Video Enhancement of People Wearing Polarized Glasses: Darkening Reversal and Reflection Reduction

Mao Ye
University of Kentucky
mao.ye@uky.edu

Cha Zhang
Microsoft Research
chazhang@microsoft.com

Ruigang Yang
University of Kentucky
ryang@cs.uky.edu

Abstract

With the wide-spread of consumer 3D-TV technology, stereoscopic videoconferencing systems are emerging. However, the special glasses participants wear to see 3D can create distracting images. This paper presents a computational framework to reduce undesirable artifacts in the eye regions caused by these 3D glasses. More specifically, we add polarized filters to the stereo camera so that partial images of reflection can be captured. A novel Bayesian model is then developed to describe the imaging process of the eye regions including darkening and reflection, and infer the eye regions based on Classification Expectation-Maximization (EM). The recovered eye regions under the glasses are brighter and with little reflections, leading to a more nature videoconferencing experience. Qualitative evaluations and user studies are conducted to demonstrate the substantial improvement our approach can achieve.

1. Introduction

Three-dimensional videoconferencing aims to capture, transmit and display people and their environments in 3D, thus creating an illusion that the remote participants are in the same room with you. Thanks to efforts in standardizing 3D-TV technologies, many of the technical components for capturing and transmitting 3D videos have become mature [19]. However, there is one significant difference between 3D videoconferencing and 3D-TV. In 3D-TV, the user only needs to consume the broadcasted content on a 3D display, which is often based on polarization or shutter glasses. In 3D videoconferencing, the user's 3D video will also be captured and sent to remote parties. If the user still wears polarized glasses or shutter glasses, the eye region will be too dark for the remote parties to tell his/her gaze orientations (Fig. 1(c)), which subsequently leads to poor communication efficiency [1].

One solution to the above issue is to develop autostereoscopic 3D displays, where no glasses is needed to perceive 3D. Such displays may include those based on lenticular screens [18], parallax barrier [11], projector arrays [2], rotating mirrors [13], eye tracking [27], etc. However, many of these solutions are still in the experimental stage, and are either too expensive, or too restrictive in users' head motion, or too limited to accommodate multiple viewers simultaneously. Currently, the majority of the 3D displays on the market are polarization or shutter glasses based. Motivated by the emerging trend of using polarized glasses for 3D viewing, in this paper, we study algorithms to enhance videos of people wearing polarized glasses for 3D videoconferencing. The algorithm could be extended to shutter glasses, which have very similar light transmission rate as polarized glasses when their shutters are in "transparency" state.

There are two main challenges that need to be addressed. First, when lights are shone on a pair of polarized glasses, only about 40% of the lights actually go through. This causes the eye region to be darkened, which is unpleasant in videoconferencing. Second, for almost every pair of polarized 3D glasses we could buy from the shelf, it is strongly reflective (Fig. 1(c)). Since the eye region has already been darkened, such reflection further deteriorates the video quality. Adding anti-glare coating to the glasses may partially resolve this issue, although it could easily cost much more than the glasses themselves.

In this paper, we present a solution that computationally perform darkening reversal and reflection reduction simultaneously. A stereo camera that also "wears" a pair of polarized glasses is adopted (Fig. 1(a)), which allows us to obtain partial images of the reflection. We then propose a novel Bayesian model to describe the imaging process of the eye region including darkening and reflection, and infer the eye region based on Classification Expectation-Maximization (EM) [4]. Qualitative evaluations and user studies are conducted to demonstrate the substantial improvement in image quality inside the eye region. To the best of our knowledge, our work is the first to investigate such practical issues for 3D videoconferencing applications. Besides, our approach can potentially be adapted to deal with more gen-

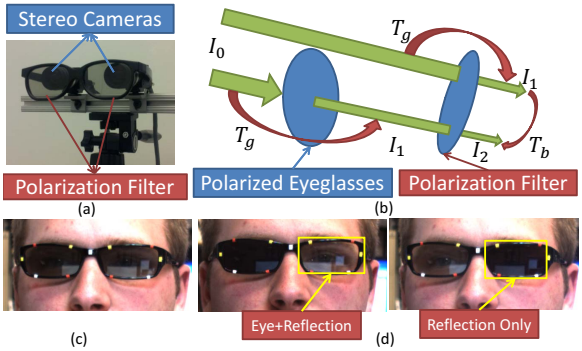


Figure 1. (a) Our setup. (b) The light loss due to polarization and our linear models (Sec. 4.2). (c) A sample image captured using regular camera without polarization filter. (d) A pair of images acquired by our stereo cameras mounted with polarization filters. Note that T_g^{-1} (Sec. 4.2) is applied to the entire images to achieve comparative global brightness as (c).

eral layer decomposition problems, where one of the layer is partially observed and noisy. One particular example is reflection layer extraction from composite images, for instance to build an environment map, where the underlying layer maybe be observed by applying a polarizer to partially filter out the reflection.

The rest of the paper is organized as follows. Related work is discussed in Sec. 2. The hardware setup is presented in Sec 3. The Bayesian model for the imaging process of the eye region is described in Sec. 4, and its inference in Sec. 5. Experimental results and conclusions are given in Sec. 6 and 7, respectively.

2. Related Work

There has been very limited amount of work that handles eyeglasses in images/videos. One interesting study is by Wu *et al*. [28], which aimed to automatically remove eyeglasses in face images. It detects eyeglasses using a boosting based detector, and then adopts a statistical analysis and synthesis approach to remove eyeglasses based on training data. Since reflection is difficult to model with a set of training examples, the algorithm does not perform well under strong reflections.

Reflection layer separation has in fact attracted a lot of research recently. However, most of the approaches focus on mirror-like wall-size objects, and are not directly applicable to our problem. We briefly review them here to provide some background and to stress the new challenges we face in our application.

The existing approaches can be grouped into three categories: physics-based, motion-based, and prior-based.

Physics-based approaches usually rely on linear polarization of the reflected images. Kong *et al*. [14] captured multiple polarized images from a single view point with different polarization angles. Under the assumption that variance of gradients of the input images is proportional to the magnitude of gradient of the reflection layer, a constrained minimization problem was formulated and solved to separate the reflection layer. Their more recent work [15] captured only three polarized images with polarizer angles separated by 45 degrees. The special relationship between images can then be utilized to extract the reflection layer. Since most 3D glasses are circularly polarized, these existing methods cannot be applied to our scenario.

Motion cues have also been used for layer extractions, some in the intensity domain [25, 24, 12, 3], while others in the gradient domain [9, 10]. Notably, the recent work by Sudipta *et al*. [24] produced visually appealing results for image-based rendering. Temporal information is crucial for motion-based approaches. However, for polarized 3D glasses, small head movement can cause large changes in reflection, making temporal information difficult to extract.

The last group of approaches incorporates various kinds of prior information about the scene in order to extract the reflection layer. For instance, Sarel and Irani [21] handled scenes with repetitive dynamic behaviors. Levin *et al*. [17] minimized the total number of edges and corners when extracting two layers. While it requires only one image, it could fail for some images. In [6], edges in the composite image were assumed to be from either layer. In some works, user assists were required for good results [16]. The statistics between the background and the reflection layers can also be assumed to be independent. Independent Component Analysis (ICA) [7] has been adopted to extract the two layers with maximal mutual information, such as [8, 5, 23]. Sarel and Irani also presented solutions for a slightly weaker assumption that the two layers are uncorrelated [20]. These methods typically still require multiple images from the same viewpoint, or object motion tracking, which will be difficult to meet in our application.

3. System Setup

We adopt a stereo camera pair to capture the user for 3D videoconferencing, as shown in Fig. 1(a). Since temporal information is unreliable in our scenario, our goal is to reverse the darkening effect and reduce the reflection solely from the spatial redundancy. To this end, we mount a pair of polarized 3D glasses on the cameras, which allows us to capture a *partial* reflection image directly due to polarization. As shown in Fig. 1(d), consider the camera with left-handed circular polarization filter on it. It will not see the user's eye with opposite polarization. Only the reflection may be seen. In contrast, the camera with right-handed circular polarization will see both the eye region and the reflection for the same eye.

In the following sections, we build a Bayesian model for the imaging process of the above setup, and present an optimization framework to enhance the eye region for our application.

4. Bayesian Imaging Model

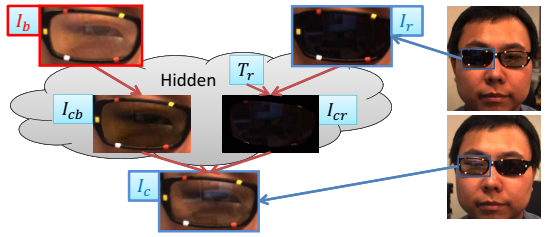


Figure 2. The Bayesian imaging model.

Let us consider the user's right eye region in Fig. 2, where the two images from the stereo cameras are shown. The top image is from one of the stereo cameras, and it contains mostly reflection, which we denote I_r. The bottom image is from another camera, and it contains both the background eye region and the reflection, which we denote I_c. Our goal is to recover the right eye region without polarized glasses for the same camera view of I_c, which we denote as I_b. In general, independence can be assumed between I_r and I_b. From a probabilistic point of view, we would like to find the Maximum-a-Posteriori probability of I_b, that is:

$$\begin{aligned} I_b &= \arg\max_{I_b} P(I_b|I_c, I_r) \\ &= \arg\max_{I_b} P(I_c|I_b, I_r)P(I_b) \end{aligned} \quad (1)$$

By examining the imaging process of the setup, we further build a graphical relationship between I_b, I_r and I_c, as shown in Fig. 2. A few hidden variables are introduced as follows:

- I_{cb}: the transmission layer of the composite image I_c;
- I_{cr}: the reflection layer of the composite image I_c;
- T_r: the mapping between the reflection layer I_{cr} and the observed reflection in alternative view I_r.

With the set of hidden variables $\mathcal{H} = \{I_{cb}, I_{cr}, T_r\}$, Eq. (1) can be re-written as:

$$\begin{aligned} I_b &= \arg\max_{I_b} \int_{\mathcal{H}} P(I_c|I_{cb}, I_{cr}, T_r, I_b, I_r) \\ &\quad \times P(I_{cb}, I_{cr}, T_r|I_b, I_r)P(I_b) \\ &= \arg\max_{I_b} \int_{\mathcal{H}} P(I_c|I_{cb}, I_{cr})P(I_{cb}|I_b)P(I_b) \\ &\quad \times P(I_{cr}|T_r, I_r)P(T_r) \end{aligned} \quad (2)$$

where the last step arises from the conditional independence assumed in Fig. 2. We next describe each component involved in the probabilistic model. Inferring I_b from the model will be explained in Sec. 5.

4.1. Image Composition

The first term on the right side of Eq. (2), $P(I_c|I_{cb}, I_{cr})$, involves the composition of two layers : the transmission layer (background layer) I_{cb} and the reflection layer I_{cr}. We apply a simple additive model for the process:

$$I_c = I_{cb} + I_{cr} + \eta_c \quad (3)$$

where $\eta_c \sim \mathcal{N}(0, \Sigma_c)$ is modeled as Gaussian noise.

4.2. Imaging Through Polarized Glasses

The second term, $P(I_{cb}|I_b)$, corresponds to the upper-left part of Fig. 2, and encodes the light loss due to polarization. For a single polarization glass, a linear color transform can model the process with acceptable performance. As shown in Fig. 1(b), we assume $I_1 = T_g \cdot I_0$, where T_g is a 3×3 matrix. However, due to the polarization filters on our stereo cameras, the transformation between I_{cb} and I_b, named T_b, is slightly different from T_g, as illustrated in Fig. 1(b). With our setup, the rest of the scene is recorded through a single polarization filter, while the eye regions behind the polarized eye-glasses are recorded through two filters. For the eye regions that can be seen, thanks to the same polarization direction of the glasses, the relative light loss with respect to the rest of the scene is smaller compared with T_g. Nevertheless, a linear model is also used:

$$I_{cb} = T_b \cdot I_b + \eta_{cb}, \quad (4)$$

where $\eta_{cb} \sim \mathcal{N}(0, \Sigma_{cb})$ is a Gaussian noise that we use to model the imperfection of the linear transform assumption. Note for a particular pair of glasses, the linear transforms T_g and T_b can be obtained through pre-calibration with a color calibration card.

4.3. Skin Color Prior

The probability $P(I_b)$ encodes prior information regarding the region to be recovered. It is possible to build a prior model for each user using many examples. In this paper, we resort to a simple skin color distribution to represent the prior probability, which can be computed from the face region in the images. For simplicity, a Gaussian distribution is assumed, i.e. $I_b \sim \mathcal{N}(\mu_b, \Sigma_b)$.

4.4. Reflection Model

The last two terms describe the upper-right subgraph of Fig. 2, which associates the observed reflection I_r in the alternative view with the reflection layer I_{cr} of the composite image I_c. Due to view differences, spatially there is a non-linear warping between these two reflection images. Moreover, since the glasses area is relatively small, part of the reflection in one view is not observable from the other, as shown in Fig. 3. Therefore, we partition the reflection layer I_{cr} into two parts: I^o_{cr} represents the part of reflection observed from the alternative view, which may be estimated from the observed reflection I_r through spatial warping. I^N_{cr} represents the part of reflection that is not observed (e.g., the region inside the yellow circle in Fig. 3(a)).

We assume that given I^o_{cr}, I^N_{cr} is independent of I_r and T_r. That is:

$$\begin{aligned} P(I_{cr}|T_r, I_r)P(T_r) &= P(I^o_{cr}, I^N_{cr}|T_r, I_r)P(T_r) \\ &= P(I^N_{cr}|I^o_{cr})P(I^o_{cr}|T_r, I_r)P(T_r) \end{aligned} \quad (5)$$

The mapping between I_r and I^o_{cr} is modeled as follows. Spatially, a warping can be applied to I_r in order to match

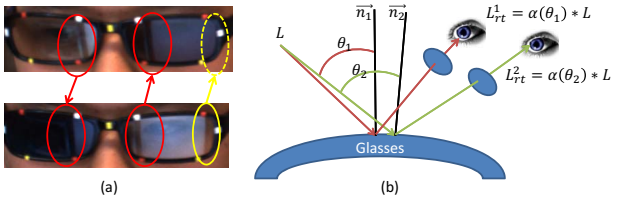


Figure 3. Reflection Model. (a) The spatial mapping between the stereo image pair. Red and yellow circles mark examples of observable and non-observable parts of the reflection layer. (b) The intensity scaling between the reflections across views.

against I^o_{cr}. Denote the warping result as I^ω_r. We have:

$$I^\omega_r(p) = I_r(x + d(p), y) \Rightarrow I^\omega_r = \omega(I_r) \tag{6}$$

where $d(p)$ is the disparity of pixel $p \triangleq (x, y)$ in the reflection image I_r; and ω abstracts the non-linear warping.

Furthermore, we introduce a per-pixel, per-channel scale factor $s = \{s^l\}_{l\in\{r,g,b\}}$ ($l \in \{r, g, b\}$ denotes image channels here), such that

$$I^o_{cr} = S \cdot \omega(I_r) + \eta_{cr}, \tag{7}$$

$$\text{where } S = diag(s^r, s^g, s^b). \tag{8}$$

Note $T_r \triangleq \{s, \omega\}$ with s and ω assumed independent to each other, and $\eta_{cr} \sim \mathcal{N}(0, \Sigma_{cr})$ is a Gaussian noise term. The notation s and its diagonal matrix form S will be used interchangeably in the remainder of the paper wherever convenient. The scale factor s is introduced for two reasons. First, the two stereo cameras may behave slightly differently, resulting in different color outputs. Second, as pointed out in [15], the light reflected off glass surfaces is partially polarized, which depends on the angle between incident light and surface normal, as illustrated in Fig 3(b). When either the scene point or the glasses moves, the scaling factor s for a single point on the glasses also changes (due to changes in $\frac{\alpha(\theta_1)}{\alpha(\theta_2)}$ in Fig. 3(b)).

Regarding the terms $P(I^N_{cr}|I^o_{cr})$ and $P(T_r)$, we will use them as regularization terms during the inference. More discussions about their probabilistic models will be given in Section 5.3 and 5.2, respectively.

5. Inference

The objective function in Eq. (2) involves a set of continuous hidden variables and a nonlinear mapping, which is difficult to solve. We adopt a variant of the Expectation-Maximization (EM) approach, namely Classification EM [4], to find an approximate solution. In Classification EM, the summation in the E-step is replaced with estimation of the modes of the hidden variables. The target variables are then estimated with the modes in the M-step. Such a process is iterated until convergence, as sketched in Table 1 for our application.

Since the reflection model involves non-linear spatial mapping ω, it is error-prone to estimate the modes of all the hidden variables simultaneously. Instead, we optimize them in turn. The spatial warping ω is the only variable that cannot be appropriately initialized; and is therefore estimated first. The other variables are initialized as follows:

$$\hat{I}_b = 0; \quad \hat{s} = \{1, 1, 1\}; \quad \hat{I}_{cb} = T_b \cdot \hat{I}_b = 0;$$
$$\hat{I}^o_{cr} = \hat{I}_c - \hat{I}_{cb} = I_c; \quad \hat{I}^N_{cr} = 0; \tag{11}$$

The overall flow of the inference process is given in Fig. 4.

Initialization: $\hat{I}_b$

Iterate until convergence:

E-step:

$$\{\hat{I}_{cb}, \hat{I}_{cr}, \hat{T}_r\} = \arg\max_{\mathcal{H}} \log(P(I_{cb}, I_{cr}, T_r|I_c, I_r, \hat{I}_b))$$
$$= \arg\max_{\mathcal{H}} \log\Big(P(I_c|I_{cb}, I_{cr})P(I_{cb}|\hat{I}_b)$$
$$\times P(I^N_{cr}|I^o_{cr})P(I^o_{cr}|I_r, s, \omega)P(s)P(\omega)\Big) \tag{9}$$

M-step:

$$\hat{I}_b = \arg\max_{I_b} \log\big(P(I_b|I_c, I_r, \hat{I}_{cb}, \hat{I}_{cr}, \hat{T}_r)\big)$$
$$= \arg\max_{I_b} \log\big(P(\hat{I}_{cb}|I_b)P(I_b)\big) \tag{10}$$

Table 1. The Classification EM framework.

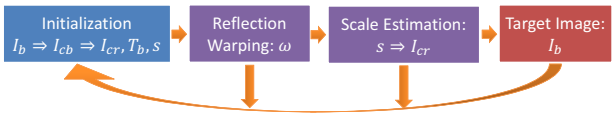


Figure 4. The flow of our inference. The three block colors indicate the three EM components: initialization, E-step and M-step.

5.1. Reflection Image Warping

By dropping the terms in Eq. (9) that are unrelated to ω, we obtain:

$$\hat{\omega} = \arg\max_{\omega} \big(\log P(\hat{I}^o_{cr}|I_r, \hat{s}, \omega)\big) + \log\big(P(\omega)\big)$$
$$\equiv \arg\min_{\omega} \big(E^\omega_D(d) + E^\omega_R(d)\big), \tag{12}$$

where the relationship between ω and d has been defined in Eq. (6). The problem above is equivalent to the traditional disparity estimation problem with stereo images [22]. The first term denotes the data cost, and the second term represents a smoothness constraint. We use graph cut as a global optimization scheme to solve the warping. Considering the specific characteristics of the data in our application, we propose to use the following data energy function based on image gradient (denoted as ∇, we use gradient of gray scale images in this work):

$$E^\omega_D(d) = \sum_{(x,y)} \Big(u(x, y)\big[\nabla I_r(x, y) - \nabla I_{cr}(x + d, y)\big]^2\Big)$$
$$\text{where} \quad u(x, y) = \log\Big[\big(\nabla I_r(x, y)\big)^2 + 1\Big] \tag{13}$$

The per pixel weight $u(x, y)$ represents relative importance of the pixels. The reasons for choosing this particular cost function and weighting scheme are:

(1) Human beings sense reflection mainly due to high frequency region of the reflection layer, e.g. edges. Therefore the pixels of the reflection layer with higher gradients are assigned with higher weight as encoded in $u(x, y)$. The logarithm is empirically designed to avoid bias arising from unexpected edges (e.g. eyeglasses frames).

(2) The initialization of $I_b = 0$ makes pixel intensities less reliable than gradients for distance measure.

In terms of smoothness constraints $E_R^\omega(d)$, the contrast-sensitive Potts model is adopted [24].

With the estimated disparities, the reflection image I_r can be spatially warped to align with the reflection layer I_{cr}, denoted as I_r^ω. In order to further improve the matching for sub-pixel alignment, we apply a filtering process as follows:

$$I_r^{\omega\prime}(p) = \sum_{q\in\mathcal{L}(p)} \mathcal{C}(I_r^\omega(q), I_{cr}(p)) \cdot I_r^\omega(q) \qquad (14)$$

where $\mathcal{L}(\cdot)$ denotes a local neighborhood (3×3 in this work); and $\mathcal{C}(I_r^\omega(q), I_{cr}(p))$ is the normalized cross correlation (NCC) between the local patch of I_r^ω at pixel q and that of I_{cr} at pixel p. 5×5 patches are used in our approach. Fig. 5 shows an example of the I_r^ω and $I_r^{\omega\prime}$ ((c) and (d) respectively). In the rest of the paper, we denote the filtered reflection as $\hat{I}_r^\omega$ for simplicity.

5.2. Scale Factors Estimation

The scale factors $s \triangleq \{s^l\}$ (l denotes image color channel) can be estimated by combining Eq. (3) and (7) with the first and fifth terms in Eq. (9):

$$\begin{aligned}\hat{s} &= \arg\max_s \log\big(P(I_c|I_{cb}, I_r, s, \hat{\omega}) + \log P(s)\big)\\ &\equiv \arg\min_s \big(E_D^s(s) + E_R^s(s)\big) \qquad (15)\end{aligned}$$

Due to the same considerations in Sec. 5.1, image gradients instead of intensities are used to define the energy terms. We focus on the gradient difference between the composite image I_c and the estimated reflection layer $\hat{I}_{cr}$, since the estimated transmission layer $\hat{I}_{cb}$ may contain reflection residue. More specifically, we define:

$$E_D^s(s) = \sum_{l\in\{r,g,b\}}\sum_p \Big(v(p)\Big[\nabla\big(I_c^l(p) - s^l(p)\hat{I}_r^{\omega,l}(p)\big)\Big]^2\Big),$$

$$\text{and } v(p) = \frac{u(p)\Big(1 + \mathcal{C}\big(\hat{I}_r^\omega(p), \hat{I}_{cb}(p)\big)\Big)}{1 + \max_p\{\mathcal{C}\big(\hat{I}_r^\omega(p), \hat{I}_{cb}(p)\big)\}},$$

$$\text{where } u(p) = \log\Big[\big(\nabla\hat{I}_r^\omega(x, y)\big)^2 + 1\Big]. \qquad (16)$$

The new weight $v(p)$ also measures the similarity between the transmission layer $\hat{I}_{cb}$ and the warped reflection image $\hat{I}_r^\omega$. If the similarity is high, there could be reflection residues in $\hat{I}_{cb}$, thus the gradient difference between the composite image and the estimated reflection shall be further reduced.

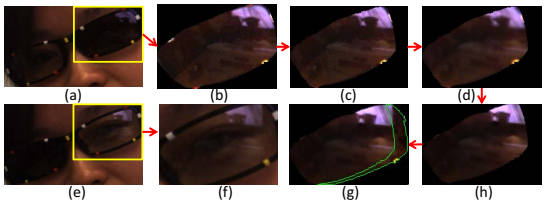


Figure 5. Recovery of the reflection layer ((b),(c),(d),(g),(h) are scaled up 3 times from raw data for visualization purpose). (a) and (e) are input image pair. (b) and (f) are the reflection image I_r and composite image I_c. (b)→(c): spatial warping (Sec. 5.1). (c)→(d): filtering (Eq. 14). (d)→(h): scale factor (Sec. 5.2). (h)→(g): hallucination of non-observable part encompassed by the green curve (Sec. 5.3).

The regularization term is formulated as follows to prevent over-fitting:

$$\begin{aligned}E_R^s(s) &= \lambda_R^s\Big(\sum_l\sum_p\sum_{q\in\mathcal{L}(p)}\big(\mathcal{F}(p, q)|s^l(p) - s^l(q)|^2\big)\\ &+ \lambda_l\sum_p\sum_{l_1,l_2\in\{r,g,b\}}|s^{l_1} - s^{l_2}|^2\Big), \qquad (17)\end{aligned}$$

$$\text{where } \mathcal{F}(p, q) = \exp\Big(\frac{\|\hat{I}_r(p) - \hat{I}_r(q)\|^2}{2\sigma^2}\Big). \qquad (18)$$

where λ_R^s and λ_l balance the relative importance of the terms. Empirically we find $\lambda_R^s = 10$ and $\lambda_l = 0.5$ are good choices.The first term forces spatial smoothness, with the function $\mathcal{F}$ designed to be the adaptive weighting kernel to partially handle the camera difference mentioned in Sec. 4.4. The second term in Eq. (17) favors consistency of the scales across channels.

5.3. Reflection Layer

The reflection layer, including both the observable and non-observable part, is updated as follows:

$$\begin{aligned}\hat{I}_{cr} &= \arg\max_{I_{cr}=\{I_{cr}^N, I_{cr}^o\}}\big(\log P(I_c|\hat{I}_{cb}, I_{cr})\\ &+ \log P(I_{cr}^N|I_{cr}^o) + \log P(I_{cr}^o|I_r, \hat{s}, \hat{\omega})\big)\\ &\equiv \arg\min_{I_{cr}}(E_{D_c}^{cr}(I_{cr}) + E_S^{cr}(I_{cr}) + E_{D_r}^{cr}(I_{cr})) \qquad (19)\end{aligned}$$

The second term assumes dependence of the non-observable part on the observable part as discussed in Sec. 4.4. In practice, we can only enforce constraints along the boundary across these two regions, i.e.:

$$E_S^{cr}(I_{cr}) = \lambda_S^{cr}\sum_{p\in\Omega(I_{cr}^o)}\sum_{q\in\mathcal{L}(p)\bigcap\Omega(I_{cr}^N)}\|\hat{I}_{cr}^o(p) - I_{cr}^N(q)\|^2 \qquad (20)$$

where $\Omega(\cdot)$ denotes the boundary region and the weight λ_S^{cr} is set to 10 in this work.

One consequence that follows is that only the first term in Eq. 19 takes effect for the inner region of the non-observable part. The fact that I_{cb}^N is initialized as zero and

updated based on the hallucinated I_{cr}^N (Eq. 23) results in non-informative and unstable update via first term without additional information. We thereby assume that this *small* region contains no critical skin details, i.e. $\nabla I_{cb}^N \approx 0$, and use penalty function in gradient domain for I_{cr}^N, while in intensity domain for I_{cr}^o:

$$E_{D_c}^{cr}(I_{cr}) = \sum_{p \in I_{cr}^o} \left\| I_c(p) - \hat{I}_{cb}(p) - I_{cr}(p) \right\|_{\Sigma_c}^2 + \sum_{p \in I_{cr}^N} \left| \nabla (I_c(p) - I_{cr}(p)) \right|^2 \quad (21)$$

where the notation $\|v\|_{\Sigma_v}$ means $v^T \Sigma_v^{-1} v$ in this paper.

Plugging in Eq. (7), the third term in Eq. (19) can be written as:

$$E_{D_r}^{cr}(I_{cr}) = \sum_{p \in I_{cr}^o} \left\| I_{cr}(p) - \hat{S}(p) \cdot \hat{I}_r^{\omega}(p) \right\|_{\Sigma_{cr}}^2 \quad (22)$$

Fig. 5 shows the procedure of reflection layer recovery, in which (g) shows the hallucinated non-observable part.

5.4. Transmission Layer

Combining Eq. (4) with the first two terms in Eq. (9), the transmission layer can be obtained as:

$$\hat{I}_{cb} = \arg\max_{I_{cb}} \left(\log P(I_c | I_{cb}, \hat{I}_{cr}) + \log P(I_{cb} | \hat{I}_b) \right)$$
$$= \arg\min_{I_{cb}} \left(\left\| I_c - \hat{I}_{cr} - I_{cb} \right\|_{\Sigma_c}^2 + \left\| T_b \cdot I_b - I_{cb} \right\|_{\Sigma_{cb}}^2 \right) \quad (23)$$

which is effectively an interpolation between $I_c - \hat{I}_{cr}$ and $T_b \cdot \hat{I}_b$. Minimizing the above objective function is equivalent to solving a linear equation. Note that during the first iteration, the second term of Eq. (23) is ignored since I_b is initialized to zero.

5.5. M-Step

With the modes of hidden variables estimated, we can proceed to recover our ultimate goal: the background image behind the glasses, by performing the M-step in Table 1:

$$\hat{I}_b = \arg\min_{I_b} \left(\left\| T_b \cdot I_b - \hat{I}_{cb} \right\|_{\Sigma_{cb}}^2 + \left\| I_b - \mu_b \right\|_{\Sigma_b}^2 \right) \quad (24)$$

This optimization is similar to that of Eq. (23). The EM process is iterated until the change of $\hat{I}_b$ is small. Normally three iterations are sufficient in our experiments.

6. Experiments

In order to demonstrate the effectiveness of our approach, we experiment on a collection of image sequences captured using the setup shown in Fig. 1(a) for qualitative evaluation. However, it is impractical to perform pixel-wise quantitative evaluation with images of subject with and without the polarized eyeglasses. Therefore, we conduct user studies instead as described in Sec. 6.3. Before describing the actual evaluation, we first present the pre- and post-processing step as follows.

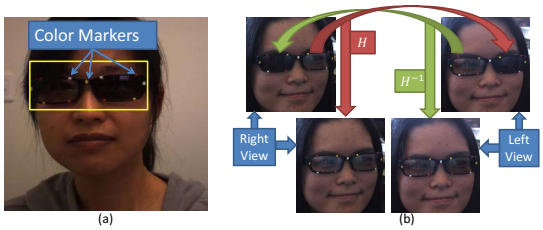


Figure 6. Pre-processing (a) and Post-processing (b). See Sec. 6.1 for details.

6.1. Pre- and Post-Processing

Eyeglasses tracking is required in order to process video data. However, tracking is not the focus of our work. We attach physical color markers on the glass frames for tracking, as shown in Fig. 6(a). The eye regions are manually defined for the first frame, and are then warped for all the rest images according to a 2D Homography calculated from the tracked markers (Fig. 6(b)). A neighborhood region around eyeglasses frame, as shown by the rectangle in Fig. 6(a), is then defined for skin detection with the eye regions excluded. Skin color pixels are identified in the HSV color space [26]; and the skin color prior (Sec. 4.3) is built upon those pixels by fitting a 3D Gaussian distribution.

Note that in the images captured with our camera setup, one of the eye regions for each view has only reflection. They are therefore not suitable for stereoscopic videoconferencing directly. As a post-processing step, we perform cross-view warping between each pair of images using 2D Homography, as illustrated in Fig. 6(b).

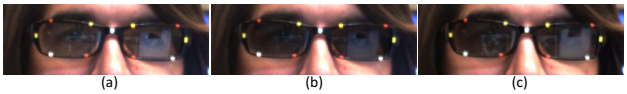
6.2. Real Scene Data Evaluation

Our approach is applied to a collection of real world image sequences of human subjects, with diversity in the following aspects:

- Subject gender: male and female;
- Subject skin color: yellow, white and black;
- Reflection: weak, moderate and strong, for example: table, monitor and sky, respectively;
- Global illumination: from relatively dark to bright.

The test data set consists of both static and dynamic scenes as detailed below.

Static Scenes: Two groups of images are acquired with regular stereo cameras (without polarization filters) and with our setup respectively. Fig. 7(a) shows the first group, in which the dark eye regions as well as the reflection substantially compromise the visual quality. Fig. 7(c) are images captured with our setup and transformed with T_g^{-1} (Sec. 4.2) in order to achieve similar overall brightness as (a) for fair comparison. Together with the noisy reflection layers, they are used as input to our algorithm. When capturing (a) and (c) with the polarization filters on and off the cameras respectively, the subjects are asked to remain as still as possible. Small movement is acceptable



(a) (b) (c)

Figure 8. Image simulation for user studies: (a) captured with our setup; (b) simulated dark eye regions; (c) captured by a camera.

for visual comparison. In Fig. 7(b), the eye regions are enhanced through T_g^{-1} from (a) to achieve darkness reversal. Nonetheless, the reflection, especially the relative strong monitor reflection, still largely deteriorates the image quality. By contrast, our approach performs both reflection reduction and darkness reversal, and achieves significant improvement as demonstrated in Fig. 7(d). Note the large differences of global illumination as well as the range of reflection strengths across images.

Dynamic Scenes: In dynamic data set, the subjects generally move freely with various head motion and eye blinking. Our results are presented in the supplemental materials. Again the substantial improvement of visual quality demonstrates the effectiveness of our approach. And the diversity in the data set validates its robustness.

Our system is currently implemented in Matlab; and the processing time for each frame is around 4s on average. Note that the computation in our approach mainly involves pixel-wise operations and linear equation solving. Therefore, GPU acceleration can be naturally implemented for real time applications.

6.3. User Studies

The dynamic data set is also used for user studies in order to evaluate the applicability of our approach to 3D videoconferencing. Ideally one would conduct the same visual comparison as in Fig. 7. However, in dynamic scenes one cannot capture two sequences of images with the subject performing exactly the same motion. Therefore, we simulate the traditional setup through the transformation $T_g \cdot T_b^{-1}$ from data captured with our setup (see Sec. 4.2). As demonstrated in Fig. 8, the simulated image (b) achieves very similar brightness in eye regions as that captured by a regular stereo camera pair (image (c)). However, as a side effect, the reflection is also weakened, which makes the user study more favorable to the traditional setup.

There are 20 college students in their 20's participated in our study. Six of the subjects are female. Each subject is presented with 9 videos. Each video consists of three sub-videos: the simulated "dark-eye" video, the one enhanced with T_b^{-1} from data captured with our setup, and the enhanced video from our method. The users are asked to rank the three sub-videos according to their preference. They are instructed to imagine that they are videoconferencing with the subject in the video. Two groups of studies are conducted: 2D videos (from one of the stereo images) and 3D anaglyph videos. The results are summarized in Table. 2.

Overall the results are very positive. Over 70% of the subjects chose results from the proposed method as the "best", for both 2D and 3D viewing. The results do show, however, in some cases the participants do prefer the darkness-reversed-only image, or even the original image. The main reason is the small amount of flickering in our results which we will discuss next.

	2D Videos			3D Videos		
	Simulated Dark Images	Darkness Reversed Only	Our Results	Simulated Dark Images	Darkness Reversed Only	Our Results
Worst	101 (56.1%)	60 (33.3%)	19 (10.6%)	107 (59.5%)	51 (28.3%)	22 (12.2%)
Medium	62 (34.5%)	85 (47.2%)	33 (18.3%)	54 (30.0%)	99 (55.0%)	27 (15.0%)
Best	17 (9.4%)	35 (19.5%)	128 (71.1%)	19 (10.5%)	30 (16.7%)	131 (72.8%)

Table 2. Summary of our user studies.

6.4. Limitations and Future Work

There are several limitations in our current approach. It can not be directly applied to polarized eyeglasses made with cheap plastic filters, because the deformation in the plastic film causes large difference in the reflection across views (see supplemental materials). Consequently, their spatial relationship can not be correctly estimated. A precise 3D modeling of the reflection surface may help here. Our current implementation estimates the reflection parameters frame by frame. The fluctuation in the estimated parameters sometime leads to flickering in the final video sequence, which causes discomfort for a small number of user study participants. In future work, we plan to model the temporal relationship of the transmission layers I_{cb} or the eye region image I_b explicitly to resolve this issue. We are also interested in applying our approach to shutter glasses, however the engineering hurdle to sync cameras with shutter glasses must be overcomed first.

7. Conclusion

We proposed a probabilistic approach for reflection reduction of polarized eyeglasses for the purpose of 3D videoconferencing, with our adapted hardware design. Our algorithm performs darkness reversal and reflection reduction effectively as demonstrated by the experiments, and substantially improves the visual quality of the images that are then used for 3D videoconferencing.

Acknowledgements We would first like to thank the anonymous reviewers for their valuable feedback. Majority of the work was conducted during the first author's internship at Microsoft Research. This work is supported in part by US NSF grant IIS-0448185, CCF-0811647, and CNS-0923131.

References

[1] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

[2] T. Balogh. Method and apparatus for producing 3-D picture. *U.S. Patent 5 801 761*, 1998.

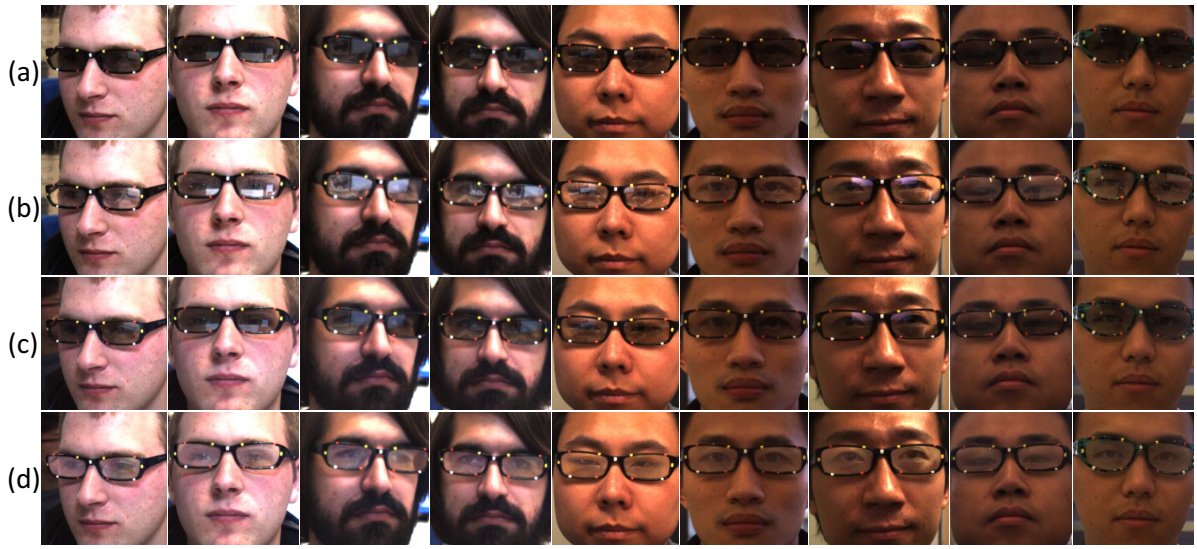


Figure 7. Images of various real-life scenes. See Sec. 6.2 for detail explanation. Note that cross-view warping is applied to the images in (c) and (d) only.

[3] J. Bergen, P. Burt, R. Hingorani, and S. Peleg. A three-frame algorithm for estimating two-component image motion. *PAMI*, 14(9):886 –896, sep 1992.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[5] A. M. Bronstein, M. M. Bronstein, M. Zibulevsky, and Y. Y. Zeevi. Sparse ICA for blind separation of transmitted and reflected images. *INTL. J. IMAGING SCIENCE AND TECHNOLOGY*, 15:84–91, 2005.

[6] Y.-C. Chung, S.-L. Chang, J.-M. Wang, and S.-W. Chen. Interference reflection separation from a single image. In *WACV*, pages 1 –6, dec. 2009.

[7] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287 – 314, 1994.

[8] H. Farid and E. H. Adelson. Separating reflections from images using independent components analysis. *Journal of the Optical Society of America*, 16:2136–2145, 1998.

[9] K. Gai, Z. Shi, and C. Zhang. Blindly separating mixtures of multiple layers with spatial shifts. In *CVPR*, june 2008.

[10] K. Gai, Z. Shi, and C. Zhang. Blind separation of superimposed images with unknown motions. In *CVPR*, 2009.

[11] J. Hamasaki. Aberration theories of lenticular and related screens. In *Proc. Int. Workshop Stereoscop. 3-D Imaging*, 1995.

[12] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *IJCV*, 12(1):5–16, Feb. 1994.

[13] A. Jones, M. Lang, G. Fyffe, X. Yu, J. Busch, I. McDowall, M. Bolas, and P. Debevec. Achieving eye contact in a one-to-many 3D video teleconferencing system. In *SIGGRAPH Emerging Technologies*, 2009.

[14] N. Kong, Y.-W. Tai, and S. Y. Shin. High-quality reflection separation using polarized images. *TIP*, dec. 2011.

[15] N. Kong, Y.-W. Tai, and S. Y. Shin. A physically-based approach to reflection separation. In *CVPR*, june 2012.

[16] A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. *PAMI*, 2007.

[17] A. Levin, A. Zomet, and Y. Weiss. Separating reflections from a single image using local features. In *CVPR*, 2004.

[18] W. Matusik and H. Pfister. 3D TV: A scalable system for real-time acquistion, transmission and autostereoscopic display of dynamic scenes. *TOG*, Aug. 2004.

[19] H. Ozaktas and L. Onural. *Three-Dimensional Television: Capture, Transmission, Display*. Springer, 2007.

[20] B. Sarel and M. Irani. Separating transparent layers through layer information exchange. In T. Pajdla and J. Matas, editors, *ECCV*. Springer Berlin Heidelberg, 2004.

[21] B. Sarel and M. Irani. Separating transparent layers of repetitive dynamic behaviors. In *ICCV*, oct. 2005.

[22] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, Apr. 2002.

[23] Y. Y. Schechner, J. Shamir, and N. Kiryati. Polarization-based decorrelation of transparent layers: The inclination angle of an invisible surface. *ICCV*, 2:814, 1999.

[24] S. N. Sinha, J. Kopf, M. Goesele, D. Scharstein, and R. Szeliski. Image-based rendering for scenes with reflections. *Proc. of SIGGRAPH*, 31(4), July 2012.

[25] R. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *CVPR*, pages 1246–, 2000.

[26] V. Vezhnevets, V. Sazonov, and A. Andreeva. A survey on pixel-based skin color detection techniques. In *IN PROC. GRAPHICON-2003*, pages 85–92, 2003.

[27] G. J. Woodgate, D. Ezra, J. Harrold, N. S. Holliman, G. R. Jones, and R. R. Moseley. Observer tracking autostereoscopic 3D display systems. In *Proc. SPIE*, 1997.

[28] C. Wu, C. Liu, H.-Y. Shum, Y.-Q. Xu, and Z. Zhang. Automatic eyeglasses removal from face images. *PAMI*, 26(3):322–336, Mar. 2004.