Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Vision and Image Understanding
Computer Vision and Image Understanding 100 (2005) 385–415
www.elsevier.com/locate/cviu

# Task oriented facial behavior recognition with selective sensing

Haisong Gu [a], Yongmian Zhang [a], Qiang Ji [b,*]

[a] *Department of Computer Science, University of Nevada, Reno NV 89557, USA*
[b] *Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, NY 12180-3590, USA*

Received 22 July 2003; accepted 25 May 2005
Available online 2 August 2005

**Abstract**

Facial behaviors represent activities of face or facial feature in spatial or temporal space, such as facial expressions, face pose, gaze, and furrow happenings. An automated system for facial behavior recognition is always desirable. However, it is a challenging task due to the richness and ambiguity in daily facial behaviors. This paper presents an efficient approach to real-world facial behavior recognition. With dynamic Bayesian network (DBN) technology and a general-purpose facial behavior description language (e.g., FACS), a task oriented framework is constructed to systematically represent facial behaviors of interest and the associated visual observations. Based on the task oriented DBN, we can integrate analysis results from previous times and prior knowledge of the application domain both spatially and temporally. With the top–down inference, the system can make dynamic and active selection among multiple visual channels. With the bottom–up inference from observed evidences, the current facial behavior can be classified with a desired confidence via belief propagation. We demonstrate the proposed task-oriented framework for monitoring driver vigilance. Experimental results demonstrate the validity and efficiency of our approach.
© 2005 Elsevier Inc. All rights reserved.

*Keywords:* Facial behavior recognition; Task oriented approach; Facial action coding system; Dynamic Bayesian networks; Selective sensing

* Corresponding author.
*E-mail address:* qji@ecse.rpi.edu (Q. Ji).

1077-3142/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2005.05.004

## 1. Introduction

A face reader often means a professional person who has the gifted ability to pick up voluntary or involuntary facial expression or feature change occurring within a very short time [14]. Since a camera can honestly record instantaneous appearance changes on faces, an automated face reading system is highly desirable from the applications such as, law enforcement, intelligent human computer interaction, etc. It also becomes possible with recent improvements in computing power, facial feature detection, and facial expression analysis technology.

The general-purpose facial expression analysis has been explored for decades. A great deal of techniques have been proposed. A recent survey of existing works can be found in [32]. In terms of accuracy and robustness, it is still hard for real-world applications. The difficulty comes from two folds. One is the richness and complexity of facial behaviors. The number of expressions occurring in daily life is more than 2000 [12]. It is hard to accurately classify them, even by an experienced human facial expression coder. The other is the variety of expression display forms. Facial behaviors can appear in the form of geometrical deformation, eye pupil movements, or furrow happenings. A single visual feature (or channel) will not be able to efficiently capture all of these changes. Fortunately, each specific application (or task) only involves several limited number of facial behaviors of interest. A task-oriented approach can be used to remove the ambiguity due to the rich expressions and variety in facial expression. Furthermore, the multiple visual channels provide the possibility to detect different forms of facial behaviors in the most efficient way. By actively selecting the most informative channel, the effective facial features can be captured so as to make the recognition highly efficient. These two techniques constitute the main contributions of our facial behavior recognition system.

The techniques for classification of facial expression displays from image sequences can be methodologically divided as neural networks [38], rule-based models [43,33], template based with minimum distance measure [13,24], discriminant analysis of feature vector [8,1], and hidden Markov model (HMM) [27]. However, except for HMM, these methods lack the sufficient expressive power to capture the spatio-temporal dependencies and uncertainties exhibited by facial expressions. Though HMM is appropriate for modeling uncertainties and real world time series, we argue that in the existing HMM-based facial expression methods, (1) each single expression display involved in the complex facial behavior change usually requires a specific topology of HMM; (2) it has to be assumed that an image sequence is welled segmented, starting from a neutral state and ending with an apex state. For a real-world facial expression, however, it is not true that the development of expression strictly follows the state of neutral, apex, and then neutral. Hence, there is a need for a method which can model a complex facial behavior while allowing fast classification.

The integration of different feature detectors in a well-grounded probabilistic theoretical framework (such as dynamic Bayesian networks) is one of hot topics and so far has been applied to various applications including human gesture understanding [4,5] and human computer interaction [34]. For facial behavior recognition, we

propose a dynamic Bayesian network (DBN) based approach, which consists of the Bayesian network and the first-order HMM. With the Bayesian network, we can fully describe the spatial relationship in facial behaviors. With a first-order HMM, the temporal dependence in the facial behavior can be captured. Compared with the pure HMM-based facial expression methods, our approach not only has a powerful ability to spatio-temporally represent complicated facial behaviors, but also simplify the computation by avoiding the usage of high-order HMM.

Facial behavior here means any activity of face or facial feature in spatial or temporal space, such as facial expressions, face pose, head motion, gaze, and furrow happening, etc. For each vision-based application, e.g., (1) monitoring vigilance levels of driver or pilot; (2) human emotion classification/understanding; or (3) nonverbal command recognition of operators in a specific environment, the task is often to identify one or several facial behaviors of interest in each application domain. We consider the facial behavior recognition as an inference process from spatio-temporal evidences incorporated with prior knowledge of application domain. In this way, the recognition can be viewed as a problem to detect a hypothesis from $N$ predefined behavior display categories (hypotheses). The robust recognition is achieved by fusion of information not only from different facial features, such as mouth, eyes, furrows, etc., at current time frame, but also from previous frames. It is crucial for a real time recognition system to make timely and efficient classification under various constraints. Typically, the facial behavior displays involve simultaneous changes in facial features in multiple regions of the face. Also the visual information can abundantly be collected from different feature regions on a face image. To save computation, it is important to select facial visual cues most likely to provide the most informative visual information for the current task. We propose a selective sensing framework, which had not been previously addressed in this field. The proposed framework is based on DBNs coupling with multiple visual channels, whereas the temporal information of facial behavior is represented by recursively keeping in memory several frames with a two temporal-granularity (or scale) mechanism. The visual cues at previous frames provides diagnostic support for inferring the hypothesis at the current frame through temporal links.

Fig. 1 shows our DBN-based block diagram. For each application, a task-oriented DBN is created with the help of the face database of application domain and a general-purpose facial behavior description language: facial action coding system (FACS) [12]. Based on the principles of FACS and the application domain knowledge, each facial behavior of interest can be disassembled up to single AUs via several layers. Each single AU can be associated with certain facial features, which are then measured by some visual sensors. With the Bayesian network, we not only infer what is the current facial behavior, but also determine which next single AU(s) is the most effective to use to detect current facial behavior in a timely and efficient manner. In the inference, there are two phases involved: detection and verification. In the detection phase, the system sets the target of facial behavior, activate the associated visual sensors, and detect a new facial behavior. In the verification phase, the detected facial behavior is verified with the current evidences and previous results. When

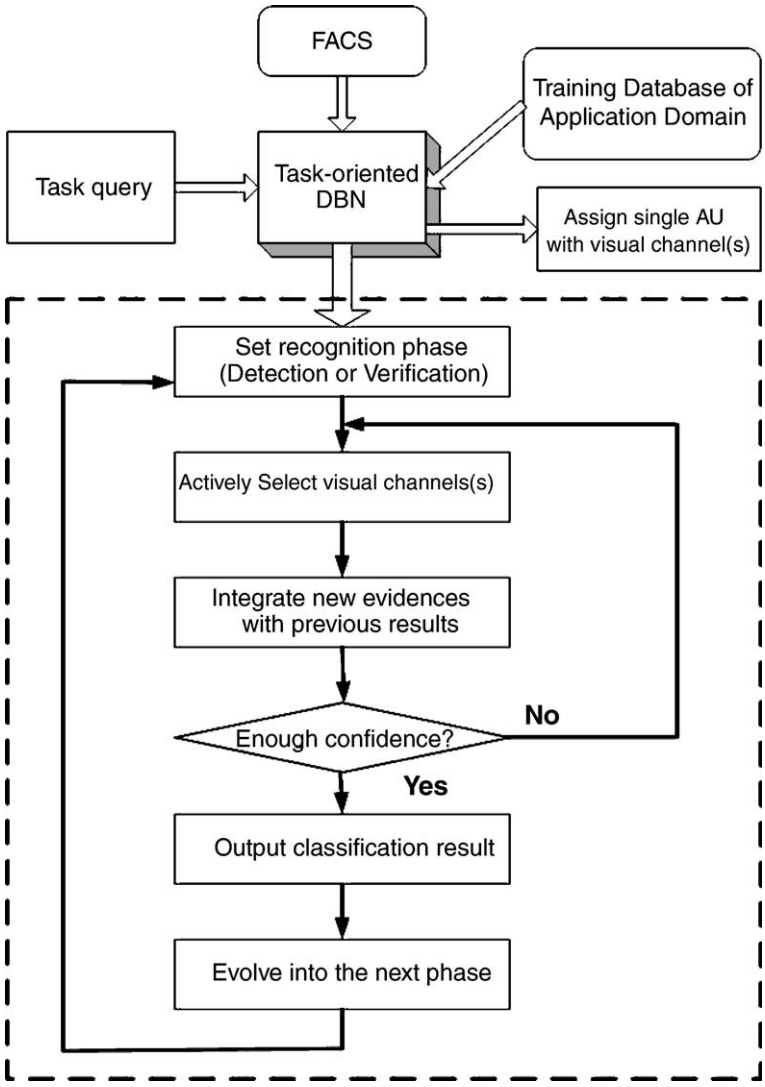



Fig. 1. The proposed DBN based facial behavior recognition with selective sensing.

the recognition confidence in each phase is high enough, the system automatically evolves into the next phase and the process repeats.

The remainder of this paper is organized as follows. In the next section, we review the previous works most relevant to our approach. Then, we present an overview of different sensing methods developed in our group for facial behavior recognition. Section 4 describes the task-oriented recognition approach. Experimental results and analysis are presented in Section 5. The final section provides conclusions, and points out future research.

## 2. Related previous works

Since early 1990's, substantial research efforts have been conducted on facial expression analysis and recognition. Pantic and Rothkrantz [32] provided a rigorous survey of existing works at that time. Technically, the early works in facial expression recognition can be broadly classified into template-based holistic spatial analysis [22,9,35], local feature-based analytic spatial analysis [26,23,40] and spatio-temporal analysis with optical flow computation [28,42,36]. However, those works suffer from several shortcomings including (1) lack of capability to recognize temporally facial expressions in image sequences; (2) using a single visual cue and therefore having difficulty handling real-world conditions such as the variation of lighting, partially occluded faces and head movements. These problems pose a number of technical challenges for later development of facial expression recognition systems.

In light of recent advances in video processing and machine vision, various methods have been proposed for recognizing facial expression either in a static image or in a video sequence. Kobayashi et al. [25] reported a real-time recognition system for emotional classification and synthesis of six prototypic emotional expressions. They use 13 vertical lines acquired by a CCD camera to cover 30 facial feature points. The brightness distribution data then feeds to a three-layer back-propagation neural network (NN) for training and classification of prototypic facial expressions. However, this technique cannot model changes of facial appearance in horizontal directions. Zhang et al. [44] presented a $680 \times 7 \times 7$ NN for six prototypic expression classification. The inputs of NN are the geometrical position of 34 facial points and 18 Gabor wavelet coefficients for each feature point; while the output is the probabilities of the examined face expression, given six prototypic emotion categories. Their technique, however, works only in the conditions of constant lighting and frontal-view faces. A common shortcoming of above approaches is that the recognition is restricted to a static face image. Static image based recognition is difficult to yield a satisfactory performance because real-world facial expression is a spatio-temporal behavior. A static facial cues are not able to describe facial expression changes overtime.

Recently, there have been several attempts to track and recognize facial expressions overtime. Yacoob et al. [43] provided an approach to the analysis and representation of facial dynamics for recognition of facial expressions from image sequences. The spatial information is based on the gradient magnitude of the intensity image and the temporal information is based on the optical flow field. Recognition of expression is based on the rules of basic actions of feature components and the rules of motion cue, which are utilized from the expression descriptions in [11,2]. The continuation work presented by Black et al. [3] uses local parameterized flow models, including an affine, a planar and a curvature model, for facial expression analysis. While the affine model and planar model both are used to represent rigid facial motions, a curvature model is augmented into affine model and to represent non-rigid motions of facial features primarily on the eyebrows and mouth. A robust regression scheme is used to estimate the parameters under the assumption of the constant brightness. To cope with large motions, a coarse-to-fine gradient-descent strategy is used.

Essa et al. [13] use a control theoretic method to extract the spatio-temporal motion energy to represent a facial motion model. The facial motion model can represent not only time dependent states, but also state evolution relationships as well. By learning "ideal" 2D motion views for each expression category, they generated the spatio-temporal templates for six different expressions. Each template is delimited by averaging the patterns of motion generated by two subjects showing a certain expression. The invariance between the motion energy template and the observed image motion energy is used as a similarity criterion to one of six expression categories. A major shortcoming of this approach is that a prior information about facial structure for each individual face is required. Since the whole face region is utilized as a dynamic face model, the proposed approach appears to be sensitive to out-of-plane head rotation and the face occlusion. Additionally, the extraction of dense flow over a whole face is relatively computational expensive.

Many early approaches to expression recognition tried to directly detect mental states (or emotion), such as happiness and sadness, etc. But, basically the mental states are not measurable. They can only be inferred from the extracted data. On the other hand, information extraction techniques from facial expression sequences, such as feature point tracking and pupil detection etc., provide quantitative measurement of facial appearances. To make a bridge between mental states and results of the information extraction, a quantitative description method for facial expression is necessary. Ekman's group [12], based on their investigation into muscular anatomy of face and facial expressions, proposed facial action coding system (FACS), an unified description language of facial expression. It has a total of 71 primitive units, called action unit (AU). Based on them, almost any expression display can be represented by AUs or a AU combination.

Many research has since been conducted for facial expression analysis based on FACS. Zhao et al. [45] proposed a $10 \times 10 \times 3$ back-propagation NN for classification of six basic facial expression. The set of training image for six basic facial expressions are selected from the emotion database of FACS [11]. As the input of NN, displacements of 10 feature points between expression image and neutral image is manually measured (perfect facial data extraction is assumed), and then the data is normalized (fuzzified) into one of eight intervals. However, the performance for recognizing the facial expression from a novel person is not reported. Pantic et al. [33] proposed a rule-based expert system for automatic recognition of multiple emotional expressions. Automatic facial data extraction is achieved by using dual face view model, one is from a frontal-view and one is from a side view. Total 30 feature points are considered from the combined face model. The redundant information generated by dual-view is used to reduce the unambiguous face geometry. The extracted facial data is converted to a set of rule descriptors for input; while the output is one of total 31 AUs-classes. Classification is performed by comparing the AU-coded description of input facial expression to AU-coded description of six basic emotional expressions acquired from the linguistic descriptions in FACS [11]. The system can classify the lower face AUs, the upper face AUs and the AU combinations. However, the assumption that each subexpression of a basic emotional expression has the same

influence on scoring that emotion category is not always realistically valid. Although their works limited to the static expression in still images, they declared that the precision of information extraction and the temporal correlation in the sequence are vital to achieve higher recognition accuracy.

In the most recent years, FACS has been employed to classify fine-grained changes in facial expression into AUs, rather than only to focus on six prototypic expressions as in most of previous works. It is a common opinion that any facial expressions may potentially be the combination of the fundamental AUs defined in FACS. Tian and Kanade [38] presented an automatic face analysis system to classify facial expressions based on both permanent facial features (brows, eyes, and mouth) and transient facial feature (deepening of facial furrows) in a nearly frontal-view face image sequence. The expression data extraction is conducted by a new approach namely multi-state face and facial component models to increase the robustness and accuracy of feature extraction. Additionally the transient features (e.g., furrow), which is often ignored in previous works, are also considered as crucial information for the recognition of some AUs. Classification is performed by using two general NNs, one for upper face AUs and one for lower face AUs. AU combination is treated as single AU. Inputs of NN are the parametric descriptions of extracted facial features. However, the proposed approach deals with a static face action, which means that only the end state of the facial movement is measured in comparison to the neutral position. It lacks the ability to represent expression state evolution relationships, and thus it is restricted to static facial expression images.

The above works [33,38] also suffer from a well recognized limitation that is the lack of temporal and detailed spatial information. Methodologically, the expression classifier used in current studies is either neural networks or rule-based approach. However, these methods lack the sufficient expressive power to capture the uncertainties, dependencies, and dynamics exhibited by facial emotional data. Lien and Kanade [27] explored hidden Markov model (HMM) for face AU classification. The input to HMM is the facial feature vector which includes the dense-flow of entire face, pre-selected facial features and transient facial line and furrows. They are respectively extracted by dense-flow extraction, facial feature tracking (feature points in the first frame of image sequence are manually marked) and high-gradient component detection. The feature vector is then quantized into a code book (a finite set of symbols), which represent various AUs modeled by HMMs. Each class of AU or AU combination has a corresponding HMM model with various orders and states. HMM topology, such as, the number of states and orders of HMM for each AU depends on the behavior of the facial movement among states. A directed link between states represents the possible inherent transition from one facial state to another, and a probability distribution is assigned to define the likelihood between them. The output is an AU with the highest probability among HMM models given a feature vector (observations), and therefore, all HMM models have to be evaluated for a given feature vector. Any new AU combination requires a new design of HMM topology for it. The proposed approach is infeasible for recognition of spontaneous facial expressions without keeping of the order of the beginning, apex and ending

duration, or for image sequences containing multiple emotional expressions. To combine the temporal, Cohen et al. [7] proposed an emotion-specific multi-level HMMs to dynamic facial expression recognition without pre-segmentation requirement. For six basic emotions, six individual HMMs are generated. Considering the hierarchical structure of natural human behavior, Oliver et al. [30] proposed to use two layered HMM to decompose the substantial parameter spaces and segment the problem into distinct layers that operate at different temporal scales so as to make the recognition more efficient.

Hoey and Little [16,17] propose to use partially observable Markov decision process (POMDP) (a variant of Bayesian networks) for modeling and recognizing facial expressions and hand gestures. Similar to the task-oriented approach we are taking, they pursue the generation of purposeful POMDPs model for recognizing certain categories of facial and gesture displays. Their work, however, focuses on learning (both the parameters and the structures) of POMDP using the so-called valued-directed method. Similar in concept to the active sensing paradigm proposed in this paper, the value-directed learning allows focusing resources only on the necessary and most important categories of human displays. Currently, we combine the principle of FACS and domain knowledge to systematically construct the structure of DNB and then use supervised training to learn its parameters. Cohen et al [6] and Sebe et al. [37] proposed to use Bayesian networks for facial expression recognition. Their work, however, focuses on learning the BN structures and parameters for facial expression classification.

The work by Kaliouby et al. [21] represents the latest research efforts attempting to understand practical mental states, instead of basic emotion displays. They mentioned the importance of multi-modal sensing, temporal constraints, and multi-level temporal abstractions to achieve a complicated expressions understanding.

Most of the previous methods focus on the single AU classification. In our approach, the AU identification is not the goal. In fact, AUs are not explicitly recognized and they are only measured through facial feature displacements. AUs are used to link feature point displacement with facial expressions based on FACS. We use FACS description to systematically create task-oriented Bayesian Network so as to efficiently realize the recognition of facial behaviors. So far almost all facial expression approaches are general-purpose and are used to recognize the basic emotion displays. Although they have the flexibility of applications, the ambiguity stemming from the generalization is very high. To improve the efficiency and accuracy of recognition, the utilization of knowledge of application domain is highly desirable. Furthermore, the facial expressions of interest are more frequent and important in a specific application than the basic emotion expressions. Specifically, we propose a novel methodology to realize these aspects by a systemic DBN generation. Also our DBN embeds an active sensing mechanism so as to automatically select the most informative sensor for the current recognition task. Our approach, based on the general description method and general inference framework, focuses on the facial behaviors important to each specific application. In this way, we expect to significantly reduce the modeling complexity and improve the recognition efficiency.

## 3. Multiple visual channels

The facial behavior is a rich resource to infer the mental state of human beings. Different mental states are often displayed in different forms of facial features. So a sensing system with multiple visual cues is required to efficiently extract the information from faces. The following subsections give a brief description on several sensing methods developed in our group [19,20,46,15].

### *3.1. IR based eye detector*

The eye pupil is the primary feature on a face. It can not only provide the indication of eye movement, but also be used to locate the face and gaze. However, it is hard to robustly detect pupils in a variety of lighting conditions and head orientations. To build a practical sensing system, we developed the Infrared based eye detector [19,46] similar to the one developed by IBM [29]. This detection system consists of an IR sensitive camera and two concentric rings of IR LEDs as shown in Fig. 2A. LEDs are arranged symmetrically around the camera optical axis so that they not only can reduce shadows generated by LEDs, but also create a strong lighting to minimize other IR radiations. This ensures the bright pupil effect under different lighting conditions. To further minimize interference from light sources beyond IR light and to maintain uniform illumination under different climatic conditions, a narrow bandpass NIR filter is attached to the front of the lens to cut lights out of the NIR range. An effective circuitry was also developed to synchronize the inner ring of LEDs and outer ring of LEDs with the even and odd fields of the interlaced image respectively so that they can be turned on and off alternately. The interlaced input image is subsequently de-interlaced via a video decoder, yielding the even and odd field images as shown in Figs. 2B and C. Basically the two fields have similar images. Significant difference happens only in the pupil areas. One image is related to the bright pupils, the other the dark pupils. The pupil detection is conducted by (1) illumination interference removal via image subtraction of odd and even images; and (2) searching the entire subtracted image to locate two bright regions with the constraints on blob size, shape, and inter-distance. We have developed a robust method for eye tracking under variable lighting conditions and facial expressions [46]. The IR sensing system can provide us with reliable information to indicate where the pu-

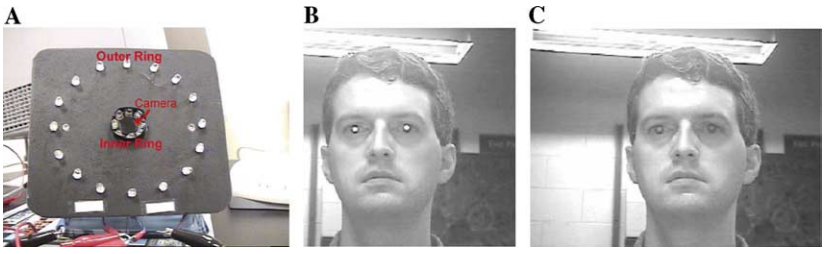



Fig. 2. (A) Hardware setup: the camera with an active IR illuminator (B) bright pupils with even field image (C) dark pupils with odd field image.

pils and the face are. Based on the pupil position, other facial features can be extracted more effectively.

### 3.2. Facial feature tracking

Most of the facial expressions are represented in the form of displacements of facial feature points. It is very important to robustly detect and track feature points on the face. With marked points in Fig. 3, we identify 22 fiducial feature points and their displacements to characterize facial behaviors. For real-world application, facial feature tracking is a very challenging issue due to the illumination changes, rapid head movements, facial expression changes and occlusion. To tackle these problems, we developed an active tracking approach [15] similar to the one developed by Wiskott et al. [41]. First, each feature point is identified by a multi-scale and multi-orientation Gabor wavelet kernel as follows:

$$\Psi(\mathbf{k},\vec{x}) = \frac{\mathbf{k}^2}{\sigma^2} e^{-\frac{\mathbf{k}^2\vec{x}^2}{2\sigma^2}} (e^{i\mathbf{k}\cdot\vec{x}} - e^{-\frac{\sigma^2}{2}}),$$

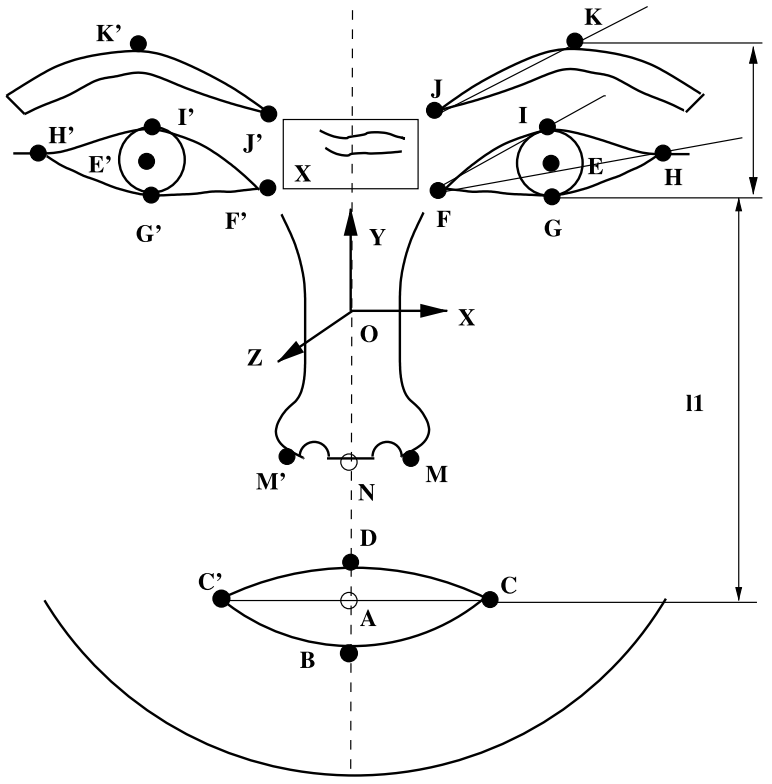



Fig. 3. The geometric relations of facial feature points and furrows.

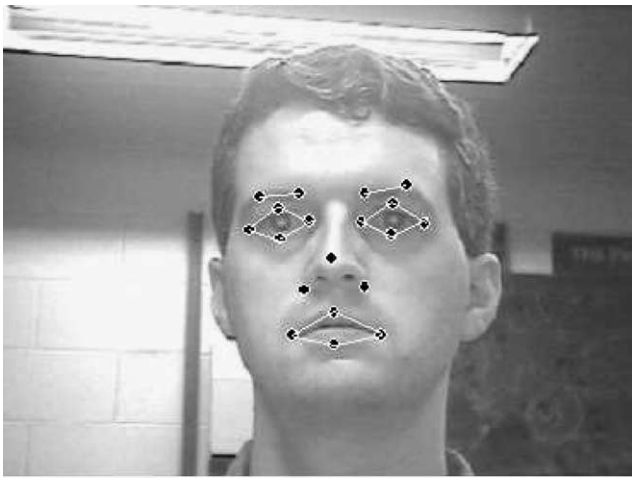


Fig. 4. Examples of feature point tracking.

where $\sigma$ is set to be $\pi$ for $128 \times 128$ images. The set of Gabor kernels consists of 3 spatial frequencies (with wavenumber: $\pi/2$, $\pi/4$, $\pi/8$), and 6 distinct orientations from 0° to 150° in the interval of 30°. For each pixel ($\vec{x}$), a set ($\Omega(\vec{x})$) of 18 Gabor coefficients in the complex form can be obtained by convolution with the Gabor kernels.

IR-based pupil detection provides the constraints on the face location and global head motion. By simultaneous use of pupil constraint and the Kalman filtering, a robust tracking for all facial features can be obtained. For the occlusion issue, a graph-based reliability propagation method is used to detect the feature and infer the correct position of the occluded feature. As a result, facial feature points and their spatial relationship are extracted in terms of local graphs, shown as in Fig. 4. Details on this work may be found in [15].

### 3.3. Furrow detection

The activation of facial muscles produces transient wrinkles and furrows perpendicular to muscular motion directions. The transient features can provide additional visual cues to support the recognition of facial expressions. Generally speaking, it is difficult to precisely set each furrow region and detect the intensity of furrow happening [39].

In our work, each facial point is identified with its physical meaning by the results of facial feature tracking. Based on the location of feature points, we can easily set up the region for each existing furrow. The presence of furrow and wrinkle within the targeted region can be determined by edge detection. The increase of edge points in a furrow region indicates the presence of furrows. Fig. 5 shows the example of furrow detection. From the results of facial feature tracking, the feature points of $J, J', F$, and $F'$ are extracted (see Fig. 3). The nose root region can be easily located as in the rectangle. Within the region, edge extraction and counting are conducted. The amount of edges are compared with one in the neutral state obtained in the ini-

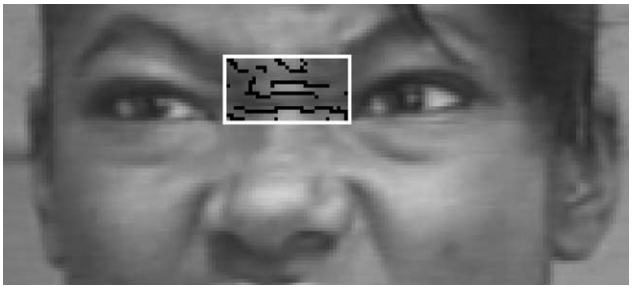


Fig. 5. Example of nasal root furrow detection.

tialization. When the current edge amount is significant more, the furrow of the nose root can be detected.

### 3.4. Head orientation estimation

Head orientation contains information about one's attention. Head pose determination is concerned with computation of the three-dimensional (3D) face orientation to detect facial behaviors such as head tilts. In a specific application, such as driver fatigue monitoring, the nominal face orientation is frontal. If the driver's head orientation in other directions for an extended period of time or occurs frequently, this indicates a certain level of fatigue. Considering the real-time application, we developed a PCA-based approach which can fast and roughly determine the head orientation from a monocular view (see [19] for details).

Our study shows that there exists a direct correlation between 3D head orientation and properties of pupils such as pupils size, inter-pupil distance, and pupils shape. Fig. 6 shows pupil measurements under different head orientations. Based on the relationship, we build a so-called pupil feature space (PFS) which is constructed by seven pupil features: inter-pupil distance, sizes of left and right pupils, intensities of left and right pupils, and ellipse ratios of left and right pupils.

Fig. 7 shows sample data projections in PFS, from which we can see clearly that there are five distinctive clusters corresponding to five face orientations (5 yaw angles). The principal component analysis (PCA) method is used to represent pose distribution. PCA analysis is primarily used for feature de-correlation and dimension reduction. To construct the PCA space, training data over all poses are used. Since exact pose estimation is not needed and that there are only a few classes of pose angles for tilt and yaw respectively, poses for the training data are labeled approximately through visual estimation. Given an input face, its pose is determined by pose classification in the PCA space.

### 3.5. Head motion estimation

By facial feature tracking, each local graph (shown in Fig. 4) represents the spatial relationship of the local facial region. Inter-frame change of local graph depicts the

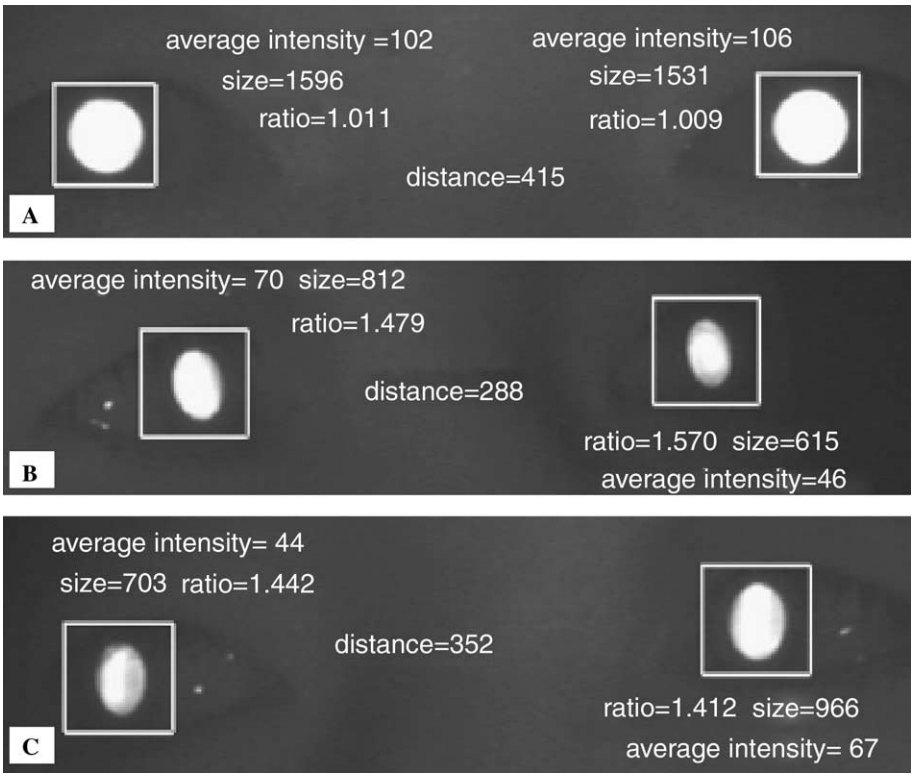



Fig. 6. Pupil images for different head orientations. (A) Frontal head position, (B) turn head left, (C) turn head right.

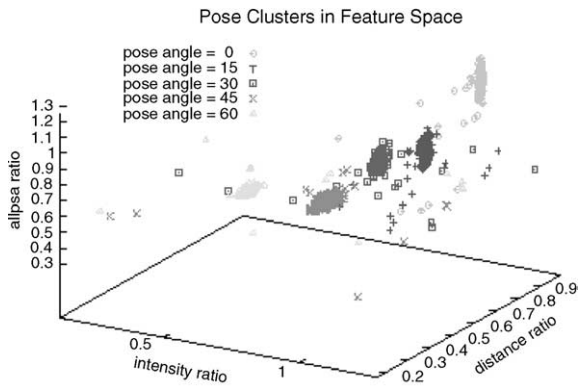



Fig. 7. Head pose clusters in pupil feature space.

displacement of each feature within the independent local region. Based on the following 2D local similarity transformation, we can determine the inter-frame motion parameters for each local region.

$$\mathbf{x}^T = \begin{pmatrix} r_{11} & r_{12} & t_1 \\ r_{21} & r_{22} & t_2 \end{pmatrix} \mathbf{x}$$

where $X$ and $X'$ are the positions of feature point before and after the transformation, respectively. $r_{ii}$s are the rotation parameters, and $t_i$s are the translation ones. By averaging the motion parameters over all the local regions, we can get in-plane global face motion. Based on obtained global similarity parameters, we get the transformed features. Then, by using the tip of nose as the center we further map them to the feature points in the current frame by a scaling transformation. The obtained zooming scale parameter can be used to determine the head motion in depth.

## 4. Facial behavior recognition

The facial behavior is the facial change occurring in spatial and temporal spaces. It is an extensive concept and covers facial expressions, face pose, head motion, gaze and furrow happening, etc. It ranges from simple facial events such as the mouth opening which can be uniquely determined by a portion of face in a snapshot, to moderate events such as smiling and talking which requires the simple spatio-temporal combination of facial displays to completely characterize, and to complex events such as the driver fatigue which needs the integration of multiple facial cues overtime.

Though there are a great deal of facial behaviors that occur in our daily life, it is believed that any facial behavior can be linguistically described by either a culture independent AU or AU combination based on FACS [12]. Here, we rely on FACS and the domain knowledge as the basis for a task-oriented modeling of facial behaviors.

### *4.1. Spatial dependence*

A specific application always focuses on some limited number of behaviors of interest. For instance, in the application of monitoring driver vigilance, the drowsiness related facial behaviors are classified so as to identify the in-vehicle driver states. The behaviors related to "Inattention," "Yawning," and "Falling asleep" are targeted. The single AUs associated with these fatigue behaviors are shown in Fig. 8. Their corresponding descriptions are presented in Table 1.

Usually, the facial behavior is the entire facial display. However, the single AU usually focuses on local feature movement. In this paper, we propose a whole-to-part model structure based on BN to relate the single AUs with the facial behavior according to the principle of FACS.

Fig. 9 shows a BN model that represents the spatial relationship of entities related to each vigilance level with local facial features and the associated visual channels. It mainly consists of 3 layers: Entire facial display, Partial facial display and Single AU. The entire display layer includes all of the behaviors (or vigilance levels) of interest. In the partial display layer, we divide an entire facial display into partial display of

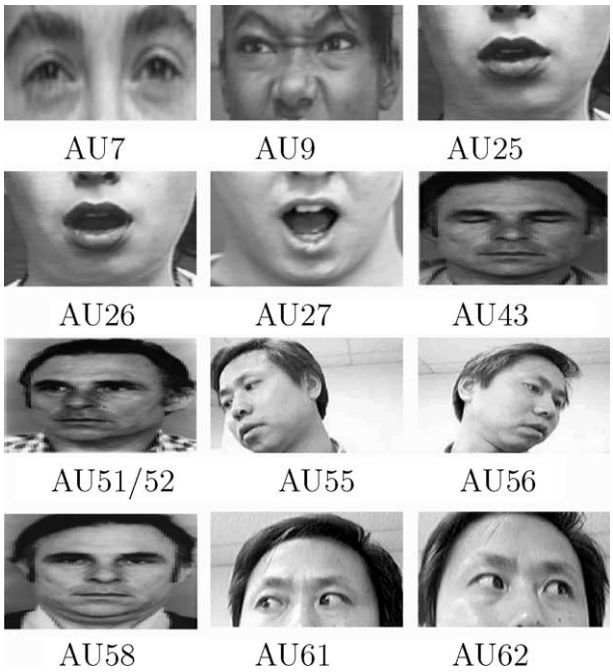



Fig. 8. Action units for components of fatigue facial behaviors (some adapted from [11]).

Table 1
Definitions of AUs in Fig. 8

| AU# | Descriptions |
|---|---|
| AU7 | Lid tightener |
| AU26 | Jaw drop |
| AU51/52 | Head turn left/right |
| AU58 | Head back |
| AU9 | Nose wrinkle |
| AU27 | Mouth stretch |
| AU55 | Head tilt left |
| AU61 | Eyes turn left |
| AU25 | Lips part |
| AU43 | Eye closure |
| AU56 | Head tilt right |
| AU62 | Eyes turn right |

local facial features, such as pupils, head trajectory, upper face and lower face. The connection between the entire and partial layers is based on each entire display and its corresponding feature appearances. In the single AU layer, all of related single AUs are included and are connected to its partial facial display based on FACS.

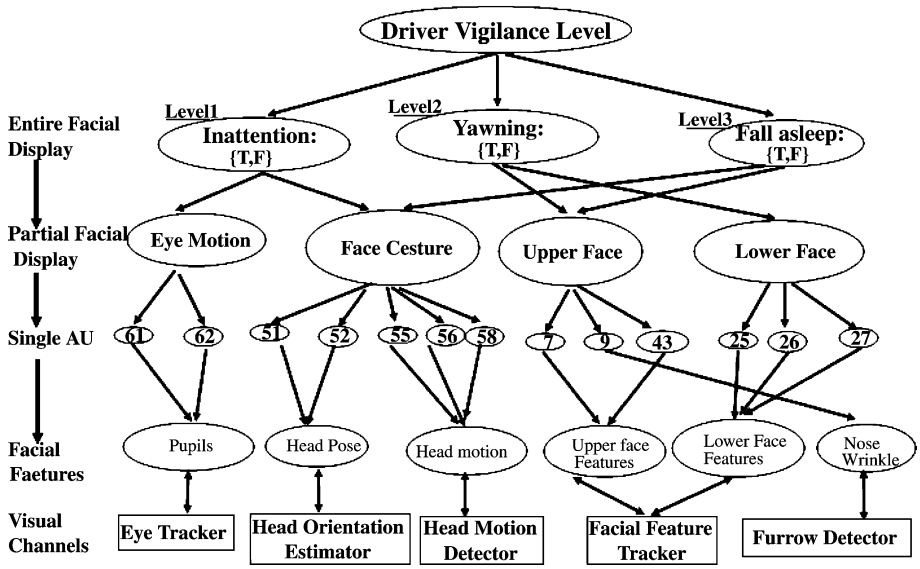



Fig. 9. BN model for driver vigilance estimation.

In Fig. 9, we also embed the visual sensors into the BN. The visual channels are not part of the Bayes net, but are used to acquire (instantiate) the corresponding visual feature nodes. The bidirectional link in the channel level is to emphasize the information control flow. One is the bottom–up in which the sensor provides the observable evidences, the other the top–down in which the sensor is activated by the inference of BN.

For our research, the BN model is constructed based on a combination of domain knowledge and the principles FACS. While the domain knowledge may be used to relate facial expression to different mental states (the first level), FACS provides a mechanism for relating facial expressions to AUs (the second level). Finally, the AUs are related to the corresponding facial features (third level). At each level, the relationships should be casual to represent them using BNs. Unfortunately, construction of BN is more of an art than science and it varies from application to application. We choose a task-oriented approach to reduce the modeling scope and complexity.

The alternative to manual BN structure specification is to learn the structure from data. But this requires a lot of training data, and the learned structures are often different from the manually specified structures. As a compromise, a BN structure may be initially specified by human, and subsequently refined through learning.

In the BN model, the top layer can be viewed as a simple Naive Bayes classifier. It consists of hypothesis variable $C$ with three mutually exclusive states $c_1, c_2, c_3$, i.e., inattention, yawning and falling asleep. The hypothesis variable with a set of attributes $A_1$, $A_2, A_3$ (corresponding to Ia, Yw, and Fa) is the parent node and each attribute corresponds to one child node. The goal for this layer is to find the probability of state $c_i$ given $A_j = a_j$:

$$\Pr(C = c_i | A_1 = a_1, A_2 = a_2, A_3 = a_3).$$

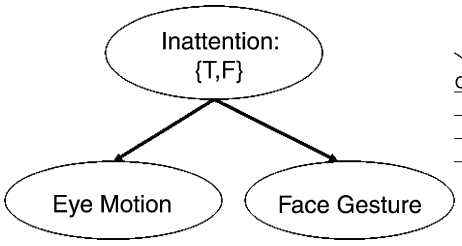
In other words, this probability represents the chance of the class state $c_i$ when each attribute variable has the value where $A_j = a_j$. When the probability of a specific state $c_i$ is highest, it is most likely that the facial behavior belongs to the $c_i$. The initial conditional probabilities between the states of hypothesis variable and the corresponding attributes are given in Table 2. These numbers are subsequently refined through a training process.

In the intermediate layers, the spatial relationships are represented by one of the typical BN connections: serial, diverging, and converging connection [18]. For instance, Fig. 10 depicts the diverging connection between *Inattention* and its child nodes. The tables on the right hand side show the initial conditional probability tables (CPT) between the parent and child nodes in this connection.

In the bottom layer, each single AU is associated with a specific displacement measurement of facial features. The displacement measurements are quantified into a set of discrete values as summarized in Table 3. We connect each single displacement measurement to the corresponding visual sensor. Table 3 gives the AUs, corresponding the quantitative feature displacements that measure the AUs, the associated visual channel and measuring method based on the feature labels in Fig. 3. For instance, $AU7$ is an Upper face local feature, which can be measured by changes of the angles $\angle IFH$ and $\angle HGF$, based on currently tracked feature points $I, F, H$, and $G$. In the initialization, all the feature points are extracted and their associated parameters are obtained as the initial parameters in the neutral state of the current subject. These initial parameters are used as the reference parameters or thresholds in the recognition processing.

Table 2
Initial probabilities

| Entire display | Ia | | Yw | | Fa | |
|---|---|---|---|---|---|---|
| | *T* | *F* | *T* | *F* | *T* | *F* |
| Inattention | 0.99 | 0.01 | 0.05 | 0.95 | 0.05 | 0.95 |
| Yawning | 0.05 | 0.95 | 0.99 | 0.00 | 0.05 | 0.95 |
| Falling asleep | 0.05 | 0.95 | 0.05 | 0.95 | 0.99 | 0.01 |




| Parent / Child | Inattention | |
|---|---|---|
| Eye | T | F |
| T | 0.9 | 0.1 |
| F | 0.1 | 0.9 |

| Parent / Child | Inattention | |
|---|---|---|
| Face | T | F |
| T | 0.75 | 0.25 |
| F | 0.25 | 0.75 |

Fig. 10. an example of diverging connections in the BN model.

Table 3
Measurement of AUs

| AU | Method | Appearance | Sensing channel |
|---|---|---|---|
| AU7 | $\angle IFH$ non-increased and $\angle HGF$ increased | Upper face | Facial tracker |
| AU9 | Wrinkle increased in $\square JFF'J'$ | Upper face | Furrow detector |
| AU25 | $\overline{DB} < T_1$ and $\overline{NA}$ non-increased | Lower face | Facial tracker |
| AU26 | $T_1 < \overline{DB} < T_2$ | Lower face | Facial tracker |
| AU27 | $\overline{DB} > T_2$ and $l$1 increased | Lower face | Facial tracker |
| AU43 | Pupils lost for a while | pupil | Eye tracker |
| AU51 | $\overleftarrow{OZ}$ turn left | Head orientation | Orientation estimator |
| AU52 | $\overleftarrow{OZ}$ turn right | Head orientation | Orientation estimator |
| AU55 | Head moves left-down | Head motion | Head motion detector |
| AU56 | Head moves right-down | Head motion | Head motion detector |
| AU58 | Head moves back | Head motion | Head motion detector |
| AU61 | $E$ and $E'$move left | pupil | Eye tracker |
| AU62 | $E$ and $E'$ move right | pupil | Eye tracker |

*Note.* $T_1$ and $T_2$ are predefined thresholds.

### 4.2. Temporal dependence

Usually, a facial behavior evolves overtime starting from the onset, the apex and the offset. Static BN (SBN) modeling of facial expression works with visual evidences and beliefs from a single time instant. It lacks the ability to express temporal relationship and dependencies in a video sequence. To overcome this limitation, we use dynamic BNs (DBN) for modeling the dynamic aspect of the facial behavior.

In Fig.11, the DBN is made up of interconnected time slices of SBNs as described in the preceding section. The relationships between two neighboring time slices are modeled by the first order hidden Markov model, i.e., random variables at temporal

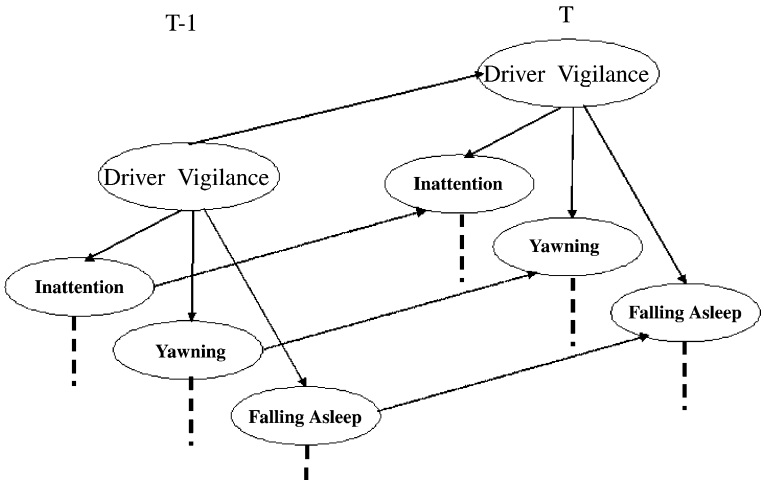



Fig. 11. DBN model in terms of temporal instance $T$.

instance $T$ are affected by observable variables at $T$, as well as by the corresponding random variables at preceding temporal instance $T-1$ only. The evolution of facial expression is defined by moving one time frame in accordance with the process of video sequences, so that the visual information at the previous time provides diagnostic support for current facial behavior. Eventually, the belief of the current hypothesis of the mental state is inferred relying on the combined information of current visual cues through causal dependencies in the current time slice, as well as the preceding evidences through temporal dependencies.

In contrast to one frame as the temporal instance used in previous DBN based sequence analysis works, we also define another time scale in the task-oriented network, that is the *phase instance*. One phase instance consisting of more than one frame is a continuous facial display, which proceeds for a certain period till enough recognition confidence is achieved. The temporal dependency in the complicated facial behavior can be modeled by the phase instance as well as the frame instance. At each frame instance, the bottom–up inference is executed to classify each current facial display. In each phase instance, the top–down inference is conducted to actively select effective visual channel(s). In the next subsection, we will provide the details about the two-time scaling mechanism for inference.

### 4.3. Purposive sensing

Active or purposive sensing decision with Bayesian has been explored in vision systems recently. Paletta et al. [31] proposed an active recognition system, where BN is used for fusing successive object hypotheses; and mutual information is used to quantify the ambiguity to determine which view to select to achieve a maximum of discrimination in recognition process. Denzler et al. [10] introduced a formalism on the basis of mutual information theory to measure the uncertainty in the state estimation process for sensor parameter selection. However, we significantly differ from their works primarily in two aspects. First, our approach concentrates on dynamic and multiple visual sensor environment; while theirs are restricted to static systems. Second, our work emphasizes on how to purposively select a subset of diverse visual channels; whereas theirs focus on how to actively select parameters of a single sensor to effectively perform the perceptual task. In summary, our work is aimed at performing active and purposive visual information from multi-sensory visual channels (feature regions) to achieve the facial expression recognition under dynamic environment (video sequences), where the world situations (facial emotions) as well as sensory observations (visual cues) vary overtime.

We embed the multiple visual channel system into a DBN framework, since DBNs provide a dynamic knowledge representation and control structure that allows sensory information to be selected and combined according to the rules of probability theory and recognition tasks targeted.

Usually to find the best feature region for facial feature extraction, we need to exhaustively search all the visual channels. However, the task-oriented approach provides an efficient method for purposive selection. As described before, in the temporal dependence of our DBN, they exist in two time scales. One is the phase instance which

usually consists of several consecutive frames, the other the frame instance. Fig. 12 depicts a vigilance-detection temporal dependence in term of phase instances. Each facial behavior recognition will proceed in two consecutive phases: detection and verification. Based on the domain knowledge of driver vigilance detection, any fatigue process starts with the Inattention behaviors. On this vigilance-task oriented system, we set the Inattention as the first facial behavior to be targeted. At first, the Inattention detection (Ia-Detection) phase is triggered. With the specified detection target in this phase, the related visual channels are selected and activated based on the top–down inference of our task-oriented BN. As vigilance intensity progresses overtime, it is possible for drivers to change the facial behaviors from "Inattention" to "Yawning" or "Falling Asleep," the "Yawning" and "Falling Asleep" are included in the detection targets in the succeeding instances. After one Inattention facial display is detected with enough confidence, the detection phase ends and the system evolves into the Inattention verification (Ia-Verification) phase. Meanwhile, the Yawning detection (Ya-Detection) phase is triggered. In this way, at each moment we can determine the entire targeted facial displays according to the current phase instance. Furthermore, from the entire facial display, a top–down BN inference is conducted to identify the associated single AUs so as to activate the corresponding visual channels. From each *frame* slice, the system checks the extracted information from activated visual channels and collects the evidences. With the observed evidences, a bottom–up BN inference is conducted to classify the current facial display. We summarize the purposive sensing in one recognition cycle as follows:

1. Set facial behaviors in the detection phase. With the top–down BN inference, identify the associated single AUs and activate the corresponding visual channels.

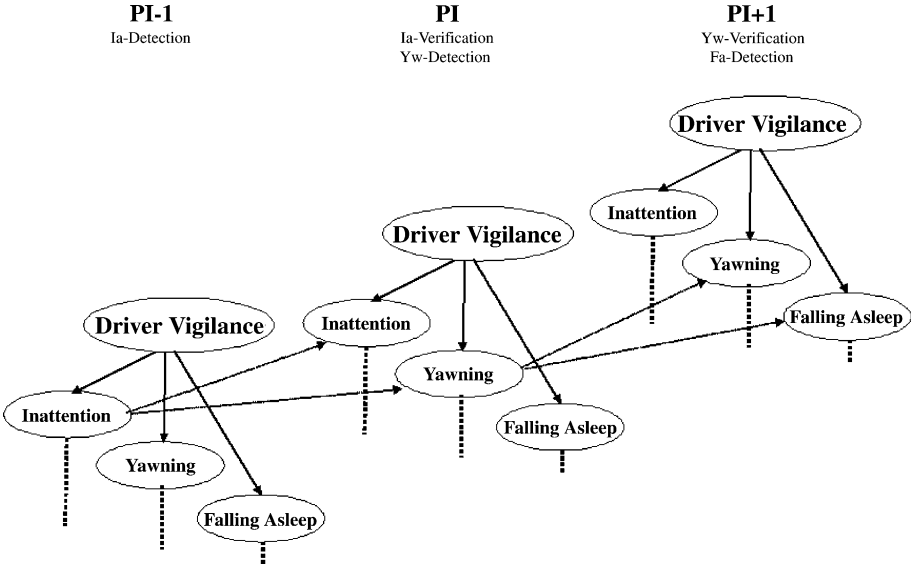



Fig. 12. DBN model in terms of phase instance.

2. In each frame slice, classify the current facial display by the bottom–up BN inference from observed evidences and seek the ONSET of the facial behavior by the curve of the posterior probability.
3. If the ONSET is identified with high confidence, evolve into the verification phase in step 4. Otherwise goto step 2.
4. Based on the targeted display(s) in the verification phase, the top–down inference is used to activate the associated visual channels.
5. From the activated visual channels, collect the observation evidences and seek the APEX of the current facial behavior frame by frame.
6. If the APEX in the verification phase is identified with high confidence, reset the recognition cycle, goto step 1. Otherwise goto step 5.

## 5. Experimental results

Currently there is few available database for mental states recognition. For each specific application, its even difficult to find the common dataset to evaluate. To validate our approach, we create a small-size dataset for the driver vigilance level detection. Subjects (total five persons) were asked to sit in a normal indoor environment and simulate several fatigue procedures from a weak level (neutral or inattention ) to a serious level (yawing and fall sleeping). Each procedure starts from front-view facial expression and lasts about 10 seconds to one minute. The facial expression video sequences (about 100–600 frames per person) were recorded using our Infrared camera system. The leave-one-out cross validation is used for training and testing. We use all of sequences except the test one to train the CPTs of DBN. Based the obtained DBN, we conduct experiments to characterize the performance of our method.

### *5.1. Static facial display classification*

Fig. 13 displays some frames of typical expressions from two different fatigue sequences. Figs. 13A, B, and C depict the neutral state, inattention and yawning state of one person respectively. Figs. 13D, E, and F show the inattention, yawning and falling asleep states in another sequence. We set prior probability for all the levels (inattention,yawning, and falling asleep) as the same to conduct the bottom–up classification from observations.

Table 4 shows observed evidences from each sensing channel and the classification result for each image. ED, OE, HD, FT, and FD in the table indicate Eye Detector, Head orientation estimator, Head motion detector, Facial tracker and Furrow detector, respectively. (Ia), (Yw), and (Fa) represent Inattention, Yawning, and Falling asleep, respectively. "$x$" means no evidence. Row (a) of Table 4 shows that there is no evidence available from all the fatigue related channels in the image of Fig. 13A. The posterior probability of each fatigue state obtaining from the Static BN inference has the same value as 0.33. It means that the person is in the neutral

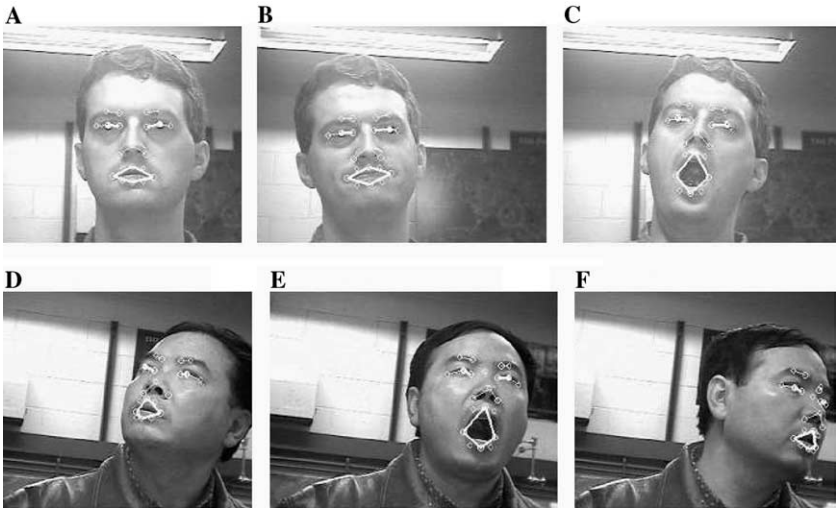



Fig. 13. Some typical facial displays in fatigue sequences with tracked features marked.

Table 4
Static classification

| No. | Observation evidences | | | | | Classification results | | |
|---|---|---|---|---|---|---|---|---|
| | ED | OE | HD | FT | FD | (Ia) | (Yw) | (Fa) |
| (a) | *x* | *x* | *x* | *x* | *x* | 0.33 | 0.33 | 0.33 |
| (b) | *x* | AU51 | *x* | *x* | *x* | **0.82** | 0.09 | 0.09 |
| (c) | *x* | *x* | *x* | AU7 AU27 | AU9 | 0.15 | **0.83** | 0.02 |
| (d) | *x* | AU52 | *x* | *x* | *x* | **0.82** | 0.09 | 0.09 |
| (e) | *x* | *x* | *x* | AU7 AU27 | *x* | 0.17 | **0.81** | 0.02 |
| (f) | *x* | AU51 | AU55 | *x* | *x* | **0.47** | 0.05 | **0.47** |

state, or the states in which there are no fatigue related facial displays. Row (b) in Table 4 represents that in Fig. 13B the evidence of AU51 is detected from the visual channel (OE). A high posterior probability is obtained for *Inattention* state. The facial behavior is classified as *Inattention*. Row (c) shows that AU7, AU27, and AU0 are detected from the visual channels FT and FD, respectively. The classification result is the *Yawning*. Rows (d) and (e) show that the corresponding classification results are *Inattention* and *Yawning* states, respectively. Row (f) shows that AU51 and AU55 are detected from channels OE and HD, respectively. The posterior probabilities of *Inattention* and *Falling asleep* are the same (0.47). At this moment, there exists ambiguity in the static facial display. The system is difficult to identify fatigue states.

The above static recognition experiments show that a good classification can easily be obtained for typical facial displays with the task-oriented BN. However, when

the ambiguity exists among the displays, or in the transition period of states, the temporal information is required to resolve the ambiguity.

### 5.2. Dynamic facial expression classification

#### 5.2.1. A moderate fatigue sequence

Fig. 14 shows a moderate fatigue sequence, which consists of samples from one 210 frame sequence at the interval of 10 frames. The clip starts with 20th frame neutral states. Then several facial behaviors of *Inattention* are performed with some neutral states at the transition moments. Finally the person opened his mouth and performed facial behaviors of *Yawning*.

Fig. 15 depicts the posterior probability curves of the three fatigue levels, obtained by the task-oriented DBN system. In the first 20 frames, there are no fatigue related evidences detected. The probability values of three levels are the same as 0.333. From frame 21 to 185, the posterior probability of *Inattention* is the highest among the

Fig. 14. One 210-frame video clip with a blended facial displays of neutral, Inattention and Yawning state.

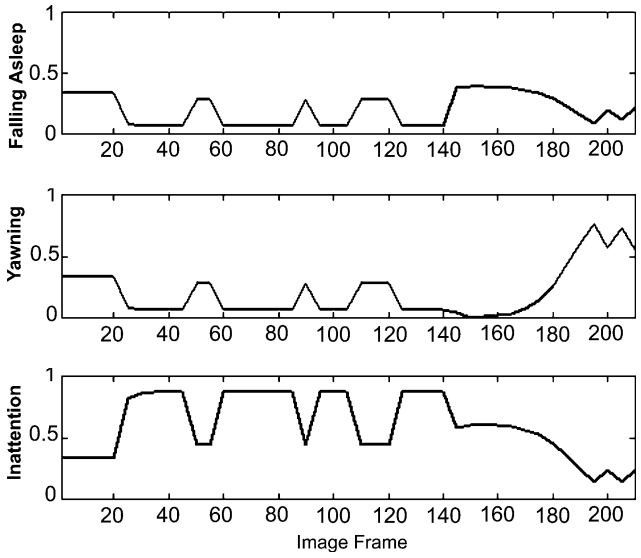



Fig. 15. The obtained posterior probability curves of three states.

three states. Within the period, the probability curve of *Inattention* goes down several times because of neutral states between different *Inattention* facial displays. At those moments, there are no fatigue related evidences observed at all. With the temporal dependence from previous classification results, the mental state of the subject is still classified as *Inattention* although the value is comparatively lower. After frame 185, the Yawning evidences (AU25, 26, 27) are consistently detected. The probability curve of *Yawning* increases gradually. During this period, the curve also reduces one time because evidences of both Yawning and Inattention are detected simultaneously at around the frame 200.

The experiments of the sequence validates that the DBN-based approach successfully integrates temporal information in the previous moments to remove the ambiguity due to multiple conflict evidences or transition periods.

Fig. 17 depicts the posterior probability values of three states obtained from the inference of DBN. We can see in this graph that the confident level goes down due to some supporting evidences missing, such as at frames 70, 80. Because of the integration of the previous results and the knowledge of application domain, the classification is still stable even at these moments.

Fig. 13F in the static experiment corresponds to the frame 110 in the sequence. The posterior probabilities of *Inattention*, *Yawning* and *Falling asleep* are 0.0079, 0.0034, and 0.9888, respectively. The temporal dependence based on DBN significantly removed the ambiguity in the static classification, shown as in row (F) of Table 4.

Fig. 18 depicted the selection of activated visual channels during the DBN-based recognition. According to the task-oriented DBN, the Inattention related visual channels, which are Head motion detector (HD), Head orientation estimator (OE), and Eye detector (ED), were firstly activated. At frame 5 the evidence of ''Inattention'' was detected from OE channel. The system started to focus on OE. At

Fig. 16. Samples of a typical fatigue sequence, which consists of ''Inattention,'' ''Yawning,'' and ''Falling asleep'' states in order.

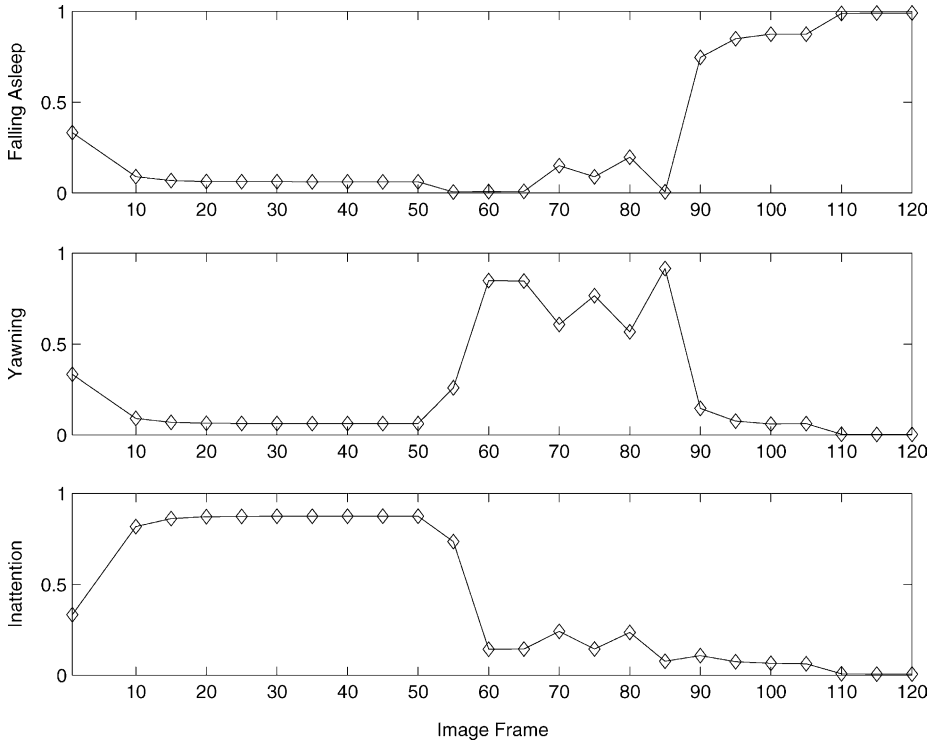



Fig. 17. The obtained posterior probability curves of three states.

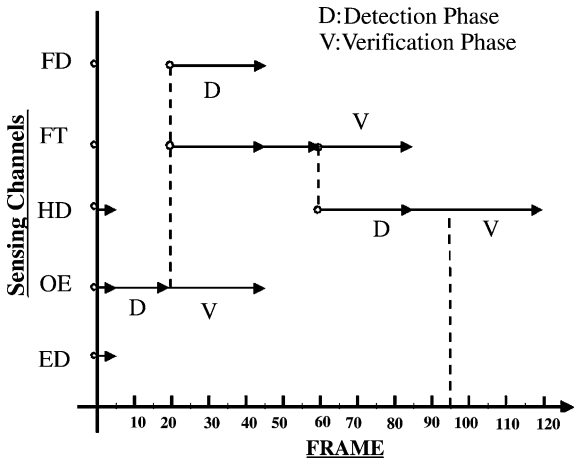



Fig. 18. Active sensor selection for the fatigue sequence.

frame 20, the ONSET of the Inattention display was detected with the curve of the posterior probability of Inattention in Fig. 17. So the system finished the detection phase of Inattention and evolved into the verification phase of Inattention. At the

same time the detection phase of Yawning was triggered and the associated sensing channels, which are Facial tracker (FT) and Furrow detector (FD) were activated. identified the APEX of the Inattention expression and verified the Inattention state. After that, at frame 45 the new evidence on Yawning was detected from the FT channel. The system focused on it. At frame 60, the ONSET of the Yawning expression was identified. The system evolved into the verification phase of Yawning and the detection phase of Falling asleep. The HD channel was activated. At frames 75 and 85, one APEX of Yawning was identified and the Yawning state was verified, respectively. From frame 85, the new evidence on Falling asleep was detected and the system focused on the HD channel. At frame 95, the ONSET of the Falling asleep was identified. The system further evolved into the verification phase of the Falling asleep. Finally the APEX of the Falling asleep was identified by the curve of the posterior probability of Falling asleep in Fig. 17.

The experimental sequence simulates a typical fatigue evolution process, with only a very short transition period. In reality, the Yawning state sometimes lasts a certain long period and does not evolve into the Falling asleep state. In this case, the activated channels will keep the same for a while. Sometimes, the Yawning expression may disappear for some reasons. In this case, the system cannot detect any evidences of Yawning for a certain period time. All three posterior probabilities tend to be low and will go to the same after a while. The fatigue states will disappear at this point. The system will reset to the detection phase for Inattention.

### *5.3. Quantitative performance evaluation and selective sensing*

In this section, we present the results from a quantitative evaluation of our method. To this end, the images sequences we mentioned were manually scored for each frame. Specifically, three probabilities are subjectively assigned to each frame to represent the probability of the image being Inattention, Yawning, and Fall Asleep respectively. These numbers, though quite subjective, serve as the ground-truth probabilities. The ground-truth numbers are subsequently compared with the probabilities produced by our system frame by frame. Fig. 19 provides the ground-truth probabilities (bold) versus the estimated probabilities for the three states for a complete sequence. The two curves apparently track each other well, but discrepancies do occur in some frames. The mean frame-wise MSQ probability difference for each state is summarized in Table 5. While subjective, this experiment, to certain degree, demonstrates validity of our approach.

To further quantitatively characterize the performance of the proposed technique, We conducted another experiment using the two sequences as previously shown in Fig. 14 and Fig. 16. Table 6 and Table 7 summarize the statistics of classifications from the sequences the two sequences respectively, compared against the ground truth by the manual inspection. The table gives the number of correctly and incorrectly classified image frames for each state, where the facial behaviors in the first row are the ground-truth expressions while the facial behavior in the first column are the recognized results. We can see that there are a relatively small number of frames misclassified, and overall performance of our system is good.

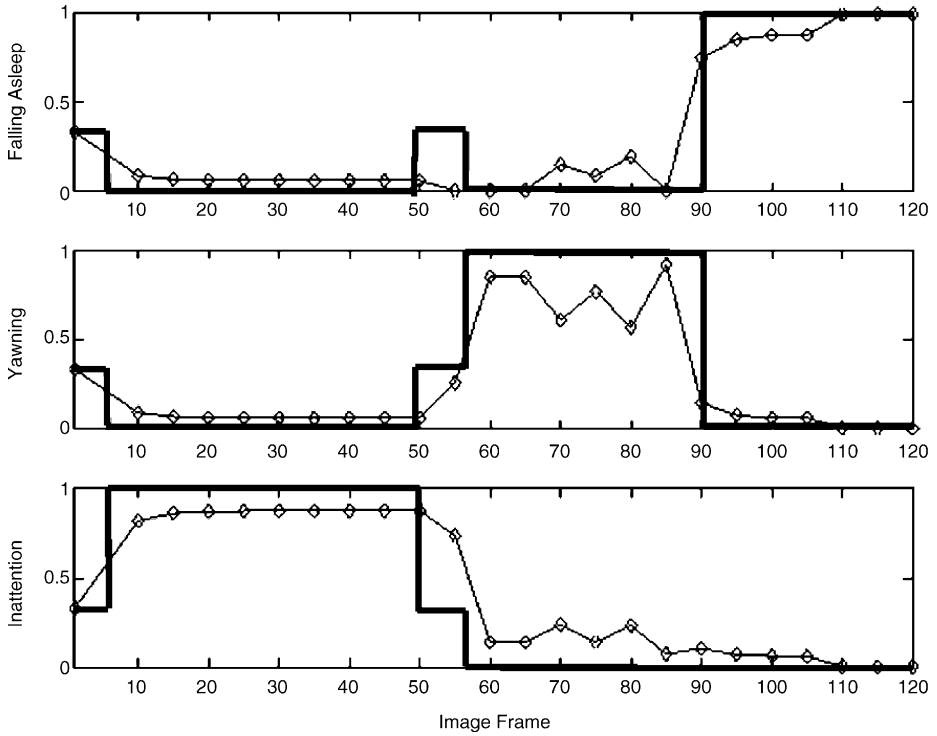



Fig. 19. Quantitative comparison of the estimated user state probabilities versus the ground-truth probabilities (bold).

Table 5
The average estimate error for each state

| | Inattention | Yawning | Falling asleep |
|---|---|---|---|
| Recognition error | 0.13 | 0.11 | 0.08 |

Table 6
The confusion table for the sequence as shown in Fig. 14

| | Inattention | Yawning | Falling asleep | Neutral |
|---|---|---|---|---|
| Inattention | 140 | 0 | 0 | 20 |
| Yawning | 0 | 30 | 0 | 0 |
| Falling asleep | n/a | n/a | n/a | n/a |
| Neutral | 0 | 0 | 0 | 20 |

To evaluate the proposed active sensing strategy, we use a segment from another image sequence for illustration. If we gather all visual channels for the information of facial feature changes each time, this will cost unnecessary computations and also cause more ambiguities in classification. We therefore choose visual channels purpo-

Table 7
The confusion table for the sequence as shown in Fig. 16

| | Inattention | Yawning | Falling asleep | Neutral |
|---|---|---|---|---|
| Inattention | 45 | 0 | 0 | 0 |
| Yawning | 0 | 30 | 5 | 0 |
| Falling asleep | 0 | 0 | 30 | 0 |
| Neutral | 5 | 0 | 0 | 5 |

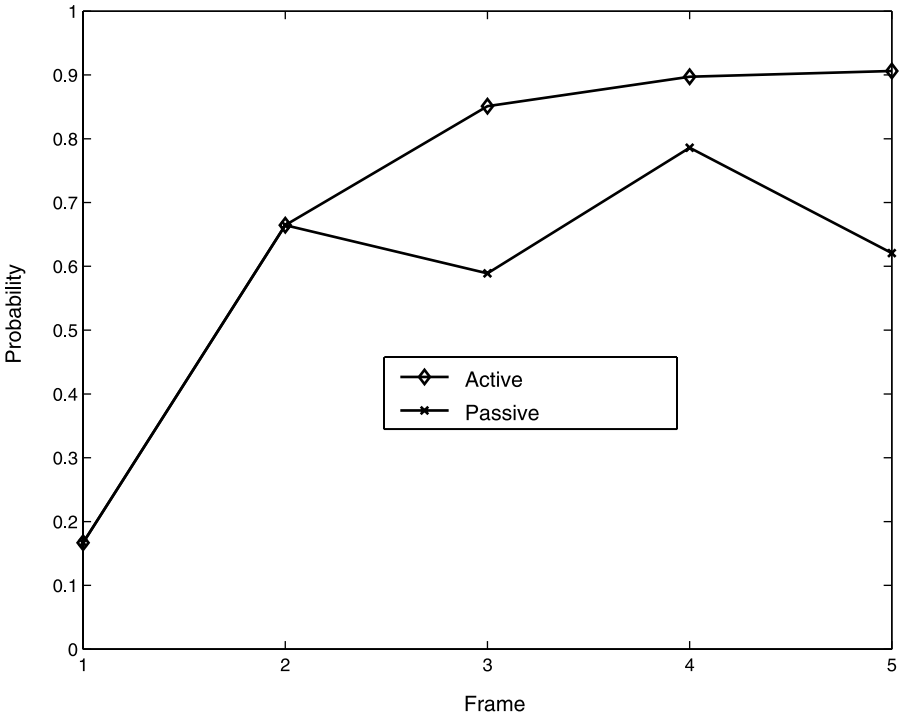



Fig. 20. Comparative result between active sensing and passive sensing (sensing all available visual channels in this example).

sively that visual cues are the most goal-relevant. In this example, we select the two most informative visual channels through the top–down inference to acquire the visual evidences for integration. The comparative result given in Fig. 20 shows that, the active sensing can achieve quickly in uncertainty reduction than the passive sensing (sensing all available visual channels in this case).

## 6. Conclusion

In this paper, we presented a practical and efficient framework for real-world facial behavior recognition. The proposed method has several favorable properties:

- This framework is based on general facial behavior description language (FACS) and dynamic Bayesian network (DBN). This approach can be used for different applications.
- The domain knowledge and previous analysis results are successfully integrated systematically to remove ambiguities in facial behavior displays.
- The purposive sensing structure among the multiple visual channels makes the recognition of facial behaviors more efficient and can be easily extended to encapsulate other new effective visual channels, such as estimation of gaze and eyelid movement, even for sensors of different modalities.

The quantitative evaluation of our framework described in this paper is inadequate. One important future work is to more throughly validate the proposed framework. This requires collecting more domain specific data under real conditions, with data labeled by experts. We are currently working on this issue. Another future issue is to improve the model parameterization method. Current method is mostly empirical, with limited training. We will work on an automated parameterization of the DBN model from the training database of each application domain.

## Acknowledgment

The research described in this paper is supported in part by a Grant from AFOSR.

## References

[1] M.S. Bartlett, B. Braathen, G.L. Littlewort-Ford, J. Hershey, J. Fasel, T. Mark, E. Smith, T.J. Sejnowski, J.R. Movellan, Automatic Analysis of Spontaneous Facial Behavior: A final Project Report. Technical Report MPLab-TR2001.08, University of California San Diego, Dec. 2001.
[2] J.N. Bassili, Emotion recognition: the role of facial movement and the relative importance of upper and lower area of the face, J. Pers. Soc. Psychol. 37 (1979) 2049–2059.
[3] M.J. Black, Y. Yacoob, Recognizing facial expression in image sequences using local parameterized models of image motion, Internat. J. Comput. Vis. 25 (1) (1997) 23–48.
[4] A. Bobick, Y. Ivanov, Action recognition using probabilistic parsing, in: Proc. IEEE Internat. Conf, Computer Vision Pattern Recognition 1998, pp. 196–202.
[5] C. Bregler, Learning and recognizing human dynamics in video sequences, in: Proc. IEEE Internat. Conf. Computer Vision Pattern Recognition, 1997, pp. 568–574.
[6] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, T. Huang. Learning bayesian network classifiers for facial expression recognition using both labeled and unlabeled data. IEEE Conf. Computer Vision Pattern Recognition, 2003.
[7] I. Cohen, N. Sebe, A. Garg, L.S. Chen, T.S. Huang, Facial expression recognition from video sequences: temporal and static modeling, Comput. Vis. Image Und. 911 (2003) 60–187.
[8] J.F. cohn, A.J. Zlochower, J.J. Lien, T. Kanade, Feature-point tracking by optical flow discriminates subtle difference in face expression, in: IEEE Internat. Conf. Automatic Face Gesture Recognition, pp. 396–401, 1998.
[9] G.W. Cottrell, J. Metcalfe. EMPATH: face, emotion, gender recognition using holos, in: R.P. Lippman (Ed.), Advances in Neural Information Processing Systems 3, 1991, 564–571.

[10] J. Denzler, C.M. Brown, Information theoretic sensor data selection for active object recognition and state estimation, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2) (2002) 145–157.
[11] P. Ekman, W.V. Friesen, Facial Action Coding System (FACS): Manual, Consulting Psychologists Press, Palo Alto, CA, 1978.
[12] P. Ekman, W.V. Friesen, J.C. Hager, Facial Action Coding System (FACS): Manual, CD Rom, San Francisco, CA, 2002.
[13] I.A. Essa, A.P. Pentland, Coding, analysis, interpretation, recognition of facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 757–763.
[14] M. Gladwell, The naked face:can you read people's thoughts just by looking at them? The New Yorker magazine 5 (2002) 38–49.
[15] H. Gu, Q. Ji, Information extraction from image sequences of real-world facial expressions, Mach. Vis. Appl. 16 (2) (2005) 105–115.
[16] J. Hoey, J. Little. Decision theoretic modeling of human facial displays, in: Proc. Eur. Conf. Comput. Vis. 2004.
[17] J. Hoey, J. Little. Value directed learning of gestures and faial displays, in: Proc. IEEE Internat. Conf. Computer Vision Pattern Recognition 2004.
[18] F.V. Jensen, Bayesian Networks and Decision Graphs, Springer-Verlag, New York, 2001.
[19] Q. Ji, X. Yang. Real-time 3D face pose discrimination based on active ir illumination, in: Internat. Conf. Pattern Recognition 2002.
[20] Q. Ji, Z. Zhu, Eye and gaze tracking for interactive graphic display, in: Smart Graphics, Hawthorne, NY, USA, 2002.
[21] R. Kaliouby, P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures, in: Proc. in CVPR Workshop on Real-Time Computer Vision for Human Computer Interaction, 2004. p. 154.
[22] M. Kato, I. So, Y. Hishnuma, O. Nakamura, T. Minami. Description and synthesis of facial expressions based on isodensity maps, in: T. Kunii (Ed.), Visual Computing, 1991, pp. 39–56.
[23] G.D. Kearney, S. Mckenzie, Machine interpretation of emotion: design of memory-based expert system for interpreting facial expressions in terms of signaled emotions (JANUS), Cogn. Sci. 17 (4) (1993) 589–622.
[24] S. Kimura, M. Yachida. Facial expression recognition and its degree estimation, in: Proc. IEEE Internat. Conf. Computer Vision Pattern Recognition. 1997, pp. 295–300.
[25] H. Kobayashi, F. Hara, Recognition of six basic facial expressions and their strength by Neural Network, in: Proc. Internat. Workshop Robot and Human Communication, 1992, pp. 381–386.
[26] H. Kobayashi, F. Hara, Facial interaction between animated 3D face robot and human beings, in: Internat. Conf. Syst., Man, Cybern. 1997, pp. 3732–3737
[27] J.J. Lien, T. Kanade, J.F. Cohn, C. Li, Detection, tracking, and classification of action units in facial expression, Internat. J. Robot. and Autonomous Syst. 31 (1997) 131–146.
[28] K. Mase, Recognition of facial expression from optical flow, IEICE Trans. E 74 (10) (1991) 3474–3483.
[29] C. Morimoto, M. Flickner, Real-time multiple face detection using active illumination, in: Proc. Fourth IEEE Internat. Conf. on Automatic Face and Gesture Recognition, 2000, pp. 8–13.
[30] N. Oliver, E. Horvitz, A. Garg. Layered representations for human activity recognition, in: Proc. IEEE Internat. Conf. on Multimodal Interfaces, 2002, pp. 3–8.
[31] L. Paletta, A. Pinz, Active object recognition by view integration and reinforcement learning, Robot. Autonomous Syst. 31 (2000) 71–86.
[32] M. Pantic, L. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2002) 1424–1445.
[33] M. Pantic, L. Rothkrantz, Expert system for automatic analysis of facial expression, J. Image Vis. Comput. 18 (11) (2000) 881–905.
[34] V. Pavlovic, A. Garg, M. Rehg. Multimodal speaker detection using error feedback dynamic Bayesian networks, in: Proc. IEEE Internat. Conf. Computer Vision and Pattern Recognition, 2000.
[35] A. Rahardja, A. Sowmya, W.H. Wilson, A Neural Network approach to component versus holistic recognition of facial expressions in images, SPIE, Intelligent Robots and Computer Vision X: Algorithms and Techniques 1607 (1991) 62–70.

[36] M. Rosenblum, Y. Yacoob, L. Davis. Human emotion recognition from motion using a radial basis function network architecture, in: Proc. IEEE Workshop on Motion of Non-Rigid and Articulated Objects, 1994, pp. 43–49.
[37] N. Sebe, M.S. Lew, I. Cohen, A. Garg, T.S. Huang, Emotion recognition using a cauchy naive bayes classifier, in: Internat. Conf. on Pattern Recognition, 2002.
[38] Y. Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 97–115.
[39] Y. Tian, T. Kanade, J.F. Cohn. Recognizing Facial Actions by Combining Geometric Features and Regional Appearance Patterns Robust lip tracking by combining shape, color and motion. Technical Report Technical Report CMU-RI-TR-01-01, Robotics Institute, Carnegie Mellon University, 2001.
[40] H. Ushida, T. Takagi, T. Yamaguchi. Recognition of facial expressions using conceptual fuzzy sets, in: Proc. Conf. Fuzzy Systems, 1993, pp. 594–599.
[41] L. Wiskott, J. Fellous, N. Kr ger, C.V. der Malsburg, Face recognition by elastic bunch graph matching, in: IEEE Trans. on Pattern Analysis Machine Intelligence, 1997, pp. 775–779.
[42] Y. Yacoob, L. Davis. Recognition facial expressions by spatio-temporal analysis, in: Proc. Internat. Conf. Pattern Recognition, 1994, pp. 747–749.
[43] Y. Yacoob, L.S. Davis, Recognizing human facial expressions from long image sequences using optical flow, IEEE Trans. Pattern Anal. Mach. Intell. 18 (6) (1996) 636–642.
[44] Z. Zhang, M. Lyons, M. Schuster, S. Akamastsu. Comparison between geometry-based and gabor wavelets-based facial expression recognition using multi-layer perception, in: Proc. Internat. Conf. Automatic Face and Gesture Recognition, 1998, pp. 454–459.
[45] J. Zhao, G. Kearney. Classifying facial emotions by backpropagation neural networks with fuzzy inputs, in: Proc. Internat. Conf. Neural Information Processing, 1996, pp. 454–457.
[46] Z. Zhu, Q. Ji, K. Fujimura, K. Lee. Combining Kalman filtering and mean shift for real time eye tracking under active ir illumination, in: Proc. Internat. Conf. Pattern Recognition, 2002.