DOI: 10.1111/cgf.13261

COMPUTER GRAPHICS *forum*
Volume 37 (2018), number 1 pp. 214–225

Enhancing the Realism of Sketch and Painted Portraits With Adaptable Patches

Yin-Hsuan Lee[1], Yu-Kai Chang[1], Yu-Lun Chang[1], I-Chen Lin[1,2], Yu-Shuen Wang[1,2] and Wen-Chieh Lin[1,2]

[1]Institute of Multimedia Engineering National Chiao Tung University, Taiwan
hsuan.cs99g@nctu.edu.tw, ttbbs542.cs97@g2.nctu.edu.tw, alan512200@gmail.com, {ichenlin, yushuen, wclin}@cs.nctu.edu.tw
[2]Department of Computer Science, National Chiao Tung University, Taiwan

Abstract

Realizing unrealistic faces is a complicated task that requires a rich imagination and comprehension of facial structures. When face matching, warping or stitching techniques are applied, existing methods are generally incapable of capturing detailed personal characteristics, are disturbed by block boundary artefacts, or require painting-photo pairs for training. This paper presents a data-driven framework to enhance the realism of sketch and portrait paintings based only on photo samples. It retrieves the optimal patches of adaptable shapes and numbers according to the content of the input portrait and collected photos. These patches are then seamlessly stitched by chromatic gain and offset compensation and multi-level blending. Experiments and user evaluations show that the proposed method is able to generate realistic and novel results for a moderately sized photo collection.

Keywords: facial modelling, matting & compositing

ACM CCS: I.3.3 [Computer Graphics]: Picture/Image Generation, I.4.3 [Image Processing and Computer Vision]: Enhancement—Registration

1. Introduction

People have relied on portraits to record appearances for thousands of years. Stylized or exaggerated portraits, such as illustrations or comics, enable artists to express their imaginations. Police continue to make use of sketch portraits to visualize the descriptions provided by witnesses. Viewers are often curious about the real appearance of the person depicted in a portrait. This is likely one reason why many people worldwide have acted in roles portraying historical people and participated in cosplay, mimicking the appearance of cartoon characters.

In this paper, we therefore investigate how to use image synthesis techniques to enhance the realism of a stylized portrait while preserving personal characteristics. We propose that this technology, termed *face realization*, is suitable for educational and entertainment usages. For instance, it can be applied to enhance the realism of historical portraits or cartoon characters. Because its input is rough sketches or paintings, users can apply this kind of technique to game avatar creation or person search for forensic usage.

A closely related topic is image style transfer. The classic image analogies [HJO*01] transformed an input image into a specific style according to the local structural mapping extracted from a pair of unfiltered and filtered examples. Recently, Wang *et al.* [WCHG13] stylized photos into paintings through stroke-based rendering. They used a pair of examples to determine the mappings between photo textures and stroke properties, such as colour, orientation and density.

These example-based techniques have demonstrated admirable results in photo stylization. However, two difficulties arise in face realization. First, people are more familiar with real faces than with paintings. A synthesis method must therefore be able to render subtle facial details. Secondly, painting styles or skills can vary widely in different paintings. To realize a portrait by learning and mapping, we must prepare numerous training pairs of photos and portraits in all required styles to determine their particular mappings.

In our work, we aim to enhance the realism of greyscale sketches and chromatically painted portraits. When inspecting these portraits,

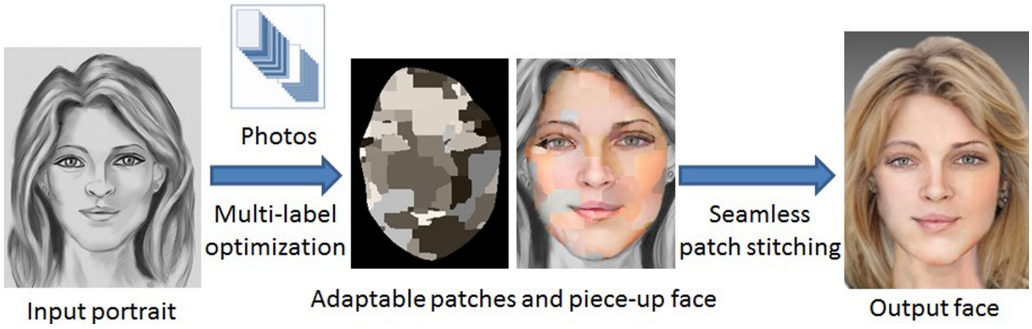

Figure 1: *Example of the enhancement of portrait realism. The proposed system extracts the adaptable patches from a moderately sized photo collection. The patch colours in the middle left image represent patches from different sources. The middle right image is the face pieced together from patches. Chromatic adjustment and seamless blending enable the realization of cartoon, sketch or painted portraits that retain their personal characteristics.*

we observed that the contours of facial features and relative shading of skin usually imply the characteristics of a subject; they should therefore be retained during enhancement. However, a portrait's realism is typically judged by the delicate textures of the facial features and skin. Hence, we find useful features in images and transfer regional textures from a moderately sized collection of facial photos.

We first attempt to divide a target portrait into regular or pre-defined regions and retrieve the most similar photo regions from a database for texture transfer. However, regular or pre-defined divisions do not always accurately capture the characteristics of a target portrait or photo. For instance, the use of rectangular divisions may separate a mouth into multiple regions and result in discontinuity during synthesis. If the entire mouth is considered as a region, a huge database will be required in order to include the wide variety of mouth appearances.

Therefore, we propose a novel facial region matching and division approach, termed *adaptable patches*, which can dynamically adjust the shapes and numbers of patches according to both the target portraits and database photos. We explain the concept with a sketched mouth example, where the leftmost one-third is shaded and the rightmost two-thirds are illuminated. If there is a facial photo I_α, of which the intensity distribution of the mouth conforms to this input sketch, we take the whole mouth as a patch and transfer the mouth texture from I_α. If there is no such photo, but there are photos I_β and I_γ, in which the mouths are entirely shaded and glossy, respectively, we prefer to divide the input mouth into two patches. The patch on the leftmost one-third is appended with the texture from the corresponding region at I_β, and the other patch on the rightmost two-thirds is appended with the regional texture from I_γ. In contrast to conventional approaches, the proposed strategy can not only capture facial characteristics but also more effectively exploit variations within data sets.

The adaptable patches are extracted from different sources, and the tones of adjacent patches can be quite different. We present chromatic gain and offset compensation to diminish chromatic gaps at patch boundaries. The aforementioned patching and compensation procedures can be formulated as pixelwise multi-labelling and regionwise quadratic optimization problems, respectively, which are automatically estimated by the proposed system.

Figure 1 shows an example of our face realization process. To demonstrate the ability of our chromatic gain and offset compensation, we include two greyscale photos in the data set. We compare our method with two related methods. Subsequent user evaluations show that our method can not only generate more realistic results but also retain characteristics of the input portraits.

2. Related Works

Tiddeman *et al.* [TSP05] studied facial image enhancement by transforming facial images through a wavelet-based Markov random field (MRF) method, which can be applied for purposes, such as facial ageing and sex reversal. Pighin *et al.* [PHL*98] combined the geometries and textures of example models in convex vector space. This type of blend shape system can generate diverse expressions according to the blending weights applied. However, high-resolution facial details can be blurred during the blending process. Nguyen *et al.* [NLEDlT08] proposed an image-based method for automatic beard shaving. They compared the parameters of beard layers by using a set of bearded and non-bearded faces. This information was used to remove the beard from an input face. Chen *et al.* [CJZW12] relighted the input portrait according to another reference image selected by users. The corresponding illumination template was then transferred to the input.

A few articles have investigated face synthesis through using whole face replacement. Blanz *et al.* [BSVS04] replaced the input facial image with another face by using a morphable face model. They not only fit the face pose but also adjusted the illumination parameters. Bitouk *et al.* [BKD*08] proposed replacing a target face by another face according to the pose, colour, lighting and blending cost. These methods have generated impressive results through whole face replacement, but they cannot be applied in novel face generation.

Other research has explored regional or local patch replacement. Mohammed *et al.* [MPK09] utilized frontal and well-aligned facial photos in their Visio-lization system. This method consists

of a global model and a local one. A base image is generated in principle component analysis (PCA) space and then divided into rectangular patches. In local fitting, a patch is chosen for replacement according to its visual consistency with the base face and existing patches in its left and upper sections. Finally, the selected patches are stitched together by modified Poisson image editing [PGB03]. Suo *et al.* [SZSC10] decomposed a face into pre-defined regions. They generated aged faces by replacing regions with samples from a large database. With sufficiently large photo sets, the regular and pre-defined region replacement methods generate convincing results. However, collecting a large and well-aligned database is time-consuming and labour-intensive. In contrast to the aforementioned methods [MPK09, SZSC10], we propose a framework that dynamically adjusts patch numbers, sizes and shapes according to different targets and database faces. The proposed method can generate varied results from a moderate amount of photo data.

For stitching two or more images together, Efros and Freeman [EF01] proposed quilting two selected image blocks along the minimum cost path at their overlap region. Capel and Zisserman [CZ98] evaluated homographies between images for mosaicing. In addition to blending overlap images, they presented a *maximum a posteriori* (MAP) method to estimate the super-resolution mosaic. Lee *et al.* [LGMt00] aligned frontal and side-view head images according to control feature lines, and merged them into a single texture map by using multi-resolution pyramids. The commonly used Poisson image editing [PGB03] takes the overlap regions as the boundary constraints and seamlessly stitches the images together by retaining their original intensity divergences. This method performs satisfactorily when the gradients at the source and target borders are similar. However, if there are abrupt colour changes, such as shadows near patch boundaries, the shadow colours might be propagated to the adjacent regions. Agarwala *et al.* [ADA*04] grouped and combined segments from multiple photos according to user-assigned sparse strokes. Their work and ours both utilize multi-label graph-cut to segment regions from multiple sources; however, the objectives are different. Our research aims to develop a means of automatically identifying the adaptable patch set that is the best fit for both the input painting and collected photo data.

For low-fidelity image enhancement and matching, the pioneer work by Freeman and Pasztor [FP99] modelled the super-resolution problem through labelling in an MRF. Their network parameters were statistically learned from training samples. Wang and Tang [WT09] presented a patch-based sketch-photo synthesis method with pairs of examples. This state-of-the-art method formulated the selection of rectangular patches in a two-layered MRF and solved the MAP. The selected patches were stitched along their minimum-error boundary cut. A multi-dictionary sparse representation model was used by Wang *et al.* [WGTL11] for example-based sketch-photo synthesis. Liang *et al.* [LSX*12] presented a synthesis method for simple line drawings. Line features were represented by the BiCE descriptor [Zit10]. Given an input sketch, K candidate photo patches were approximated by locality-constrained linear coding. The result photo patches were selected based on MRF with the neighbouring compatibility function considered. Johnson *et al.* [JDA*11] used a large collection of photographs to add detail to computer-rendered realistic images. They used a mean-shift co-segmentation algorithm to match CG image regions with photographs, and then transferred colour, tone and texture.

Shrivastava *et al.* [SMGE11] presented a method of matching paintings or sketches to real photographs. To conduct a cross-domain search, they performed pattern matching mainly on discriminative objects. Klare and Jain [KJ10] discussed various distances for sketch-to-photo matching based on SIFT features [Low04]. Their experiments demonstrated that the direct matching (L_2-$norm$) of sketch and photo features is at least as accurate if not more so than that of matching through common representation derived from training pairs. The fusion of the two distances is slightly more favourable than the direct matching.

In a reversal of methods, Kim *et al.* [KSL*08] converted real faces into sketches. They presented an automatic method of analysing a tone and a line map to produce stipple renderings from photographs. Wu *et al.* [WTLL13] adapted the stippling style from real paintings by Seurat. Their statistical colour model enabled the conversion of an image to a pointillist painting. Zhao and Zhu [ZZ11] collected photos of real faces and painting strokes provided by artists. For a given input photo, they found the closest face in terms of geometry and colour, and then warped the corresponding strokes for painting synthesis. Chen *et al.* [CLR*04] presented a facial sketch rendering method based on prototype pairs of photos and sketches. They divided a face into multiple components and used the k-nearest neighbours (KNN) algorithm for photo-to-sketch synthesis of components. Furthermore, because hairstyles are diverse and are not structured in the same regular way that faces are, they synthesized the hair independently from the face. We also considered these characteristics and our system separated each image into hair, face and neck layers.

3. Pre-processing and Overview

Unlike related work synthesizing for specific types of paintings, our goal is to realize painted or sketch portraits. We chose not to require training pairs of photos and paintings for all styles. We collected a moderate amount of photos from the Internet to form the database. Each photo has a clear face with a sufficient number of pixels. The Stacked Trimmed Active Shape Model (STASM) [MN08] is applied to extract the sparse feature points from photos. The positions of these feature points can be interactively improved by users.

The flowchart of the proposed framework is shown in Figure 2. Before proceeding towards the main stages, we perform several pre-processing operations. Because the viewpoints and facial feature contours of the collected photos are different from those of the input portrait, we adapt perspective projection warping [Sze06] to preliminarily align the viewpoints. Perspective warping uses corresponding point pairs to evaluate the least square homography, but the warped feature points may not exactly align with the points of the input painting. We therefore apply local field warping to amend the slight offsets. Figure 3 demonstrates the alignment of faces to the input portrait. The supplementary document provides further details on feature points and local warping.

The main stages of the proposed framework are *realization* and patch *stitching*. The goal of the realization stage is to form a pieced-up face with photo patches according to the input portrait. The

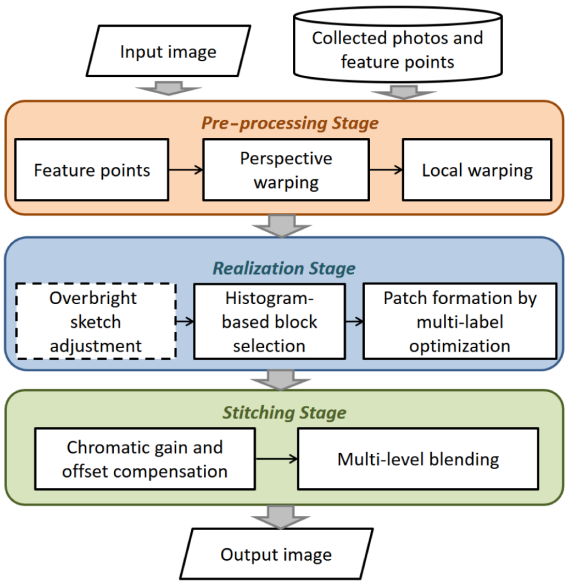

Figure 2: *Flow chart of the proposed system.*

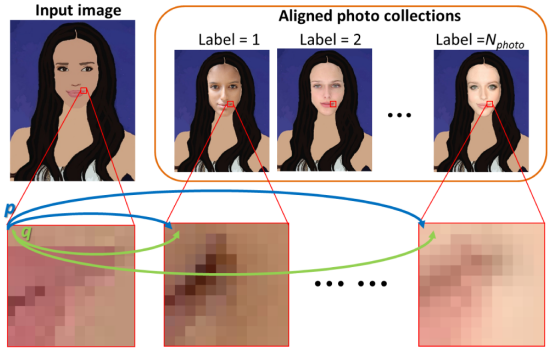

Figure 3: *Aligned database photos to fit the input face. After alignment, an input pixel p can be associated with one appropriate pixel at the same position in the photo collection. For instance, the blue arrows represent possible matches for p; the green arrows represent possible matches for q.*

middle of Figure 1 shows an example. The first step in realization is brightness adjustment for an overbright sketch input. Our system then promptly filters out less relevant parts in photo sets according to colour histograms. The last step of realization is to find the adaptable patches through optimization.

The intensity gaps between adjacent patches are removed in the stitching stage. Chromatic gain and offset compensation is presented for border gap reduction, where the three-channel colour gains and means of the each extracted patch are optimally adjusted. Multi-level blending [BA83] is then used to make the boundary seamless while retaining the high-frequency textures.

Through the realization and stitching stages, the proposed method enhances the realism of the input portrait paintings, including greyscale sketches, cartoons and coloured portraits. The details of these two stages are described in Sections 4 and 5.

4. Realization with Adaptable Patches

The core of our face realization, adaptable patch formation (Subsections 4.3 and 4.4), is based on the appearances of the input paintings and collected photos. We found that brightness normalization for input sketches (Subsection 4.1) can improve the synthesis results.

4.1. Overbright sketch adjustment

In a few greyscale sketches, artists drew only the facial contours and left the facial skin blank or excessively bright. This causes the system to match only the high-brightness regions in photos. Therefore, we analyse the intensity histogram of a sketched face. If the mean intensity of the face is higher than the average mean of the photo data by one standard deviation, our system automatically adjusts its intensity to conform to the intensity limitations, as follows:

$$k = \left\lfloor \frac{m - M}{\sigma} \right\rfloor, \qquad m' = m - k \times \sigma, \tag{1}$$

where m is the intensity mean of the input sketch face, M and σ are the average mean and standard deviation from photo data, $\lfloor\ \rfloor$ is the floor operator, k is the adjusting scale with respect to σ and m' is the mean after adjustment. This step is only applied to greyscale sketch inputs, and is also an option provided to users.

4.2. Histogram-based block selection

The major realization procedure is to concurrently select the best-fitting photo content from multiple sources. Its computation time is proportional to the number of photos involved in the optimization. To make the example data set expandable and keep the computation time feasible, highly relevant photos and their partial regions, called *blocks*, must be selected before patch optimization.

Assume that the width of the input face is w and w_p is one-third of w. A block size is specified to be $w_p \times w_p$ pixels. When a block is placed at every half w_p pixels in both x and y directions, the input face can be covered by overlapping blocks. The residual pixels around the bounding box of the face are attached to the closest blocks. As shown in Figure 4, for each block of the input face, we select 30 similar blocks from the aligned faces of the photo set. The similarity between two blocks is measured according to greyscale histograms for sketches and colour histograms for colour paintings. A pixel on the input face can be covered by four overlapping blocks at most, and each block has 30 similar photo blocks. Hence, one pixel on the input face has at most 120 candidate photo pixels from similar photo blocks. If we want to formulate the candidate selection as a pixelwise labelling problem, we need 120 label indices for each pixel. The next task is to find the criteria regarding pixel labels that can help us form adequate patches for the piece-up face.

4.3. Optimal patch formation

An intuitive approach is to find the labels with pixelwise minimum intensity (colour or greyscale) differences from the input. However,

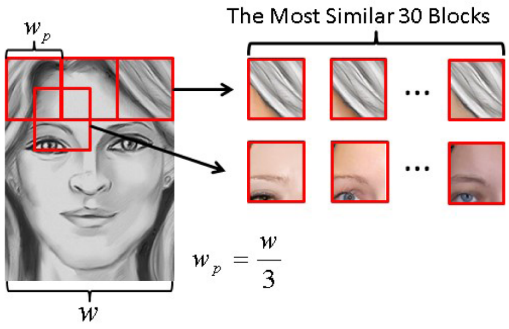

Figure 4: *Relevant region filtering by block selection. An input face is covered by overlapping blocks. Each block of the input face matches the 30 most similar blocks from the aligned photo faces according to their histograms.*

if we consider only the intensity differences between the input and corresponding pixels in photo data, the results will be highly similar to the input painting, with little enhancement. In addition, when the minimum-difference labels in a local region are inconsistent, the transferred colours in proximity can be diverse, and it introduces additional disturbance in the image.

Instead of finding the best-fitting pixels individually, we find the patches in which the shapes are best fitted to the local appearances of the input and certain photos. In other words, we further enforce local consistency of labels to ensure that the local texture can be transferred from fitted photos onto the input face. We formulate this task as a multi-labelling problem in a conditional random field (CRF). The input image I is represented by a weighted graph $G = (P, N)$. Each node $p \in P$ in the graph G stands for one pixel of the input face. The pairwise adjacent pixels are linked by graph edge $(p, q) \in N$. A four-connected neighbourhood is used in our case. Our preliminary objective function is formulated as follows:

$$O(f) = \sum_{p \in P} D_p(f_p) + \sum_{p,q \in N} S_{p,q}(f_p, f_q), \tag{2}$$

where $\sum D_p$ and $\sum S_{p,q}$ are the data and smoothness (local coherence) terms described in the following section.

4.4. Penalty terms in the objective function

4.4.1. *Data term*

Intensity features (*colour or greyscale*) The data term in Equation (2) sums up the intensity differences between the pixels of the input painting and the labelled (referred) photos. A pixel p on the input face has a variable label f_p, which represents an index of photo blocks covered on p. The source photo ID of block f_p is denoted by F_p. The data term of a pixel p with label f_p can be defined as

$$D_p(f_p) = \frac{\|I_{\text{input}}(p) - I_{F_p}(p)\|_1}{w_c}, \tag{3}$$

where $I_{\text{input}}(p)$ and $I_{F_p}(p)$ denote the intensities (colour or greyscale according to the input portrait) at location p in the input portrait and the aligned photo with the ID F_p, respectively. The division of weight w_c normalizes the colour difference range to [0, 1] (255 for greyscale and 765 for colour images). During the minimization of Equation (2), the data term requires an intensity of pixel p at the referred photo ID F_p similar to that of pixel p in the input. The selection of an inadequate f_p incurs a large penalty.

Edge features In a sketch or painted portrait, artists usually use salient strokes or significant intensity changes to emphasize facial features. Hence, in addition to evaluating the intensity differences listed in Equation (3), we consider an additional edge feature for the data comparison. Including the edge feature makes the patch formation process place greater emphasis on the fittingness of contours and characteristics of facial features. It can also mitigate edge discontinuity when we composite a facial feature from multiple sources.

We apply the Canny detector to find the edges of the input face and all database photos, as shown in Figure 5. The edge intensity of a pixel marked by the Canny edge detector is initially set to 255; those of other pixels are set to zero. Next, we smoothly diffuse the edge intensities. Currently, our system applies a 3×3 Gaussian filter 10 times; it can be replaced with a larger filter or other diffusion methods. The propagated edge intensities provide shift tolerances in edge matching. The new data term with the edge feature becomes

$$D_p(f_p) = \frac{\|I_{\text{input}}(p) - I_{F_p}(p)\|_1}{w_c} + \rho \frac{\|E_{\text{input}}(p) - E_{F_p}(p)\|_1}{w_c}, \tag{4}$$

where ρ is a weight (1 in our case) indicating the proportion of edge differences in the data term. E_{input} and E_{F_p} are the propagated edge maps of the input and the photo of ID F_p.

Figure 6(b) provides an example of the enhanced face without the edge feature in the data term. The eyebrow resembles the input but its contour is distorted and edges are discontinuous. As shown in Figure 6(c), when the edge feature is included in the objective function, the contours are smoother and conform to the input.

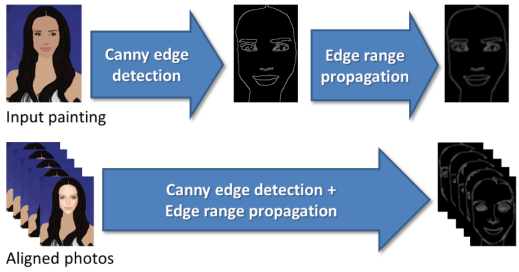

Figure 5: *Propagated Canny edge values as an additional feature. In addition to three colour channels (greyscale for sketches), we include propagated edge contours in the data term.*

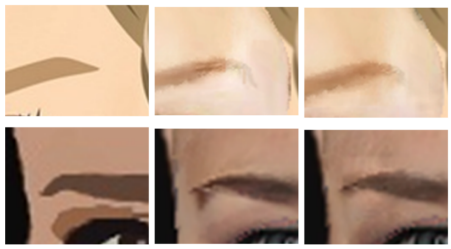

Figure 6: *Realism enhancement with the edge contour feature in the data term. (a) The input painting. (b) The enhanced result without edge features. (c) The enhanced result with edge features.*

4.4.2. *Smooth term*

The smooth term in Equation (2) is used to maintain the label consistency within a local region. It penalizes two neighbouring pixels p and q if their colours are similar in their indexed photos, but they are assigned different labels. f_p and f_q denote the variable labels of input pixels p and q, and their corresponding photo IDs are F_p and F_q, respectively.

For two adjacent pixels p and q, we assume that the penalty (cost) of edge (p, q) should be non-zero if the photo ID $F_p \neq F_q$, and this penalty should be bidirectionally equal. In addition, when the indexed (photo) colours of p and q are similar, we prefer to choose an identical source for local affinity; otherwise, a larger penalty is added. The smooth term in the objective function is defined as

$$S_{p,q}(f_p, f_q) = \begin{cases} 0, & \text{if } F_p = F_q, \\ \lambda \exp\left(-\frac{\|I_{F_p}(p) - I_{F_q}(q)\|_1}{w_c}\right), & \text{otherwise,} \end{cases} \quad (5)$$

where λ is a weight that controls the local affinity. A larger λ implies a stronger constraint on local coherence, and the patch sizes become larger and the influence of the input colour decreases. The constant w_c is applied to range normalization again. In our system, the default value of λ is 0.2. It is boosted to 2.0 when the pixel p and q belong to eye mask regions determined by eye feature points.

4.4.3. *Symmetry term*

The retrieved patches that optimize Equation (2) are satisfactory in most cases. However, human eyes are sensitive to symmetry in features such as the eyes and mouth. Because the minimum of Equation (2) does not address this characteristic, an asymmetrical appearance may be generated. For instance, the two eyes in Figure 7(c) are retrieved from different sources. Therefore, we apply the third type of penalty, called the *symmetry term*, to reinforce this characteristic.

After the feature points of the input face are estimated (Section 3), symmetry edges are applied to connecting nodes of predefined symmetry feature points in graph G. For instance, as shown in Figure 8, an additional edge between eye corners p and r renders two nodes special neighbours. The symmetric neighbourhood is

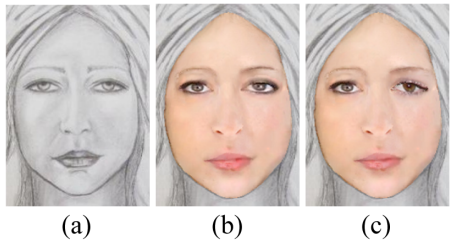

Figure 7: *Effect of the symmetry term in optimization. (a) The input face; (b) the enhanced face with the symmetry term and (c) the enhanced face without the symmetry term. The patch stitching has been applied to (b) and (c).*

denoted by M. The new objective function with the symmetry term becomes

$$O(f) = \sum_{p \in P} D_p(f_p) + \sum_{(p,q) \in N} S_{p,q}(f_p, f_q) + \sum_{(p,r) \in M} U_{p,r}(f_p, f_r). \quad (6)$$

For $(p, r) \in M$, the symmetry term of our objective function is defined as

$$U_{p,r}(f_p, f_r) = \begin{cases} 0, & \text{if } F_p = F_r, \\ \mu \exp\left(-\frac{\|I_{F_p}(p) - I_{F_r}(r)\|_1}{w_c}\right), & \text{otherwise,} \end{cases} \quad (7)$$

where f_r and F_r are the block index and source photo ID of pixel r. μ is a weight about the strength of symmetric penalty, and the setting of μ is the same as λ. The constant w_c enables range normalization. For a pair of symmetry nodes p and r, if their labels are identical, we consider them to be inherently symmetric and no penalty is enforced. If they have different labels, the symmetry penalty is added according to the colours in their indexed sources.

After solving Equation (6) by the alpha-expansion method [BVZ01], we approximate the optimal labels of pixels on the input face. Adjacent pixels with an identical source ID are grouped as a patch. The shapes, sizes and numbers of these adaptable patches can be dynamically fitted according to the characteristics of the input face and photo data.

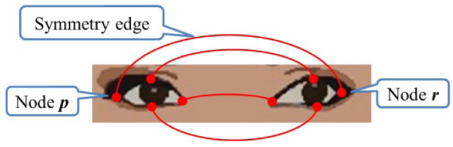

Figure 8: *Symmetry edges link symmetric feature points at eye or lip corners. These edges enforce symmetric penalties if the linked nodes p and r are from different sources.*

5. Seamless Patch Stitching

Because photo collections comprise multiple subjects whose illumination and makeup differs, facial photos in databases have a variety of colour tones. In the stitching stage, we apply chromatic gain and offset compensation and multi-level blending to seamlessly stitch the extracted patches.

5.1. Chromatic gain and offset compensation

Brown and Lowe [BL07] presented a gain compensation method to narrow the intensity gaps between overlapping panoramic images. Because the scene captured in a set of panoramic images is identical, the differences in intensity in overlap regions result from changes in the aperture and time of exposure. By contrast, our patches are derived from different sources. Both intensities and hues can vary. Our early trials demonstrated that adjusting the lightness gain for each patch is not always sufficient to narrow down the colour differences in our cases. Therefore, we concurrently adjust three gain variables and three offset variables of all colour channels of each patch. The procedure is as follows.

First, for each patch, the original region is expanded (by five pixels in our cases), causing adjacent patches to overlap. An index pair of two adjacent patches is denoted by $(i, j) \in H$, where H is a set including all pairwise patches with overlap regions. Assume that $H = (h_1, \ldots, h_n)$ contains n pairs. The indices within a pair h_k are denoted by h_k^i and h_k^j. The overlap region of a pair of patches h_k is denoted by $r(h_k)$. The proposed system then computes the mean vectors of colour intensities of patch h_k^i and h_k^j in their overlap region $r(h_k)$. These two mean vectors are represented by $m(h_k^i)$ and $m(h_k^j)$.

Each expanded patch has a three-channel gain vector a and a three-channel offset vector b. For a patch h_k^i, if we simply fit its mean colour to one of its adjacent patches h_k^j, the difference in $r(h_k)$ is reduced. However, if this patch also belongs to another pair h_l, the difference in $r(h_l)$ could increase. Hence, we must concurrently consider all the gain and offset vectors of all patches. The chromatic gain and offset compensation involves finding the optimal a and b of all patches that can minimize the sum of mean colour differences for all overlapping pairs h_k. This concept can also be formulated as an objective function:

$$
\begin{aligned}
O_{cmp} &= G(A, b) + kC(A, b), \\
G(A, b) &= \sum_{h_k \in H} \Big| A\left(h_k^i\right) m\left(h_k^i\right) + b\left(h_k^i\right) - A\left(h_k^j\right) m\left(h_k^j\right) \\
&\qquad - b\left(h_k^j\right) \Big)\Big|^2 \\
C(A, b) &= \left(\sum_i |A(i) - I_{3x3}|^2 + \sum_i |b(i)|^2 \right), \qquad (8)
\end{aligned}
$$

where $A(h_k^i)$ is the 3×3 diagonal matrix accounting for chromatic gains, and their diagonal elements are $a(h_k^i)$. The first term G in Equation (8) evaluates the sum of the adjusted colour mean differences between overlapping pairs. The second term C keeps the gain and offset vectors close to their original values. k is a constant weight that controls the strength of this constraint. k is 1 in our setting. $I_{3\times3}$ is an identity matrix. Equation (8) is a quadratic least square equation that can be solved by a linear system. The optimized gain and offset vectors are then used to adjust the colour of each patch. The third column of Figure 9 shows the adjusted patches by chromatic gain and offset compensation.

Figure 9: *Seamless patch stitching. (From left to right) The input sketch, the piece-up face, the face after chromatic gain compensation and the face after multi-level blending. Chromatic gain and offset compensation can even narrow the gaps between patches in greyscale and colour photos.*

5.2. Multi-level blending

We further use multi-level blending to smooth the patch boundaries. A Laplacian pyramid is used to divide the overlap regions defined in Subsection 5.1 into multiple levels. The proposed system then blends the low-pass images within overlap regions to smooth the colour gaps. Facial details can be recovered by restoring the high-pass images. The blending weights are proportional to the reciprocal of a small constant plus the square distance from the patch centre. Please refer to [BL07, BA83] for the blending details.

6. Experiments and Discussion

This section demonstrates and compares the results of the proposed method with those generated by related methods. It also discusses the advantages and limitations of these methods. Please refer to our supplementary document for additional experiments regarding other characteristic labels. The document further compares the proposed method with a dedicated sketch-to-photo method.

6.1. Data collection and implementation details

The photo set was collected from publicly shared albums on the Internet. Our main targets were photos of young women. Two hundred photographs were collected for the face and neck database. We found that the hair styles in these 200 photos could only be categorized into a few types. To include more diverse hair styles, we separately collected 90 photographs for the hair data set. Our system utilized several public libraries. The STASM [MN08] library was used for facial feature point extraction. The Catmull–Rom spline

was applied to indicate the contours of hair and facial features. To solve multi-label optimization, we used the graph-cut optimization library provided by Veksler and others [BVZ01, KZ04, BK04]. We also applied the OpenCV library for image processing. The experiments were performed on a desktop with 3.1 GHz CPU and 3.5 GB memory. On average, it took 13 minutes to optimize an input image with 400 × 600 pixels. The computation time could be shortened with further optimized coding or parallel computation.

6.2. Experimental results and comparisons

The proposed algorithms were applied to enhance diverse types of input portraits from the Internet, including monochromatic sketches, coloured cartoons and paintings. In most cases, we did not know the identities or real faces of the subjects. For monochromatic inputs, we used the greyscale intensities and propagated edge values in our data term; for coloured inputs, the three-channel intensities and edge values were applied in the data term. By contrast, in the smooth and symmetry terms, the colour intensities of indexed photos were used for both monochromatic and coloured inputs. Figure 10 shows examples of the overbright sketch adjustment described in Subsection 4.1. The second column of Figure 11 shows our results for sketch, cartoon and painted portraits. More results are provided in the supplemental video.

We compared our results with those enhanced by two methods that are described later. The database for these methods is identical to ours; likewise, these photos were aligned by perspective and local warping during pre-processing. In addition to RGB colour channels, we included the propagated edge feature mentioned in Subsection 4.4.1 to improve the results of the compared methods. Figure 12 shows an example of the influence of our proposed edge feature on a compared method. The use of an identical feature set helped us to clarify the effectiveness of different algorithms.

The two compared methods are *best-matched-face* and *regular-patch-based* methods. The best-matched-face method finds the most similarly aligned face from the database according to intensity and edge features. This method can be regarded as face enhancement through whole face replacement. The regular-patch-based method

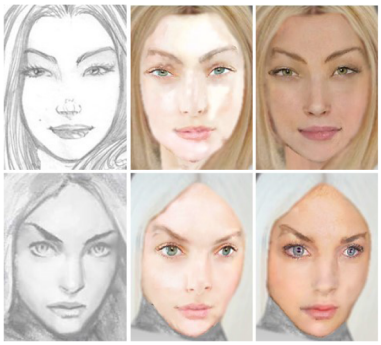

Figure 10: *Automatic adjustment for the overbright inputs. The first column shows the input portraits. The second column shows the results without brightness adjustment and the last column shows the results after adjustment.*

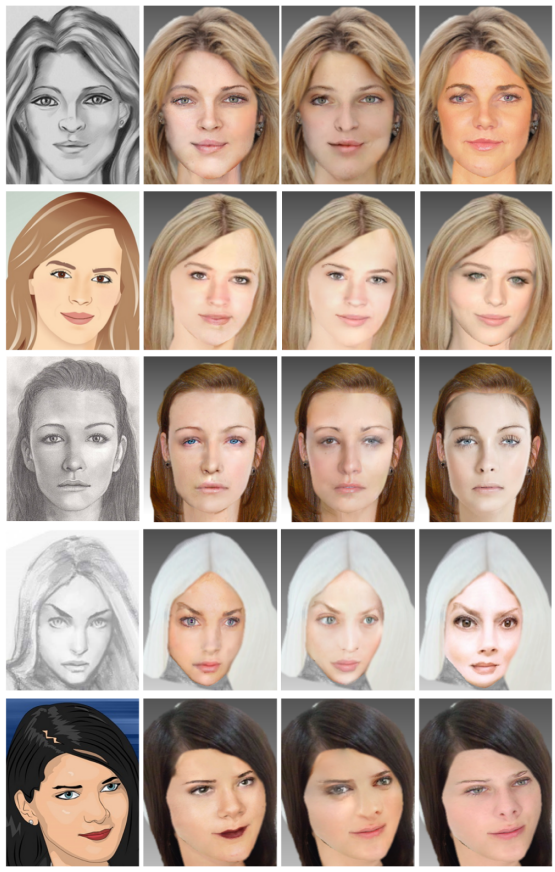

Figure 11: *Realizing faces from cartoons, sketches and stylized paintings. The first to the fourth columns are, respectively, the* ***input portraits****, results of* ***the proposed*** *method, results of the* ***regular-patch-based*** *method and results of the* ***best-matched face*** *method.*

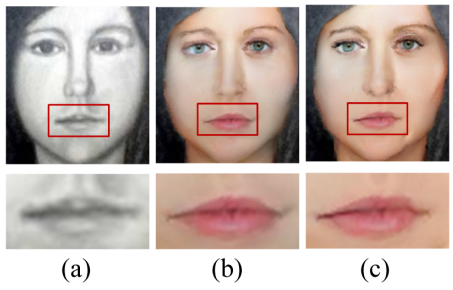

Figure 12: *Influence of the proposed edge feature on the compared regular-patch-based method. (a) The input sketch. (b) Result of the regular-patch-based method with edge features. (c) Result of the regular-patch-based method without edge features.*

sequentially finds the rectangular photo patches that are the closest matches for the input portraits and photo patches that have been attached in the left and upper sections. The patch symmetry of facial features is also considered. Except for the newly included edge feature, the implementation is nearly identical to the synthesis stage proposed in the original Visio-lization method [MPK09].

Nevertheless, in Visio-lization, all extracted photo patches are stitched in a single layer. We found that hair colours can occasionally be propagated to the face regions during Poisson image editing. In addition, the hairlines are usually blurry or discontinuous. Therefore, in our implementation, we separated an input head into three layers: face, neck and hair. The images of these layers were estimated separately and then superposed from the bottom to the top. This three-layer strategy was also applied to the proposed and best-matched face methods. These three enhancement methods can be directly applied to the face and neck layers. By contrast, the hair synthesis involves more complex occlusion and illumination. To focus on the face enhancement problem, we applied the best-matched hairstyle to the hair layer of all three methods. The results of the best-matched-face and regular-patch-based methods are shown in the third and the fourth columns of Figure 11.

6.3. Discussion of the three methods

The best-matched-face method is the most intuitive means of enhancing face realism. Because this method captures the whole warped texture from an identical source, it retains high realism with few stitching artefacts. However, the results obtained by this method can only reach rough similarity to the input. Important local features, such as contours or shading of the eyes and lips, might be completely different.

The regular-patch–based method is an extension of the synthesis component of the Visio-lization method [MPK09]. It was originally applied to face synthesis involving a large frontal photo database. For a moderately sized data set, the sizes and locations of regular patches must be carefully designed. If the patch size is too large, discontinuity may occur at patch boundaries. By contrast, a patch that is too small can make the enhanced faces blurry after stitching. The results of this method have higher similarity to the input portrait than those enhanced by the best-match-face method. However, the usage of regular patches could decrease the realism of the image.

By contrast, the proposed method utilizes optimization methods for patch formation and stitching. It can capture the characteristics of the input portrait and has fewer stitching defects. To verify the effectiveness of the proposed method, we conducted user evaluations of the enhanced results of the three methods.

6.4. User evaluations

In conventional image processing or compression applications, peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are commonly used to evaluate the quality of processed images. However, for face realization applications, the input paintings are stylized and sometimes exaggerated, and the colours on a painting are restricted by the pigments. Even with ground truth images, these two measures are still unsuitable for evaluating realized faces. Therefore, we asked users to evaluate the results of the three methods.

Twenty-four volunteers aged 20–25 years old participated in the evaluation of 18 sets of test data. Each test data set contained an input portrait and three enhanced results obtained by different methods. For each test set, we simultaneously showed the input face and the enhanced results. The three results were placed in random order on the screen. Volunteers did not know the enhancement method for each result. A volunteer was required to rate two qualities of each face on a scale of 1 to 10 (2: very poor; 4: poor; 6: acceptable; 8: good; 10: perfect). The first quality was the *similarity* between the input face and the enhanced face. This score measured how well the personal characteristics of an input portrait were retained after enhancement. The second quality was the *realism* of the resulting face, which measured the enhancement quality and was inversely related to the degrees of defects detected by users. We also required volunteers to rank their overall preferences for the three results. The evaluation time for each volunteer was approximately one half-hour to an hour. Table 1 demonstrates the means and standard deviations of the three methods in similarity and realism evaluations. Figure 13 shows the first, second and third preference vote counts for the three methods.

Table 1: *Scores of user studies.*

Methods	Score mean	Standard deviation
Similarity of *The Proposed*	7.200	0.645
Similarity of *Regular-Patch–Based*	6.072	0.727
Similarity of *Best-Matched*	5.778	1.076
Realism of *The Proposed*	7.150	0.598
Realism of *Regular-Patch–Based*	6.161	0.758
Realism of *Best-Matched*	6.800	0.937

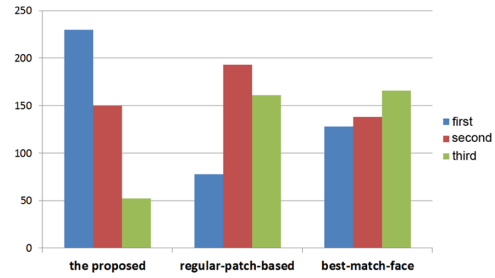

Figure 13: *Vote counts for three methods according to overall preference ranking. Twenty-four volunteers participated in the study of 18 test data sets. The blue, red and green bars represent the amounts of first, second and third preference votes, respectively. From left to right: the proposed method, regular-patch-based method and best-matched-face method.*

The results of the proposed method received the highest average scores in both similarity and realism, as well as the highest number of first-preference votes. The standard deviations of the scores were also smaller. This means our method involves a more consistent enhancement process than the other two methods. We estimated whether the user evaluation scores were statistically different through paired two-tailed t-tests and p values. For the similarity scores, the t and p values between our scores and those of the regular-patch–based and the best-matched-face methods were $t(17) = 5.21$, $p < 0.001$ and $t(17) = 4.06$, $p < 0.001$, respectively. Our similarity scores significantly outperformed the comparative scores. For the realism scores, there was a highly significant

difference between the proposed and regular-patch-based methods, with $t(17) = 4.71$, $p < 0.001$. However, the realism difference between our method and the *best-matched-face* method was less significant, with $t(17) = 1.15$, $p > 0.1$. The voting data of each test set are provided in the supplementary document.

6.5. Advantages, uniqueness and limitations

Unlike related sketch-to-photo synthesis methods [WT09, WGTL11, LSX*12] that require training pairs of photos and sketches in a dedicated style, the proposed framework requires only a modest amount of photo data to effectively enhance portraits. The realization process is mainly formulated as a graph-based multi-labelling problem. Labelling in MRFs or CRFs has been developed in several classic works for various purposes, such as super-resolution [FP99], disparity evaluation [BVZ01] and interactive photomontage [ADA*04]. Our intention is to find patches of variable numbers and shapes to effectively explore the variations within the photo collection. Facial characteristics such as symmetry features can also be embedded within the graph. The pixelwise labelling mitigates the block effects in the regular-patch–based method [MPK09, WT09]. In some cases with obvious viewpoint changes, the perspective warping cannot handle new areas of occlusion or disocclusion. The optimized patch formation can implicitly avoid extracting data from such regions if their tones or edges are different from those of the input portrait.

We included the propagated edge feature with three-channel colour intensities as the pixel features. This alleviates discontinuous contours and matches a greater amount of detail variation. When the edge feature was applied to the comparative methods, it also improved their enhanced results. The proposed chromatic gain and offset compensation reduces the patch gaps from photos of various sources. Figure 14 shows an example of the piece-up face by stitching together 57 different sources. Figure 14(a) represents patches of different sources by different colours, and (b) shows the sources used in this case. These photo sources are generally different from the input sketch. Two of them are in greyscale. However, our enhanced face is highly realistic and retains the personal characteristics of the input sketch, as shown in Figure 1. We applied our method to two additional data sets [WT09, MB98] and demonstrated our results and the results of [WT09] in the supplementary document. Even though we did not use the training sketches, some of our results may be comparable to those generated by a dedicated sketch-to-photo method.

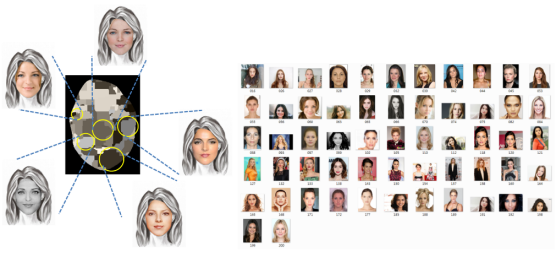

Figure 14: *Adaptable patches from multiple sources. (a) An example of a piece-up face and their aligned sources. (b) The 57 source photos were extracted from a set of 200 photos for the piece-up face in (a). Please refer to Figure 1 for the input and the enhanced image.*

Our system has a few limitations. First, it is a data-driven method. Even though we can find more fitting patches, the variety of enhanced faces is still constrained by the database. Secondly, we treat the input portrait as the foundation. If the facial structure or shading of the input painting is unusual, satisfactory results might not be generated. Third and finally, we do not explicitly estimate the camera view angles, and the operations are performed in the image space. Currently, we can process faces with view angles of up to 20°. This limitation could be overcome with facial data that incorporates 3D information (e.g. data from depth sensors).

7. Conclusion and future work

Enhancing the realism of 'unreal' facial paintings is an interesting but difficult task. These sketches or painted portraits are usually stylized and are created using various techniques. This paper proposes forming adaptable patches according to the content of the input image and collected photos. The numbers and shapes of these patches are dynamically adjusted by multi-label optimization. These patches are then seamlessly stitched by the chromatic gain and offset compensation and blending. User evaluations show that the proposed method is able to provide realistic results, similar to the subject, with a photo collection of moderate size. Possible future research directions include the application of texture analysis such as [LLC15] for salient feature extraction and the extension of directional texture synthesis such as [JFA*15] for complicated hair or beard synthesis.

Acknowledgements

The authors would like to acknowledge the providers of all the images and photographs published on the Internet. These images are for academic and non-commercial usage only. The authors also thank the volunteers who participated in the user evaluations. This paper was partially supported by the Ministry of Science and Technology, Taiwan, under grant no. MOST 104-2221-E-009-129-MY2.

References

[ADA*04] Agarwala A., Dontcheva M., Agrawala M., Drucker S., Colburn A., Curless B., Salesin D., Cohen M.: Interactive digital photomontage. *ACM Transactions on Graphics 23*, 3 (2004), 294–302.

[BA83] Burt P. J., Adelson E. H.: A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics 2*, 4 (1983), 217–236.

[BK04] Boykov Y., Kolmogorov V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*, 9 (2004), 1124–1137.

[BKD*08] Bitouk D., Kumar N., Dhillon S., Belhumeur P., Nayar S. K.: Face swapping: Automatically replacing faces in photographs. *ACM Transactions on Graphics 27*, 3 (2008), 39:1–39:8.

[BL07] Brown M., Lowe D. G.: Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision 74*, 1 (2007), 59–73.

[BSVS04] Blanz V., Scherbaum K., Vetter T., Seidel H.-P.: Exchanging faces in images. *Computer Graphics Forum 23*, 3 (2004), 669–676.

[BVZ01] Boykov Y., Veksler O., Zabih R.: Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 11 (2001), 1222–1239.

[CJZW12] Chen X., Jin X., Zhao Q., Wu H.: Artistic illumination transfer for portraits. *Computer Graphics Forum 31*, 4 (2012), 1425–1434.

[CLR*04] Chen H., Liu Z., Rose C., Xu Y., Shum H.-Y., Salesin D.: Example-based composite sketching of human portraits. In *NPAR '04: Proceedings of the 3rd International Symposium on Non-Photorealistic Animation and Rendering* (Annecy, France, 2004), ACM, pp. 95–153.

[CZ98] Capel D., Zisserman A.: Automatic mosaicing with super-resolution zoom. In *CVPR '98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Santa Barbara, CA, 1998), IEEE Computer Society, pp. 885–891.

[EF01] Efros A. A., Freeman W. T.: Image quilting for texture synthesis and transfer. In *Proceedings of International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (Los Angeles, CA, USA, 2001), ACM, pp. 341–346.

[FP99] Freeman W. T., Pasztor E. C.: Learning low-level vision. In *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2* (Kerkyra, Corfu, Greece, 1999), IEEE Computer Society, pp. 1182–1189.

[HJO*01] Hertzmann A., Jacobs C. E., Oliver N., Curless B., Salesin D. H.: Image analogies. In *Proceedings of International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (Los Angeles, CA, USA, 2001), ACM, pp. 327–340.

[JDA*11] Johnson M. K., Dale K., Avidan S., Pfister H., Freeman W. T., Matusik W.: Cg2real: Improving the realism of computer generated images using a large collection of photographs. *IEEE Transactions on Visualization and Computer Graphics 17*, 9 (2011), 1273–1285.

[JFA*15] Jamriška O., Fišer J., Asente P., Lu J., Shechtman E., Sýkora D.: Brushables: Example-based edge-aware directional texture painting. *Computer Graphics Forum 34*, 7 (2015), 257–268.

[KJ10] Klare B., Jain A. K.: Sketch-to-photo matching: A feature-based approach. In *Proceedings of SPIE 7667, Biometric Technology for Human Identification* (Orlando, Florida, US, 2010), vol. VII, p. 766702.

[KSL*08] Kim D., Son M., Lee Y., Kang H., Lee S.: Feature-guided image stippling. *Computer Graphics Forum 27*, 4 (2008), 1209–1216.

[KZ04] Kolmogorov V., Zabin R.: What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*, 2 (2004), 147–159.

[LGMt00] Lee W., Gu J., Magnenat-thalmann N.: Generating animatable 3D virtual humans from photographs. *Computer Graphics Forum 19*, 3 (2000), 1–10.

[LLC15] Lin I.-C., Lan Y.-C., Cheng P.-W.: SI-Cut: Structural inconsistency analysis for image foreground extraction. *IEEE Transactions on Visualization and Computer Graphics 21*, 7 (July 2015), 860–872.

[Low04] Lowe D. G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*, 2 (Nov. 2004), 91–110.

[LSX*12] Liang Y., Song M., Xie L., Bu J., Chen C.: Face sketch-to-photo synthesis from simple line drawing. In *APSIPA '12: Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (Hollywood, CA, USA, 2012), pp. 1–5.

[MB98] Martinez A., Benavente R.: The AR face database. CVC Technical Report No. 24, June 1998.

[MN08] Milborrow S., Nicolls F.: Locating facial features with an extended active shape model. In *Proceedings of European Conference on Computer Vision* (Marseille, France, 2008), Springer, pp. 504–513.

[MPK09] Mohammed U., Prince S. J., Kautz J.: Visio-lization: Generating novel facial images. *ACM Transactions on Graphics 28*, 3 (2009), 57:1–57:8.

[NLEDlT08] Nguyen M. H., Lalonde J.-F., Efros A. A., Dela Torre F.: Image-based shaving. *Computer Graphics Forum 27*, 2 (2008), 627–635.

[PGB03] Pérez P., Gangnet M., Blake A.: Poisson image editing. *ACM Transactions on Graphics 22*, 3 (2003), 313–318.

[PHL*98] Pighin F., Hecker J., Lischinski D., Szeliski R., Salesin D. H.: Synthesizing realistic facial expressions from photographs. In *ACM SIGGRAPH 1998* (1998), ACM, pp. 75–84.

[SMGE11] Shrivastava A., Malisiewicz T., Gupta A., Efros A. A.: Data-driven visual similarity for cross-domain image matching. *ACM Transactions on Graphics 30*, 6 (2011), 154:1–154:9.

[Sze06] Szeliski R.: Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Vision 2*, 1 (2006), 1–104.

[SZSC10] Suo J., Zhu S.-C., Shan S., Chen X.: A compositional and dynamic model for face aging. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 3 (2010), 385–401.

[TSP05] Tiddeman B., Stirrat M., Perrett D. I.: Towards realism in facial image transformation: Results of a wavelet MRF method. *Computer Graphics Forum 24*, 3 (2005), 449–456.

[WCHG13] Wang T., Collomosse J. P., Hunter A., Greig D.: Learnable stroke models for example-based portrait painting. In *Proceedings of British Machine Vision Conference (BMVC)* (Bristol, UK, 2013), pp. 9–13.

[WGTL11] Wang N., Gao X., Tao D., Li X.: Face sketch-photo synthesis under multi-dictionary sparse representation framework. In *Proceedings of Sixth International Conference on Image and Graphics ICIG* (Hefei, Anhui, China, 2011), pp. 82–87.

[WT09] Wang X., Tang X.: Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 11 (Nov. 2009), 1955–1967.

[WTLL13] Wu Y.-C., Tsai Y.-T., Lin W.-C., Li W.-H.: Generating pointillism paintings based on Seurat's color composition. *Computer Graphics Forum 32*, 4 (2013), 153–162.

[Zit10] Zitnick C. L.: Binary coherent edge descriptors. In *Proceedings of European Conference on Computer Vision (ECCV): Part II* (Heraklion, Crete, Greece, 2010), Springer-Verlag, pp. 170–182.

[ZZ11] Zhao M., Zhu S.-C.: Portrait painting using active templates. In *NPAR '11: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering* (2011), ACM, pp. 117–124.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web site:

Figure S1: Feature points and alignment of faces and necks.

Figure S2: Feature points and alignment of hair.

Figure S3: Experiments with the CUFS data set.

Figure S4: Experiments for targets with glasses or beards in the AR Face data set.

Figure S5: Experiments with the AR Face data set.

Video Data S1