Avatar-mediated Face Tracking and Lip Reading for Human Computer Interaction

Xiaozhou Wei Lijun Yin
Department of Computer Science
SUNY at Binghamton
Binghamton, NY 13902
xwei, lijun@cs.binghamton.com

Zhiwei Zhu Qiang Ji
Department of ECSE
Rensselaer Polytech Institute
Troy, NY 12180
zhuz@rpi.edu, qji@ecse.rpi.edu

ABSTRACT

Advanced human computer interaction requires automatic reading of human face in order to make the computer interact with human in the same way as human-to-human communication. We developed an automatic face tracking and lip reading system through a 3D face avatar to facilitate HCI applications in speech learning, emotional state monitoring, and non-verbal human computer interface design. The system implements a novel active face feature tracking algorithm with an uncalibrated camera. The 3D face pose is estimated and tracked by a Kalman filter-based matching process with a dynamic face model updating and constraint. The obtained facial motion parameters are transferred to an individualized 3D face avatar. As a result, a person's lip shape or expressions can be cloned to the animated 3D face avatar, by which all lip shapes from the same speech of different subjects can be easily compared and measured. This real time system targets the automatic facial expression analysis and synthesis for the next generation of HCI design.

Categories and Subject Descriptors: I.3.6[Computer Graphics] : Methodology and Techniques- [Interaction techniques] I.4.9[Image Processing and Computer Vision]: Applications

General Terms: Algorithms

Keywords: Face tracking, lip reading, avatar, animation

1. INTRODUCTION

Rather than simply responding to user commands, the advanced human computer interaction (HCI) should be able to detect and track the user's emotional, motivational, cognitive and task states, and use this knowledge to initiate the communication. Face is the most communicative element that can signify all these states which can be identified. For example, monitoring eye gaze could give signal of fatigue, lip-reading could assist children and deaf people on interactive speech learning. In medical practice traditional speech recognition technique could help translate the doctor's dictation to texts automatically. However, the recognition rate is degraded dramatically due to the low voice quality and accents. Incorporating lip reading with dictation can greatly improve the accuracy of speech-to-text translation.

In this paper we present an avatar-mediated face tracking and lip reading system, which is outlined in Figure 1. The system is fully automatic and non-intrusive, it tracks face pose, lip shape, gaze, eye openness and other expression features simultaneously in real time and create animation on any general 3D face model. Two major function units which are essential in HCI developing are well integrated in this system. The first unit is a robust and real time face feature tracking unit which can work under a variety of imaging conditions. The second unit is real time face modelling and animation unit, which uses limited number of feature parameters to create realistic animation.

Lip tracking methods have been well developed for years. Deformable templates and active contours based methods are two among them. Based on detected lip shape, further speech recognition could be carried out. As only a few parameters describe the position of the mouth uniquely, the allocation of mouth shape and positions to words could be made with the help of Hidden Markov models[12]).

Face pose estimation techniques can be categorized mainly into appearance-driven and model-based approaches [7, 8]. The appearance-driven approach requires a significant number of training data to enumerate all the possible appearances of features. The model based approach [7, 8] assumes that the knowledge of a specific object is available. The requirement of frontal facial views and constant illumination limited its application. Our technique uses an active tracking technique to tackle the issues of variable illuminations, rigid(head motion) and non-rigid(local skin deformation) feature tracking, and the self-occlusion of features.

Facial animation rules must be carefully designed to produce a realistic facial expression for avatar communication. Performance-driven facial expression generation relies on the accurate detection of facial features. Physical based animation method and elastically deformable models show realistic results [4], however it is time-consuming. Image-based model fitting methods use 3D morphable model to synthesize the impressive facial appearances [1, 2, 5]). Its analysis-by-synthesis loop for error energy minimization requires intensive computations. Expressions could be copied from one model to another [3] with modification-based meth-

MM'04, October 10-16, 2004, New York, New York, USA.

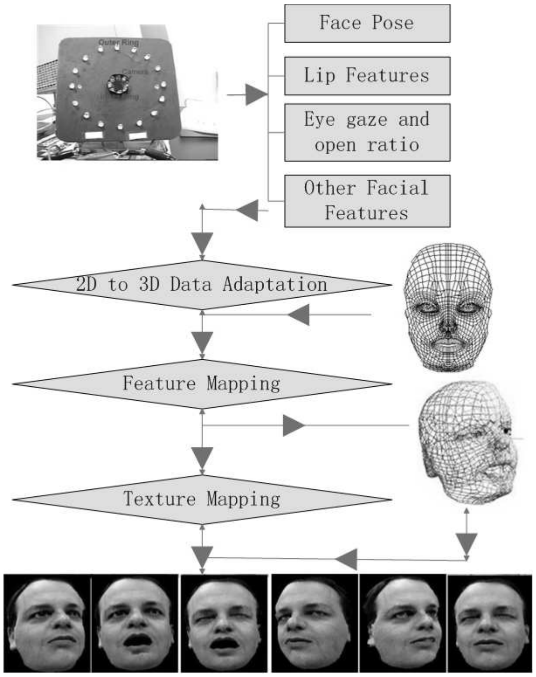

Figure 1: HCI System Diagram

ods. However, due to the lack of fine modelling on certain expression-rich areas, it is hard to track and create lip movement and eye opening given the input of 2D motion vectors. In this paper we propose Kalman filter based tracking approach and 2D to 3D parameter transferring based animation method to realize the real-time interaction system. The setup of system shows in Figure 2.

The organization of this paper is as follows: In section 2, we will overview the developed tracking and lip reading system. The individualization of motion parameters and the corresponding animation rule will be described in section 3, followed by experimental results and evaluations in section 4. Finally the concluding remarks will be given in section 5.

2. VIDEO TRACKING BASED ON SINGLE UNCALIBRATED CAMERA

Our motion tracking unit is composed of two components: face pose tracking and feature detection and tracking. We will give detail description of them in following sections respectively.

2.1 Face pose tracking

The eye pupil is selected as it is a significant feature and is able to provide strong and reliable constraints for initializing the tracking system. The active facial sensing system consists of an IR sensitive camera and two concentric rings of IR LEDs. The pupils can be easily extracted from the difference of the two images produced by different ring. The eye tracking result is further improved by a mean shift eye tracker. The shapes of the pupils are extracted by a deformable template method. The ratio of pupil ellipse axes is used to characterize the percentage of eye closure, the detailed algorithm is described in [9].

The face pose tracking system performs 2D face detection and 3D face pose tracking simultaneously rather than treating them separately. As have mentioned above, eyes have been detected in the first step, so based on eyes a fronto-parallel face model can be created to represent the initial tracked face. The detected face region will serve as the initial 3D planar face model so that tracking can be carried out based on this model. The 3D face pose is then tracked in 3D space using Kalman filter, starting from the front-parallel face pose.

Face pose is characterized by a rotation matrix R resulted from successive Euler rotations of the camera frame around its XYZ axis. Three rotation angles ω, ϕ, κ along X,Y,Z axis respectively. A scale factor λ, is used to represent the distance from face to camera, and form the projection matrix. Therefore the face pose is characterized by these four parameters and its estimation can be expressed as determination of $(\omega, \phi, \kappa, \lambda)$.

Face detection and pose estimation are done first via a matching optimization algorithm. As Kalman Filtering provides the predicted face position and 3D face pose, we only need to search for a face position and 3D pose. Thereby insuring that the detected face view position image can best match the projected face view image for that pose. The matching criterion is based on the sum of squared difference errors over all the image pixels within the region of interest:

$$E_{matching} = \sum_{i=1}^{N} (I_p(i) - I_c(i))^2 \qquad (1)$$

where I_p is the reconstructed face view image, I_c is the current image, N is the total pixel number, and i denotes the ith pixel. $I_p(i)$ is projected from reference image $I_r(c, r)$ via a mapping function f

$$I_p = f(I_r, \alpha) \qquad (2)$$

where α consists of the face pose parameters. By finding the local optimal face pose parameter we can find the projected face image that best matches the face in the current image frame.

The tracking strategy requires the face model to be updated continuously in order to accommodate face change due to variations in illumination, face aspect, and facial expression. In order to maintain tracking performance it is necessary to dynamically adapt the face model to keep them consistent with the face while the face appearance is changing. Thus to update the face template we need to include a portion of the initial template into every template. For an arbitrary frame k, after the best match is found, the new face template for frame $k+1$ is updated to be the combination of the initial face template and the face model generated from the new frame. Experiments prove that the accuracy is greatly improved by dynamic face model updating and physical eye locations constraints.

2.2 Feature detection and tracking

We have selected twenty-two facial features to track. Eight features surrounding the lip are tracked, which are shown in Figure 3. Meanwhile three features for each eye, two features for each eyebrow, and four features for nose are selected to track respectively. We use the multi-scale and multi-orientation Gabor wavelet to represent each feature. To identify each facial feature at the initial frame, 18 Gabor coefficients are used to represent each feature pixel and its vicinity [10], a set of 18 Gabor coefficients $\Omega(\vec{x})$ is obtained

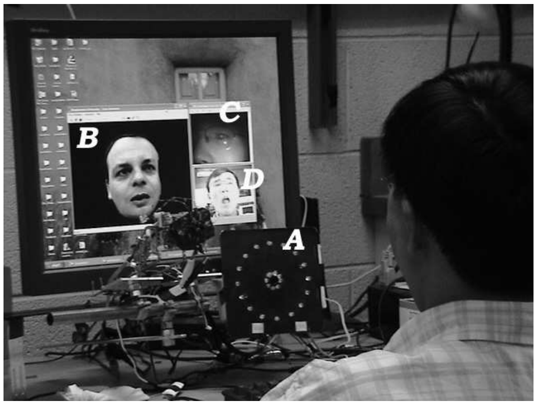

Figure 2: System Setup. A: IR LED rings and IR active camera; B: Animated avata; C: Tracked pupil; D: Tracked face in video

by the convolution operation:

$$\Omega(\vec{x}) = \int I(\vec{x}')\Psi[\mathbf{k},(\vec{x}-\vec{x}')]d^2\vec{x}' = (c_1, c_2, ..., c_{18})^T \quad (3)$$

where 2D Gabor kernels are defined as:

$$\mathbf{\Psi}(\mathbf{k},\vec{x}) = (\mathbf{k}^2/\sigma^2)e^{-\mathbf{k}^2\vec{x}^2/2\sigma^2}(e^{i\mathbf{k}\cdot\vec{x}} - e^{-\sigma^2/2})$$

$\sigma = \pi$ is set for 128×128 images. The set of Gabor kernels consists of 3 spatial frequencies (with wave number k: $\pi/2, \pi/4, \pi/8$) and 6 distinct orientations from 0^o to 180^o in the interval of 30^o. With the assumption of a smooth motion of each facial feature, Kalman filter can be used to further help facial feature tracking. The motion state of each feature at each frame (time instance) can be formulated by state model: $\mathbf{S_{t+1}} = \mathbf{\Phi S_t} + \mathbf{W_t}$, where $\mathbf{\Phi}$ is the transition matrix and $\mathbf{W_t}$ models system perturbation. The state vector $\mathbf{S_t}$ at time t is characterized by its position and velocity, i.e., $\mathbf{S_t} = (\vec{P}, \vec{V})$.

In order to handle the case of mis-tracking or self-occlusion, the detected features are verified and inferred by both qualitative and quantitative evaluations. If the wrong feature is detected by the verification, data obtained from previous frame is used [6].

3. AVATAR ANIMATION

Creating realistic animation based on features obtained from video is an essential part of the interaction between human and computer. To transfer the source motion parameters (MP) to target animation parameters (AP), the individual 3D face avatar is first created by our face modeling algorithm. This algorithm is performed based on the two views of an individuals face. We have developed an algorithm [11] to reliably identify features on the profile such as chin tip, mouth, nose tip, and nose bridge. After the process of static face modelling, we marked the regions in mouth, nose, eye, pupils, eyebrows and define the area-of-influence by these regions.

3.1 Estimation of animation parameter

In order to obtain APs, we need to process source MPs by a series of transformations. Because we have very limited number of feature points(22), we have to define rules which can realistically animate the non-feature vertices. The numbers of tracked feature points are not only limited, but also scattered irregularly upon a large surface. Some non-feature points among or far from feature points can't be influenced appropriately. For example, the out-end points of eyebrows, and points on nose. For those regions that are either located in moving-active area(eyebrow), or have sparsely scattered feature points inside, we want to add additional feature points to help further recursive inference procedure. For the eyebrows, we add an out-end point by connecting the inner-end and center eyebrow point and extend it to position where center point is at the middle of the inner and outer-end points. The motion parameter obey following rules: $MP_{inner} + MP_{outer} = 2 \times MP_{center}$ Two feature points(MP_{a1}, MP_{a2}) each are also added between nose top(MP_t) and left nose corner(MP_{lc}), and nose top and right nose corner(MP_{rc}). For the left side we have: $MP_{a1} = \ell_1 \times MP_t + \ell_2 \times MP_{lc} \quad and \quad MP_{a2} = \ell_2 \times MP_t + \ell_1 \times MP_{lc}$, ℓ_1 and ℓ_2 are weights assigned, typical values are 1/3 and 2/3 as two points are inserted. Same rule apply to right side of nose.

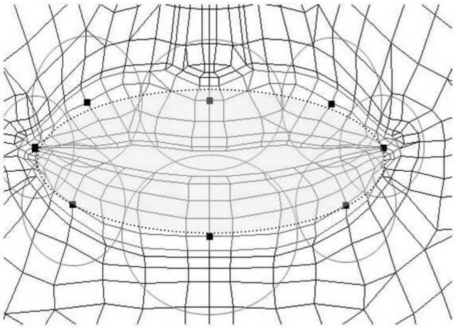

Figure 3: Parabolic model super imposed on Region definition of lips on the mesh, black point represent tracked features

The motion of each feature point on a face can be modelled as a local skin deformation plus a global head affine transformation. Given a feature point $\vec{v_o}$ with the front view of a source subject, its corresponding point $(\vec{v}_p)$ after the local deformation due to expressions and global motion due to the head pose change is formulated as : $\vec{v}_p = T \cdot R \cdot \vec{v}_f, \quad \vec{v}_f = \delta\vec{v} + \vec{v}_o$ where T, R are translation and rotation matrices, $\vec{v}_o$ is a feature point with neutral expression of the front view. $\delta\vec{v}$ is the deformation applied onto $\vec{v}_o$, which is also the approximation of APs. The animation data obtained is further adapted to apply onto the avatar model. Except the 22 feature control feature points, the 3D animation vector V of non-feature vertex i (denoted as V^i) can be derived by the following equation: $V^i = \sum_{j=1}^{N}(\omega(d_{i,j}) \cdot V_k^j)$. $d_{i,j}$ is the distance between vertex i and vertex j, and $d_{i,j}, j = 1, 2, \ldots, N, (e.g.\ N = 3)$ have been arranged in increasing order. $\omega(d_{i,j})$ is a weight function which will assign a larger weight value to $d_{i,j}$ who has smaller value.

3.2 Lip shape modelling and animation

As mentioned previously, eight feature points around lips are detected directly by our approach. We divide the two lips into 8 regions, and each region is controlled by one feature point. The mouth contour can now be derived easily with an interpolated curve, such as parabolic curve (Figure 3). After the estimation of animation parameters, the

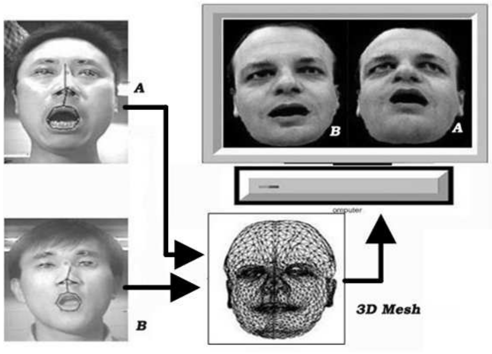

Figure 4: Lip Shape Cloning

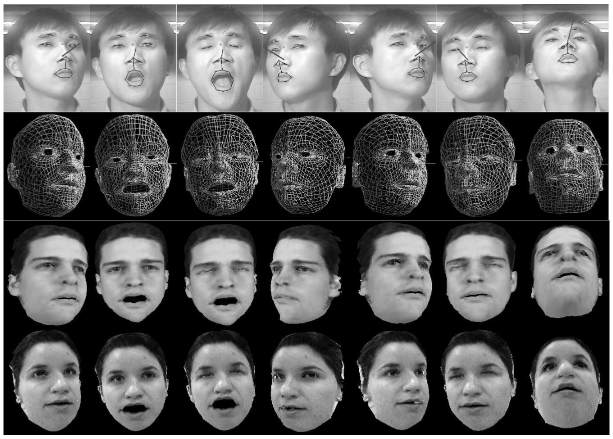

Figure 5: Experiment Results, first row: images extracted from video, second row: animated face model mesh, Third and fourth row: Animated individual models with texture mapped

curve can be animated by applying these parameters on our predefined regions.

Our system can clone the speaker's mouth shape to the avatar on the lip area. In the scenario of two speakers, their mouth shapes can be cloned to the same avatar, which is shown in the monitor (Figure 4). Then two shape of same speech from two person can be compared or even measured by comparing the eight points on the two models.

4. RESULT

Our non-intrusive tracking and animation system works in a fully automatic fashion. Experiment result in Figure 5 shows the good visual quality on created expressions which are cloned from the first row of Figure 5.

The tracking speed is about 15 frames/second on a double Pentium IV 3G machine, and the animation speed is 30 frames/second on a Pentium IV 2.4G PC. Also as the tracked data is small in size, it enables remote face expression be rebuilt in any environment. This is ideal for human computer interaction in far distance.

Experiments show that the system works reliably under rotation angle between -30 to 30 degree. However it fails when the rotation angle is beyond this limit, or there is heavy occlusion. In the case of a tracking failure, the system will be reset and resume tracking shortly.

5. CONCLUSIONS AND FUTURE WORK

The system is demonstrated to work well under various illumination conditions, even with reflections of eye glasses on subjects. Our experiments show that it is feasible to create face animation and expressions with very limited number of facial features. This enable us to further develop a more user-friendly HCI interface. Such a system could be potentially applied as a non-intrusive expression and pose reader to assist the existing emotion identifier or lie detector for law enforcement or security.

In the future we will investigate more reliable feature tracking techniques without the use of IR devices. Our system will be extended to perform real-time speech-to-text translation based on the tracked lips.

Acknowledgement

The authors would like to thank the NSF (IIS-0414029) and FSOR for the support of this work.

6. REFERENCES

[1] K. Hiwada, A. Maki, A. Nakashima, "A Real-Time Face Tracking System Based on Morphable 3D Model Fitting", ICCV, 2003

[2] Sami Romdhani, Thomas Vetter, "Efficient, Robust and Accurate Fitting of a 3D Morphable Model", ICCV, 2003

[3] Jun-yong Noh, Ulrich Neumann, "Expression Cloning", *SIGGRAPH01*, pp277-288.

[4] D. Terzopoulosa and K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models." *IEEE Trans. PAMI*, 15(6), 1993

[5] Volker Blanz, Thomas Vetter, "A Morphable Model for the Synthesis of 3D Faces", *SIGGRAPH*, 1999.

[6] Haisong Gu, Qiang Ji, Zhiwei Zhu, "Active Facial Tracking for Fatigue Detection", *IEEE Workshop on Appl. of Computer Vision*, December , 2002

[7] Feng Jiao, Stan Li, Heung-Yeung Shum, Dale Schuurmans, "Face Alignment Using Statistical Models and Wavelet Features", *Proc. of IEEE CVPR*, 2003

[8] Z. Zhang, "Feature-based facial expression recognition: Experiments with a multi-layer perception". *Technical report INRIA* ,(335),1998

[9] Z.Zhu and Q. Ji, Kikuo Fujimura and Kuang-chih Lee, "Combining Kalman Filtering and Mean Shift for Real Time Eye Tracking Under Active IR Illumination", *International Conference on Pattern Recognition*, August 11-15, 2002.

[10] T. Lee, "Image representation using 2D Gabor wavelets", *IEEE Trans. PAMI*, 18(10):959-971, 1996

[11] H. Ip and L. Yin, "Constructing 3DIndividualized Head Model From Two Orthogonal Views", *The Visual Computer - the International Journal in Computer Graphics*, 12(5):254-266, Springer-Verlag, 1996.

[12] Pei Yin, Irfan Essa, James M. Rehg, "Asymetrically Boosted HMM for Speech Reading", *IEEE Proceeding on CVPR, Vol II:755-761*, 2004.