Adaptive, Selective, Automatic Tonal Enhancement of Faces

Hrishikesh Aradhye
Google, Inc.
1600 Amphitheatre Ave.
Mountain View, CA, 94043

George D. Toderici
Google, Inc.
1600 Amphitheatre Ave.
Mountain View, CA, 94043

Jay Yagnik
Google, Inc.
1600 Amphitheatre Ave.
Mountain View, CA, 94043

ABSTRACT

This paper presents an efficient, personalizable and yet completely automatic algorithm for enhancing the brightness, tonal balance, and contrast of faces in thumbnails of online videos where multiple colored illumination sources are the norm and artifacts such as poor illumination and backlight are common. These artifacts significantly lower the perceptual quality of faces and skin, and cannot be easily corrected by common global image transforms. The same identifiable user, however, often uploads or participates in multiple photos, videos, or video chat sessions with varying illumination conditions. The proposed algorithm adaptively transforms the skin pixels in a poor illumination environment to match the skin color model of a prototypical face of the same user in a better illumination environment. It leaves the remaining non-skin portions of the image virtually unchanged while ascertaining a smooth, natural appearance. A component of our system automatically selects such a prototypical face for each user given a collection of uploaded videos/photo albums or prior video chat sessions by that user. We present several human rating studies on YouTube data that quantitatively demonstrate significant improvement in facial quality using the proposed algorithm.

Categories and Subject Descriptors

I.4.3 [**Image Processing and Computer Vision**]: Enhancement

General Terms

Algorithms

Keywords

Facial Enhancement, Image Quality, Color Transfer

1. INTRODUCTION AND PRIOR WORK

A large fraction of videos uploaded on YouTube are recorded by lay persons under uncontrolled capture conditions, poor illumination, and using inexpensive webcams. This content, although home-made, does have the potential to be virally popular. Small but noticeable improvements in video and/or thumbnail quality may go a long distance towards gathering the initial critical mass of eyeballs. Thumbnail image quality is particularly critical as the decision to watch the video is largely based on the thumbnail. This paper presents one such algorithm that is automatic, personalizable, and yet simple enough to be used on millions of images and videos.

The emphasis of the proposed algorithm is on the color, brightness and contrast adjustment specifically of faces. Most commercial photo editing software tools provide the ability to automatically restore natural white balance. Such "white patch" family of approaches is restricted to a single illuminant source or multiple sources of the same type, which is inadequate for our purpose. Global white, however, balancing often has little effect on facial skintone quality, since (a) the dominant illumination in the room is often different from the light source near the face such as an LCD screen or a table lamp, and (b) the face occupies a relatively small area of the image.

To handle multiple illuminations, Hsu et. al. [1] present a white balance technique for scenes with exactly two light types that are assumed to be provided by the user. Lischinski and colleagues [4] developed a scribble interface which lets users segment images into regions lit by a single type of light. While these approaches may be feasible for personal photo albums, the user input requirement will not scale up to repositories of video data with hundreds of millions of videos.

In contrast, the proposed work does not restrict itself to any specific number or nature of illumination sources and requires no user input. It builds upon recently published work in image color transfer [5]. They transformed the pixel color as a linear combination of colors of these best-match pixels in all swatches in the image weighted by the inverse Eucledian distance. We use a PCA-based parametric color model and extend the Eucledian weighting used in [5] [7] [6] to a trilateral scheme using color and luminance distance in addition to spatial distance. This is the same concept as bilateral image filtering. It allows us to prevent erroneous recoloration of the background in spite of its complexity and the nature/number/color of illumination sources lighting the background. We select the best reference face automatically given a collection of images of the same person. Unlike [6][3], our process is computationally lightweight and completely automatic.

We propose a novel yet simple metric of face skintone quality to automatically select one or more prototypical faces given a collection of prior photographs and/or video frames involving the person of interest. We transform the facial and body skintone of a poorly lit face to better match that of the prototypical face(s). This approach is thus adaptive to a specific person, or a group of persons of a specific ethnicity or skintone category, if so modeled. This paper work addresses a well-identified commercial need not adequately

MM'09, October 19–24, 2009, Beijing, China.

addressed by the state of the art cameras, editing software, and published literature. We demonstrate the utility of this approach on a collection of real-world imagery selected from YouTube corpus.

2. PROPOSED ALGORITHM

This section provides the technical details of our approach. The subscript src refers to the source (input) image; tgt is the target image with a prototypical face; dst is the intermediate destination image; and out is the final output image.

2.1 Adaptive Prototypical Face Selection

We now discuss an automatic mechanism to pick the target image for a given person. We assume that a candidate set of images known to include the face of the person in question is available. The premise is that at least some of these images were captured under better illumination conditions and thus could be used for facial enhancement in the above scheme. Images with faces that are too bright (average normalized intensity 0.7) or too dark (average normalized intensity < 0.4) were eliminated. For each pixel within the facial bounding box of the remaining faces, we compute the probability of skin given its RGB values. To this end, we use skin and non-skin 16-kernel Gaussian Mixture Models (GMMs) [2]. We contend that the GMM cluster centers likely represent prototypical human skintones under good illumination. Consequently, assuming that all faces in our candidate set are of the same person, we select the candidate face with the highest average skin probability to be the prototypical face for the person under consideration.

The first row of figure 1 provides some example results of this prototypical face selection procedure on a YouTube video blog, with the thumbnails sorted in decreasing order of their GMM skin score described above. The thumbnail with the best score (Figure 1(1)) is correctly selected as the prototypical face selected for this user. The same person is seen in a variety of illumination conditions and backgrounds. The pose of the subject and the distance from the camera may also vary. Different videos may have been recorded using more than one camera. Some may have been post-processed using a movie editing software whereas some may have been uploaded as captured. In spite of these challenges, our procedure is provides reasonable results. Note that blue-tinged faces correctly get low scores.

For application domains where it is not possible to obtain a reliable candidate set, we assume a pre-set target face of a Caucasian female for all users. Note that a mismatch of ethnicities, genders and/or skintone categories between the source and target face is possible in this scenario. Our results indicate that tonal quality is improved in spite of such a mismatch, in part due to conservative parameters that modulate the extent of recoloration, and in part because of common characteristics of human skin distribution independent of ethnicity [2]. Further research is needed to minimize the likelihood of such a mismatch.

2.2 Facial Color Modeling Using PCA

Analogous to the approach by Xiao and Ma [6], we represent both source and target images by applying PCA in the RGB color space on pixels within the region of interest, which for our purposes is the facial bounding box. We a thus obtain a 3×3 PCA projection matrix V, a 3×3 diagonal matrix D comprised of eigenvalues, and a 1×3 mean vector μ for both source and target facial bounding boxes. For each 1×3 RGB vector I_{src} in the source image, the destination vector I_{dst} can be computed as:

$$I_{dst}(x,y) = (I_{src}(x,y) - \mu_{src})\, V_{src} D_{src}^{-1} D_{tgt} V_{tgt}^{-1} + \mu_{tgt} \quad (1)$$

To achieve accurate correspondence between source and target ellipsoids in the RGB space, we ensure that (a) matrices V and D correspond to decreasing order of eigenvectors, and (2) R component of each eigenvector in V is positive. If the R component of any eigenvector is negative, we multiply the eigenvector by -1, i.e., rotate it by 180°.

2.3 Trilateral Weighting

The non-skin portions of the source and target image could be completely different and thus should not be matched. To selectively enhance the facial regions to the fullest extent possible without introducing any artificial coloration of non-skin regions, we propose a new trilateral weighting scheme based on spatial and color distances and intensity.

Spatial Distance Weight: Modeling the face in the image as a vertical ellipse, the normalized distance of any pixel (x, y) from the center of the face can be modeled as:

$$d_s(x,y) = 2\sqrt{\frac{(x-x_c)^2}{W^2} + \frac{(y-y_c)^2}{H^2}} \quad (2)$$

where (x_c, y_c) is the center and (W, H) are the width and height of the facial bounding box in the source image. We compute an analogous distance from the center of the body region, assumed to be centered at $(x_c, y_c + 3H)$ and of dimensions $(2W, 2.5H)$. At any pixel, we take the minimum of the face and body distances. Admittedly, this modeling of the spatial positioning of the body region is simplistic but works for a majority of imagery in the domain of interest. It also lowers the likelihood of pixels on off-white walls in indoor lighting to be classified as skin.

We compute a spatial weight by modulating this distance by a tanh nonlinearity for sharper saturation:

$$w_s(x,y) = 0.5 + 0.5 tanh\left(-\alpha_s \left(d_s(x,y) - \beta_s\right)\right) \quad (3)$$

where $\alpha_s = 0.75$ and $\beta_s = 1.25$ are parameters. These default values were manually determined in the early stages of this work and were kept unchanged since then.

Color Distance Weight: Spatial distance alone is not sufficient as non-skin objects within close vicinity of the face and body (scarfs, clothes, eyeglasses, hair) should not get recolored. We thus formulate a color distance as:

$$d_c(x,y) = (I_{src}(x,y) - \mu_{src})\, V_{src} D_{src}^{-1} \quad (4)$$

We compute a color weight by using d_c in place of d_s in Equation 3 with $\alpha_c = 1.1$ and $\beta_c = 2.0$.

Grayscale Weight: To ensure that true black pixels in the image do not become gray, we formulate an gray value weight defined as a linear ramp.

$$d_g(x,y) = \begin{cases} 0 & \text{if } G_{src}(x,y) <= G_l \\ \frac{G_{src}(x,y) - G_l}{G_u - G_l} & \text{if } G_l < G_{src}(x,y) <= G_u \\ 1 & \text{if } G_u < G_{src}(x,y) \end{cases} \quad (5)$$

where $G_{src}(x,y)$ is the normalized gray value of the source image at pixel (x,y) with $G_l = 0.1$ and $G_u = 0.5$.

Weight Smoothing and Combination: We synthesize a combined weight $w = \lambda w_s w_c w_g$, where λ is the degree of enhancement, which is the overall control parameter for the facial enhancement. In our experiments, λ was set to be 0.5 which made the enhancement more subtle but still noticeable. The weight map w is then smoothed with a Gaussian kernel of radius 5.

We are now in a position to compute the output image:

$$I_{out} = (1 - w_{smooth})\, I_{src} + w_{smooth} I_{dst} \quad (6)$$

It is possible to enhance (or reduce) tonal differences by increasing or reducing the largest eigenvalue: $D_{tgt}(1,1) = \gamma D_{tgt}(1,1)$, where γ is a contrast enhancement factor. In our experiments, γ was set to be 1.5.

3. RESULTS AND DISCUSSION

Image enhancement is, by its nature, subjective. We conducted rigorous human rating experiments to be able to quantify the extent of image quality improvements over the original image, and an automatically white-balanced image using Picasa.

3.1 Datasets and Evaluation Setup

All experiments used the same default parameters provided in Section 2 and included only those images that had detectable faces. Our first dataset consists of thumbnails of top 5 results of the top 500 most popular English queries in the US on YouTube. This dataset is broad with unconstrained content. Nearly 20% of these thumbnails had detectable faces. All performance statistics presented below were computed only on those images with detectable faces. We refer to this dataset as *top500*. Our second dataset (*vlog1000*) includes top 1000 search results for the query "vlog" (short for "video blog"). Nearly 56% of these thumbnails had detectable faces. We constructed a third dataset (*user-db*) by randomly selecting up to one of the over-exposed and up to one of their under-exposed thumbnails of these bloggers.

Each pair of before enhancement/after enhancement images in a side-by-side setup was rated by five different human annotators not affiliated with the authors. Annotators were located worldwide and used their personal monitors with their own color characteristics. Each annotator rated a random subset no bigger than 25 pairs. Left-right placement in the side-by-side display was randomly determined for each pair. Possible set of annotations was "left image is better", "right image is better", or "the difference is not significant". These ratings were mapped to a numerical rating +1 if the enhanced image is deemed better than the original, -1 if deemed worse, and 0 if the difference was not significant. Accumulating across five raters, each pair got an integer score between -5 to +5.

3.2 Performance Comparison

A summary of human rating results is provided in Table 1. Columns 7 and 8 in Table 1 provide the fraction of image pairs in each experiment where all five raters agreed. Analogously, columns 5 and 6 provide fraction of image pairs with rating equal to or lower than -2 and equal to or greater than +2, respectively.

Experiments 1 and 2 compared the quality of original thumbnail against the quality of face-enhanced thumbnail with a default prototype image of a Caucasian female. It can be seen that the fraction of image pairs where the enhancement had a positive impact on image quality is far higher than cases where the impact was negative. Experiment 3 replaced the source image by the output of Picasa's "I'm Feeling Lucky" button applied on the source image. The ratings continued to be high, which underscores the value added by our face-selective image enhancement on top of global color adjustment at the image level. Global white balance lowers the need for face-specific enhancement for a small subset of images, specifically those images with only one noticeable illumination source. Thus, ratings for Experiment 3 were lower than Experiment 2.

Experiments 4 and 5 compared face-enhanced images with two different target faces. Faces in one set of images were enhanced with a one-for-all default face of a Caucasian female. Faces in the other set of images were enhanced with a personalized prototypical face selected automatically using the algorithm described in Section 2.1. This validates the strength of our prototype face selection algorithm in spite of a challenging dataset.

3.3 Example Results

The second row of figure 1 provides a set of face-enhanced thumbnails. These results were obtained using the prototype face selected automatically as per the algorithm described in Section 2.1. The extent of improvement in facial tonal quality is noticeable, with or without global white balance (e.g., Picasa's "I'm Feeling Lucky" transform). Note that that the algorithm succeeds in preserving the original coloration of non-skin objects in most cases, even when objects such as clothing and sunglasses are close to or within the facial region. Figure 2 provides more example results with a variety of YouTube users with different ethnicities and video capture environments: such as an LCD screen or a desk lamp in addition to indoor lighting (column 1), scenes with backlight, where the dominant light source is behind the camera (column 2), scenes with too little illumination (column 3), and, scenes with over-exposed faces (column 4). Some of the reasons for erroneous or unnatural coloration were incorrect selection of prototypical face, coloration bleeding into flesh-toned background close to the face, and contrast enhancement of chroma noise in the source image so that noise becomes more noticeable.

4. CONCLUSION

This work addresses an established commercial need for enhancement of facial quality in consumer applications such as online video or photo sharing and video conferences/chats. It demonstrates a novel facial quality enhancement approach that is selective to faces, can be personalized to each individual user, and is completely automatic. It can effectively mitigate common causes for unnatural facial coloration or lack of facial contrast and detail in common consumer scenarios while making no assumptions about the number/nature/color of illumination sources in the scene and the placement of the subject relative to the light source. We demonstrate the performance of the algorithm with several example images, as well as with extensive human rating studies. The algorithm is computationally efficient and is extensible to video processing. We believe that this approach could be a valuable addition to commercial photo and video processing applications.

5. REFERENCES

[1] E. Hsu, T. Mertens, S. Paris, S. Avidan, and F. Durand. Light mixture estimation for spatially varying white balance. *ACM Trans. Graph.*, 27(3), 2008.

[2] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *IJCV*, 46(1), 2002.

[3] N. Joshi. *Ph.D. Dissertation: Enhancing Photographs using Content-Specific Image Priors*. 2008.

[4] D. Lischinski, Z. Farbman, M. Uyttendaele, and R. Szeliski. Interactive local adjustment of tonal values. In *SIGGRAPH*, 2006.

[5] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5), 2001.

[6] X. Xiao and L. Ma. Color transfer in correlated color space. In *Proc. ACM Int. Conf. on Virtual reality continuum and its applications*, 2006.

[7] L. Yin, J. Jia, and J. Morrissey. Towards race-related face identification: Research on skin color transfer. *Automatic Face and Gesture Recognition, IEEE Int. Conf. on*, 0, 2004.

Table 1: Summary of Human Rating Results

Expt	Dataset	Thumb A	Thumb B	Rating ≤ -2	$\geq +2$	$= -5$	$= +5$
1	top500	orig	orig + this work (one-for-all target)	11%	42%	1%	8%
2	vlog1000	orig	orig + this work (one-for-all target)	9%	53%	1%	14%
3	vlog1000	orig + Picasa	orig + Picasa + this work (one-for-all target)	12%	43%	$< 1\%$	10%
4	vlog1000	orig + this work (one-for-all target)	orig + this work (automatic target)	8%	60%	1%	21%
5	user-db	orig + this work (one-for-all target)	orig + this work (automatic target)	12%	50%	1%	14%

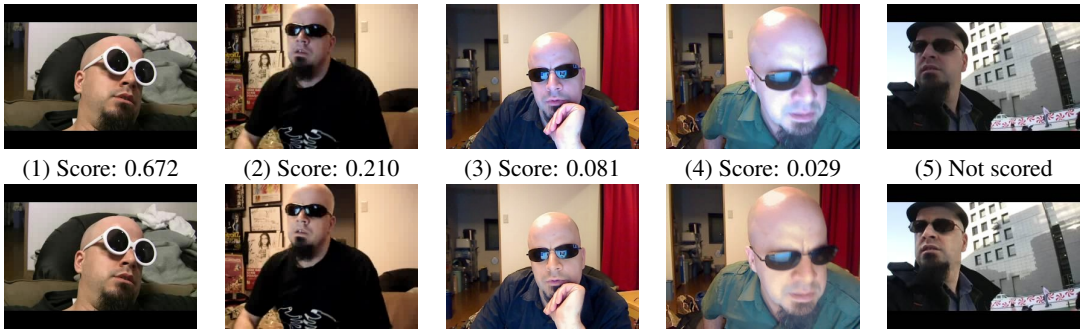

Figure 1: Adaptive Prototypical Face Selection Example: YouTube User ElevenColors. Row 1: original thumbnails sorted by decreasing scores. Faces deemed too dark or too bright are labeled *Not scored*. The top-left thumbnail (no. 1: score 0.672) was correctly selected as the prototypical face for this user. The second row shows face enhancement results with this automatically selected target image. Only a fraction of processed images are shown due to space restrictions.

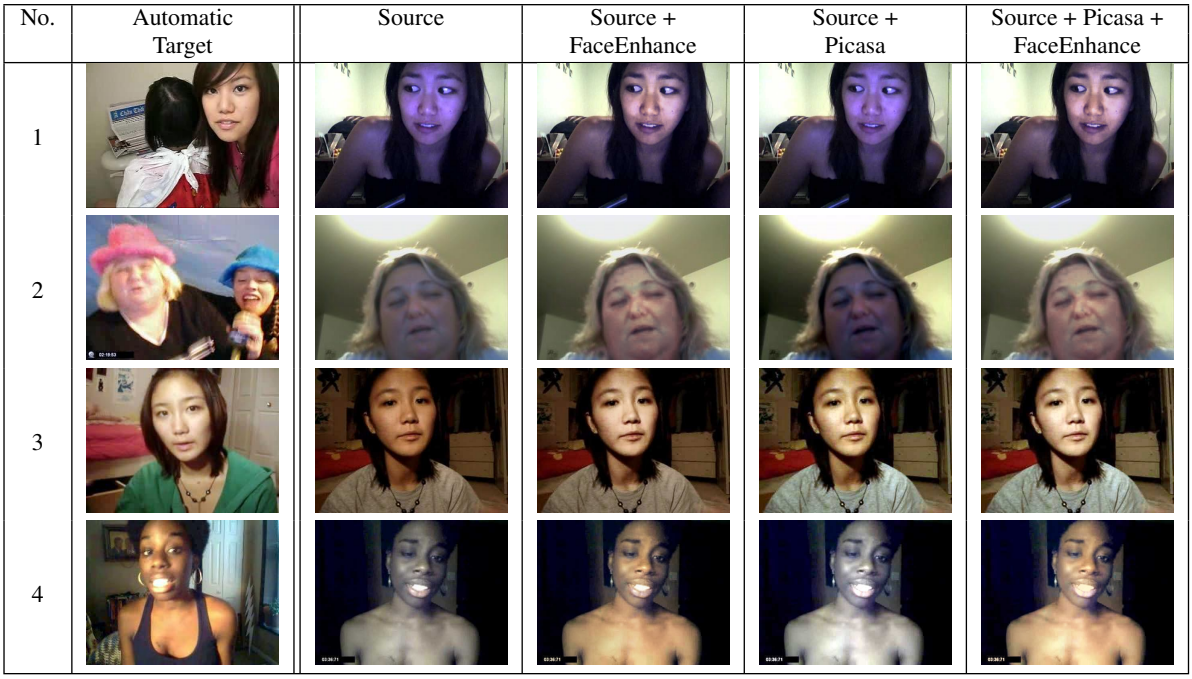

Figure 2: Face Enhancement Example Results. Each row corresponds to a unique YouTube user. The leftmost column shows the prototype image automatically selected from the videos uploaded that user (Section 2.1). The second column shows thumbnail for a separate video by the same user and is used as the source image for enhancement. The remaining columns show different enhancement results.