LEARNING-BASED PERCEPTUAL IMAGE QUALITY IMPROVEMENT FOR VIDEO CONFERENCING

Zicheng Liu, Cha Zhang and Zhengyou Zhang

Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

ABSTRACT

It is well known that in professional TV show filming, stage lighting has to be carefully designed in order to make the host and the scene look visually appealing. The lighting affects not only the brightness but also the color tone which plays a critical role in the perceived look of the host and the mood of the stage. In contrast, during video conferencing, the lighting is usually far from ideal thus the perceived image quality is low. There has been a lot of research on improving the brightness of the captured images, but as far as we know, there has not been any work addressing the color tone issue. In this paper, we propose a learning-based technique to improve the perceptual image quality during video conferencing. The basic idea is to learn the color statistics from a training set of images which look visually appealing, and adjust the color of an input image so that its color statistics matches those in the training set. To validate our approach, we have conducted user study and the results show that our technique significantly improves the perceived image quality.

1. INTRODUCTION

Online video conferencing has attracted more and more interest recently due to the improvements in network bandwidth, computer performance and compression technologies. Webcams are getting popular in home users, allowing them to chat with families and friends over the Internet for free. However, the qualities of webcams vary significantly from one to another. One common issue is that webcams tend to perform poorly under challenging lighting conditions such as low light or back light. In the past several years, we have seen a lot of progress on brightness and contrast improvement in both hardware design and image processing. Nowadays most commercially available webcams are equipped with automatic gain control. Some webcams even perform real time face tracking and use the face area as the region of interest to determine the gain control parameters. There has been a lot of image processing techniques being developed that improve the brightness and contrast of the captured images [8, 1, 4, 7].

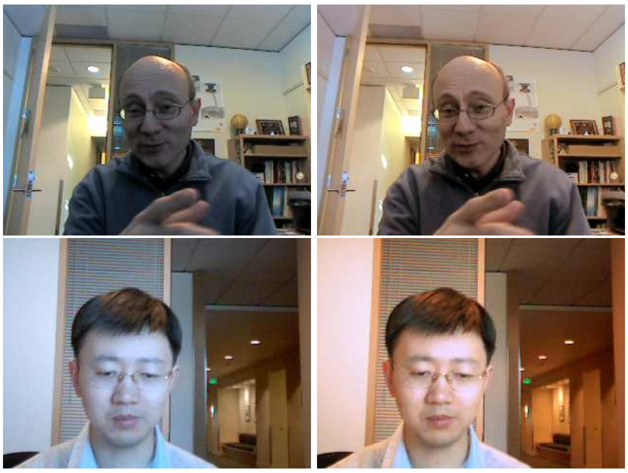


Fig. 1. Images on the left are input frames captured by two different webcams. Images on the right are the enhanced frames by the proposed method.

Exposure, however, is not the only factor that affects the perceptual quality of a webcam video. As shown in Fig. 1, the images on the left are frames directly obtained from an off-the-shelf webcam. While the exposure of these frames are both very good, they look pale and unpleasant. Instead, the images on the right are much more appealing to most users. In TV show filming, it is no secret that stage lighting has to be carefully designed to make the images look visually appealing [3]. The lighting design involves not only the brightness but also the color scheme, because the color tone is essential in the perception of the look of the host and the mood of the stage.

Since color tone is more subjective than brightness and contrast, it is difficult to come up with a quantitative formula to define what is a visually appealing color tone. Therefore, existing approaches on example-based image enhancement often requires a user to select an example image. For instance, Reinhard et al. [6] proposed a color transfer technique to impose one image's color characteristics on another. The user needs to select an appropriate target image for a given input image to conduct the color transfer. Qiu [5] proposed to apply content-based image retrieval technology to obtain a set of image clusters. Given an input image, its color can be modified by selecting an appropriate cluster as target. They showed that this technique works well for color enhancement of outdoor scenes when the user is available to select an appropriate cluster. However, they did not address the perceptual issue of face images and it is not clear how their technique can be applied to video conferencing scenario. Existing approaches on color balancing based approaches either use a color reference object or try to identify pixels with presumed colors (e.g. white and black) [9], and adjust the color scheme accordingly to restore the original color. These approaches did not address the problem of what is a visually appealing skin tone.

In this paper, we propose a data driven approach for video enhancement in video conferencing applications. The basic idea is to use a set of professional-taken face images as training examples. These images are often fine-tuned to make sure the skin-tone and the contrast of the face regions are visually pleasing. Given a new image, we adjust its color so that the color statistics in the face region is similar to the training examples. This procedure automates the enhancement process and is extremely efficient to compute. We report our user study experiments which show that the proposed method can significantly improve the perceived video quality.

2. LEARNING-BASED COLOR TONE MAPPING

At the core of our algorithm is what we call *learning-based color tone mapping*. The basic idea is to select a set of training images which look good perceptually, and build a Gaussian mixture model for the color distribution in the face region. For any given input image, we perform color tone mapping so that its color statistics in the face region matches the training examples.

Let n denote the number of training images. For each training image I_i, we perform automatic face detection [10] to identify the face region. For each color channel, the mean and standard deviation are computed for all the pixels in the face region. Let $v_i = (m_1^i, m_2^i, m_3^i, \sigma_1^i, \sigma_2^i, \sigma_3^i)^T$ denote the vector that consists of the means and standard deviations of the three color channels in the face region. The distribution of the vectors $\{v_i\}_{1 \leq i \leq n}$ are modelled as a mixture of Gaussians. Let m denote the number of mixture components. Let (μ_j, Σ_j) denote the mean vector and covariance matrix of the j'th Gaussian mixture component, $j = 1, ..., m$.

Given any input image, let $v = (m_1, m_2, m_3, \sigma_1, \sigma_2, \sigma_3)^T$ denote the means and standard deviations of the three color channels in the face region. Let $D_j(v)$ denote the Mahalanobis distance from v to j'th component, that is,

$$D_j(v) = \sqrt{(v - \mu_j)^T \Sigma_j^{-1} (v - \mu_j)}. \qquad (1)$$

The target mean and standard deviation vector for v is defined as a weighted sum of the Gaussian mixture component centers μ_j, $j = 1, ..., m$, where the weights are inversely proportional to the Mahalanobis distances. More specifically, denoting $\bar{v} = (\bar{m}_1, \bar{m}_2, \bar{m}_3, \bar{\sigma}_1, \bar{\sigma}_2, \bar{\sigma}_3)^T$ as the target mean and standard deviation vector, we have

$$\bar{v} = \sum_{j=1}^{m} w_j * \mu_j \qquad (2)$$

where

$$w_j = \frac{1/D_j(v)}{\sum_{l=1}^{m} 1/D_l(v)}. \qquad (3)$$

After we obtain the target mean and deviation vector, we perform color tone mapping for each color channel to match the target distribution. For color channel c, $c = 1, 2, 3$, let $y = f_c(x)$ denote the desired tone mapping function. In order to map the average intensity from m_c to $\bar{m}_c$, $f_c(x)$ needs to satisfy

$$f_c(m_c) = \bar{m}_c. \qquad (4)$$

In order to modify the standard deviation σ_c to match the target $\bar{\sigma}_c$, we would like the derivative at m_c to be equal to $\frac{\bar{\sigma}_c}{\sigma_c}$, that is,

$$f_c'(m_c) = \frac{\bar{\sigma}_c}{\sigma_c}. \qquad (5)$$

In addition, we need to make sure $f_c(x)$ is in the range of [0,255], that is,

$$\begin{aligned} f_c(0) &= 0 \\ f_c(255) &= 255 \end{aligned} \qquad (6)$$

A simple function that satisfies these constraints is

$$f_c(x) = \begin{cases} 0 & x < 0 \\ \frac{\bar{\sigma}_c}{\sigma_c}(x - m_c) + \bar{m}_c & 0 \leq x \leq 255 \\ 255 & x > 255 \end{cases} \qquad (7)$$

The drawback of this function is that it saturates quickly at the low and high intensities. To overcome this problem, we instead fit a piecewise cubic spline that satisfy these constraints. To prevent quick saturation at the two ends, we add constraints on the derivatives $f_c'(0)$ and $f_c'(255)$ as follows:

$$\begin{aligned} f_c'(0) &= 0.5 * \tfrac{\bar{m}_c}{m_c} \\ f_c'(255) &= 0.5 * \tfrac{255 - \bar{m}_c}{255 - m_c} \end{aligned} \qquad (8)$$

Given the constraints of Equations 4, 5, 6, and 8, we can fit a piecewise cubic spline that satisfy these constraints [2]. The details are omitted here.

Figure 2 (a) shows a tone mapping curve generated by using equation 7. Figure 2 (b) shows the curve generated by using cubic splines.

Note that the color tone mapping function is created based on the pixels in the face region, but it is applied to all the pixels in a given input image.

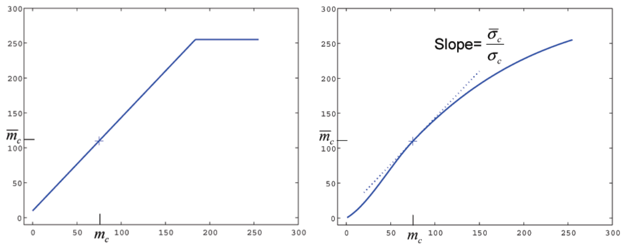


Fig. 2. Left: Tone mapping curve by using linear function. Right: Tone mapping curve by using piecewise cubic splines.

3. APPLICATION TO VIDEO SEQUENCES

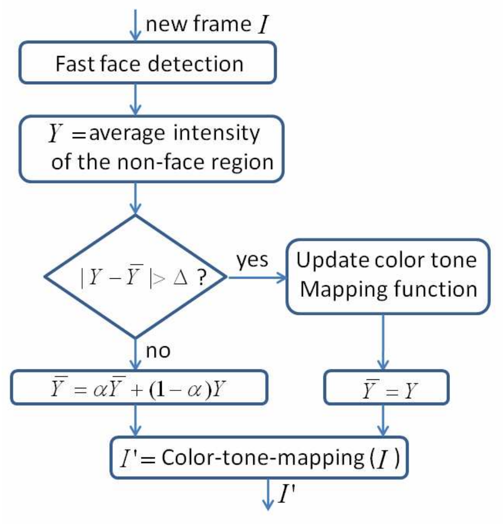
During video conferencing, the overall image intensity changes due to automatic gain control, lighting change, and camera movement. We need to update the color tone mapping function when such change occurs. For this purpose, we have developed an *overall image intensity change detector*. It works by keeping track of the mean of the average intensity of the non-face region in the history frames. Given an new frame, if its average intensity in the non-face region differs from the accumulated mean by an amount larger than a predefined threshold, it is decided that there is an overall image intensity change.

Figure 3 shows a flow chart of the algorithm where Y is average intensity of the non-face region of the new frame, $\bar{Y}$ is the accumulated mean of the frames in the history, Δ is a user defined threshold for determining whether there is an overall image intensity change, and α is a parameter that controls the update speed of $\bar{Y}$. If the current frame does not have overall intensity change, $\bar{Y}$ is updated from the current frame Y. Otherwise, we reset $\bar{Y}$ to be equal to Y.

4. EXPERIMENT RESULTS

We collected approximately 400 celebrity images from the web as our training data set . It covers different types of skin colors. We use an EM algorithm to construct a Gaussian mixture model with 5 mixture components. Figure 4 shows a sample image for each class. Please note that these images are not necessarily the most representative images in their classes. Figure 5 shows the mean colors of the five classes.

To reduce computation, we perform overall intensity change detection every two seconds. The color tone mapping is performed on RGB channels. When an overall image intensity change is detected, the color tone mapping function is recomputed for each color channel, and it is stored in a lookup table. In the subsequent frames, color tone mapping operation becomes a simple table look up which is extremely fast. Our system uses only 5% CPU time for 320×240 video with frame rate of 30 frames per second on a 3.2 GHz machine.



new frame I
Fast face detection
Y =average intensity of the non-face region
$|Y-\overline{Y}|>\Delta$?
yes
Update color tone Mapping function
no
$\overline{Y}=\alpha\overline{Y}+(1-\alpha)Y$
$\overline{Y}=Y$
I'= Color-tone-mapping (I)
I'

Fig. 3. Flow diagram of the learning-based color tone mapping algorithm applied to video sequences.

We have tested our system on a variety of video cameras and different people. We will report our user study result in the next section. Figure 1 shows an example of learning-based color tone mapping. The images on the left are input frames from two different video sequences. The images on the right are the results of learning-based color tone mapping. We can see that the resulting images are perceived as having colored lights being added to the scene that makes the scene look slightly brighter, but more importantly, having a warmer tone.

5. USER STUDY

We collected 16 teleconferencing videos with 8 different cameras, including those made by Logitech, Creative, Veo, Dynex, Microsoft, etc. Each video sequence lasts about 10 seconds. A field study was conducted, where the users were asked to view the original video and the enhanced video side by side (as shown in Figure 1). The orders of the two videos are randomly shuffled and unknown to the users in order to avoid bias. After viewing the video, the users shall give a score to each video, within the range of 1–5, where 1 indicates very poor quality, 5 indicates very good quality, and a score of 3 is considered acceptable. A total of 18 users responded in our test. All the users said they used LCD displays to watch the videos.

Figure 6 shows the average scores of the 18 users for the 16 video sequences. It can be seen that the enhanced video outperforms the original video in almost all the sequences. The average score of the original video is merely 2.55, which is below the acceptable level, while the pro-

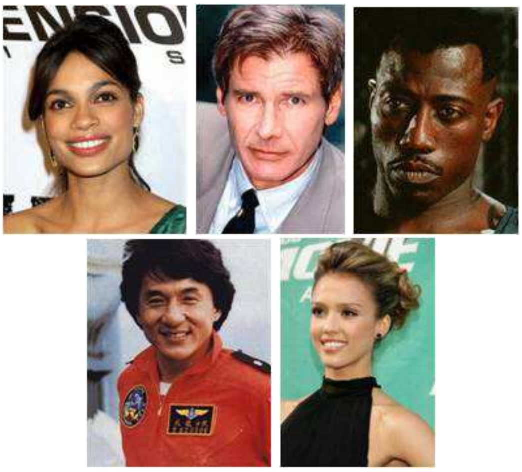


Fig. 4. A sample image from each class.

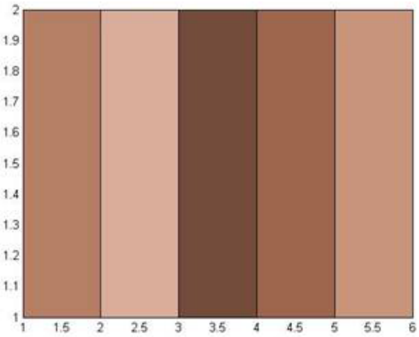


Fig. 5. Mean colors of the five classes.

posed method has a score 3.30, which is now above the acceptable level. The t-test score between the two algorithms for the 16 sequences is 0.001%, which shows that the difference is statistically significant.

6. CONCLUSION

We have presented a novel technique, called learning-based color tone mapping, to improve the perceptual image quality for video conferencing. Compared to existing image enhancement techniques, the novelty of our technique is that we adjust not just the brightness, but more importantly, the color tone. We have tested our system on a variety of webcam devices and conducted user study. The user study results show that our technique significantly improves people's perception on the video quality.

7. REFERENCES

[1] S. A. Bhukhanwala and T. V. Ramabadram. Automated global enhancement of digitized photographs. *IEEE Transactions on Consumer Electronics*, 40(1), 1994.

Sequence ID	Original Video	Enhanced Video
1	2.89	3.50
2	2.94	3. 33
3	1.94	2.67
4	1.61	2.83
5	2.11	3.00
6	3.22	4.17
7	3.11	3.72
8	1.89	2.72
9	2.11	2.17
10	1.94	2.00
11	2.11	3.44
12	3.28	3.72
13	2.65	4.18
14	2.33	3.72
15	3.00	3.89
16	3.72	3.67
Average	**2.55**	**3.30**

Fig. 6. User study results.

[2] G. Farin. *Curves and surfaces for CAGD: A practical guide*. Academic Press, 1993.

[3] N. Fraser. *Stage Lighting Design: A practical Guide*. Crowood Press, 2000.

[4] G. Messina, A. Castorina, S. Battiato, and A. Bosco. Image quality improvement by adaptive exposure correction techniques. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 549–552, Amsterdam, The Netherlands, July 2003.

[5] G. Qiu. From content-based image retrieval to example-based image processing. *University of Nottingham Technical Report: Report-cvip-05-2004*, May 2004.

[6] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 34-41, September/October 2001.

[7] F. Saitoh. Image contrast enhancement using genetic algorithm. In *IEEE International Conference on SMC*, pages 899–904, Amsterdam, The Netherlands, October 1999.

[8] C. Shi, K. Yu, J. Li, and S. Li. Automatic image quality improvement for videoconferencing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.

[9] C. T. Swain and B. G. Haskell. Color balance for video conferencing. In *IEEE International Conference on Image Processing (ICIP)*, 1997.

[10] P. Viola and M. Jones. Robust real-time object detection. In *Second International Workshop on Statistical and Computational Theories of Vision*, Vancouver, July 2001.