OPEN ACCESS
IJOICT | SUBMISSION
ISSN XXXX-XXXX
SOCJ.TELKOMUNIVERSITY.AC.ID/IJOICT
INTL. JOURNAL ON ICT VOL. XX, NO. XX, JUNE 2016
PP. XX-XX. DOI:XXX

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

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

Abstract

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

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

Received on xxx, accepted on xxx, published on xxx

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 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 this 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 this process is labeling.

Hidden Markov Models are a widely used method for sequence labeling. This method was developed by Rabiner for speech labeling. But its implementation has been widely used in many areas such as Bioinformatics [1], fisheries [2], meteorology [3], and health [4]. 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 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 [5]. This is the shortage of HMM because this calculation becomes impractical to represent data with some interacting features. To overcome this problem, conditional model is the best choice [6].

A well known conditional model for sequence labeling is Conditional Random Fields (CRF). CRF is able to 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 external 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. This 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 and Liefeng Bo [7]. This model is a combination from Conditional Random Fields and Neural Networks. Neural networks are useful for learning complex nonlinear relationship. This 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 developed from LDCRF. LDCRF with Neural Networks is called Latent-Dynamic Conditional Neural Fields (LDCNF) [8]. LDCNF consists of two layers, namely the layer of gate for learning non-linear relationship and dynamic layer of intrinsic structure. Therefore, this study proposes 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.

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

II. LITERATURE REVIEW

In this 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:

$X=x_1,x_2,x_3,x_4,\ldots..x_n$

$Y=y_1,y_2,y_3,y_4,\ldots..y_n$

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

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

Where

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

θ is 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 a number of hidden-state to define the structure intrinsic so probability of LDCRF for input X to label Y as follows:

$$p\,(y|x) = \frac{\exp\,[\sum_{i=1}^{k}\sum_{j=1}^{m}\theta_i\varphi_i(x, j, h_j, h_{j-1})]}{\sum_{l'}\exp\,[\sum_{i=1}^{k}\sum_{j=1}^{m}\theta_i\varphi_i(x, j, h_j, h_{j-1})]} \quad [2]$$

Feature function in LDCRF only occurs between 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 intermediate layer acts as a gate function. Thus the probability for input, X to label Y is defined as follows:

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

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:

$$p\,(y|x) = \frac{\exp\,[\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{g=1}^{n}\tau(\alpha_g\varphi_i(x,j,y_j,y_{j-1}))]}{\sum_{l'}\exp[\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{g=1}^{n}\tau(\alpha_g\varphi_i(x,j,y_j,y_{j-1}))]} \quad [4]$$

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 [1], [2]. In general, the number of clusters and its centroid are initialized at the beginning. 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 in order to reach the objective function or value of the expected error.

III. DATASET AND SYSTEM ARCHITECTURE

A. Dataset

Performance of methods will be tested by gesture phase segmentation. This 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 Microsoft Kinect Sensor [9]. From the dataset is provided, this 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 this study, the timestamp is not used.

B. System Architecture

System Architecture used in this study is based on research conducted by Fabio Tamburini, Chiara Bertini, Pier Marco Bertinetto for prosodic Prominence detection [10]. The illustration is shown in Fig. 1.

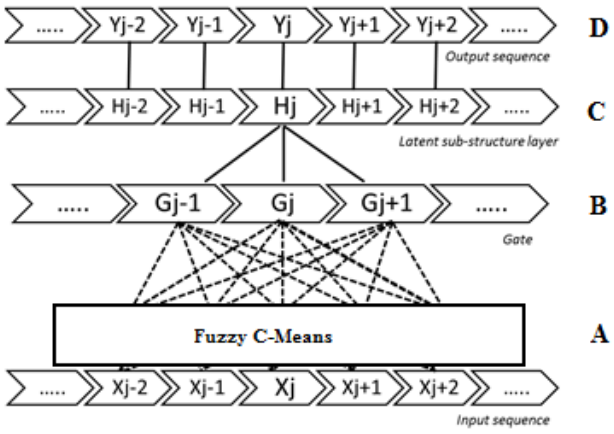

Fig. 1. Architecture

There are no intermediate layers in system architecture of CRF. This is different from three bsic 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 the 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 number of fold that will be used. In this study, we use three-Cross-validation so there will 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.

A. Number of clusters

Fuzzy C-Means clustering groups the objects into several clusters. This cluster number is initialized at the beginning, so that the necessary experiments are done to find the optimal number of clusters for gesture recognition. 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 this experiment, fuzziness degree is 1.05.

From the results in Table 2, it can be known that CRF best performance is obtained when the number of clusters used is four which has the highest sensitivity, but lowest running time and the worst performance occur in the use of eight clusters. Performance decreased with the increase in the number of clusters cause a decrease in sensitivity, but the increase in running time. However, despite the downward trend, but by using seven clusters is better than using six clusters.

Implementation using LDCRF has a sensitivity of CRF between 0.32 to 0.49 so that CRF is better than LDCRF and also running time of LDCRF requires more time ranging from 647.70 to 899.49 seconds. Here it can be seen that the use of hidden-state does not provide effective process. On the other hand the gate function of intermediate layers gives a significant result compared to the hidden-state in LDCRF. Sensitivity for the gesture recognition increased to range from 0.39 until 0.81. Significant results can be seen that it has less running time than CRF and LDCRF. This 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 CRF and LDCRF, the highest performance is obtained when the number of clusters used is 4 and the running time is only 36.17 seconds. Trend decrease in sensitivity occurs at the time of adding the number of clusters used. Performance gate function also proved capable of raising sensitivity that ranges between 0:40 to 0.83 but the increase in running time also occurs in LDCNF due to the mechanism of a merger between the CNF and LDCNF.

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

TABLE II
NUMBER OF CLUSTERS

Methods	Cl	Sensitivity	Running Time
CRF	4	0.47	35.06
	5	0.46	35.55
	6	0.33	80.00
	7	0.38	56.58
	8	0.35	82.73
LDCRF	4	0.49	647.70
	5	0.41	670.58
	6	0.32	778.99
	7	0.52	899.49
	8	0.32	873.34
CNF	4	0.81	36.17
	5	0.76	36.22
	6	0.38	77.65
	7	0.48	52.15
	8	0.39	74.24
LDCNF	4	0.83	253.58
	5	0.78	241.96
	6	0.38	311.99
	7	0.49	335.91
	8	0.40	310.68

B. Fuzziness

To be able in finding the optimal degree fuzziness 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. This test is only implemented four clusters according to the results already obtained previously. From the results in the Table 3 is known that the implementation of the CRF that increasing degrees of membership then an upside sensitivity and also the increase in running time. By using 1.05, sensitivity has the highest sensitivity and to raise it to 1.1, sensitivity is reduced by half. This means that the increase is not necessary because it makes the performance of CRF be decreased. If the hidden-state was added as an intermediate layer in LDCRF, then the best performance is still obtained by using 1.05 as the degree of membership. When compared with the CNF, running time of the gesture recognition only has less difference than CRF even when the degree of membership that is set is 1.2 to 1.4, running time CNF is less than the CRF but sensitivity has a significant difference compared to CRF.

On the other hand, LDCNF has the highest performance in this implementation when the degree of membership used was 1.05. Sensitivity and running time for this degree was 0.83 and 253.38. Improved degree of membership has a trend decline in sensitivity, but improved running time.

TABLE III
FUZZINESS

Methods	FD	Sensitivity	Running Time
CRF	1.05	0.47	35.06
	1.1	0.29	35.55
	1.2	0.17	80.00
	1.3	0.05	56.58
	1.4	0.09	82.73
LDCRF	1.05	0.49	647.70
	1.1	0.29	670.58
	1.2	0.15	778.99
	1.3	0.06	899.49
	1.4	0.07	873.34
CNF	1.05	0.81	36.17
	1.1	0.65	36.22
	1.2	0.54	77.65
	1.3	0.37	52.15
	1.4	0.31	74.24
LDCNF	1.05	0.83	253.58
	1.1	0.65	241.96
	1.2	0.53	311.99
	1.3	0.38	335.91
	1.4	0.31	310.68

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 the Table 4. From the table it is known that the use by using only the basic classifiers only to achieve sensitivity of 0.10 to 0.29. The worst performance was shown by LDCRF which has the lowest sensitivity and the highest running time. This indicates that the use of hidden-state is not effective for this 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.

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. This 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 sensitivity is obtained FLDCNF.

TABLE IV
COMPARISON

Methods	Sensitivity	Running Time
CRF	0.10	105.23
LDCRF	0.07	1540.53
CNF	0.19	47.27
LDCNF	0.29	606.06
FCRF	0.47	35.06
FLDCRF	0.49	647.70
FCNF	0.81	36.17
LFDCNF	0.83	253.58

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 really 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 proven effective to find the right feature subset so the accuracy increased but the execution time decreased with this feature subset. The other hand, the combination of Fuzzy C-Means Clustering as Clustering and the base classifiers, give the better performance than without the use of Fuzzy C-Means Clustering. If no fuzzy, basic classifiers have the sensitivity ranged between 0.10 to 0.29 and running time ranged between 47.27 to 1540.53 seconds. Meanwhile, with fuzzy, sensitivity ranged from 0.47 to 0.83 and the running time decreases ranged from 35.06 to 647.70 seconds. This indicates the use of Fuzzy C-Means Clustering can filter the feature to discover new optimal features in the classification process. The discovery of the optimal features has an 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

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

[2] 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.

[3] 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.

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

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

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

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

[8] 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.

[9] 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.

[10] 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.