OPEN ACCESS
ISSN 2356-5462
socj.telkomuniversity.ac.id/ijoict

Intl. Journal on ICT
Vol. 2, Issue. 2, December 2016. pp. 1-10
doi:10.21108/ijoict.2016.22.124

Fuzzy Latent-Dynamic Conditional Neural Fields for Gesture Recognition in Video

Intan Nurma Yulita [#*1], Mohamad Ivan Fanany [#2], Aniati Murni Arymurthy [#3]

 # Faculty of Computer Science, Universitas Indonesia
Kampus UI Depok, Depok 16424, Indonesia

[*]Department of Computer Science, Universitas Padjadjaran
Jalan Raya Bandung Sumedang, Km. 21 Sumedang 45363, Indonesia

Corresponding Authors:

[1] intan.nurma@unpad.ac.id
[2] ivan@cs.ui.ac.id
[3] aniati@cs.ui.ac.id

Abstract

With the explosion of data on the internet led to the presence of the big data era, so it requires data processing to get the useful information. One of the challenges is the gesture recognition the video processing. Therefore, the study proposes Latent-Dynamic Conditional Neural Fields and compares with the other family members of Conditional Random Fields. For improving the accuracy, these methods are combined by using Fuzzy Clustering. From the results, it can be concluded that the performance of Fuzzy Latent-Dynamic Conditional Neural Fields are the highest. Also, the combination of the basic classifiers and Fuzzy C-Means Clustering has the higher than the original ones. The evaluation is tested on a temporal dataset of gesture phase segmentation.

Keywords: Fuzzy Clustering, Gesture Recognition, Latent-Dynamic Conditional Neural Fields

I. Introduction

THE development of the Internet is rapidly increasing since 1990. It has led to an explosion of data many times with the presence of the social media era. Data with all kinds of formats from the text, audio, and video either structured or not has been uploaded on the internet. Even very large sized data is very fast growing exponentially every second. Data with the condition often is known as big data. It is a trend which attracted much attention of researchers to study it. One of the challenges in big data is processing of sequential data. One of the interesting tasks in the process is labeling.

Hidden Markov Models are a widely used method for speech labeling. However, its implementation has been extensively used in many areas such as Bioinformatics (Yoon, 2009), fisheries (Spampinato & Palazzo, 2012), meteorology (Lambert et al., 2003), and health (Cooper & Lipstich, 2004). Input HMM in the form of observational data with one dimension of observational data made through vector quantization. Quantization commonly performed using the k-means clustering. The many features of the data will determine the number of dimensions K-means clustering. The final result of K-means clustering is centroids (center of clusters). Observational data are obtained by selecting the closest of the centroid to the data. To determine the label of a sequence of observations, HMM uses joint probability based on a calculation of all the possibilities of the observational sequence (Zhang, 2012). It is the shortage of HMM because the calculation becomes

Received on Aug 2016. Accepted on Nov 2016

impractical to represent data with some interacting features. To overcome the problem, the conditional model is the best choice (Lafferty et al., 2001).

A well known conditional model for sequence labeling is Conditional Random Fields (CRF). CRF can combine the features of a complex sequence of observational data which does not require the assumption of non-independence among features. Labeling of CRF is based on the outer structure of interacting labels. Further, CRF is developed to be Hidden-state Conditional Random Fields which using the intrinsic structure of the sequence of observation. Its mechanism causes the performance of conditional models to decrease. Therefore, Latent-dynamic Conditional Random Fields combines intrinsic and extrinsic structure. Despite their success in labeling, but they still fail to learn complex nonlinear relationship.

One of the possible solutions is Neural Conditional Fields (CNF). CNF is developed by Jiang Peng et al. (Peng et al., 2009). The model is a combination from Conditional Random Fields and Neural Networks. Neural networks are useful for learning complex nonlinear relationship. Its function is added to be a part of CNF. It is implemented through the gates at the intermediate level layer. CNF is developed from CRF, but it can be prepared from LDCRF. LDCRF with Neural Networks is called Latent-Dynamic Conditional Neural Fields (LDCNF) (Levesgue et al., 2013). LDCNF consists of two layers, namely the layer of the gate for learning non-linear relationship and dynamic layer of the intrinsic structure. Therefore, the study proposed LDCNF for gesture recognition and compared with the other family members of Conditional Random Fields: CRF, LDCRF, and CNF. These methods are basic classifier for the recognition.

For improving the accuracy, the study also proposes clustering to be combined with the basic classifier. The clustering is used as a filtering, which captures interesting feature subset to be the input to the basic classifier. Fuzzy C-Means clustering is selected for the study because the method finds the subset without the loss information which may be raised during the process.

II. LITERATURE REVIEW

In the study, the four basic classifiers are used, namely Conditional Random Fields, Latent-Dynamic Conditional Random Fields, Neural Conditional Fields, and Latent-Dynamic Conditional Neural Fields. Also, these classifiers are combined by using Fuzzy C-Means Clustering. Each method will be described as follows.

A. Conditional Random Fields (CRF)

X is the vector of input sequence while Y is a vector of label sequence. Both are defined as follows:

$$X=x_1,x_2,x_3,x_4,\ldots..x_n \tag{1}$$

$$Y=y_1,y_2,y_3,y_4,\ldots..y_n \tag{2}$$

Both vectors have the same length. The probability of input X to label Y based on the following calculation:

$$p\,(y|x) = \frac{\exp\,[\sum_{i=1}^{k}\sum_{j=1}^{m}\theta_i\varphi_i(x,j,y_j,y_{j-1})]}{\sum_{l'}\exp[\sum_{i=1}^{k}\sum_{j=1}^{m}\theta_i\varphi_i(x,j,y_j,y_{j-1})]} \tag{3}$$

Where

$\varphi(x,j,y_j,y_{j-1})$ is a feature function on the current position (j).

θ is the weight of feature function.

Feature function may not only get from two positions of labels (y_j, y_{j-1}) but can be defined based on size of window. However, if the window size is too large, allowing unable to obtain feature function. The weights of the feature function obtained through the training data. The mechanism is generally done via gradient ascent.

B. Latent-Dynamic Conditional Random Fields (LDCRF)

The difference between CRF and LCRF is the intermediate layer consisting of some hidden-state to define the structure intrinsic so the probability of LDCRF for input X to label Y as follows:

$$p\ (y|x) = \frac{\exp\ [\sum_{i=1}^{k}\sum_{j=1}^{m}\theta_i\varphi_i(x,j,h_j,h_{j-1})]}{\sum_{l'}\exp[\sum_{i=1}^{k}\sum_{j=1}^{m}\theta_i\varphi_i(x,j,h_j,h_{j-1})]} \quad (4)$$

Feature function in LDCRF only occurs between the input and the intermediate layer.

C. Conditional Neural Fields (CNF)

To map a non-linear relationship is complex then the neural network is placed as an intermediate layer on the CNF. The middle layer acts as a gate function. Thus the probability for input, X to label Y is defined as follows:

$$p(y|x) = \frac{\exp[\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{g=1}^{n}\tau(\alpha_g\varphi_i(x,j,y_j,y_{j-1}))]}{\sum_{l'}\exp[\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{g=1}^{n}\tau(\alpha_g\varphi_i(x,j,y_j,y_{j-1}))]} \quad (5)$$

Where

τ is a gate function with weight for every gate, α_g.

D. Latent-Dynamic Conditional Neural Fields (LDCNF)

LDCNF has two intermediate layers consisting of several hidden states. The first layer aims to represent the intrinsic structure, and secondly to represent the complex nonlinear relationship. The calculation of probability for input, X to label, Y defined as follows:

$$p(y|x) = \frac{\exp[\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{g=1}^{n}\tau(\alpha_g\varphi_i(x,j,y_j,y_{j-1}))]}{\sum_{l'}\exp[\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{g=1}^{n}\tau(\alpha_g\varphi_i(x,j,y_j,y_{j-1}))]} \quad (6)$$

E. Fuzzy C-Means Clustering

Clustering is a K-Means that implement Fuzzy Logic. An object becomes a member of any cluster, but it has different degrees of membership. In general, the number of clusters and its centroid is initialized at the beginning. The degree of membership of an object to each cluster is calculated based on the distance of the object to each centroid. Furthermore, by using the degree of membership of these objects, the centroids are updated. These changes further affect the degree of membership of each object so that the calculation process 3s repeated. It continued to reach the objective function or value of the expected error.

III. RESEARCH METHOD

A. *Dataset*

The performance of methods will be tested by gesture phase segmentation. The dataset can be downloaded from the UCI repository. The data set consisted of temporal data from the segmentation gesture phase, which collected by the School of Art, Sciences and Humanities, University of Sao Paulo, Brazil uses The Microsoft Kinect Sensor (Madeo et al., 2013). From the dataset is provided, the study only uses three data which comprised of 1747, 1073, and 1111 frames. The data have 20 attributes that consist of six positions (left hand, right hand, head, spine, left wrist and right wrist), the coordinates (x, y, x) for each position, timestamp, and phase (rest, preparation, stroke, hold, and retraction). For the study, the timestamp is not used.

B. *System Architecture*

System Architecture used in the study is based on research conducted by Fabio Tamburini et al. for Prosodic Prominence Detection (Tamburini et al., 2014). The illustration is shown in Fig. 1.

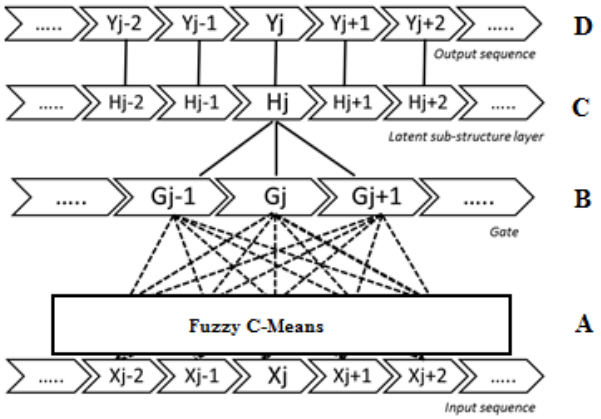

Fig. 1. Architecture

There are no intermediate layers in the system architecture of CRF. It is different from the other three basic classifiers. LDCRF and CNF have one intermediate layer, but the intermediate layer of LDCRF consists of hidden state, while the CNF is composed of the gate. Intermediate layer located between the input and output layer. LDCNF is a combination of LDCRF and CNF, so it has two intermediate layers consisting of a latent-dynamic layer and gate level. The input data will be processed by the gate before latent-dynamic layer. Unlike the case with implementations using Fuzzy, clustering is the first process. Overall difference methods are described in Table 1.

TABLE I
METHODS

No	Methods	Process
1	CRF	Input → D
2	LDCRF	Input → C →D
3	CNF	Input → B →D
4	LDCNF	Input → B →C →D
5	FCRF	Input → A →D
6	FLDCRF	Input → A → C →D
7	FCNF	Input → A → B →D
8	FLDCNF	Input → A → B →C →D

IV. Results And Discussion

The analysis was conducted by comparing among the basic classifiers and also the combination with their fuzzy filtering. Testing scheme is done through 3-cross validation. Validation by using cross validation divides the data into k subsets. K is the number of fold that will be utilized. In the study, we use three-Cross-validation, so there will be three iterations/rounds of testing. The performance of each method was tested for gesture phase labeling. The performance is based on sensitivity and execution time. The result of each test is shown in the next sub-chapter.

TABLE II
NUMBER OF CLUSTERS

No	**Methods**	**Cl**	**Sensitivity**		**Running Time**	
			Average	**Stdev**	**Average**	**Stdev**
1	FCRF	4	0.47	0.31	35.06	11.52
2	FCRF	5	0.46	0.40	35.55	7.39
3	FCRF	6	0.33	0.41	80.00	11.48
4	FCRF	7	0.38	0.44	56.58	11.31
5	FCRF	8	0.35	0.46	82.73	6.16
6	FLDCRF	4	0.49	0.27	647.70	30.97
7	FLDCRF	5	0.41	0.45	670.58	43.90
8	FLDCRF	6	0.32	0.41	778.99	204.99
9	FLDCRF	7	0.52	0.46	899.49	49.52
10	FLDCRF	8	0.32	0.43	873.34	302.75
11	FCNF	4	0.81	0.10	36.17	7.90
12	FCNF	5	0.76	0.15	36.22	8.13
13	FCNF	6	0.38	0.43	77.65	13.60
14	FCNF	7	0.48	0.37	52.15	12.78
15	FCNF	8	0.39	0.44	74.24	4.55
16	FLDCNF	4	0.83	0.08	253.58	78.30
17	FLDCNF	5	0.78	0.14	241.96	20.90
18	FLDCNF	6	0.38	0.43	311.99	7.99
19	FLDCNF	7	0.49	0.37	335.91	138.99
20	FLDCNF	8	0.40	0.43	310.68	33.63

A. Number of clusters

Fuzzy C-Means clustering groups the objects into several clusters. The cluster number is initialized at the beginning. To find the optimal number of clusters for gesture recognition, the necessary experiments are done. The mechanism of the experiment carried out by changing the initial parameters for the number of clusters with fuzziness degree that is used is fixed. In the experiment, fuzziness degree is 1.05.

From the results in Table 2, the FCRF best performance is obtained when the number of clusters is four, which gives the highest sensitivity, but lowest running time. The worst performance occurs in the use of eight clusters. Performance decreased with the increase in the number of clusters as a decrease in sensitivity, but the increase in running time. However, despite the downward trend, using seven clusters is better than using six clusters.

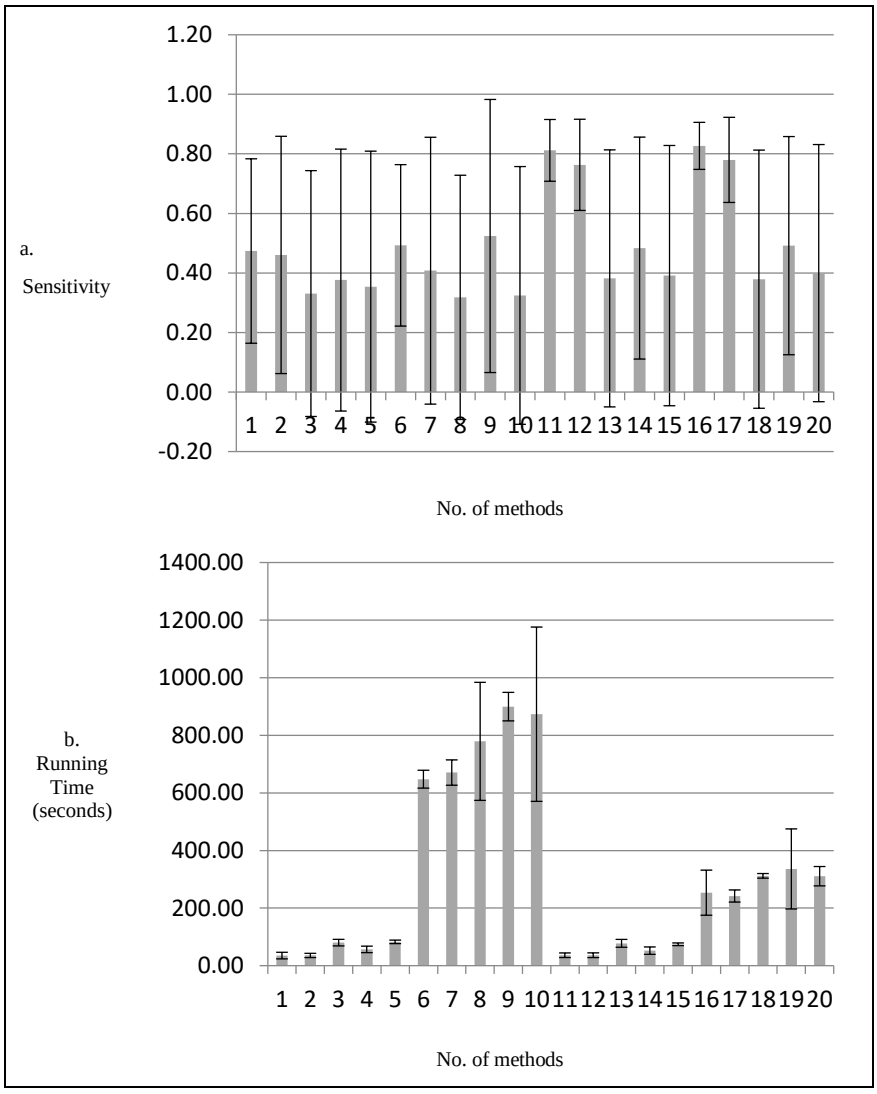

Fig. 2. Number of clusters

The FLDCRF implementation gives a sensitivity of FCRF from 0.32 to 0.49 so that FCRF is better than FLDCRF in term of sensitivity. The running time of FLDCRF requires more time ranging from 647.70 to 899.49 seconds. Here it can be seen that the use of hidden-state does not deliver an effective process. On the other hand, the gate function of intermediate layers gives a significant result compared to the hidden-state in FLDCRF. Sensitivity for the gesture recognition increased to a range from 0.39 until 0.81. Significant results can be seen that it has less running time than FCRF and FLDCRF. It happens because the features of dataset find the optimal subset in the intermediate layer. So that the fewer features used for the recognition process to reduce the running time required. Also, by using FCRF and FLDCRF, the highest performance is obtained when the number of clusters used is 4 and the running time is only 36.17 seconds. Decreasing trend in sensitivity occurs at the time of adding the number of clusters used. Performance gate function also proved capable of raising sensitivity that ranges from 0.40 to 0.83. But the increase in running time also occurs in FLDCNF due to the mechanism of a merger between the FCNF and FLDCNF.

On the other hand, the error bars with standard deviation (Stdev) is shown in Fig.2. The biggest uncertainty in testing is when FCRF is implemented for eight clusters, and FLDCRF for seven clusters. While the smallest occurred during the implementation of FCNF and FLDCNF for four clusters. When compared to the required running time, the highest deviation occurs on the use of FLDCRF, and the lowest on FCNF.

Overall, the highest performance obtained if all four methods only using four clusters. It is different from the number of classes of gesture recognition dataset used. The difference indicates that the two categories of the dataset have characteristics that are very close together. Setting the number of clusters as well as the number of classes of dataset turned out to cause a decrease in sensitivity.

TABLE III
FUZZINESS

No	Methods	FD	Sensitivity		Running Time	
			Average	Stdev	Average	Stdev
1	FCRF	1.05	0.47	0.31	35.06	11.52
2	FCRF	1.1	0.29	0.30	37.08	11.62
3	FCRF	1.2	0.17	0.14	44.17	15.10
4	FCRF	1.3	0.05	0.04	46.16	10.78
5	FCRF	1.4	0.09	0.10	44.09	10.54
6	FLDCRF	1.05	0.49	0.27	647.70	30.97
7	FLDCRF	1.1	0.29	0.17	542.73	87.01
8	FLDCRF	1.2	0.15	0.13	460.94	60.28
9	FLDCRF	1.3	0.06	0.05	432.17	75.51
10	FLDCRF	1.4	0.07	0.08	403.65	105.82
11	FCNF	1.05	0.81	0.10	36.17	7.90
12	FCNF	1.1	0.65	0.16	32.51	10.90
13	FCNF	1.2	0.54	0.28	36.40	10.48
14	FCNF	1.3	0.37	0.38	38.70	12.03
15	FCNF	1.4	0.31	0.30	43.61	10.63
16	FLDCNF	1.05	0.83	0.08	253.58	78.30
17	FLDCNF	1.1	0.65	0.18	203.79	48.03
18	FLDCNF	1.2	0.53	0.27	203.92	25.41
19	FLDCNF	1.3	0.38	0.37	207.70	25.63
20	FLDCNF	1.4	0.31	0.30	205.14	34.26

B. Fuzziness

To be able in finding the optimal fuzziness degree (FD) is done by changing the parameter ranges from 1.05 to 1.4. If its implementation using membership degree is one, then clustering is equal to k-means clustering.

The test is only implemented four clusters according to the results already obtained previously. Results in Table 3 show that in the implementation of the FCRF. The increasing degrees of membership will decrease sensitivity and increase in running time. By using 1.05, sensitivity has the highest sensitivity and to raise it to 1.1. The sensitivity is reduced by half. It means that the increase is not necessary because it makes the performance of FCRF be decreased. If the hidden-state was added as an intermediate layer in FLDCRF, then the best performance is still obtained by using 1.05 as the degree of membership.

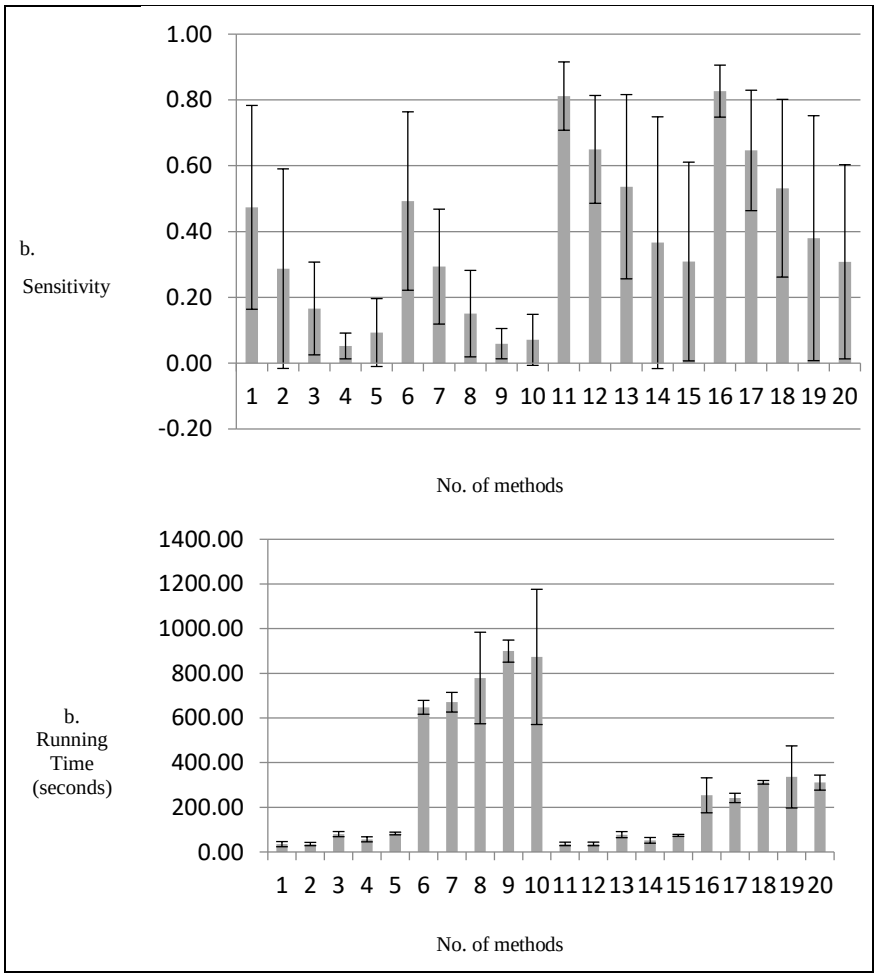

Fig. 3. Fuzziness degree

When compared with the FCNF, running time of the gesture recognition only has less difference than FCRF even when the degree of membership that is set is 1.2 to 1.4, running time FCNF is less than the FCRF, but sensitivity has a significant difference compared to FCRF.

On the other hand, FLDCNF has the highest performance in the implementation when the degree of membership used was 1.05. Sensitivity and running time for its degree was 0.83 and 253.38. Improved degree of membership has a trend decline in sensitivity but improved running time. Fig. 3 shows the error bars with the standard deviation (Stdev) for the experiment. FLDCNF and FCNF have the highest performance for fuzziness degree of 1.05. On the other hand, the standard deviation is too high. The increase of the fuzziness level of both methods leads to a decrease in sensitivity, and the standard deviation. Different conditions occur in FCRF and FLDCRF.

C. Comparison with and without Fuzzy

Performance comparison between the basic classifiers (CRF, LDCRF, CNF, and LDCNF) and combined with the use of fuzzy (FCRF, FLDCRF, FCNF, and FLDCNF) are shown in Table 4. From the table, it is known that the use by using only the basic classifiers only to achieve a sensitivity of 0.10 to 0.29. The worst performance was shown by LDCRF which has the lowest sensitivity and the highest running time. It indicates that the use of hidden-state is not effective for the gesture recognition

The use of gate function in the intermediate layer is proven effective in improving the performance. The combination of hidden-state and gate function gives the highest sensitivity despite the running time needed to reach 606.06 seconds and the standard deviation of 460.43.

Implementation of fuzzy and the basic classifiers using four clusters and membership degree is 1.05. From the table, it can be seen that the performance of CRF increased from 0.10 became 0.47 and the running time is becoming increasingly decreased to 35.06. It is because the number of features that are used less and it causes the efficiency of running time. The performance of FLDCRF, FCNF, and FLDCNF also experienced an increase in sensitivity and a decrease in running time. The highest of sensitivity is obtained FLDCNF.

The performance of the combination with Fuzzy is also increasingly seen in Fig. 4. Sensitivity is always greater than the basic classifiers. The sensitivity obtained is above 0.4 although FLDCRF has a large standard deviation. As for running time, the use of hidden-states from LDCRF, LDCNF, FLDCRF, and FLDCNF have greater running time. However the deviation error becomes smaller with the use of Fuzzy of those methods.

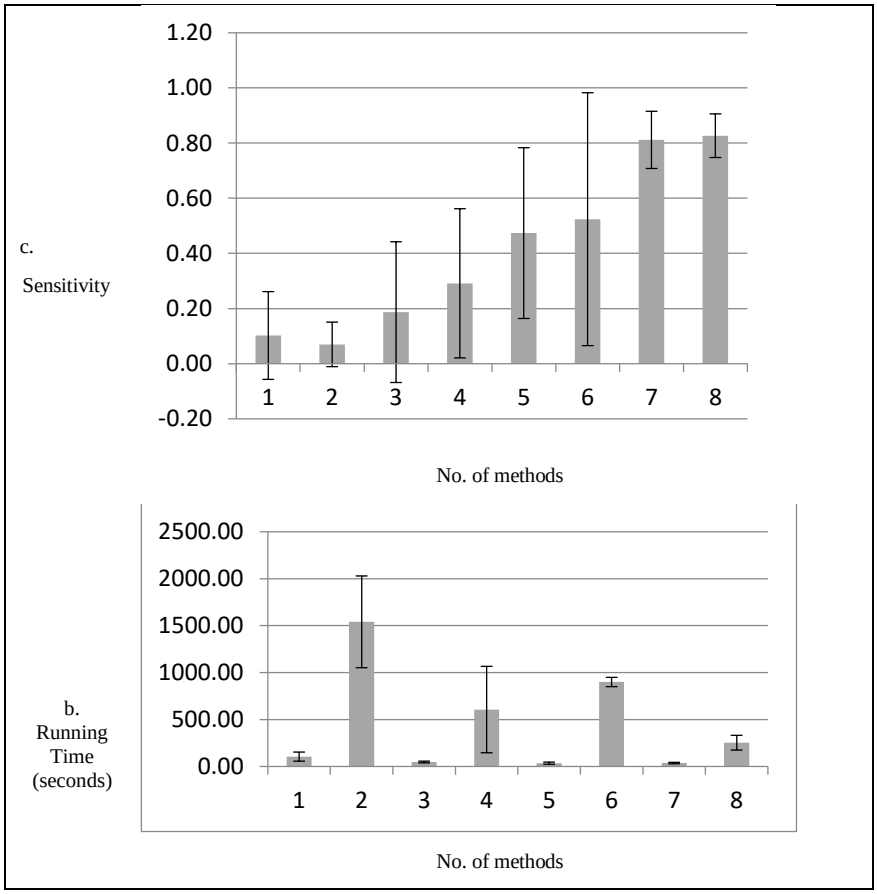

Fig. 4. Comparison

TABLE IV
COMPARISON

No	Methods	Sensitivity		Running Time	
		Average	Stdev	Average	Stdev
1	CRF	0.10	0.16	105.23	48.57
2	LDCRF	0.07	0.08	1540.53	488.55
3	CNF	0.19	0.26	47.26	9.49
4	LDCNF	0.29	0.27	606.06	460.43
5	FCRF cl=4, w=1.05	0.47	0.31	35.06	11.52
6	FLDCRF cl=7, w=1.05	0.52	0.46	899.49	49.52
7	FCNF cl=4, w=1.05	0.81	0.10	36.17	7.90
8	FLDCNF cl=4, w=1.05	0.83	0.08	253.58	78.30

V. Conclusion

From the study that has been done, it is known that the hidden state variables are not always effective and efficient for sequence labeling. It depends on the characteristics of the dataset. If the features have a strong correlation, the performance of a method that uses hidden state variables in the intermediate layer will be superior. Also, it is known that the gate function of CNF and LDCNF proved effective to find the right feature subset. So the accuracy increased but the execution time decreased with the feature subset. The other hand, the combination of the Fuzzy C-Means Clustering as Clustering and the base classifiers, give better performance than without the use of the Fuzzy C-Means Clustering. If no fuzzy, basic classifiers have the sensitivity ranged between 0.10 to 0.29 and running time ranged from 47.27 to 1540.53 seconds. Meanwhile, with fuzzy, sensitivity varied from 0.47 to 0.83 and the running time decreases ranged from 35.06 to 647.70 seconds. It indicates the use of the Fuzzy C-Means Clustering can filter the feature to discover new optimal features in the classification process. The discovery of the optimal features has the advantage to decrease the required running time.

ACKNOWLEDGMENT

The Author thanks to the Indonesian Endowment Fund for Education (LPDP) and Machine Learning and Computer Vision Laboratory, Universitas Indonesia that contributed and supported the study

REFERENCES

Ben Cooper, and Marc Lipsitch. 2004. The Analysis of Hospital Infection Data Using Hidden Markov Models. Biostatistics(2004),5,2,Pp.223–237.

Byung-Jun Yoon. 2009. Hidden Markov Models and their Applications in Biological Sequence Analysis. Current Genomics vol.10 page 402-415

C. Spampinato, S. Palazzo. 2012. Hidden Markov Models for Detecting Anomalous Fish Trajectories in Underwater Footage. 2012 IEEE International Workshop on Machine Learning for Signal Processing, Santander, Spain.

Jian Peng, Liefeng Bo, and Jinbo Xu. Conditional neural fields. In Proceedings of Neural Information Processing Systems (NIPS). 2009.

John D. Lafferty, Andrew McCallum, Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence. In ICML 2001.

Levesque, J.C., Morency, L.P. and Gagné, C., "Sequential emotion recognition using Latent-Dynamic Conditional Neural Fields", in Proc. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Shanghai, 2013, 1–6.

Madeo, R. C. B. ; Wagner, P. K. ; PERES, S. M.. A Review of Temporal Aspects of Hand Gesture Analysis Applied to Discourse Analysis and Natural Conversation. International Journal of Computer Science and Information Technology, v. 5, p. 1-20, 2013b.

Martin F. Lambert, Julian P. Whiting, Andrew V. Metcalfe. 2003. A Non-parametric Hidden Markov Model for Climate State Identification. Hydrology and Earth System Sciences, 7(5), 652-667.

Tamburini, F., Bertini, C., & Bertinetto, P. M. (2014). Prosodic prominence detection in Italian continuous speech using probabilistic graphical models. In Proceedings of Speech Prosody (SP-2014), Dublin, Ireland, pp. 285–289.

Zhang, S., 2012. Fuzzy-based latent-dynamic conditional random fields for continuous gesture recognition. Optical Engineering, 51(6), p.067202.