on human body orientation
* Continuity and stability of the signal.
* Reduce the dimension of the signal while maintaining the character of the motion.

---

* Hierarchy of skeleton joint
* Torso PCA frame
* Computation for each degree



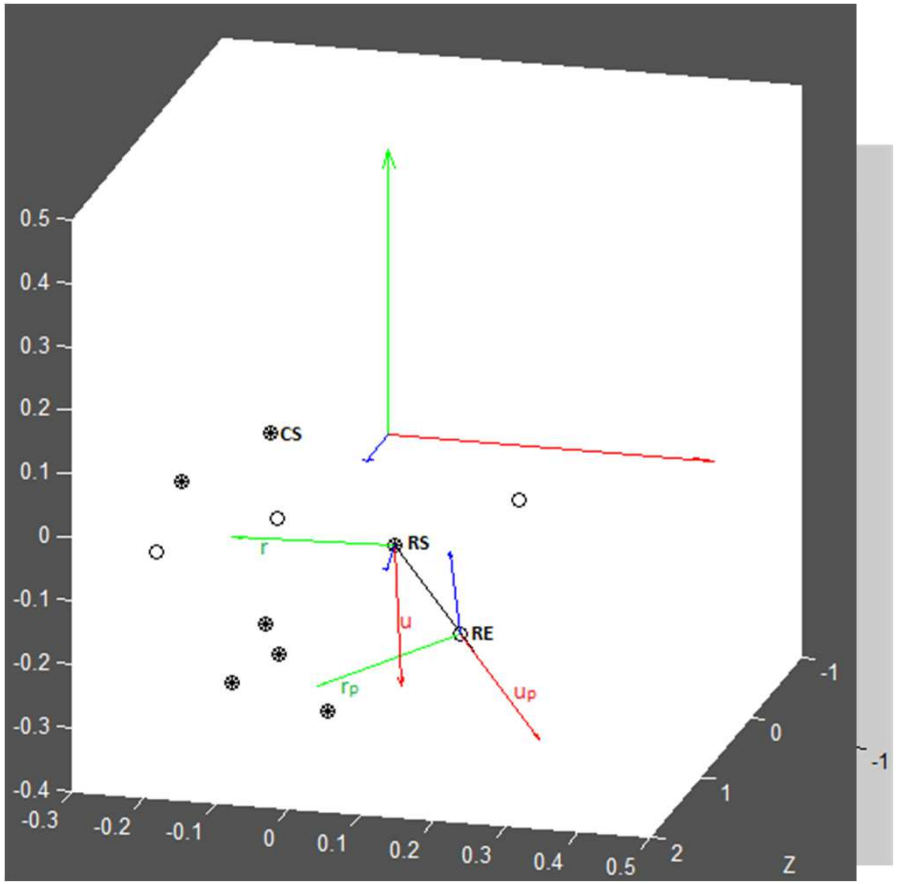
Human Pose Representation

For the scope of body poses which involves up to second-degree joints,

* Upper body poses are represented by an 8-tuple
$$P_{u,2} = \left(LE_{\varphi}, LE_{\theta}, LS_{\varphi}, LS_{\theta}, RS_{\theta}, RS_{\varphi}, RE_{\theta}, RE_{\varphi}\right).$$
* Lower body poses are represented by the 6-tuple
$$P_{l,2} = \left(LK_{\theta}, LH_{\varphi}, LH_{\theta}, RH_{\theta}, RH_{\varphi}, RK_{\theta}\right).$$

For the scope of body poses which involves up to third-degree joints,

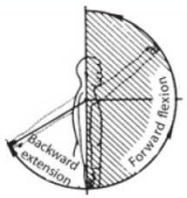
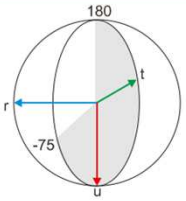
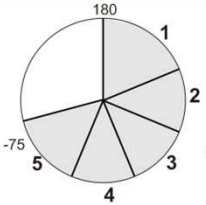
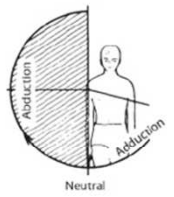
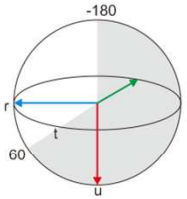
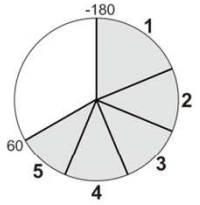
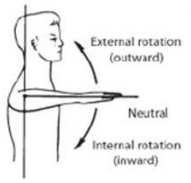
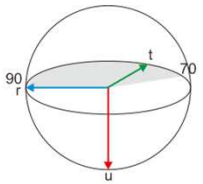
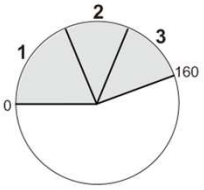
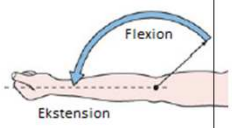
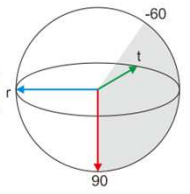
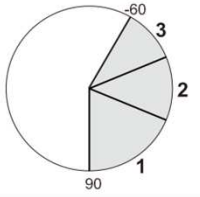
* Upper body poses are represented by an 12-tuple
$$P_{u,3} = \begin{pmatrix} LW_{\varphi}, LW_{\theta}, LE_{\varphi}, LE_{\theta}, LS_{\varphi}, LS_{\theta}, \dots \\ RS_{\theta}, RS_{\varphi}, RE_{\theta}, RE_{\varphi}, RW_{\theta}, RW_{\varphi} \end{pmatrix}$$
* Lower body poses are represented by the 12-tuple
$$P_{l,3} = \begin{pmatrix} LA_{\phi}, LA_{\theta}, LK_{\theta}, LH_{\phi}, LH_{\varphi}, LH_{\theta}, \dots \\ RH_{\theta}, RH_{\varphi}, RH_{\phi}, RK_{\theta}, RA_{\theta}, RA_{\phi} \end{pmatrix}$$
* Head poses are represented by the 3-tuple
$$H = \left(H_{\varphi}, H_{\theta}, H_{\phi}\right)$$



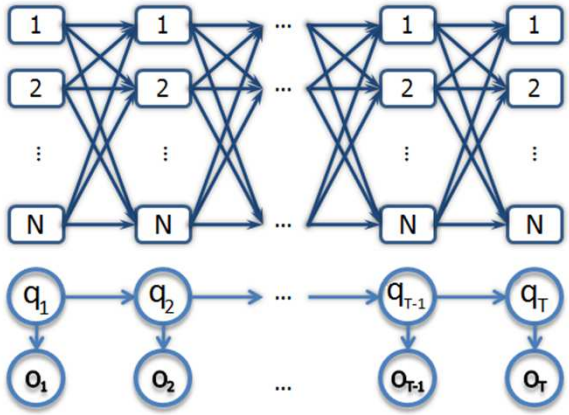
Pose Vector Quantization based on Range of Movement

Left Arm	Right Arm
$-60 \leq LS_{\varphi} \leq 180$	$-60 \leq RS_{\varphi} \leq 180$
$-75 \leq LS_{\theta} \leq 180$	$-180 \leq RS_{\theta} \leq 75$
$20 \leq LE_{\varphi} \leq 180$	$0 \leq RE_{\varphi} \leq 160$
$-60 \leq RE_{\theta} \leq 90$	$-60 \leq RE_{\theta} \leq 90$

* Pose vector $P_{u,2}$ is quantized
* All possible pose's configuration involving up to second-degree joints on one arm are 225 poses.
* All possible pose's configuration on two arms are $225 \times 225 = 50625$ poses.

Joint Angle	Physical Representation	Range of Joint Angle	Directional Codewords
RS_{φ}			
RS_{θ}			
RE_{φ}			
RE_{θ}			

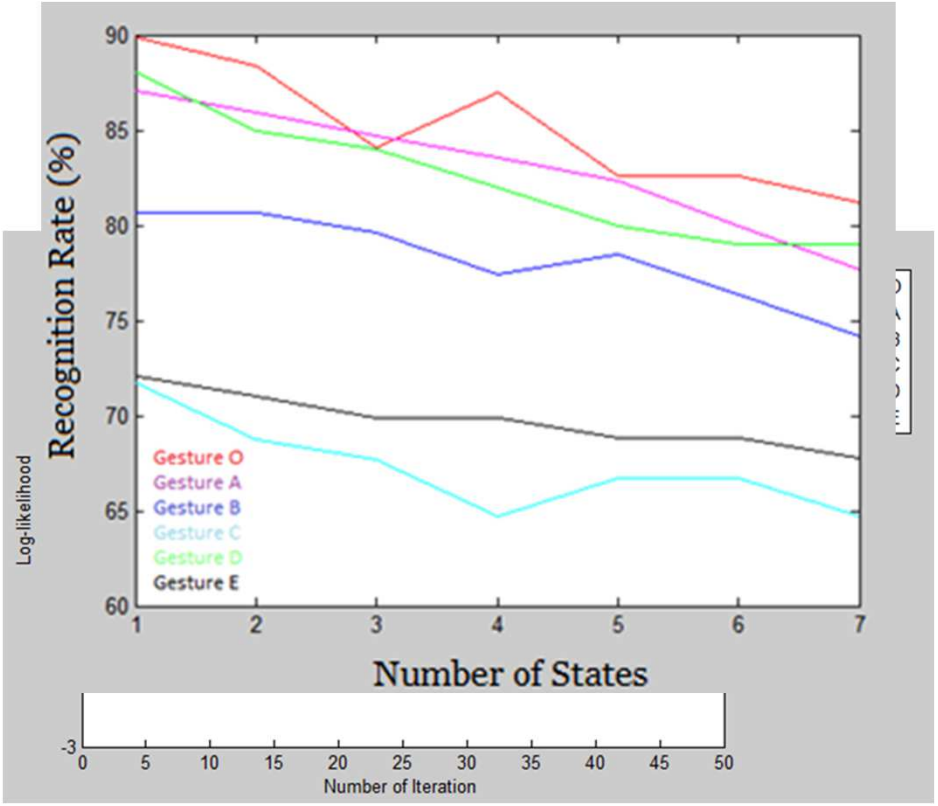
Learning using Baum Welch Algorithm



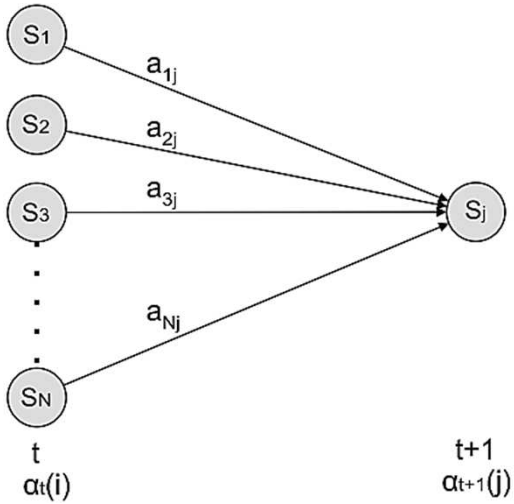
$\bar{\pi}_i$ = frekuensi harapan di state S_i pada saat $t = 1$ adalah $\gamma_1(i)$

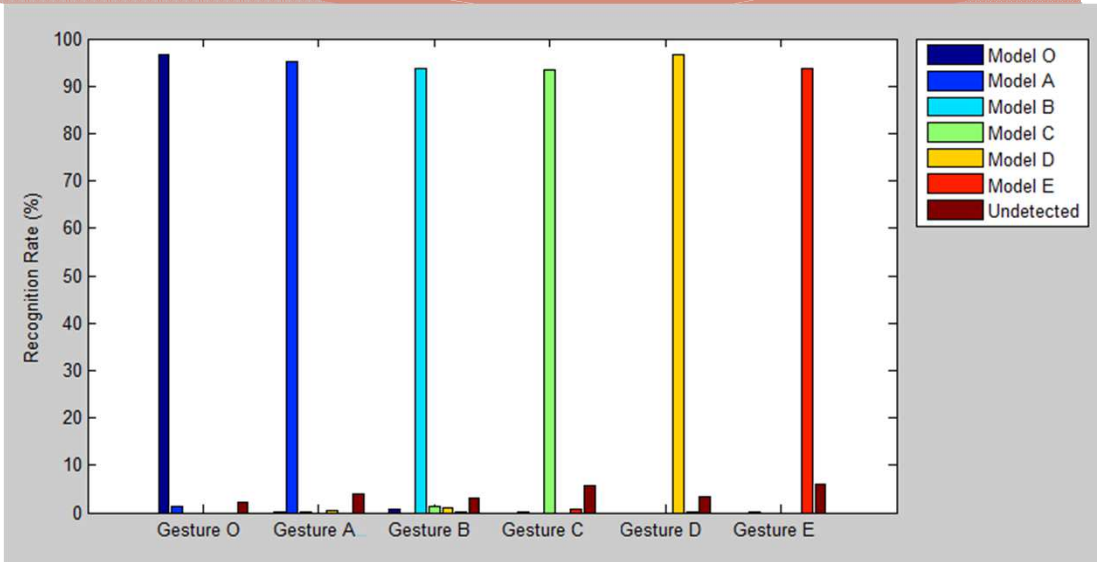
$$\bar{a}_{ij} = \frac{\text{freq harapan transisi datang } S_i \text{ menuju } S_j}{\text{freq harapan transisi datang dari } S_i} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\bar{b}_j(k) = \frac{\text{freq harapan berada di } S_j \text{ dan simbol observasi } v_k}{\text{freq harapan berada di state } S_j} = \frac{\sum_{\substack{t=1 \\ O_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j)}$$



Classification using Forward Algorithm





* $\alpha_1(i) = \pi_i b_i(O_1)$
* $\alpha_{t+1}(j) = \left[\sum_{j=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(O_{t+1})$
* $P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$
  * $1 \le t \le T,\ 1 \le i \le N$

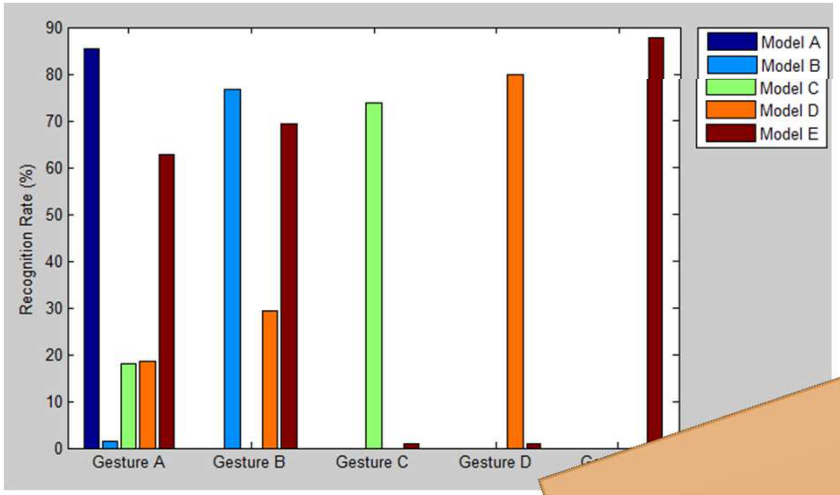
Gesture	Keterdeteksian Gesture oleh Model (%)						Tidak terdeteksi
	O	A	B	C	D	E	
O	96.64	1.29	0.00	0.00	0.00	0.00	2.07
A	0.22	95.27	0.22	0.00	0.43	0.00	3.87
B	0.79	0.00	93.69	1.18	0.99	0.20	3.16
C	0.00	0.16	0.00	93.33	0.00	0.81	5.69
D	0.00	0.00	0.00	0.00	96.52	0.18	3.30
E	0.00	0.20	0.00	0.00	0.00	93.79	6.01

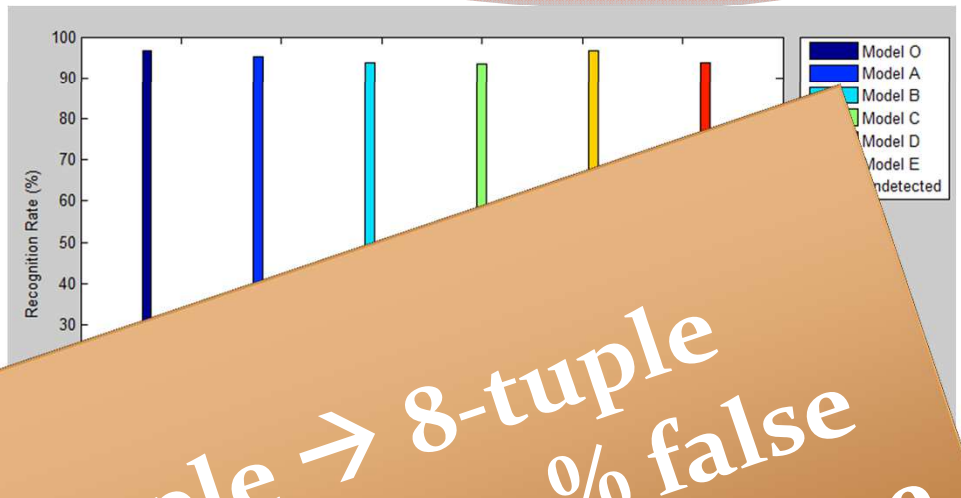
Classification using Forward Algorithm

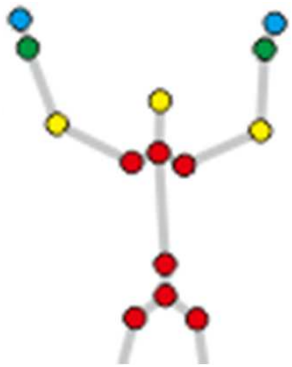
Gesture	Detected asTrue (%)	Detected as False Positive (%)	Detected as False Negative (%)
O	96.64	1.01	1.29
A	95.27	1.49	0.87
B	93.69	0.22	3.16
C	93.33	1.18	
D	96.52		
E	93.79		

Accuracy = 94.87 %

Effect of using Skeleton Representation and Max-likelihood Thresholding





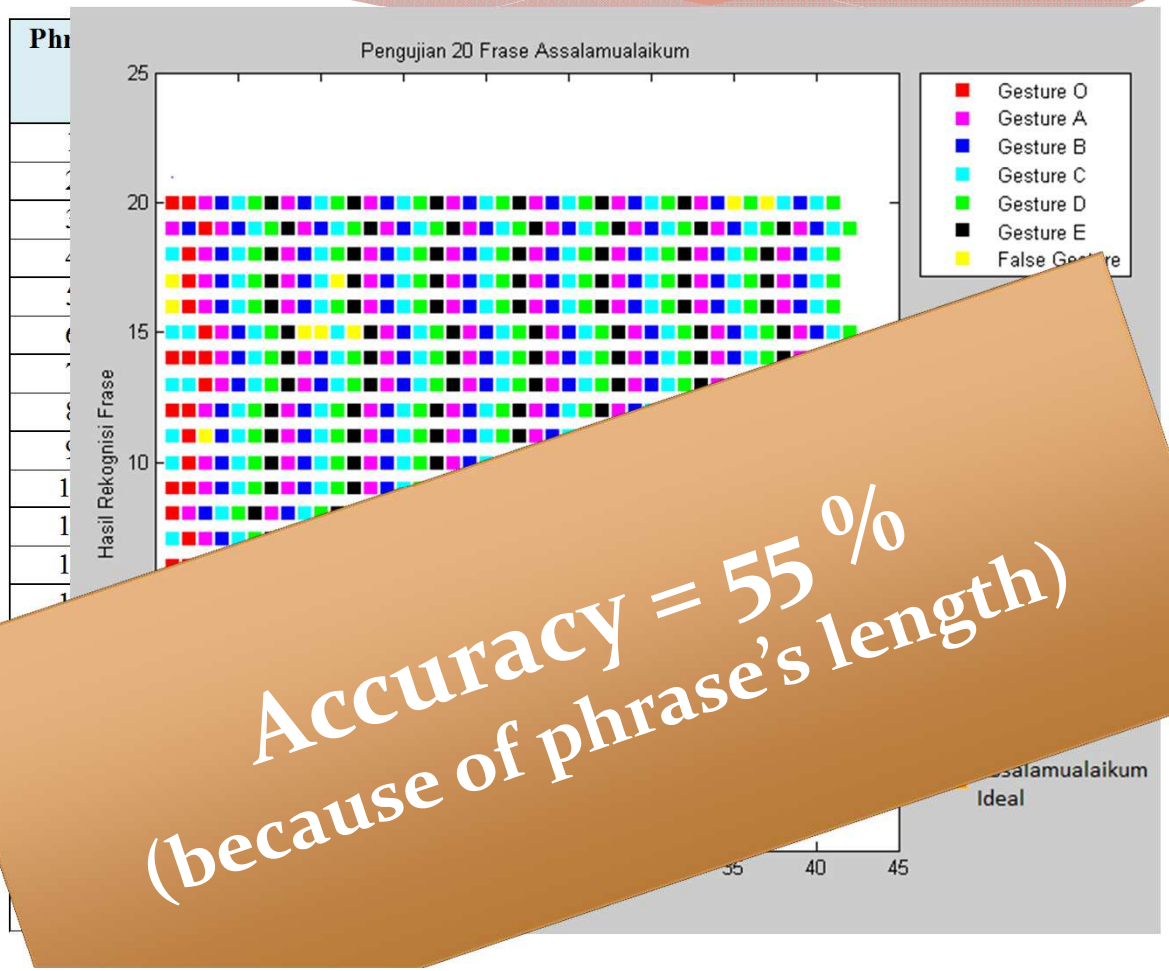


33-tuple → 8-tuple
10 % false→ 0.22 % false
80.7 % true→ 94.87 % true

$\varphi, RE_\theta, RE_\varphi).$

Dance Phrase Recognition

* Twenty of complete recording data of “Assalamualaikum” dance phrase are collected.
* Segmented using time window given by the system



Accuracy = 55 %
(because of phrase’s length)

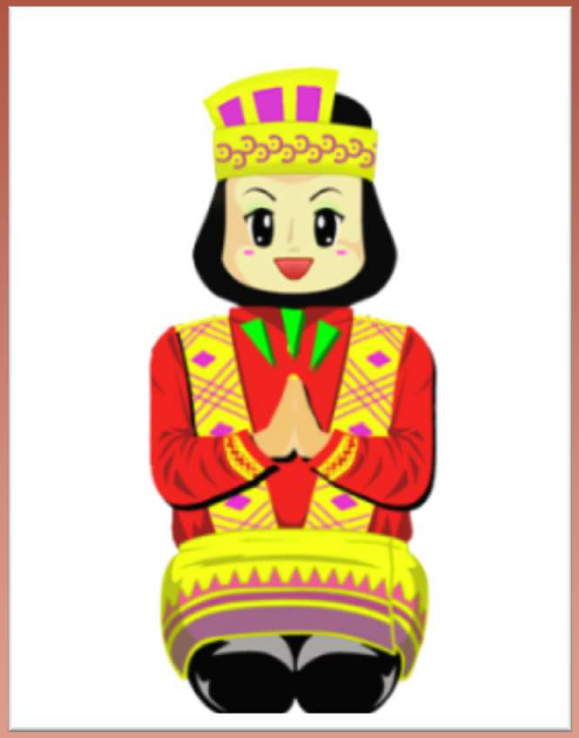
Realtime Implementation

Conclusion

* Finite state machine $\mathcal{L} = (\boldsymbol{S}, \boldsymbol{I}, \boldsymbol{P}, \boldsymbol{O}, \boldsymbol{f}, \boldsymbol{e}, \boldsymbol{s_0}, \boldsymbol{S_t})$ can be used to model the whole dance; dance gestures cast as discrete states and phrase as a sequence of gestures.
  * Hidden Markov Model can be used to model the real whole dance
* The most discriminative feature to represent all of dance pose are angular skeletal representation $P_{u,2} = \left(RS_{\varphi}, RS_{\theta}, RE_{\varphi}, RE_{\theta}, LS_{\varphi}, LS_{\theta}, LE_{\varphi}, LE_{\theta}\right)$ that is quantized based on range of movement.
* HMM are an effective and efficient method of both recognizing and classifying dance gestures involving several joints.
* Learning and recognition system for six of dance's gesture classes from the phrase "Assalamualaikum" has been implemented.
  * The classifier system has an accuracy of 94.87% for single gesture
  * It has an accuracy of 55% for complete phrase of 40 gestures.

Suggestion for Next Research

* Observation of the dance can be expanded up to lower body, and/or expanded to third-degree joints.
  * Additional inertial sensors for capturing position and orientation of third-degree joints due to Kinect sensor can not detect it.
* Skeleton representation can be deepened to also consider the dynamic aspects
  * to implement it into dancing humanoid robot.
* For wider dance movement, the segmentation process must be independent from wearable sensors.
  * Segmentation process of dance movement can use gesture spotting method based on Viterbi algorithm [16].



Thanks

Wassalamualaikum
wa Rahmatullah wa Barakatuh