Probabilistic Inference of Facial Action Unit by Dynamic Bayesian Network for Image Sequence

Yee Koon Loh
Intelligent Biometric Group
School of Electrical and Electronic Engineering,
Engineering Campus, Universiti Sains Malaysia
Pulau Pinang, Malaysia
lyk10_ee045@student.usm.my

Shahrel A. Suandi
Intelligent Biometric Group
School of Electrical and Electronic Engineering,
Engineering Campus, Universiti Sains Malaysia
Pulau Pinang, Malaysia
shahrel@usm.my

Abstract—*Facial expressions play important role in many applications especially in video surveillance and face identification. The movements of the facial muscles convey the emotional state of an individual. However, reliable expression recognition by machine is still a challenge. This paper proposes a probabilistic framework by Dynamic Bayesian Network (DBN) to study the relationship among Action Units (AUs) that are involved in expressions appearing in an image sequence from video.The image sequence involves the evolution of expressions from neutral to apex and back to neutral which gives very important information to study the temporal changes of AUs. The image sequence as training data is used to learn parameters for DBN which gives transition and observation model of the network. A set of test data is used to be input to the network as evidence to infer the presence or absence of the AU in the image. Experiments are carried out to show the performance of the network which are modeled to represent the true relationship among the AUs.*

Keywords—Facial Action Unit; Dynamic Bayesian Network; Probabilistic inference

I. INTRODUCTION

Facial expression recognition has been an active topic among researchers for over two decades. It is a very challenging task due to a lot of difficulties and limitations. Facial expression is a complex study due to different expressions for different people and the expressions differ at different age. Computer vision techniques have been widely used to recognize facial expression from images and video but due to pose, illuminations and complicated combination offacial expression, a stable facial expression recognition system has yet to be developed.

Facial Action Coding System(FACS) has been widely used to study the facial expression in computer-human interaction.FACS is the most objective and comprehensive coding system in the behavioural sciences. Ekman and Keltner[1] developed the system which describes facial expressions in terms of Action Units (AUs). FACS consists of 46 AUs which are primarily related to facial muscle movements. However, this system requires human to go through the learning and training processes to become expert in coding FACS. FACS is developed for coding by hand, using human experts. It is currently performed by trained experts who make perceptual judgement of video sequences. The limitations is that the system requires approximately 100 hours to train a person to make these judgements reliably and pass a standardized test for its reliability[2].

Due to richness of facial expressions, it is important to learn the relationship among AUs. It is hard to find that a facial expression is determined by a single AU. The AUs are related to each other as it involves the muscles of human face. In this paper, the relationship of 14 main AUs that were involved in 6 basic expressions (anger, disgust, fear, happiness, sadness, surprise) as described by FACS are studied and modeled using Dynamic Bayesian Network (DBN).

II. RELATED WORKS

Facial expression analysis can be studied in two types: appearance-based and geometric-based. Overview of various approaches like optical flow method, local binary patterns, pyramid of histogram of gradient (PHOG) and local phase quantisation (LPQ) method and facial action coding system (FACS) were reported in survey done by Lonare et. al[3]. Challenges in face expression recognition in real life applications such as pose, illumination and occlusion have been issued in [4]. The survey gives also an overview of the methodology to be followed for facial expression recognition.

Moore et. al [5] carried out experiments to investigate how pose affects facial expression recognition. The effects of pose on facial expression recognition using variations of LBPs at different resolutions and different grid sampling sizes were investigated. Mahoor et. al [6] proposed sparse learning approach for AU combination classification. Gabor features were extracted at the location of facial landmark points, which have been extracted using Active Appearance Model (AAM) to represent facial images. A dictionary was developed to recognize the combination of facial AUs using L1-norm minimization.

Yan Tong et. al [7] proposed a unified facial action model based on the DBN to systematically discover and learn such relationships, and then combine them with the image observations to perform a robust and reliable recognition of spontaneous facial action. A system to recognize facial action unit by exploiting their semantic and dynamic relationships using Dynamic Bayesian Network (DBN) is proposed in [8].

978-1-4799-8389-6/15/$31.00

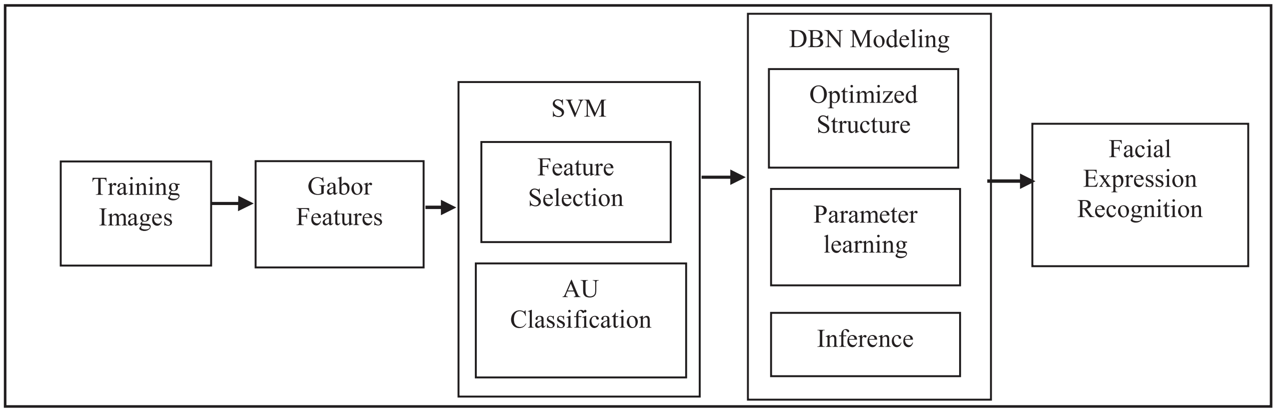

Fig 1. Overall system.

Probabilistic relationships among various AUs are presented and temporal changes in facial action units are accounted.

III. System Overview

Facial expression recognition from images consists of different block of processes. Facial features have to be extracted from images and classified according to AUs. This paper focuses on DBN modeling of AU recognition which will be used in facial expression recognition in future work. Cohn-Kanade [9] database has been used as training images to get the important facial features for feature selection and AU classification. Gabor features are computed from the input images in 5 scales and 8 orientations to represent the image. All the features are then used as input to Support Vector Machine (SVM) [10] to train the data to get a model for each AU. The model is used to classify the test data which will be used as evidence for DBN inference. Overall system is shown in Fig. 1. The following section will describe the procedure for DBN modeling.

IV. Dynamic Bayesian Network

Dynamic Bayesian network (DBN) is a Bayesian network that represents a temporal probability model. The topology of connections between successive slices of time and between the state and evidence variables are specified. Let AU_i be the query variable, $\boldsymbol{O}$ be the set of evidence variables, o_i be the observed values for them, and let $\boldsymbol{Y}$ be the set of remaining unobserved variables (the past states). The query is to calculate posterior distribution of AU_i given o_i, is defined as in Eq. (1),

$$P(AU_i \mid o_i) = \alpha P(AU_i, o_i) = \alpha \sum_y P(AU_i, o_i, y) \quad (1)$$

where summation is over all possible combinations of values y of the unobserved past states $\boldsymbol{Y}$ and $\alpha = 1 / P(o_i)$ is the normalisation factor (probability of evidence). Variables AU_i, $\boldsymbol{Y}$ and $\boldsymbol{O}$ constitute the complete set of variables. i denotes 14 AUs involved. The optimized BN structure in [11] has been used in our DBN model with additional observed nodes added for each AU nodes. Therefore each time slice of DBN consists of 28 nodes, i.e 14 AUs nodes denoted as {AU1,AU2...,AU27}are hidden nodes to be inferred and each of these nodes are linked to an observed node. The observed nodes are denoted as {O1,O2,...,O27}. From time slice *t-1* to *t*, there are only two arcs that show transition from different AU in time slice *t-1* to time-slice *t* which are AU2 to AU5 and AU12 to AU6 as reported in [8] which show significant relationship in the database. Besides that, each AU at time slice *t-1* is connected to AU at time slice *t*. For example, $AU1_{t-1}$ to $AU1_t$. For ease of visualization purpose, these arcs are not shown in the figure. DBN structure is shown in Fig. 2 to show temporal relationship from time-slice *t-1* to time-slice *t*.

A. Parameter Learning

After constructing the DBN structure for AUs, it is necessary to learn parameters from the training data for the DBN. Besides learning the Conditional Probability Table (CPT) for each slice, the transition probabilities between slices are learned. The structure is partial observability, which means some nodes are sometimes hidden, the likelihood does not decompose. In this case, Expectation-Maximization algorithm (EM) [12] is used to find local maximum. EM iteratively maximizes the expected complete-data log-likelihood, which does decompose into a sum of local terms. This process sets the parameters in a DBN to their Maximum Likelihood (ML) or Maximum a Posteriori (MAP) values using batch EM. After learning the parameters, the DBN is ready to be used to do inference on the desired nodes.

B. DBN Inference

DBN inference is implemented to get the posterior probability of AU given the evidence nodes which tells us the AU is likely to present or absent in the image. Then, AU recognition is done by comparing the output with ground truth of the AU in the image.

DBN inference is commonly known as “unroll” DBN for desired number of time steps and treat as a static BN. Evidence at appropriate times are applied and then apply the

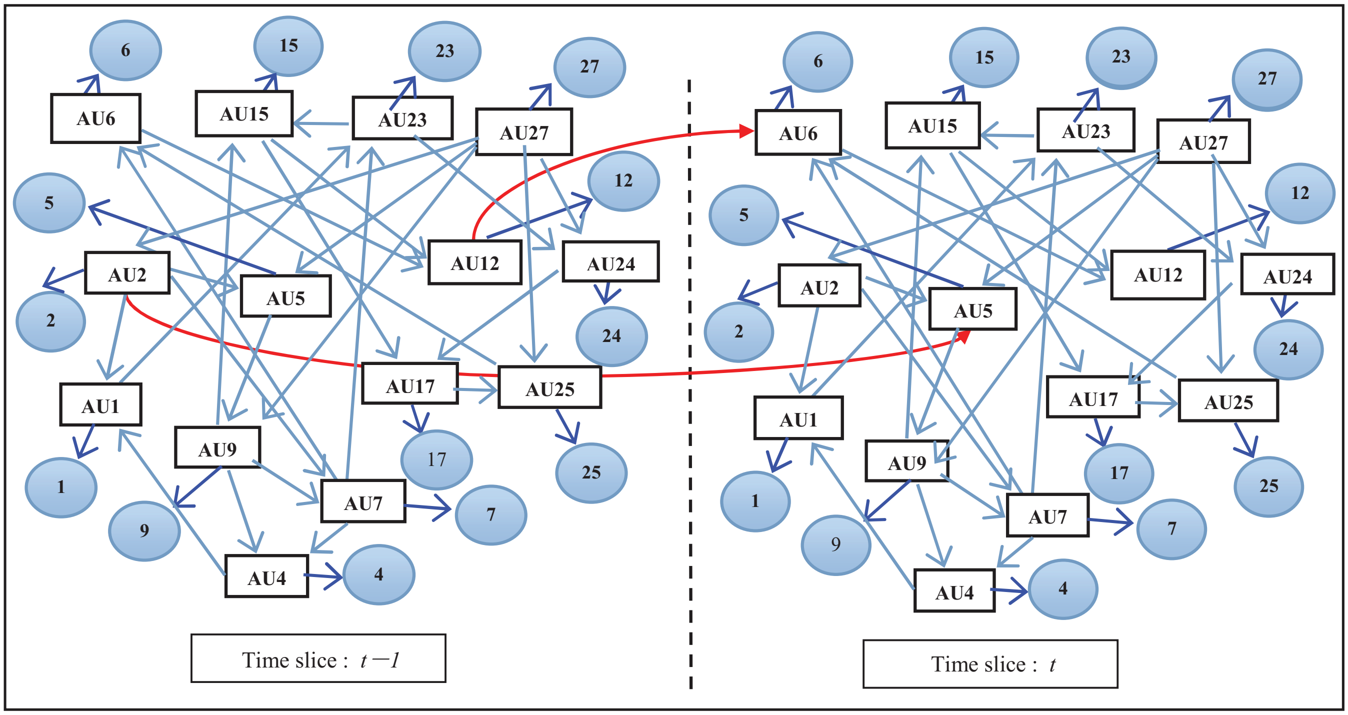

Fig. 2 : DBN Modeling for AU recognition. Square nodes denote the hidden variables, shaded circles denote observation node for each AU. Red arc represent temporal link of different AU at time slice *t-1* to time slice *t*. For illustration purpose, AU at time slice *t-1* to same AU at time slice is not shown. For example, AU1$_{t-1}$ to AU1$_t$ is not shown.

technique of exact inference on any time slice of static BN. Unrolling DBN has its disadvantage for our work due to large number of nodes in the structure. With many time steps, unrolled DBN becomes huge and may cause the inference algorithms to run out of memory or take too long.

A DBN represents a process that is stationary and Markovian. The node relationships within a time slice t and the transition function from time slice t to time slice $t+1$ do not depend on t. The transition function depends only on the immediately preceding time slice and not on any previous time slices (e.g., no arrows go from time slice t to timeslice $t+2$). Therefore, the set of nodes in a time slice d-separate the past from the future. A DBN can be represented using just a 2-slice Temporal Bayesian Network (2TBN) that represents the first two time slices of a process and do inference using just this structure too.

Hence, in this work, junction tree algorithm is used to compute DBN inference process. Kevin Murphy's Interface algorithm [13] which runs junction tree algorithm as a subroutine is employed. The process is described briefly as below:

1) Before time is incremented:

– Get current junction tree (if time = 0, this is J_1; otherwise, this is J_t).

– Update beliefs in current junction tree if necessary.

– Get α_t by marginalizing out-clique potential in current junction tree down to outgoing interface potential.

2) Increment of time

3) After time is incremented:

– Get current junction tree (this will always be J_t since time > 0).

– Multiply α_t onto in-clique potential in current junction tree. The out-clique potential is the only quantity we need to keep when we advance from time step to time step.

In Bayesian concept learning, the posterior of each AU is simply the likelihood times the prior, normalized. Likelihood is computed by considering all the evidences and then marginalization can be done on inferred AU nodes.

V. Experiment Results

A. SVM classification

AU classification is first done on Cohn-Kanade database using SVM. 70% of images by subjects are used as training data while the rest of the subjects are used as testing data. For each sequence of images, the last frame which contained the expression apex by the subject was taken as positive examples. Negative examples consisted of all apex frames that did not contain the target AU plus neutral images obtained from the first frame of each sequence of images. However, the number of positive examples of each AU is not consistent. Table 1 shows the number of examples for each AU and the classification result on test data. (Accuracy: Percentage of the positive and negative examples are classified correctly. Hit:

true positive rate and FA: false alarm rate). The average accuracy for all AUs achieve 91% ,with a positive rate of 66% and a false-alarm rate of 6% for the 14 AUs. The predicted class of each image (either 1 for presence or 0 for absence) for each AU are used as evidence for DBN inference. The results show reliability of SVM to generate evidence for DBN inference later.

TABLE I. SVM CLASSIFICATION RESULT FOR EACH AUS. ACC: PERCENTAGE OF CORRECTLY CLASSIFIED EXAMPLES. HIT: TRUE POSITIVE RATE. FA: FALSE ALARM RATE.

AU	Name	N	Acc (%)	Hit (%)	FA (%)
1	Inner brow raise	143	90	64	5
2	Outer brow raise	91	95	64	0
4	Brow corrugator	153	83	82	16
5	Upper lid raise	78	95	58	1
6	Cheek raise	103	85	68	12
7	Lower lid tight	105	85	62	11
9	Nose wrinkle	50	94	62	3
12	Lip corner pull	112	98	84	0
15	Lip corner depress	74	94	65	4
17	Chin raise	157	90	75	6
23	Lip tighten	38	95	33	1
24	Lip press	43	94	53	3
25	Lips part	294	84	86	16
27	Mouth stretch	75	97	73	0
	Average		91	66	6

B. *DBN inference*

For DBN learning process, RPI ISL[8] database is used as training data to get the parameters for the DBN. Cohn-Kanade database has limited number of images and not all images in a sequence is labeled. Only the last image of a sequence is labeled as it is expression apex. Therefore, it is used as test data for classification purpose. RPI ISL database consists of 6 subjects showing multiple expressions with 899 images for each sequence. It shows evolution of an expression from neutral to apex and back to neutral. It is complied with our aim to study the temporal changes from frame to frame in an image sequence. Results are shown in Table 2. Average recognition rate for all AUs is 67%. There are 4 AUs that shows low recognition rate which are AU5, AU9, AU12 and AU17. If we refer back to the structure in Fig. 2, we notice that AU5 is linked to AU9. These two AUs have low hit rate due to low number of examples in training data for facial features in SVM. This can be improved by adding more positive examples to train the AUs in feature selection process. On the other hand, AU12 and AU17 share a parent node. It shows us that AU12 and AU17 are not likely to appear with AU15.

TABLE II. AU RECOGNITION BY DBN

AU	Accuracy (%)
1	90
2	95
4	79
5	5
6	84
7	85
9	17
12	17
15	94
17	10
23	95
24	94
25	84
27	91
Average	67

VI. CONCLUSION AND FUTURE WORK

A DBN is constructed based on two AU labeled database to study the relationship of AUs. Since this is a probabilistic modeling, inference on AU nodes are much rely on the prior of the nodes, likelihood and number of data. The number of training data for each AU is not consistent which affect the inference result because every node is inter-connected in the network. More experiments will be carried out to study the relationship for AU for video/image sequence. Our aim is to confirm a structure that represents the relationship among AUs for image sequence. In future work, AU combination for facial expression will be studied. The structure will be revised with different factors that can improve the robustness of the system. Temporal changes and link of AUs in an image sequence leads to different facial expression and it is important for video surveillance system.

ACKNOWLEDGEMENT

This research was funded by Universiti Sains Malaysia Research University Grant (1001/PELECT/814208).

REFERENCES

[1] P. Ekman and W. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement." Consulting Psychologists Press, Palo Alto, CA, 1978.

[2] Bartlett, Marian Stewart, et al. "Automatic recognition of facial actions in spontaneous expressions." Journal of Multimedia 1.6 (2006): 22-35.

[3] Lonare, Ashish, and Shweta V. Jain. "A Survey on Facial Expression Analysis for Emotion Recognition." International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 12, 2013.

[4] Khatri, Nidhi N., Zankhana H. Shah, and Samip A. Patel. "Facial Expression Recognition: A Survey." International Journal of Computer Science and Information Technologies, Vol. 5 (1) , pp.149-152, 2014.

[5] Moore, S., and R. Bowden. “Local binary patterns for multi-view facial expression recognition.” Computer Vision and Image Understanding 115.4 (2011): 541-558.
[6] Mahoor, Mohammad H., et al. “Facial action unit recognition with sparse representation.” Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on. IEEE, 2011.
[7] Yan Tong, Wenhui Liao, Zheng Xue and Qiang Ji, “A Unified Probabilistic Framework for Spontaneous Facial Activity Modeling and Understanding.” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.
[8] Yan Tong, Wenhui Liao and Qiang Ji, “Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships.” IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 29, No. 10, October 2007.
[9] Kanade, T., Cohn, J. F., & Tian, Y. (2000). “Comprehensive database for facial expression analysis.” Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), Grenoble, France, 46-53.
[10] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM : a library for support vector machines.” ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.
[11] Yee Koon Loh , Shahrel A. Suandi , “Optimized Structure for Facial Action Unit Relationship Using Bayesian Network.” World Journal of Computer Application and Technology, 2 , 133 - 138, 2014.
[12] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm.” Journal of the Royal Statistical Society. Series B (Methodological) :1-38, 1977.
[13] Kevin Murphy, “Dynamic Bayesian Networks: Representation, Inference and Learning.” PhD thesis, U.C. Berkeley, 2002.
[14] Tian, Yingli, Takeo Kanade, and Jeffrey F. Cohn. “Facial expression recognition.” Handbook of face recognition, Springer London, 2011, 487-519.
[15] Zavaschi, Thiago HH, et al. “Fusion of feature sets and classifiers for facial expression recognition.” Expert Systems with Applications 40.2 (2013): 646-655.