

# Using Stochastic Gradient-Descent Scheme in Appearance Model Based Face Tracking

Zhidong Li<sup>#\*1</sup>, Jing Chen<sup>#\*2</sup>, Adrian Chong<sup>\*3</sup>, Zhenghua Yu<sup>#4</sup>, Nicol N. Schraudolph<sup>#5</sup>

<sup>#</sup> Program of Making Sense of Data, NAL Lab, NICTA  
Level 4, 223 Anzac Parade, Kensington, NSW, 2051, Australia

<sup>1</sup> Zhidong.Li@nicta.com.au

<sup>2</sup> Jing.Chen@nicta.com.au

<sup>\*</sup> School of Computer Science and Engineering,  
the University of New South Wales, Australia

**Abstract**— Active appearance model (AAM) has been widely used in face tracking and recognition. However, accuracy and efficiency are always two main challenges with the AAM search. The paper therefore proposes a fast appearance-model based 3D face tracking algorithm to track a face appearance with significant translation, rotation, and scaling activities by using stochastic meta-descent (SMD) optimization scheme to accelerate the appearance model search and to improve the tracking efficiency and accuracy. The proposed algorithm constructs an active face appearance model by using several semantic landmark points extracted from each frame and then processes the appearance model search to approximate the model translating, rotating, and scaling by using the SMD filter to minimize the appearance difference between the current model and the new observation. We compare the results with both a conventional AAM and a Camshift filter and find that our algorithm outperforms both two in terms of efficiency and accuracy in tracking a fast moving, rotating, and scaling face object in a video sequence.

## I. INTRODUCTION

Tracking human body, hand, and face has captured huge research interests in the past decades and been widely used in human behaviour analysis, face recognition, and human-computer interaction etc. However, capability of fast and robust tracking at 3D human body and its components with significant appearance variation is still an ongoing research topic [1, 4, 7]. Appearance model, on the other hand, is an active approach applied in the model-based object tracking especially in tracking pedestrian, face, and gesture etc. The appearance-model approach uses a parameterised model to approximate the object appearance and controls the model translating, rotating, and scaling activities by updating those model parameters [2, 11-12]. However, the efficiency and accuracy are always two trade-off issues and are considered as main benchmarks in evaluating the performance of an appearance model based object tracking.

An appearance model usually comprises of two components, i.e. constructing a parameterised appearance model, and then updating the model parameters to make the model approaching to the new observation [2]. When constructing an appearance model, it usually requires that the constructed model shall be able to represent and describe the

appearance information including location and orientation of the object well by using limited number of parameters to approximate the object appearance without losing key appearance information. This requires that the features used in constructing the model shall be rich and accurate enough in describing the object appearance. For instance, a human face is usually modelled and represented by those landmark feature points such as eyes, nose, mouth, and ears etc that can describe the motion, orientation, and even the expression of a human face well. Generally, features used to model an object appearance include the low-level features, e.g. colour, edge, and shape etc. [11], the invariant features such as the scale-invariant feature transform (SIFT) etc [12], and the semantic features such as eye and mouth features of a face which required certain prior training process. On the other hand, given an object appearance model and a new observation, it also requires that the model can be efficiently manipulated by iteratively adjusting those model parameters to have the model maximally matching the current observation, which can also be understood as tracking the object by minimizing the spatial appearance difference between the landmark feature points of the appearance model and the new observation.

In this paper, aiming at tracking a human face with significant movement and rotation, we propose an appearance-model based face tracking system to track a human face in a 3D space. There are two main contributions from this paper. First, the paper propose a 3D face appearance model by using nine semantic landmark-feature points to represent the pose, orientation, and scale of a face model. Second, the paper introduces the SMD to accelerate the appearance model search and to improve the search accuracy and finally improves the tracking efficiency. We compare the performance of our algorithm with a conventional AAM model and a Camshift filter and find that our proposed algorithm is able to work more robustly, efficiently, and accurately over the other two especially at tracking a face model with high motion and rotation activities.

## II. PREVIOUS WORK

In the past decades, there has been huge amount of research working at model based face tracking. In year 1996, Basu,

Essa and Pentland [1] used an ellipsoid model plus optical flow approach to track head motion in 3D space, but the scheme is concerned by its tracking speed and sensitivity to the noise in the input video sequence. Then, there has been much evolution since that at the model-based

face tracking till the introduction of two typical examples of model-based approaches, i.e. Active Shape Models (ASM) and Active Appearance Models (AAM) developed by Cootes and Taylor [2, 7]. Both ASM and AAM use some low-level features, such as shape, texture, edge, and SIFT etc features to construct the object appearance model [2, 11, 13]. However, the accuracy and search speed are always two main concerns that have been attracting lots of research at improving the efficiency of the ASM search and AAM search [10-13]. Meanwhile, there have been lots of developments in tracking strategy itself, such as Kalman filters, particle filters [3], and mean-shift tracker [4, 9]. Obviously there is a trend of combining an AAM/ASM with a particle filter or a mean-shifter filter but the integrated framework seems can only use some simple low-level features for its appearance model [11]. Credit to its low computational cost and the parameter-free nature, the mean-shift tracker has been widely used in those strict time-constrained applications such as real-time object tracking and image segmentation etc. However, a mean-shift tracker uses simple feature histogram as well as its inputs and is very difficult to handle those complicated and high-dimension landmark features as an appearance-model based approach can offer. Therefore, to compete with the fast convergence speed that a mean-shift tracker can offer while promising its capabilities in tracking objects based on more complex features, improving the efficiency of an appearance model search becomes very critical in advertising these technologies into a practical application.

In 2005, Bray et al. presented an optimization algorithm by using Stochastic Meta-Descent (SMD) that is able to online adjust its iteration step size and thus to achieves a substantially high convergences speed [5-6]. One of the very successful applications of SMD is using SMD in optimizing the gradient-descent based search in tracking the activities of an artificial hand. The application demonstrates the capability of the SMD in optimizing gradient-based search by using more than 100 features points which is really difficult to achieve by other schemes.

In the paper, we construct an ellipsoidal human face appearance model with nine landmark-features points, e.g. eyes points, mouth points, forehead points, and chin point etc, detected from the image; then we introduce the Stochastic Meta-Descent (SMD) scheme into our algorithm to accelerate the appearance-model search aiming to minimize the appearance difference between the current appearance model and the new observation. The SMD itself is a gradient-based online optimization method that can quickly adjust the descent

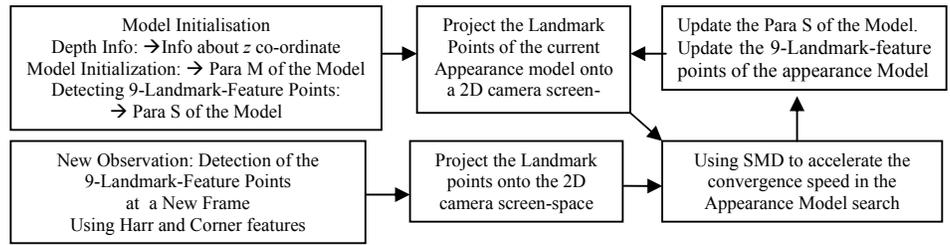


Fig. 1. OUTLINE OF THE PROPOSED FACE TRACKING ALGORITHM

step size according to the gradient distribution and thus achieve a fast convergence. From the final simulation results, we found that the proposed scheme is relatively efficient and accurate in tracking a 3D face with continuous and remarkable translating, rotating, and scaling activities when compared with the conventional AAM and the Camshift algorithm.

The following parts are organized as, we will first introduce the construction of an appearance model based 3D face model, then the SMD scheme will be illustrated in details before introducing how to make it working with an appearance model search. Finally, some testing results will be given followed by the analysis and conclusion.

### III. ALGORITHM DESCRIPTION

The outline of the proposed face-tracking algorithm is shown in Figure 1. The whole algorithm comprises of detecting landmark points, constructing a 3D face appearance model, and on-line approaching the new observation through appearance-model search. The main concepts of the proposed algorithm are to construct a 3D face appearance model by an ellipsoid model with several semantic landmark points and to manipulate the face model through the model parameters. The landmark points are used to determine the location and pose etc of the face appearance model. Given a new observation at the landmark points, we use SMD to accelerate the appearance-model search that aims to minimize the appearance distance between the landmark points of the current appearance model and the observation by adjusting the model parameters. A tracking is thus fulfilled by iteratively and continuously process the appearance-model search and update.

#### A. 3D Face Appearance Model

In the paper, we use an ellipsoidal model that was ever successfully employed by Basu et al in tracking a head with significant movements over a large number of sequences [1]. The ellipsoidal intrinsic model itself is represented by a parameter set  $M$  that indicates the radii of the ellipsoid in the  $x$ ,  $y$ , and  $z$  coordinates accordingly. Given that the model has freedom in translating, rotating, and scaling in a 3D space, we use another varying parameter set  $S$  to denote the location of the ellipsoid centre and the rotations of the ellipsoid in three-dimensional space, i.e.  $x$ ,  $y$ , and  $z$  co-ordinates. Finally, to map each co-ordinate point  $x_m$  residing on the surface of the ellipsoid model to a projected 2D camera-based measurement  $x_c$ , we use a model function  $f$  to denote such a translation

from a real 3D ellipsoid-model space to the projected screen-space that is measured by a single camera,

$$x_c = f(x_m, S, M) \quad (1)$$

To provide the landmark-feature points with freedom in manipulating the face appearance model, in the face appearance model we use up to 9 semantic landmark-feature points to indicate the position of the forehead (2 points), eyes (4 points), mouth (2 points), and chin (1 point), and to represent the model appearance. Thus, problem of tracking at the 3D ellipsoid-model becomes as continuously observing the 9 semantic landmark features in the consecutive image frames and having the model translating, rotating, and scaling in a 3D space to minimize the appearance difference between the current appearance model and the new observation by adjusting the model parameters.

### B. Detection of the 9 landmark-Feature Points

To get the 9 semantic landmark points used to construct the ellipsoid face appearance model, we used the simple Harr and corner features introduced in Viola and Jones algorithm [8] to detect areas containing face, mouth, and eyes. Then, we used 2 points, i.e. left and right corners of the mouth, to represent the position of the mouth; 2 pairs of the inner and outer corners' points of the 2 eyes to represent the positions of the left and right eyes, and 2 forehead points (up-left and up-right points of the detected face rectangle) and one chin point (down-middle point of the detected face rectangle) to denote the scale and position of the face appearance model. Figure 2 shows an example of the 9 landmark points used to represent the 3D face appearance model. This model was found very efficient in manipulate any face motions with low computation cost.

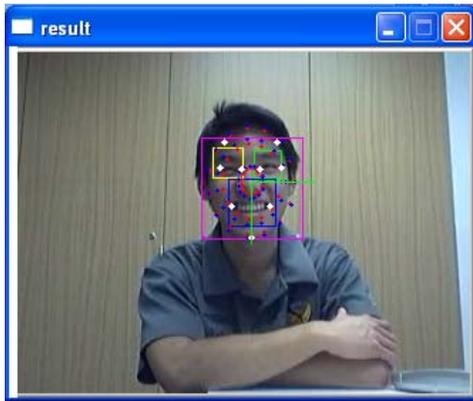


Fig. 2. Ellipsoid face appearance model based on the 9 landmark-feature points (in white colour)

### C. Model Initialization

The initialisation of the model comprises of setting the parameter set  $M$  used to decide the initial model scale, and the parameter set  $S$  used to decide the initial position and orientation of the model. In the paper, we assume that we have some prior knowledge about the initial depth information

between the camera and the centric point of the appearance model so that we are able to setup the model parameters in the  $z$  coordinate (in the future this can be simply implemented by using stereo cameras to obtain the depth information of the initial face appearance model). Once the face appearance model is initially setup, we use the scale information to set the radii parameter of the ellipsoid model in  $x$  and  $y$  coordinates and determine the parameter set  $M$  of the appearance model. Then, the nine landmark-feature points detected from the current frame are used to setup the parameter set  $S$  and to obtain the initial information about the location and orientation of the face appearance model.

### D. Using SMD Optimization in Appearance-Model Search

Problem of face tracking now becomes as simple as detecting those 9 landmark-feature points in each frame followed by processing an appearance-model search to minimize the appearance difference between the current appearance model and the new observations. To speed up such an appearance-model search, the SMD is applied aiming to accelerate the convergence speed in the search iteration.

On contrary to the conventional gradient descent that is slow in convergence and has the shortage of not being able to find the global minimum in a complex function and being easy to diverge to an infinite, the SMD, based on the gradient descent approach, updates the model parameters by adopting online adaptive step variation scheme in each iteration to reduce the required iterations amount which leads to a faster convergence to a global minimum. In SMD, it observes the sign of the autocorrelation of the gradient between adjacent iterations, and then optimizes the iteration step size by decreasing step sizes when the gradient oscillates (i.e. where  $g_i \cdot g_{i-1} < 0$ ) and increasing step sizes when the gradient is consistent. In addition, since step sizes is parameter-specific, all model parameters can therefore reach their optimal independently and individually and thus to accelerate the whole process. Besides, stochastic sub-sampling can also be used to avoid spurious local minimum. That is, by randomly choosing sample model points during iteration, the local minimum will change at iteration to ensure that the algorithm converges towards the global minimum instead of the local one and the model converges towards the correct final position.

The cost (error) function of the SMD seeks to minimise the appearance distance between the current appearance model and the new observation of the new frame. The process is, given a new observation, we project the nine landmark-feature points in 3D space, denoted as  $x_m$ , into a 2D screen by equation (1); then we calculate the observation appearance based on the projection set  $x_c$ . Let  $x_p$  denote the current model appearance, the cost function can be expressed by

$$E = \sum (\|x_p - x_c\|)^2 \quad (2)$$

Where  $\|\cdot\|$  denotes the  $L_2$ -norm. Obviously, the tracking problem here now becomes to minimize the cost function by iteratively manipulating the 3D ellipsoid-model in translation, rotation, and scaling so that the projection of those 3D 9

landmark-feature points on the 2D screen finally introduces a minimal appearance difference between the current appearance model and the new observation. In the simulation, we also found that using the feature points of the eyes and mouth only is very effective to account for the model rotation, while using all 9 feature points is better in model translation.

To calculate the gradient of the cost function  $E$  in the state space of poses  $S$ , let  $J_K$  denote the Jacobian of a function  $K: \mathbb{R}^n \rightarrow \mathbb{R}^m$ , i.e. the  $m \times n$  matrix of partial derivatives of the  $m$  outputs of  $K$  with respect to the  $n$  inputs. The gradient  $\mathbf{g}_i$  of the error function  $E$  at iteration  $i$  is calculated by backwards chaining of derivatives, where  $^T$  denotes the matrix transpose.

$$\mathbf{g}_i = \partial E / \partial \mathbf{s} = \partial E / \partial f(x,y) \cdot \partial f(x,y) / \partial \mathbf{s} = \mathbf{J}_E^T \mathbf{J}_F^T \mathbf{J}_{(x,y)}^T \quad (3)$$

Convergence of gradient descent to the minimum can be accelerated significantly by scaling the gradient of each parameter using its own step size. The model parameter set  $S$  is then updated via equation (4), where  $i$  denotes the  $i^{\text{th}}$  iteration and  $\mathbf{a}$  is the step size.

$$\mathbf{s}_{i+1} = \mathbf{s}_i - \mathbf{a}_i \cdot \mathbf{g}_i \quad (4)$$

The step size of each individual parameter is updated by observing the autocorrelation of the gradient as mentioned before. Step sizes shall remain positive to ensure the gradient descent approach converging to the minimum, and also shall incorporate the history of the gradient autocorrelation to achieve minimal computational cost. Therefore, the step sizes shall be updated via the following formula,

$$\mathbf{a}_i = \mathbf{a}_{i-1} \cdot \max(1/2, 1 + \mu \mathbf{v}_i \cdot \mathbf{g}_i) \quad (5)$$

Where  $\mu$  is a scalar meta-step size for the adaptation of  $\mathbf{a}$ ,  $\mathbf{v}_i$  represents the effect of all past step sizes on the new parameter values and is calculated via,

$$\mathbf{v}_{i+1} = \lambda \mathbf{v}_i + \mathbf{a}_i \cdot (\mathbf{g}_i - \lambda \mathbf{H}_i \mathbf{v}_i) \quad (6)$$

where  $0 \leq \lambda \leq 1$  governs the time scale over which the step size history is taken into account, and  $\mathbf{H}_i$  denotes the instantaneous Hessian at iteration  $i$ .

Therefore, in iteration, the SMD optimisation calculates the gradient descent, updates gradient step sizes; updates the model parameter set  $S$ ; and finally updates SMD  $\mathbf{v}$  vector. After iteration, the updated model parameter set  $S$  indicates the new location and orientation of the appearance model and gradually the appearance model will converge towards the new observation. By repeating such process in each new frame, the face model can be continuously updated and the face model can be thus continuously tracked.

#### IV. EXPERIMENTAL RESULTS AND ANALYSIS

To evaluate the efficiency and accuracy of the proposed algorithm, we run the proposed algorithm, a conventional AAM, and a Camshift algorithm referred by the OpenCV [10] on the same testing video sequences. All testing sequences are in 320x240 resolution size and 15 frames/sec. The initial SMD parameters are set as  $\mathbf{a}_0 = \{10, 20, 30, 40, 50, 60\}$ ,  $\mu=0.05$ , and  $\lambda=0$ . We used HSV (4:2:2) in measuring the feature value of each individual landmark-feature point in the appearance

model. Figure 3 shows several comparisons between our proposed scheme and the Camshift tracker in terms of tracking accuracy at kinds of face motions such as face translation (movement without rotation), rotation, translation plus rotation, and scaling up and down. All testing results shows that the Camshift tracker gets lost easily in the testing scenarios especially for those containing significant rotation and motion, while our proposed algorithm robustly has a better tracking accuracy with translation being extremely accurate. The rotation values of the proposed algorithm were accurate as well even when combined with a remarkable scaling factor. Another testing was conducted by comparing the speed of appearance model search between a conventional AAM and our proposed scheme, of which results are list in the Table I. The measurement figures in the Table I are the mean values of the testing results from all video sequences covering all intensity levels of translation, rotation, and scaling. In addition to that our proposed algorithm can speed up the convergence, we also find that the conventional AAM search is easy to either diverge to infinite or converge to a local minimum when the new observation is far from its current appearance model due to lack of adaptation in its step size during iteration. Therefore, introducing the SMD into the appearance-model search can achieve a better convergence not only in its speed and but also in its accuracy when compared with a conventional AAM search. (Note: video of testing results will be provided upon request)

TABLE I  
COMPARISON OF MODEL SEARCH SPEED BETWEEN A CONVENTIONAL AAM AND OUR PROPOSED ALGORITHM (MAXIMAL ITERATION NUMBER IS SET TO 100 FOR EACH SEARCH)

Methods	Mean value of the Iteration amounts used to get convergence for face tracking			
	Translation	Rotation	Scaling	Translation +Rotation
Conventional AAM	21.81	8.25	38.81	Diverge to Infinite
our proposed scheme	13.81	5	20	81.8

#### V. CONCLUSION AND FUTURE WORKS

The paper proposed a 3D appearance-model based face tracking algorithm by using semantic landmark-feature points to construct a face appearance model and using stochastic meta-descent (SMD) optimization scheme to accelerate the appearance model search, aiming to improve both efficiency and accuracy of the face tracking. The testing results demonstrate our original expectation to some extends by comparing with some of the state of the arts.

While the proposed algorithm works quite efficiently in tracking a single face object by constructing a single face appearance model, however, we also noticed that there is still

some work ahead if employing our proposed algorithm in a multi-faces tracking case, especially in the case when two face models occlude each other which may break the grouping of the landmark-feature points. This will be one of topics of our future research.

#### ACKNOWLEDGMENT

National ICT Australia (NICTA) is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

#### REFERENCES

- [1] S. Basu, I. Essa, and A. Pentland, "Motion regularization for model-based head tracking," ICPR Intl. Conf. Pattern Recognition, Vienna, Austria, August 25-30 1996.
- [2] T.F. Cootes, and C.J. Taylor, "Statistical models of appearance for medical image analysis and computer vision," SPIE Proc. Medical Imaging, 2001.
- [3] M. Isard and A. Blake, "Condensation-Conditional Density Propagation for Visual Tracking," International Journal of Computer Vision, vol.29, no.1, pp.5-28, 1998.
- [4] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," IEEE Trans. on Pattern Analysis and Machine Intelligence, pages 564-577, May 2003.
- [5] M. Bray, E. Koller-Meier, P. Müller, N. N. Schraudolph, and L. Van Gool, "Stochastic optimisation for high-dimensional tracking in dense range maps," IEE Proc. Vision Image Signal Processing 152(4):501-512, August 2005.
- [6] M. Bray, E. Koller-Meier, N. N. Schraudolph, and L. Van Gool, "Fast stochastic optimisation for articulated structure tracking," Image Vision Computing 25(3):352-364, March 2007.
- [7] L. Brown, "3D head tracking using motion adaptive texture-mapping," IEEE Proc.Intl. Conf. Computer Vision and Pattern Recognition, 2001.
- [8] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features," IEEE Proc. Intl. Conf. Computer Vision and Pattern Recognition, 2001.
- [9] R. T. Collins, "Mean-shift blob tracking through scale space," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2003.
- [10] Intel OpenCV library <http://www.sourceforge.net/projects/opencvlibrary>.
- [11] Swaminathan, G., Venkoparao, V., Bedros, S, "Multiple appearance models for face tracking in surveillance videos," AVSS 2007 IEEE Conference on, 5-7 Sep 2007.
- [12] Liang Liu; Yunhong Wang; Tieniu Tan, "Online Appearance Model Learning for Video-Based Face Recognition," CVPR 2007 IEEE Conference on, 17-22 June 2007.
- [13] Daijin Kim; Jaewon Sung, "A Real-Time Face Tracking using the Stereo Active Appearance Model," Image Processing 2006 IEEE International Conference on, 8-11 Oct 2006.

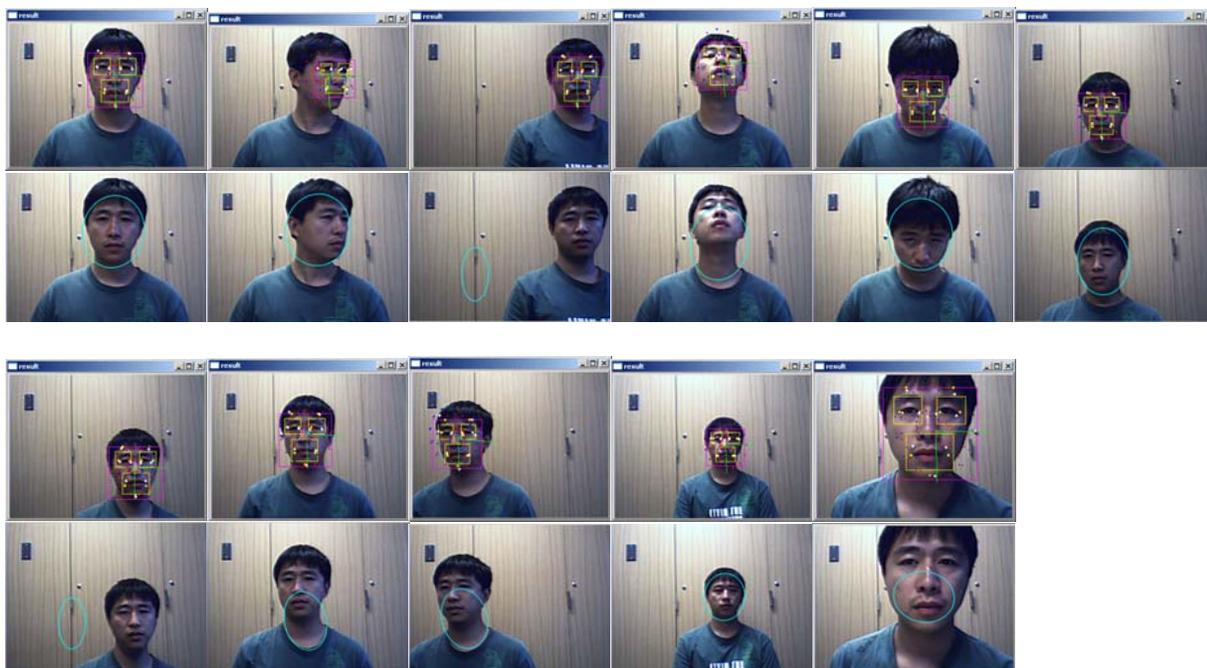


Fig. 3. Comparisons between the proposed algorithm and the Camshift in terms of the accuracy when tracking a 3D face with significant translating, rotating, and scaling. The first/third rows are results from the proposed algorithm while the second/fourth rows are results from the Camshift filter. It can be seen that the proposed algorithm can track the moving/rotating/scaling face more stably and accurately (processing frame rate = 15fps) than a Camshift filter does.