Inferring Facial Action Units with Causal Relations

Yan Tong, Wenhui Liao, Qiang Ji
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, NY 12180-3590, USA
{*tongy2, liaow, jiq*}*@rpi.edu*

Abstract

A system that could automatically analyze the facial actions in real time have applications in a number of different fields. However, developing such a system is always a challenging task due to the richness, ambiguity, and dynamic nature of facial actions. Although a number of research groups attempt to recognize action units (AUs) by either improving facial feature extraction techniques, or the AU classification techniques, these methods often recognize AUs individually and statically, therefore ignoring the semantic relationships among AUs and the dynamics of AUs. Hence, these approaches cannot always recognize AUs reliably, robustly, and consistently.

In this paper, we propose a novel approach for AUs classification, that systematically accounts for relationships among AUs and their temporal evolution. Specifically, we use a dynamic Bayesian network (DBN) to model the relationships among different AUs. The DBN provides a coherent and unified hierarchical probabilistic framework to represent probabilistic relationships among different AUs and account for the temporal changes in facial action development. Under our system, robust computer vision techniques are used to get AU measurements. And such AU measurements are then applied as evidence into the DBN for inferencing various AUs. The experiments show the integration of AU relationships and AU dynamics with AU image measurements yields significant improvements in AU recognition.

1 Introduction

Ever since Ekman and Friesen [4] developed Facial Action Coding System (FACS), various methods have been proposed to automatically identify action units (AUs). A system that can recognize AUs in real time without human intervention can be applied in many application fields, including automated tool for behavioral research, videoconferencing, affective computing, perceptual human-machine interfaces, 3D face reconstruction and animation, and so on.

In general, these previous work can be summarized into five groups: facial motion extraction based on optical flow [21][5], difference image based methods [6], statistical model based methods [12][11], specific feature based methods [14][17][15][23], and machine-learning based method [1]. These methods differ either in the feature extraction techniques, or the AU classification techniques, or both. However, their common point is only attempting to classify each AU independently and statically, but ignoring the semantic relationships among AUs and the dynamics of AUs. Hence, these approaches cannot always recognize AUs reliably, robustly, and consistently due to the richness, ambiguity, and dynamic nature of facial actions. This motivates us to exploit the relationships among the AUs as well as the temporal development of each AU in order to improve AU recognition.

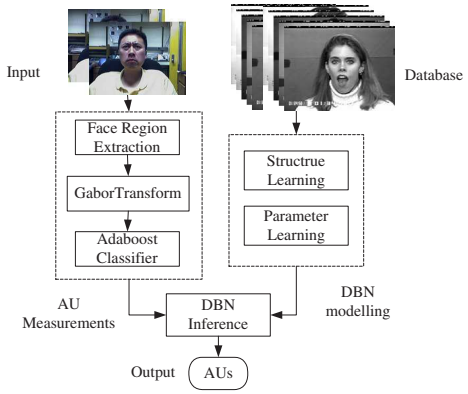


Figure 1. The flowchart of the real-time AU recognition system.

Although each AU is contracted by different facial muscles, it is rare that single AU occurs alone in spontaneous facial behavior. Instead, some patterns of AU combinations appear frequently to express natural human emotions. Thus, some of the AUs would appear simultaneously for the most time. On the other hand, it is nearly impossible for some AUs appearing together. Furthermore, AUs usually change over time. Such relationships can be well modeled with a

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

dynamic Bayesian Network (DBN). The DBN is capable of accounting for uncertainty in the AU recognition process, representing probabilistic relationships among different AUs, modeling the dynamics in facial action development, and providing principled inference solutions. In addition, with the dependencies coded in the graphical model, it can handle situations where some measurements for certain AUs are missing.

In this paper, a fully automatic system is proposed for real-time recognition of action units in real-world environment. Figure 1 gives the flowchart of our automatic AU recognition system. First, the face and eyes are automatically detected in live video. The face region is aligned based on the eye positions detected and is convolved with a set of multi-scale and multi-orientation Gabor filters. After that, a confidence score is obtained by an AdaBoost classifier for each AU. These scores are then discretized to be used as evidence into the dynamic Bayesian network for AU inference. In order to correctly model the relationships among different AUs, advanced learning techniques are applied to learn both the structure and parameters of the BN based on the prior structure and parameters. The experiments show that the AU recognition accuracy rate can be greatly improved with such a dynamic BN.

2 Related Work

Generally, the previous work in detection face action units can be summarized into five groups. The first group is to extract facial motion through optical flow [21, 5]. The extraction of dense flow is slow and also has the shortcoming of sensitivity to inaccurate image alignment and motion discontinuities. The second group is the method based on difference image [6]. This kind of method is critically dependent on the accurate alignment between the reference neutral face and the target face.

The third group includes the methods based on statistical models. Lanitis et. al [12] utilize statistical learning of images by an Active Appearance model. The face deformation due to facial expression is characterized by a set of appearance parameters, which could be used to determine the expression. Kapoor et. al [11] further use the shape parameters to identify each AU by a SVM classifier. The statistical model-based techniques are limited by the linear assumption of PCA.

The fourth group includes the specific feature based methods. This group recognizes the AUs by a two-stage feature-based approach. First, the features related to corresponding AUs are extracted, either by tracking the permanent features (i.e. feature points), or detecting transient features (i.e. wrinkles and furrows), or both. Then, the AUs are recognized by different recognition engines.

Lien et. al [14] extract the features by feature point tracking, dense flow tracking with PCA and edge detection. Each AU is classified by a discriminant classifier or a Hidden Markov Model. Tian et. al [17] utilize a set of multi-state facial component models for feature point tracking and edge detector for furrow detection, but require manual intervention in the initial frame. Neural networks are used to recognize a specific AU or AU combination. Pantic and Rothkrantz [15] extract the facial features by a hybrid feature detector on a dual view (frontal and profile) face model. A rule-based method is used for AU coding. Valstar et. al [19] use a particle filter with factorized likelihoods for tracking facial feature points. The AUs are classified by a probabilistic actively learned SVM. Instead of using geometric positions of the feature points, the multi-scale and multi-orientations Gabor wavelet coefficients extracted at the feature points are employed in [24][18] as the feature representation. Furthermore, hybrid systems are proposed [3] by combining the techniques of holistic approach based on PCA, extracting features, and optical flow measurement. This group of techniques requires designing special-purpose features for each AU. Thus the accuracy of the AU coding depends on how reliably and accurately the facial features are extracted.

The last group consists of machine-learning based methods. Bartlett et. al [1] investigate general purpose learning mechanisms based on machine-learning techniques for data-driven AUs recognition. This method has the advantage of detecting the changes both in geometrical positions of features and in face appearance (i.e. wrinkle, bulges) simultaneously. The best recognition performance is obtained through SVM classification on a set of Gabor wavelet coefficients selected by AdaBoost. However, this kind of approach requires a good face alignment, which relies on reliably and accurately localizing eye positions.

In addition to the above approaches, there are three groups using Bayesian networks for facial expression classification. Cohen et. al [2] use Gaussian Tree-Augmented Naive Bayes (TAN) to learn the dependencies among different facial features in order to classify facial expressions. However, due to TAN's structure limitations, it cannot handle more complex relationships between facial features as well as temporal changes. Zhang and Ji [23] exploit BN for classifying six basic facial expressions with a dynamic fusion strategy. Gu and Ji [7] use the similar idea for facial event classification, such as fatigue. In their BNs, AUs are modeled as hidden nodes connecting facial features and facial expressions. Still, they don't consider the relationships among AUs as well as the dynamics of each AU. And AUs are not recognized explicitly in their models.

Our system differs from the cited ones in that it models probabilistic relationships among different AUs, and accounts for the temporal changes in facial action development with a dynamic Bayesian network. Advanced learning

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

techniques are used to learn both the structure and parameters of the BN from the the training database. To the best of our knowledge, we are the first to exploit the relationships among AUs so that to improve AU recognition in addition to advanced computer vision techniques.

3 AU Measurement Extraction

3.1 Face and Eyes Detection

The position, size, and orientation of the face region are estimated using the knowledge of eye centers, which are obtained through a boosted eye detector. Given the knowledge of eye centers, the face region is normalized and scaled into a 64×64 image. Eye detection is performed using a binary classifier [20] that is formed from the linear combination of geometric Haar features. The Haar feature set is comprised of over $50,000$ individual features, which is significantly larger than the actual amount of pixels in a training image. The feature selection and classifier construction is done simultaneously by using AdaBoost. The classifier is trained with over $40,000$ positive samples and hundreds of thousands of negative samples. Finally, the boosted classifier utilizes nearly 1000 Haar features, with a final positive rate of nearly 99%, and a final false rate of $10^{-7}\%$.

3.2 AU Measurements

Given the normalized images, we extract a measurement for each AU through a general purpose learning mechanism based on Gabor feature representation and AdaBoost classification. Such approaches can be generalized to recognize any AU in the presence of other AUs given a training data set. For this work, the magnitudes of a set of multi-scale and multi-orientation Gabor wavelets as in [13] are used as the feature representations, similar to the work by Bartlett et. al [1]. Gabor wavelet based feature representation has the psychophysical basis of human vision, and achieves robust performance for expression recognition and feature extraction under illumination and appearance variations. Instead of extracting the Gabor wavelet features at specific feature points, the whole normalized face region is convolved by a set of Gabor filters at 5 spatial frequencies and 8 orientations. Thus $8 \times 5 \times 64 \times 64 = 163840$ Gabor wavelet features are used as the individual features for AdaBoost.

An AdaBoost classifier is employed to obtain the measurement for each action unit. AdaBoost is not only a good feature selection method, but also a fast classifier. The feature selection and classifier construction is performed simultaneously by AdaBoost. At first, each training sample is initialized with an equal weight. Then, weak classifiers are constructed using the individual Gabor features during AdaBoost training. In each iteration, the single weak classifier with the minimum weighted error is selected to be linearly combined to form the final classifier with a weight proportional to the minimum error. The samples are then reweighted based upon the performance of the selected weak classifier, and the process is repeated. AdaBoost forces the classifier to focus on the most difficult samples in the training set, and thus it results in a very efficient classifier. To increase the speed of the actual algorithm, the final classifier is broken up into a series of cascaded AdaBoost classifiers. These individual classifiers are sequentially connected to form a final cascade classifier. A large majority of the negative samples will be removed in earlier cascades. Therefore it results in a much faster real-time classifier. The final classifier utilizes around 200 Gabor features.

However, this machine-learning based approach treats each AU independently and each frame individually, and relies on the robustness and accuracy of face region alignment. In order to model the dynamics of AUs and deal with the image uncertainty, we utilize a DBN for AU inference. Thus the output of AdaBoost classifier is discretized and used as the evidence for subsequent BN inference.

4 AU Inference with a Dynamic Bayesian Network

4.1 AU Relationship Analysis

As discussed before, due to the richness, ambiguity, and dynamic nature of facial actions, as well as the image uncertainty and individual discrepancy, current computer vision methods cannot perform feature extraction reliably, which limits AUs recognition accuracy. However, if we could consider the relationships among different AUs and model such relationships into a framework, it will compensate the disadvantages of vision techniques and thus improve the recognition accuracy rate. For subsequent discussion, Table 1 summarizes a list of commonly occurring AUs and their meanings, although the proposed system is not restricted to recognizing these AUs given the training data.

In a spontaneous facial behavior, it is rare that only a single AU appears alone. Groups of AUs usually appear together to show meaningful facial expressions. For example, the happy may involve AU6 (cheek raiser) + AU12 (lip corner puller), the surprise may involve AU1 (inner brow raiser)+AU2 (outer brow raiser)+AU5 (upper lid raiser)+AU26 (jaw drop), and the fear may involve AU1 (inner brow raiser)+AU2 (outer brow raiser)+AU4 (brow lowerer)+AU5 (upper lid raiser)+AU25 (lips part). Furthermore, some of the AUs would appear simultaneously for the most time. For example, it is difficult to do AU2 (outer brow raiser) without adding AU1 (inner brow raiser) since the activation of AU2 tends to pull the inner eyebrow as

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

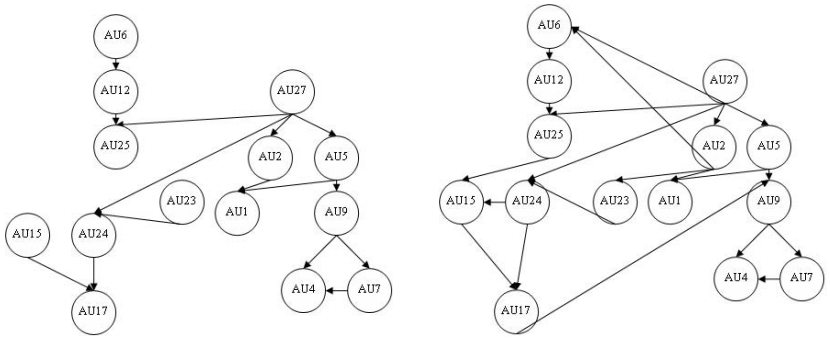


Figure 2. (a) The prior BN for AU modeling before learning; (b) The learnt BN from the training data.

AU1  Inner brow raiser	AU2  Outer brow raiser	AU4  Brow Lowerer	AU5  Upper lid raiser	AU6  Cheek raiser
AU7  Lid tighten	AU9  Nose wrinkle	AU12  Lip corner puller	AU15  Lip corner depressor	AU17  Chin raiser
AU23  Lip tighten	AU24  Lip presser	AU25  Lips part	AU27  Mouth stretch	

Table 1. A list of action units.

well. AU27 (mouth stretch) rarely appears without AU25 (lips part). On the other side, it is nearly impossible for some AUs appearing together, such AU23 (lip tighten) and AU27 (mouth stretch).

We construct a Bayesian network (BN) to model and learn such relationships. A BN is a directed acyclic graph (DAG) that represents a joint probability distribution among a set of variables. In a BN, nodes denote variables and the links among nodes denote the conditional dependencies among variables. The dependencies are characterized by a conditional probability table (CPT) for each node. A BN is manually constructed as shown in Figure 2(a) to model the relationships among the 14 AUs shown in Table 1.

4.2 Learning Model Structures

After analyzing the relationships among the AUs, we get an initial BN structure as shown in Figure 2(a). Although it is our best guess based on the analysis, it may not be correct enough to reflect the true relationships. Therefore it is necessary to use large amount of training data to "correct" our "guess" with a structure learning algorithm.

The structure learning algorithm first defines a score that describes the fitness of each possible structure B_s to the observed data. Suppose we have a domain of discrete variables $U = \{X_1, ..., X_n\}$, and a database of cases $D = \{C_1, ..., C_n\}$. Then, a score can be defined as:

$$\begin{aligned} Score(B_s) &= \log p(D, B_s) \\ &= \log p(D|B_s) + \log p(B_s) \end{aligned} \quad (1)$$

The two terms are actually the likelihood (the first term) and the prior probability (the second term) of the structure B_s. Instead of giving an equal prior to all possible structures, we assign a high probability to the prior structure in Figure 2(a). And the prior of any other network is decreased depending on the deviation between this network and the provided prior network: $p(B_s) = c\kappa^{\delta}$, where c is a normalization constant, $0 < k \leq 1$, and δ is the number of nodes in the symmetric difference between B_s and the provided prior structure [9].

To compute the likelihood $\log p(D|B_s)$, the Bayesian information criterion (BIC) [16] has been used:

$$\log p(D|B_s) \approx \log p(D|\hat{\theta}_{B_s}, B_s) - \frac{d}{2}\log(N) \quad (2)$$

where θ_{B_s} are the network parameters, $\hat{\theta}_{B_s}$ is the maximum likelihood estimate of θ_{B_s}, d is the number of logically independent parameters in B_s, and N is taken from the size of D. Thus the second term actually measures the structure complexity.

Having defined a score, the next step is to identify a network structure with the highest score with a searching algorithm. Thus the structure learning becomes an optimization problem: find the structure that maximizes the score. We apply iterated hill-climbing here. First, we apply local search until a local maximum is reached. Then, we randomly perturb the current network structure, and repeat the process for some manageable number of iterations. It helps to avoid getting stuck at a local maximum.

The learnt structure is shown in Figure 2(b). Basically, the learnt structure keeps all of the links in the initial structure, and several links have been added, such as AU27 to

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

AU6, AU2 to AU6, AU25 to AU15, and AU2 to AU23, which are reasonable to reflect facial expressions. For example, AU27 (mouth stretch) would mostly be contracted in a surprise, while AU12 (lip corner puller) would mostly happen in a smile. These two action units are seldom present together. Overall, the learnt structure reflects the data pattern in the training data set better.

4.3 A Dynamic Bayesian Network

In the structure learning, we only focus on the static BN. However, it lacks the ability to express temporal dependencies between the consecutive occurrences of certain AUs in image sequences. AU is an event that develops over time. It can therefore be better recognized from a sequence of observations over time instead of from a snapshot. We thus use a dynamic BN to model the dynamic aspect of AUs.

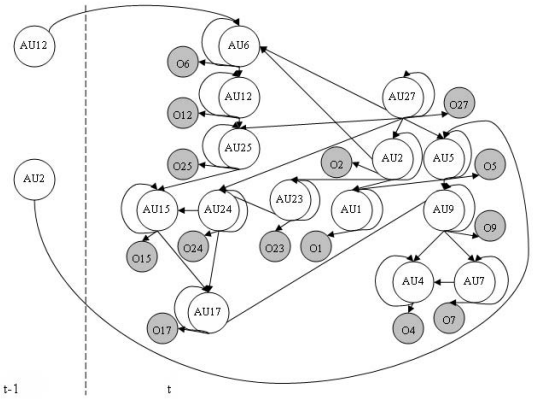


Figure 3. The Dynamic BN for AU modeling. The self-arrow at each AU node indicates the temporal relationship of a single AU from the the previous time frame to the current time frame. The arrow from AU_i at time $t-1$ to AU_j ($j \neq i$) at time t indicates the temporal relationship among different AUs. The shaded circles indicate the measurements for each AU.

In general, a dynamic BN is made up of interconnected time slices of static BNs, and the relationships between two neighboring time slices are modeled by a Hidden Markov model, i.e., random variables at time t are influenced by variables at time t, as well as by the corresponding random variables at time $t-1$ only. Each time slice is used to represent the snapshot of an evolving temporal process at a time instant. These time slices are interconnected by the arcs joining particular temporal variables from two consecutive slices. Specifically, in the proposed framework, the inter-slice arc from an AU node at time $t-1$ to that at time t depicts how an AU develops over time.

Figure 3 shows the dynamic BN with the visual measurement nodes. For presentation clarity, we use the self-arrows to indicate the temporal relationship of a single AU from the previous time frame to the current time frame. The arrow from AU_i at time $t-1$ to AU_j ($j \neq i$) at time t indicates the temporal relationship among different AUs.

Furthermore, we associate each AU node with a measurement node, which is indicated with a shaded circle in Figure 3. The links between them model the measurement uncertainty of the computer vision techniques in Section 3. From the point view of a BN, the measurement nodes are regarded as observed nodes, which provide evidence in the inference procedure, and the AU nodes are hidden nodes, whose states need to be inferred from the DBN given the evidence. We will talk about the inference procedure later.

Therefore, such a DBN is capable of accounting for uncertainty in the AU recognition process, representing probabilistic relationships among different AUs, modeling the dynamics in AU development, and providing principled inference solutions. In addition, with the dependencies coded in the graphical model, it can handle the situations where some data entries for certain AUs are missing. For example, even if some AU measurements are not available with the computer vision techniques, it is still possible to infer the states of the corresponding AUs using other available measurements.

4.4 Learning Model Parameters

A complete DBN consists of two parts, one is structure, and another is parameter (conditional probability table (CPT) associated with each node). As we presented in Section 4.2, in the structure learning algorithm, for each structure, the parameters are estimated with the Maximum Likelihood estimate (MLE). However, MLE doesn't consider the influence of prior information. Furthermore, more parameters need to be learnt considering the temporal links and the links with the measurement nodes. Thus, first, we set up the prior probabilities for each node in the DBN as shown in Figure 3 based on our data analysis, and then learn the parameters with the Bayesian estimate method. We first present how to learn parameters for a static BN, and then extend to a DBN.

Let θ_{ijk} indicate a probability parameter for a BN with the structure B_s, where i ranges over all the variables (nodes in BN), j ranges over all the possible parent instantiations for variable X_i, and k ranges over all the instantiations for X_i itself. Therefore, x_i^k represents the kth state of variable X_i, and $pa^j(X_i)$ is the jth configuration of the parent nodes of X_i. In other words, $\theta_{ijk} = p(x_i^k | pa^j(X_i), B_s)$. The goal of the learning is to maximize the posterior distribution $p(\theta|D, B_s)$(MAP) [8]. With the built-in conditional independence in a BN,

$$p(\theta|D, B_s) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\theta_{ij}|D, B_s) \quad (3)$$

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

where n is the number of variables in the BN, q_i is the number of the parent instantiations for X_i, and $p(\theta_{ij}|D, B_s)$ can be represented by Dirichlet distributions:

$$p(\theta_{ij}|D, B_s) = Dir(\alpha_{ij1} + N_{ij1}, ..., \alpha_{ijr_i} + N_{ijr_i}) \quad (4)$$

where α_{ijk} reflects the prior beliefs about how often the case $X_i = k$ and $pa(X_i) = j$, N_{ijk} reflects the number of cases in D for which $X_i = k$ and $pa(X_i) = j$, and r_i is the number of all the instantiations of X_i.

This learning process considers both prior probability and likelihood, so that the posterior probability is maximized. Since the training data is complete, it can be actually explained as a counting process, which results in the following updating formula for the probability distribution parameters:

$$\theta_{ijk} = \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}} \quad (5)$$

where $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$.

Learning the parameters in a DBN is actually similar with learning parameters in a static BN. Based on the Markov assumption, in the learning procedure, we regard the DBN as a two-slice static BN, thus the similar learning procedure applies. The only difference is that more parameters need to be learnt, and the training data need to be divided into set of time sequences.

4.5 AU Inference

In this section, we describe how to infer the probabilities of different AUs given the AU measurements obtained with the computer vision techniques in Section 3.

Dynamic inference estimates the AUs at each time step t. Let AU_i^t indicates the node of AU i at time step t, $e^t = \{e_1^t, e_2^t, ..., e_{27}^t\}$ be the measurements for the 14 AUs at time step t. Thus, the probability of AU_i^t given the available evidence $e^{1:t}$ is:

$$p(AU_i^t|e^{1:t}) = c_i p(AU_i^t|e^{1:t-1}) p(e^t|AU_i^t, e^{1:t-1}) \quad (6)$$

where c_i is the normalization constant, $p(e^t|AU_i^t, e^{1:t-1})$ is the likelihood. Based on the Markov assumption, $p(e^t|AU_i^t, e^{1:t-1}) = p(e^t|AU_i^t, e^{t-1})$. And $p(AU_i^t|e^{1:t-1})$ is the prediction, which can be obtained as:

$$p(AU_i^t|e^{1:t-1}) = \sum_{AU_i^{t-1}} [p(AU_i^t|AU_i^{t-1}, e^{1:t-1}) p(AU_i^{t-1}|e^{1:t-1})] \quad (7)$$

where $p(AU_i^{t-1}|e^{1:t-1})$ is already inferred at time step $t-1$. $p(AU_i^t|AU_i^{t-1}, e^{1:t-1})$ is the transition model for AU_i^t. We assume $p(AU_i^t|AU_i^{t-1}, e^{1:t-1}) = p(AU_i^t|AU_i^{t-1})$.

In this way, we obtain the posterior probability of each AU with the observed measurements. In the system, we adapt the Variable Elimination (VE) inference algorithm [22] into the dynamic BN for efficient inference.

5 Experimental Results

5.1 Facial Action Units Database

The dynamic facial action unit recognition system is trained on FACS labeled images from two databases. The first database is Cohn and Kanade's DFAT-504 database [10] consisting of more than 100 subjects. This database is collected under controlled illumination and background and has been widely used for evaluating facial action unit recognition system. The results on the Cohn and Kanade's database will be used to compare with other published methods. The second database consists of 42 image sequences from 10 subjects containing the target AUs, coded by a FACS trained expert. Subjects are instructed to perform the single AUs and AU combinations as well as the 6 basic expressions. The database is collected under real-world condition with unconstrained illumination and background as well as moderate head motion. The results on the second database is intended to demonstrate the robustness in real-world applications. Overall, the combining database consists of 14,000 images from 111 subjects.

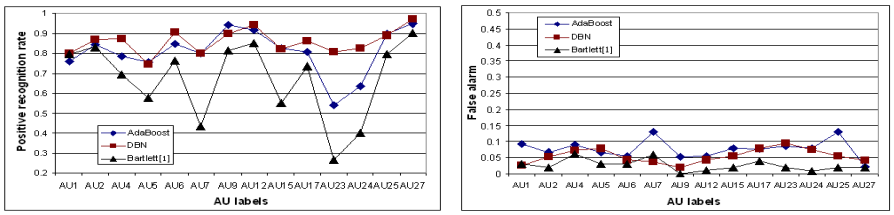
5.2 Evaluation on Cohn and Kanade Database

To evaluate the recognition accuracy of the system, it is evaluated on the Cohn and Kanade database using leave-one-subject-out cross-validation for 14 target AUs. The AdaBoost classifier and the DBN are trained on all of the data but leave one subject out for testing. For AdaBoost classifier training, the positive samples are chosen as the training images containing the target AU at different intensity levels, and the negative samples are selected as those images without the target AU regardless the presence of other AUs. The action unit labels corresponding to each frame are used in both of the static and dynamic Bayesian network learning.

Figure 4 shows the performance for generalization to novel subjects in Cohn and Kanade Database. The three curves demonstrate the performance among using AdaBoost classifier alone, using dynamic BN, and Bartlett's method [1] (reported on their website). With only AdaBoost classifiers, our system achieves an average recognition rate 91.2% with positive rate 80.7% and false alarm rate 7.7% for 14 AUs, where the average rate is defined as the percent of examples recognized correctly. With the use of the dynamic BN, the system achieves the overall average recognition rate of 93.3% with a significant improvement in positive rate to 86.3% and improved false alarm rate of 5.5%.

As shown in Figure 4, for AUs that can be well recognized by the AdaBoost classifier, the improvement by using DBN is not that significant. However, for the AUs that are difficult to be recognized by the AdaBoost classifier, the improvements are impressive, which exactly demonstrates the benefit of using the DBN. For example, recognizing AU23

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE



(a) (b)

Figure 4. Comparison of AU recognition results on novel subjects in Cohn-Kanade database using the AdaBoost classifier, dynamic BN, and the results of Bartlett [1] respectively: (a) average positive rates; (b) average false. alarm

(lip tighten) and AU24 (lip presser) is difficult, since the two actions occur rarely and the appearance changes caused by these actions are relatively subtle. Fortunately, the co-presence probability of these two actions is very high, since they are contracted by the same set of facial muscles. By employing such relationship in the DBN, the positive recognition rate of AU23 increases from 54% to 80.6% and that of AU24 increases from 63.4% to 82.3%. Similarly, by employing the co-absence relationship between AU25 (lips part) and AU15 (lip corner depressor), the false alarm rate of AU25 reduces from 13.3% to 5.7%.

In summary, the average performance of our system is equal to or better than the previously reported systems. Compared with the system performance reported by Bartlett et. al (overall average recognition rate 93.6%), our system achieves a similar average recognition rate (93.33%). However, with a marginal drop of false alarm rate 2.95%, our system significantly increases the positive rate from 70.1% (in Bartlett's method) to 86.3%. Tian et. al [17] achieves a similar performance however manual intervention is required in the initial frame with neutral expression. Valstar et. al [19] report a 84% average recognition rate on the Cohn-Kanade database while training on another database. Kapoor et. al [11] obtain an 81.2% of recognition rate on 5 upper AUs in Cohn-Kanade database with hand marked pupil positions.

5.3 Experiment Results under Real-world Condition

In the second set of experiments, the system is trained on all of the training images from the Cohn-Kanade database and 38 image sequences with 6 subjects from the second database. The remaining four image sequences with 4 different subjects from the second database are used for testing. This experiment intends to demonstrate the system robustness for real-world environment. The system performance is reported in Figure 5. The average recognition rate of the DBN is 93.27% with an average positive rate of 80.8% and a false alarm rate of 4.47%.

Figure 6 shows an image sequence where the system output (the probabilities) of multiple AUs changes over time. In the beginning, the subject gradually raises the eyebrows and upper eyelids, and then gradually lowers the eyebrows and relaxes eyelids, but raises the upper lids finally. Correspondingly, the values of the system output of AU1 (inner eyebrow raise) and AU2 (outer eyebrow raise) first increase, and then decrease, while the value of AU5 (upper lid raise) changes at the pattern of increase-decrease first, and finally increases again as shown in the right chart of Figure 6. Since there are individual discrepancies with respect to the magnitude of action units, it is difficult to determine the absolute intensity of a given subject. Our dynamic modeling of facial action units can more realistically reflect the evolution of a spontaneous facial emotion, and thus can extract the relative intensity changes of the action units.

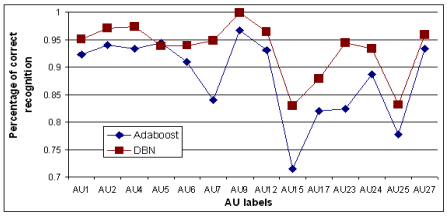


Figure 5. Comparison of average AU recognition rates on novel subjects under real-world circumstance using the AdaBoost classifier and dynamic BN respectively.

6 Conclusions and Future Work

In this paper, we propose a novel approach for AU classification that systematically accounts for relationships among AUs and their temporal evolution. Specifically, we use a dynamic Bayesian network to model the relationships among different AUs. The DBN provides a coherent and unified hierarchical probabilistic framework to represent probabilistic relationships among different AUs, and accounts for the temporal changes in facial action development. Under our system, robust computer vision techniques are used to obtain AU measurements. And such AU mea-

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

Figure 6. Automatic AU recognition: the left images show an image sequence in which three AUs (AU1,AU2, and AU5) change over time; the right image shows the probabilities of involved AUs in the corresponding frames for the left sequence.

surements are then applied as evidence into the DBN for inferring various AUs. The experiments show the integration of AU relationships and AU dynamics with AU image measurements yields significant improvements in AU recognition. Our system will be extended to recognize more AUs as well as facial event.

References

[1] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. *Proc. of CVPR05*, 2:568–573, 2005.

[2] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91:160–187, 2003.

[3] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Trans. on PAMI*, 21(10):974–989, 1999.

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

[5] I. Essa and A. Pentland. Coding, analysis, interpretation and recognition of facial expression. *IEEE Trans. on PAMI*, 19(7):757–763, 1997.

[6] B. Fasel and J. Luettin. Recognition of asymmetric facial action unit activities and intensities. *Proc. of ICPR00*, 1:1100–1103, 2000.

[7] H. Gu and Q. Ji. Facial event classification with task oriented dynamic bayesian networkg. *Proc. of CVPR04*, 2:870–875, 2004.

[8] D. Heckerman. A tutorial on learning with bayesian networks. *Technical Report MSR-TR-95-06, Microsoft Research*, pages 1–40, 1995.

[9] D. Heckerman, A. Mamdani, and M. Wellman. Real-world applications of bayesian networks. *Comm. of the ACM*, 38(3):24–68, March 1995.

[10] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. *Proc. of FG00*, pages 46–53, 2000.

[11] A. Kapoor, Y. Qi, and R. W. Picard. Fully automatic upper facial action recognition. *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003.

[12] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Trans. on PAMI*, 19(7):743–756, 1997.

[13] T. Lee. Image representation using 2d gabor wavelets. *IEEE Trans. on PAMI*, 18(10):959–971, October 1996.

[14] J. J. Lien, T. Kanade, J. F. Cohn, and C. Li. Detection, tracking, and classification of action units in facial expression. *Journal of Robotics and Autonomous System*, 31:131–146, 2000.

[15] M. Pantic and L. J. M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Trans. on SMC-Part B: Cybernetics*, 34(3):1449–1461, June 2004.

[16] Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[17] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. on PAMI*, 23(2):97–115, February 2001.

[18] Y. Tian, T. Kanade, and J. F. Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. *Proc. of FG02*, pages 229–234, May 2002.

[19] M. F. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. *Proc. of CVPRW'05 on Vision for Human-Computer Interaction*, June 2005.

[20] P. Viola and M. Jones. Robust real-time face detection. *International J. of Computer Vision*, 57(2):137–154, May 2004.

[21] Y. Yacoob and L. S. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Trans. on PAMI*, 18(6):636–642, June 1996.

[22] N. L. Zhang and D. Poole. A simple approach to bayesian network computations. In *Proc. of the Tenth Canadian Conf. on Artificial Intelligence*, Banff, Alberta, Canada, May 16-22 1994.

[23] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. on PAMI*, 27(5):699–714, May 2005.

[24] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. *Proc. of FG98*, pages 454–459, 1998.

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE