SPARSITY-BASED JOINT GAZE CORRECTION AND FACE BEAUTIFICATION FOR CONFERENCING VIDEO

Xianming Liu[1,2], *Gene Cheung*[2], *Deming Zhai*[1], *Debin Zhao*[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, China
[2]National Institute of Informatics, Japan

ABSTRACT

A well-known problem in video conferencing is gaze mismatch. Instead of relying exclusively on online captured data for rendering, a recent work first trains offline dictionaries using a large image database of movie and TV stars to learn "beautiful" features. During real-time conferencing, one can then simultaneously correct gaze and beautify the subject's facial components in single images by seeking sparse linear combination of pre-trained dictionary atoms for face reconstruction. Extending on this work, we focus on joint gaze correction / face beautification for video. First, we define a large search space invariant to scale, shift and rotation for facial feature beautification based on SIFT. We then address two practical issues unique to video: i) how beautified results can be temporally consistent across group of pictures (GOP), and ii) how blinking eyes can be beautified even though the training database contains only open-eye facial images. Experimental results show that our method achieves the desired temporal consistency, and the blinking process is smooth and natural.

Index Terms— Gaze correction, face beautification, sparse coding

1. INTRODUCTION

Video conferencing is now widely used across the globe via available tools like Skype, Google Hangout, etc. However, the user experience enabled by existing tools is still inferior to live face-to-face communication. One of the glaring problems is *gaze mismatch* [1]: a capturing web camera is typically located above or below a display monitor, while the conference subject looks at his counterpart rendered at the screen center. This means that the participant's gaze direction is not aligned with the camera's line-of-sight, and thus the parties cannot converse eye-to-eye. See Fig. 1(a) for an illustration.

To address the gaze mistmatch problem, previous solutions [1, 2, 3] typically redraw the eyes or the entire face from a virtual viewpoint as observed from the screen center, using the online camera-captured facial image(s) as reference. Inherently an inverse imaging problem, often the re-rendered eyes / faces are not natural looking due to insufficient textural / structural information in the reference image(s).

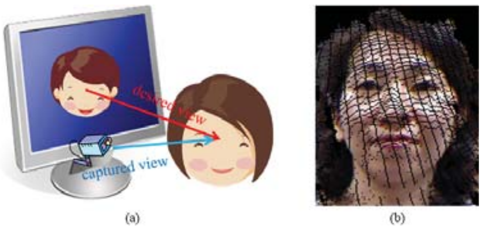

Fig. 1. (a) Illustration of gaze mismatch problem, where the camera is located below the display. (b) DIBR-synthesized facial image.

Instead of relying exclusively on online captured data for rendering, a recent work [4] proposed to first train offline dictionaries using a large image database of movie and TV stars to learn "beautiful" features. During real-time conferencing—within a unified dual sparse coding framework—one can then jointly correct gaze *and* beautify the subject's facial components (*e.g.*, eyes or eyebrows) by seeking sparse linear combinations of pre-trained dictionary atoms for face reconstruction. The overriding premise is that *given there exists uncertainty in gaze correction (typical in inverse imaging problems), the rendered facial image should err on the side of more beautiful faces*. While convincing gaze-corrected and beautified facial images were presented [4], the optimization was designed for single images, and how the process can be applied for video was not discussed.

In this paper, we address the specific challenges of performing joint gaze correction / face beautification for video. First, we define a large search space invariant to scale, shift and rotation for facial feature beautification based on SIFT [5]. We then address two practical issues unique to video: i) how beautified results can be temporally consistent across group of pictures (GOP), and ii) how blinking eyes can be properly beautified even though the training database contains only open-eye facial images. Experimental results show that our method achieves the desired temporal consistency, and the blinking process is smooth and natural.

The outline of the paper is as follows. We first overview related works in Section 2. We define our joint gaze correction / face beauification problem in Section 3. We discuss our proposal for temporal consistency and blinking eye beautification in Section 4. Finally, experimental results and conclusion are presented in Section 5 and 6, respectively.

978-1-4673-7314-2/15/$31.00 ©2015 IEEE

2. RELATED WORK

Recently, researchers proposed to jointly correct gaze and beautify the rendered human face in a single process [4]. Prior to the start of a video conference session, two dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are trained offline separately using two large image datasets: one with face images of the intended conference subject, the other with images of "beautiful" human faces—collection of frontal face images of Asian movie and TV stars from age 20 to 40. During actual video conference, a Kinect camera placed at the bottom of a display captures texture and depth images of the conference subject in realtime, which are used as reference to synthesize gaze-corrected viewpoint images of the subject as observed from the screen center via *depth-image-based rendering* (DIBR) [6]. However, the synthesized images suffer from missing pixels due to self-occlusion and insufficient pixel sampling in the captured view, as well as rendered pixel errors due to rounding to pixel grid and depth estimation error. See Fig. 1(b) for an example of DIBR-synthesized facial image.

The key technical contribution in [4] is then the simultaneous completion of the DIBR-synthesized image and beautification of facial components (*e.g.* eyes and eyebrows) via a unified dual sparse coding framework. Specifically, given the two offline trained dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, [4] jointly searches for two sparse code vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$—one is sparse in the first dictionary and explains the available DIBR-synthesized pixels, and the other is sparse in the second dictionary[1] and matches well with the first vector up to a restricted linear transform $\mathbf{L}$. The transform $\mathbf{L}$ here limits the amount of beautification performed, so that the beautified face is still unmistakably the original conference subject—this is called the *recognizability constraint*. Mathematically, the objective function can be formulated as:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{L}} \|\mathbf{x}-\mathbf{\Phi}\boldsymbol{\alpha}\|_2^2+\lambda_0\|\boldsymbol{\alpha}\|_1+\mu\,\|\mathbf{\Phi}\boldsymbol{\alpha}-\mathbf{L}\mathbf{\Psi}\boldsymbol{\beta}\|_2^2+\lambda_2\|\boldsymbol{\beta}\|_0. \quad (1)$$

where μ and λ's are weighting parameters that trade off between the fidelity term, and the recognizability term and sparsity terms, respectively.

3. PROBLEM DEFINITION

In [4], the recognizability constraint restricts the types of linear transformations $\mathbf{L}$ (*e.g.*, scaling and rotation) performed for different facial components. However, the limitation placed on the types of permissible linear transformations is too restrictive. In this paper, we redefine the recognizability constraint by leveraging on the well known notion of invariant feature transforms like *scale-invariant feature transform* (SIFT) [5] that are to some extent invariant to scaling, rotation and translation (none of which affect recognizability when small changes are made). Specifically, given a reconstructed facial component $\mathbf{\Phi}\boldsymbol{\alpha}$ using atoms in the first dictionary $\mathbf{\Phi}$ and a candidate beautified component $\mathbf{\Psi}\boldsymbol{\beta}$ using atoms in the second dictionary $\mathbf{\Psi}$, we enforce an upper bound on the distance between them in a feature space:

$$\|\, S(\mathbf{\Phi}\boldsymbol{\alpha})-S(\mathbf{\Psi}\boldsymbol{\beta})\,\|_2^2\le\gamma, \quad (2)$$

where $S(\cdot)$ is a function that maps a vector in pixel domain to a vector in a chosen feature space, and γ is a parameter that specifies the desired degree of recognizability. Using l_2-norm of feature vector difference as metric for recognizability is consistent with the object retrieval literature [7]. With the new constraint, we can reformulate the objective function as:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|\mathbf{x}-\mathbf{\Phi}\boldsymbol{\alpha}\|_2^2+\lambda_1\|\boldsymbol{\alpha}\|_0+\mu\,\|S(\mathbf{\Phi}\boldsymbol{\alpha})-S(\mathbf{\Psi}\boldsymbol{\beta})\|_2^2+\lambda_2\,\|\boldsymbol{\beta}\|_0\,. \quad (3)$$

The posed optimization can be solved efficiently via local quadratic approximation of $S()$ and *iteratively reweighted least squares* (IRLS) minimization [8]. Specifically, $S(\mathbf{\Phi}\boldsymbol{\alpha})$ and $S(\mathbf{\Psi}\boldsymbol{\beta})$ are approximated by quadratic functions of $\mathbf{\Phi}\boldsymbol{\alpha}$ and $\mathbf{\Psi}\boldsymbol{\beta}$ constructed from solutions $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ computed in previous iteration t, and l_0-norms $\|\boldsymbol{\alpha}\|_0$ and $\|\boldsymbol{\beta}\|_0$ are replaced by weighted l_2-norms, where the weights are chosen to promote sparsity. See [8] for details of IRLS.

4. TEMPORAL CONSISTENCY & EYE BLINKING

4.1. Temporal Consistency

One practical problem of joint face reconstruction and beautification for conferencing video is how to maintain temporal consistency. That is, the beautified facial components should be similar for successive frames.

We divide the video into multiple groups of pictures (GOPs), each containing T frames. Only facial components in the first frame of a GOP are beautified using our proposed optimization; subsequent frames in the GOP reuse the beautified results and simply update the rendered locations of the components based on detected *landmarks* [9] (to be defined shortly) in the original reconstructed face. In order to generate temporally consistent beautified results across GOPs, while being adaptive to possible changing captured image content (*e.g.*, lighting, shadows, etc), we introduce a new temporal consistency term in the optimization objective:

$$\begin{aligned}\min_{\{\boldsymbol{\alpha}_{t+T},\boldsymbol{\beta}_{t+T}\}} &\|\mathbf{x}_{t+T}-\mathbf{\Phi}\boldsymbol{\alpha}_{t+T}\|_2^2+\lambda_1\|\boldsymbol{\alpha}_{t+T}\|_0\\ &+\mu_1\,\|S(\mathbf{\Phi}\boldsymbol{\alpha}_{t+T})-S(\mathbf{\Psi}\boldsymbol{\beta}_{t+T})\|_2^2+\lambda_2\|\boldsymbol{\beta}_{t+T}\|_0\\ &+\mu_2\,\|S(\mathbf{\Psi}\boldsymbol{\beta}_t)-S(\mathbf{\Psi}\boldsymbol{\beta}_{t+T})\|_2^2,\end{aligned} \quad (4)$$

where $\mathbf{x}_{t+T}$ is the observed face vector in the current GOP, $\boldsymbol{\alpha}_{t+T}$ and $\boldsymbol{\beta}_{t+T}$ are the two sparse codes with respect to $\mathbf{\Phi}$ and $\mathbf{\Psi}$ respectively, and $\mathbf{\Psi}\boldsymbol{\beta}_t$ is the beautified result in the last GOP. The last term states that the new beautified result $\mathbf{\Psi}\boldsymbol{\beta}_{t+T}$ should not deviate from the previous result $\mathbf{\Psi}\boldsymbol{\beta}_t$ by much in terms of feature space distance.

[1] Assuming that each atom in the second dictionary is a "beautiful" facial component (*e.g.*, eye) learned from the training image set, a sparse linear combination of atoms is also by extension beautiful. (A non-sparse combination will just be an average eye and hence not beautiful.)

Note that in (4) we require the solution $\mathbf{\Psi}\boldsymbol{\beta}_{t+T}$ to be close to *both* the current reconstructed face $\mathbf{\Phi}\boldsymbol{\alpha}_{t+T}$ *and* the previous beautified face $\mathbf{\Psi}\boldsymbol{\beta}_t$ in feature space. We can therefore simplify (4) by requiring only the solution $\mathbf{\Psi}\boldsymbol{\beta}_{t+T}$ to be close to a mixture of $\mathbf{\Phi}\boldsymbol{\alpha}_{t+T}$ and $\mathbf{\Psi}\boldsymbol{\beta}_t$:

$$\begin{aligned} \min_{\{\boldsymbol{\alpha}_{t+T},\boldsymbol{\beta}_{t+T}\}} & \|\mathbf{x}_{t+T} - \mathbf{\Phi}\boldsymbol{\alpha}_{t+T}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}_{t+T}\|_0 \\ & +\mu_1 \left\|S\left(F(\boldsymbol{\alpha}_{t+T})\right) - S(\mathbf{\Psi}\boldsymbol{\beta}_{t+T})\right\|_2^2 + \lambda_2\|\boldsymbol{\beta}_{t+T}\|_0, \end{aligned} \tag{5}$$

where $F(\cdot)$ is a fusion function defined as:

$$F(\boldsymbol{\alpha}_{t+T}) = \lambda(\mathbf{\Phi}\boldsymbol{\alpha}_{t+T}) + (1-\lambda)(\mathbf{\Psi}\boldsymbol{\beta}_t). \tag{6}$$

(5) is then in the same form as previous optimization.

4.2. Eye Blinking

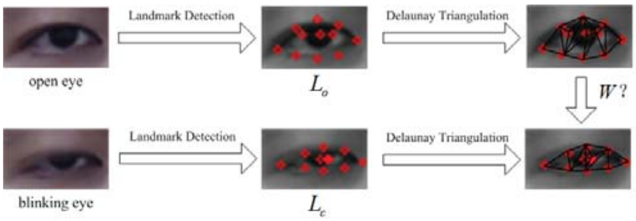

Fig. 2. The procedure of deriving transformation function

Blinking is an involuntary rapid closing and opening of the eyelid. It is an essential function that distributes moisture across the eye and remove irritants from the surface of the cornea. Blinking is especially problematic for face beautification in video, because the training image database does not contain facial images with blinking eyes.

In response we propose a two-step procedure, where the landmarks of the original captured eyes are first detected, so that the appropriate warping function W can be derived to transform a beautified open eye to a beautified blinking eye. Landmarks are feature points that outline the shapes and characteristics of the eyes. Let E_c be the original constructed blinking eye, called the *destination*, and E_o be the corresponding open eye, called the *source*. Landmarks are first detected in the source and destination eyes. As illustrated in Fig. 2, we identify 11 landmarks using a procedure described in [9]. We denote the landmark coordinates of E_o and E_c as $L_o = \{x_i, y_i\}_{i=1}^{11}$ and $L_c = \{x'_i, y'_i\}_{i=1}^{11}$, respectively. The problem is then how to obtain a suitable transformation function W which maps each landmark in E_o to E_c.

Using the detected landmarks as key feature points, we define *triangular meshes* over the points for the open eye E_o and the blinking eye E_c, respectively. Doing so means we have triangle-to-triangle correspondences. Then each triangle is warped separately from source to destination through a parametric transformation:

$$x'_i = a_1x_i + b_1y_i + c_1, y'_i = a_2x_i + b_2y_i + c_2. \tag{7}$$

The above process is a six-parameter affine transformation to map three endpoints in a triangle to another three endpoints in another triangle. The six unknown parameters can be effectively derived through six linear equations provided by endpoints of two corresponding triangle in L_o and L_c. After deriving $\{W_i\}$, it can be directly performed to the beautified open eye to obtain a beautified blinking eye. The algorithm flow is shown in **Algorithm 1**.

Algorithm 1 Algorithm of Blinking Eye Beautification

```
Input:
    The current blinking eye $E_c$, its closet open eye $E_o$, and
    the beautified eye of $E_b$;
Output: A beautified blinking eye $E_f$.
Procedure:
  Mapping Function Learning
    Landmarks detection for $E_c$ and $E_o$;
    Define a triangular mesh over landmarks;
      – Construct same mesh in both eyes;
      – Establish triangle-to-triangle correspondences;
    Warp each triangle separately from $E_o$ to $E_c$;
      – Learn an affine mapping function $W$ for each triangle
      pair;
  Blinking Eye Beautification
    Landmarks detection for $E_b$;
    Define a triangular mesh over landmarks;
    Warp each triangle separately from $E_b$ to $E_f$ using corresponding mapping function $W$.
```

5. EXPERIMENTAL RESULTS

In this section, experimental results are presented to demonstrate the performance of our proposed joint gaze-correction and beautification system for conference video.

First, we examine the effect of beautification on individual video frame. The results include two male test samples and two female samples. As shown in Fig. 3, we see that the beautified images in the second row are more attractive than the original images in the first row. In particular, beautified subjects have enlarged eyes, and the male subjects have more pronounced eyebrows. Note that in each of these examples, the differences between the original face and the beautified one are quite subtle, and thus the resemblance between the two faces is unmistakable. Yet the subtle changes clearly have a noticeable impact on the attractiveness of these faces.

We now examine the issues of temporal consistency and eye blinking in video. In Fig. 4, we show the results of two test videos. By comparison of successive frames of these two test videos, we can observe that the proposed method achieves desired temporal consistency for video conferencing. Moreover, our method achieves satisfactory results for frames containing blinking eyes. The blinking process is smooth and natural. The reviewers are invited to examine the demo video provided as supplemental materials. In the demo video, the right side is the beautified result.

6. CONCLUSION

We consider joint gaze correction and face beautification for conference video. First, we redefine the search space of fa-

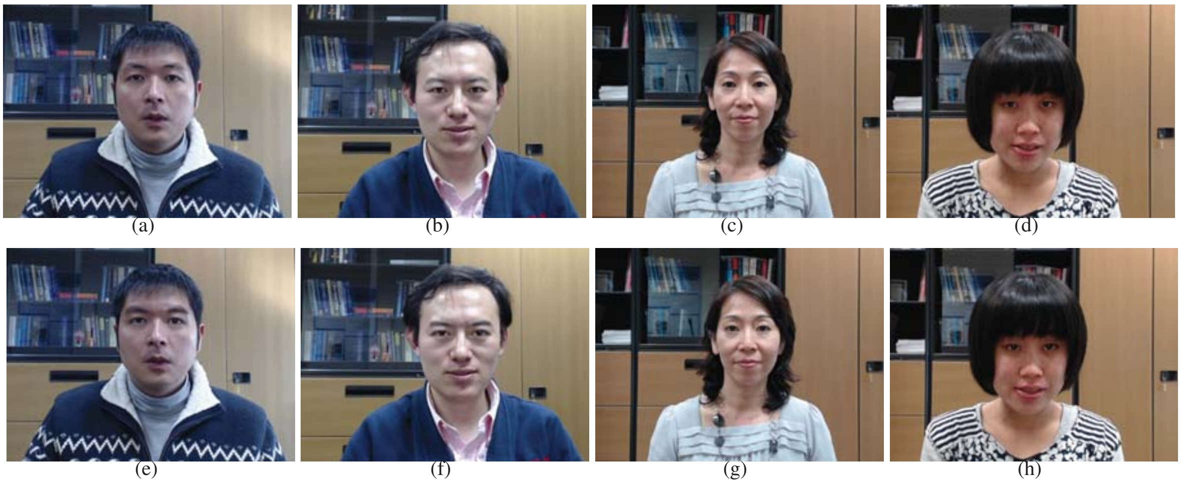

Fig. 3. Results of face reconstruction and beautification. (a)-(d) are the reconstruction results, (e)-(h) are the corresponding beautified results.

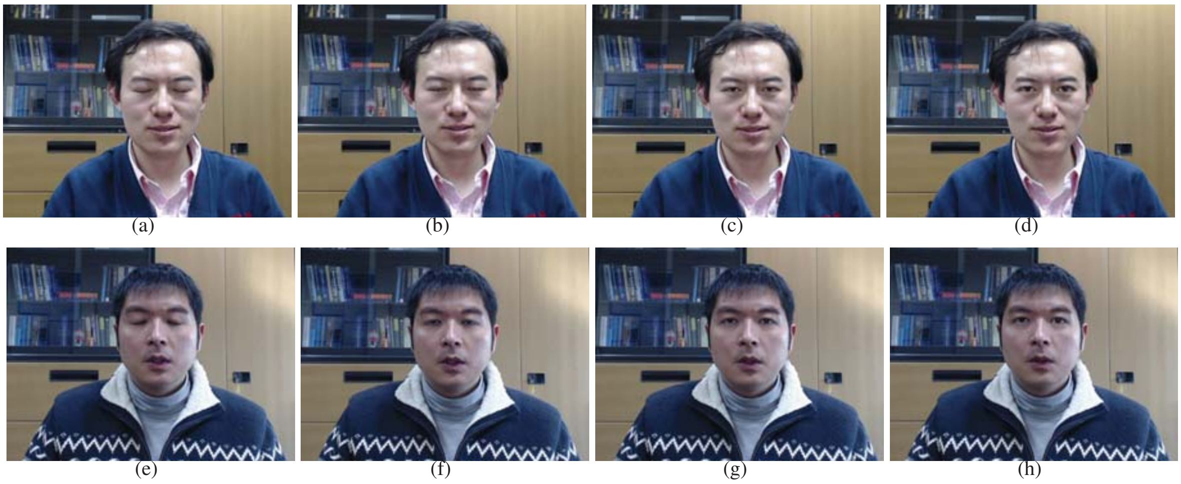

Fig. 4. Temporal consistency test for a blinking process. (a)-(d) are the successive frames in Test Video 1, (e)-(h) are the successive frames in Test Video 2.

cial component beautification to allow for a larger feasible solution set, and propose a fast optimization procedure for this new formulation. We then address two practical issues unique to video: temporal consistency and eye blinking. Experimentations show consistent and natural looking faces with enhanced attractiveness for video can be synthesized.

7. ACKNOWLEDGEMENTS

This work is supported by the Major State Basic Research Development Program of China (973 Program 2015CB351804), the National Science Foundation of China under Grants 61300110.

8. REFERENCES

[1] J. Gemmell, k. Toyama, C. L. Zitnick, T. Kang, and S. Seitz, “Gaze awareness for video conferencing: A software approach,” in *IEEE Multimedia*, 2000, vol. 7, no.4, pp. 26–35.

[2] A. Blake P.H.S. Toor A. Criminisi, J. Shotton, “Gaze manipulation for one-to-one teleconferencing,” in *IEEE ICCV 2003*.

[3] R. Yang and Z. Zhang, “Eye gaze corection with stereovision for video-teleconferencing,” in *IEEE Trans. on Pattern Analysis and Maching Intelligence*, July 2004, vol. 26, no.7, pp. 956–960.

[4] X. Liu, G. Cheung, D. Zhai, D. Zhao, H. Sankoh, and S. Naito, “Joint gaze-correction and beautification of DIBR-synthesized human face via dual sparse coding,” in *IEEE ICIP 2014*.

[5] D. Lowe, “Object recognition from local scale-invariant features,” in *IEEE ICCV 1999*.

[6] D. Tian, P.-L. Lai, P. Lopez, and C. Gomila, “View synthesis techniques for 3D video,” in *Applications of Digital Image Processing XXXII, Proceedings of the SPIE*, vol. 7443 (2009).

[7] Y. Yang, B. Geng, Y. Cai, A. Hanjalic, and X.-S. Hua, “Object retrieval using visual query context,” in *IEEE Trans. on Multimedia*, December 2011, vol. 13, no.6, pp. 1295–1307.

[8] I. Daubechies, R. Devore, M. Fornasier, and S. Gunturk, “Iteratively re-weighted least squares minimization for sparse recovery,” in *Commun. Pure Appl. Math*, 2009.

[9] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in *IEEE CVPR 2012*.