Memory-Based Particle Filter for Tracking Objects with Large Variation in Pose and Appearance

Dan Mikami, Kazuhiro Otsuka, and Junji Yamato

NTT Communication and Science Laboratories

Abstract. A novel memory-based particle filter is proposed to achieve robust visual tracking of a target's pose even with large variations in target's position and rotation, i.e. large appearance changes. The memory-based particle filter (M-PF) is a recent extension of the particle filter, and incorporates a memory-based mechanism to predict prior distribution using past memory of target state sequence; it offers robust target tracking against complex motion. This paper extends the M-PF to a unified probabilistic framework for joint estimation of the target's pose and appearance based on memory-based joint prior prediction using stored past pose and appearance sequences. We call it the Memory-based Particle Filter with Appearance Prediction (M-PFAP). A memory-based approach enables generating the joint prior distribution of pose and appearance without explicit modeling of the complex relationship between them. M-PFAP can robustly handle the large changes in appearance caused by large pose variation, in addition to abrupt changes in moving direction; it allows robust tracking under self and mutual occlusion. Experiments confirm that M-PFAP successfully tracks human faces from frontal view to profile view; it greatly eases the limitations of M-PF.

1 Introduction

Visual object tracking, one of the most important techniques in computer vision [1], is required for a wide range of applications such as automatic surveillance, man-machine interfaces [2,3], and communication scene analysis [4]. Target tracking has still been acknowledged as a challenging problem because the target's appearance changes greatly due to pose variation, occlusion, illumination change, etc. For example, when an object rotates, its visible surface gradually becomes invisible, i.e. self-occlusion, and hidden surfaces becomes visible. Mutual occlusion, the interjection of another object between the target and the camera, makes the target's visible surface invisible. Also, the target tracker needs to handle complex motion, such as when the moving direction abruptly reverses, which can occur with occlusions in real world situations.

Bayesian filter-based trackers have been acknowledged as a promising approach; they represent a unified probabilistic framework for sequentially estimating the target state from an observed data stream [5]. At each time step, the Bayesian filter computes the posterior distribution of the target state by using

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part III, LNCS 6313, pp. 215–228, 2010.

observation likelihood and the prior distribution. One variant, the particle filter, has been widely used for target tracking. It represents probability distributions of the target state by a set of samples, called particles. Particle filter can potentially handle non-Gaussian distribution and nonlinear dynamics/observation processes; this contributes to robust tracking. For object tracking, an example of target state is the position and orientation of the target.

We proposed the memory-based particle filter (M-PF) as an extension of the particle filter [6]. M-PF eases the Markov assumption of PF and predicts the prior distribution based on target's long-term dynamics using past history of the target's states. M-PF realized robustness against abrupt object movements and quick recovery from tracking failure without explicit modeling of target's dynamics. However, M-PF employs the same observation process as the traditional PF. The visual tracker in [6] uses a single template representing frontal face, which is built at tracker initialization. Therefore, the M-PF-based tracker can handle face rotation only so long as the initial frontal face remains visible; [6] suggests that the horizontal limit is 50 degrees.

This paper extends M-PF to a unified probabilistic framework for joint estimation of target's position/pose and its appearance based on memory-based joint prior distribution prediction using stored past pose-appearance pairs. We call it the Memory-based Particle Filter with Appearance Prediction (M-PFAP). The appearance of an object varies with its pose. By predicting appearance from pose, M-PFAP enables robust tracking against changes in appearance. A memory-based approach is proposed to generate the joint prior distribution of pose and appearance; the complex relationship between them is not explicitly modeled. M-PFAP can robustly handle the large changes in appearance caused by large pose variation, in addition to abrupt changes in moving direction; it allows robust tracking under self and mutual occlusion. To the best of our knowledge, M-PFAP is the first pose tracker that handles pose-appearance relationship as a probabilistic distribution and that simultaneously predicts future pose and appearance in a memory-based approach. As the tracking target, this paper focuses on the face and we implement the M-PFAP-based face pose tracker. Experiments confirm that M-PFAP successfully tracks human faces from frontal view up to profile view, i.e. 90 degrees horizontally; it far exceeds the limits of M-PF.

This paper is organized as follows. Section 2 overviews related works, Sect. 3 proposes M-PFAP, and Sect. 4 describes face pose tracking based on M-PFAP, experiments, and results. Finally, Sect. 5 gives our conclusions.

2 Related Works

2.1 Template Matching-Based Tracking and Template Update

Template matching has been widely employed for visual target tracking; the template represents the target's appearance from the camera's view. The target position is the best-matched position of the template on the input image. To cope with appearance change, the template is updated repeatedly over time [7,8]. However, error in the estimates of position/pose yields erroneous templates

and error accumulates, which results in tracking failure. It is called "drift". To suppress drift, two approaches have been proposed.

The first approach uses pose-invariant features extracted from the target. The tracker of Matthews et al. [9] employs a set of invariant features from multiple views of the target object; the tracker can keep track of the target even when its pose changes. Jepson et al. [10] proposed a WSL model which uses separate models for Stable, Wandering, and Lost situations; these models are mixed to predict the target appearance by using the EM algorithm. Zelniker et al. [11] combined multiple features according to e.g. illumination condition. These methods can be used only for position estimation, not for pose estimation.

The second approach is template updating through adaptive criteria. Morency et al. [12,13] and Ross et al. [14] proposed methods that use an initial template as a supplement to avoid error accumulation; both the initial template and updated template are used for matching. However, the use of the initial template limits the pose range possible. In the example of [13], a frontal face is used as the initial template, and the horizontal rotation angle in their experiment was up to 50 degrees. Lefèvre et al. [15] used view-based templates obtained online. Their approach is to generate templates from not only frontal views but also from profile views. This allows an appearance model to be generated by interpolation, not by extrapolation. However, the trackable angle range is restricted by the profile views.

M-PFAP provides a new approach to handling the large appearance changes caused by pose change. It handles pose-appearance relationship as a probabilistic distribution, and estimates pose and appearance simultaneously in the Bayesian filter framework by using the memory-based approach.

2.2 Memory-Based Particle Filter (M-PF)

M-PF [6] realized robust target tracking without explicit modeling of target's dynamics even when a target moves quickly.

Fig.1 outlines M-PF. M-PF keeps temporal sequence of past state estimates $\widehat{\mathbf{x}}_{1:T} = \{\widehat{\mathbf{x}}_1, \cdots, \widehat{\mathbf{x}}_T\}$ in memory. Here, $\widehat{\mathbf{x}}_{1:T}$ denotes a sequence of state estimates from time 1 to time T, and $\widehat{\mathbf{x}}_t$ denotes a pose estimate at time t. M-PF assumes that the subsequent parts of past similar states provide the good estimates of the current future.

M-PF introduced Temporal Recurrent Probability (TRP), which is a probability distribution defined in the temporal domain and indicates the possibility that a past state will reappear in the future. To predict the prior distribution, M-PF starts with TRP modeling. It then conducts temporal sampling based on TRP. The sampled histories are denoted by blue dots in Fig.1. It retrieves the corresponding past state estimates for each sampled time step, which are denoted by pink dots in Fig.1. After that, considering the uncertainty in the state estimates, each referred past state is convoluted with kernel distributions (light green dist. in Fig.1), and they are mixed together to generate the prior distribution (green dist. in Fig.1). Finally, a set of particles is generated according to the prior distribution (black dots in Fig.1). M-PF-based face pose tracker

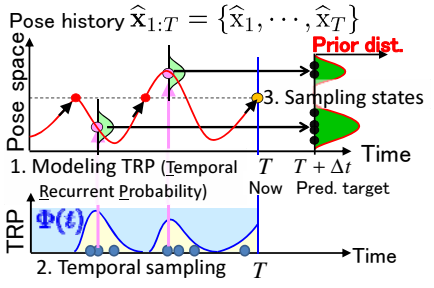


Fig. 1. M-PF employs past state sequences to predict a future state. First, it calculates the reoccurrence possibility of past state estimates (TRP). Past time steps are then sampled based on TRP. Past state estimates corresponding to the sampled time steps are combined to predict prior distribution. M-PF enables the implicit modeling of complex dynamics.

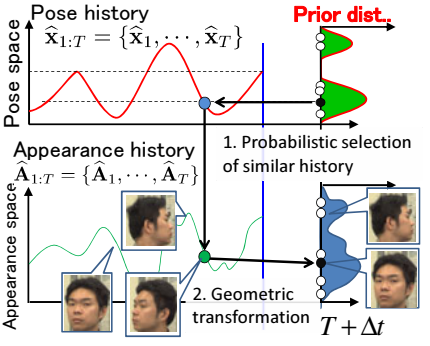


Fig. 2. M-PFAP extends M-PF [6] to realize robustness against large changes in pose. We focus on the fact that the pose-appearance relationship is not one-to-one but stochastic. The key extension from M-PF is prediction of joint prior distribution of pose and appearance.

in [6] estimates the position and rotation at each time step. M-PF uses the same observation process as traditional PF, which uses a single template built at initialization. This yields the 50 degree face rotation limit noted in [6].

M-PFAP extends M-PF. It adds appearance prior distribution prediction to M-PF for enabling handling of large appearance changes while keeping the merits of M-PF; robustness against abrupt movements and recoverability from tracking failure.

3 Memory-Based Particle Filter with Appearance Prediction (M-PFAP)

3.1 Formulation of M-PFAP

In this section, we define M-PFAP by extending the Bayesian filter formulation. The Bayesian filter consists of two processes, *update* and *prediction*, as

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) = k_t \cdot p(\mathbf{z}_t|\mathbf{x}_t) \cdot p(\mathbf{x}_t|\mathbf{z}_{1:t-1}), \tag{1}$$

$$p(\mathbf{x}_{t+1}|\mathbf{z}_{1:t}) = \int p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}) p(\mathbf{x}_{1:t}|\mathbf{z}_{1:t}) d\mathbf{x}_{1:t}, \tag{2}$$

where k_t is a normalization term, $\mathbf{z}_{1:t} = \{\mathbf{z}_1, \cdots, \mathbf{z}_t\}$ and $\mathbf{x}_{1:t} = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$ denote a sequence of observation vectors and that of state vectors from time 1 to t, respectively. Equation (1) corresponds to the update process that computes the posterior distribution of the target state, and (2) corresponds to the prediction process, which calculates the prior distribution for the next time step.

M-PF replaced the prediction process in (2) with memory-based prior prediction as written in (3).

$$p(\mathbf{x}_{t+\Delta t}|\mathbf{z}_{1:t}) := \pi(\mathbf{x}_{t+\Delta t}|\widehat{\mathbf{x}}_{1:t}, \Delta t). \tag{3}$$

M-PF obtains the prior distribution at time $t + \Delta t$ from the history of state estimates $\widehat{\mathbf{x}}_{1:t}$ and the lead time Δt.

M-PFAP adds appearance as the state vector in addition to position and rotation. Hereafter, $\mathbf{X}_t = (\mathbf{x}_t, A_t)$ denotes state vector at time t, where $\mathbf{x}_t$ and A_t denote the position/rotation and the appearance at time t, respectively. Examples of appearance include a set of feature points and corresponding gray levels. The posterior distribution and the prior distribution of M-PFAP are defined below.

$$p(\mathbf{X}_t|\mathbf{z}_{1:t}) = k_t \cdot p(\mathbf{z}_t|\mathbf{x}_t, A_t) \cdot p(\mathbf{x}_t, A_t|\mathbf{z}_{1:t-1}), \tag{4}$$

$$p(\mathbf{X}_{t+\Delta t}|\mathbf{z}_{1:t}) = \pi(\mathbf{X}_{t+\Delta t}|\widehat{\mathbf{X}}_{1:t}, \Delta t) = \pi(\mathbf{x}_{t+\Delta t}, A_{t+\Delta t}|\widehat{\mathbf{x}}_{1:t}, \widehat{A}_{1:t}, \Delta t), \tag{5}$$

where $\widehat{\mathbf{X}}_{1:t} = \{(\widehat{\mathbf{x}}_1, \widehat{A}_1), \cdots, (\widehat{\mathbf{x}}_t, \widehat{A}_t)\}$ denotes the sequence of pairs of estimated pose $\widehat{\mathbf{x}}_t$ and appearance $\widehat{A}_t$ at time t. $\widehat{A}_{1:t} = \{\widehat{A}_1, \cdots, \widehat{A}_t\}$ denotes the sequence of appearances from time 1 to time t. We define the joint prior distribution of pose and appearance described in (5) as follows, by introducing a conditional probability of a future appearance given a future pose $\pi(A_{T+\Delta t}|\mathbf{x}_{t+\Delta t}, \widehat{\mathbf{x}}_{1:t}, \widehat{A}_{1:t})$ and a past history of appearance and pose $(\widehat{\mathbf{x}}_{1:t}, \widehat{A}_{1:t})$.

$$\text{Equation (5)} = \pi(\mathbf{x}_{t+\Delta t}|\widehat{\mathbf{x}}_{1:t}, \widehat{A}_{1:t}, \Delta t) \cdot \pi(A_{t+\Delta t}|\mathbf{x}_{t+\Delta t}, \widehat{\mathbf{x}}_{1:t}, \widehat{A}_{1:t}, \Delta t), \tag{6}$$

$$:= \pi(\mathbf{x}_{t+\Delta t}|\widehat{\mathbf{x}}_{1:t}, \Delta t) \cdot \pi(A_{t+\Delta t}|\mathbf{x}_{t+\Delta t}, \widehat{\mathbf{x}}_{1:t}, \widehat{A}_{1:t}). \tag{7}$$

The first part of (7) corresponds to prior distribution of pose $\mathbf{x}_{t+\Delta t}$. It assumes that the pose at Δt time in the future $\mathbf{x}_{t+\Delta t}$ is independent of the past history of appearance $\widehat{A}_{1:t}$, in other words, the dynamics of object movement are independent of the past appearance history. The last part of (7) corresponds to the conditional probability for a given pose, $\mathbf{x}_{t+\Delta t}$; it assumes that the appearance at Δt time in the future $A_{t+\Delta t}$ depends on the pose at time $T + \Delta t$, $\mathbf{x}_{T+\Delta t}$, and is independent of lead time Δt. The first part of (7) is prior distribution of pose; i.e. it equals the prior distribution of M-PF.

To define the conditional probability for a given pose, M-PFAP assumes that the main determinant of appearance is pose. Note that there is no deterministic one-to-one correspondence between them, i.e. significant uncertainty exists in the relationship. This assumption is based on the following observations. First, when the object rotates, its visible surface gradually becomes invisible and vice versa. However, appearance is also influenced by various factors such as illumination change and non-rigid deformation. Moreover, the explicit modeling of the appearance changes caused by pose is difficult, because appearance exhibits complex dynamics of high dimensionality.

Based on the above assumption, M-PFAP represents the relationship between pose and appearance as a probability distribution, like in Fig.3. In Fig.3, the

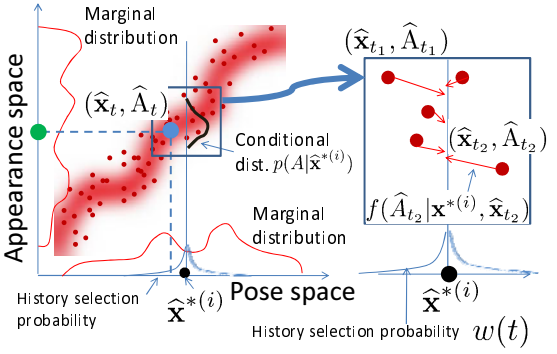


Fig. 3. An illustration of pose-appearance relationship; what we want to compute here is the conditional prior distribution, $p(A|\mathbf{x}^{*(i)})$, for given the pose $\mathbf{x}^{*(i)}$. We approximately represent the conditional distribution using past pose-appearance pairs. Each selected past pair is geometrically transformed, $f(\cdot)$, to compensate the difference between desired pose and selected pose.

horizontal axis and vertical axis denote pose space and appearance space, respectively. As seen in Fig.3, for a given pose, there is a distribution of possible appearances and vise versa. M-PFAP handles such uncertainty within the M-PF framework. It exploits the long-term history of the target state to predict complex prior distribution. This paper proposes a memory-based algorithm that jointly predicts appearance and pose.

3.2 Algorithm of the M-PFAP

M-PFAP sequentially estimates the target position and pose by repeating the posterior distribution estimation in (4) and the prior distribution prediction in (7). In the prior prediction step, M-PFAP predicts a joint prior distribution of pose and appearance. In the posterior prediction step, the observation likelihood of each particle is calculated by using the appearance estimated in the prior prediction step. Then, point estimates of pose and appearance are obtained from their joint posterior distribution. Next, the pair of pose and appearance is added to the history. Prior distribution prediction, posterior distribution estimation, and accumulation of history are described below.

Prior distribution prediction in M-PFAP
M-PFAP generates a set of particles, $\{(\mathbf{x}^{*(1)}, A^{*(1)}), \cdots, (\mathbf{x}^{*(N)}, A^{*(N)})\}$, which represents a joint prior distribution of pose and appearance, by using a memory-based mechanism and the stored history of them.

We focus on the fact that the joint distribution of pose and appearance defined by (7) is the product of the prior distribution of pose and the conditional probability of appearance (conditioned by pose). Therefore, we employ two step solutions. In Step-1, a set of particles that represents a prior distribution of pose

is created by the previous M-PF, which is described in Fig.1 and Sect.2.2. In Step-2, the appearance that corresponds to each particle is generated. The process is shown in Fig.2 and Fig.3. In Fig.2, the current time is denoted by T, and prediction target is Δt time future state. Here, we assume we already have the history of pose $\widehat{\mathbf{x}}_{1:T}$ and appearance $\widehat{A}_{1:T}$. Each step is detailed below.

Step-1. Generating pose prior samples

Step-1 generates a set of particles $\{\mathbf{x}^{*(1)}, \cdots, \mathbf{x}^{*(N)}\}$ that represents the prior distribution of pose at time $T + \Delta t$ in the same manner as M-PF. This step corresponds to the first part of (7), $\pi(\mathbf{x}_{T+\Delta t}|\widehat{\mathbf{x}}_{1:T}, \Delta t)$.

Step-2. Prediction of appearance prior

Step-2 uses random sampleing according to $\pi(A_{T+\Delta t}|\mathbf{x}_{T+\Delta t}, \widehat{\mathbf{x}}_{1:T}, \widehat{A}_{1:T})$, which is the last part of (7), to generate a set of appearance samples. It generates appearances $\{A^{*(1)}, \cdots, A^{*(N)}\}$ corresponding to particles $\{\mathbf{x}^{*(1)}, \cdots, \mathbf{x}^{*(N)}\}$, that are obtained from Step-1. Here, what we want to compute is the conditional prior distribution $p(A|\mathbf{x}^{*(i)})$ for given the pose $\mathbf{x}^{*(i)}$. The basic idea is that the appearance distribution can be obtained as a mixture of past appearances whose associated poses are similar to pose condition $\mathbf{x}^{*(i)}$. Based on the idea, first, past pose-appearance pairs are sampled (Step-2-1), and then the past sampled appearances are geometrically transformed to fill in the gap between the desired pose $\mathbf{x}^{*(i)}$ and past sampled pose $\widehat{\mathbf{x}}_t$ (Step-2-2). We define the conditional appearance distribution as

$$p(A|\mathbf{x}^{*(i)}) := \frac{1}{\alpha} \sum_{t=1}^{T} w(t) \cdot \delta(A - f(\widehat{A}_t|\mathbf{x}^{*(i)}, \widehat{\mathbf{x}}_t)), \tag{8}$$

where $f(\cdot)$ denotes the geometric transformation, $\delta(\cdot)$ denotes the delta function, $w(t)$ denotes the weight which is determined by the difference between $\mathbf{x}^{*(i)}$ and $\widehat{\mathbf{x}}_t$, and α is the normalization factor to make $\int p(A|\mathbf{x}^{*(i)})dA = 1$. Random sampling with weight $w(t)$ based on (8) generates the appearance prior distribution. We name weight $w(t)$ the history selection probability. This is defined in the temporal domain based on pose similarity; the higher the similarity between a pose in the history $\widehat{\mathbf{x}}_t$ and that of the target particle, $\mathbf{x}^{*(i)}$, becomes, the higher the history selection probability becomes. The uncertainty that exists in the appearance-pose relationship can be represented as random sampling from the past history. We expect that the mixture of past appearances well reflects the uncertainty in the appearance-pose relationship. This approach is simple but effective; it does not need explicit modeling or stochastic learning

Step-2-1. Sampling history

This step samples a past history of pose that is similar to the particle $\mathbf{x}^{*(i)}$, denoted by a black dot in the upper part of Fig.2 and in Fig.3 on the horizontal line. More specifically, this paper samples one past pose history, $\widehat{\mathbf{x}}_t, t \sim w(t)$, this is because we use enough samples, $\mathbf{x}^{*(i)}, (i = 1, \cdots, N)$, to create sufficient diversity in the appearance distribution. The sampled history is denoted by a blue dot in the upper

part of Fig.2 and Fig.3. As the history selection probability, this paper employs function $w(t)$; this makes the probability proportional to the inverse of the Euclidian distance between the pose of target particle $\mathbf{x}^{*(i)}$ and that of history entry $\widehat{\mathbf{x}}_t$, $(t < T)$.

$$w(t) = \beta / \sqrt{(\widehat{\mathbf{x}}_t - \mathbf{x}^{*(i)})^T \cdot (\widehat{\mathbf{x}}_t - \mathbf{x}^{*(i)})}, \tag{9}$$

where, β is a normalization factor to realize $\sum_{t=1}^{T} w(t) = 1$.

Step-2-2. Appearance prediction

Considering the gap between the pose of sampled $\widehat{\mathbf{x}}_t$ and the target pose $\mathbf{x}^{*(i)}$, Step-2-2 predicts the appearance $A^{*(i)}$ by geometrically transforming $\widehat{A}_t$ based on the pose difference as written in

$$A^{*(i)} = f(\widehat{A}_t | \mathbf{x}^{*(i)}, \widehat{\mathbf{x}}_t). \tag{10}$$

Here, we assume that the local appearance difference caused by the small difference in pose can be well predicted by local geometric transformation. See Sect. 4 for more details.

Posterior distribution estimation

As in (4), posterior distribution is defined by multiplying the prior distribution by the likelihood function for the observation at time step t. In the PF approach, the posterior distribution is represented by weighted particles. The weight is calculated by using a likelihood function for given input image. This function is calculated based on the matching error between the appearance and input images.

In contrast to M-PF, which uses a fixed appearance model, M-PFAP uses predicted appearance in the prior distribution for each particle.

Accumulation of history

At each time T, M-PFAP stores pose-appearance pairs $\widehat{\mathbf{X}}_T = (\widehat{\mathbf{x}}_T, \widehat{A}_T)$. From the particle set that represents the joint posterior distribution of appearance and pose, the point estimates of pose and appearance are calculated. For pose, weight averaging is used, and appearance estimates are obtained from the latest input image by using the target's pose estimates and rough shape model on the current image frame. See Sect.4 for more details.

4 Implementation of Face Pose Tracker

We create a variant of the Sparse Template Condensation Tracker (STCTracker) [16], by using M-PFAP to implement particle filtering. Figure 4 shows the flowchart of the implemented face pose tracker. This section describes some details.

Pose parameter

Target position and pose are described by a vector of seven dimensions, $\mathbf{x} = (m_x, m_y, s, r_r, r_p, r_y, l)$; 2-DOF translation, scale, 3-DOF rotation, and an illumination coefficient.

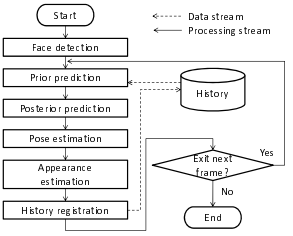


Fig. 4. Flowchart of face pose tracking by using M-PFAP

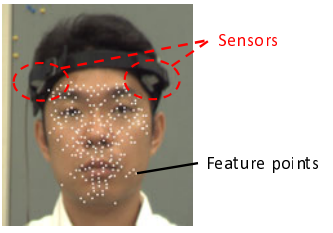


Fig. 5. Interest points deployed for initial face

Sparse template representation of an appearance

M-PFAP employs sparse template matching, which uses a sparse template as the appearance model, as same as [16]. The sparse template consists of a sparse set of pixels within the target region. Here, appearance A is denoted by $\{(u_{x(1)}, u_{y_{(1)}}, u_{z(1)}, b_{(1)}), \cdots, (u_{x(M)}, u_{y_{(M)}}, u_{z(M)}, b_{(M)})\}$, where M denotes the number of interest points, $(u_{x(i)}, u_{y_{(i)}}, u_{z(i)})$ denotes the 3-D position of an interest point, and $b_{(i)}$ denotes its gray level. The matching error is calculated as the sum of differences between the gray levels of the interest points and those of the corresponding points in the input image. Figure 5 shows the 250 interest points selected in the initialization step. These points are selected from edge sides and from minimum or maximum points among 8 neighbor pixels.

Geometric transform for predicting appearance

As written in Step-2-2, M-PFAP uses geometric transformation to bridge the gap between the sampled pose $\widehat{\mathbf{x}}_t$ and the target pose $\mathbf{x}^{*(i)}$. As the geometric transformation, M-PFAP uses 3-DOF rotation. It transforms interest point's corrdinate $[\widehat{u}_x, \widehat{u}_y, \widehat{u}_z]^T$ into desired pose, $[u_x^{*(i)}, u_y^{*(i)}, u_z^{*(i)}]^T$, as in (11).

$$\left[u_x^{*(i)} \; u_y^{*(i)} \; u_z^{*(i)} \right]^T = R(\mathbf{x}^{*(i)}) R'(\widehat{\mathbf{x}}_k) \left[\widehat{u}_x \; \widehat{u}_y \; \widehat{u}_z \right]^T, \tag{11}$$

where $R(\cdot)$ and $R'(\cdot)$ denotes rotation matrix and inverse matrix of $R(\cdot)$, respectively. Additionally, illumination change is assumed by uniform changes in gray levels of a set of interest points. The gray level $b^{*(i)}$ corresponding to interest point $[u_x^{*(i)}, u_y^{*(i)}, u_z^{*(i)}]^T$, is obtained by $b^{*(i)} = v \cdot \widehat{b}$, where $v \sim N(1, \sigma^2)$. $N(1, \sigma^2)$ denotes normal distribution with mean 1 variance σ^2.

Adding pose and appearance into history

At each time step, M-PFAP stores a pose-appearance pair. The pose estimate $\widehat{\mathbf{x}}_T$ is obtained as the point estimate of marginal posterior distribution of pose. Appearance $\widehat{A}_T$, which is a set of three dimensional interest points and corresponding gray levels in this paper, is obtained from the point estimate of the pose and the latest input image.

M-PFAP employs two steps to obtain a new appearance $\widehat{A}_T$; interest point detection and depth information extraction. First, interest points are detected

from the input image. Then, corresponding depth values of interest points are extracted from a rough shape model. As the rough shape model, we used a laser-scanned averaged head shape (not a tracked person's model).

M-PFAP stores pose-appearance pairs only when the tracking is stable to prevent erroneous pairs from being stored. The stability of tracking is judged by the maximum likelihood of particles. The maximum likelihood works well in most cases, however, it is not perfect, and erroneous pairs may become held in memory. If, however, the erroneous pairs are in the minority in memory, the stochastic sampling from all past memory yields few erroneous samples and the majority of samples are valid. This condition ensures that M-PFAP does not suffer explosive error growth, which is a serious weakness of the traditional template update scheme.

Additionally, to suppress memory usage and retrieval time, M-PFAP employs the data structuring process. It stores a new pose-appearance pair only when there are no pairs whose pose are very similar to the new pose.

5 Experiments and Results

Experiments in this paper targets face pose. This section describes the experimental environment, the details of the experiments, and the results.

5.1 Experimental Environment

We used PointGreyResearch's FLEA, a digital color camera, to capture 1024 × 768 pixel-size images at 30 frames per second. The tracking processes use only gray images converted from color images. A magnetic-based sensor, Polhemus FASTRAK was used to obtain quantitative ground truth data. The rotation angles, pitch, roll, and yaw roughly correspond to shaking, nodding, and tilting actions, respectively. As shown in Fig.5, two sensors were attached to both temples of the subject. The number of particles was set to 2000.

Table 1 summarizes the proposed method and baseline methods. We employed three baseline methods, all based on M-PF; the first one is the original M-PF, it uses only one template without updating; the second one (LT) updates the template and uses the latest template; the last one (NN) updates the template and uses the template nearest to the target pose.

Table 1. Comparison between proposed method and comparative methods

	template updating	selection criterion of templates
ProposediM-PFAPj	accumulating history	Probabilistic selection
M-PF	No	initial template
LT	Yes	latest template
NN	Yes	nearest template

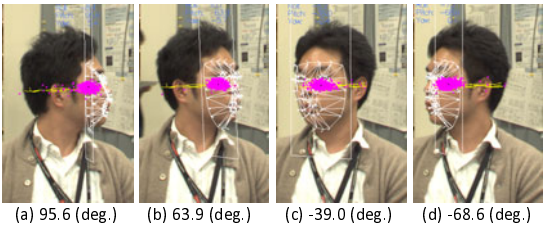


Fig. 6. An example of face tracking; the proposed method can track against large appearance changes

5.2 The Effective Tracking by the M-PFAP

To verify the effectiveness of M-PFAP, we used a test video sequence that included a head that rotated from frontal view to profile view (=90 deg. in horizontal direction). The target video includes profile faces. Figure 6 shows the result of M-PFAP. The snapshots in Fig.6 are listed in time order from left to right. In Fig.6, the white mesh represents the estimated position and rotation, and the dots located around center of the face denote the prior distribution of face pose, which only represent positions. The initial template surface is almost invisible in profile view as in Fig.6 (a) and (d) ; old trackers that use only frontal view template cannot track the face anymore. In contrast, M-PFAP successfully tracked the face in profile view.

5.3 Quantitative Evaluation of Tracking Accuracy

Three types of video were employed for this quantitative evaluation. Video-1 included a wide range of moderate rotations. This video was used for evaluating basic performance. Video-2 included abrupt movements, such as abrupt reverse of moving direction and abrupt shaking of the head. This video is used to verify that M-PFAP mirrors the robustness against such abrupt motion of M-PF. Video-3 included occlusions such as the rotating head being hidden by a moving arm. Occlusion recovery is another merit of M-PFAP inherited from M-PF.

Fig.7 shows the tracking results of Video-1. The horizontal axis denotes time and the vertical axis denotes horizontal rotation angle. Figure 7 shows that the M-PFAP output closely followed the ground truth. M-PF, on the other hand, became unstable when the rotation angle exceeds about 60 degrees. NN and LT could not track correctly; they had worse performance than M-PF. We consider that the updated template included errors and so the tracking drifted.

Fig.8 shows snapshots during tracking of the target face moved from left to right abruptly (Video-2). The snapshots are listed from left to right in time order. It was tracked correctly. Absolute average errors of face pose tracking against Video-1 and Video-2 and variances are shown in Table 2. The proposed method yielded improved tracking performance in both videos.

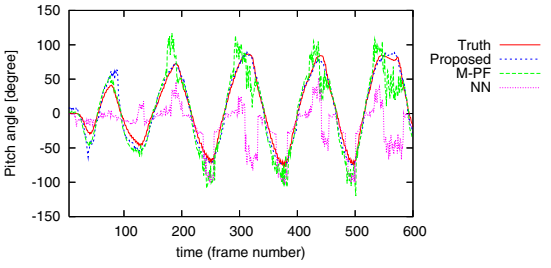


Fig. 7. Head rotation angle in horizontal direction

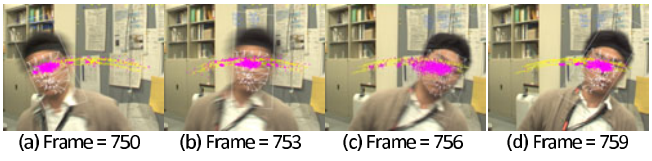


Fig. 8. Snapshots of tracking the abrupt movements

Video-3 included occlusions. Snapshots of Video-3 during tracking are shown in Fig. 9. In this scene, the face turns from right to left, at the same time, an arm moves from top to bottom causing an occlusion; the face turns and shifts during the occlusion; so the face poses before and after occlusion are completely different. Additionally, the profile face appears immediately after the occlusion; it can not be tracked by the initial template. M-PFAP could recover tracking even in this severe situation.

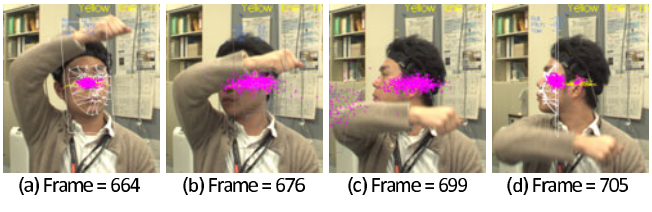


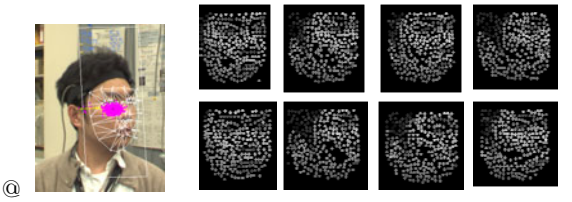
Fig. 9. Snapshots of occlusion recovery

5.4 Past Appearance Used for Appearance Prediction

Fig.10(b) shows the history entries that were selected to estimate the pose prior distribution of Fig.10(a). It can be observed that many entries were used for appearance prediction. Each entry includes error to some extent. By using a number of entries to estimate appearance, M-PFAP prevents the tracking from accumulating errors and from drifting.

Table 2. Absolute average errors [degree] in horizontal rotation; values in blacket show corresponding variances

	Proposed	M-PF	NN	LT
Video-1	7.1 (35.0)	15.9 (182.7)	35.0 (1316.5)	38.7 (1448.1)
Video-2	8.5 (108.8)	14.3 (117.0)	16.7 (684.7)	31.2 (698.0)



(a) Target face (b) Past appearance used for prediction

Fig. 10. Pose estimation target and past appearance used for the appearance prior prediction

6 Summary and Future Works

This paper proposed M-PFAP; it offers robust visual tracking of the target's position and pose. M-PFAP is an extension of M-PF and represents a unified probabilistic framework for the joint estimation of target position/pose and its appearance based on memory-based joint prior prediction using stored past pose and appearance sequences. Quantitative evaluations confirmed that M-PFAP successfully tracks human faces in frontal view up to profile view, i.e. 90 degree horizontal rotation; it thus completely overcomes the limitation of M-PF.

Future works include the following two points. First, we consider how to handle appearance change due to illumination change. Among the various illumination changes, the current implementation of M-PFAP realizes robustness against uniform illumination change since the state vector employs an illumination coefficient. Also, M-PFAP potentially can handle non-uniform illumination change by accumulating pose-appearance pairs under gradual changes in illumination. We are going to evaluate the limits of robustness against various illumination conditions and achieve further robustness.

Second, we will tackle GPU implementation. Our current CPU-based M-PFAP does not work in real-time. We consider that M-PFAP suits GPU acceleration, because it is an extension of M-PF and supports parallel processing as does M-PF. M-PF processing was made 10 times faster by GPU implementation. For the GPU implementation, more effective way of storing memory to save memory usage and to save retrieval time should be considered.

References

1. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. PAMI 25, 564–577 (2003)
2. Bradski, G.R.: Computer vision face tracking for use in a perceptual user interface. In: Proc. IEEE Workshop Applications of Computer Vision, pp. 214–219 (1998)
3. Tua, J., Taob, H., Huang, T.: Face as mouse through visual face tracking. CVIU 108, 35–40 (2007)
4. Otsuka, K., Araki, S., Ishizuka, K., Fujimoto, M., Heinrich, M., Yamato, J.: A realtime multimodal system for analyzing group meeting by combining face pose tracking and speaker diarization. In: Proc. ACM ICMI, pp. 257–264 (2008)
5. Gordon, N., Salmond, D., Smith, A.F.M.: Novel approach to non-linear and non-Gaussian Bayesian state estimeation. IEE Proc.F:Communications Rader and Signal Processing 140, 107–113 (1993)
6. Mikami, D., Otsuka, K., Yamato, J.: Memory-based particle filter for face pose tracking robust under complex dynamics. In: Proc. CVPR, pp. 999–1006 (2009)
7. Papanikolopoulos, N., Kosla, P., Kanade, T.: Visual tracking of a moving target by a camera mounted on a robot: A combination of control and vision. IEEE Trans. Robot. Autom. 9, 14–35 (1993)
8. Black, M., Yacoob, Y.: Recognizing facial expressions in image sequence using local parameterized models of image motion. IJCV 25, 23–48 (1997)
9. Matthews, I., Ishikawa, T., Baker, S.: The template update problem. IEEE Trans. PAMI 26, 810–815 (2004)
10. Jepson, A.D., Fleet, J.D., El-Margaghi, T.F.: Robust online appearance models for visual tracking. IEEE Trans. PAMI 25, 1296–1311 (2003)
11. Zelniker, E.E., Hospedales, T.M., Gong, S., Xiang, T.: A unified bayesian framework for adaptive visual tracking. In: Proc. BMVC, pp. 100–200 (2009)
12. Morency, L.P., Rahimi, A., Darrell, T.: Adaptive view-based appearance models. In: Proc. CVPR, pp. 803–810 (2003)
13. Morency, L.P., Whitehill, J., Movellan, J.: Monocular head pose estimation using generalized adaptive view-based appearance model. Image and Vision Computing (2009), doi:10.1016/j.imavis.2009.08.004
14. Ross, D., Lim, J., Lin, R.-S., Yang, M.-H.: Incremental learning for robust visual tracking. IJCV 77(1-3), 125–141 (2008), doi:10.1007/s11263-007-0075-7
15. Lefèvre, S., Odobez, J.: View-based appearance model online learning for 3d deformable face tracking. In: Proc. VISAPP (2010)
16. Lozano, O.M., Otsuka, K.: Real-time visual tracker by stream processing. Journal of VLSI Signal Processing Systems (2008), doi:10.1007/s11265-008-0250-2