DYNAMIC FACE APPEARANCE MODELING AND SIGHT DIRECTION ESTIMATION BASED ON LOCAL REGION TRACKING AND SCALE-SPACE TOPO-REPRESENTION

Shaun J. Canavan and Lijun Yin
Department of Computer Science, State University of New York at Binghamton, Binghamton, NY.

ABSTRACT

Dynamic modeling of facial appearances and sight directions are demanded for HCI and multimedia applications. Traditional approaches for face tracking and eye tracking from 2D videos do not involve explicit facial modeling. In this paper, we propose to use an explicit 3D model to model the dynamic facial appearance as well as the eye shape to estimate the viewing direction. We apply active appearance models for local region tracking, and use a scale-space topographic representation for frame model instantiation. The individualized 3D models across video sequences allow us to estimate the iris viewing orientation dynamically. The proposed framework has been realized and tested in a person-independent fashion for AAM tracking and model instantiation using a single camera.

1. INTRODUCTION

Facial appearances and sight orientations can be modeled in a 3D space. Existing 3D dynamic imaging systems [6] [13] require a rigorous setup (e.g., short range of capture, user intervention for calibration of multiple cameras, lengthy pose-processing, and strict user cooperation, etc.), thus limiting their applications for human computer interaction. In this paper, we present a system to model facial dynamic appearance and eye sight direction using a single video camera. We create dynamic 3D models from tracking information obtained from active appearance models and scale-space topographic features, and map them to a 3D space to create a 3D representation for each frame of a face video. We model both the 3D facial region and 3D iris region dynamically and explicitly, allowing an accurate estimation of eye sight directions through a dynamic video. The system framework is outlined in Figure 1.

To model a face and its iris in a dynamic 3D space, feature tracking is the first step needed for the topographic model creation. Our system allows the user to either track the entire face or the subject's eye region separately. The user can select which model they would like to use (face or eye). Here we use an active appearance model (AAM) [2] to track 459 feature points which are defined in the facial region. Since the subsequent eye modeling requires a multi-scale space topographic representation and multi-size surface patch fitting for topographic label classification, we need to restrict the region of interest for efficient computation. Therefore we further track 8 landmarks to determine the region of interest for eyes.

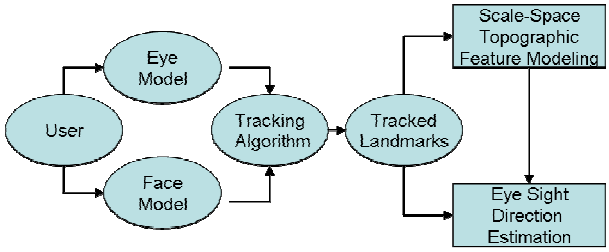

Figure 1: system diagram

Extended from our previous work on topographic analysis for facial feature modeling [11] [12], we propose a new scale-space topographic feature representation approach to model the dynamic facial appearance and iris sphere explicitly. We use a 3D geometric reference model (including a 3D facial surface mesh and a 3D eye mesh) to model individual faces and individual eyes. A multi-step dynamic mesh adaptation method is applied on both facial regions and eye regions to instantiate the model across video sequences. Note that unlike the conventional methods [4] [1][5][7][8] for eye tracking and eye gaze estimation, which have used 2D holistic based approaches or local component based approaches, we estimate the eye viewing direction through the explicit 3D iris modeling. This allows for more flexible and reliable eye sight detection under various poses, expressions, and imaging conditions. The rest of the paper will describe the components for tracking and modeling separately.

2. TRACKING WITH PERSON-INDEPENDENT AAM

Active appearance models were introduced by Cootes *et al.* [2]. It consists of two separate types of models; one is the variation of the face shape, the other is the variation of the gray level of that shape. These two models are combined together to create a statistical appearance model. During the training phase the user manually selects landmarks that correspond to the most important features on each of the images that will be used for training. After the landmarks are

selected each of the landmarks from the images in the training set are warped to match the mean shape. Each set of landmarks are represented as a vector and PCA is applied to them. This can be approximated by the following formulas: $x = \bar{x} + Q_s c_s$ for shape and $g = \bar{g} + Q_g c_g$ for texture. In the shape formula $\bar{x}$ is the mean shape, Q_s represents the modes of variation and c_s defines the shape parameters. In the texture formula $\bar{g}$ is the mean gray level. Q_g represents the modes of variation and c_g defines the texture parameters. In various works pertaining to active appearance models 95% - 98% of the variance is usually kept. To conduct our experiments we chose to retain 95% of the variance.

To track the entire face 459 landmark points are used that cover the entire face (Figure 2 (b)). To create a training model where each image contained 459 landmarks would be a cumbersome and time consuming process. To alleviate this challenge we select 92 key points in each of the training set images (Figure 2 (a)). We then interpolate to the required 459 points to track and eventually create the 3D model. The interpolation is done using a Catmull-Rom spline.

(a) (b) (c)

Figure 2: (a) original 92 key points; (b) interpolated 459 points; (c) 8 points selected for eye region

To track the eye region the model consists of 8 key points around the eyes (Figure 2 (c)). The points create a "boxed in" region around both of the eyes. This region allows us to set the ROI for a separate eye tracking and eye model creation.

3. SCALE-SPACE TOPOGRAPHIC 3D MODELING

3.1. Dynamic 3D appearance face modeling

Given the feature points tracked, we apply a reference model to align with the tracked points. However, in order to create a 3D model representation for each individual frame, and to estimate the eye sight orientation, we deform the reference model into the non-rigid (non-feature) regions of the face. To do so, we extend our previous work based on an adaptive mesh [12] to a hierarchical topographic scale-space. Here we used three-levels of topographic representations with coarse, medium, and fine structures respectively.

We treat a face image as a topographic terrain surface, and each pixel can be categorized into one of the twelve primitive surface features[10]. The composition of these basic primitives provides a fundamental representation of different skin surface details. Based on the topographic primal sketch [10] we have developed a topographic face labeling approach to represent and model facial surfaces, and created individual face models by adjusting a generic model [12]. Here is the brief overview of our existing approach. Given an input image, we can determine the topographic feature on each pixel location using a surface patch approximation approach [10]. A continuous surface *f(x,y)* is used to fit the local *N* by *N* patch centered at *(x,y)* with the least square error. The topographic label is classified according to the extrema values of the second directional derivative of the surface. After obtaining the first-order and second order derivatives at (x; y), we can construct a 2 by 2 Hessian matrix [10]. The feature labeling is based on the values of eigenvalues and eigenvectors, and the gradient magnitude [12].

The results of topographic labeling represent different levels of feature details, depending on the variance of the Gaussian smoothing function ($\boldsymbol{\sigma}$) and the fitting polynomial patch size (**N**) (both $\boldsymbol{\sigma}$ and **N** are known as *scales*). The topographic label map associated with the scales is defined as *topographic scale-space*. The existing applications of topographic analysis are limited in a "still" topographic map with a selected scale. As we know, every label may represent various features in a specific image. Various features (e.g., features of the human face) may be "screened out" with various "optimal" scales. A small scale could produce too much noise or fake features. A large scale may cause the loss of important features. Our previous work also shows that too many fake features could cause the model adaptation to be distracted. More seriously, it could make the adaptation unstable, even causing it to not converge. Too few features will not attract the generic model into the local facial region with expected accuracy. Due to the difficulty to select an "optimal" scale, here we propose to represent the facial features in the topographic scale space, and modeling faces in a hierarchical structure from a coarse level, to a medium level, and a fine level. Such a procedure will ensure the stable convergence of the dynamic mesh to the face region with a constraint of the upper level topographic space, thus resulting in an accurate estimation of 3D facial appearances and their sight directions.

In our modeling process in the topographic scale-space domain, the dynamic meshes are moved by not only the 2-D external force (*e.g.* topographic gradient) but also the depth force (*e.g.* topographic curvature) for model deformation in *multiple scales*. Here we take the model as a dynamic structure in which the elastic meshes are constructed from nodes connected by springs. The external forces of the nodes are used to link the dynamic mesh to the observed face image data. The motion for the dynamic node system is formulated by a second-order differential equation [9], where the node motion is driven by both internal force (e.g., mesh spring stiffness and topographic gradients) and the external force (e.g., topographic curvature and the topographic labels.)

The model adaptation process is performed by three stages: a coarse adaptation onto the coarse scale of the topographic map, a medium scale adaptation onto a medium topographic map, and a fine adaptation onto the fine scale of the topographic map. The three stages employ the similar adaptation algorithm as described in [12], except for additional constraints assigned to each level of adaptation. Specifically, the second stage (medium level) requires the node motion in the restricted local topographic region which has been defined by the coarse topographic map, and the node motion for the fine level adaptation is restricted in the regions which have been defined in the medium topographic map. This strategy will prevent the mesh from distraction, and thus result in a stable adaptation. As a result, the mesh can distribute itself in both salient feature areas and facial surface "wave" areas.

3.2. 3D Iris modeling and 3D sight direction estimation

Extending the topographic analysis of face features, we applied a scale-space topographic context to conduct an eye model adaptation within the eye region. The procedure is the same as the face model creation procedure as described in Section 3.1. After mapping a model onto the eye region, we can project a ray from the center of the eyeball sphere to the iris center to estimate the eye sight direction. The two 3-D points: centre of eye-ball (*Pb*) and centre of pupil (*Pc*) are illustrated in Figure 3 (upper row). The line linking the two points represents the direction of the eye sight. Note that given the four 3D points obtained from two eye-corners, pupil center (*Pc*), and an arbitrary point on the iris boundary, the eye-ball sphere parameters, center *Pb* and radius *r*, can be uniquely determined.

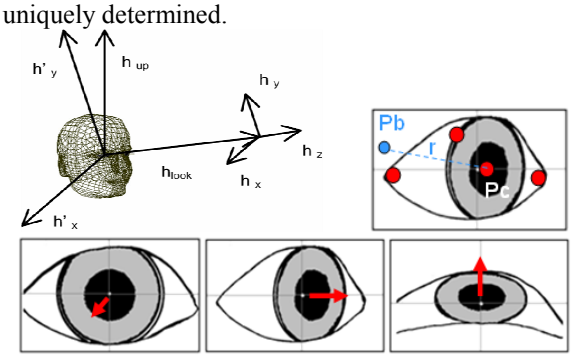

Figure 3: Upper row: eye sight direction and the eye-ball sphere determination based on four points; Lower row: eye sight directions in different views.

4. EXPERIMENTAL RESULTS

In order to test the accuracy of our system we used three different cameras with different resolutions to capture and track our data. We tested our system with a low, medium, and high resolution setting. For our low resolution tests we used a Logitech QuickCam Orbit AF with a resolution of 320x240 (as shown in Figure 5). We used a Sony network camera SNC-RZ30N with a resolution of 640x480 for our medium resolution tests (as shown in Figure 6). Finally, for our high resolution tests we used a Di3D [6] capturing system that creates texture images with a resolution of 1040x1392 (as shown in Figure 4).

Although the eye region is contained in the entire face, we found that it is beneficial for us to track the eye region separately. Since the only information that we need is where the landmarks are located, we have found it easier to only select the landmarks around the eyes instead of extracting this information from the face. Also, there are instances where we found it difficult to successfully track a subject's face but we were able to track the eye region. We believe that this is due to our use of a person-independent active appearance model. Gross *et al.* [3] noted that it is harder to fit a generic AAM compared to a person specific AAM due to the high dimensionality of the shape model.

4.1. Evaluations

In order to evaluate the accuracy of the geometric shape of our created models, we used the 3D dynamic range scans [13] captured from Di3D imaging system [6] as the ground-true data for comparison (Figure 4).

The ground true face model contains 35,000 vertices; our created model has about 2,900 vertices. We used both 3D range model scans and our generated models (300 frame models), and manually selected 92 feature points on each model in areas of mouth, facial contour, nose sides, nose bridge, eyes, eyebrows and cheek. After normalizing all the models into a range of (-50, +50) in three coordinates of x, y and z, we calculate the mean square error (MSE) between the two sets of 3D surface feature points. The result shows that the average MSE of 300 frames models is 6.74. This is much less than the MSE (=12.7) when we compare the coarse models to the range models. In addition, the estimated eye directions from our generated models are also compared to the eye directions of the range models. Among 300 frames, 249 frames show less than 5 degree difference between two data sets.

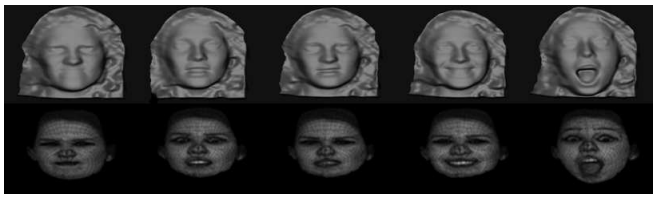

Fig 4: High-resolution video example. Upper: range scans as ground-truth; Lower: our generated models (3D meshes overlapped on textures).

There are three major advantages of the proposed 3D model based approach: (1) the modeling procedure relies on the multiple-scale model adaptation in a global face space rather than very few individual points in local facial regions. It is more resistant to image noises under various imaging conditions; (2) the three-levels of topographic features allow the face and eye representations in a high level of detail, and (3) the eye sphere estimation is based on the four points including the eye center and eye corners and excluding the

eyelid points. It has certain robustness to occlusion from eyelids. Unlike other conventional 2D tracking systems, our 3D model based eye sight estimation does not require any calibration of cameras.

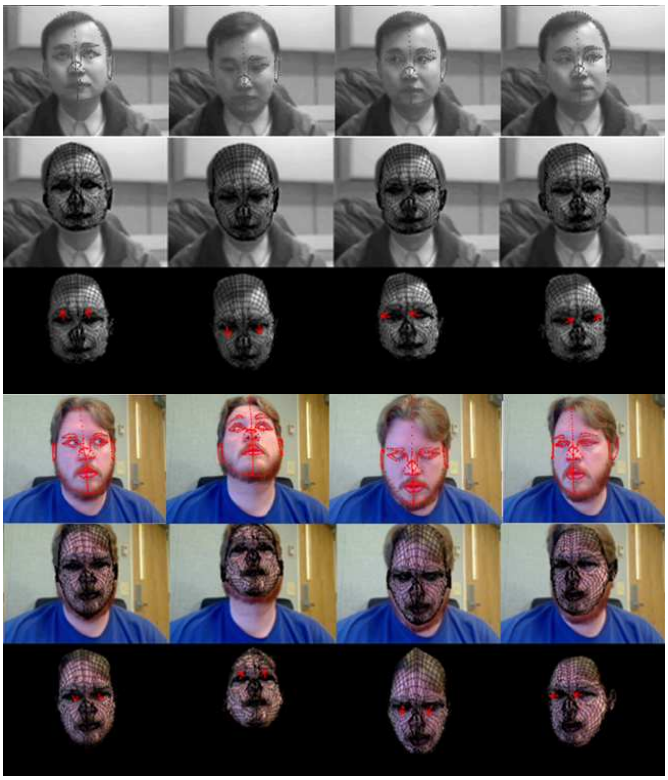

Fig 5: Low-resolution videos for two subjects. Upper three rows: tracked feature points; generated models; and detected eye sight directions (shown as red arrows). Lower four rows illustrate the results of a second subject.

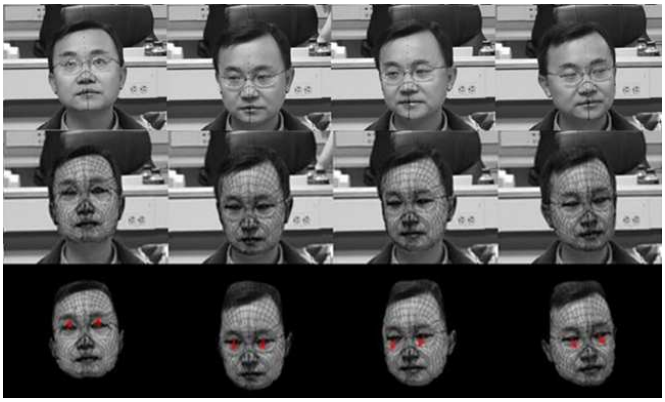

Fig 6: Fig 4: Medium resolution videos for one subject. From top to bottom: tracked feature points; generated models; and detected eye sight directions (shown as red arrows).

5. CONCLUSION AND FUTURE WORK

We have presented a scale-space topographic modeling approach to model the dynamic facial appearance and eye sight directions. The experimental results are encouraging. While we are able to track face movements and eye sight orientations under various resolutions, backgrounds and expressions, the tested pose changes are still in a small range. Our future work is to improve the tracking algorithm in order to handle the case of larger pose changes. For example, one method is to extend our tracking system by including multiple views of faces. We will further evaluate the performance by comparing our approach to the other existing 2D based approaches. In addition, we will further develop a variable mesh resolution approach with a smaller number of tracking points in order to realize a real time application.

6. ACKNOWLEDGEMENT

This material is based upon the work supported by the NSF under grants IIS-0541044, IIS-0414029, Air Force Research Lab, and the NYSTAR's James D. Watson Investigator Program

7. REFERENCES

[1] S. Amarnag, R. S. Kumaran, and J. N. Gowdy, "Real time eye tracking for human computer interfaces", *Proc. of IEEE International Conference on Multimedia and Expo*, 2003.

[2] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23 no. 6. pp. 681-685, June 2001.

[3] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models", *Image and Vision Computing*, vol. 23, no. 11, pp. 1080-1093, 2005.

[4] T. Ishikawa, S. Baker, I. Matthews, T. Kanade, "Passive driver gaze tracking with active appearance models", *Proceedings of the 11th World Congress on Intelligent Transportation Systems*, 2004.

[5] Q. Ji, H. Wechsler, A. Duchowski, and M. Flickner, "Eye detection and tracking", *Computer Vision and Image Understanding*, 98(1), 2005.

[6] Di3D Inc. www.di3d.com

[7] J. Magee, M. Scott, B. Waber, and M. Betke, "EyeKeys: a real-time vision interface on gaze detection from a low-grade video camera", *Computer Vision and Pattern Recognition Workshop*, 2004.

[8] T. Takegami, T. Gotoh, S. Kagei, and R. Minamikawa, "A Hough based eye direction detection algorithm without on-cite calibration". *IEEE Trans. on PAMI*, 20(10): 2001.

[9] D. Terzopoulos and K.Waters. Analysis-synthesis of face image seq. using physical-anatomical models. *IEEE Trans. PAMI*, 1993.

[10] O. Trier, T. Text, and A.K. Jain. Recognition of digits in hydrographic maps: binary vs topographic analysis. *IEEE Trans. on PAMI*, 19(4), 1997.

[11] J. Wang, L. Yin, and J. Moore, "Using geometric property of topographic manifold to detect and track eyes for human computer interaction", *ACM Trans. on Multimedia Computing, Communication Applications*, 3(4): 1-19, 2007.

[12] L. Yin and K. Weiss. Generating 3d views of facial expressions from frontal face video based on topographic analysis. In *ACM Multimedia 2004*, p360-363.

[13] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, A high-resolution 3D dynamic facial expression database, *IEEE International Conference on Face and Gesture Recognition*, 2008.