Combining Stochastic and Deterministic Search for Pose-Invariant Facial Expression Recognition

Shiro Kumano, Kazuhiro Otsuka, Junji Yamato,
Eisaku Maeda, and Yoichi Sato
Institute of Industrial Science, The University of Tokyo
{kumano,ysato}@iis.u-tokyo.ac.jp
NTT Communication Science Laboratories
{otsuka,yamato}@eye.brl.ntt.co.jp
maeda@cslab.kecl.ntt.co.jp

Abstract

We propose a novel method for pose-invariant facial expression recognition from monocular video sequences that combines stochastic and deterministic search processes. We use the simple face model called variable-intensity template, which can be prepared with very little time and effort. We tackle the two issues found in previous work on the variable-intensity template: low accuracy in head pose estimation, and assumption violations due to external intensity changes such as illumination change. We mitigate these issues by introducing the deterministic approach into the stochastic approach implemented as a particle filter. Our experiment demonstrates significant improvements in recognition performance for horizontal and vertical head orientations in the range of $\pm$40 degrees and $\pm$20 degrees, respectively, from the frontal view.

1 Introduction

To realize sophisticated human-computer interaction systems and to automatically analyze human conversation structure [12], a number of image-based facial expression recognition methods have been proposed. Most of them adopt the frontal-face assumption [1, 9, 14, 16]: The image shows a nearly frontal view of the user's face and the user does not rotate the head significantly. This assumption is, however, often violated in real situations such as multi-party conversations, where people will often turn their faces to look at other participants. Hence, unless a head-mounted self-shot camera is allowed, e.g. [13], we must simultaneously handle the variations in head pose as well as facial expression changes.

For handling out-of-plane head rotations, the face model[1] should be accurate enough to reliably separate a change in face appearance into facial pose and expression components. In other words, the use of inaccurate face models degrades the accuracy of both

[1] It refers to a set of a face shape model and facial expression model in this paper.

head pose estimation and facial expression recognition. Unfortunately, it is not easy to generate an accurate face model for each user because the intra-personal variations in human face shape and expression are nontrivial; simply preparing one general model is not satisfactory. There are two main approaches to resolving this issue: creating an accurate person-specific face model and utilizing a general face model with facial features that well handle some error.

Gokturk et al. [6] and Wang et al. [18] generate a person-specific model by using stereo cameras and a 3D digitizer, respectively. Accordingly, this approach cannot be applied to monocular video sequences. Dornakia and Davoine [5] adequately fit a general face model to each user by shifting multiple control points manually, a task that is too expensive to be practical. Lucey et al. [10] reconstruct the user's face model from just monocular images by a structure-from-motion technique. However, in their experiments, the recovered face model rather degraded the recognition rates.

Some methods try to handle out-of-plane head rotations by using robust facial features without accurate face shapes. Black and Yacoob [3], and Tong et al. [15] utilize optical flow and wavelet coefficients in facial part regions as features, respectively. Although the plane shapes they use are convenient for calculating dense features, they cannot correctly express the large change in facial appearance caused by significant out-of-plane head rotation.

Kumano et al. [8] recognize facial expressions from changes in the intensities of a set of sparse interest points fixed on a cylinder. Their interest points are located away from the edges of facial parts, to detect the shift of neighboring facial parts as well as to alleviate the impact of the misallocation caused by approximation error in the facial shape. Their model, the variable-intensity template, is advantageous since it can be prepared with little time and effort. To estimate facial pose and expression simultaneously, they use a particle filter, which maintains their posterior probability density function given input face images with a set of multiple hypotheses stochastically generated. This makes the estimation easy to avoid local maxima and to recover from temporary disturbance.

The method in [8], however, still has two shortcomings: First, the stochastic search generally requires a large number of hypotheses to accurately estimate the target states. Furthermore, since a rough shape, a cylinder, is used, the most likely state may be significantly different from the actual state. Second, their method cannot handle large changes in face intensity such as those caused by illumination changes or vertical head rotations.

To overcome these two shortcomings, two key features are newly introduced in this work: For the first problem, the stochastic approach of particle filters is combined with the deterministic approach of a gradient method. This makes the estimation robust and effective. First, an approximate estimator is robustly provided by the particle filter. It is then seeded into the maximum likelihood estimation (MLE) that is efficiently solved by the gradient method. To make the search meaningful, we utilize an average face model as the face shape. For the second problem, the intensities of the interest points are deterministically adjusted by an iteratively reweighted least squares technique, without increasing the dimension of the stochastic search space.

Consequently, this paper has two explicit advantages over the work in [8]:

1. Higher performance in the estimation of both facial pose and expression.

2. Illumination-invariability.

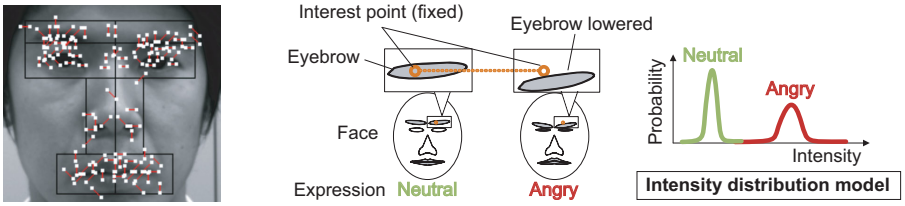


Figure 1: (Left) An example of a set of interest points $\mathcal{P}$: Small white rectangles represent interest points. (Right) Intensity distribution model $\mathcal{I}$: The intensity distributions of interest points, described as normal distributions, change with facial expression.

Our method can be considered as one of regularization in maximum likelihood estimation, where a prior term is adopted. In general regularization, the weight of prior is gradually reduced, while our framework first considers prior density and likelihood as equal terms, and the prior is ignored thereafter. To the best of our knowledge, at least for facial expression recognition, this is the first attempt to introduce a posterior density obtained by a stochastic search as a prior for gradient-based maximum likelihood estimation. For example, Chang and Ansari [4] guide all hypotheses along the gradient directions of the likelihood function by a mean shift algorithm. This process is beneficial if the task is just object tracking and the likelihood function is very smooth. However, such a likelihood function does not suit for the discrimination of facial expressions.

The remainder of this paper is organized as follows. First, facial expression recognition based on the variable-intensity template is briefly introduced in Section 2. Section 3 describes our proposed method. Then, in Section 4, experimental results are given. Finally, a summary and future work are given in Section 5.

2 Stochastic estimation with variable-intensity templates

In this section, we briefly overview the method in [8] as well as several modifications introduced in this work. The system flow with the variable-intensity template consists of two stages. First, a variable-intensity template is prepared for each person from just a general face shape and one frontal face image for each facial expression (hereinafter referred to as training images). Second, the facial pose and expression in a test sequence are estimated.

2.1 Variable-intensity templates

The variable-intensity template [8] consists of three components: a rigid face shape model $\mathcal{M}$, a set of interest points $\mathcal{P}$, and an intensity distribution model $\mathcal{I}$. The variable-intensity template is basically person-specific except for the shape model. This achieves high recognition performance. As the face shape model $\mathcal{M}$, unlike the cylinder used in [8], we use an average face model[2]. The fitting of the face model is described in Section 2.1.1.

[2]We used an average head dummy of Japanese young males, published by Digital Human Research Center, Advanced Industrial Science and Technology (AIST), *http://www.dh.aist.go.jp/research/centered/facedummy/*.

The set of interest points, $\mathcal{P} = \{p_i\}_{i=1}^{N_p}$, consists of sparse multiple points fixed on the shape model $\mathcal{M}$, as shown in the left part of Fig.1. The variable p_i denotes the image coordinates of the *i*-th interest point in the training image, and N_p denotes the number of interest points. These interest points are extracted as dipoles that straddle the edges of facial parts, such as eyebrows, in the neutral expression image. Details are provided in [8].

The intensity distribution model $\mathcal{I}$ describes how the interest point intensities vary with different facial expressions. As shown in the right part of Fig.1, the interest point intensities change strongly due to the shift of neighboring facial parts. This property enables us to recognize facial expressions from observed interest point intensities. The interest point intensities are assumed to be independent of each other, and to follow a normal distribution in all facial expressions. Hereinafter, the mean and standard deviation of the *i*-th interest point for facial expression category $e(\in \{1,\ldots,N_e\})$ are denoted by $\mu_i(e)$ and $\sigma_i(e)$, respectively.

2.1.1 Training of variable-intensity template

Starting with just the general face shape and the training face images, or one frontal face image for each facial expression, the variable-intensity template is automatically prepared for each person. In the training images for the same subject, the face is assumed to be stationary between all facial expressions[3]. We actually captured the face of each user without any head movement during the demonstration of facial expressions. We set the intensity mean $\mu_i(e)$ to be the value recorded from the training image labeled with expression e at image coordinate p_i. Furthermore, we assume that standard deviation σ_i is proportional to the mean and empirically set it as $\sigma_i = \mu_i/3$.

We fit the average face shape to each user as follows: (1) fit the center of the shape model to the center of face region in the training image (neutral expression) as detected by the method of [17], (2) stretch the shape model in the horizontal and vertical directions to match both face width and height; stretching in the depth direction uses the scaling factor given as the square root of the product of vertical and horizontal scaling factors.

2.2 Estimation of posterior probability density function

The joint posterior probability density function (pdf) of facial pose and expression at time t given all face images up to that time, $I_{1:t}$, is recursively represented as follows:

$$p(h_t, e_t|I_{1:t}) \propto L(h_t, e_t|I_t) \int p(h_t|h_{t-1}) \sum_{e_{t-1}} P(e_t|e_{t-1}) p(h_{t-1}, e_{t-1}|I_{1:t-1}) dh_{t-1} \qquad (1)$$

where $L(h, e|I)$ denotes joint likelihood of head pose h and facial expression e for image I (described in Section 2.3), and the facial pose state h and expression state e follow first order Markov processes; h_t and e_t are assumed to be conditionally dependent given image I_t. The facial pose state consists of the following six continuous variables: the two-dimensional translation in the image plane, $[x\ y]^{\mathrm{T}}$, three-dimensional rotation angles (pitch, θ_x, yaw, θ_y, and roll, θ_z), and scale s. The posterior pdf in Eq. (1) is stochastically estimated in the frame work of a particle filter [7].

[3]The position and size of the faces may vary for different persons.

The particle filter approximates the pdf as a set of weighted hypotheses called particles. Each particle expresses a state and its weight w: $\{h_t^{(l)}, e_t^{(l)}, w_t^{(l)}\}_{l=1}^{N}$ and $\sum_l w_t^{(l)} = 1$, where N denotes the number of particles. In our case, $w_t^{(l)} \propto L(h_t^{(l)}, e_t^{(l)} | I_t)$.

For the head motion model $p(h_t|h_{t-1})$, rather than the simple random walk model used in [8], we utilize an adaptive random walk model where the system noise increases as the head moves more significantly: $h_t = h_{t-1} + g(v_{t-1})$, where v denotes the velocity of head pose, and $g(v)$ is a zero-mean multivariate Gaussian process with covariance that varies according to $|v|$. For covariance adaption, we follow [11] except that we assume that the head pose components are independent of each other, and we don't use a higher-order motion model to avoid overfitting. With regards to facial expression, we set $P(e_t|e_{t-1})$ to be equal for all expression combinations.

Estimators of facial pose and expression, $\widetilde{h}_t$ and $\widetilde{e}_t$, are calculated as their expectations of marginal density: $\widetilde{h}_t = \sum_l w_t^{(l)} h_t^{(l)}$ and $\widetilde{e}_t = \arg\max_e \sum_l w_t^{(l)} \delta_e(e_t^{(l)})$, where $\delta_e(e')$ is the indication function such that $\delta_e(e') = 1$ if $e = e'$, and $\delta_e(e') = 0$ otherwise.

2.3 Likelihood function

We calculate the likelihoods based on the difference between observed intensities and the intensity distribution model $\mathcal{I}$ where each point intensity is expressed as a normal distribution:

$$L(h,e|I) = \prod_{i=1}^{N_p} \frac{1}{\sqrt{2\pi}\sigma_i(e)} \exp\left[-\frac{1}{2}\rho\left(d(I_i(h), \mathcal{I}_i(e))\right)\right], \tag{2}$$

where $I_i(h)$ denotes the intensity in image I at the position of the i-th interest point under head pose h, $q_i(h)$, and $\mathcal{I}_i(e)$ is the intensity distribution model for the i-th interest point in facial expression e. Image coordinate $q_i(h)$ is calculated by a weak-perspective projection of the three-dimensional coordinates of the i-th interest point on shape model coordinate system, x_i. The coordinate x_i is obtained by orthogonal projection of image coordinate p_i onto shape model $\mathcal{M}$.

The distance $d(\cdot,\cdot)$ is defined as follows:

$$d(I_i, \mathcal{I}_i(e)) = \begin{cases} \dfrac{\gamma_i I_i - \mu_i(e)}{\sigma_i(e)}, & \text{if visible} \\ d_o, & \text{otherwise (occluded)} \end{cases} \tag{3}$$

where γ_i denotes an intensity adjustment factor (described in Section 3.2). If the interest point whose surface normal obtained from the shape model is not pointing toward the camera, it is considered to be occluded and is given a constant distance d_o.

The function $\rho(\cdot)$ in Eq. (2) denotes a robust function. Unlike the discontinuous function in [8], we use a continuously differentiable function, the Geman-McClure function, to calculate its derivatives (see Section 3.1): $\rho(\xi) = c \cdot \xi^2/(1+\xi^2)$, where c is a scaling factor. This makes the estimation more robust against noise such as imaging noise and large position shifts due to shape model error.

3 Combining stochastic and deterministic search

We introduce two kinds of deterministic search based on maximum likelihood estimation into the particle-filter-based stochastic search. The targets of the deterministic search are to efficiently enhance the estimation performance, and to adjust face intensity to offset the effects of external factors such as illumination changes.

3.1 Estimation improvement

To efficiently improve the estimation, we utilize the approximate estimators robustly obtained by the particle filter as a seed for the deterministic search. Moreover, we apply the deterministic search only for head pose estimation. That is, we immediately adopt the facial expression recognized by the particle filter, $\widetilde{e}_t$, as the final expression estimator. This is because, with the variable-intensity template, the facial expression recognition is often easier than accurately aligning head pose, due to the definition of interest points in the vicinity of facial parts. Simply enhancing the facial pose estimation will also improve the facial expression recognition by reducing the wasteful particles.

Based on Eq. (2), MLE for head pose is simplified as follows:

$$\widehat{h}_t = \arg\max_h L(h|\widetilde{e}_t, I_t) \tag{4}$$

$$= \arg\min_h \sum_i \rho_{i,t} \tag{5}$$

where ρ_i denotes $\rho(d(I_i(h), \mathcal{I}_i(e)))$. To solve this equation, we utilize a gradient method, the quasi-Newton method in this paper. We start the gradient method from two initial guesses, or seeds: One is the current estimator by the particle filter, $\widetilde{h}_t$. The other is the expectation given only the adjacent (just prior) MLE. With our random walk model (Section 2.2), the expectation equals the estimator of immediate prior, $\widehat{h}_{t-1}$. These two seeds are effective for quick motions and slow motions, respectively.

These seeds are individually updated as

$$\widehat{h}^{(m)} = \widehat{h}^{(m-1)} - \alpha \cdot \nabla \sum_i \rho_i \tag{6}$$

where m is the iteration step, and $\alpha(>0)$ is the learning factor. The j-th component of the gradient vector $\nabla \sum_i \rho_i$ is transformed into $\partial / \partial h_j (\sum_i \rho_i) = \sum_i \partial \rho_i / \partial h_j$. According to Eq. (3),

$$\frac{\partial \rho_i}{\partial h_j} = \frac{\partial \rho_i}{\partial d_i} \frac{\partial d_i}{\partial I_i} \left(\frac{\partial I_i}{\partial X_i} \frac{\partial X_i}{\partial h_j} + \frac{\partial I_i}{\partial Y_i} \frac{\partial Y_i}{\partial h_j} \right) \tag{7}$$

where $\partial I_i / \partial X_i$ and $\partial I_i / \partial Y_i$ are image gradients of the image I_t at image coordinates $q_i(h) = [X_i \; Y_i]^{\mathrm{T}}$. Finally, in the two updated estimators, the more likely one is selected as the final head pose estimator.

3.2 Intensity adjustment

We adjust the intensity of interest points observed in the input image to handle changes in intensity itself, e.g. illumination changes or vertical head rotation. Assuming that the

rate of change in interest point intensity is uniform in small facial sub-blocks, we define the intensity adjustment factor as MLE in Eq. (2), given facial pose h and expression e:

$$\widehat{\gamma}_b = \arg\min_{\gamma_b} \sum_{i \in \mathcal{P}_b} \rho_i \tag{8}$$

where γ_b and $\widehat{\gamma}_b$ represent the intensity adjustment factor for facial sub-block b and its MLE, respectively, and $\mathcal{P}_b (\subset \mathcal{P})$ denotes the set of interest points belonging to sub-block b. In practice, we divide the face into four facial blocks, left eyebrow and eye / right eyebrow and eye / left parts of nose and mouth / right parts of nose and mouth. This robust regression problem can be efficiently solved by using an iteratively reweighted least squares algorithm [2]. Although this uniform intensity change assumption is not strictly valid, the small adjustment error does not severely disturb facial expression recognition. The reason is that the interest points defined in the vicinity of facial parts yield significant differences in intensity between facial expressions.

The intensity adjustment factor is calculated with each change in head pose h and/or expression e, that is, the generation of new particles (Section 2.2), or the update of head pose estimator with the gradient method in Eq. (7). Because the intensity is adjusted frame by frame, rapid changes in illumination can be handled.

4 Experimental Results

To evaluate the robustness of our method against out-of-plane head rotations, we performed two types of tests on video sequences: In Test 1, subjects exhibited multiple facial expressions with the head fixed in one of horizontal or vertical directions relative to the camera: horizontally -40, -20, 0, 20 and 40, vertically -20, 0 and 20 (degrees). In Test 2, a subject freely changed horizontally and vertically orientations of the head. The target facial expressions were neutral, angry, sad, surprise and happy, or $N_e = 5$. Nine subjects, seven males and two females in their 20s to 40s, participated in Test 1 with horizontal head rotation once. Four of these males also participated in Test 1 with vertical head rotation once. One of them participated in Test 2 once[4].

Grayscale video sequences with a size of 512×384 pixel were captured at 15 fps for each subject. The number of particles was set to 1,500, and the processing time was about 50 ms/frame on a Core 2 Extreme processor at 3.00GHz with 2.0GB RAM.

4.1 Test with fixed head direction (Test 1)

In Test 1, the subject demonstrated five facial expressions one by one with the head fixed in horizontal or vertical directions relative to the camera for a duration of 60 frames followed by a 60 frame interval, according to instructions (used as ground truths) displayed on a monitor. The recognition rates of facial expression were calculated without using the first 20 frames of each expression just after the instruction was displayed, because of the time lag between the instruction and the exhibition of the facial expression.

Figure 2 shows some successful estimation results of facial poses and expressions in Test 1. Table 1 shows the average facial expression recognition rates of the nine and

[4]Video sequences showing the results in Test 1 and 2 are available from *http://www.hci.iis.u-tokyo.ac.jp/kumano/papers/BMVC2008/*.

Figure 2: Some estimation results of facial poses and expressions in Test 1: The facial expression category in the upper part of each image denotes the recognized one.

Methods	IA	Av	Gr	Total	Frontal	Horizontal ±20	Horizontal ±40	Vertical ±20
Kumano et al. [8]				79.3	95.2	91.6	79.3	51.0
Intensity Adjustment	√			88.3	95.5	94.0	86.6	77.3
Average face shape	√	√		90.4	96.4	94.1	85.2	85.7
Proposed method	√	√	√	91.7	96.3	95.1	87.3	88.3

IA: Intensity adjustment, Av: The use of Average face shape model,
Gr: Head pose estimation with gradient method.

Table 1: Comparison of average recognition rates (%) of facial expressions in Test 1 between the work in [8] and the proposed method.

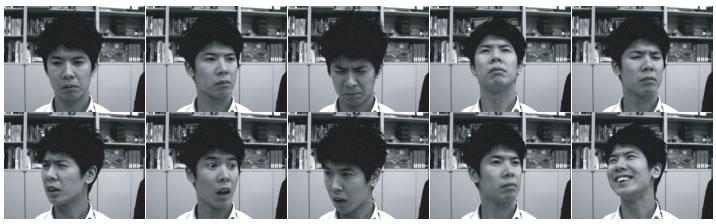
Methods	σ_x [pixel]	σ_y [pixel]	σ_{θ_x} [degree]	σ_{θ_y} [degree]	σ_{θ_z} [degree]	σ_s (no unit)
Kumano et al. [8]	1.6×10^{1}	1.7×10^{1}	9.1	8.3	9.9×10^{-1}	1.4×10^{-2}
Intensity Adjustment	1.1×10^{1}	1.4×10^{1}	7.4	6.3	8.6×10^{-1}	1.5×10^{-2}
Average face shape	9.2	7.5	3.8	4.9	6.2×10^{-1}	1.6×10^{-2}
Proposed method	6.8	3.9	2.0	3.5	4.1×10^{-1}	1.1×10^{-2}

Table 2: S.D. of each head pose component in a frontal-face video sequence from Test 1.

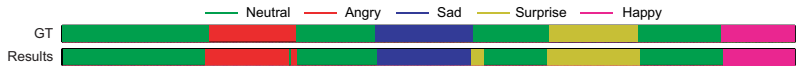
four tests for horizontal and vertical head orientations, respectively. The average rate in the range of ±40 degree yaw angles and ±20 degree pitch angles exceeded 90(%); the recognition rate decreased as the head rotation angle increased. The proposed method outperformed the three other methods. The intensity adjustment and the use of the average shape mainly increased the recognition rate with vertical head rotation, while the gradient method enhanced the performance in almost all head directions. These results suggest that the gradient-based head pose estimation effectively improves the stochastic search. The effect of the deterministic search can also be seen in Table 2, which lists the standard deviations of each head pose component in a frontal-face video sequence from Test 1. In this case, small standard deviations mean that head pose estimation is stable. Our method demonstrates the most stable estimation.

4.2 Test with free head rotations (Test 2)

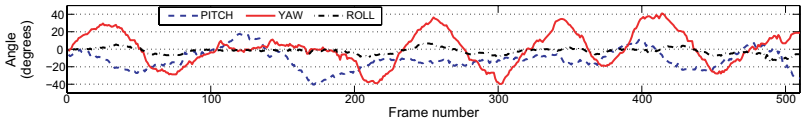
In Test 2, the subject freely demonstrated five facial expressions one by one while rotating the head. Key frames of the video sequence, and the estimated results of facial expres-



(a) Input video sequence (from upper left to lower right, frame number 1, 80, 130, 180, 200, 250, 373, 400, 450, 510).



(b) Ground truth (upper) and recognition results (lower) of facial expression.



(c) Estimation results of facial pose (horizontal axis equals to that of (b)).

Figure 3: Key frames on the test video sequence and its estimation results in Test 2.

sion and head rotation angles in each frame are shown in Fig.3. The ground truth of the facial expression at every frame was labeled by the subject. Figure 3 shows that facial expressions were recognized correctly in almost all frames even though the head orientation varied significantly. We consider the reason for the few mistakes at the end of angry and sad expressions is that we have only the intensity models for the fully exposed facial expressions.

5 Summary and future work

The method proposed in this paper combines stochastic and deterministic search methods for estimating facial pose and expression. The distinct advantage of our method is to efficiently achieve the robust and accurate estimation, despite the use of the simple face model, the variable-intensity template, which can be prepared very easily. In our experiment, five facial expression categories were recognized with overall accuracy of 91.7% for horizontal and vertical facial orientations in the range of ± 40 degrees and ± 20 degrees, respectively, from the frontal view.

A key topic in future research is learning more about what is happening when facial expressions change. First, the most significant challenge is to recognize subtle spontaneous facial expressions. To this end, we would like to apply unsupervised learning with an incremental clustering technique, and to estimate the strength of facial expressions from changes in interest point intensity by referring to optical flow estimation. Second, we would like to introduce the dynamics of facial expressions. Third is a more advanced interest point extraction method. Adding or relocating interest points after observing target expressions is likely to improve the recognition performance.

References

[1] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *J. Multimedia*, 1(6):22–35, 2006.

[2] A. Beaton and J. Tukey. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, 16(2):147–185, 1974.

[3] M. Black and Y. Yacoob. Recognition facial expressions in image sequences using local parameterized models of image motion. *Int. J. Computer Vision*, 25:23–48, 1997.

[4] C. Chang and R. Ansari. Kernel particle filter for visual tracking. *IEEE Signal Process. Lett*, 12:242–245, 2004.

[5] F. Dornaika and F. Davoine. Simultaneous facial action tracking and expression recognition in the presence of head motion. *Int. J. Computer Vision*, 76(3):257–281, 2008.

[6] S. B. Gokturk, C. Tomasi, B. Girod, and J. Bouguet. Model-based face tracking for view-independent facial expression recognition. In *IEEE Int. Conf. Automatic Face and Gesture Recognition*, pages 287–293, 2002.

[7] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int. J. Computer Vision*, 29(1):5–28, 1998.

[8] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. In *Asian Conf. Computer Vision*, volume 1, pages 324–334, 2007.

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

[10] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. Torre, and J. Cohn. AAM derived face representations for robust facial action recognition. In *IEEE Int. Conf. Automatic Face and Gesture Recognition*, pages 155–160, 2006.

[11] K. Oka and Y. Sato. Real-time modeling of face deformation for 3D head pose estimation. In *IEEE Int. W. Analysis and Modeling of Faces and Gestures*, pages 308–320, 2005.

[12] K. Otsuka, H. Sawada, and J. Yamato. Automatic inference of cross-modal nonverbal interactions in multiparty conversations. In *Int. conf. Multimodal Interfaces*, pages 255–262, 2007.

[13] M. Pantic and L. Rothkrantz. Expert system for automatic analysis of facial expression. *Image and Vision Computing*, 18:881–905, 2000.

[14] Y. L. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.

[15] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007.

[16] M. F. Valstar and M. Pantic. Combined Support Vector Machines and Hidden Markov Models for modeling facial action temporal dynamics. In *ICCV-HCI*, pages 118–127, 2007.

[17] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[18] J. Wang, L. Yin, X. Wei, and Y. Sun. 3D facial expression recognition based on primitive surface feature distribution. In *IEEE Computer Vision and Pattern Recognition*, pages 1399–1406, 2006.