Active Facial Tracking for Fatigue Detection

Haisong Gu, Qiang Ji†, Zhiwei Zhu
Dept. of Computer Science,University of Nevada Reno
Dept. of ECSE, Rensselaer Polytechnic Institute†
haisonggu@ieee.org
qji@ecse.rpi.edu

Abstract

The vision-based driver fatigue detection is one of the most prospective commercial applications of facial expression recognition technology. The facial feature tracking is the primary technique issue in it. Current facial tracking technology faces three challenges: (1) detection failure of some or all of features due to a variety of lighting conditions and head motions; (2) multiple and non-rigid object tracking; and (3) features occlusion when the head is in oblique angles. In this paper, we propose a new active approach. First, the active IR sensor is used to robustly detect pupils under variable lighting conditions. The detected pupils are then used to predict the head motion. Furthermore, face movement is assumed to be locally smooth so that a facial feature can be tracked with a Kalman filter. The simultaneous use of the pupil constraint and the Kalman filtering greatly increases the prediction accuracy for each feature position. Feature detection is accomplished in the Gabor space with respect to the vicinity of predicted location. Local graphs consisting of identified features are extracted and used to capture the spatial relationship among detected features. Finally, a graph-based reliability propagation is proposed to tackle the occlusion problem and verify the tracking results. The experimental results show validity of our active approach to real-life facial tracking under variable lighting conditions, head orientations, and facial expressions.

1. Introduction

Faces as the primary part of human communication have been a research target in computer vision for a long time. Many applications such as Human-Computer Interaction (HCI), interactive gaming, and driver fatigue detection, etc. heavily rely on facial expression recognition technology. The driver fatigue detection is considered as one of the most prospective commercial applications of automatic facial expression recognition [11]. Automatic recognition(or analysis) of facial expression consists of three levels of tasks: face detection, facial expression information extraction, and expression classification.

In these tasks, the information extraction is the main issue for the feature-based facial expression recognition from an image sequence. It involves detection, identification, and tracking facial feature points under different illuminations, face orientations, and facial expressions. The tracking results will eventually determine the classification quality and application performance. Current techniques for facial feature tracking can be divided into model-free and face model based ones. The model free methods[9, 10, 2] assuming one 3D rigid motion with local shape deformation can directly obtain 3D object motion and easily be applied to different objects. But the face consists of several parts independently controlled by their own facial muscles[4], a single 3D rigid motion is hard to accurately represent a wealth of expression changes. The feature points often get lost and drifted due to rapid head motion, light variation and intensity degenerated intensity[2]. The post processing is usually required to identify each tracked feature. The face model based methods[1, 12] have the advantage of integrating available knowledge on a specific object and application. The feature identification and detection are obtained simultaneously. Some common assumptions in previous face related works were : frontal facial views, the constant illumination, and the fixed lighting source. Unfortunately these assumptions are not realistic. In the application of real world facial expression understanding, we have to consider at least three issues: (1) capturing the

full features in a variety of lighting conditions and head motion; (2) multiple and nonrigid object tracking; and (3) the self-occlusion of features.

The eye pupil as a key feature on the face is often used as the starting point for expression information extraction. IR (Infra-Red) based eye tracking techniques[5, 13] have been proved to be a stable approach to detect pupils and face location. In this paper, we developed a IR based sensing system to stably detect the pupil positions under variable lighting conditions. These pupil positions provide strong and reliable constraints for detection and tracking of other facial features.

Kalman filtering is a useful means for object tracking. It can impose a smoothing constraint on the face motion. For the trajectory of each facial feature, this constraint removes random jumping due to the uncertainty in the image. We combine the Kalman filtering with the pupil motion to predict the current location of each feature. By doing so, we not only obtain a smooth trajectory for each feature, but also catch the rapid head motion. Given the predicted feature position, the multi-scale and multi-orientation Gabor wavelet method [7, 8, 12] is used to detect each facial feature in the vicinity of the predicted location. The detection in the Gabor space provides an accurate and fast solution for multi-feature tracking.

The facial expression is looked as a pattern change of the entire face. It is important for us to not only track each single feature, but also identify the facial feature and capture the spatial relationship between features. The Gabor wavelet based method is used to identify each feature in the tracking initialization. The Gabor coefficients are updated at each frame to adaptively represent the feature profile change. These updated coefficients (or profile) are used as the template to match the feature on the ongoing frame. The updating approach commonly used in adaptively tracking works very well when no occlusion or self-occlusion happens. Considering the free head motion in which the head can turn around from the frontal view to the side view or vice verse, the self-occlusion often makes the tracker to fail because a random or arbitrary profile is assigned to the occluded feature. In this paper a graph-based and reliability propagation method is developed to tackle this problem and refine the tracking results. The basic idea is to use the related reliable features to verify and infer unstable features. By Active in the paper, it does not just mean to use active IR illumination, the more important thing is to actively make use of the reliable and believable information to achieve the highly accurate tracking.

The organization of the paper is as follows: In the next section we outline the active sensing system. Section 3 provides our pupil-guided and Kalman filtering based approach to multi-feature tracking. Section 4 presents tracking refinement with reliability propagation. The experimental results for fatigue facial expressions are provided in Section 5. Finally conclusions are in Section 6.

2. IR Active Facial Sensing

Our active facial sensing system consists of an IR sensitive camera and two concentric rings of IR LEDs as shown in Fig. 1 (a). The circuitry was developed to synchronize the inner ring of LEDs and outer ring of LEDs with the even and odd fields of the interlaced image respectively. The interlaced input image is subsequently de-interlaced via a video decoder, producing the even and odd field images as shown in Figure 1 (b) and (c). Basically the two fields are the similar images. Significant differences happen on the pupil areas. One image is related to bright pupils, the other dark pupils. So the pupils are easy to be detected and tracked from the difference of the two images[6, 13].

(a) (b) (c)

Figure 1. (a) Hardware setup: the camera with an active IR illuminator (b) bright pupils in an even field image (c) dark pupils in an odd field image

The active sensing system not only provides us the reliable information to indicate where the face and the pupils are, but also the dark image sequence (odd fields) presents a normal grayscale sequence in almost any lighting conditions, which allows us to use the conventional methods to extract and track other facial features.

3. Facial Feature Tracking

3.1. Feature-based Facial Representation

The facial features around eyes and mouth represent the most important spatial patterns composing

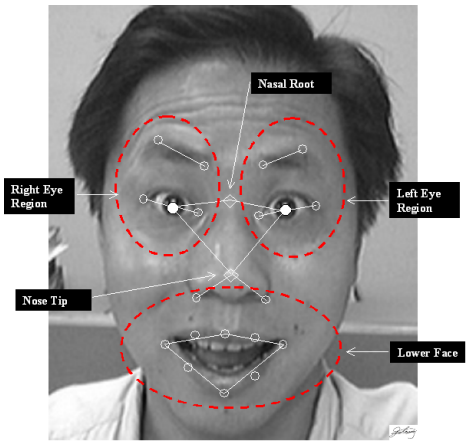

Figure 2. The facial features and local graphs

the facial expression display. Generally these patterns (or local graphs) with their spatio-temporal changes and the synchronization relationships can describe almost the facial expressions. For the fatigue detection, in which there are only limited facial expression displays, the facial features around eyes and the mouth contain enough information to capture these limited expressions. So here, we use 22 fiducial features around eyes and the mouth, and three local graphs as the facial model (shown in Fig.2).

The feature-based tracking approach, especially with the Gabor wavelet, has the psychophysical basis of human vision [3] and can achieve a good performance [8, 12] for facial expression recognition. In the paper we use the multi-scale and multi-orientation Gabor wavelet to represent each feature. The 2D Gabor kernels used are as follows:

$$\Psi(\mathbf{k}, \vec{x}) = \frac{\mathbf{k}^2}{\sigma^2} e^{-\frac{\mathbf{k}^2 \vec{x}^2}{2\sigma^2}} (e^{i\mathbf{k}\cdot\vec{x}} - e^{-\frac{\sigma^2}{2}}) \qquad (1)$$

where $\sigma = \pi$ is set for 128x128 images. The set of Gabor kernels consists of 3 spatial frequencies (with wavenumber k: $\pi/2, \pi/4, \pi/8$), and 6 distinct orientations from 0^o to 180^o in the interval of 30^o. 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.

$$\begin{aligned} \Omega(\vec{x}) = & \quad (m_1 e^{i\phi_1}, m_2 e^{i\phi_2}, ..., m_{18} e^{i\phi_{18}})^T \\ = & \quad \int I(\vec{x}^{'}) \Psi[\mathbf{k}, (\vec{x} - \vec{x}^{'})] d^2 \vec{x}^{'} \end{aligned}$$

These coefficients can be used to represent this pixel and its vicinity [7]. Here this coefficient set is not only used to detect and identify each facial feature at the initial frame, but also conduct tracking processing as the adaptive template of each feature.

3.2. Kalman Filter with Pupil Constraints

The facial feature tracking is a crucial task because the accurate tracking is the necessary base for successful expression classification. The face is a typical nonrigid object. The feature profiles and the relationships among features can independently change according to the facial expressions. Tracking for these features is a tough issue. It involves three basic problems. (1) Easy to get lost due to the rapid head motion; (2) feature random jumping due to the noise; and (3)the self-occlusion when the head turns around up to the profile view. For the first two problems, we develop a pupil-guided Kalman filtering solution. It mainly consists of feature prediction and feature detection. For the occlusion issue we will address in the section **4**.

3.2.1 Feature prediction The pupil positions from our active sensing provide the reliable information, which indicates where the face roughly locates and how the head changes entirely. This kind of information allows the system to catch facial features even under rapid head movement.

On the other hand, the Kalman filter is a well-known tracking method. It puts a smooth constraint on the motion of each feature. For each feature, its motion state at each time instance (frame) can be characterized by its position and velocity. Let (x_t, y_t) represent its pixel position and (u_t, v_t) be its velocity at time t in x and y directions. The state vector at time t can therefore be represented as $S_t = (x_t\ y_t\ u_t\ v_t)^T$. The system can therefore be modelled as

$$\mathbf{S}_{t+1} = \Phi \mathbf{S}_t + \mathbf{W}_t \qquad (2)$$

where Φ is the transition matrix and $\mathbf{W}_t$ represents system perturbation.

We further assume that a feature detector estimates the feature position ($\mathbf{O}_t = (\hat{x}_t, \hat{y}_t)^T$) at time t. Therefore, the measurement model in the form needed by the Kalman filter is

$$\mathbf{O}_t = H\mathbf{S}_t + \mathbf{V}_t \qquad (3)$$

where H is the measurement matrix and $\mathbf{V}_t$ represents measurement uncertainty. Given the state model in equation (2) and measurement model in equation (3) as well as some initial conditions, the state vector $\mathbf{S}_{t+1}$, along with its covariance matrix

Σ_{t+1}, can be updated using the system model and measurement model. Based on the state model in equation (2), the position prediction of each feature ($P^k = (x^k, y^k)^T$) can be obtained. Meanwhile we also get the error covariance matrix Σ_{t+1} to represent the uncertainty of the current prediction from Kalman filtering.

By combining the head motion with the Kalman filtering, we can obtain an accurate and robust prediction of feature location in the current frame, even under rapid head movement. The final predicted position for each facial feature

$$\widehat{P}_{t+1} = P^f_{t+1} + e^{-\Sigma_{(x,y)}}(P^k_{t+1} - P^f_{t+1}) \qquad (4)$$

where the entire head motion $P^f (= (x^f, y^f)^T)$ is the average of pupil motions between two consecutive frames. $\Sigma_{(x,y)}$ is the 2×2 sub-matrix with the 1st and 2nd diagonal entries of the covariance matrix Σ_{t+1}.

3.2.2 Feature detection The above feature prediction from previous frames provides a small area centered at each predicted position. Usually the Kalman filter based tracker use the error covariance matrix to limit the area size. The searching process within the area is used to detect the optimal position. But in tracking for a mass of features, this processing is the time-consumption and is not acceptable for the real-time implementation. Here a fast detection method is used.

For the pixel ($\vec{\mathbf{x}}$)in the small vicinity of the predicted position($\vec{\mathbf{x}}^{'}$), the phase shift of the Gabor coefficients $\Omega(\vec{\mathbf{x}})$ from $\vec{\mathbf{x}}^{'}$ can approximately be compensated by the terms $\vec{d} \cdot \vec{k_n}$. The $\vec{d}$ indicates the displacement from the predicted position($\vec{\mathbf{x}}^{'}$). So the phase-sensitive similarity function of these two pixels can be

$$\mathbf{S} = \frac{\sum_n m_n m_n^{'} cos(\phi_n - \phi_n^{'} - \vec{d} \cdot \vec{k_n})}{\sqrt{\sum_n m_n^2 \sum_n m_n^{'2}}} \qquad (5)$$

where m_n and ϕ_n indicate the amplitude and phase in the complex Gabor coefficients $\Omega(\vec{\mathbf{x}})$, respectively.

The similarity function can be approximated by its Taylor expansion as :

$$\mathbf{S} \approx \frac{\sum_n m_n m_n^{'}[1 - 0.5(\phi_n - \phi_n^{'} - \vec{d} \cdot \vec{k_n})^2]}{\sqrt{\sum_n m_n^2 \sum_n m_n^{'2}}} \qquad (6)$$

By maximizing the above function, we can get the optimal displacement vector of feature position.

$$\vec{d}_{\,o} = \frac{1}{\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx}} \begin{pmatrix} \Gamma_{yy} & -\Gamma_{yx} \\ -\Gamma_{xy} & \Gamma_{xx} \end{pmatrix} \begin{pmatrix} \theta_x \\ \theta_y \end{pmatrix} \qquad (7)$$

if $\Gamma_{xx}\Gamma_{xy} - \Gamma_{xy}\Gamma_{yx} \neq 0$, with

$$\begin{aligned} \theta_x &= \Sigma_n m_n m_n^{'} k_{nx}(\phi_n - \phi_n^{'}) \\ \theta_y &= \Sigma_n m_n m_n^{'} k_{ny}(\phi_n - \phi_n^{'}) \\ \Gamma_{xx} &= \Sigma_n m_n m_n^{'} k_{nx} k_{nx} \\ \Gamma_{xy} &= \Sigma_n m_n m_n^{'} k_{nx} k_{ny} \\ \Gamma_{yx} &= \Sigma_n m_n m_n^{'} k_{nx} k_{ny} \\ \Gamma_{yy} &= \Sigma_n m_n m_n^{'} k_{ny} k_{ny} \end{aligned}$$

Basically, the phase-sensitive similarity function can only determine the displacements up to half wavelength of the highest frequency kernel, which would be ± 2 pixel area centered at the predicted position for $k = \pi/2$. But this range can be increased by using low frequency kernel. Currently a three level coarse to fine approach is used, which can determine up to ± 8 pixel displacement. For each feature only three displacement calculations are needed to determine the optimal position, which dramatically speeds up the detection processing and makes the real-time implementation possible.

4. Tracking Refinement

4.1. Facial Model Extraction

The face is looked as the spatial pattern composed of facial features. The entire pattern not only represents the spatial relationship among facial features, but also can be used to verify detected facial features. When facial tracking is considered as the object tracking task with multiple features, it involves two issues. One is the nonrigid object tracking, the other is self-occlusion. So far many research works conducted feature tracking under (1) the head motion without facial expression change, or (2) different expression displays with only frontal views. They focused on one of the above two issues. But in real-life facial expression tasks, both issues should be taken into consideration. We have to deal with the profile change of each feature due to different expressions and the self-occlusion due to the head motion from the frontal view to the side view, or vice verse. In order to tackle these issues, we propose a graph-based and reliability propagation approach.

To accurately locate facial features under the two conditions mentioned above, here a facial feature is not looked as an isolated point. Each feature is related to one or all of three local graphs: left eye, right eye and lower face(shown in Fig.2). With the known pupil positions and a Common Facial Model, the position of each feature at the initial frame will be extracted and identified in the tracking initialization. From these extracted features, their Gabor coefficients and spatial relationships are used to create a Personalized Facial Model for the current face.

During tracking, the Gabor coefficients as the template (or profile) of each feature are updated frame by frame in order to handle the change due to different expressions. With the detected features and their spatial relationship, the local graphs are also updated.

4.2. Reliability Propagation

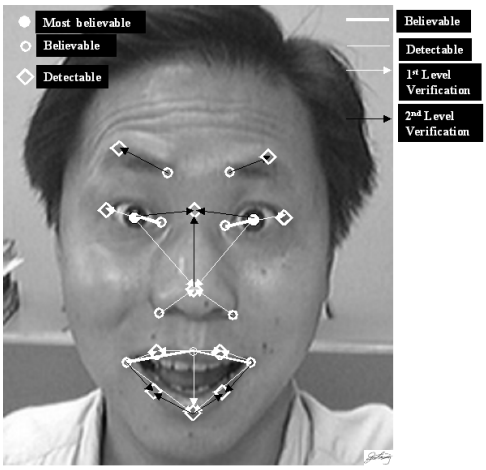

Figure 3. The verification pass

In a nonrigid object tracking, the profile updating at each frame is a reasonable way to handle the profile change. However, when a specific feature is occluded, the updated profile goes out of meaning because in the current frame there is no detectable visual information for updating. This is a difficult issue. First of all, we have to update the profile frame by frame so as to handle the intensity change due to oblique face orientations. On the other hand, some nonsense profile is accepted as the template to conduct the ongoing tracking. Furthermore, it is difficult to detect when and where the occlusion happened if we only focus on a single facial feature. Here a refinement method based on reliability propagation is proposed.

According to the feature property and their positions on the face model, we divide all of the facial features into 3 types: Most believable(two pupils), Believable and Detectable as indicated by filled circles, circles and diamonds respectively in Fig.3. We also create the spatial relationship between them and assign one of two types (Believable and Detectable) for each connection.

For each Detectable feature, based on the related connection with Most believable or Believable features, its detected position will be verified in the arrow direction shown in Fig.3.

Since the propagation levels and the connection strengths between features are different for different Detectable features, a verification strategy is created for all the Detectable features in the reliability order(shown in Fig.4).

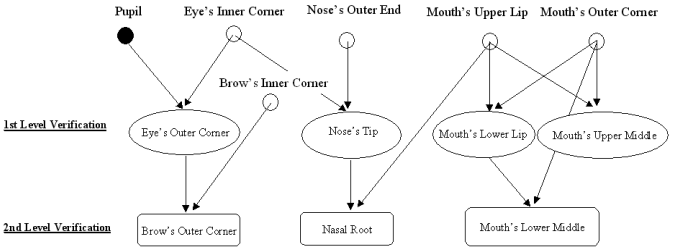

Figure 4. The verification strategy

It consists of two level verifications. In the first level, only Most believable and Believable features are used to infer the Detectable features. In the second, the modified Detectable features also join in the verification. The confidence level of the verification will go down in the order. In the reliability propagation, the qualitative and quantitative evaluations are conducted. The qualitative evaluation checks the correctness of the spatial relationships, such as whether the two corner points of the eye are on the different sides of the pupil, etc. The quantitative one compares the extracted relationship data(aspect ratios and joint angles) with the corresponding data in the Personalized Facial Model. If a wrong feature is detected by the evaluation, the prediction is conducted with the known relationship data obtained in the tracking initialization.

5. Experimental Results

Fig. 5 shows the 449 frame sequence of a person in fatigue. The person in the scene performed several common facial expressions of driver drowsing. He firstly yawned from the neutral state, then moved the head rapidly from the frontal view to the large side view and back in the opposite direction, rapidly raised the head up and lowered it down, finally returned to the neutral state. During the head motion, the facial expression changes dramatically. The long sequence includes typical tracking issues, such as feature intensity variation due to oblique face orientations, rapid head motion and self-occlusion, etc. The IR sensor provided the even and odd sequences, corresponding to bright and dark pupil images, respectively. The odd sequence with detected pupil positions are used as input for facial tracking.

Figure 5. The fatigue expression sequence

Table 1. Comparison of three approaches

Approaches	Total features	Losing features	Accuracy ratio
Simple tracking	9856	854	91.3 %
Pupil-Kalman	8960	214	97.6 %
Pupil-Kalman with propagation	8960	77	99.1 %

Table 1 reports the results by the pupil guided and Kalman filtering based method without, and with reliability propagation. For comparison, we also shows the result from linear motion model based tracking (called simple tracking). There are 22 facial features in each frame. In the simple tracking, the pupils were also detected and tracked. So the total feature number is 9856(= 22 × 448). There were 854 features incorrectly tracked in the whole sequence. Table 2 shows the details.

Table 2. Failure in the simple tracking

Features	Failure occurring Periods [Start frm., End frm.]	Losing Features
LP	[54,234], [270,343]	253
LB(2)	[75,121]	92
MF(8)	[400,433]	264
OCR	[188,218]	30
RNE	[400,433]	33
OCL	[75,198],[378,433]	178
LNE	[399,403]	4

The detection failure happened on the Left Pupil (LP), the Left eyebrow (2 points)(LB), the Mouth Features (8 points)(MF),the Outer Corner of the Right eye(OCR), the Right Nose End(RNE), the Outer Corner of the Left eye(OCL) and the Left Nose End(LNE). The reasons of these failures come from (1) unstable and degenerated intensity information (LP, LB), (2) rapid head motion (MF, OCR, RNF), (3) self-occlusion (OCL,LNF)

With the constraints from pupil positions and Kalman filtering, the tracking failure reduced significantly (Table 3). Since the pupil position was provided, we remove them from the total tracking number. So the total feature number is 8960. The tracking failure due to the rapid head motion on MF, RNE and OCR were improved.

Table 3. Failure with the constraints

Features	Failure Occurring Periods [Start frm., End frm.]	Losing Features
LB(2)	[75,95]	40
OCL	[75,110], [373,488]	150
LNE	[398,422]	24

With the reliability propagation, the further improvement for the self-occlusion issue has been reached (Table 4).

Table 4. Failure after refinement

Features	Failure Occurring Periods [Start frm., End frm.]	Losing Features
LB(2)	[75,95]	40
OCL	[398,411]	13
LNE	[398,422]	24

One of the self-occlusion case was shown in Fig.6. Fig6(a) displays the self-occlusion of the Left eye at No. 70 frame. Fig6(b) and (c) show the tracked OCL at No. 87 frame without and with verification, respectively. The tracking failure due to the occlusion on the outer corner of left eye(OCL) was corrected by the reliability propagation. Video demos of tracking results for the whole sequence and other subjects are available at http://www.cs.unr.edu/~*zhu_z/Demo/demo.html*.

6 Conclusion

In this paper, we proposed the active approach to facial tracking of real-life facial expressions. The accuracy improvements stem from: (1) active sensing, which allows us to robustly detect pupils and the head motion; (2) combination of the Kalman filtering with the head motion to accurately predict the features locations; (3)the use of Gabor wavelet for fast feature detecting; (4) the reliability propagation based on spatial relationships to handle oc-

(a) (b) (c)

Figure 6. (a) the occlusion on OCL at No. 70 frame (b) the result without (c) the result with verification

clusion to refine the tracking results. The extracted local graphs and their spatio-temporal relationships will be used to conduct the expression classification. It is one of our future research.

7 Acknowledgements

This work is supported by a grant from US Air Force office of Scientific Research.

References

[1] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. ICCV, 1995.

[2] F. Bourel, C. C. Chibelushi, and A. A. Low. Robust facial feature tracking. MBVC, 2000.

[3] J. Daugman. Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. IEEE Trans. ASSP, 36:1169–1179, 1988.

[4] P. Ekman. Facial Expressions. Handbook of Cognition and Emotion. John Wiley Sons Ltd, New York, 1999.

[5] Haro, Antonio, M. Flickner, and I. Essa. Detecting and tracking eyes by using their physiological properties, dynamics, and appearance. IEEE CVPR, pages 163–168, 2000.

[6] Q. Ji and X. Yang. Real time visual cues extraction for monitoring driver vigilance. Proc. of International Workshop on Computer Vision Systems, July 2001.

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

[8] B. Manjunath, R. Chellappa, and C. von der Malsburg. A feature based approach to face recognition. IEEE CVPR, pages 373–378, 1992.

[9] C. Tomasi and T. Kanade. Detection and tracking of point features. Carnegie Mellon University Technical Report, (CMU-CS-91-132), April 1991.

[10] L. Torresani and C. Bregler. Space-time tracking. ECCV, 2002.

[11] H. Veeraraghavan and N. P. Papanikolopoulos. Detecting driver fatigue through the use of advanced face monitoring techniques. Technique Report of ITS Institute, (CTS 01-05), September 2002.

[12] Z. Zhang. Feature-based facial expression recognition: Experiments with a multi-layer perceptron. Technical report INRIA, (335), 1998.

[13] Z. Zhu, Q. Ji, K. Fujimura, and K. Lee. Combining kalman filtering and mean shift for real time eye tracking under active ir illumination. proc. of ICPR, pages 373–378, August 2002.