Gesture Recognition using Latent-Dynamic based Conditional Random Fields and Scalar Features

I N Yulita[1*], M I Fanany[2] and A M Arymurthy[2]

[1]Departemen Ilmu Komputer, Universitas Padjadjaran, Jl. Raya Bandung – Sumedang Km. 21, Sumedang 45363, Indonesia
[2]Fakultas Ilmu Komputer, Universitas Indonesia, Kampus UI Depok, Depok16424, Indonesia

Email: intan.nurma@unpad.ac.id

Abstract. The need for segmentation and labeling of sequence data appears in several fields. The use of the conditional models such as Conditional Random Fields is widely used to solve this problem. In the pattern recognition, Conditional Random Fields specify the possibilities of a sequence label. This method constructs its full label sequence to be a probabilistic graphical model based on its observation. However, Conditional Random Fields can not capture the internal structure so that Latent-based Dynamic Conditional Random Fields is developed without leaving external dynamics of inter-label. This study proposes the use of Latent-Dynamic Conditional Random Fields for Gesture Recognition and comparison between both methods. Besides, this study also proposes the use of a scalar features to gesture recognition. The results show that performance of Latent-dynamic based Conditional Random Fields is not better than the Conditional Random Fields, and scalar features are effective for both methods are in gesture recognition. Therefore, it recommends implementing Conditional Random Fields and scalar features in gesture recognition for better performance

1. Introduction

The need for segmentation and labeling of sequence data appears in several fields [1]. Many methods were developed to solve this problem. One of them is Conditional Random Fields [2]. This technique has been implemented for the recognition of the object [3], speech [4], and relational learning [5]. The advantage of this technique is its ability in processing the knowledge through the use of a series of sequences to form a model which can overcome the bias of the data. If the data has a bias that is the resemblance between several labels, then Conditional Random Fields will pay attention to the output/label sequence.

However, Conditional Random Fields did not use intrinsic structure from its observation so that further developed state Hidden-state Conditional Random Fields [6,7]. But Hidden-state Conditional Random Fields would eliminate the ability to capture the dynamics of extrinsic between their labels by CRF [8]. Thus, it developed into Latent-Dynamic Conditional Random Fields to merge the two mechanisms [9].

On the other hand, the recording of data gesture is usually a point of articulation which is every point consist of several coordinate axes. Therefore, it is important to use only important features for its modelling. This process is can be done by using feature extraction. Thus, this paper offers the use of

scalar features to improve the accuracy of Latent-Dynamic Conditional Random Fields on gesture recognition. Its use reduces the number of features so make modeling time will be reduced**.**

2. Experimental Method

In this chapter will describe methods that will be used in gesture recognition is Conditional Random Fields, Latent-Dynamic Conditional Random Fields, and Scalar Features. Also, it will be explained the used dataset and the flowchart of the proposed method.

2.1. Method

Features on the same position with different coordinate axes will be changed to a scalar value by using Euclidean distance. The calculation is as follows:

$$s_{ij} = \sqrt{x_{ij}{}^2 + y_{ij}{}^2 + z_{ij}{}^2} \tag{1}$$

Where

x_{ij}, y_{ij}, z_{ij} are the three values of the coordinates of the data for the same position

s_{ij} is a scalar value at a specific position

Illustration of the scalar features is described in Figure 1. If the data have coordinates in the x, y, and z coordinates, then the scalar features are the calculation of Euclidean distance.

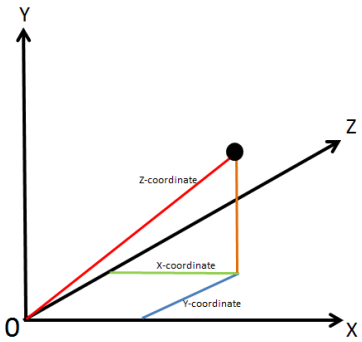

Figure 1. Three coordinate points of X, Y, and Z produce a scalar feature

These scalar features become the input of gesture recognition. In this study, the recognition is also processed by using the sequential information among labels/classes of the dataset. Thus, to capture both factors: the features of the data and the interacting labels, this study proposes LDCRF which it can capture both internal structure patterns from the features of the data, and internal structure patterns of the interacting labels[10]. In contrast to CRF, LDCRF also adds some hidden state variables that models internal structure patterns. These hidden state variables is placed in latent sub-structure layer. This layer serves to filter the input sequence. The other hand, HCRF also uses this layer but it ignores the external structure of the interaction of the label. So the lack of CRF and HCRF are resolved by building of the model of LDCRF.This method is a combination of CRF and HCRF. The modelingof LDCRF is presented in Figure 2,the architecture used is based on research conducted by Fabio Tamburini, Chiara Bertini, Pier Marco Bertinetto for prosodic Prominence detection [11]

LDCRF is one type Probabilistic Graphical Models that represent the model through a probability distribution [12]. The label is determined based on the maximum value of the probability as follows:

$$p(y|x) = \frac{\partial}{\sum_{lr} \partial} \tag{2}$$

$$\partial = \exp\left[\sum_{a=1}^{b} \sum_{c=1}^{d} w_a f_a(x, c, h_c, h_{c-1})\right] \tag{3}$$

Where
$f_a$is feature function
$w_a$is weight of feature function
$h_c$is hidden state

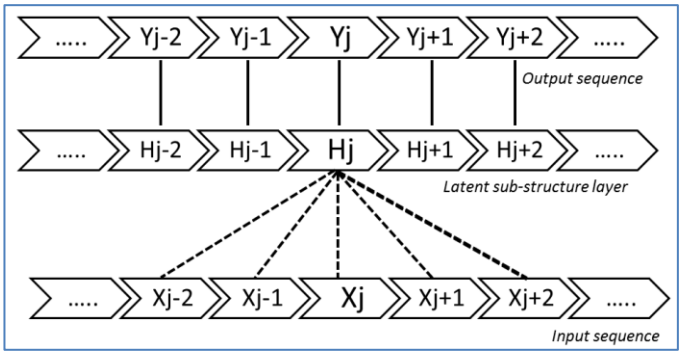

Figure 2. The architecture of Latent-Dynamic Conditional Random Fields

The weights of the feature function are obtained from the training data using optimization techniques. Some widely used optimization technique is Conjugative Gradient and Broyden-Fletcher-Goldfarb-Shanno. While the feature function is obtained by latent sub-structure layer. A feature function defines a value (0 or 1) of the sequence x, sequence c of two hidden state being observed. For example the value of feature function is 1 if the the position is 3 and also the hidden state 3rd and 4th are the representation of the label "rest" and "preparation".

2.2. Flowchart of the Recognition

The recognition process of this research is based on scheme that it will be illustrated in Figure 3.

Figure 3. The Proposed Method

In this study, thedata used aresegmentation of gesture phase from some videos, which was developed by the University of Sao Paulo [13]. In each of these segmentation presented six points of articulation that is left hand, right hand, left wrist, right wrist, head, and spine, which eachposition have coordinates of x, y, and z, as illustrated in Figure 4. The purpose of gesture recognition is the prediction of gesture phase (rest, preparation, stroke, hold, and retraction).

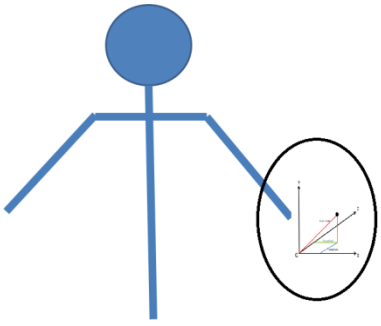

Figure 4. The skeleton of dataset

Then the dataset is preprocessed in the first step. This preprocessing includes feature selection. Through preprocessing, only the relevant features are used for building modeling. From the dataset, they have the feature of timestamp. The preprocessing will eliminate this feature because this feature is not helpful to build the proposed modelling.

The second step is feature extraction. The obtained features are six scalar features of the 18 features. This process is done by using the calculation of scalar features. Then these x scalar features are modeled using Latent-Dynamic Conditional Random Fields. From these resulting models, the maximum probability of data to the label will be the final label.

3. Result and Discussion

Some experiment tests have been done using the dataset in this research. The evaluation scheme isbased on 3-cross-validation and calculation of the accuracy and execution time for both methods. The test results are described in the next section.

2.3. Accuracy

Effectiveness in the process of recognition will be tested based on the percentage of the accuracy so that it can evaluate the performance of the resulting model to the data being tested.The accuracy is presented by %.

Table 1. The Accuracy of this gesture recognition

	Without Scalar Features		With Scalar Features	
Accuracy	CRF	LDCRF	CRF	LDCRF
First Iteration	50.14	54.14	57.75	54.61
Second Iteration	50.00	50.00	78.08	71.08
Third Iteration	64.14	50.07	64.70	68.06
Average	54.85	51.40	66.84	64.59

From Table 1, it can be seen that the LDCRF implementation without scalar features of the first iteration gets a higher accuracy than CRF. But accuracy is lower in the second and third iterations so that the average accuracy is also lower than the CRF. While the implementation of the scalar features, LDCRF the first and second iteration is lower than the CRF. But in the third iteration, higher accuracy. Similarly, the implementation without the scalar features, accuracy LDCRF on average lower than CRF. This indicates that the addition of hidden layer LDCRF no more effective for the gesture recognition even worsens the performance of the external structure.

Also, the table 1 is known that the use of scalar features can boost recognition accuracy of both CRF and LDCRF movement. In the first iteration, the accuracy of CRF increased 7.61% and amounted 0.47 LDCRF%. hile in the second iteration, respectively 28.08% and 21.08% and in the third iteration of 0.56% and 7.99%. So that the average increase was 11.99% and 13.19%. With the increasing accuracy, it can be seen that the effective use of scalar features for gesture recognition.

2.4. Running Time

Another evaluation parameter is therunning time in order to know how long it takes to produce a model in the recognition process. The execution time is presented by seconds.

Table 2. The running time of this gesture recognition

Running Time	Without Scalar Feature		With Scalar Feature	
	CRF	LDCRF	CRF	LDCRF
First Iteration	115.06	1671.98	48.67	969.42
Second Iteration	49.00	1891.90	68.23	1106.89
Third Iteration	52.71	3035.23	66.82	1154.95

Average	72.26	2199.70	61.24	1077.09

In Table 2, it can be seen that the implementation without the scalar features, running time of CRF and LDCRF are 115.06 and 1671.98 seconds. The running time is decreased by more than half in the second iteration of CRF but a slight increase on the third iteration. This is in contrast with LDCRF for the second iteration; there is an increase almost be doubled in the third iteration. Based on the calculation of average it was found that LDCRF takes three times greater than the CRF.

With the use of scalar features, the running time of CRF and LDCRF in the first iteration have decreased by about 50%. But in the second and third iteration, CRF increased slightly, but LDCRF sharp decline. From the calculation of averages, CRF and LDCRF decrease running time when used the scalar features.

4. Conclusion

From the results of experiments that have been done, it is known that CRF is superior to LDCRF. This indicates that the use of hidden state will not necessarily increase the accuracy of labeling. Regarding the running time, the LDCRF takes than the CRF. For faster labeling, the CRF is selected for our gesture recognition. The use of scalar features proven to be effective for our gesture recognition as shown by higher accuracy and lower running time

5. References

[1] Lafferty J D, McCallum A and Pereira F C N 2001 *Conditional random fields: Probabilistic models for segmenting and labeling sequence data* I*CML* (San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.)

[2] Sutton C and McCallum A 2011 *An Introduction to Conditional Random Fields* (Foundations and Trends in Machine Learning)

[3] Quattoni A,Collins M and Darrell T 2004 *Conditional random fields for object recognition* (MIT Press) pp. 1097–104

[4] Charles S and Andrew M 2006 *An Introduction to Conditional Random Fields for Relational Learning* (MIT Press) chapter 4 p 93–128

[5] Zweig G and NguyenP 2009 *A segmental CRF approach to large vocabulary continuous speech recognition* (IEEE ASRU Workshop)

[6] Gunawardana A, Mahajan M,Acero A, and Platt J C 2005 *Hidden conditional random fields for phone classification* (InInterspeech) p 1117–20

[7] Quattoni A, Wang S, Morency L P, Collins M and Darrell T 2007 *Hidden conditional random fieldsIEEE Trans. Pattern Anal. Mach. Intell.***29** 1848–52

[8] Tong Y,Chen R and Gao J 2015 *Hidden State Conditional Random Field for Abnormal Activity Recognition in Smart Homes*

[9] Boonsuk S, Suchato A, PunyabukkanaP C, Wutiwiwatchai N and Thatphithakkul 2014 *Mathematical Problems in Engineering*

[10] Wittner C, Schauerte B and Stiefelhagen R 2015*What's the point? Frame-wise Pointing Gesture Recognition with Latent-Dynamic Conditional Random Fields*

[11] Rahimi A M, Ruschel R, and Manjunath B S 2016 *UAV Sensor Fusion with Latent-Dynamic Conditional Random Fields in Coronal Plane Estimation CVPR*

[12] Tamburini F, Bertini C and Bertinetto P M 2014 *In Proceedings of Speech Prosody* (SP-2014) 285–9

[13] MadeoR C B, Wagner P K and Peres S M 2013 *International Journal of Computer Science and Information Tech.* **5** 1-20

Acknowledgments
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.