REAL-TIME FACE SWAPPING IN VIDEO SEQUENCES: MAGIC MIRROR

Nuri Murat Arar[1], Fatma Güney[1], Nasuh Kaan Bekmezci[1], Hua Gao[2] and Hazım Kemal Ekenel[1,2,3]
[1]*Department of Computer Engineering, Bogazici University, Istanbul, Turkey*
[2]*Institute for Anthropomatics, Karlsruhe Institute of Technology, Karlsruhe, Germany*
[3]*Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey*
{*nuri.arar, fatma.guney1, kaan.bekmezci*}*@boun.edu.tr, hua.gao@kit.edu, ekenel@itu.edu.tr*

Keywords: Face Swapping, Face Blending, Face Replacement, Face Detection, Active Appearance Model, Face Tracking

Abstract: Magic Mirror is a face swapping tool that replaces the user's face with a selected famous person's face in a database. The system consists of a user interface from which the user can select one of the celebrities listed. Upon selection, model fitting automatically starts using the results of face detection and facial feature localization. Model fitting results in a set of points which describe the estimated shape of user's face. By using the shape information, user's face is replaced with the selected celebrity's face. After some post processing for color and lighting adjustments are applied, final output is displayed to the user. The proposed system is able to run in real-time and generates satisfactory face swapping which can be applied for face de-identification in videos or other entertainment applications.

1 INTRODUCTION

Advances in image analysis gave rise to applications of face processing techniques in several areas. Face replacement studies gains importance due to the increasing publicity of personal photographs. Besides, studies in face swapping and morphing produce the most intriguing systems regarding facial image analysis. In this study, we present such an interesting system for fully-automatic face swapping in video sequences. The system performs face swapping on the images acquired by a video capturing device, which is why we call it as Magic Mirror. User interacts with the system via a user interface. By using this interface, user is presented with a list of celebrities and requested to select the face for swapping. A snapshot of the user interface and a sample output of the system can be seen in Fig. 1.

Magic Mirror can be used as a tool to amuse people by swapping their faces with famous people's faces from politics, sports, and entertainment sectors such as movie stars, popular musicians, or simply swapping friends' faces with each other. It is possible to insert Magic Mirror in popular chat applications as a plug-in option for video chats. Magic Mirror might also be used for personal security purposes such as identity protection, face de-identification, against increasing availability of visual media.

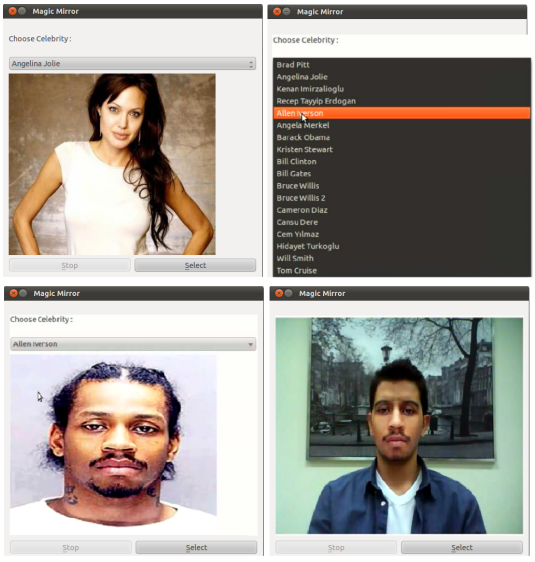

Figure 1: Example snapshots that shows how the system executes via the user interface of the system.

There are a few studies regarding both face swapping and face de-identification. Existing studies (Bitouk et al., 2008) and (Blanz et al., 2004), present the technique for face swapping on individual images. The main contribution of our study is the efficient use and effective combination of available methods for face detection, facial feature localization, face modeling, and face tracking in order to realize a fully

automatic, real-time face replacement system. We achieved a plausible performance in spite of continuity and short execution time constraints caused by real-time video processing. We used Active Appearance Model (AAM) (Cootes et al., 2001) to model the face, therefore, our study does not put a constraint on similarity of pose angles as in (Bitouk et al., 2008). In (Blanz et al., 2004), the performance of the system strictly depends on manual initialization. Also, 3D modeling of faces requires more complicated implementation issues. On the other hand, our study presents a fully automatic way of swapping faces successfully using a simpler approach. Additionally, the work in (Gross et al., 2006) achieves face de-identification using AAM in images. The main difference of this work from ours is that they used AAM to produce different appearances by changing model parameters to de-identify faces, whereas, we use AAM to obtain face contour in different images, align and swap them for de-identification.

The rest of the paper is organized as follows; Section 2 introduces related work and Section 3 explains a detailed description on the proposed method. Experimental results and discussions are given in Section 4. Finally, Section 5 concludes the paper.

2 RELATED WORK

There are both 2D and 3D studies related to face swapping, face replacement or face de-identification.

The work in (Blanz et al., 2004) presents a system that exchanges faces across large differences in viewpoint and illumination in images. It allows user to replace faces in images for a hair-style try-on. They fit a morphable 3D model to both input and target face images by estimating shape, pose and the direction of the illumination. The 3D face reconstructed from the input image is rendered using pose and illumination parameters obtained by the target image. This system requires manual initialization for accurate alignment between the model and the faces images.

The work in (Gross et al., 2006) introduces a framework for de-identifying facial images. The majority of the privacy protection schemes currently used in practice, rely on ad-hoc methods such as pixelation or blurring of the face, they, instead, combine a model-based face image parameterization with a formal privacy protection model. They change the identifying points on the face according to the protection model. They utilize AAM in order to model the faces and find the identity information on them.

The method proposed in (Bitouk et al., 2008) automatically replaces faces in images. They construct a large library of face images which are extracted from images obtained by the internet using a face detection software and aligned to a common coordinate system. Their replacement is composed of three stages: First they detect a face on the input image and align it to the coordinate system in order to find candidate faces which are similar to the input face in terms of pose and appearance. Then, they adjust pose, lighting, and color properties of the candidate face images according to the input image, and perform swapping with the input image. They rank the resulting face replacements according to a match distance and display the top ranked ones as results.

3 METHODOLOGY

3.1 System Overview

We propose a face swapping mechanism in video sequences by both combining some known computer vision techniques, such as face detection and AAM, and providing a face alignment procedure. We used modified census transform (MCT) feature-based face detection (Kublbeck and Ernst, 2006) and Active Appearance Model (AAM) to locate and model the face and its attributes. For the face alignment procedure, we use piecewise-affine warping. The basic steps of our face swapping approach are shown in Fig. 2. Mainly, there are two independent execution mechanisms; offline and online processes. The offline process is a one time procedure to build an AAM used during the online process. The AAM is built from a set of landmarked faces and this model is loaded each time the real-time system is started.

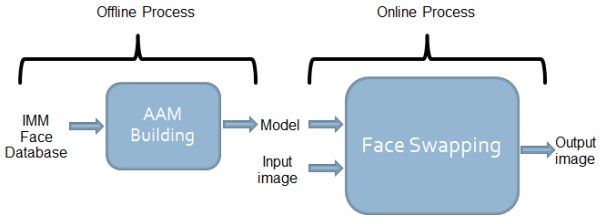

Figure 2: System overview

The online process is executed each time the system starts and the internal steps including face detection and tracking, AAM fitting, face alignment, face swapping and post processing are sequentially performed. That is, results of the detection step are used for the construction of initial shape required in AAM modeling step and alignment through warping produces the coordinates used in swapping. The online process starts with a welcome screen in which a

user is asked to select a face from the list of celebrity faces. This face chosen by the user is called the target face. Then, face detection and AAM fitting is performed on the target face. The best fitted shape for the target face is obtained. Then, video frames are captured from a video camera which are called as input frames. The similar procedure explained above is applied on the input frames. That is, face detection is performed on the input frames to locate the faces. For each input frame, face detection results are fed into AAM fitting step and an estimated face shape is obtained. After face detection and AAM fitting steps, we have the best fitted shapes of both input face and target face, but, for a successful swapping, we need them to be in the same scale on a common coordinate system. Then, we warp the target shape to the input shape for the exact alignment. We swap two faces by exchanging the pixel intensities of aligned target face with corresponding intensities of the input face using shape information of both faces. After swapping, in order to enhance the reality of output images, we apply post-processing by computing output image as a weighted combination of input and target face for color and lighting adjustment. After post-processing, the output image is displayed to the user. Fig. 3 shows an example sequence of input image, target image and the output image, respectively.

Figure 3: Input image (left), target image (middle) and output image (right).

3.2 Face Detection using MCT

We use MCT frontal face detector, which advocates the use of inherently illumination invariant image features for object detection, to locate the face in video frames. The features convey only structural object information which are computed from a 3×3 pixel neighborhood using the modified census transform. For classification, this feature set is used with a four-stage classifier cascade. Each stage classifier is a linear classifier, which consists of a set of lookup-tables of feature weights. Detection is carried out by scanning all possible analysis windows of a particular size. Each window is classified as either background or face. Detector has four classifiers of increasing complexity. Each stage has the ability to reject current analysis window or pass it to the next stage according to the outcome of a thresholding operation. As a result of four stage classifier, the detector determines the location and scale of the face on the image. Using the same algorithm we can train detectors for individual facial components such as eyes and mouth. An example of detection output is displayed in Fig. 4.

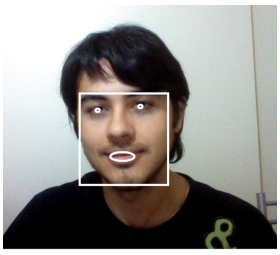

Figure 4: Outputs of face detection step: bounding face rectangle, mouth and eyes.

3.3 Active Appearance Model

AAM is known as a deformable model which defines the statistical modeling of shape and texture of a known object and its fitting algorithm. Many face related studies use AAM to model the shape of the face. Shape is defined by a set of points on the common coordinate system of the training set used. First, a mean shape and modes of variation is learned from training data by applying PCA. Then, by changing the mean shape through different variations, which are defined in modes of variation, various instances of shape can be obtained. Texture information is a statistical model of the gray-level image and captured by sampling with a suitable image warping function defined by the shape parameter.

3.3.1 AAM Building

AAM building process is an offline process performed only once before the system starts to operate. We used the IMM face database (Nordstrm et al., 2004) containing 240 annotated monocular images of 40 different human faces in order to build the model. Each image in the database is annotated with 58 landmarks for defining the shape of the face (Fig. 5). AAM is trained with these annotated images.

Figure 5: Sample original image in IMM face database (left) and annotated image (right).

3.3.2 AAM Fitting

AAM fitting is used to obtain shape of both target and input faces. AAM fitting is an iterative process of adjusting model parameters to fit the model to unseen face images (Cootes et al., 2001). An initial shape for the detected face is constructed by using the facial features detected in the first step. First, an affine transform is estimated according to the facial features on the input image and the model shape. Then, estimated transformation is performed on the model shape so that the differences in locations between the feature points on the model shape and the input image are minimized (Gao, 2008). The result of AAM fitting is an optimized model shape which best approximates the face on the current image Fig. 6. This optimization is performed using the simultaneous inverse compositional algorithm as stated in (Gross et al., 2005).

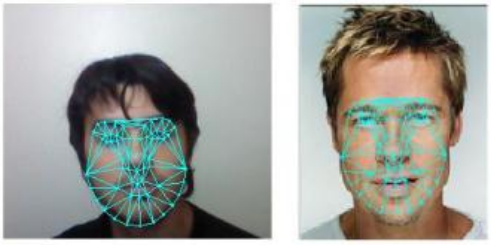

Figure 6: AAM fitting results of input image (left) and target image (right).

3.3.3 Tracking with AAM

Face detection is the most expensive step in terms of time constraints. Since this application aims to do face swapping in real time, using a tracker instead of performing face detection on each frame saves a lot of time. Based on the fact that face in the current frame cannot be too far away from the face found in the previous frame, a tracking mechanism using AAM can be developed in this context. AAM fitting is performed on each frame by using the fitting results, i.e., shape and appearance parameters of previous frame as the initialization of current frame (Gao, 2008).

Face is located using the face detector in the first frame, after that, best fit shape of previous frame is directly fed to the AAM as the initial shape of current frame. An error measure that yields low value for good face model fitting and high value for poor fitting is defined to indicate the quality of AAM fitting. Whenever the error is high, it indicates that the tracking or fitting failed and a reinitialization is required to continue tracking on the following frames. Reinitialization is carried out by detecting face and feeding the initial shape constructed from it to AAM fitting step.

3.4 Face Alignment

Faces registered in the system are extracted in various sizes and orientations. Input frames coming from video camera also differ in size and orientation, because user may not stand in front of the the camera at constant distance and position. In order to perform directly swapping of input and target faces, an alignment step is necessary to bring various sized and positioned face images into a common coordinate system with a particular size. For the alignment process, firstly we bring RGB colored face of the chosen celebrity by using its best fit shape and its original image into a coordinate system so that we obtain the aligned target face. Then, we bring RGB colored face of the input frame into a coordinate system so that we obtain the aligned input face. To do this, a piecewise-affine warping is applied with triangulated mesh. Pixels inside each triangle are determined by bi-linear interpolation. At this point, these faces are aligned in terms of their positions but their sizes are different and also their shapes are different in terms of width, height and the locations of points. Finally, we warp the aligned face of the target face to the aligned input face. This warping process resizes the target face in proportion to input face and warps its points to achieve the alignment (Fig. 7).

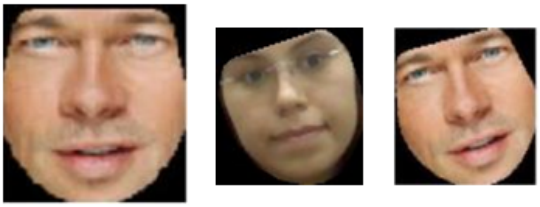

Figure 7: Target image aligned to a common coordinate system (left), aligned input image (middle) and target image aligned to the aligned input image (right).

3.5 Face Swapping

After the alignment process, the size, orientation and the shape of input face and target face correspond in common coordinate system. At this point, the relative locations of the faces on the images are known, but to complete the swapping process, the transformation between these locations needs to be calculated. In order to calculate this transformation, firstly, bounding box of shapes are located. Then, function which matches the upper left corner of the bounding box of the target shape with the upper left corner of the bounding box of the input shape is computed. The transformation is obtained by applying this function to all pixels of target face. Lastly, corresponding pixel

intensity values are copied to the output image. The results of face swapping step are shown in Fig. 8.

Figure 8: Input image (left), target image (middle) and face swapping result (right).

3.6 Face Blending as Post Processing

After face swapping process, face blending is performed in order to smooth the transitions on face contour. Face blending is the process of combining input and target faces' intensities with changing weights according to pixel coordinates. On the face contour, the weight of the input face is higher. The weight of target face is increased while getting closer to the center of output face. Fig. 9 depicts the effectiveness of the post processing step.

Figure 9: Face swapping without post processing (left) and face swapping with post processing (right).

4 RESULTS

Some qualitative results produced by the system are shown in Fig. 10 and 11 . The results are obtained by selecting frames from the output real-time video of the system. Fig. 12 shows frames taken from a 6 seconds video stream produced by the system in order to show the robustness of the face alignment to different size, rotation and translation transformations during the video. Please note that, it is difficult to quantitatively assess the performance of the developed system, since it requires a significant amount of facial feature point annotations. As a matter of fact, evaluation of face replacement results are mostly qualitative by the problem's nature. An example of quantitative evaluation is employed in (Bitouk et al., 2008) by performing a user study testing people's ability to distinguish between real images of faces and those generated by their system. However, this evaluation strategy cannot be applied here, since we deal with video sequences rather than individual images and require user to select the target image beforehand.

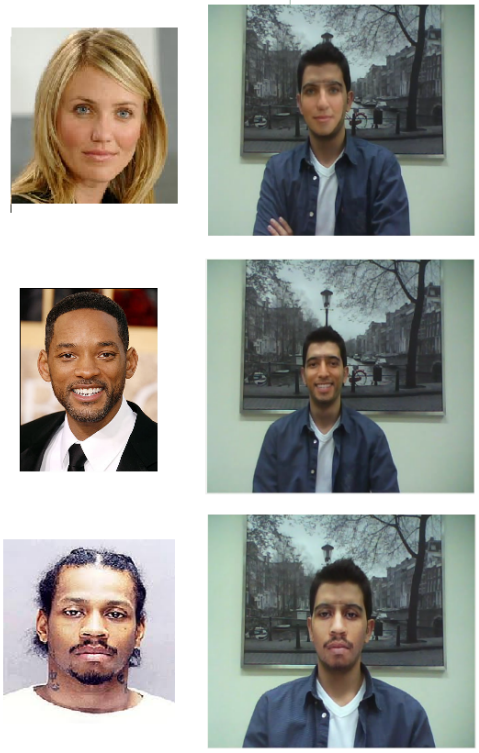

Figure 10: Some results of the system: target images (left) and output images (right).

The system runs in real-time by processing faces of 110×110 average spatial resolution in 640×480 sized frames. To obtain quantitative results somehow and evaluate the real-time property of the system, we measured average execution times of each internal step. The results of the experiments conducted on an Intel Core i5 2.5 Ghz computer are shown in Table 1.

Internal Steps	Average Execution Time (in ms)
Detection	173.22
Tracking	17.12
AAM fitting	0.05
Alignment	21.66
Swapping and Blending	2.91

Table 1: Average execution times of internal steps.

Face detection is the most time consuming process of the system. With the use of the tracking mechanism, face detection is not performed in each frame. However, there is a trade-off between response time and realistic face swapping. For lower values of the error measure on model fitting, re-initialization of the tracker occurs more frequently for better face model

Figure 11: Example output sequence, each row shows three screen shots belonging to a different celebrity selection. For each frame in a row, note the differences in position, angle, and size.

fitting results. Each re-initialization means a new face detection, therefore, average processed frame rate decreases. Average processed frame rate can be increased by applying higher values of the error measure which results in poor fitting of faces. In our demo sessions, we observed that poor fitting is a more noticeable deficiency compared to low response time. Especially, when lightening is insufficient in the environment, fitting error increases and tracker becomes useless. In such cases, even using continuous detection for tracking produces more realistic results.

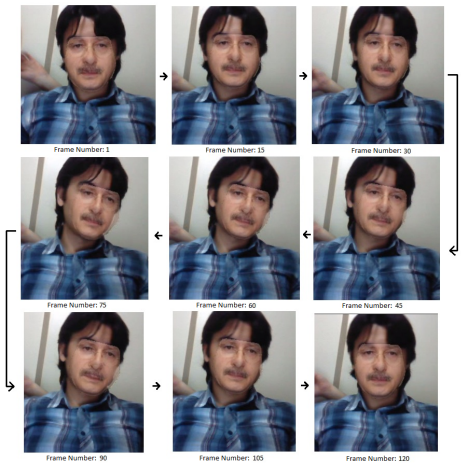

Figure 12: Sample frames from a 6 seconds output video.

5 CONCLUSION

In this study, we developed a complete and fully-automatic system to swap the face of the user with a face selected by the user via the user interface of the program. The images are acquired by a video capturing device and face swapping process is applied on these images. The system have a quite simple and user-friendly interface for users to select a celebrity from a list. After selecting the celebrity, the system performs face swapping between the user and the celebrity without requiring any manual initialization.

We used MCT face detector to locate frontal faces and MCT eye and mouth detectors for the localization of facial features. Detection results are fed to the AAM fitting mechanism as initial shape positions. After the AAM fitting process, best fits for the input and target faces are generated. Using these shapes, a piecewise-affine warping is applied with triangulated mesh in which target shape is warped into the input shape. Face swapping is performed by obtaining a transformation function between aligned target face and input face. Finally, we apply a post-processing by using a weighted combination of intensity values to blend in target face and input face. The system produces qualitatively nice results and is able to run in real-time by processing frames of 640×480 spatial resolution.

REFERENCES

Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P. N., and Nayar, S. K. (2008). Face swapping: Automatically replacing faces in photographs. *ACM Transactions on Graphics*, 27(3):1–8.

Blanz, V., Scherbaum, K., Vetter, T., and Seidel, H.-P. (2004). Exchanging faces in images. *Computer Graphics Forum*, 23(3):669–676.

Cootes, T., Edwards, G., and Taylor, C. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685.

Gao, H. (2008). Face registration with active appearance models for local appearance-based face recognition. Master thesis, Universität Karlsruhe.

Gross, R., Matthews, I., and Baker, S. (2005). Generic vs. person specific active appearance models. *Image Vision Computing*, 23(11):1080–1093.

Gross, R., Sweeney, L., Torre, F., and Baker, S. (2006). Model-based face de-identification. *22nd IEEE International Conference on Data Engineering, (ICDE)*, 27(3):161–168.

Kublbeck, C. and Ernst, A. (2006). Face detection and tracking in video sequences using the modified census transformation. *Image Vision Computing*, 24(6):564–572.

Nordstrm, M. M., Larsen, M., Sierakowski, J., and Stegmann, M. B. (2004). The imm face database - an annotated dataset of 240 face images. Technical report, Technical University of Denmark.