AUGMENTED REALITY MIRROR FOR VIRTUAL FACIAL ALTERATIONS

Vlado Kitanovski, Ebroul Izquierdo

Multimedia and Vision Research Group, Queen Mary, University of London, UK
{vlado.kitanovski, ebroul.izquierdo}@eecs.qmul.ac.uk

ABSTRACT

We present a system for virtual mirror experience that performs attentive facial geometric alterations in augmented reality. The virtual mirror is simulated using commonly available PC with webcam that capture, process and display video in real-time. High realism is obtained by considerate 3D-aware warping of the 2D captured video. A Kalman-based real-time face tracker is used for 3D head pose estimation and accurate facial features localization. The 3D face model used is adapted to the person in front of the mirror by utilizing active shape models for facial landmarks detection, followed by *z*-depth progressive refining. Geometric adjustments are performed on 3D face vertices while the 2D warping is calculated utilizing the location of back-projected-to-2D face model vertices. The evaluation of our system shows that realistic facial modifications can be rendered in scenarios that correspond to typical usage of a real mirror.

Index Terms— Augmented reality, virtual mirror, real-time facial features tracking, image warping

1. INTRODUCTION

The general concept of mediated reality refers to any environment where the view of reality is modified by adding, removing or modifying scene objects. Depending on the dominant environment and modification type, mediated reality condenses to different points on the reality-virtuality continuum [1] e.g.: the well known *virtual reality* – refers to completely synthetic environments, *augmented reality* – refers to augmenting the real-world environment with computer-generated objects; *augmented virtuality* – refers to adding real objects into computer-generated environments etc. Our work is related to modifying human faces in real environments while preserving their realistic appearance, which compared to all of the non-real-world environments, is closest to what is usually referred as augmented reality.

More concretely, in this paper, we present a real-time system that turns a computer display into a virtual mirror, where the user can see his/her face with applied desired geometric (shape) modifications. The potential usage of this system includes applications like virtual face beautification or visualizing outcomes of facial surgery. An important objective in the design of our system was to find a good compromise that meets the following requirements as much as possible: highly non-intrusive behavior, real-time performance, highly realistic appearance and low-cost implementation.

Several authors have used the concept of virtual mirror for applications that involve virtual face modifications. Cullinan and Agamanolis [2] designed responsive virtual mirror to create a shared environment for use in video teleconferencing. Ushida et al. [3] designed a virtual mirror using half-silvered mirror, camera and projector. They used this mirror for simulating younger/older appearance of the user's face. Their equipment is massive, the mirror display is hazy due to the non-ideal half-silvered mirror and the processing unit is not aware of the 3D world. Iwabuchi et al. [4] designed computer-augmented mirror to aid the facial makeup application. The mirror uses two cameras, physical markers placed on the face and infrared sensors to offer additional functionalities of a real mirror like zooming, panning, or changing the lighting. However, their system doesn't apply any virtual modification; it can be used only as a tool to aid the real makeup process. Darrell et al. [5] presented a virtual mirror that uses two cameras for localization of faces in the captured video. They perform arbitrary 2D face warping to create entertaining and funny impression of the mirror being aware of persons' presence. Makino et al. [6] designed virtual mirror that captures the 3D shape by analyzing the reflection from 3 light sources with a photometric stereo technique. Their system is a bit intrusive and lacks both realism and accurate facial features localization.

For our virtual mirror, we use very simple equipment set – a PC with a webcam. Facial features are tracked in real-time using user-customized 3D face model that is generated in a non-intrusive way. User's head movements in front of the mirror are utilized to progressively refine and improve the 3D face model. The details of the whole process are explained in section 2. In order to maximize the realistic impression from the mirror, we perform the user-desired facial modifications by means of 2D warping, as explained in section 3. Section 4 presents evaluation results, which are followed by concluding remarks and future research directions in section 5.

2. 3D FACE TRACKING AND MODELING

The user of our system is supposed to be able to make virtual shape modifications of the face, like changing the size, the form or the position of his/her lips, nose, and eyebrows. In order to achieve proper rendering of any applied modification, accurate tracking and accurate 3D head pose estimation is needed. There are many different approaches for 3D facial features tracking in monocular videos, but in general, they can be classified into three main categories: feature-based, appearance-based and combined. While statistical appearance models, despite intensive training, are not the best solution for tracking facial features under different facial expressions or lighting conditions, feature-based approaches are more robust regarding these issues, but however, the latter ones may suffer from drifting and accumulating tracking errors [7]. We are using a feature-based approach as we believe that having more control in terms of tracker's deterministic behavior is somewhat more preferable for our application. Extended Kalman Filter (EKF) [8] is fed with template matching results to estimate the state $\mathbf{b} = [\mathbf{r}^T, \mathbf{t}^T, \boldsymbol{\tau}_a^T]^T$ of our 3D face model - the pose (rotation and translation) and the facial animation parameters $\boldsymbol{\tau}_a$ (that define how the face model vertices displace under some common facial expressions) . Similar approach can be found in [7], [9], where authors are estimating only the pose and the rigid 3D face model. Estimating the z-depth of the 3D face model, however, is not too critical for accurate tracking and shouldn't be performed all the time during tracking. Separate EKF is used in the beginning phase only to refine the initialized z-depth of our model. After certain criterion is met, the second EKF is simply turned off. The 3D face model used is initialized in the beginning of the tracking process, when the user places his face in a frontal position with respect to the camera. Unlike the uncomfortable manual landmarks/3D model initialization found in [9-10] we use the Appearance Shape Model (ASM) trained on frontal faces [11-12] to automatically detect 68 facial landmarks as shown on Fig. 1. Additional 49 points are added using interpolation while preserving facial symmetry as shown on the left image in Fig. 2. All of the points are then assigned predefined z coordinates (taken from generic 3D face model) to initialize our 117-vertices user-customized face model (Fig. 2, right).

Figure 1: Automatic facial landmark localization using ASM.

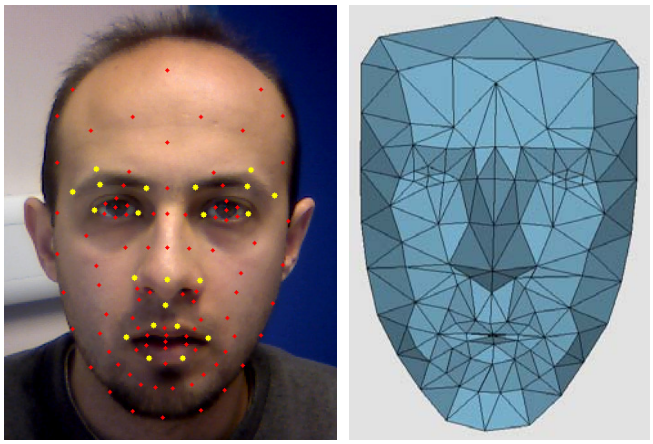

Figure 2: Frontal face with facial landmarks (left); Corresponding 3D model (right).

Relation between the 3D world and 2D image is established using the camera perspective model, with the coordinate origin set to be in the pinhole camera, the x-y plane parallel to the image plane and the z axis passing through the image centre. In this way, only one intrinsic parameter is needed – the focal distance f whose value is roughly approximated and then refined during tracking, as explained later. The EKF is initialized at the starting frame, with $\mathbf{r} = [0, 0, 0]^T$ (user's face pose is frontal to the camera), the translations along x and y axis are calculated as the distance between model's and image centers, while the z-translation is roughly approximated. At each time step, the 3D model is rendered first using the estimated parameters $\mathbf{b}$ and the first-frame texture, and then square templates are extracted around the N=22 tracked points (shown in yellow on Fig. 2). In order to improve performance under changing lighting conditions or facial animation, these templates are alpha-blended with corresponding templates extracted from the previous frame to produce the final templates. The latter ones are used for matching within a search window in the next frame. Larger search window results in increased robustness to high-speed head movements but also increases the computational time. Zero-mean normalized cross-correlation is used to calculate the matching score. The center of the search window for each tracked point in the next frame n+1 is calculated using a second-order linear model (assuming uniform time steps):

$$\hat{x}_{n+1} = x_n + (x_n - x_{n-1}) + \frac{1}{2}(x_n - 2x_{n-1} + x_{n-2}) \qquad (1)$$

In equation (1), $\hat{x}_{n+1}$ is the estimated x coordinate of the search window centre for the next frame. The same model is used in y direction. Considering that the facial movements at high frame rate are smooth, this approach is quite accurate and requires considerably less calculations than using the EKF for the same purpose (which would require several matrix multiplications for predicting $\mathbf{b}$ and projecting the tracked points). The measurement noise covariance matrix, $\mathbf{R}$, is used to tell the EKF which tracked

points should be considered more when estimating the state for the incoming frame. This matrix is diagonal and for each tracked point i, it is updated with the following value:

$$R(i,i) = \frac{1+Cd}{M_c} \cdot W(\mathbf{r},i) \quad (2)$$

In the last equation, C is a constant; d is the distance between the estimated centre and the actual best match; M_c is zero-mean normalized cross correlation between the matching templates, while $W(\mathbf{r},i)$ are predefined weighting factors used to suppress points (by increasing the measuring error covariance) that, according to the face pose $\mathbf{r}$, might be occluded. In this way, the EKF is "made aware" about potential occlusions under head rotation, or eventual bad matches. The relation matrix $\mathbf{H}$, that relates the 2D tracked points with the state $\mathbf{b}$, is linearized around the current state parameters using the camera model. The other covariance matrices are initialized to the identity matrix.

At the beginning stage of the tracking, a second EKF is used to refine the model's z-depth to fit better the particular user's face. As estimating the z coordinate for all vertices would be impossible in real-time, only those k vertices that correspond to the tracked points on the left half of the face are estimated (we assume that the face is symmetric in the z direction). The z coordinates of all other vertices are corrected accordingly using linear interpolation. In this stage, the focal distance f is also estimated, so the state vector is $\mathbf{s} = [z_1, z_2, \ldots z_k, f]^T$. As z-depth and focal distance don't vary in time, they are estimated until the total relative change $\Delta\mathbf{s}$ of the state vector in the last consecutive F frames falls below experimentally obtained threshold:

$$\Delta\mathbf{s} = \frac{\sum_{i=1}^{F} abs(\mathbf{s}_i - \mathbf{s}_{i-1})}{\sum_{i=1}^{F} \left(abs(\mathbf{r}_i - \mathbf{r}_{i-1}) + abs(\mathbf{t}_i - \mathbf{t}_{i-1})\right)} < T_s \quad (3)$$

When the condition in relation (3) is met, the second EKF is turned off and the tracker continues to use the corrected 3D vertices and focal distance.

3. VIRTUAL FACE MODIFICATION

In order to get realistic mirror impression, warping of the original camera video (the ideal mirror view) is used, as opposed to rendering the complete 3D scene consisted of the user's face and the background – in which case, high realism is infeasible due to non-ideal segmentation and modeling. Video 2D warping is performed according to the user's desire for slight modification of his/her lips, nose or eyebrows. Few authors [13-14] have used this approach to apply realistic facial modifications towards improving facial attractiveness in portrait images. We are not dealing with images, but however we can still consider the face as relatively planar surface and use 2D warping to render small 3D changes on the face. This especially holds if it is taken into account that "a mirror is functional only when you look at it", which means common usage of a mirror implies about-frontal face poses. In our system, we have implemented a dozen of different modifications including: lips reduction / augmentation, lips shape alteration, thinning / thickening eyebrows, and nose reduction / augmentation. Each of them is manipulated by a separate pair modification unit–modification control vector, and they are performed on the static 3D face model in the same linear way as the local facial animation while making facial expressions. Assuming that the coordinates of the static model vertices are stored in the column-vector $\mathbf{g}_s$, the final face model is given by:

$$\mathbf{g} = \mathbf{g}_s + \mathbf{M}\boldsymbol{\tau}_m + \mathbf{A}\boldsymbol{\tau}_a \quad (4)$$

In the last equation, columns of $\mathbf{M}$ are the modification units; $\boldsymbol{\tau}_m$ is user-managed modification-control vector that specifies the amount of each separate modification. In the same manner, $\mathbf{A}$ and $\boldsymbol{\tau}_a$ are the local facial action units matrix and facial action control vector, respectively. After the final model is obtained using eq. (4), it is rotated and translated using $\mathbf{r}$ and $\mathbf{t}$, and then back-projected on the image plane to form 2D mesh. This back-projected mesh, together with the one obtained without applying any modification ($\boldsymbol{\tau}_m = 0$) define a triangular mesh warp that establishes affine mapping from each triangle in the source (unmodified) mesh to the corresponding triangle in the destination mesh. This type of warping has several advantages, e.g. GPU exploitation and spatially-controlled warping – triangles are independently warped without being influenced of other non-neighboring triangles [14]. The main disadvantage – C^0 continuity at the edges, is not a serious problem for our application as the facial modification performed is usually minimal, and always within allowed range. Apparently, this approach cannot achieve realistic face rendering under arbitrary rotations of the head. To suppress distortions that occur when the face is rotated far from its frontal position, adaptive modification units are used. Concretely, the matrix $\mathbf{M}$ is made dependent on the current rotation $\mathbf{r}$ so that almost no warping is performed for those 2D vertices that are projections of occluded 3D vertices. Similar $\mathbf{r}$-dependent weighting function as in (2) is used to alter particular column of $\mathbf{M}$ and thus suppress unwanted warping distortion for the given modification unit.

4. EVALUATION RESULTS

In this section we present the evaluation results of our virtual mirror. It was implemented using C/C++ code and run with 25 fps on Pentium D 3.2 GHz PC with a webcam and onboard graphic card. Intel's OpenCV and the OpenGL library are used for handling the webcam and the GPU-based warping. ASM Library [12] is used for automatic landmarks detection in frontal faces. The system captures and displays 640x480 video while the 3D face tracking algorithm works on the down-sampled 320x240 resolution.

Figure 3: Virtual mirror output examples. Upper row: original; Lower row: augmented nose and lips under different rotations.

Examples of performed nose and lips augmentation are shown on Fig. 3. Figure 4 shows lips shape modification while talking. As long as the warping is within certain limits that correspond to feasible facial alterations, rendered faces look realistic and don't suffer visible distortions. The 3D tracker is robust to head movements of moderate-speed, moderate illumination changes during single tracking session, and head rotations within the ranges: ±55°, ±35° and ±30° for rotations around x, y and z axis, respectively. Most users were satisfied with the achieved realism of rendered alterations. While giving satisfactory results for alterations dominant in x and y direction, our virtual mirror cannot render properly significant z alterations. Although these alterations are not so common, achieved realism can be improved by using a model with more vertices. A demo video showing mirror's output can be seen at: http://www.elec.qmul.ac.uk/mmv/vlado/vmexpl.htm.

5. CONCLUSION

In this paper, we have presented our real-time system that turns a computer display into a virtual mirror for visualizing shape alterations of the facial features. Realistic impression is achieved by employing intelligent piece-wise 2D warping of the real world video. The warps are computed using accurate 3D tracking data for the facial features. Evaluation tests showed that almost no visual distortions are noticeable when the mirror's response is relevant to the user.

As the mirror's realistic response depends directly on the tracking data accuracy and tracker initialization, the future work may be focused on automatic estimation of the 3D initial head pose so that tracker initialization can be performed on any nearly-frontal face pose. Besides trading tracker robustness for complexity, other space for possible further improvements is introducing denser meshes for improved rendering of z-direction dominated modifications.

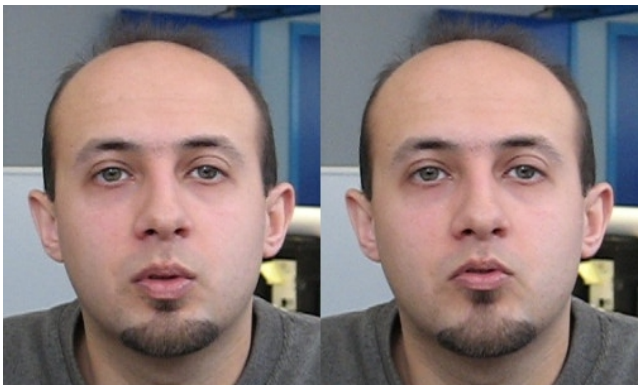

Figure 4: Lips shape alteration while talking. Left: original frame; Right: modified frame.

6. REFERENCES

[1] P. Milgram, F. Kishino, "A Taxonomy of Mixed Reality Displays" *IEICE Trans. on Information Systems*, No.12, 1994.

[2] C. Cullinan, S. Agamanolis, "Reflexion: A Responsive Virtual Mirror for Interpersonal Communication" *in proc. UIST 2002*, ACM Press 2003.

[3] K. Ushida, Y. Tanaka, T. Naemura, and H. Harashima, "i-mirror: An Interaction/Information Environment Based on a Mirror Metaphor Aiming to Install into Our Life Space" *in proc. ICAT 2002*, Tokyo, December 4-6, 2002.

[4] E. Iwabuchi, M. Nakagawa, and I. Siio, "Smart Makeup Mirror: Computer-Augmented Mirror to Aid Makeup Application" *in proc. ICHCI 2009*, San Diego, July 19-24, 2009.

[5] T. Darrell, G. Gordon, J. Woodfill, and M. Harville, "A Virtual Mirror Interface using Real-time Robust Face Tracking" *in proc. ICFGR 1998*, Nara, April 14-16, 1998.

[6] T. Makino, T. Nakaguchi, N. Tsumura, and Y. Miyake, "Virtual Mirror Based on 3D Shape Reconstruction and Real-Time Face Tracking" *in proc SPIE*, vol. 5670-30, January 2005.

[7] J. Ahlberg and F. Dornaika, "Parametric Face Modeling and Tracking" *chap in Handbook of Face Recognition*, Springer, 2005.

[8] Bishop, G.,Welch, G "An Introduction to the Kalman Filter" *in proc SIGGRAPH,* Course 8, Chapel Hill, NC, 2001.

[9] J. Strom, "Model-Based Real-Time Head Tracking" *EURASIP Journal on Applied Signal Processing*, vol. 10, 2002.

[10] M. Chaumont and B. Beaumesnil, "Robust and Real-Time 3D-Face Model Extraction" *in proc. ICIP 2005*, Genova, 2005.

[11] S. Milborrow and F. Nicolls, "Locating Facial Features with an Extended Active Shape Model" *in proc. ECCV 2008*, Marseille, October 12-18 2008.

[12] Y. Wei, "Research on Facial Expression Recognition and Synthesis", *Master thesis*, Nanjing University, February 2009.

[13] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski, "Data-driven Enhancement of Facial Attractiveness" *in proc. SIGGRAPH 2008*, Los Angeles, August 11-15, 2008.

[14] S. Melacci, L. Sarti, M. Maggini, and M. Gori, "A Template-based Approach to Automatic Face Enhancement" *Journal of Pattern Analysis & Applications*, vol. 13, Issue 3, August 2010.