AUTOMATIC SCENE RELIGHTING FOR VIDEO CONFERENCING

Stephen Mangiat and Jerry Gibson

Electrical and Computer Engineering Department
University of California, Santa Barbara, CA 93106
{smangiat, gibson}@ece.ucsb.edu

ABSTRACT

This paper describes a new method to automatically improve scene lighting for video conferencing by learning the photometric mapping between a lower exposure and desired exposure created using High Dynamic Range (HDR) imaging techniques. Once the mapping is learned in a calibration step, it can be used to transform all subsequent images, effectively producing higher dynamic range video without ghosting artifacts. A stereo algorithm is also described that allows multiple exposures to be taken at every frame, useful if the lighting of a scene changes significantly. Results show that this is an effective way to improve face lighting and therefore the overall experience of video conferencing.

Index Terms— High Dynamic Range Video, Radial Basis Functions, Face Lighting, Video Conferencing

1. INTRODUCTION

Though video conferencing (VC) usage has risen dramatically in recent years, widespread adoption is still hindered by several technical limitations. One major issue that detracts from the user experience is poor lighting on the face, due to the low dynamic range of VC camera sensors. When the background is very bright the face becomes underexposed, and likewise overexposed in darker rooms. Yet, even an auto-exposure algorithm that utilizes face detection fails to correctly expose the entire face if the lighting within the face itself is highly varying, as is often the case in rooms with sunlit windows. As a result, high-end VC installations have very specific room lighting requirements.

Addressing the face lighting problem is paramount for mobile video conferencing, which can take place outdoors in high dynamic range (HDR) conditions. The mobile device scenario also makes a software solution that can work with cheap sensors preferable. Many algorithms to automatically enhance the exposure of images for VC have been proposed. The method described in [1] first establishes "important" regions within an image, such as skin, and determines a correct exposure based on these regions. Skin detection here requires the development of a skin color model, and [2] includes a way to build this model at runtime and thus adapt to current lighting conditions. Similarly, a "desirable" skin color model can be learned using a database of faces, as in [3]. After the skin areas are detected, their average gray value is used to perform global exposure correction using a function that estimates how the camera sensor converts exposure into a pixel value ([1],[4]).

Some disadvantages of these approaches are that they rely on a skin color model and use a general equation to estimate the response of the camera sensor and transform colors. These equations cannot be used to recover the color information of saturated (white) pixels. A more accurate solution used to image HDR scenes using low dynamic range sensors has been well studied and described in [5]. Here, the camera response curve is first estimated and multiple images of the same scene taken at different exposures are combined to form an HDR radiance map. This map can then be viewed on low dynamic range displays using tone mapping algorithms such as those described in [6] .

The methods used in [5] and [6] work well with static scenes, but are not directly applicable to video or moving scenes because it is impossible to combine multiple exposures without exact correspondences between the images. A method described in [7] addresses this by taking successive frames with alternating high and low exposures, and creates a radiance map by compensating for the global and local motion between frames. Consequently, the quality of the output is limited by the quality of the registration, and occlusions and other regions with poor correspondence can lead to ghosting artifacts.

The method proposed in this paper eliminates ghosting and the need to alternate exposure at every frame by combining radiance map recovery [5] and tone mapping [6] with a learned photometric mapping, as described in [8]. Specifically, during a calibration step at the beginning of a video chat (or any time initiated by the user), low and high exposure frames are taken and combined to form an image with "optimal" color balance. During this short calibration (minimum two frames), it can be assumed that the scene is static, and therefore image registration is not needed. Once the HDR im-

This research has been supported by the California Micro Program, Applied Signal Technology, Cisco, Sony Ericsson and Qualcomm, Inc., and by NSF Grant Nos. CCF-0429884, CNS-0435527, and CCF-0728646.

age is obtained, the photometric mapping between the lower exposure and this desired image can be learned using Radial Basis Functions (RBFs). The advantage here is that if future frames are taken at the same lower exposure, the same mapping can be used to transform them without any additional information. Furthermore, imaging the scene at a lower exposure ensures that none of the face will be overexposed and color information is retained.

In what follows, Section 2 will provide a summary of the HDR imaging and tone mapping processes used to create the desired scene lighting for the single camera case. Then, a brief discussion of RBFs in Section 3 is followed by a description of both a single camera and stereo camera implementation in Section 4. Finally, we will discuss our initial results and some conclusions.

2. HDR TONE MAPPING

Multiple exposures of a scene are often used to reconstruct a high dynamic range radiance map, which is then displayed using tone mapping. The method described in [5] combines various exposures according to a weighting function derived from a learned camera response curve. For simplicity, our implementation uses only two exposures (low and high), and weights the lower exposure twice as much.

Once the two exposures are combined, a global tone mapping procedure outlined in [6] is used to map all pixels back into displayable range. First, the average log-luminance is calculated by

$$\overline{L}_w = \exp\left(\frac{1}{N}\sum_{x,y}\log(\delta + L_w(x,y))\right), \tag{1}$$

where L_w is the luma component or Y channel in YCbCr, and N is the number of pixels. Next, the entire image is scaled according to

$$L(x,y) = \frac{a}{\overline{L}_w} L_w(x,y) \tag{2}$$

so that $\overline{L}_w$ maps to a desired key value a, which we set to .5. A contrast enhancement is also performed by compressing high luminances with

$$L_d(x,y) = \frac{L(x,y)\left(1 + \frac{L(x,y)}{L_{white}^2}\right)}{1 + L(x,y)}, \tag{3}$$

where L_{white} is the smallest luminance mapped to white, normally set to the maximum luminance in the radiance map. Because colors can become desaturated when imaging at low and high exposures, we also subtly increase the color saturation of the tone mapped image.

3. RADIAL BASIS FUNCTIONS

Radial basis functions are next used to learn the photometric mapping between the lower exposure image and the tone mapped HDR image. The advantage of learning this mapping is that it can be applied to other frames that do not have multiple exposure information. Non-parametric regression using RBFs has been used in other contexts such as image colorization and seamless mosaicking [8]. The goal is to estimate a function f, given a training set of input-output mappings $x_i \rightarrow y_i$ where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$. This leads to the minimization of [8]

$$H(f) = \sum_{i=1}^{N}(f(x_i) - y_i)^2 + \lambda\phi(f), \tag{4}$$

which is composed of a data fitting term and a regularizing term $\phi(f)$. The general solution can be derived as

$$f(x) = \sum_{i=1}^{N} w_i h(x, x_i) + \sum_{j=1}^{q} d_j \psi_j(x). \tag{5}$$

Here $h(x, x_i)$ are the radial basis functions, which depend only on the radial distance from the centroid ($h(x, x_i) = h(\|x - x_i\|)$) and ψ_j are the basis functions of their null space. As in [8], we use Gaussian RBFs, whose null space is empty, and the weights can then be solved with

$$w = (H + \lambda I)^{-1} y, \tag{6}$$

where H is an $N \times N$ matrix with $h_{ij} = h(\|x - x_i\|)$, I is an identity matrix, and λ is the regularization parameter. Because N is very large, the number of basis functions is restricted to a much smaller number M , to form an $N \times M$ matrix $\hat{H}$. Without regularization, the weights estimation reduces to a Linear Least Squares Estimation problem, such that

$$w = \hat{H}^{\dagger} y, \tag{7}$$

where $\hat{H}^{\dagger}$ represents the pseudo-inverse of $\hat{H}$. The advantage again is that the weights only need to be estimated once (during calibration), assuming that the lighting in the scene remains fairly constant. The drawback though is that for the rest of the frames, the distance between each pixel and each centroid must be calculated to create a new $\hat{H}$. This process is also repeated three times, for each RGB channel.

4. IMPLEMENTATION

A customizable stereo camera, shown in Fig. 1, was created using two Point Grey Research Firefly cameras. The small size of the rig allows it to simulate a stereo camera on a handheld device. Firefly cameras allow for fast exposure control by varying the shutter speed at every frame. Wide-angle lenses are also used so that the face can be adequately imaged from an arms length.

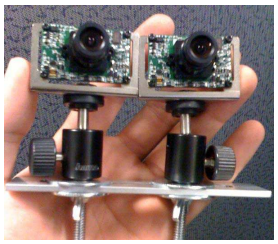


Fig. 1. Handheld Stereo Rig

4.1. Single Camera

Initially, a face aware auto-exposure algorithm can be used to find the maximum exposure at which the amount of over-exposed pixels within the face is below a low threshold, but for the purposes of this paper this step is performed manually. Then during the calibration step, a much longer exposure is taken and combined with the previous frame. After performing HDR tone mapping as described in Section 2, an image with the desired color balance on both the face and background is formed. Using Gaussian radial basis functions, we then learn the mapping between the original low exposure and this HDR image. The number of RBFs is set to 20 and their centroids are initialized randomly. For the duration of the video conference, the exposure is kept at the original low level, and all images are transformed using this mapping.

4.2. Stereo Camera

A stereo camera implementation allows different exposures to be taken at every time instant. This is useful for mobile VC, where the colors and global lighting might significantly change. After calibrating and rectifying the stereo rig [9], the left and right cameras are run at low and high exposures respectively. Using a simple normalized correlation stereo matching algorithm, dense correspondences are estimated and the right image is warped to the left image. To improve processing speed, this matching can be done on subsampled images without a significant degradation in output quality.

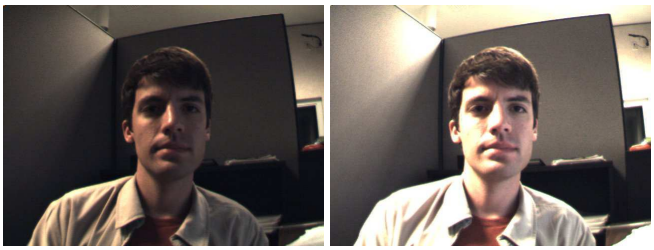
After stereo matching, the warped high exposure image is combined with the lower exposure to create the desired color balance again using the HDR tone mapping technique. However because errors in stereo matching are highly likely, the images must be masked to remove outliers from the training set of input-output mappings. In order to reduce the influence of patches with poor correspondence, only pixels with a correlation greater than .8 are used. Furthermore, we calculate the perceptually weighted color distance between corresponding pixels in the lower exposure and the tone mapped image using

$$CD = \sqrt{\left(2 + \frac{\overline{R}}{256}\right)\Delta R^2 + 4\Delta G^2 + \left(2 + \frac{(255-\overline{R})}{256}\right)\Delta B^2}, \quad (8)$$

where $\overline{R}$ is the average red value [10]. Pixels with a color distance greater than a threshold are thus discarded (in addition to those with poor stereo matching) to further reduce the effect of color aberrations.

5. RESULTS

In this section, we show some results for both the single camera and stereo camera implementations. First, in Fig. 2 we show successive frames (640x480) at low and high exposure taken with a single camera. In Fig. 2 (a) there are only a few saturated pixels in the face, however the shadows make the face much too dark. Alternatively in Fig. 2 (b), the scene is much brighter, but the tradeoff is that the left side of the face becomes saturated. This unbalanced and unnatural lighting is distracting during video conferencing and detracts from the experience.



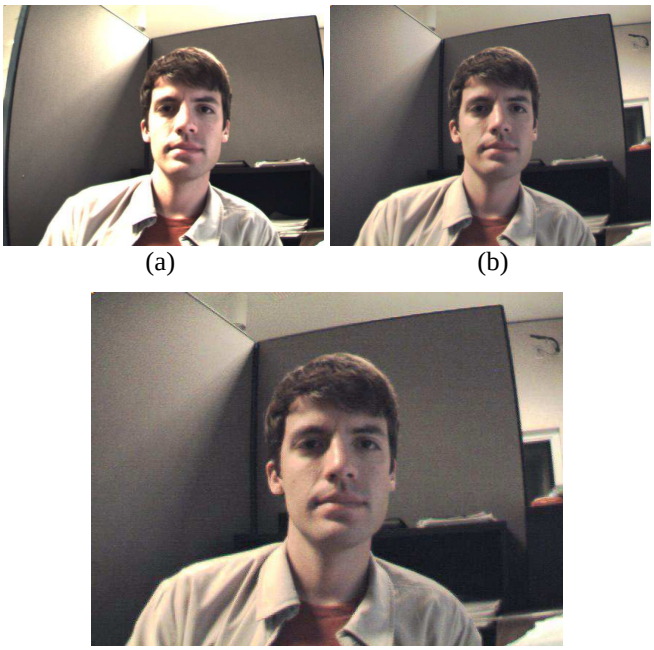
(a) (b)

Fig. 2. Single Camera Calibration Images: (a) Low exposure. (b) High exposure.

For comparison purposes only, Fig. 3 (a) shows the same time instant, but imaged with the second camera of the stereo rig running with auto-exposure turned on. Here it is clear that auto-exposure fails to correctly light the entire face, due to the unbalanced lighting. Fig. 3 (b) then shows the output of the HDR tone mapping procedure described in Section 2. This image achieves a pleasing balance between the low and high exposures in Fig. 2. The low exposure and this tone mapped image are passed into the RBF algorithm as x and y respectively. Once the mapping weights are learned, Fig. 3 (c) shows the result of applying this mapping to the low exposure image in Fig. 2 (a).

There is a slight reduction in image quality of the output compared to the desired image in Fig. 3 (b), as the colors are less vivid and noise has become more noticeable. This is due to the fact that the lower signal is being amplified along with noise. Yet overall the results are convincing, even when using a small number of RBF centroids. The entire face is now exposed at an acceptable level, and the background lighting remains consistent. Any increase in noise may not be noticeable after the images are compressed for transmission.

Similarly, Fig. 4 shows some results for the stereo implementation. After stereo matching is used to map the right image to the left, they are combined and masked to remove outliers as shown in Fig. 4 (c). It is clear that regions with



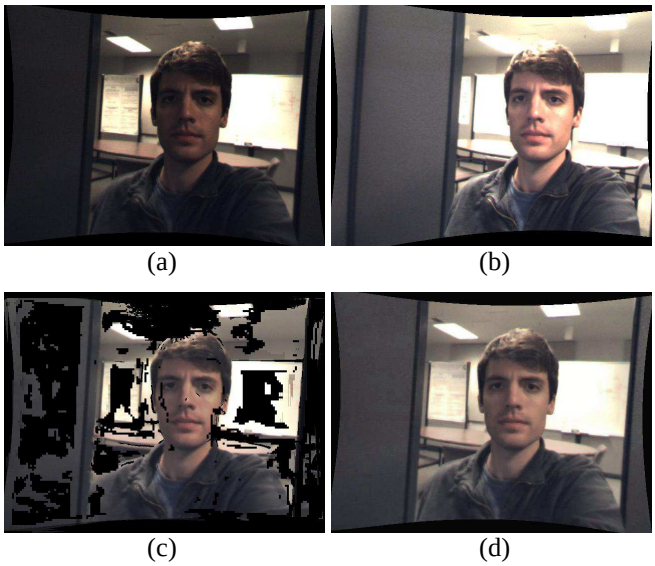
(a) (b) (c)

Fig. 3. (a) For comparison, the other camera in the stereo rig captures with auto-exposure on. (b) HDR tone mapped image created to produce "desired" color balance. (c) The enhanced output image, after mapping the low exposure to (b) using RBFs.

little texture have been removed due to poor correspondence. Finally, Fig. 4 (d) shows the output after the left image has been enhanced using the mapping learned with RBFs. The results are shown for rectified images, but the original images can be transformed with the same mapping. This shows that despite not having exact correspondence between exposures for the entire image, the method is still successful in finding an accurate mapping for all pixels. Not only is the face correctly exposed, but the burn out effect of lights on the ceiling is eliminated.

6. CONCLUSIONS

We have outlined a general purpose method to automatically increase dynamic range and enhance the lighting of video exposed with low dynamic range sensors. It is especially suitable for video conferencing because the scene color information remains fairly constant. For scenarios where this might not be the case, such as mobile VC, a stereo implementation can be used. The use of radial basis functions is key to eliminating the need for alternating exposures as well as the ghosting artifacts common in HDR video. Future work may incorporate a face aware auto-exposure algorithm and more adaptive radiance map construction and RBF parameter control. Methods may also be pursued to reduce computational overhead while maintaining output quality.



(a) (b) (c) (d)

Fig. 4. (a) Rectified Left Camera. (b) Rectified Right Camera. (c) Masked tone mapped image. (d) Enhanced Left Camera Image.

7. REFERENCES

[1] G. Messina, A. Castorina, S. Battiato, and A. Bosco, "Image quality improvement by adaptive exposure correction techniques," *ICME '03*, vol. 1, pp. I–549–52 vol.1, July 2003.

[2] Cuizhu Shi, Keman Yu, Jiang Li, and Shipeng Li, "Automatic image quality improvement for videoconferencing," *ICASSP '04*, vol. 3, pp. iii–701–4 vol.3, May 2004.

[3] Zicheng Liu, Cha Zhang, and Zhengyou Zhang, "Learning-based perceptual image quality improvement for video conferencing," *ICME '07*, pp. 1035–1038, July 2007.

[4] S.A. Bhukhanwala and T.V. Ramabadran, "Automated global enhancement of digitized photographs," *IEEE Transactions on Consumer Electronics*, vol. 40, no. 1, pp. 1–10, Feb 1994.

[5] Paul E. Debevec and Jitendra Malik, "Recovering high dynamic range radiance maps from photographs," in *SIGGRAPH '08*, New York, NY, USA, 2008, pp. 1–10, ACM.

[6] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, 2002.

[7] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski, "High dynamic range video," in *SIGGRAPH '03*, New York, NY, USA, 2003, pp. 319–325, ACM.

[8] Marco Zuliani, Luca Bertelli, and B. S. Manjunath, "An automatic method to learn and transfer the photometric appearance of partially overlapping images," *IEEE ICIP*, pp. 493–496, Oct. 2008.

[9] Zhengyou Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," *ICCV '99*, vol. 1, pp. 666–673 vol.1, 1999.

[10] T Riemersma, "Colour metric," http://www.compuphase.com/cmetric.htm, Dec. 2008.