Vol.2009-CVIM-167 No.1
2009/6/9

Joint Estimation of Facial Pose and Expression Using Variable-Intensity Templates

SHIRO KUMANO[†1,*1] and YOICHI SATO[†1]

In this paper, we propose a method for pose-invariant facial expression recognition from monocular video sequences. The advantage of our method is that, unlike existing methods, our method uses a simple model, called the variable-intensity template, for describing different facial expressions. This makes it possible to prepare a model for each person with very little time and effort. Variable-intensity templates describe how the intensities of multiple points, defined in the vicinity of facial parts, vary with different facial expressions. By using this model in the combined estimation framework of a particle filter and a gradient method, our method is capable of robustly and effectively estimating facial poses and expressions simultaneously. Experiments demonstrate the effectiveness of our method. A recognition rate of over 90% is achieved for all facial orientations, horizontal, vertical, and in-plane, in the range of ± 40 degrees, ± 20 degrees, and ± 40 degrees from the frontal view, respectively.

1. Introduction

Facial expression recognition is attracting a great deal of attention because of its usefulness in many applications such as human-computer interaction and the analysis of conversation structure[34]. Most existing methods for facial expression recognition assume that the person in the video sequence does not rotate the head strongly and that the image shows a nearly frontal view of the face[1),5),12),17),21),22),26),40)]. However, in situations such as multi-party conversations (e.g. meetings), people will often turn their faces to look at other participants. Hence, unless a stressful head-mounted camera is used, e.g. 36), we must simultaneously handle the variations in head pose as well as facial expression changes.

The major approach to correctly recognizing facial expressions in a face image containing head movements is to prepare a three-dimensional facial shape model of the user's neutral expression (rigid model) and its deformation model for other facial expressions (non-rigid model). The shape model and the facial expression model are together referred to as the face model in this paper. In this approach, facial pose variations are described by globally translating and rotating the shape model in three-dimensional space, and facial expression changes are described by locally deforming the shape model according to the facial expression model. This is often called the geometry-based approach.

Existing methods require an accurate face model, because image variations caused by facial expression change are often smaller than those caused by facial pose change. Accordingly, the use of inaccurate face models degrades the accuracy of both the facial pose and the expression estimates because those two components cannot be differentiated reliably. One method generates a face model for each user, a person-specific model, by using stereo cameras[14)]. Accordingly, this approach cannot be applied to monocular video sequences. Other methods utilize a general face model which can be applied to arbitrary users[7),50)]. However, generating an accurate general deformable model is not so easy in practice. It has been reported that general models cannot cover the large interpersonal variations of face shape and facial expression expected with sufficient accuracy[15)].

Another approach, the region-based approach, has been reported to be relatively robust for small out-of-plane head rotations. In this approach, a two-dimensional (plane)[3),42)] or three-dimensional rigid shape model, e.g. a cylinder in 27), is prepared. Facial expressions are recognized using regional features such as optical flow[3),27)], and Gabor-wavelet coefficients[42)]. They first estimate the head pose, and then calculate regional features for the aligned face. However, the use of rough shape models such as plane or cylinder degrades the accuracy of the facial expression estimates, because large out-of-plane head rotations often cause image appearance variations that are more significant than those yielded by facial expression changes. In addition, the optical flow estimation is easily violated by illumination changes and non-rigid motion. The failure of optical flow estimation directly disturbs facial expression recognition.

Motivated by these problems, we propose a novel *point intensity-based* approach for facial expression recognition. Specifically, we propose *variable-intensity templates* for the following reasons:

(1) Monocular video capture systems are supported.
(2) A face model for each person can be easily prepared.

†1 The University of Tokyo
*1 Presently with NTT Communication Science Laboratories

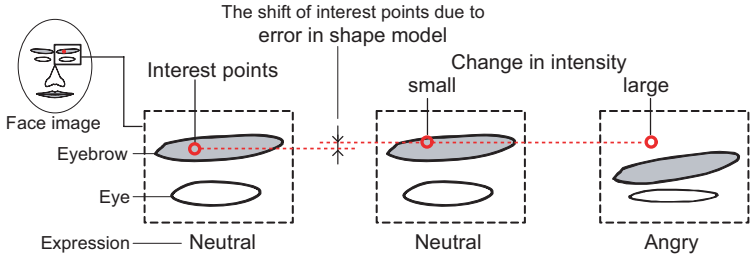


Fig. 1 Our method absorbs errors in shape models and recognizes facial expressions by treating the changes in intensity of multiple points defined around facial parts.

(3) Facial expressions can be estimated even with a large change in facial pose.
Our idea is to describe the change in facial expression not as the deformation of a shape model or as optical flow in facial part regions but as just a change in the intensities of multiple points fixed on a rigid shape model. The variable-intensity template makes it unnecessary to estimate non-rigid motion caused by facial expressions.

The variable-intensity template consists of three components: a rigid shape model, a set of interest points, and an intensity distribution model. As the shape model, we use an average face shape generated from the shapes of many people. The interest points are sparsely defined away from the edges of facial parts such as eye corners, in a frontal and neutral expression face image. The intensity distribution model is the facial expression model that describes how interest point intensity varies for different facial expressions. The set of interest points and the intensity distribution model are person-specific.

The mechanism and effect of the use of the variable-intensity template are as follows. The intensities of the interest points vary significantly due to the movements of the facial parts created when showing facial expressions (see **Fig. 1**). Focusing on this characteristic, we recognize facial expressions by matching the model intensities prepared in advance against the observed intensities. We want a system that is robust against head pose variations even if the shape model used produces some approximation error. To achieve this, we must handle the problem that the interest points will be shifted from their actual positions due to the shape model error as the out-of-plane head rotation angles increase. Hence, we site the interest points away from the edges of facial parts to keep the change in intensity small when the face is rotated as shown in Fig. 1.

Our contribution[⋆1] is that we propose a facial expression recognition method for varying facial poses based on the key idea that facial expressions can be correctly recognized by knowing how the interest point intensities vary with the facial expressions. The main advantage of our method is that a face model for each person is created simply by capturing frontal face images of the person making the target facial expressions. Thus it is more efficient that existing methods that need measurement instruments[14], time-consuming manual intervention processes[7], large amounts of training data[42] and so on.

Furthermore, we implement a novel combined search algorithm of a particle filter and a gradient method to simultaneously estimate facial poses and expressions, where the variable-intensity template is utilized as a face model. Many reearches[7),14),25)] showed that the particle filter can realize a robust estimation. However, the stochastic search generally requires a large number of hypotheses to accurately estimate the target states, or to reach a peak of probability density function (pdf) in this work. This shortcoming can be overcome by combining the deterministic approach of a gradient method with the particle filters. This makes the estimation robust and effective.

The remainder of this paper is organized as follows. First, 3 describes our proposed method. Then, in 4, experimental results are given. Finally, a summary and future work are given in 5.

2. Related work

While most existing facial expression recognition systems assume that the person in the video sequence does not rotate the head significantly and that the image shows a nearly frontal view of the face[1),5),12),17),22),26),40)], some methods consider head pose variations. Facial pose and expression vary independently, and, image variations caused by facial expression change are often smaller than those caused by head movement.

Hence, the precise separation of these two components is needed to achieve correct recognition of facial expressions. Unfortunately, the highly non-linear effect on face appearance in the image due to large out-of-plane head rotations is not easy to handle. In addition, only a few video databases containing non-frontal faces have been published, e.g. Face Video Database[⋆2], so it is difficult to compare the recognition ability

⋆1 A part of this work appeared in International Journal of Computer Vision (IJCV)[25], British Machine Vision Conference (BMVC)[24], and Asian Conference on Computer Vision (ACCV)[23].
⋆2 Face Video Database of the Max Planck Institute for Biological Cybernetics in Tuebingen:

of existing methods quantitatively. Given these issues, pose-invariant facial expression recognition can be said to be an underdeveloped research area.

In this section, we briefly overview the facial expression recognition methods that allow some out-of-plane head rotations with regard to their face shape models. Excellent reviews of a number of recent studies assuming near-frontal faces can be found in 11), 35), 37), 41).

In what follows, we divide the existing methods into the following five groups with regards to the shape model used: (a) face shapes directly measured, (b) face shapes recovered from images, (c) general face shapes, (d) parametric shapes, and (e) shape-free approach. These shape models, except for group (b) and (e), are detailed below in descending order of their approximation accuracy. Group (a-d) are divided into two classes by the following three kinds of classifications: measured models (a) versus non-measured models (b-d), person-specific models (a,b) versus person-independent models (c,d), and deformable models (a-c) versus rigid models (d).

(a) *Face shapes directly measured*
The most conclusive way to make the estimation robust against head rotations is to directly measure the actual three-dimensional shape of each user with an instrument other than monocular camera. Gokturk et al.[14] use a stereo cameras to accurately measure the three-dimensional positions of 19 facial feature points defined on the user's face and their deformations due to change in facial expression. In the test stage, they classify facial expressions using deformation parameters estimated by using an optical-flow-like technique. Their experimental results suggest that their method can recognize facial expressions in a variety of head poses. In the field of head tracking, there is a similar approach to handle facial expression variations, e.g. 33). However, these methods cannot be applied to monocular video sequences, because they require stereo cameras to measure the three-dimensional positions of feature points in the initial frame of test video sequences. Dispensing with images, the systems of 39), 44) correctly recognize facial expressions regardless of head pose from face range data acquired by a three-dimensional digitizer.

http://vdb.kyb.tuebingen.mpg.de/.

(b) *Face shapes recovered from images*
Some methods try to recover the actual three-dimensional shape of the user from a monocular video sequence by using a structure-from-motion technique. Xiao et al.[45] propose 2D+3D Active Appearance Models (AAM) where first the positions of the set of feature points in the video sequence are tracked by using a common two-dimensional AAM and then the three-dimensional face shape is recovered by using a non-rigid structure-from-motion technique. Lucey et al.[30] apply 2D+3D AAM to recognize facial expressions. However, in their experiments, the recovered three-dimensional shapes were inaccurate which rather degraded the recognition rates.

(c) *General face shapes*
Some methods utilize models that can be applied to arbitrary users, i.e. person-independent models. Cohen et al.[6] utilize the Piecewise Bezier Volume Deformation (PBVD) tracker which tracks feature points defined on a generic three-dimensional wire-frame model of the face. They classify facial expressions using its motion parameters as the temporal cues. Dornaika et al.[7] describe a system that uses a existing deformable face mesh model; its deformation parameters are related to action units. They estimate facial expressions in the framework of a particle filter[18], after estimating the head pose by a gradient-based method. However, generating an accurate general deformable model is not so easy in practice. Gross[15] suggest that such person-independent models cannot cover the large interpersonal variations in face shape and facial expression expected with sufficient accuracy.

(d) *Parametric shapes*
Some methods utilize simple rigid shapes rather than complicated deformable shapes. Black et al.[3], Liao et al.[27], and Tong et al.[42] approximate the human face as plane (the first two) and cylinder (the last), respectively. As the feature, Black et al.[3] and Liao et al.[27] utilize the optical flow in the face region, while Tong et al.[42] use Gabor wavelet coefficients. However, using the simple parametric shape causes large error in the alignment of the face. Hence, facial expressions at large head rotation angles tend to be misrecognized in the absence of a mechanism for handling the face misalignment. Tong et al.[42] introduced Dynamic Bayesian Network (DBN) inference, where the relationships between action units are modeled. Their experiments suggest that this framework

can compensate some misclassification of action units. However, it is reasonable to expect that it cannot handle heavy face misalignments caused by large out-of-plane head rotations. Furthermore, note that optical flow estimation is easily corrupted by illumination changes and non-rigid motion.

(e) *Shape-free approach*
Hu and his colleagues developed a facial expression recognizer that does not utilize any face shape model[16]. By utilizing Support Vector Machines (SVMs), they discriminated horizontal head orientations (discretized at 15 degree or larger angle interval) and facial expressions sequentially or simultaneously. Although their system can handle faces in the range of frontal- to profile-view, the discretization error for head pose in real situations, where the head pose is continuous, is expected to degrade the facial expression recognition. Note that preparing a large number of classifiers (each of which covers a small angle range) is undesirable in terms of processing time and memory resources. Furthermore, they assume that the face regions in input data are precisely extracted. This, however, remains difficult for current computer vision technologies.

Overall, to the best of our knowledge, there is no work that matches all of our goals:

- Monocular video capture systems are supported.
  The approach that directly measures face shapes (a) does not have this advantage.
- A face model for each person can be easily prepared.
  The general model-based approach (c) and shape-free approach (e) fail to achieve this advantage because of model complexity.
- Facial expressions can be estimated even with a large change in facial pose.
  The recovered face shape model-based approach (b) and the parametric model-based approach (d) have not achieved this end.

The method proposed herein is a novel *point intensity-based* approach that can realize all of the above goals. We describe a change in facial expression not as the deformation of the shape model or as optical flow in facial part regions, but just as a change in the intensities of local face regions. In addition, none of the existing papers, other than 16), 44), describe how the recognition rate of facial expressions varies with out-of-plane head rotation angle. Hence, pose-invariability of their methods cannot be quantified while the problem of the difference in test data remains. In this paper, we also describe

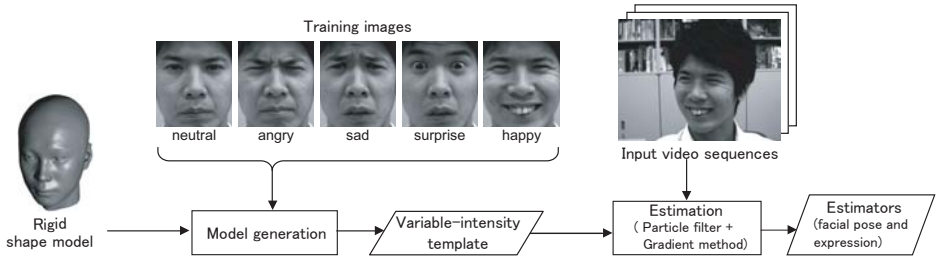


Fig. 2 System flow chart.

experiments that assess the impact of head pose on the recognition rate.

3. Proposed method

Our method consists of two stages as shown in **Fig. 2**. First, a variable-intensity template for each person is generated from just one frontal face image for each facial expression (referred to as the training images in this paper). Second, facial pose and expression are simultaneously estimated in a combined search framework of a particle filter and a gradient method based on their likelihood for the template.

3.1 Variable-intensity template

The variable-intensity template $\mathcal{M}$ consists of the following three components:

$$\mathcal{M} = \{\mathcal{S}, \mathcal{P}, \mathcal{N}\} \quad (1)$$

where $\mathcal{S}$, $\mathcal{P}$, and $\mathcal{N}$ denote a rigid face shape model, a set of interest points, and an intensity distribution model, respectively. The intensity distribution model describes the intensity distribution of each interest point for different facial expressions.

The variable-intensity template is generated from a set of training images and the rigid face shape model. The training image set consists of one face image for each facial expression, $\{\boldsymbol{g}_e\}_{e=1}^{N_e}$, as shown in Fig. 2, where $\boldsymbol{g}_e$ denotes the image labeled with facial expression $e \in \{1, \cdots, N_e\}$, and N_e is the number of target expressions. In particular, the training image for neutral expression is denoted by $\boldsymbol{g}_{\mathrm{NEU}}$. The face in each image $\boldsymbol{g}$ is frontal, and virtually fixed. The way to generate such training images is described in 3.6.

As the face shape model $\mathcal{S}$, we use the average face shape model[⋆1] shown in Fig. 2 in

⋆1 *Average head dummy of Japanese young males* contains over 100,000 polygons without texture.

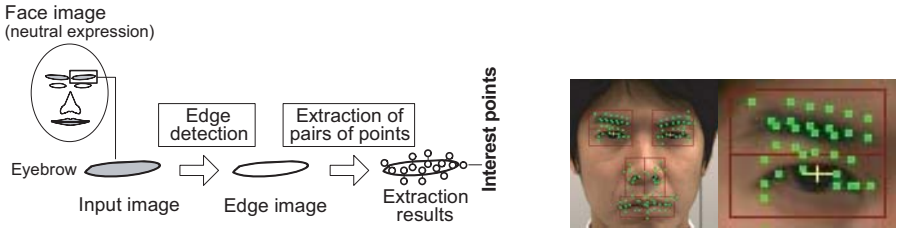


Fig. 3 Left: The method used to extract interest points $\mathcal{P}$. Right: Example of a set of paired interest points $\mathcal{P}$, shown as 128 green dots. The large rectangles represent the boundaries of facial parts. The right side shows the expanded right eyebrow and eye regions.

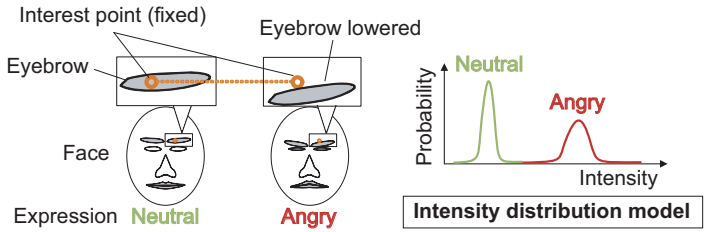


Fig. 4 Intensity distribution model $\mathcal{N}$: The intensity distributions of interest points, described as normal distributions, indicate the change in facial expressions. The colors in the right part correspond to the interest points in the left part.

this paper. The approach used to fit it to each user is described in 3.6.2.

3.1.1 Set of interest points $\mathcal{P}$

The multiple interest points are sparsely defined away from the edges of facial parts such as eye corners in the training image in neutral expression $\boldsymbol{g}_{\text{NEU}}$, as shown in the right part of **Fig. 3**. All interest points are fixed on the rigid shape model $\mathcal{S}$. The set of interest points is described as $\mathcal{P} = \{\boldsymbol{p}_i\}_{i=1}^{N_p}$, where $\boldsymbol{p}_i$ denotes the image coordinates of the i-th interest point in the training image $\boldsymbol{g}_{\text{NEU}}$, and N_p denotes the number of interest points. The set of interest points $\mathcal{P}$ is different for each subject.

An interest point constitutes a pair of points (herein called *paired interest points*)[31] that straddle and are centered on the edge, to detect the bidirectional motions of the facial parts. The paired interest points are sparsely selected in the four facial part regions (eyebrows, eyes, nose, and mouth) in the training image, $\boldsymbol{g}_{\text{NEU}}$. These facial part regions are detected as rectangular boundaries by using a cascaded AdaBoost detector based on Haar-like features[43]. Examples of these pairs are shown in the left part of Fig. 3: The total number of interest points, N_p, for each subject is set to be 128 (eyebrows: 20×2, eyes: 12×2, nose: 24, and mouth: 40). This number was decided after conducting preliminary evaluations of the performance metrics of accuracy and processing speed. The details of the interest point selection are described in 25).

3.1.2 Intensity distribution model $\mathcal{N}$

The intensity distribution model describes how the interest point intensity varies for different facial expressions. As shown in **Fig. 4**, the interest point intensity changes strongly due to the shift of its associated facial part. Focusing on this property, we recognize facial expressions from the changes in observed interest point intensities.

The observed intensity of each interest point is varied by small position shifts of the interest point. This is caused by error in the shape model as well as the errors in the intensity adjustment, described in 3.3. Such variation is represented as a normal distribution in our method. The intensity distribution model $\mathcal{N}$ is described as: $\mathcal{N} = \{\mathcal{N}_i(e)\}_{i=1,e=1}^{N_p,N_e}$, where $\mathcal{N}_i(e)$ denotes a normal distribution of the intensity of the i-th interest point for expression e. We set the mean of each normal distribution, $\mu_i(e)$, to be the value recorded from the training image labeled with expression e, $\boldsymbol{g}_e$, at image coordinate $\boldsymbol{p}_i$. Furthermore, we assume that standard deviation is proportional to its mean, or $\sigma_i(e) = k\mu_i(e)$, and set the proportionality factor k to be 0.3.

3.2 Likelihood of facial poses and expressions

In our method, the facial pose and expression are simultaneously estimated based on their likelihood given the observed intensity of interest points with the intensity distribution model. Their maximum likelihood estimators (MLEs) are calculated by a combined search algorithm of a particle filter (3.4) and a gradient method (3.5).

The likelihood is defined based on the difference between observed intensities and the intensity distribution model $\mathcal{N}$ where each point intensity is expressed as a normal distribution:

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

It is published by Digital Human Research Center, Advanced Industrial Science and Technology, *http://www.dh.aist.go.jp/research/centered/facedummy/*.

(1) Initialize $N_b \times N_b$ weight matrix $\boldsymbol{W} = \mathrm{diag}\,[1, \cdots, 1]^{\mathrm{T}}$, where N_b denotes the number of elements in set $\mathcal{P}_b$.
(2) Estimate current intensity adjustment factor $\gamma^{(m)}$ using the previous weight matrix $\boldsymbol{W}^{(m-1)}$:

$$\gamma^{(m)} = (\boldsymbol{I}_b^{\mathrm{T}} \boldsymbol{W}^{(m-1)} \boldsymbol{I}_b)^{-1} \boldsymbol{I}_b^{\mathrm{T}} \boldsymbol{W}^{(m-1)} \boldsymbol{\mu}_b \quad (4)$$

where m denotes the iteration step number, and $\boldsymbol{I}_b$ and $\boldsymbol{\mu}_b$ denote $N_b \times 1$ vectors of observed intensity and mean of intensity distribution model divided by standard deviation, $\boldsymbol{I}_b = [\cdots, I_i/\sigma_i, \cdots]^{\mathrm{T}}$ and $\boldsymbol{\mu}_b = [\cdots, \mu_i/\sigma_i, \cdots]^{\mathrm{T}}, i \in \mathcal{P}_b$.
(3) Update weight matrix $\boldsymbol{W}$:

$$W_{ii}^{(m)} = w\left(\frac{\gamma^{(m)} I_i - \mu_i}{\sigma_i}\right) \quad (5)$$

where W_{ii} denotes the i-th diagonal element of weight matrix $\boldsymbol{W}$, and $w(\cdot)$ denotes the weight function related to the robust function $\rho(\cdot)$ as $w(x) = (d\rho(x)\,/dx)/x$.
(4) Steps 2 and 3 are repeated until convergence.

Fig. 5 Intensity adjustment based on the iteratively reweighted least squares technique.

where $p(\boldsymbol{I}|\boldsymbol{h}, e)$ denotes the joint likelihood of head pose $\boldsymbol{h}$ and facial expression e for image $\boldsymbol{I}$. Observed intensity $I_i(\boldsymbol{h})$ denotes the intensity in image $\boldsymbol{I}$ at the position of the i-th interest point under head pose $\boldsymbol{h}$, $\boldsymbol{q}_i(\boldsymbol{h})$. The image coordinate $\boldsymbol{q}_i(\boldsymbol{h})$ is calculated by a weak-perspective projection of the three-dimensional coordinates of the i-th interest point on shape model coordinate system, $\boldsymbol{x}_i$. The coordinate $\boldsymbol{x}_i$ is obtained by orthogonal projection of image coordinate $\boldsymbol{p}_i$ onto shape model $\mathcal{S}$.

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

$$d(I_i, \mathcal{N}_i(e)) = \frac{\gamma_i\, I_i - \mu_i(e)}{\sigma_i(e)} \quad (3)$$

where γ_i denotes an intensity adjustment factor (described in 3.3). 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 the Geman-McClure robust function[13]: $\rho(\xi) = c \cdot \xi^2/(1 + \xi^2)$, where c is a scaling factor. This makes the estimation more reliable against noise such as imaging noise and large position shifts due to shape model error.

3.3 Intensity adjustment

We adjust the intensity of interest points observed in the input image, I_i, 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. 3 given facial pose $\boldsymbol{h}$ and expression e:

$$\widehat{\gamma}_b = \arg\min_{\gamma_b} \sum_{i \in \mathcal{P}_b} \rho_i \quad (6)$$

where ρ_i denotes $\rho(d(I_i(\boldsymbol{h}), \mathcal{N}_i(e)))$. Variable 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 (IRLS) 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.

As shown in **Fig. 7**, the intensity adjustment is performed with each change in head pose $\boldsymbol{h}$ and/or expression e, that is, the generation of new particles in the particle filter (3.4), or the update of head pose estimator with the gradient method in Eq. 14. This frame-by-frame adjustment makes it possible to handle rapid changes in illumination.

3.4 Estimation by using a particle filter

The estimation of the proposed method consists of two steps. First, the joint posterior probability density function (pdf) of facial pose and expression given input face images is calculated by using a particle filter[18]. Next, the facial pose estimator is improved by a gradient method as described in 3.5.

The joint posterior pdf of facial pose and expression at time t given all face images up to that time, $\boldsymbol{I}_{1:t}$, is recursively represented as follows:

$$p(\boldsymbol{h}_t, e_t|\boldsymbol{I}_{1:t}) = \alpha p(\boldsymbol{I}_t|\boldsymbol{h}_t, e_t) \int p(\boldsymbol{h}_t|\boldsymbol{h}_{t-1}) \sum_{e_{t-1}} P(e_t|e_{t-1}) p(\boldsymbol{h}_{t-1}, e_{t-1}|\boldsymbol{I}_{1:t-1}) d\boldsymbol{h}_{t-1} \quad (7)$$

where the facial pose $\boldsymbol{h}_t$ and expression e_t are assumed to be conditionally dependent given the image $\boldsymbol{I}_t$, following individual first order Markov processes as shown in **Fig. 6**.

The head pose $\boldsymbol{h}_t$ consists of the following six continuous variables: the coordinate of

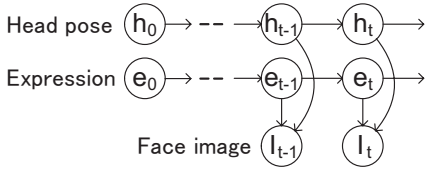


Fig. 6 Dynamic Bayesian Network for head poses, facial expressions, and face images.

the center of the template on the image plane, three-dimensional rotation angles (yaw, pitch, and roll), and scale. For the head motion model, $p(\boldsymbol{h}_t|\boldsymbol{h}_{t-1})$, we utilize a random walk model where the components of head pose are considered to be independent each other. The system noise for each component is created by a zero-mean Gaussian process. The variances of the system noises were empirically decided. With regards to facial expression, we set $P(e_t|e_{t-1})$ to be equal for all expression combinations in this paper. Any expression transition matrix can be utilized in our framework, but it is difficult to obtain the actual transition. Accordingly, in this paper, the simple distribution models are assumed.

The posterior pdf in Eq. 7, unfortunately, cannot be calculated exactly, because parameters of facial pose $\boldsymbol{h}_t$ are continuous, and their distributions are complex due to the non-linearity of the projection function, occlusion judgment etc. Hence, we use a particle filter[18], which approximates the pdf with a set of weighted hypotheses, or particles, stochastically generated. This makes the estimation easy to avoid local maxima and to recover from temporary disturbance. Each particle expresses a state and its weight. In our case, the state in the search space and weight of the l-th particle are expressed as $[\boldsymbol{h}_t^{(l)}, e_t^{(l)}]$ and $\omega_t^{(l)}$, respectively. The weight $\omega_t^{(l)}$ is proportional to the likelihood:

$$\omega_t^{(l)} \propto p(\boldsymbol{I}_t|\boldsymbol{h}_t^{(l)}, e_t^{(l)}) \tag{8}$$

where $\sum_l \omega_t^{(l)} = 1$ is satisfied.

The estimator of head pose by the particle filter, $\widetilde{\boldsymbol{h}}_t$, is defined to be the expectation of its marginal pdf, $p(\boldsymbol{h}_t|\boldsymbol{I}_{1:t})$, while the recognized facial expression $\widetilde{e}_t$ is the expression that maximizes the marginal probability $p(e_t|\boldsymbol{I}_{1:t})$:

$$\widetilde{\boldsymbol{h}}_t = \sum_l \omega_t^{(l)} \boldsymbol{h}_t^{(l)} \tag{9}$$

$$\widetilde{e}_t = \arg\max_e \sum_l \omega_t^{(l)} \delta_e(e_t^{(l)}) \tag{10}$$

where the function $\delta_e(e')$ is the indication function such that $\delta_e(e') = 1$ if $e = e'$, and $\delta_e(e') = 0$ otherwise.

3.5 Estimation improvement by a gradient method

The particle filter can realize a robust estimation by using multiple hypotheses. However, the stochastic search generally requires a large number of hypotheses to accurately estimate the target states. To overcome this shortcoming, the deterministic approach of a gradient method is combined with the stochastic approach of particle filters. This makes the estimation robust and effective.

The deterministic search is applied only for the estimation of the continuous head pose. In contrast, for the discrete facial expression category, we immediately adopt the expression recognized by the particle filter, $\widetilde{e}_t$, as the final 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:

$$\widehat{\boldsymbol{h}}_t = \arg\max_{\boldsymbol{h}} p(\boldsymbol{I}_t|\boldsymbol{h}, \widetilde{e}_t) \tag{11}$$

$$= \arg\min_{\boldsymbol{h}} \sum_i \rho_{i,t}. \tag{12}$$

To solve this equation, we utilize a gradient method.We start the gradient method from two initial guesses, or seeds: One is the current estimator by the particle filter, $\widetilde{\boldsymbol{h}}_t$. The other is the expectation given only the adjacent (just prior) MLE. With our random walk model (3.4), the expectation equals the estimator of immediate prior, $\widehat{\boldsymbol{h}}_{t-1}$. These two seeds are effective for quick motions and slow motions, respectively.

These seeds are individually updated as

$$\widehat{\boldsymbol{h}}^{(m)} = \widehat{\boldsymbol{h}}^{(m-1)} - \alpha \cdot \nabla \sum_i \rho_i \tag{13}$$

where m is the iteration step, and $\alpha(> 0)$ is the learning factor. The j-th component

for t=1:T
(1) Calculate joint posterior pdf, $p(\boldsymbol{h}_t, e_t|\boldsymbol{I}_{1:t})$, by the particle filter $\cdots$ 3.4
- Generate hypotheses[18], $\{\boldsymbol{h}_t^{(l)}, e_t^{(l)}\}$
- Intensity adjustment
$\widehat{\gamma}_b = \arg\min_{\gamma_b} \sum_{i\in\mathcal{P}_b} \rho_i \cdots$ Eq. 6
- Weight (Evaluate) each hypothesis
$w_t^{(l)} \propto p(\boldsymbol{I}_t|\boldsymbol{h}_t^{(l)}, e_t^{(l)}) \cdots$ Eq. 8
- Calculate expectation of the posterior pdf
$\widetilde{\boldsymbol{h}}_t = \sum_l \omega_t^{(l)} \boldsymbol{h}_t^{(l)}$, $\widetilde{e}_t = \arg\max_e \sum_l \omega_t^{(l)} \delta_e(e_t^{(l)}) \cdots$ (Eq. 9, 10)

(2) Search MLE by the gradient method $\cdots$ 3.5
while $|\widehat{\boldsymbol{h}}^{(m)} - \widehat{\boldsymbol{h}}^{(m-1)}| >$ threshold
- Intensity adjustment
$\widehat{\gamma}_b = \arg\min_{\gamma_b} \sum_{i\in\mathcal{P}_b} \rho_i \cdots$ Eq. 6
- Update estimator
$\widehat{\boldsymbol{h}}^{(m)} = \widehat{\boldsymbol{h}}^{(m-1)} - \alpha \cdot \nabla \sum_i f_i\left(\widehat{\boldsymbol{h}}^{(m-1)}, \widetilde{e}, \boldsymbol{I}\right) \cdots$Eq. 13
end
end

Fig. 7 Estimation flow of the proposed method

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,

$$\frac{\partial \rho_i}{\partial h_j} = \frac{\partial \rho_i}{\partial d_i}\frac{\partial d_i}{\partial I_i}\left(\frac{\partial I_i}{\partial X_i}\frac{\partial X_i}{\partial h_j} + \frac{\partial I_i}{\partial Y_i}\frac{\partial Y_i}{\partial h_j}\right) \tag{14}$$

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

3.6 Other processing

3.6.1 Generating training images

We prepare training images, $\boldsymbol{g}$, where the iris centers are laid on a horizontal line. Such training images are generated from source face images (referred to as training source images) that contain a human face in one of the target facial expressions. We assume that the face in the training source images is stationary between neutral expression and non-neutral expressions for the same subject. The position and size of the faces, however, may be different for each person. The training images are the training source images that are rotated in-plane until iris centers fall on a horizontal line. The iris centers are detected in the eye regions by utilizing Fast Radial Symmetry[29),48)]. This rotation is calculated with each training source image only for neutral expression. The non-neutral expression images are rotated by the same rotation matrix for the neutral expression image of the same person.

3.6.2 Shape model fitting

We fitted the average face shape to each user as follows: (I) fit the center of the shape model to the center of face region in the training image in neutral expression, $\boldsymbol{g}_{\mathrm{NEU}}$, detected by the method of 43), (II) stretch the shape model in the horizontal and vertical directions to match both face width and height; stretching in the depth direction used the scaling factor given as the square root of the product of vertical and horizontal scaling factors.

4. Experimental Results

To evaluate the estimation performance of the proposed method, we performed three experiments: The objective of the first test was to evaluate the accuracy of our system in recognizing facial expressions when the face does not face the camera. The second test evaluates the performance of our system with person-independent strategy. The third test evaluates the performance of the system in head pose tracking.

At the same time, to evaluate the effectiveness of our interest points (paired interest points described in 3.1.1), we prepared two other kinds of variable-intensity templates: those with randomly allocated points and those with points defined on the edges (these points are hereinafter referred to as random points and edge points, respectively). These points are also defined in each of the four facial part regions as shown in **Fig. 8**.

In the following experiments, the recognition rates were calculated as the ratio between the number of frames wherein the estimated expression matched the ground truth to the total number of target frames. In this paper, we ran our system offline to evaluate the recognition rates of facial expression. 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. Currently, the identity of a person is given a priori.

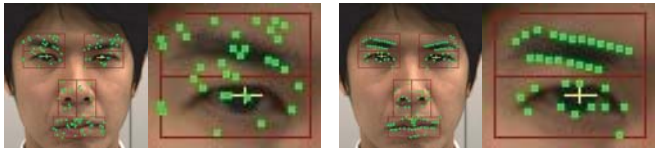


Fig. 8 Example of two kinds of set of interest points: randomly allocated (left) and defined on edges (right). These points are used for the comparison with our interest points shown in Fig. 3.

4.1 Performance evaluation for Non-frontal-view faces

Unless the true face shape model is used, the interest points are misaligned due to error in the shape model. Such shift increases as the head is more rotated in out-of-plane. Accordingly, it is important to evaluate the robustness of the system against head rotation. Unfortunately, to the best of our knowledge, there is no facial expression database that contains video sequences where the face moves together with various changes in facial expression. Hence, we generated such video datasets and used them to evaluate our system.

4.1.1 Our original facial expression datasets

Our datasets contain the following two kinds of video sequences: *fixed-pose dataset* and *free-pose dataset*. Both datasets contain facial expression categories of neutral, angry, sad, surprise and happy. Both sets include training images and test video sequences for each subject. The training images were captured immediately prior to the capture of the test video sequences. That is, each couple of training images and test sequences was captured in the same situation on the same day. These video sequences were captured by the same IEEE 1394 XGA (1024×768 pixel) color camera at 15 fps. In this paper, we utilized these video sequences after converting them into grayscale 512×384 pixel sequences.

In the fixed-pose dataset, subjects exhibited multiple facial expressions with the head fixed in one of three directions relative to the camera: $0/\pm 20/\pm 40$ degree in yaw (horizontal), $0/\pm 20$ degree in pitch (vertical), or $0/\pm 20/\pm 40$ degree in roll (in-plane) (see **Fig. 9**-10). Yaw, pitch, and roll sequences contain nine subjects (seven males and two females in their 20s to 40s), four subjects (four males of the nine subjects), and one subject (one male of the four males), respectively. For each direction, each subject provided one sequence.

All sequences start with neutral expression face looking toward the camera, then move to one direction relative to the camera. Finally, the subjects create five facial expressions one by one without moving their head, each for a duration of 60 frames followed by a 60 frame interval with relaxed face as indicated by the instructions displayed on a monitor, which are used as truth of the facial expression.

The free-pose dataset, on the other hand, contains one sequence of the subject, who is also contained in the fixed-pose dataset. The subject freely shows five facial expressions one by one while shaking the head left and right.

To capture the training images, we gave the subjects the following instructions: (I) face the head directly at the camera during the capture process, (II) show facial expressions one by one as requested by the instructions with category name on the monitor, (III) press a key on a keyboard, when the target expression was being expressed, to have a facial image captured as the training image.

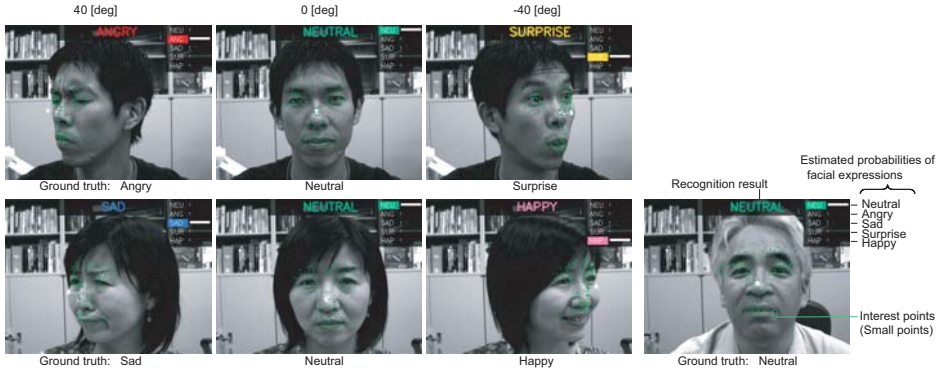
4.1.2 Evaluation with fixed-pose dataset

First, by using the fixed-pose dataset, we evaluate how accurately our system can recognize facial expressions for various head orientations[⋆1]. In calculating the recognition rates, we excluded the first 20 frames of each expression just after the instruction was displayed, because of the time lag between the instruction and the creation of the facial expression.

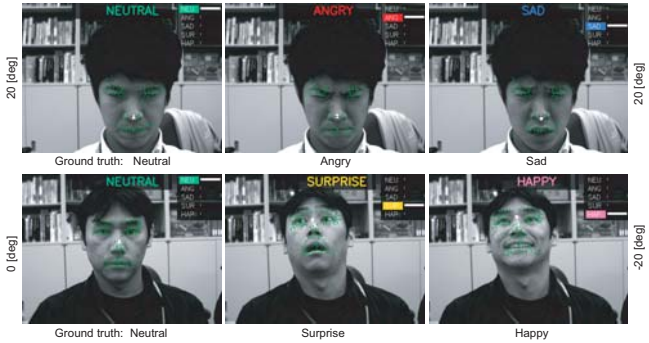
Figure 9 shows some estimation results of facial poses and expressions by using our paired interest points for horizontal, vertical and in-plane head rotations, respectively. They show the facial poses and expressions were correctly estimated for all subjects.

Table 1 shows a comparison of our paired interest points and random points with the average facial expression recognition rates for each target head rotation angle. For out-of-plane (yaw and pitch) rotations, the recognition rate decreased as the head rotation angle increased. Nevertheless, our paired interest points are more robust against the out-of-plane rotations, than the random points. For in-plane direction, our method recognized all facial expressions perfectly, because the misalignment of interest points due to error in the shape model does not occur with in-plane rotation. Accordingly, we evaluated it just for one subject. Our paired interest points achieved at least 90(%) in average for all rotations, yaw, pitch and roll.

⋆1 The results in 4.1.2 and 4.2 were obtained by the proposed method without the gradient method described in 3.5.



(a) Horizontal head rotations.



(b) Vertical head rotations.

Fig. 9 Some estimation results of facial poses and expressions with fixed-pose dataset (Horizontal): The expression in the upper part of each image denotes the recognized result. The width of each bar in the upper right part of each image denotes the expectation of posterior pdf for each facial expression, $P(e_t|\boldsymbol{I}_{1:t})$. Small points on each face denote interest points.

Table 2 shows the confusion matrix for the average recognition rates with the paired interest points. These results suggest that, for many subjects, the difference in face appearance between sad expression and both neutral and angry expressions is smaller than for other expression combinations, accordingly, they were sometimes confused with each other.

On the other hand, with the edge interest points, head tracking was lost for many video sequences. The reason for this seems to be that such interest points are too sensitive to misallocation of interest points. The misallocation is caused by errors in the shape model and slight differences in facial expression between the test sequences and training images.

Fig. 10 Some estimation results of facial poses and expressions with the fixed-pose dataset (In-plane rotation): The expression in the upper part of each image denotes the recognized result.

Table 1 Average recognition rates of facial expressions for a variety of head orientations with fixed-pose dataset by comparing two types of interest points: Pair: our paired interest points. Random: points randomly defined in each facial part region.

Point type	Total	Angle [deg]				
		-40	-20	0	20	40
Yaw (Horizontal): nine sequences.						
Pair	92.3	83.3	94.3	95.4	95.9	92.5
Random	90.1	90.2	92.1	95.4	96.1	76.7
Pitch (Vertical): four sequences.						
Pair	94.0	N/A	87.0	97.5	97.6	N/A
Random	86.4	N/A	74.1	98.1	87.0	N/A
Roll (In-plane): one sequence.						
Pair	100.0	100.0	100.0	100.0	100.0	100.0
Random	100.0	100.0	100.0	100.0	100.0	100.0

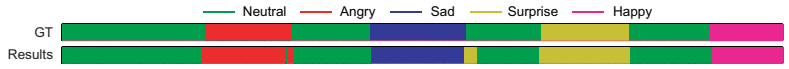
unit is (%)

Table 2 Average confusion matrix of facial expressions with fixed-pose dataset for horizontal and vertical directions: GT and RCG denote ground truth and recognition, respectively. The overall recognition rate is 93.1[%].

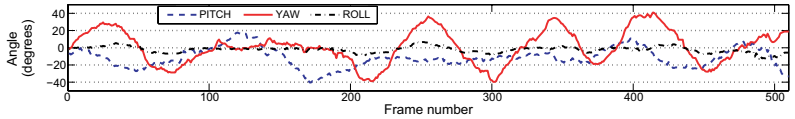
GT \ RCG	Neutral	Angry	Sad	Surprise	Happy
Neutral	88.9	1.5	7.9	1.6	0.1
Angry	0.4	97.6	0.9	1.1	0.0
Sad	3.2	8.0	85.4	3.3	0.1
Surprise	0.0	0.0	4.2	95.8	0.0
Happy	0.2	1.1	0.0	1.1	97.7

unit is (%)

(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)).

Fig. 11 Key frames on the test video sequence and its estimation results in the free-pose dataset.

4.1.3 Evaluation with free-pose dataset

We next evaluated our system with the free-pose dataset, where the subject changes both the facial expression and the head's horizontal orientation.

Five frames of the video sequence of the free-pose dataset and the estimated results of facial expression and pose in each frame[⋆1] are shown in (a), (b) and (c) of **Fig. 11**, respectively. The ground truth of the facial expression in every frame was hand-labeled by the subject. Figure 11(b) shows that facial expressions were recognized correctly in almost all frames. A quantitative evaluation for head pose is given in 4.3.

4.2 Performance evaluation with person-independent strategy

The main advantage of our system is that our system makes it possible to generate the variable-intensity template for each person with very little time and effort. As demonstrated in 4.1.2, by using our person-specific model, we can achieve the high facial expression recognition performance. Such a person-specific strategy is useful for applications where the number of users is small. However, some applications require that it can be applied to arbitrary users without such training for each user. Accordingly, we evaluate the potential of our system to be extended to the person-independent strategy.

To compare our system (with person-independent strategy) against existing methods, we utilized a widely used public facial expression database, the Cohn-Kanade DFAT-504 database[19]. The Cohn-Kanade database consists of FACS-coded[9] video sequences of 104 subjects, starting from a neutral expression and ending in the peak of the facial expression. For each person, there are on average 8 frames for each expression. The desired facial displays were described and modeled before the material was recorded.

Although there are a variety of facial expression sequences coded with Action Units[9], or components of facial expression, in the Cohn-Kanade database, they are not labeled with facial expression categories. Hence, we hand-labeled the last frame of them with one of the non-neutral expressions, following the FACS rule[10]. We excluded about one third of sequences that were not labeled with six basic facial expressions[8]: angry, sad, surprise, happy, fear, and disgust. Next, we excluded about one half of the remaining sequences where the stationary head assumption (described in 3.6) was violated, that is, where the head moved while the facial expressions were being demonstrated. 129 video sequences were finally selected (13, 20, 14, 43, 12, and 27 for angry, sad, surprise, happy, fear, and disgust expressions, respectively) from the 53 subjects. All initial frames of the selected 129 sequences were inevitably labeled as neutral. All labeled frames, i.e. the first and last frames in each selected video sequence, were used as training image candidates. That is, our target facial expressions in the person-independent strategy are seven categories: neutral and six basic facial expressions.

The average facial expression recognition rates were calculated as follows. First, we tracked (estimated facial pose and expression) over all video sequences for each subject in the database. Next, we picked up the recognition results of the first frames and the last frames for all tracked sequences. Thus, the total number of the target frames for neutral expression is the sum of those for the non-neutral expression (each of the six categories started with a neutral expression). Finally, we calculated the recognition rate separately for each of the seven expressions.

⋆1 A part of video sequences for the result with both the fixed-pose dataset and the free-pose dataset are available from *http://www.hci.iis.u-tokyo.ac.jp/˜kumano/papers/IJCV2008/*.

Fig. 12 The average face image for each facial expression, $\overline{\boldsymbol{g}}_e$, learned from the Cohn-Kanade DFAT-504 database. Each average face image consists of the average facial part images. $\overline{\boldsymbol{g}}_{e,p}$.

Learning stage

We developed a fully automatic system for learning a person-independent, or general, variable-intensity template. Given training images, it proceeds as follows.

First, facial parts are detected in the rotationally aligned training image (3.6) in neutral expression of each subject. The face part regions in non-neutral expressions are set to be the same as those of the same subject in neutral expression. Hereinafter, these detected face images are just referred to as training images, and those in expression e for the j-th subject are denoted as $\boldsymbol{g}_{j,e}$, especially those in neutral expression also as $\boldsymbol{g}_{j,\mathrm{NEU}}$.

Next, all training images of each subject are normalized to remove the effect of global lighting, skin color, etc. The normalization is given by $\tilde{\boldsymbol{g}}_{j,e} = (\boldsymbol{g}_{j,e} - \alpha^{(j)}\mathbf{1})/\beta^{(j)}$. The normalization factors for the j-th subject, $\alpha^{(j)}$ and $\beta^{(j)}$, are calculated as the mean and variance of intensity, respectively, in the rectangle in $\boldsymbol{g}_{j,\mathrm{NEU}}$ that encloses eyebrow and mouth regions.

Then, average face images of each facial expression e, $\overline{\boldsymbol{g}}_e$, are generated individually for every facial part p with the normalized images $\tilde{\boldsymbol{g}}_{j,e}$ for all subject j. This average images of each facial part p are hereinafter referred to as average facial part images, and denoted as $\overline{\boldsymbol{g}}_{e,p}$. An example of the average facial part image $\overline{\boldsymbol{g}}_{e,p}$ is shown in **Fig. 12**.

Finally, a set of interest points is selected in the average facial part images $\overline{\boldsymbol{g}}_{e,p}$ as described in 3.1.1. A sample of the resulting set of interest points is shown in the left side of **Fig. 13**.

Testing stage

In the test stage, the variable-intensity template for the person is prepared by using the average face images $\overline{\boldsymbol{g}}_e$.

First, the facial parts are detected in the initial frame of the test video sequence, after rotating the frame until the iris centers fall on a horizontal line using the approach described in 3.6 (hereafter this rotated image is just referred to as the initial frame).

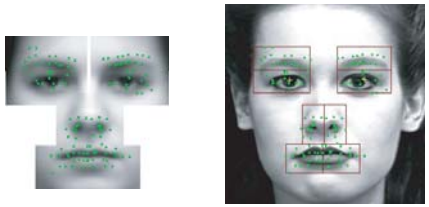


Fig. 13 Left: Example of set of interest points defined in the average facial part images $\overline{\boldsymbol{g}}_{e,p}$ that are learned with the Cohn-Kanade DFAT-504 database. Right: The set of interest points $\mathcal{P}$ generated by projecting the average facial part image into the initial frame, or the training image for neutral expression $\boldsymbol{g}_{j,\mathrm{NEU}}$, of a target subject.

Fig. 14 Sample recognition results with the set of interest points $\mathcal{P}$ shown in the right part of Fig. 13.

Next, the coordinates of the i-th interest point in the initial frame, $\boldsymbol{p}_i$, is calculated by scaling and translating its coordinates in the average facial part images $\overline{\boldsymbol{g}}_{e,p}$. This mapping adapts the average facial part image to the same facial part region in the initial frame both in size and position. The right part of Fig. 13 shows the mapping results.

Finally, the mean intensity in the intensity distribution in each facial expression $\mu(e)$ is restored via denormalization of the average facial part images: $\widehat{\boldsymbol{g}}_{e,p} = \beta^{(j)}\,\overline{\boldsymbol{g}}_{e,p} + \alpha^{(j)}\mathbf{1}$, where $\widehat{\boldsymbol{g}}_{e,p}$ is the restored (training) image, $\alpha^{(j)}$ and $\beta^{(j)}$ are calculated as described in the above test stage process, where the face region in the initial frame is used as $\boldsymbol{g}_{j,\mathrm{NEU}}$. Finally, the mean intensity $\mu_i(e)$ is set to be the intensity of the restored training image, $\widehat{\boldsymbol{g}}_{e,p}$, at coordinates $\boldsymbol{p}_i$.

Recognition performance

We evaluated the recognition performance achieved by this person-independent strategy, based on the leave-one-subject-out cross-validation. The results with our paired interest points and those with random points are shown in **Table 3** (also in **Fig. 14**) and **Table 4**, respectively. The overall recognition rate was about 60% with our paired interest points,

Table 3 Confusion matrix with the Cohn-Kanade DFAT-504 database by using our paired interest points with person-independent strategy. The overall recognition rate is 59.8[%].

GT \ RCG	N	A	Sd	Sp	H	F	D
Neutral (N)	82.9	5.7	1.9	1.6	0.3	5.7	1.9
Angry (A)	23.1	30.7	7.7	7.7	7.7	7.7	15.4
Sad (Sd)	20.0	15.0	35.0	5.0	5.0	10.0	10.0
Surprise (Sp)	0.0	0.0	0.0	100.0	0.0	0.0	0.0
Happy (H)	5.0	3.7	0.0	2.5	73.8	2.5	12.5
Fear (F)	16.7	0.0	8.3	8.3	33.4	25.0	8.3
Disgust (D)	3.8	19.2	3.8	0.0	0.0	1.9	71.3

unit is (%)

Table 4 Confusion matrix with the Cohn-Kanade DFAT-504 database by using randomly defined points with person-independent strategy. The overall recognition rate is 70.2[%].

GT \ RCG	N	A	Sd	Sp	H	F	D
Neutral (N)	82.1	4.1	4.6	3.8	0.0	1.6	3.8
Angry (A)	7.7	69.2	0.0	0.0	0.0	15.4	7.7
Sad (Sd)	15.0	5.0	70.0	5.0	0.0	5.0	0.0
Surprise (Sp)	0.0	0.0	0.0	100.0	0.0	0.0	0.0
Happy (H)	0.0	3.7	0.0	2.5	83.8	5.0	5.0
Fear (F)	25.0	0.0	8.3	8.3	33.4	25.0	0.0
Disgust (D)	9.6	3.8	3.8	3.8	0.0	17.3	61.7

unit is (%)

and about 70% with the random points. In both results, fear expression recorded significantly lower recognition rates than the other expressions. The average recognition rate excluding fear expression with random points is about 80%.

We were encouraged by these results with the person-independent strategy, even though these recognition rates are lower than those with our person-specific strategy in 4.1.2 and those of state-of-the-art methods such as 22), 28), 38), 47), 49) (over 90% for all methods$^{\star 1}$). Note that, our original target is the person-specific strategy, and the objective of this evaluation was to assess if our approach has the potential to be extended to the person-independent version.

The main difference between the person-specific strategy and the person-independent strategy is whether the interest points are correctly allocated. With the person-specific strategy, the paired interest points are correctly allocated directly to the face image of the target subject, as show in Fig. 3. On the other hand, with the person-independent strategy, we locate the paired interest points in the average faces learned with the database. Hence, without precise facial part detection, these interest points are likely to be misallocated in the actual face image for each subject, as shown in **Fig. 15**. Consequently, the recognition rates with person-independent strategy seem to be significantly degraded. By improving the facial part detection, the performance of our paired interest points with the person-independent strategy is expected to outperform those with the random points.

The interest point misallocation affects more severely for the facial expressions that produce similar appearances each other. For example, as shown in Fig. 12, angry, sad and disgust expressions resemble each other, especially in the eyebrow regions. Inevitably, the eyebrow regions are likely to yield large localization errors, especially with our rectangle-based detector, due to their large interpersonal variation in position, size, and shape. Furthermore, misrecognition of some neutral expressions, in both Table 3 and Table 4, also seems to be caused by the facial part mislocalization. The neutral eyebrow image in the right part of Fig. 15 associates angry and disgust expressions (lowered eyebrow).

We suggest that the low recognition rates for fear expression are caused by the large interpersonal difference in the display of fear expression in the database, especially in the eyebrow regions (the mouth region is chiefly determiner of the fear expression). Consequently, the eyebrow regions in average face image in Fig. 12 are significantly blurred compared to the other expressions. Such a variety of fear expressions are hardly to be recognized correctly with the degenerated intensity distribution model. Consequently, many fear expressions are mistakenly recognized as happy expression, where the mouth region resembles those in fear expression. This issue could be mitigated by modifying our method to divide the intensity distribution model into upper and lower face regions (c.f. Action Units).

4.3 Performance evaluation for head tracking

Finally, we evaluate the performance of the proposed method for head tracking, separated from those for facial expression recognition, by using the Boston University (BU) face tracking database[4]$^{\star 2}$. The BU face tracking database contains sets of 30[fps] QVGA (320×240 pixel) MPEG compressed video sequences. These sequences were

$\star$1 The neutral expression was not one of their targets in 49).

$\star$2 *http://www.cs.bu.edu/groups/ivc/HeadTracking/*

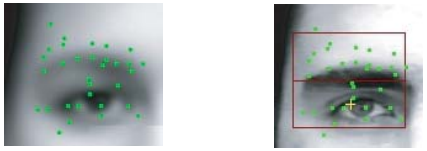


Fig. 15 Example of misallocation of interest points in eyebrow region. Left: interest points defined in the average face. Right: those allocated in the actual face image. Both images shows neutral expression. The interest points on the eyebrow in the average image do not so in the actual image.

Fig. 16 Sample tracking results on BU face tracking database.

captured under uniform illumination or varying illumination together with head pose recorded by a magnetic sensor. Subjects were asked to perform free head motion, which included translation and rotation. We use the set of video sequences captured under uniform illumination. Moreover, the BU database does not contain any frames with distinctive facial expression. Accordingly, we generated the intensity distribution model that describes only neutral expression.

Table 5 shows the mean absolute angular error of our head pose estimation module for yaw, pitch, and roll together with the results of two existing methods listed in 32). Our system is ranked between them. Table 5 also shows that vertical head pose detection is more accurate than horizontal detection.

These results demonstrated that the performance of head pose estimation with the paired interest points almost matches that with random points and bests that with edge points. The cause of the low accuracy with edge points seems that they are sensitive to the misalignment caused by shape model error and to the slight change in neutral expression. Compared to the random points, little effect of the use of the paired interest points is found in Table 5.

4.4 Discussion

This paper evaluated the estimation performance of the proposed method in three aspects: facial expressions in non-frontal-view with person-specific strategy, person-independent strategy, and head tracking accuracy.

Table 5 Comparison with mean absolute error in head pose estimation between the proposed method, its variations in interest points, and existing methods. The accuracies of these two existing methods are taken from 32).

Methods	Errors [deg]		
	yaw	pitch	roll
Pair	7.0	4.4	2.8
Edge	11.9	8.2	7.6
Random	7.1	4.7	3.2
Cascia et al.[4)]	3.3	6.1	9.8
Xiao et al.[46)]	3.8	3.2	1.4

The results with person-specific strategy using our dataset demonstrate that our method (person-specific) can achieve the high performance for facial expression recognition even with a large head rotation in almost all orientations. This is because it avoids the problem of personal differences in terms of the geometric arrangement of facial parts and facial expressions. These results suggest that locating interest points away from the edges of facial parts can suppress the severe impact of interest point shift produced by out-of-plane rotation of the head with shape model error.

By enhancing interest point selection module, the performance of our system would be increased. First, the interest point selection module in this paper uses only a neutral expression image. However, selecting salient points for each facial expression is expected to increase the recognition performance. Second, the head tracking will become more stable by more accurately fitting the face shape model.

Our experiments also demonstrate the potential of our approach in supporting person-independent applications. The performance of the person-independent version is expected to be improved by enhancing the facial part detection module using techniques such as active contours[20)] or profile-based matching[17)], i.e. shifting each paired point along the line connecting the interpair points to be centered on the edge.

Based on these results with both strategies, we believe that our person-independent version could handle large head rotations as well as the person-specific version. Compared to the existing deformation-based facial expression recognition methods, our approach has the large advantage that facial expression models can be learned just from monocular images, that is, many facial expression databases can be utilized. Further-

more, unlike the dense-feature-based approach (such as optical-flow-based one), our approach can introduce any three-dimensional face shape model to improve the robustness against large head rotations, without increasing the computational cost.

5. Summary and future work

In this paper, we presented a particle filter-based method for estimating facial pose and expression simultaneously; it uses a novel face model called the variable-intensity template. Our method has the distinct advantage that a face model for each person can be prepared very easily with a simple step. Our method recognized the five facial expressions with overall accuracy of 95.4% for horizontal, vertical and in-plane facial orientations in the range of ± 40, ± 20, and ± 40 degrees from the frontal view, respectively.

We intend to tackle the recognition of subtle spontaneous facial expressions. To this end, we would like to apply unsupervised learning with an online clustering technique, and to estimate the intensity of facial expressions from changes in interest point intensity by referring to optical flow estimation.

References

1) Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I. and Movellan, J.: Automatic Recognition of Facial Actions in Spontaneous Expressions, *Journal of Multimedia*, Vol.1, No.6, pp.22–35 (2006).
2) Beaton, A. and Tukey, J.: The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data, *Technometrics*, Vol.16, No.2, pp.147–185 (1974).
3) Black, M. and Yacoob, Y.: Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion, *International Journal of Computer Vision*, Vol.25, pp.23–48 (1997).
4) Cascia, M.L., Sclaroff, S. and Athitsos, V.: Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, No.4, pp.322–336 (2000).
5) Chang, Y., Hu, C., Feris, R. and Turk, M.: Manifold based analysis of facial expression, *Image and Vision Computing*, Vol.24, No.6, pp.605–614 (2006).
6) Cohen, I., Sebe, N., Garg, A., Chen, L.S. and Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling, *Computer Vision and Image Understanding*, Vol.91, No.1-2, pp.160–187 (2003).
7) Dornaika, F. and Davoine, F.: Simultaneous Facial Action Tracking and Expression Recognition in the Presence of Head Motion, *International Journal of Computer Vision*, Vol.76, No.3, pp.257–281 (2008).
8) Ekman, P. and Friesen, W..: *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expresssions*, NJ: Prentice Hall. (1975).
9) Ekman, P. and Friesen, W.V.: *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press (1978).
10) Ekman, P., Friesen, W.V. and Hager, J.C.: *FACS Investigator's Guide. A Human Face*, A Human Face (2002).
11) Fasel, B. and Luttin, J.: Automatic facial expression analysis: Survey, *Pattern Recognition*, Vol.36, pp.259–275 (2003).
12) Fasel, B., Monay, F. and Gatica-Perez, D.: Latent semantic analysis of facial action codes for automatic facial expression recognition, *In Proceedings of the ACM SIGMM international workshop on Multimedia information retrieval*, pp.181–188 (2004).
13) Geman, S. and McClure, D.E.: Statistical methods for tomographic image reconstruction, *Bulletin of the International Statistical Institute*, Vol.LII, pp.5–21 (1987).
14) Gokturk, S.B., Tomasi, C., Girod, B. and Bouguet, J.: Model-Based Face Tracking for View-Independent Facial Expression Recognition, *In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp.287–293 (2002).
15) Gross, R., Matthews, I. and Baker, S.: Generic vs. person specific Active Appearance Models, *Image and Vision Computing*, Vol.23, No.11, pp.1080–1093 (2005).
16) Hu, Y., Zeng, Z., Yin, L., Wei, X., Zhou, X. and Huang, T.S.: Multi-View Facial Expression Recognition, *In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition* (2008).
17) Huang, C.-L. and Huang, Y.-M.: Facial expression recognition using model-based feature extraction and action parameters classification, *Journal of Visual Communication and Image Representation*, Vol.8, No.3, pp.278–290 (1997).
18) Isard, M. and Blake, A.: Condensation – conditional density propagation for visual tracking, *International Journal of Computer Vision*, Vol.29, No.1, pp.5–28 (1998).
19) Kanade, T., Cohn, J. and Tian, Y.-L.: Comprehensive Database for Facial Expression Analysis, *In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp.46–53 (2000).
20) Kass, M., Witkin, A. and Terzopoulos, D.: Snakes: Active contour models, *International Journal of Computer Vision*, Vol.1, No.4, pp.321–331 (1988).
21) Koelstra, S. and Pantic, M.: Non-rigid registration using free-form deformations for recognition of facial actions and their temporal dynamics, *In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition* (2008).
22) Kotsia, I. and Pitas, I.: Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines, *IEEE Transactions on Image Processing*, Vol.16, No.1, pp.172–187 (2007).
23) Kumano, S., Otsuka, K., Yamato, J., Maeda, E. and Sato, Y.: Pose-Invariant Facial Expression Recognition Using Variable-Intensity Templates, *In Proceedings of Asian Conference on Computer Vision*, Vol.1, pp.324–334 (2007).
24) Kumano, S., Otsuka, K., Yamato, J., Maeda, E. and Sato, Y.: Combining Stochastic and

Deterministic Search for Pose-Invariant Facial Expression Recognition, *In Proc. British Machine Vision Conference*, No.2 (2008).
25) Kumano, S., Otsuka, K., Yamato, J., Maeda, E. and Sato, Y.: Pose-Invariant Facial Expression Recognition Using Variable-Intensity Templates, *International Journal of Computer Vision*, Vol.83, No.2, pp.178–194 (2009).
26) Lanitis, A., Taylor, C.J. and Cootes, T.F.: Automatic Interpretation and Coding of Face Images Using Flexible Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp.743–756 (1997).
27) Liao, W.-K. and Cohen, I.: Belief Propagation Driven Method for Facial Gestures Recognition in Presence of Occlusions, *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, pp.158–163 (2006).
28) Littlewort, G., Bartlett, M.S., Fasel, I.R., Susskind, J. and Movellan, J.R.: Dynamics of facial expression extracted automatically from video, *Image and Vision Computing*, Vol.24, No.6, pp.615–625 (2006).
29) Loy, G. and Zelinsky, A.: Fast Radial Symmetry for Detecting Points of Interest, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 8, pp. 959–973 (2003).
30) Lucey, S., Matthews, I., Hu, C., Ambadar, Z., Torre, F. and Cohn, J.: AAM Derived Face Representations for Robust Facial Action Recognition, *In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp.155–160 (2006).
31) Matsubara, Y. and Shakunaga, T.: Sparse Template Matching and Its Application to Real-time Object Tracking, *IPSJ Transactions on Computer Vision and Image Media*, Vol.46, No.9, pp.17–40 (In Japanese) (2005).
32) Murphy-Chutorian, E. and Trivedi, M.M.: Head Pose Estimation in Computer Vision: A Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence (To be appeared)* (2008).
33) Oka, K. and Sato, Y.: Real-time Modeling of Face Deformation for 3D Head Pose Estimation, *In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp.308–320 (2005).
34) Otsuka, K., Sawada, H. and Yamato, J.: Automatic inference of cross-modal nonverbal interactions in multiparty conversations: ”who responds to whom, when, and how?” from gaze, head gestures, and utterances, *In Proceedings of the International Conference on Multimodal Interfaces*, pp.255–262 (2007).
35) Pantic, M. and Bartlett, M.: *Machine Analysis of Facial Expressions*, pp.377–416, I-Tech Education and Publishing (2007).
36) Pantic, M. and Rothkrantz, L.: Expert system for automatic analysis of facial expression, *Image and Vision Computing*, Vol.18, pp.881–905 (2000).
37) Pantic, M. and Rothkrantz, L. J.M.: Automatic analysis of facial expressions: the state of the art, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, pp.1424–1445 (2000).
38) Sebe, N., Lew, M.S., Sun, Y., Cohen, I., Gevers, T. and Huang, T.S.: Authentic facial expression analysis, *Image and Vision Computing*, Vol.25, No.12, pp.1856–1863 (2007).
39) Tang, H. and Huang, T.S.: 3D Facial Expression Recognition Based on Properties of Line Segments Connecting Facial Feature Points, *In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition* (2008).
40) Tian, Y.L., Kanade, T. and Cohn, J.F.: Recognizing Action Units for Facial Expression Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.23, No.2, pp.97–115 (2001).
41) Tian, Y.-L., Kanade, T. and Cohn, J.: *Facial expression analysis*, Springer (2005).
42) Tong, Y., Liao, W. and Ji, Q.: Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.29, No.10, pp.1683–1699 (2007).
43) Viola, P. and Jones, M.: Rapid object detection using a boosted cascade of simple features, *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.511–518 (2001).
44) Wang, J., Yin, L., Wei, X. and Sun, Y.: 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution, *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1399–1406 (2006).
45) Xiao, J., Baker, S., Matthews, I. and Kanade, T.: Real-Time Combined 2D+3D Active Appearance Models, *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.2, pp.535 – 542 (2004).
46) Xiao, J., Moriyama, T., Kanade, T. and Cohn, J.: Robust Full-Motion Recovery of Head by Dynamic Templates and Re-registration Techniques, *International Journal of Imaging Systems and Technology*, Vol.13, pp.85–94 (2003).
47) Yang, P., Liu, Q., Cui, X. and Metaxas, D.N.: Facial Expression Recognition Based on Dynamic Binary Patterns, *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2008).
48) Zhang, W., Chen, H., Yao, P., Li, B. and Zhuang, Z.: Precise Eye Localization with AdaBoost and Fast Radial Symmetry, *In Proceedings of the International Conference on Computational Intelligence and Security*, Vol.1, pp.725–730 (2006).
49) Zhao, G. and Pietikainen, M.: Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.29, No.6, pp.915–928 (2007).
50) Zhu, Z. and Ji, Q.: Robust Real-Time Face Pose and Facial Expression Recovery, *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.681–688 (2006).