Mobile Computational Photography 2013: Introduction to the JEI Focal Track Presentations

Todor Georgiev
Qualcomm Inc.
355 East Trimble Road
San Jose, California 95131
E-mail: todorg@qualcomm.com

Andrew Lumsdaine
Indiana University
215 Lindley Hall
150 South Woodlawn
Bloomington, Indiana 47401
E-mail: lums@cs.indiana.edu

In addition to the usual conference presentations, the 2013 Mobile Computational Photography conference includes a "focal track" of peer-reviewed papers that appear in a special section of the *Journal of Electronic Imaging*. Here, we introduce these papers, using an extract from the Editorial[5] accompanying the JEI issue.

Many of the capabilities of mobile computational photography will likely leverage plenoptic (a.k.a. lightfield) camera capabilities. In the mobile setting, these will need to be built using micro-optic techniques, either arrays of miniaturized cameras or with arrays of microlenses. Wafer-level cameras, built using semiconductor processes, will become a key sensor technology. In their paper "Resolution and sensitivity of wafer-level multi-aperture cameras," Oberdorster and Lensch[1] present an analysis of some of the ensemble optical properties of wafer-level cameras, with particular attention to controlling aberrations.

Algorithmically, obtaining large-camera capabilities out of mobile computational platforms (particularly those based on plenoptic camera ideas) will require new processing approaches and algorithms. As advances in plenoptic rendering continue to be made, being able to effectively estimate depth (disparity) in a scene is emerging as a critical need. Krishnamurthy and Rastogi[2] develop an approach to depth estimation that is particularly well-suited to plenoptic imagery in their paper "Refinement of depth maps by fusion of multiple estimates."

On the one hand, mobile computational photography is about cameras. But these devices are much more than simply cameras: they are multipurpose mobile computing platforms that include technological features such as GPS, accelerometers, touch screens, etc. Many of these technologies can be leveraged and brought to help provide higher quality (and innovative) photographic capabilities. One such application is presented by Sindelar and Sroubek.[3] Their paper "Image deblurring in smartphone devices using built-in inertial measurement sensors" uses the accelerometers and gyroscopes in a smartphone to determine the motion trajectory while a photo is taken, allowing the blur caused by that motion to be removed from the picture.

Finally, in considering a hand-held device as a powerful computational imaging platform one can also consider other capabilities to add to the device to provide a more compelling user experience, such as a projector. In the paper "Compensating specular highlights for non-Lambertian projection surfaces," Kao et al.[4] describe a portable platform that includes both camera and projector. With these two devices in the same platform, the camera can be used in closed-loop fashion to correct (and augment) the projected image. In this paper, Kao et al. address the issue of compensating for specular highlights in particular.

Multimedia Content and Mobile Devices, edited by David Akopian, et al., Proc. of SPIE-IS&T Electronic Imaging,
SPIE Vol. 8667, 866701 · © 2013 SPIE-IS&T · CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2004110

References

1. A. Oberdörster and H. P. A. Lensch , "Resolution and sensitivity of wafer-level multi-aperture cameras," *J. Electron. Imag.* **22**(1), 011001 (2013). http://dx.doi.org/10.1117/1.JEI.22.1.011001
2. B. Krishnamurthy and A. Rastogi, "Refinement of depth maps by fusion of multiple estimates," *J. Electron. Imag.* **22**(1), 011002 (2013). http://dx.doi.org/10.1117/1.JEI.22.1.011002
3. O. Šindelář and F. Šroubek, "Image deblurring in smartphone devices using built-in inertial measurement sensors," *J. Electron. Imag.* **22**(1), 011003 (2013). http://dx.doi.org/10.1117/1.JEI.22.1.011003
4. C.-T. Kao, T.-H. Huang, H. Lee, and H. H. Chen, "Compensating specular highlights for non-Lambertian projection surfaces," *J. Electron. Imag.* **22**(1), 011004 (2013). http://dx.doi.org/10.1117/1.JEI.22.1.011004
5. T. Georgiev, A. Lumsdaine, and S. Goma "Special Section Guest Editorial: Mobile Computational Photography," J. Electron. Imag. 22(1), 010901 (2013). http://dx.doi.org/10.1117/1.JEI.22.1.010901.

Resolution and sensitivity of wafer-level multi-aperture cameras

Alexander Oberdörster
Fraunhofer Institute for Applied Optics and Precision Engineering
Albert-Einstein-Straße 7, D-07749 Jena, Germany
E-mail: alexander.oberdoerster@iof.fraunhofer.de

Hendrik P. A. Lensch
Eberhard Karls University Tübingen
Computer Graphics Group Sand 14, D-72076 Tübingen, Germany

Abstract. *The scaling limits of multi-aperture systems have been widely discussed from an information-theoretical standpoint. While these arguments are valid as an upper limit, the real-world performance of systems for mobile devices remains restricted by optical aberrations. We argue that aberrations can be more easily controlled with certain architectures of multi-aperture systems, especially those manufactured on wafer scale (wafer-level optics, WLO). We complement our analysis with measurements of one single- and one multi-aperture WLO camera. We examine both sharpness and sensitivity, giving measurements of modulation transfer function and temporal noise, and showing that multi-aperture systems can indeed reduce size without compromising performance.* *[DOI: 10.1117/1.JEI.22.1.011001]*

1 Introduction

In multi-aperture optics, a single optical system is replaced by an array of optical channels side by side. In a single-aperture system, its focal length and pixel pitch determine the rate it samples object space. In a multi-aperture system, this relation can be broken up by interlacing the views of adjacent channels so that they supersample object space. After capture, the channel microimages are assembled digitally to obtain a continuous image. Using this principle, the system thickness can be reduced while keeping sampling of object space constant. When the optics is considered to be diffraction-limited, however, either sensitivity or effective resolution has to be sacrificed.[1]

On the other hand, when system thickness is reduced, lens dimensions are reduced along with it. multi-aperture systems are often realized with micromanufacturing techniques, which are more accurate for lenses with small diameters and sags, leading to better optical performance.

We examine the balance of these two effects using the electronic cluster eye (eCLEY)[2] as one example of a multi-aperture system. The eCLEY uses supersampling to reduce system thickness and lens dimensions. Additionally, the total field of view of the system is divided; each channel only images a small field of view.

After reviewing related work in this area (Sec. 2), we discuss performance scaling and manufacturing issues in Sec. 3. Next, we treat the effects of image reconstruction on sharpness and noise (Sec. 4). We then compare the theoretical results to the actual performance of the eCLEY using measurements of the modulation transfer function (MTF) and the temporal noise in Sec. 5. Finally, we compare the MTF with a state-of-the-art single-aperture camera manufactured with wafer-level optics.

2 Related Work

An early small multi-aperture system was TOMBO,[3] which uses a low number of identical channels with the same viewing direction. The same principle has also been applied to macroscopic infrared focal plane arrays for remote sensing applications.[4] Flexible laboratory setups such as the Stanford large camera array have been valuable to investigate possible applications and configurations of multi-aperture systems, as well as yielding practical insights on how to calibrate these systems.[5] The eCLEY, in contrast, is specifically designed for precise and cost-effective manufacturing with microfabrication techniques and contains unique channels with different viewing directions.

Supersampling with multi-aperture systems is a natural extension of super-resolution from video sequences. Park et al.[6] have conducted a comprehensive review of existing methods. Registration techniques as well as reconstruction algorithms have been adapted to multi-aperture systems, for example by Nitta et al.[7] and Kanaev et al.[8,9] However, for images from real-world systems, simple shift-and-add schemes preceded by calibration with sub-pixel accuracy have remained popular, for example as reported by Kitamura et al.[10] An extended version of this type of scheme is also used for reconstructing images from the eCLEY.[11]

Independent of the applied reconstruction algorithm, the theoretical performance limits of thin optical systems were comprehensively investigated by Haney.[1] He concludes that multi-aperture systems with reduced length can only match the performance—sensitivity and resolution—of single-aperture systems at a significant increase in footprint.

Paper 12337SSP received Sep. 3, 2012; revised manuscript received Nov. 21, 2012; accepted for publication Dec. 4, 2012; published online Jan. 31, 2013.

Measurements of both sensitivity and resolution from experiments are rare. Figures for peak signal-to-noise ratio comparing ground truth with a simulation are stated most frequently, along with example images from the actual system. Portnoy et al.[4] give contrast measurements for a single frequency along with the signal-to-noise ratio.

We provide an analysis of the sensitivity and the resolution of multi-aperture, systems. We confirm our theoretical model with measurements of the MTF and the temporal noise of a specific system, the electronic cluster eye, which is described in Sec. 3.

3 Scaling in Multi-Aperture Systems

In this section, we discuss scaling in general multi-channel systems. As we will see with the example of the eCLEY, there are two aspects to any multi-aperture configuration that have different impacts on system volume and performance.

The eCLEY is based on the principle of interlaced tiles, as introduced in Ref. 2. Each optical channel of the eCLEY has a small field of view (FOV) and a unique viewing direction. The FOVs of adjacent channels overlap, together creating a larger FOV (Fig. 1). Their viewing directions are carefully tuned, so that pixels of one channel sample object space inbetween pixels of the adjacent channels (Fig. 2). In practice, one pixel does not have a discrete viewing direction; it integrates light over a solid angle. The implications are discussed in Sec. 3.2.

These two aspects of the concept serve specific purposes:

- *Segmenting* the system FOVs into smaller channel FOVs reduces the field each channel has to image. Aberrations can be controlled with a less complex lens system, reducing cost and making manufacturing easier and less prone to degradation because of tolerances.
- *Interleaving* the tiles achieves supersampling of object space and is responsible for reducing the effective focal length of the system, which is the lower limit to thickness.

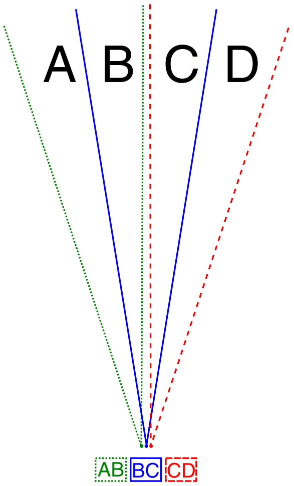


Fig. 1 The field of view of the (eCLEY) is segmented into multiple channels, each viewing a part of the total FOV. The FOVs of the channels overlap. Here, three channels are shown in different colors, in one dimension. The actual eCLEY has 17 × 13 channels.

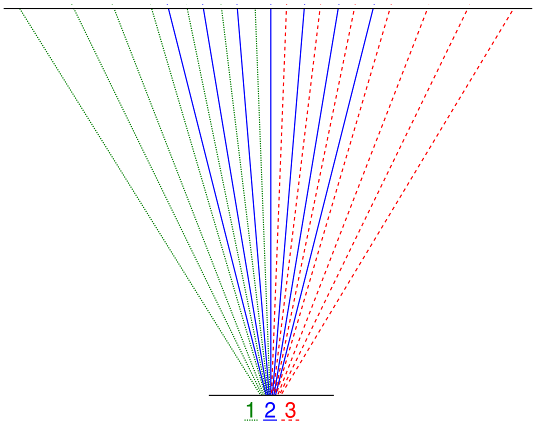


Fig. 2 The viewing directions of the eCLEY channels are adjusted carefully so that the pixels of one channel sample object space between the pixels of the adjacent channel. Three channels shown in one dimension, with seven pixels for each channel. The eCLEY has 39 × 39 pixels per channel.

Both aspects act in concert to decrease lens diameters. Interleaving achieves this goal directly, because at the same F-number, a smaller focal length leads to smaller lens diameters. Segmentation achieves the same goal indirectly, as less complex lens systems tend to have smaller lens diameters: The further away a lens is from the aperture stop, the larger it has to be to avoid vignetting of marginal rays. The more lenses a system has, the larger the axial extent of the system, leading to large lenses far away from the aperture.

We now investigate how multi-aperture systems compare to single-aperture systems in terms of light collection efficiency, resolution and physical size. For better clarity, we treat the effects of segmentation and supersampling separately.

3.1 *Light Collection*

First, we determine the light collection efficiency of a single-aperture system. Consider a setup with a scene emitting the radiance L, a lens with diameter D and effective focal length f, and an image sensor [Fig. 3(a)]. The sensor has the extent $w \times h$, divided into $n_x \times n_y$ pixels with a pitch of p_x. From the image plane, the lens subtends a solid angle of

$$\Omega = \frac{\pi(D/2)^2}{f^2}.$$

As the aperture takes on the radiance of the scene,[12] the sensor receives an irradiance of

$$I = \eta_{\text{lens}} \cdot \Omega \cdot L,$$

with the lens having an optical transmittance of η_{lens}. Each pixel integrates I over its photosensitive area, collecting a radiant flux of

$$\Phi_{\text{pix}} = \gamma_{\text{pix}} \cdot p_x^2 \cdot I,$$

where γ_{pix} is the fill factor of the pixel. The total flux collected by the sensor is

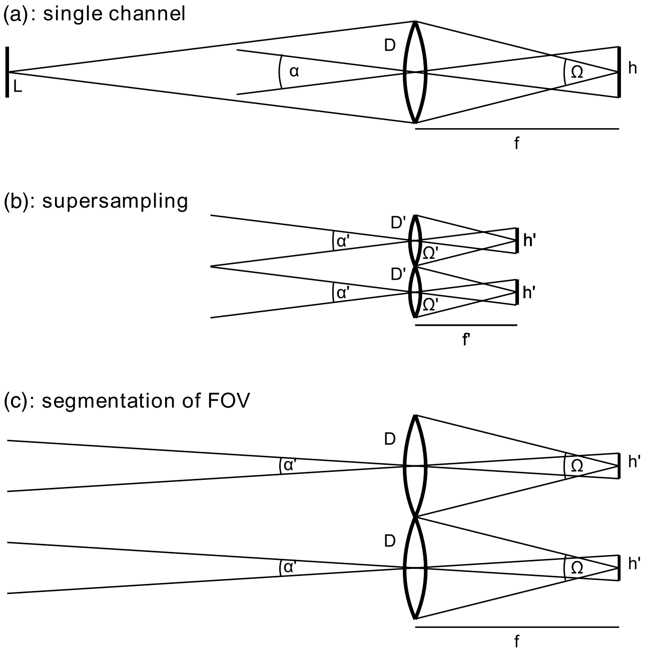


Fig. 3 Geometric properties of a single-aperture system (a) and two multi-aperture systems, supersampling (b) and segmenting (c) with lens diameter *D*, focal length *f*, image height *h*, and field of view *α*. The lens subtends an angle Ω at the image plane.

$$\Phi_{\text{tot}} = n_{\text{x}} n_{\text{y}} \cdot \Phi_{\text{pix}},$$

neglecting effects such as distortion or vignetting.

We discuss a supersampling multi-aperture system next [Fig. 3(b)]. To decouple sampling rate from pixel pitch, the single optical system is replaced by $N \times N$ channels side by side, with the supersampling factor *N*. In case of the eCLEY, the supersampling factor *N* is 2, though the number of channels is higher because the FOV is segmented.

Each channel is a scaled version of the original system (see Ref. 2), so $f' = f/N$, $D' = D/N$ and each system retains the F-number of the original camera. Therefore, in each system, $\Omega' = \Omega$. Consequently, each pixel records the same flux Φ_{pix} as in the single-aperture case. As the system samples the same FOV (solid angle) as the original system with the same sampling rate, the total amount of samples—or pixels—stays the same. Therefore,

$$\Phi'_{\text{tot}} = \Phi_{\text{tot}}.$$

Next, we segment the FOV α of the camera into $M \times M$ channels [Fig. 3(c)]. In case of the eCLEY, *M* is 8 horizontally and 6 vertically. The geometry of each of these channels is identical to that of the original optical system: Both *f* and *D* stay the same. The FOV of each channel is limited, however, by reducing the image size in each channel. The viewing direction of the channel is selected by introducing a lateral offset between optical system and image. Again neglecting distortion and vignetting, in each channel, partial FOV is α/M and image size is $w/M \times h/M$. As sampling rate and pixel size are kept the same, each channel now uses $n_{\text{x}}/M \times n_{\text{y}}/M$ pixels. If either distortion or vignetting are not corrected in the optical system, they affect single-aperture and multi-aperture systems in the same way.

Because the focal length is still f and the aperture diameter is still D, Ω remains the same and $\Phi'_{\text{pix}} = \Phi_{\text{pix}}$. As the total amount of pixels in the system does not change, $\Phi'_{\text{tot}} = \Phi_{\text{tot}}$.

In summary, both segmenting and supersampling multi-aperture systems collect the same amount of light as single-aperture systems, as long as the *F*-number and the total photodetector area are kept constant.

3.2 *Sharpness*

In this section, we will first examine the effects of supersampling on image sharpness. Segmentation of the FOV will be of relevance in the course of the discussion.

By using supersampling, a digital camera can be made thinner without sacrificing sampling rate in object space and without requiring a smaller pixel pitch. To retain actual optical resolution in object space along with sampling rate, however, the MTF of the channels in image space has to keep up with the sampling rate.

Supersampling multiplies the image plane sampling frequency f_S and the Nyquist frequency f_{Ny} by a factor of N. Therefore, the MTF should now show significant modulation up to $f'_{\text{Ny}} = N \cdot f_{\text{Ny}}$. Consequently, it has to improve considerably.

The MTF of a camera is the product of the MTF of the lens and the sensor, where the sensor MTF consists of a geometrical component and a component resulting from crosstalk between pixels:

$$\text{MTF}_{\text{S}} = \text{MTF}_{\text{O}} \cdot \text{MTF}_{\text{G}} \cdot \text{MTF}_{\text{C}}.$$

MTF_{G} describes spatial integration over the photodetector. For square photosites, it is the Fourier transform of the rect function with the width of the photosensitive area p_{p}:

$$\text{MTF}_{\text{G}}(f) = \sin c(\pi p_{\text{p}} f).$$

The pixel pitch p_{x} stays the same. While we are still free to choose a smaller p_{p}, light sensitivity decreases with photosensitive area, or p_{p}^2. Therefore, we assume MTF_{G} to be constant.

Crosstalk depends on the chief ray angle of light incident on the sensor and on sensor technology. Neither of them changes for multi-aperture systems. Therefore, MTF_C is constant as well.

The burden of improving the system MTF is therefore placed entirely on the optical MTF. As described by Lohmann,[13] if an optical system is scaled by the factor $1/N$, the area of an image point A_p scales as

$$A_p(N) = \lambda^2 F^2 + \left(\frac{1}{N}\right)^2 \bar{\xi}^2, \tag{1}$$

for light of the wavelength λ and with aperture stop number F and the lateral aberration ξ ($\bar{\xi}^2$ is its Gaussian moment).

When diffraction is neglegible, the diameter of an image point is

$$d_P(N) \propto \sqrt{A_p(N)} \approx \frac{1}{N} \bar{\xi}$$

and the resolution limit therefore scales linearly with the size of the system. Supersampling with *N* scales each individual optical channel of the multi-aperture system by $1/N$. Point diameters are therefore scaled by $1/N$.

Segmenting the FOV also has beneficial effects. Many of the Seidel aberrations depend on field height h.[12] Field curvature and astigmatism, for example, increase with h^2. Therefore, segmenting the FOV into M parts reduces aberrations accordingly.

However, quantifying the benefit exactly is not possible so easily. The well-known scaling laws for Seidel aberrations only apply to imaging with a single lens. In practice, aberrations are partially corrected with multilens systems, whose behaviour is more complex. This is true even for low-cost mass-market cameras for mobile devices. With a certain amount of correction, higher-order aberrations cannot be neglected any more; these aberrations also defy description by simple scaling laws.

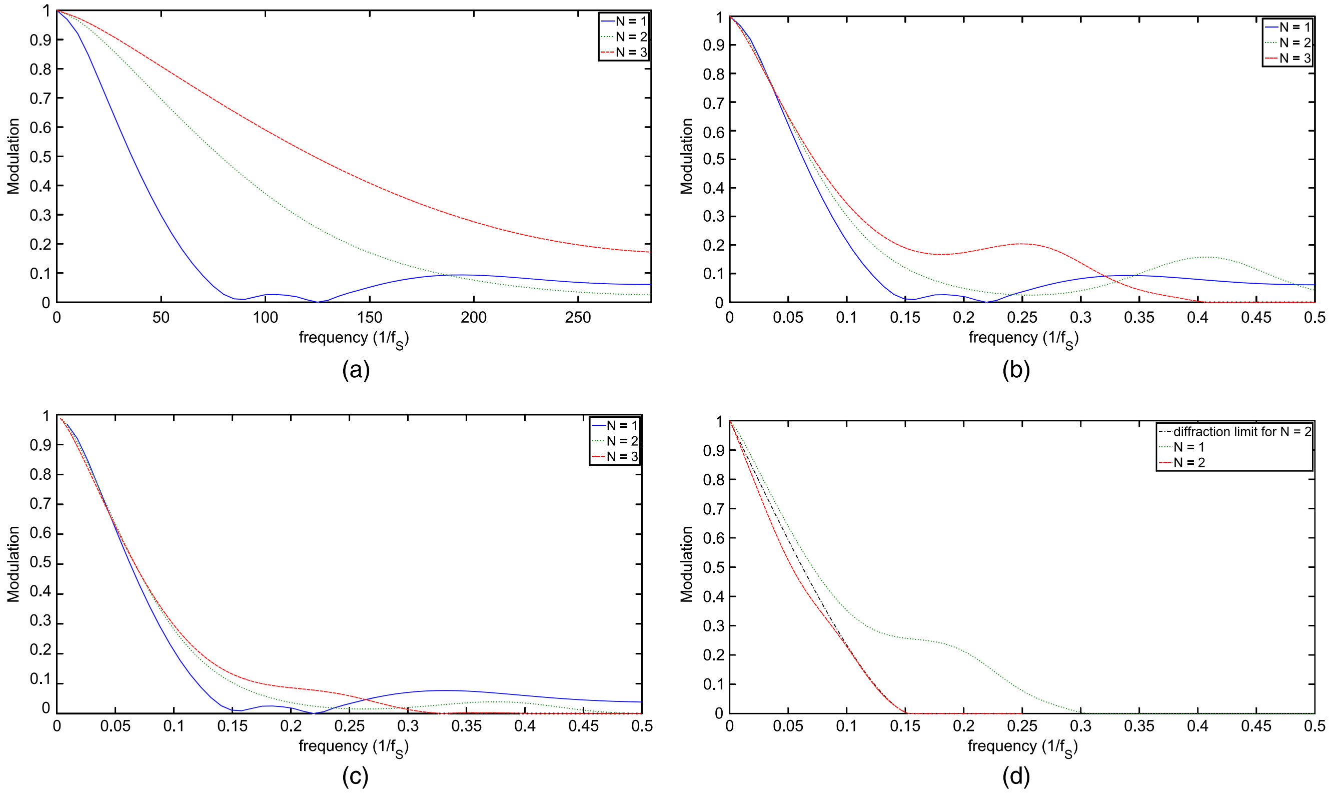
In conclusion, optical MTF is indeed improved considerably by scaling. This is necessary to retain optical resolution in object space. As an example on how this works out, Fig. 4 shows the MTF of a system with aberrations ($N = 1$) and the effect of scaling down this system ($N > 1$). First, only optical MTF is plotted on an absolute frequency axis (a). Optical MTF is improved as expected for increasing N. However, when the f axis is normalized to the sampling frequency f_S, which scales with N, improvement is less apparent (b). When we include pixel MTF, system MTF is similar for all N (c). Therefore, object-space sharpness of the supersampled systems is comparable to the original system. Increasing N further still improves optical MTF, but pixel MTF cancels this gain.

Enhancement to the optical MTF itself is limited by diffraction, which is independent of system scaling. This is illustrated in Fig. 4(d). Here, we used the same optical system as in Fig. 4(a), but scaled it down by 4, so f is now 1 mm. Again, optical MTF is improved for $N = 2$, but improvement is limited by diffraction (dashed line). The resulting object space sharpness for $N = 2$ is lower than the sharpness of the original system.

3.3 *Manufacturing Tolerances*

When manufacturing a lens system, deviations in lens curvatures, distances, decenter and tilt degrade the system performance. When scaling a lens system down, deviations have to be smaller as well, or performance is compromised. As a simple example, consider a single thin lens with focal length f and diameter D positioned so that it focuses light from point P onto an image plane (Fig. 5). When the lens is moved from its correct image plane distance f by a deviation Δs, defocus leads to an image point diameter $d_P = \Delta s \cdot D/f$. The smaller the system, the smaller d_P has to be to retain sharpness. Accordingly, Δs has to be smaller as well.

The same is true for the focal length of the lens: For a plano-convex lens, f is proportional to lens radius R,[12] so a deviation ΔR leads to a new focal length f' with $\Delta f = f - f'$. Δf effectively is a defocus shift Δs, leading to an image point diameter analogous to a lens shift.



(a)

(b)

(c)

(d)

Fig. 4 Scaling effects for a system with two optical surfaces and $f = 4$ mm, F2.4, simulated on-axis MTF curves. (a) Optical MTF improves for scaling the optical system down. Supersampling factors $N = 1$ to 3; $N = 1$ is the original system. Plotted on absolute frequency axis (cycles/mm). (b) The same curves plotted on a frequency axis relative to the image-space sampling frequency f_S, which scales with N. (c) Optical MTF multiplied with geometrical pixel MTF. (d) Improvement in optical MTF is limited by diffraction effects, shown by scaling the system down by a factor of 4.

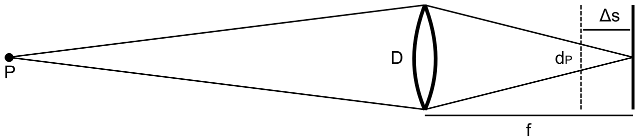


Fig. 5 Image point diameter d_P of a defocused optical system with focal length f and lens diameter D, with the image plane moved by Δs.

For perspective, with current pixel technology, $d_P < 2\ \mu$m is desirable. This requires a focus shift of less than $2d_P = 4\ \mu$m.

The ability to meet the required tolerances depends on the technology that is used to manufacture and assemble the lens components. A suitable technology for multi-aperture systems is wafer-level optics (WLO), as multiple lenses side by side are manufactured and aligned in parallel. Assembly of the lens components can be achieved with the required micron precision.[14]

Critical, however, is precision during replication of the lens components. Lenses are manufactured from certain polymers by molding and ultraviolet curing. During hardening, these materials shrink significantly. The amount of shrink is proportional to the lens volume. Lens volume grows with the square of the lens radius and linearly with lens sag. Therefore, small lenses with low sags are preferable. Molding tools are adjusted to anticipate shrink; however, shrink has a certain spread that is proportional to shrink itself. The hardened lenses therefore still have form deviations that scale with N^3.

Using a multi-aperture architecture—either supersampling or segmenting—decreases lens diameters. For a supersampling factor of N, lens diameter decreases by N, lens sag also decreases by N and we can expect deviations to decrease with the cube of N. Segmenting the FOV also reduces lens diameters, having a similar effect on deviations.

In conclusion, while tolerances have to be tighter for scaled-down lens systems, the fact that small lenses can be manufactured with less shrink makes it easier to meet these tolerances. Therefore, sharpness of actual, mass-manufactured camera systems can benefit significantly from a multi-aperture architecture. This result contradicts the theoretical analysis in Sec. 3.2, which suggested that multi-aperture systems can at best reach a performance comparable to single-aperture systems.

3.4 *Volume*

We already established that system thickness is reduced by supersampling. In some applications, however, total system volume is more relevant than thickness. Therefore, we now examine how multi-aperture system volume V' compares to that of a single-aperture system V. We again treat the two different architectures (supersampling and segmented FOV) separately. In both cases, we first derive the footprint of the system. It is given by either sensor footprint A_{sens} or total aperture area A_{tot}, depending on which one is larger. In the single-aperture case, A_{tot} is simply the area of the single system aperture. The values for the multi-aperture system are A'_{sens} and A'_{tot}, which is now the sum of all individual aperture areas A'. Next, we derive system height. In both cases, system height scales with effective focal length f. To f, a part of the optical system thickness h_{opt} is added, depending on system complexity and placement of the principal planes. We disregard the thickness of the image sensor, sensor carrier and casing, as these values are small compared to the focal length and are not affected by the system architecture.

Supersampling: As noted in Sec. 3.1, neither the pixel pitch nor the total number of pixels on the sensor change. Therefore, $A'_{\text{sens}} = A_{\text{sens}}$. This is also true for A'_{tot}:

$$A'_{\text{tot}} = N^2 A' = N^2 \pi \left(\frac{D'}{2}\right)^2 = \pi \left(\frac{D}{2}\right)^2 = A_{\text{tot}},$$

assuming circular apertures with diameter D. Therefore, system footprint stays the same. f, in contrast, is reduced by a factor of N. As the lens dimensions all scale with N, $h'_{\text{opt}} = h_{\text{opt}}/N$. Therefore, $h'_{\text{tot}} = h_{\text{tot}}/N$ and $V' = V/N$. In conclusion, a supersampling system is not only thinner, but also has less volume than a single-aperture system.

Segmentation of FOV: Again, pixel size and number stay the same, so $A'_{\text{sens}} = A_{\text{sens}}$. However, the single-aperture with area A_{tot} is now replaced with M copies of the original aperture. Aperture area therefore is increased:

$$A'_{\text{tot}} = M \cdot A_{\text{tot}}.$$

The proportion of A_{tot} to A_{sens} in a camera is approximately the proportion of the corresponding lengths:

$$\frac{D_{\text{sens}}}{D} = \frac{2f\ \tan\frac{\alpha}{2}}{\frac{f}{N}} = 2N\ \tan\frac{\alpha}{2}.$$

Miniaturized cameras tend to have a large FOV. If we assume $N = 2.8$ and $\alpha = 70°$, $D_{\text{sens}}/D \approx 4$. Therefore, the sensor width is larger than the lens diameter and system footprint is given by A_{sens} for $M \leq 4$.

Effective focal length is not affected. Reduced optical system complexity in each channel, however, decreases h'_{opt} slightly. Therefore, system volume V' is smaller than V for moderate segmentation of FOV, but increases with M^2 for large M.

This analysis does not consider additional volume consumed by the system casing, structures for suppressing stray light or walls separating channels. The latter are needed to prevent crosstalk between channels. In current systems such as the eCLEY, structures for crosstalk suppression do consume a considerable amount of space between channels. They therefore increase the total volume of the system and lead to unused areas on the image sensor. For reducing this waste of space and sensor area, very thin vertical or slanted walls have to be manufactured. Techniques for cheaply fabricating these structures are currently being developed.

4 Reconstruction

In the last section, the theoretical and practical scaling characteristics of multi-aperture systems were discussed. In the next section, these characteristics are verified with measurements. To compare the analysis with the measurements, we have to consider that in a multi-aperture system, a multitude of images have to be combined into a continuous image. This image reconstruction step has effects on image sharpness and alters the noise characteristics of the system. In principle, neither can be improved without negatively affecting the other. As the focus of this publication lies on the scaling characteristics of multi-aperture systems per se, we do not attempt an exhaustive analysis of this topic. Instead, we quantify the effects of a single, simple reconstruction scheme, a shift-and-add algorithm with Gaussian interpolation. In this case, the effect is a decrease in noise and a loss in sharpness.

4.1 *Algorithm*

We treat each recorded pixel as a measurement of the light incident on the camera from a specific direction. The pixel viewing directions are derived from the model of the optical system; it includes effects such as geometric distortion. We intersect each of these pixel viewing rays with a virtual focal plane (Fig. 6). The intersection points of viewing rays and focal plane form a two-dimensional cloud of measurements, an irregular sampling of the scene (irregular because of parallax and geometric distortion of the channels; Fig. 7). To render an image from this point cloud, we create a regular sampling of the scene by interpolation. For each pixel of the target image, contributions from the nearest measurement points available are added, weighted with the distance from the measurement coordinate to the target pixel (Fig. 8).

From the distance r, the weight $W_{x,y,i,j}$ of the neighbor j, contributing to the target pixel at coordinates x and y, is calculated as

$$W_{x,y,i,j} = e^{-w\cdot r^2_{x,y,i,j}}, \qquad (2)$$

where w is an adjustable filter width. The weights $W_{x,y,i,j}$ are normalized so that $\sum W_{x,y,i,j} = 1$.

The algorithm is presented in full in Ref. 11.

4.2 *Sharpness*

Interpolation can be treated as a spatial filter. Calculating the Fourier transform of the filter kernel yields the MTF of the interpolation operation. The interpolation kernel is the continuous version of Eq. (2), the Gaussian

$$K(r) = e^{-wr^2},$$

again with the filter width w and the distance from target pixel to measurement coordinate r, in units of pixels. The effect is a loss in modulation at higher frequencies.

4.3 *Noise*

Each target pixel is calculated from the weighted mean of ν measurements. If a value V is calculated as the weighted sum of measurements $m_{x,y,i,j}$ with equal uncertainties σ_{m},

$$\sigma_{\mathrm{V}} = \sigma_{\mathrm{m}} \cdot \frac{1}{\sqrt{\sum W^2_{x,y,i,j}}},$$

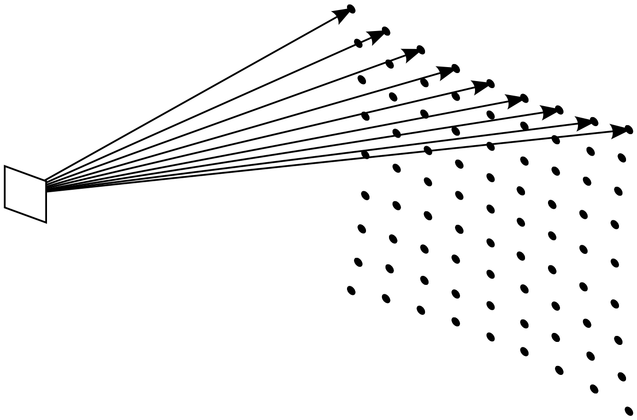


Fig. 6 Viewing directions of the pixels of one channel of a multi-aperture system (arrows), intersected with a virtual focal plane (points). The points show how this channel samples object space on a specific focal plane.

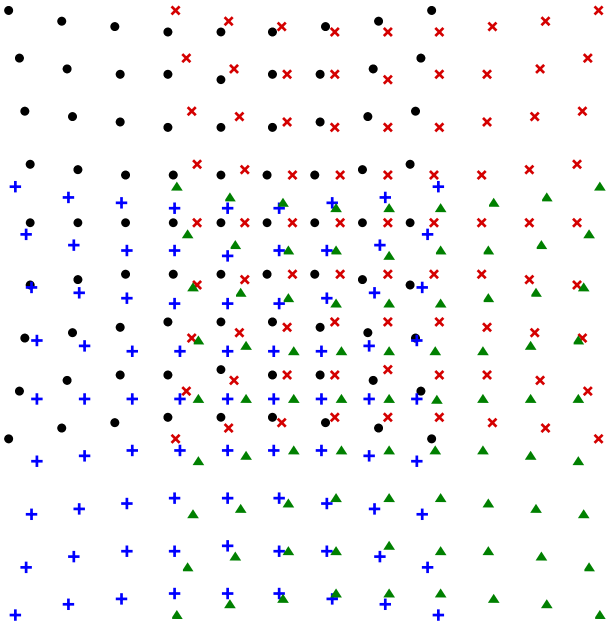


Fig. 7 Placing the measurement of multiple channels (four in this case, shown in different colors) on a common focal plane according to the distances of channels and focal plane creates an irregular point cloud.

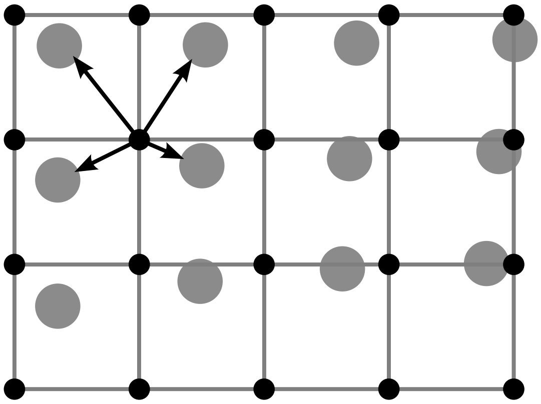


Fig. 8 To render an image from an irregular point cloud, a regular grid is overlaid on the point cloud. Each of the grid intersections represents a target pixel in the image. Each of the target pixels (black) is calculated by interpolating the nearest measurements from the point cloud (grey).

with the weights $W_{x,y,i,j}$. The weights are different for each target pixel. They depend on the distance of the measurement to the target pixel r and on the filter width w.

For the following analysis, we first assume uniform density of measurements. In one extreme case, the target pixel is exactly on top of a single source pixel. Choosing a filter width of $w = 2$ and setting $r = 0$ in Eq. (2), $W'_i = 1$ (not normalized yet). Four other pixels are at the distance of $r = 1$ pixel, yielding $W'_i = 0.13$. Four further pixels are at a distance of $r = \sqrt{2} = 1.4$, yielding $W'_i = 0.02$. Normalizing yields contributions of $W_i = 0.63$, 0.08 and 0.01, respectively. Noise is consequently reduced by a factor of $1/\sqrt{0.63^2 + 4 \cdot 0.08^2 + 4 \cdot 0.01^2} = 1.54$. In the other extreme, the target pixel is exactly between four source pixels, each contributing equally. No other pixels contribute significantly due to their large distance. Noise is decreased by $\sqrt{4} = 2$.

In conclusion, noise is decreased by a factor of about 1.54 to 2.

5 Results

In this section, to support the conclusions of the last section, we compare one state-of-the-art single-aperture WLO camera, the OmniVision CameraCube, with a WLO multi-aperture system, the eCLEY. To verify sharpness, we directly compare the MTF of these systems. Direct comparison of the sensitivity of the two cameras is not useful, as they employ different sensors with different pixel pitches (1.75 μm versus 3.2 μm). The route taken is described next.

5.1 *Sensitivity*

According to theory, the eCLEY should have the same sensitivity as a single-aperture camera with the same aperture F3.7. For verification, we took an image of a uniformly lit target with the image sensor used in the eCLEY, with a single-aperture 16-mm lens (Schneider Cinegon) attached and set to F3.7. The same target was also recorded with an eCLEY. To avoid linearity issues, the exposure time t_{exp} was adjusted so that both cameras recorded roughly the same mean value (DN) on the target area. The values recorded and the corresponding exposure times t_{exp} were:

	Cinegon	**eCLEY**
Value	146	144
t_{exp}	3.3 ms	4.2 ms

The longer exposure time for the eCLEY suggests a lower sensitivity (by a factor of 0.77). We suspected that this discrepancy is caused by the way the eCLEY objective is attached to the sensor. The clear epoxy filling the gap between objective and substrate has a refractive index close to that of the per-pixel microlenses on the sensor, thereby rendering them ineffective.

We validated our suspicion by attaching a plane glass to one half of the sensor, again filling the air gap with epoxy. We then recorded the same target area with the treated sensor, again imaging the target area with the Cinegon lens set to F3.7. We measured values of 140 on the sensor half without plane glass and 110 on the other half, yielding a relative sensitivity $\gamma = 0.71$. This figure also gives us an estimate on the relative area of the photosensor on each pixel. We assume a perfect efficiency of the pixel microlenses and set the fill factor of the sensor pixels to $\eta_{\mathrm{pix}} = 0.71$.

Sensitivity of the eCLEY consequently has to be adjusted by a factor of 1.40, yielding an adjusted relative sensitivity of about $\gamma = 1.1$, higher than the single-aperture lens. The new discrepancy is most likely caused by losses due to internal reflections in the Cinegon lens, which has more air-glass surfaces than the eCLEY objective.

Note that the loss in sensitivity due to the loss of the per-pixel microlenses is not inherent to multi-aperture systems or WLO. The attenuation can be avoided by replacing the bottom substrate with a spacer layer that introduces an air gap between optics module and sensor.

5.2 *Noise*

As illustrated in Sec. 4.3, the reconstruction scheme that we use interpolates measurements, which should reduce noise. To verify this claim, we first established the image noise of the sensor used in the eCLEY.

To this end, we recorded 100 images of a scene with a wide dynamic range, using the eCLEY. The recorded images contain microlens images with all values in the dynamic range of the camera, from 0 to 255. As we are interested in temporal noise, we evaluated the temporal behaviour of each pixel. For each of them, the mean and the standard deviation were calculated. Pixels were then distributed into bins of integer values according to their mean. The resulting distribution of standard deviation over image signal is plotted on a log-log scale in Fig. 9.

We proceeded to process each of the recorded images with our reconstruction algorithm, creating continuous images from the raw images. Filter width w was set to 2.0. These processed frames were characterised pixel by pixel as before, yielding another distribution of standard deviations, this time including reconstruction. This distribution is also plotted in Fig. 9.

Comparing the plots shows that noise is attenuated by a factor of 2.0, being at the top end of our prediction from Sec. 4.3 and validating our model of the reconstruction algorithm.

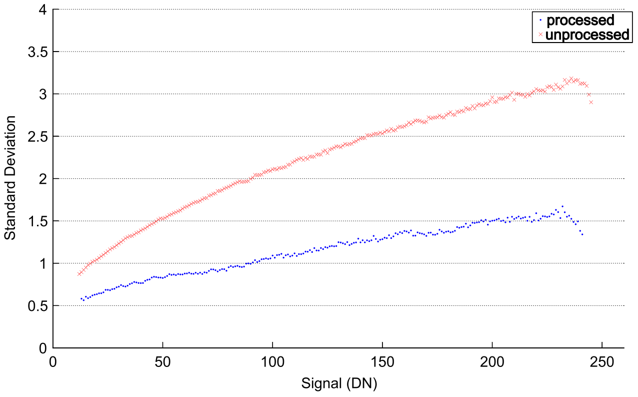


Fig. 9 Temporal pixel noise (standard deviation) of the sensor in the eCLEY, unprocessed microlens image and reconstructed image (processing filter width $w = 2.0$).

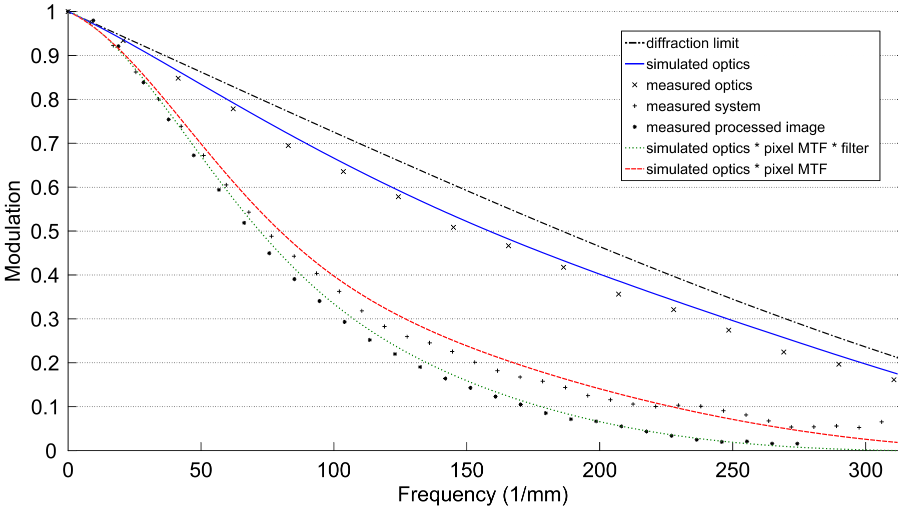


Fig. 10 MTF of the eCLEY. From the diffraction limit downwards, one additional component is added to the simulation for each curve. The complete system MTF is plotted at the bottom, with measurements confirming the model at several steps.

5.3 *Sharpness*

With the results from the previous Secs. 5.1 and 5.2, we have a complete model of the eCLEY transfer function. Figure 10 shows simulations of all components. On top, the diffraction limit for F3.7 is plotted. The optical MTF of the eCLEY central channel is quite close to this limit. It is calculated from a ZEMAX model of the eCLEY objective lens.

Next, the contribution of the sensor is multiplied with the optical MTF. From Sec. 5.1, we assume square photodiodes with a width of $\sqrt{\gamma} \cdot 3.2\ \mu\text{m} = 2.7\ \mu\text{m}$. The crosstalk was modeled as a Gaussian and fitted to results from Refs. 15 and 16.

Finally, the reconstruction step is considered by multiplying the filter kernel K to optical and sensor MTF, with filter width $w = 2.0$.

To validate this model, we measured three steps of the image formation process: The optical MTF of a single eCLEY channel, the MTF of a single channel including sensor and the MTF of the complete system, including reconstruction. Each measurement was carried out with the slanted-edge method.[17]

The measurements (also plotted in Fig. 10) match the predicted MTFs quite closely, validating our model of the eCLEY. In summary, we have shown that

- Microlens arrays can be manufactured with low tolerances, so that they closely match the simulated performance;
- the image sensor plays a significant part in the total MTF of a supersampling multi-aperture system, because the photodiodes are larger than the virtual pixel pitch; and
- the reconstruction algorithm reduces noise at the expense of reduced sharpness.

Note that no calibration was necessary to align the microlens images in the reconstruction step. The distributions of

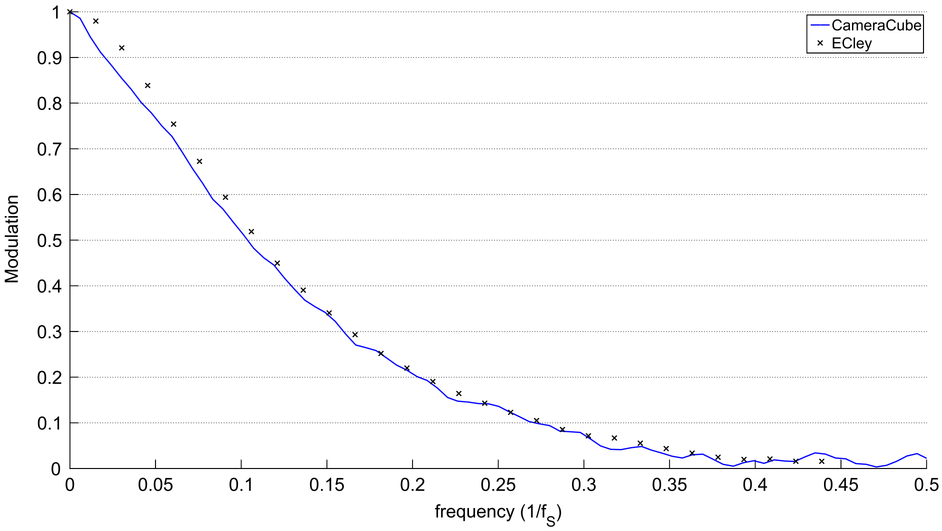


Fig. 11 MTF of OmniVision CameraCube and eCLEY. System MTFs plotted relative to the sampling frequency of each system.

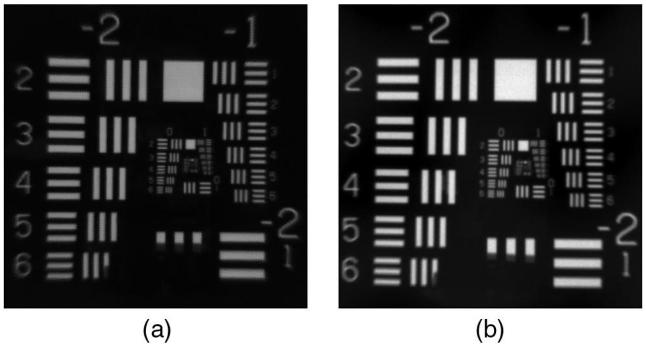


Fig. 12 Photographs of a backlit USAF test target with the eCLEY (a) and the OmniVision CameraCube (b). For comparable results, all postprocessing, sharpening and noise reduction was disabled in both cases.

the pixels on the virtual focal plane were taken directly from the ZEMAX model. This fact demonstrates the manufacturing and alignment precision of the microlens array.

Finally, we compared the MTFs of the eCLEY and an OmniVision CameraCube. Figure 11 shows the complete system MTF of both systems. We normalized the frequency axis on the image-space sampling frequency of each camera, which is $1/1.75\ \mu\text{m} = 571$ cycles/mm for the CameraCube and $2 \cdot 1/3.2\ \mu\text{m} = 625$ cycles/mm for the eCLEY. The eCLEY exhibits comparable sharpness at reduced total track length.

Figure 12 compares two shots of an USAF test target recorded with the eCLEY and the CameraCube. These photographs also demonstrate similar sharpness for both systems.

5.4 *Volume*

Despite having a larger pixel pitch (3.2 μm instead of 1.75 μm), the eCLEY has a shorter track length than the CameraCube (1.4 mm instead of approximately 2.2 mm). This is the result of 2× supersampling in the eCLEY ($N = 2 \times 2$, in x and y), which cuts total track length in half. Additionally, the eCLEY has only one optical surface per channel instead of the two surfaces of the CameraCube,[18] which also reduces thickness.

Footprint, on the other hand, is larger for the eCLEY, being 6.8×5.2 mm compared to 3.2×2.8 mm. This 4× increase in footprint is partly due to the larger pixel pitch, partly a result of the segmentation of the FOV ($M = 7 \times 4$, in x and y), as deduced in Sec. 3.4.

6 Conclusion

We provided an analysis of the sensitivity, resolution and volume of two types of multi-aperture systems. Compared to single-aperture cameras, systems which supersample object space significantly reduce volume at constant sensitivity. Matching the resolution is challenging, but possible for low supersampling factors N in cases where the optical system is not diffraction limited. Systems that segment the FOV increase footprint and volume, but simplify the optical system, which helps reducing track length. Both principles can be used in tandem to design cameras with lower track length and sufficient sharpness, as demonstrated with our measurements of the eCLEY.

In this analysis, we assumed monochrome sensors without color filter arrays (CFAs). For a sensor with CFA, the color channels are traditionally undersampled, potentially leading to aliasing. This is a favorable premise for a supersampling multi-aperture system: with $N = 2$, aliasing can be avoided and, at the same time, track length can be halved. Extending the discussion of this publication to color systems therefore is a promising direction.

Finally, plenoptic cameras are in essence also multi-aperture systems. In the focused plenoptic camera, multiple channels view overlapping parts of an intermediate, demagnified image of the subject. Each channel has a limited field of view; the sampling patterns of the channels are interleaved so that the intermediate image is supersampled. This translates into increased resolution, however, only when the combined MTF of objective lens, microlens array and sensor is sufficiently large.

In multi-aperture and plenoptic cameras, filtering can regain sharpness at the price of increased noise. This is traditionally the subject of superresolution algorithms. Work in this area has focused on aligning the multiple views of the subject accurately and robustly, with the required subpixel resolution. When the optical system is manufactured with sub-micron precision, good alignment can be already be achieved from the geometry of the design. Similarly, the transfer function can be simulated with useful precision. To examine whether the available data is sufficient to increase sharpness without introducing artifacts would be another interesting topic.

References

1. M. W. Haney, "Performance scaling in flat imagers," *Appl. Opt.* **45**(13), 2901–2910 (2006).
2. A. Brückner et al., "Thin wafer-level camera lenses inspired by insect compound eyes," *Opt. Express* **18**(24), 24379–24394 (2010).
3. J. Tanida et al., "Thin observation module by bound optics (TOMBO): concept and experimental verification," *Appl. Opt.* **40**(11), 1806–1813 (2001).
4. A. Portnoy et al., "Design and characterization of thin multiple aperture infrared cameras," *Appl. Opt.* **48**(11), 2115–2126 (2009).
5. V. Vaish et al., "Using plane + parallax for calibrating dense camera arrays," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Vol. 1, pp. I-2–I-9, IEEE (2004).
6. S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Process. Mag.* **20**(3), 21–36 (2003).
7. K. Nitta et al., "Image reconstruction for thin observation module by bound optics by using the iterative backprojection method," *Appl. Opt.* **45**(13), 2893–2900 (2006).
8. A. V. Kanaev et al., "TOMBO sensor with scene-independent superresolution processing," *Opt. Lett.* **32**(19), 2855–2857 (2007).
9. A. V. Kanaev et al., "Analysis and application of multiframe superresolution processing for conventional imaging systems and lenslet arrays," *Appl. Opt.* **46**(20), 4320–4328 (2007).
10. Y. Kitamura et al., "Reconstruction of a high-resolution image on a compound-eye image-capturing system," *Appl. Opt.* **43**(8), 1719–1727 (2004).
11. A. Oberdörster et al., "Correcting distortion and braiding of micro-images from multi-aperture imaging systems," *Proc. SPIE* **7875**, 78750B (2011).
12. W. J. Smith, *Modern Optical Engineering: The Design of Optical Systems*, McGraw-Hill, New York (1990).
13. A. W. Lohmann, "Scaling laws for lens systems," *Appl. Opt.* **28**(23), 4996–4998 (1989).
14. R. Völkel et al., "Technology trends of microlens imprint lithography and wafer level cameras (WLC)," in *Conf. Micro-Optics*, Brussels, Belgium (2008).
15. J. Chen et al., "Imaging sensor modulation transfer function estimation," in *Proc. 17th IEEE Int. Conf. Image Process.*, pp. 533–536, IEEE, Hong Kong (2010).
16. X. Jin et al., "Sensitivity and crosstalk study of the zero gap microlens used in 3.2 μm active pixel image sensors," *Microelectron. Eng.* **87**(4), 631–634 (2010).
17. S. E. Reichenbach, S. K. Park, and R. Narayanswarmy, "Characterizing digital image acquisition devices," *Opt. Eng.* **30**(2), 170–177 (1991).
18. R. Fraux, "Omnivision's VGA wafer-level camera," *3D Packaging* **22**, 26–27 (2012).

Alexander Oberdörster is a researcher at the Fraunhofer Institute for optics and precision engineering in Jena, Germany. He graduated in physics from the University of Düsseldorf and the Fraunhofer ISE in Freiburg. At the ISE and at Spheron VR AG, he studied and designed devices for measuring optical scattering properties of surfaces (BRDFs). Moving on to the Fraunhofer IIS in Erlangen, he developed cameras for digital cinematography and related technologies. Currently, he is working on image processing algorithms for multi-aperture imaging systems. His research interests include computational photography, non-uniform sampling and reconstruction and the measurement of light fields.

Hendrik P. A. Lensch holds the chair for computer graphics at Tübingen University. He received his diploma in computer science from the University of Erlangen, in 1999. He worked as a research associate at the computer graphics group at the Max-Planck-Institut für Informatik in Saarbrücken, Germany, and received his PhD from Saarland University, in 2003. He spent two years (2004 to 2006) as a visiting assistant professor at Stanford University, USA, followed by a stay at the MPI Informatik as the head of an independent research group. From 2009 to 2011, he was a full professor at the Institute for Media Informatics at Ulm University, Germany. In his career, he received the Eurographics Young Researcher Award 2005, was awarded an Emmy-Noether-Fellowship by the German Research Foundation (DFG) in 2007 and received an NVIDIA Professor Partnership Award in 2010. His research interests include 3-D appearance acquisition, computational photography, global illumination and image-based rendering, and massively parallel programming.

Refinement of depth maps by fusion of multiple estimates

Balaji Krishnamurthy
Anubha Rastogi
Adobe Systems India Pvt Ltd
I-1A, Sector 25A
Noida, 201307, India
E-mail: kbalaji@adobe.com

Abstract. *Computing depth maps from a stereo pair is a well-studied problem of computer vision, and a large number of methods and cost functions have been proposed. The methods have different strengths and weaknesses and different error characteristics. That is, a pixel could be assigned an erroneous depth value by some methods, but other methods could assign the correct depth value to the same pixel. We describe a method that can make use of multiple depth estimates of low quality and fuse them, by trying to retain the correct depth values and rejecting the incorrect depth values, in order to obtain a more accurate result. We observe that depth values of pixels located in smooth areas of a depth estimate and depth values which survive left-right cross validation tend to be more accurate. Our method makes use of a reliability criterion-based upon the smoothness and cross-validation of the depth estimates that allows us to patch the estimates together and obtain a higher quality result. © 2013 SPIE and IS&T.* [DOI: 10.1117/1.JEI.22.1.011002]

1 Introduction

Computing depth maps from a stereo pair is a well-studied problem of computer vision, but a completely satisfactory solution is still elusive.[1,2] The methods proposed are either computationally demanding or the quality of the depth maps is low. Many of the proposed methods for obtaining high quality depth maps from a stereo pair formulate the problem in terms of finding the global minimum of an appropriate energy function. Global optimization techniques like cooperative optimization,[3] graph cut,[4–6] or belief propagation[7] are used to minimize the energy. However, the global optimization techniques are computationally demanding, especially as the number of depth labels increase. Good results have been obtained with some local methods[8–10] and some hybrid methods[11] too, but implementing them in an efficient manner is a challenge due to the large and adaptive support windows that are required.

There has also been a significant amount of progress in real-time or near real-time methods, with steadily improving quality and performance.[12,13] Many different matching cost functions have also been proposed.[14,15] Due to the availability of a large number of stereo matching strategies with different error characteristics, it is fruitful to study if it may be possible to blend the results of the existing methods and produce a depth map of higher quality. This aspect has not received much attention in the recent years. Our results show that fusion or blending methods can be used to significantly enhance the quality of real-time depth estimation methods.

In this paper, we propose a method for fusing multiple depth estimates, which makes use of a reliability criterion-based upon the smoothness of the depth estimates.

2 Related Work

Some work on depth map fusion has been done in the area of multiview three-dimensional reconstruction. In Ref. 16, several depth estimates for a reference view are obtained by projecting the depth maps of each acquired view onto the reference view. The weighted average of the depth estimates at each pixel, based on the confidence values of the depth estimates at each view, is used to combine the estimates. In Ref. 17, the authors propose a confidence measure of a depth estimate at each pixel, based on the shape of the cost aggregation curve. They also propose a variation of the weighted average method of Ref. 16 based on their confidence measure. These methods were originally proposed for fusion of depth estimates obtained from different views, based on a single depth estimation algorithm. To the best of our knowledge, there has not been any study on whether these methods can be adapted to the case where the depth estimates arise from the application of different cost functions or algorithms.

Reference 18 uses depth estimates obtained using different sizes of support windows in a variational framework to produce a piecewise smooth depth map and a piecewise smooth approximation of the input images. In Ref. 19, the authors implicitly perform a fusion of depth values obtained on the basis of different color segmentations of the input image. They use a voting framework in which the Epanechnikov kernel[20] is used to find the mode of the different depth estimates.

If a confidence value can be attached to each depth estimate, fusion can also be achieved by using the confidence values as data terms in a global optimization framework like graph cuts[4] or belief propagation.[21] While a global optimization framework may yield good results, it is computationally expensive. We restrict our study to local methods for depth map fusion.

3 Combining Multiple Depth Estimates

We assume that the input images are rectified and the disparity range $[d_{\min}, d_{\max}]$ is known. Let $D_m = \max(|d_{\min}|, |d_{\max}|)$ be the largest absolute disparity. The images are assumed to be in RGB color space, and the color values have been

Paper 12339SSP received Aug. 28, 2012; revised manuscript received Nov. 22, 2012; accepted for publication Dec. 4, 2012; published online Jan. 7, 2013.

normalized to lie in the [0,1] range. In this paper, we use the terms disparity and depth interchangeably.

Let d^i_{rv} and d^i_{sv}, with $i = 1, \dots, m$ be m initial depth maps for the reference view and the second view, respectively. These estimates could have been obtained by using different methods. We also allow for the fact that the depth estimates may not be complete and some pixels may not have a valid depth estimate. We wish to obtain better quality depth maps $\tilde{d}_{rv}$ and $\tilde{d}_{sv}$ for the reference view and the second view.

The presence of a significant variation in depth without a variation in color between two neighboring pixels signals an error in the depth estimate since neighboring pixels of similar color are likely to belong to the same object. However, neighboring pixels may have significant variation in color, without a variation in depth since an object can have more than one color.

The presence of many such errors in a neighborhood of a pixel indicates that the depth estimate at that location is unreliable. The depth estimate usually looks very noisy in such unreliable areas. We make use of this fact in the fusion process by giving a lower weight to such noisy areas and a higher weight to smooth areas.

In the following equations, we will use Iversonian bracket notation,[22] where given a predicate **L**, [**L**] evaluates to 1 if a predicate **L** is true and 0 otherwise. Let us consider a depth estimate d^i_{rv} of the reference image R. We say that there exists a depth variation between two pixels **p** and **r** if the difference in their depth estimates is larger than a threshold δ

$$\mathcal{D}^i_{rv}(\mathbf{p},\mathbf{r}) = [|d^i_{rv}(\mathbf{p}) - d^i_{rv}(\mathbf{r})| >= \delta]. \qquad (1)$$

We say that there exists a color variation between two pixels **p** and **r** if the difference in their color values is larger than a threshold τ

$$\mathcal{C}^i_{rv}(\mathbf{p},\mathbf{r}) = [\|R(\mathbf{p}) - R(\mathbf{r})\| >= \tau]. \qquad (2)$$

Let $\mathcal{N}(\mathbf{p})$ be a set of pixels with a valid depth estimate in a neighborhood around **p**. Let n be the size of $\mathcal{N}(\mathbf{p})$. Then we define the smoothness of the depth estimate d^i_{rv} at **p** as

$$S^i_{rv}(\mathbf{p}) = -\frac{1}{n}\sum_{\mathbf{j}\in\mathcal{N}(\mathbf{p})}\frac{1}{s_{\mathbf{j}}}\sum_{\mathbf{k}\in\mathcal{M}(\mathbf{j})}\mathcal{D}^i_{rv}(\mathbf{j},\mathbf{k})(1-\mathcal{C}^i_{rv}(\mathbf{j},\mathbf{k})), \qquad (3)$$

where $\mathcal{M}(\mathbf{j})$ is the set of four immediate neighbors of pixel **j** with a valid depth estimate and $s_{\mathbf{j}}$ is the number of pixels in $\mathcal{M}(\mathbf{j})$. The smoothness $S^i_{rv}(\mathbf{p})$ is a value between 0 and 1. It represents the fraction of pixels in a neighborhood of **p** that do not exhibit a variation in depth without a variation in color.

The availability of depth estimates for the second view provides us another set of depth estimates for each pixel of the reference view, since we can project the depth of a pixel in the second view to the reference view. The aforementioned criterion for smoothness of a depth estimate applies to the depth estimate of the second view as well.

Further, if a depth estimate d^i_{rv} of the reference view at a pixel **p** is cross validated by any of the m estimates d^j_{sv} of the second view, it increases the confidence of the estimate d^i_{rv} at **p**. Let $d = d^i_{rv}(\mathbf{p})$. Let $\mathbf{q} \mapsto \mathbf{p}$ denote that a pixel **q** from the second view projects to **p** in the reference view, via the depth estimate d^j_{sv} of the second view. We can define the cross-validation of the estimate d^i_{rv} at **p** as

$$\mathcal{V}_{\mathbf{p}}(d^i_{rv}) = \begin{cases} 1 & \text{if}\left(\sum_{j=1}^{m}\sum_{\mathbf{q}\mapsto\mathbf{p}}[|d^j_{sv}(\mathbf{q}) - d| < \delta]\right) > 0 \\ 0 & \text{otherwise} \end{cases}. \qquad (4)$$

Note that more than one pixel **q** from the second view can project to **p** via a depth estimate due to errors or presence of occlusions. The inner summation in the condition of Eq. (4) takes this fact into account.

The smoothness notion and the cross-validation described above allows us to attach a reliability value for each possible depth value $d_{\min} \le d \le d_{\max}$ at each pixel **p** of the reference image in the following way

$$\mathcal{R}_{\mathbf{p}}(d) = \sum_{i=1}^{m}[d^i_{rv}(\mathbf{p}) = d]S^i_{rv}(\mathbf{p}) + \sum_{j=1}^{m}\sum_{\mathbf{q}\mapsto\mathbf{p}}[d^j_{sv}(\mathbf{q}) = d]S^j_{sv}(\mathbf{q}) + \alpha\sum_{i=1}^{m}[d^i_{rv}(\mathbf{p}) = d]\mathcal{V}_{\mathbf{p}}(d^i_{rv}), \qquad (5)$$

where α is a weighting factor.

The reliability values $\mathcal{R}_{\mathbf{p}}(d)$ for each disparity d are accumulated and smoothed. The smoothing operation we use is given by

$$\mathcal{G}_{\mathbf{p}}(d) = \sum_{i=-\sigma}^{\sigma}\frac{1}{1+|i|}\mathcal{R}_{\mathbf{p}}(d+i), \qquad (6)$$

where σ is a parameter that is used to control the extent of smoothing that is to be performed. The depth d for which $\mathcal{G}_{\mathbf{p}}(d)$ is maximized is chosen as the depth value $\tilde{d}$ for pixel **p**, provided $\mathcal{G}_{\mathbf{p}}(\tilde{d})$ is greater than a threshold κ. Otherwise, we do not assign any depth value to pixel **p**.

3.1 *Parameter Selection and Sensitivity*

In our implementation, the parameter values used are $\delta = D_m/32$, $\tau = 0.05$, $\alpha = 1$, $\kappa = m$, and $\sigma = D_m/17$. We arrive at these values from natural considerations as described below.

The parameter δ of Eq. (1) is the threshold at which two pixels can be said to have different depths. The value of this parameter can be naturally thought of as a function of the disparity range. If the range is small (say 16 disparities), a difference of 1 is a significant depth difference whereas when the disparity range is large, the threshold at which the disparity difference gets significant is higher.

The parameter τ of Eq. (2) is the threshold at which two pixels are said to have different colors. We have chosen a very conservative value of 0.05 (which is 12.75 on an 8 bit per pixel grayscale image). This will ensure that all color differences are certainly captured.

The parameter α in Eq. (5) controls the relative contribution of the smoothness term and the confirmation terms. We have chosen a natural value $\alpha = 1$ that weights these terms equally.

The value κ is a threshold below which we do not output a disparity value. We have set it as m (the number of initial estimates). The smoothed reliability values range from 0 to $3m$, and the value of κ is at one third of the range. Decreasing

Table 1 Comparison of fusion algorithms. The table shows the percentage of pixels of the covered area differing from the ground truth by more than 1. The inputs to the fusion algorithm are the results of the ELAS,[12] constant space belief propagation,[23] and horizontal and vertical scan line methods.

Fusion method	Aloe	Cones	Teddy	Venus	Reindeer	Art	Books	Moebius
Median of depths	14.78	14.18	17.43	5.23	26.49	38.80	25.60	24.17
Median of costs	14.86	14.76	16.80	4.60	25.60	23.5	25.38	24.29
Confidence fusion	06.07	06.95	09.37	02.28	16.24	15.64	15.03	15.34
Our method	03.79	04.59	07.45	01.98	12.26	16.14	13.41	11.41

κ increases the coverage at the expense of increased errors, while increasing it decreases coverage at the expense of increasing precision.

The value of parameter σ in Eq. (6) is the smoothing kernel radius. Reliability values are computed for every disparity value. They cluster around a few disparity values. Smoothing allows us to take this clustering into account and not be misled by significant but isolated reliability values. For instance consider a situation where reliability values for a pixel $\mathbf{p}$ are clustered at $\mathcal{R}_{\mathbf{p}}(10) = 1$, $\mathcal{R}_{\mathbf{p}}(11) = 5$, $\mathcal{R}_{\mathbf{p}}(12) = 6$, $\mathcal{R}_{\mathbf{p}}(13) = 2$, and $\mathcal{R}_{\mathbf{p}}(35) = 7$. Looking at reliability values alone will make us choose the wrong disparity value of 35, while clearly the disparity value $d = 12$ is a much better choice. The value of σ has been set equal to twice the δ value. For 60 disparities $\sigma = 3.5$.

We have not fine tuned the values for any particular dataset to avoid the risk of over fitting. The results are not highly sensitive to variations in the parameter values. With the

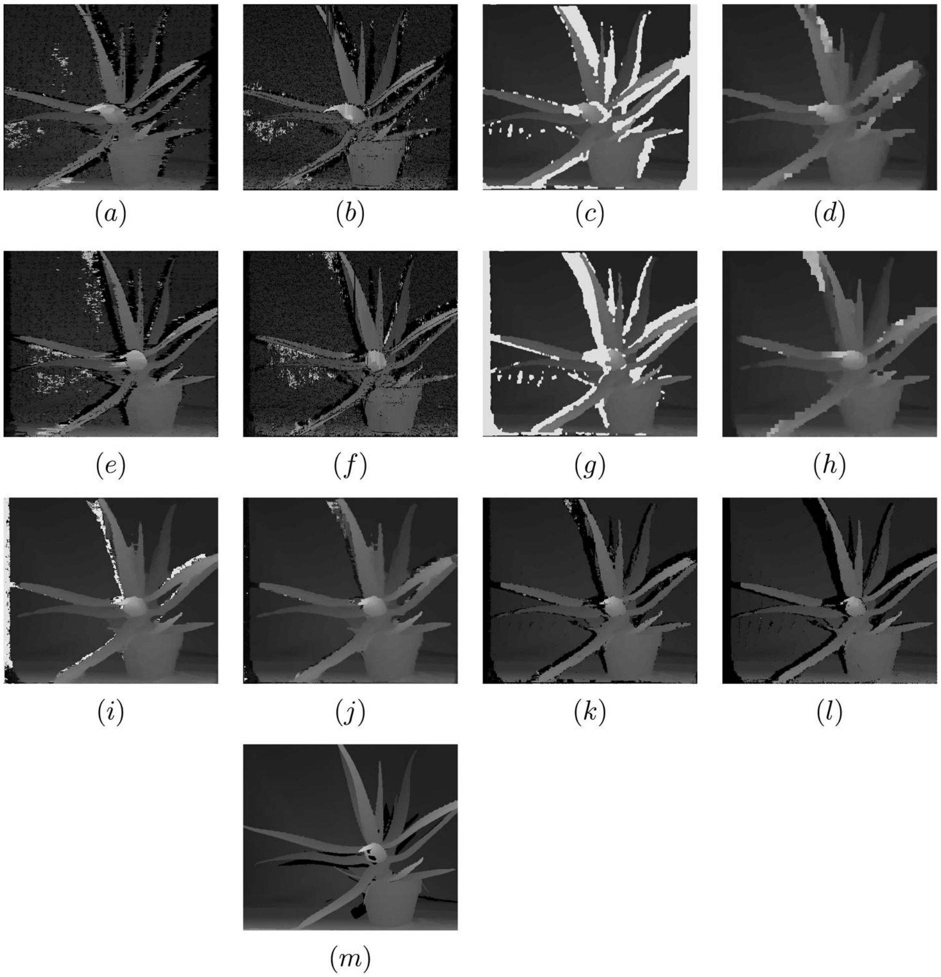


Fig. 1 A comparison of the various fusion methods on the aloe image. The images of the first row are depth estimates of the right view obtained by (a) horizontal scan line, (b) vertical scan line, (c) ELAS,[12] and (d) constant space belief propagation.[23] The second row (e) to (h) are the corresponding left view estimates. The images in the third row are the results of the fusion methods (i) median depth, (j) median of confidence, (k) the method of Ref. 17, and (l) our method. The ground truth image (m) is given in the last row.

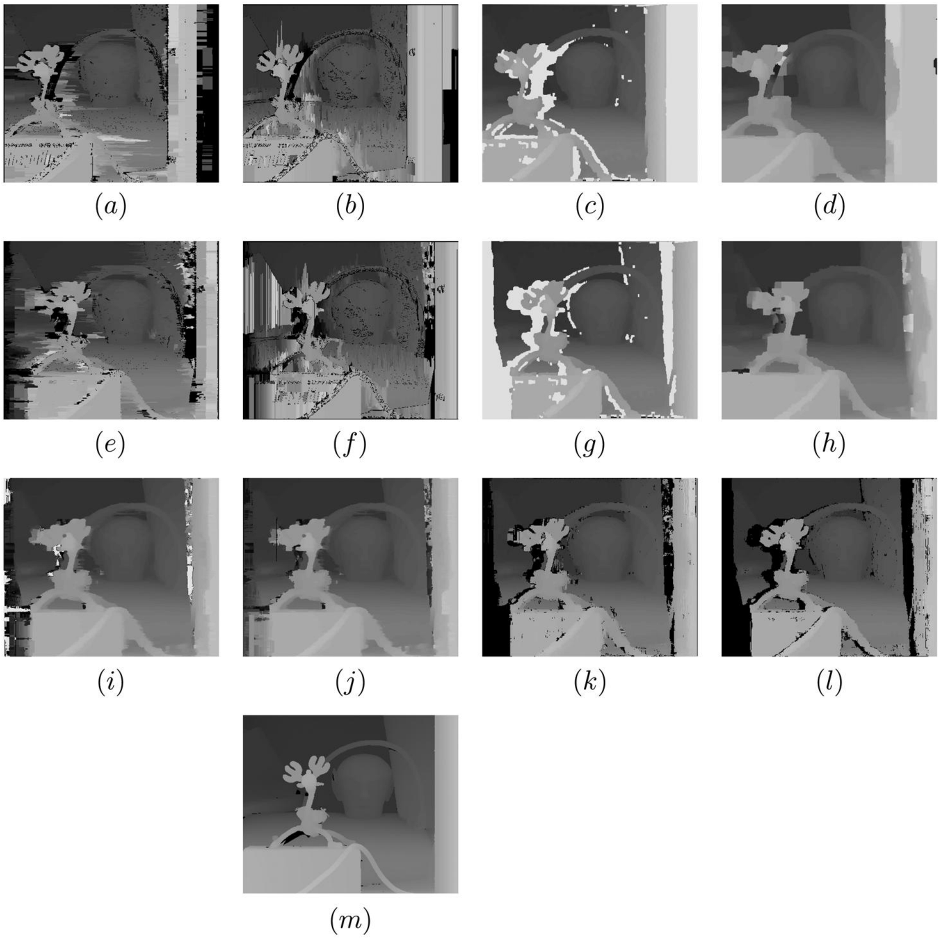


Fig. 2 A comparison of the various fusion methods on the reindeer image. The images of the first row are depth estimates of the right view obtained by (a) horizontal scan line, (b) vertical scan line, (c) ELAS,[12] and (d) constant space belief propagation.[23] The second row (e) to (h) are the corresponding left view estimates. The images in the third row are the results of the fusion methods (i) median depth, (j) median of confidence, (k) the method of Ref. 17, and (l) our method. The ground truth image (m) is given in the last row.

chosen values, our method consistently outperforms the confidence fusion method[17] described in Sec. 4 and varying the values gives slightly better performance in some images while some others fare a bit poorer.

4 Evaluation

We have studied how our method fares in combining the outputs of a few different real-time stereo algorithms, as well as its ability to fuse the outputs of different window-based, winner-take-all methods. The algorithms we used in our study are:

- The efficient stereo matching algorithm of Ref. 12.
- The constant space belief propagation algorithm of Ref. 23.
- A scan line-based stereo matching algorithm. Each scan line is segmented into areas of homogeneous color. Each segment is assigned a single disparity. Instead of the full disparity range, we first consider a restricted set of the disparities previously assigned to the neighboring segments and determine the disparity value with the smallest per pixel matching cost in the restricted set. If it is less than a threshold τ_c, we assign the value of that disparity to the entire segment. If not, we choose the disparity in the full range, that minimizes the per pixel matching cost of the entire segment. We run this algorithm independently on horizontal scan lines as well as vertical scan lines. We have used a value of $\tau_c = 0.05$.

Apart from the above algorithms, we also consider the winner-takes-all outputs of adaptive window-based methods.[24] We have used window sizes of 5×5, 7×7, and 9×9. Apart from L_2 costs, we also consider the census transform[25] using a 9×7 window size.

Our choice of the algorithms and cost functions have been motivated by the presence of fast real-time implementations of the same.[12,23,26]

We have compared our method with the following fusion methods:

- *Confidence-based fusion*. We implemented the confidence-based fusion method described in Ref. 17. We use the confidence measure $C(x)$ of a depth estimate d_0 at a pixel x as described in Eq. (1) of Ref. 17, namely

$$C(x) = \left(\sum\nolimits_{d\neq d_0} e^{-(c(x,d)-c(x,d_0))^2/\sigma^2}\right)^{-1}, \qquad (7)$$

where $c(x, d)$ is the matching cost for depth d at x. In our implementation, we normalize all costs to lie between 0 and 1. The depth estimates for the reference view as well as the second view are used in the fusion process by rendering the depth estimates of the second view on to the reference view. We have set the value of the parameters $\sigma = 120$ as given in Ref. 17. In the paper, the authors do not output a fused depth value if it is supported by fewer than C_{Thres} number of input depth maps. They use a value of $C_{\text{Thres}} = 5$ when 15 input depth maps are provided. Accordingly, we use the threshold value C_{Thres} to be $1/3$ of the total number of input depth estimates.

- *Median-value-based fusion*. At each pixel, the depth estimates of the reference view and the rendered depth estimates of the second view onto the reference view are accumulated and the median value is chosen as the fused depth value. We have also considered using the depth corresponding to the maximum value of the confidence measures of the depth estimates using Eq. (7).

5 Results

We have tested our fusion method on various images of the Middlebury dataset, as well as other real world images. It is able to significantly decrease the errors in depth estimates as compared to the inputs. We tabulate the errors in the fused depth map as a percentage of the pixels for which a disparity estimate has been output by the fusion algorithms, as well as the percentage of image area that finally gets a fused depth estimate. We also performed better than the methods we compared against. The fusion methods tabulated from top to bottom are

- *Median of depths*. The median value of all the input depth estimates at each pixel is used as the fused depth.
- *Median of costs*. The input depth estimates at each pixel are sorted and the depth with the highest confidence estimate is chosen as the fused depth.
- *Confidence fusion*. The method of Ref. 17 as explained in Sec. 4.
- *Our method*.

In Table 1, we show the error in the entire image including the occluded regions. The error is computed as the percentage of pixels of the covered area differing from the ground truth by more than 1. The inputs were the result of various real-time stereo matching methods as described in Sec. 4. Our method can be seen to perform better than the others. The input depth maps had considerable amount of error. Figures 1–4 provide a visual indication of the errors in the inputs and the results obtained using the different fusion methods. Table 2 shows the percentage of image area covered by the fusion methods.

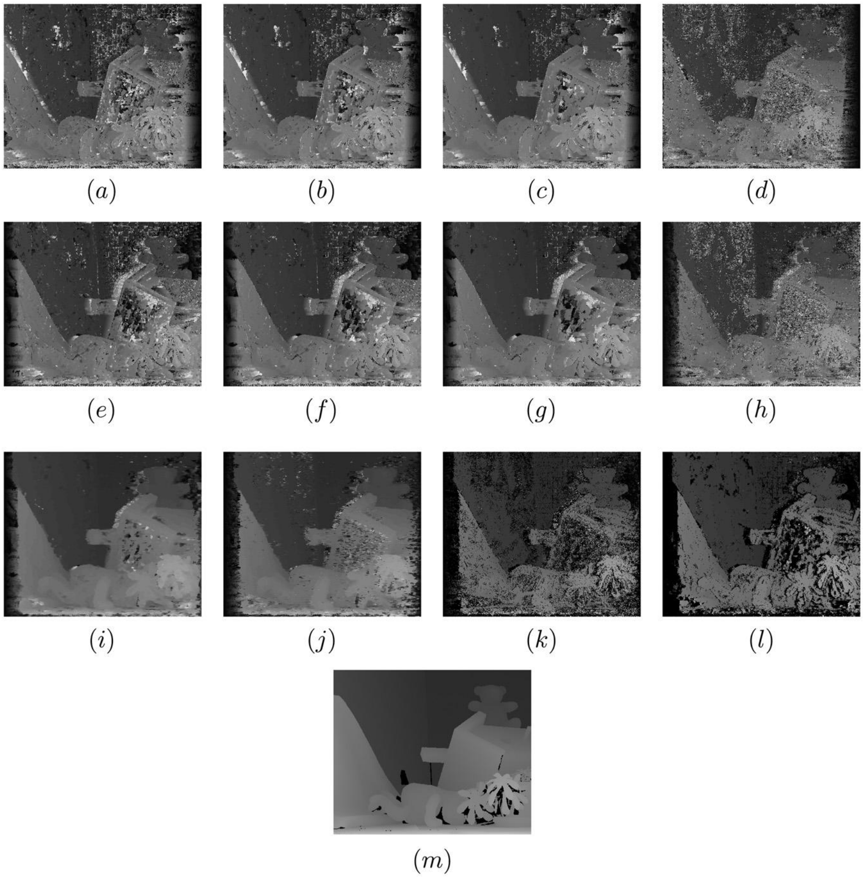


Fig. 3 A comparison of the various fusion methods on the teddy image. The images of the first row are depth estimates of the right view obtained by (a) WTA L_2 cost with 2×2 window, (b) WTA L_2 cost with 3×3 window, (c) WTA L_2 cost with 4×4 window, and (d) WTA census cost 9×7. The second row (e) to (h) are the corresponding left view estimates. The images in the third row are the left view results of the fusion methods (i) median depth, (j) median of confidence, (k) the method of Ref. 17, and (l) our method. The ground truth image (m) is given in the last row.

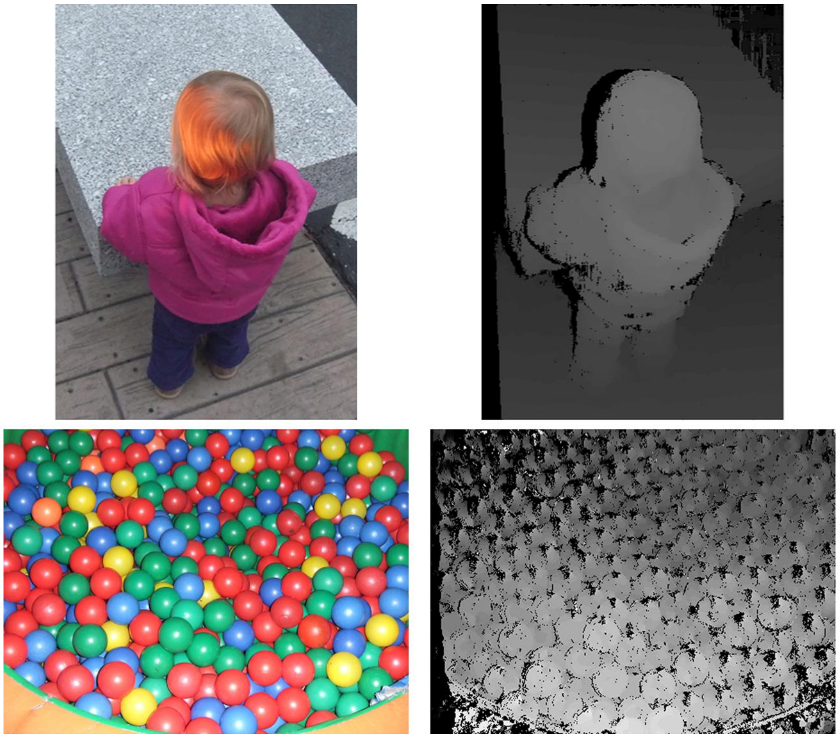


Fig. 4 Our results on some real world images. The image in the top row is from the dataset of Ref. 27.

In Tables 3 and 4, we show the error and coverage for the nonoccluded areas alone. In this case, too, our algorithm performs better than the others.

In Tables 5–8, we show the percentage error and coverage for the full image and the nonoccluded regions only. The input depth maps in this case are the results of winner-takes-all adaptive window-based cost aggregation as described in Sec. 4. The inputs are much noisier than in the previous case, yet fusion algorithms are able to extract reasonably good depth maps. Our algorithm performs much better than others even in this noisy case.

Table 2 Comparison of fusion algorithms. The table shows the percentage of pixels of the covered area that have a valid depth estimate. The inputs to the fusion algorithm are the results of the ELAS,[12] constant space belief propagation,[23] and horizontal and vertical scan line methods.

Fusion method	Aloe	Cones	Teddy	Venus	Reindeer	Art	Books	Moebius
Median of depths	99.91	99.95	99.97	99.95	99.97	99.89	99.99	99.92
Median of costs	99.95	99.90	99.96	99.97	99.94	89.47	99.92	99.98
Confidence fusion	86.63	88.20	89.44	96.45	85.70	73.11	86.35	86.28
Our method	84.31	87.01	87.22	96.42	81.85	70.14	85.52	84.20

Table 3 Comparison of fusing algorithms. The table shows the percentage of pixels of the nonoccluded covered area, differing from the ground truth by more than 1. The inputs to the fusion algorithm are the results of the ELAS,[12] constant space belief propagation,[23] and horizontal and vertical scan line methods.

Fusion method	Aloe	Cones	Teddy	Venus	Reindeer	Art	Books	Moebius
Median of depths	06.21	04.99	09.26	02.23	13.34	24.49	17.00	14.58
Median of costs	06.35	05.86	08.67	01.90	13.08	18.27	16.66	14.60
Confidence fusion	04.17	04.74	07.23	01.49	11.37	11.75	13.88	12.36
Our method	03.25	03.19	06.24	01.24	9.58	13.91	12.80	09.56

Table 4 Comparison of fusing algorithms. The table shows the percentage of pixels of the nonoccluded covered area that have a valid depth estimate. The inputs to the fusion algorithm are the results of the ELAS,[12] constant space belief propagation,[26] and horizontal and vertical scan line methods.

Fusion method	Aloe	Cones	Teddy	Venus	Reindeer	Art	Books	Moebius
Median of depths	84.53	85.24	87.48	95.88	82.01	76.41	88.35	86.34
Median of costs	84.53	85.24	87.47	95.88	82.00	76.26	88.35	86.32
Confidence fusion	81.83	82.49	84.71	94.75	79.82	69.11	84.62	82.53
Our method	81.28	82.75	83.88	95.06	78.78	67.82	84.39	81.85

Table 5 Comparison of fusing algorithms. The table shows the percentage of pixels of the covered area differing from the ground truth by more than 1. The inputs to the fusion algorithm are the results of the winner-takes-all adaptive window methods.

Fusion method	Aloe	Cones	Teddy	Venus	Reindeer	Art	Books	Moebius
Median of depths	17.26	22.63	25.69	18.95	31.59	42.19	37.93	30.04
Median of costs	18.36	23.92	30.36	31.45	34.40	39.54	37.02	28.16
Confidence fusion	11.69	17.42	24.73	19.07	21.20	25.96	28.65	21.20
Our method	05.08	10.55	12.70	10.08	15.14	21.38	22.55	13.78

Table 6 Comparison of fusing algorithms. The table shows the percentage of pixels of the image area having a valid depth estimate. The inputs to the fusion algorithm are the results of the winner-takes-all adaptive window methods.

Fusion method	Aloe	Cones	Teddy	Venus	Reindeer	Art	Books	Moebius
Median of depths	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Median costs	99.41	99.57	99.48	99.81	99.23	99.64	99.57	99.80
Confidence fusion	85.19	74.53	71.85	71.43	71.34	69.83	68.34	77.33
Our method	81.30	73.24	71.37	79.98	71.69	67.93	67.62	72.74

Table 7 Comparison of fusing algorithms. The table shows the percentage of pixels of the nonoccluded covered area differing from the ground truth by more than 1. The inputs to the fusion algorithm are the results of the winner-takes-all adaptive window methods.

Fusion method	Aloe	Cones	Teddy	Venus	Reindeer	Art	Books	Moebius
Median of depths	08.00	14.39	18.17	16.27	18.81	27.85	30.79	21.00
Median of costs	08.38	15.46	23.31	29.12	21.79	22.92	29.52	18.22
Confidence fusion	06.34	12.55	20.44	17.37	13.39	16.00	24.21	14.50
Our method	04.27	09.16	11.63	09.34	12.97	18.14	21.69	12.00

Table 8 Comparison of fusing algorithms. The table shows the percentage of pixels of the nonoccluded image area having a valid depth estimate. The inputs to the fusion algorithm are the results of the winner-takes-all adaptive window methods.

Fusion method	Aloe	Cones	Teddy	Venus	Reindeer	Art	Books	Moebius
Median of depths	84.53	85.24	87.48	95.88	82.01	76.41	88.35	86.34
Median of costs	84.20	85.01	87.29	95.79	81.82	76.21	88.18	86.22
Confidence fusion	77.84	67.60	66.03	69.48	64.37	61.60	63.89	70.69
Our method	78.37	69.62	68.84	78.81	69.44	64.88	66.46	70.76

6 Discussion

We have described a method for fusing several depth maps in order to output a depth map of higher quality than the inputs. We have demonstrated that our method performs robustly on several standard stereo pairs as well as unconstrained real world images. The algorithm is simple to implement and can be used along with real-time methods to significantly improve the quality of the depth maps. Simple post processing techniques like segmentation followed by plane fitting can lead to better coverage and subpixel refinement.

The success of our proposed method indicates that further study in this direction may be profitable.

References

1. D. Marr and T. Poggio, “Cooperative computation of stereo disparity,” *Science* **194**(4262), 283–287 (1976).
2. D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *IJCV* **47**(1–3), 7–42 (2002).
3. Z.-F. Wang and Z.-G. Zheng, “A region based stereo matching algorithm using cooperative optimization,” in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE Computer Society, Anchorage, AK (2008).
4. Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(11), 1222–1239 (2001).
5. V. Kolmogorov and R. Zabih, “Computing visual correspondence with occlusions via graph cuts,” Technical Report, Ithaca, NY (2001).
6. V. Kolmogorov and R. Zabih, “Multi-camera scene reconstruction via graph cuts,” in *Proc. 7th European Conf. on Computer Vision-Part III*, pp. 82–96, Springer-Verlag, London, UK (2002).
7. A. Klaus, M. Sormann, and K. Karner, “Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure,” in *Proc. 18th International Conf. on Pattern Recognition—Volume 03*, pp. 15–18, IEEE Computer Society, Washington, DC (2006).
8. A. Hosni et al., “Local stereo matching using geodesic support weights,” in *16th IEEE International Conf. on Image Processing*, pp. 2093–2096, IEEE Press, Cairo, Egypt (2009).
9. S. Mattoccia, F. Tombari, and L. Di Stefano, “Stereo vision enabling precise border localization within a scanline optimization framework,” in *Proc. 8th Asian Conf. on Computer Vision—Volume Part II*, pp. 517–527, Springer-Verlag, Berlin, Heidelberg (2007).
10. D. Mukherjee, G. Wang, and Q. Wu, “Stereo matching algorithm based on curvelet decomposition and modified support weights,” in *IEEE International Conf. on Acoustics Speech and Signal Processing*, pp. 758–761, IEEE Press, Dallas, TX (2010).
11. H. Hirschmuller, “Stereo vision in structured environments by consistent semi-global matching,” in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 401–406, IEEE Press (2006).
12. A. Geiger, M. Roser, and R. Urtasun, “Efficient large-scale stereo matching,” in *Asian Conf. on Computer Vision*, Queenstown, New Zealand, pp. 25–38 (2010).
13. S. Kosov, T. Thormählen, and H.-P. Seidel, “Accurate real-time disparity estimation with variational methods,” in *Proc. 5th International Symposium on Advances in Visual Computing: Part I*, pp. 796–807, Springer-Verlag, Berlin, Heidelberg (2009).
14. S. Birchfield and C. Tomasi, “A pixel dissimilarity measure that is insensitive to image sampling,” *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(4), 401–406 (1998).
15. P. Mordohai, “The self-aware matching measure for stereo,” in *ICCV*, pp. 1841–1848, IEEE Press, Kyoto (2009).
16. B. Curless and M. Levoy, “A volumetric method for building complex models from range images,” in *Proc. 23rd Annual Conf. on Computer Graphics and Interactive Techniques*, pp. 303–312, ACM, New York, NY (1996).
17. P. Merrell et al., “Real-time visibility-based fusion of depth maps,” in *International Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE Press, Rio de Janeiro (2007).
18. T. Pock, C. Zach, and H. Bischof, “Mumford-shah meets stereo: Integration of weak depth hypotheses,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE Press, Minneapolis, MN (2007).
19. G. Gales, A. Crouzil, and S. Chambon, “A region-based randomized voting scheme for stereo matching,” in *Proc. 6th International Conf. on Advances in Visual Computing—Volume Part II*, pp. 182–191, Springer-Verlag, Berlin, Heidelberg (2010).
20. H. Chen and P. Meer, “Robust computer vision through kernel density estimation,” in *Proc. of the 7th European Conf. on Computer Vision—Part I*, pp. 236–250, Springer-Verlag, London, UK (2002).
21. J. Sun, N.-N. Zheng, and H.-Y. Shum, “Stereo matching using belief propagation,” *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(7), 787–800 (2003).
22. D. E. Knuth, “Two notes on notation,” *Am. Math. Monthly* **99**(5), 403–422 (1992).
23. Q. Yang, L. Wang, and N. Ahuja, “A constant-space belief propagation algorithm for stereo matching,” in *CVPR*, pp. 1458–1465, IEEE Press, San Francisco, CA (2010).
24. T. Kanade and M. Okutomi, “A stereo matching algorithm with an adaptive window: theory and experiment,” *IEEE Trans. Pattern Anal. Mach. Intell.* **16**(9), 920–932 (1994).
25. R. Zabih and J. Woodfill, “Non-parametric local transforms for computing visual correspondence,” in *Proc. Third European Conf.—Volume II on Computer Vision*, pp. 151–158, Springer-Verlag, London, UK (1994).
26. M. Humenberger et al., “A fast stereo matching algorithm suitable for embedded real-time systems,” *Comput. Vis. Image Understand.* **114**(11), 1180–1202 (2010).
27. B. Price and S. Cohen, “Stereocut: consistent interactive object selection in stereo image pairs,” in *2011 IEEE International Conf. on Computer Vision*, pp. 1148–1155, IEEE Press, Barcelona (2011).

Biographies and photographs of the authors are not available.

Image deblurring in smartphone devices using built-in inertial measurement sensors

Ondřej Šindelář
Charles University in Prague
Faculty of Mathematics and Physics
Prague, Czech Republic
Email: ondrasindelar@gmail.com

Filip Šroubek
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 4, CZ-182 08, Praha 8, Czech Republic

Abstract. *Long-exposure handheld photography is degraded with blur, which is difficult to remove without prior information about the camera motion. In this work, we utilize inertial sensors (accelerometers and gyroscopes) in modern smartphones to detect exact motion trajectory of the smartphone camera during exposure and remove blur from the resulting photography based on the recorded motion data. The whole system is implemented on the Android platform and embedded in the smartphone device, resulting in a close-to-real-time deblurring algorithm. The performance of the proposed system is demonstrated in real-life scenarios. © 2013 SPIE and IS&T [DOI: 10.1117/1.JEI.22.1.011003]*

1 Introduction

Blur induced by camera motion is a frequent problem in photography mainly when the light conditions are poor. As the exposure time increases, involuntary camera motion has a growing effect on the acquired image. Image stabilization (IS) devices that help to reduce the motion blur by moving the camera sensor in the opposite direction are becoming more common. However, such hardware remedy has its limitations, as it can compensate only for motion of a very small extent and speed. Deblurring the image offline using mathematical algorithms is usually the only choice we have in order to obtain a sharp image. Motion blur can be modeled by convolution, and the deblurring process is called deconvolution, which is a well-known ill-posed problem. In general, the situation is even more complicated, since we usually have no or limited information about the blur shape.

We can divide the deconvolution methods into two categories: methods that estimate the blur and the sharp image directly from the acquired image (blind deconvolution) and methods that use information from other sensors to estimate the blur (semi-blind deconvolution).

Over the last few years, blind deconvolution has experienced a renaissance. The key idea of new algorithms belonging to the first category is to address the ill-posedness of blind deconvolution by characterizing the image prior to using natural image statistics and by a better choice of estimators. A frantic activity started with the work of Fergus et al.,[1] who applied variational Bayes to approximate the posterior by a simpler distribution. Other authors[2,3,4,5] stick to the "good old" alternating maximum a posteriori estimation approach, but by using ad hoc steps, which often lack rigorous explanation, they converge to a correct solution. Levin et al. in Refs. 6 and 7 proved that a proper estimator matters more than the shape of priors. They showed that marginalizing the posterior with respect to the latent image leads to the correct solution of the blur. The marginalized probability can be expressed in a closed form only for simple priors that are, e.g., Gaussian. Otherwise approximation methods such as variational Bayes[8] or the Laplace approximation[9] must be used. Complex camera motion often results in blur that is space-variant, i.e., the blur is a function of a position vector. As a rule, the space-variant blur cannot be expressed by an explicit formula, but in many cases it has a special structure that can be exploited. If only one type of camera motion is considered (e.g., rotation), we can express the degradation operator as a linear combination of basis blurs (or images) and solve the blind problem in the space of the basis, which has much lower dimension than the original problem. Whyte et al.[10] considered rotations about three axes up to several degrees and described blurring using three basis vectors. For blind deconvolution, they used an algorithm analogous to Ref. 1 based on marginalization over the latent sharp image. Gupta et al.[11] adopted a similar approach, replacing rotations about x and y axes by translations. State-of-the-art blind-deconvolution algorithms achieve sometimes awesome results. However, their main limitation is that they work only in specific situations, they are prone to local extrema, and they are computationally very demanding.

The second category of deconvolution algorithms (semi-blind) tries to overcome these drawbacks by using information about the camera motion from other sources. One possibility is to acquire a pair of images: one correctly

Paper 12326SSP received Aug. 22, 2012; revised manuscript received Dec. 20, 2012; accepted for publication Dec. 28, 2012; published online Feb. 4, 2013.

0091-3286/2013/$25.00 © 2013 SPIE and IS&T

exposed but blurred and one underexposed (noisy) but sharp image. Then we can apply multichannel blind deconvolution methods, which are better posed, as was proposed for example in Refs. 12, 13, and 14. Another possibility is to attach an auxiliary high-speed camera of lower resolution to estimate the point-spread function (PSF) using for example optical flow techniques.[15,16] Many devices, such as modern smartphones, are now equipped with inertial sensors (gyroscopes and accelerometers) that can give us accurate information about camera motion. If we are able to reconstruct camera path, then we can recover blur and perform nonblind image deblurring. This idea was recently described by Joshi et al., in Ref. 17, but they have designed an expensive measuring apparatus consisting of a digital single-lens reflex camera and a set of inertial sensors and perform image deblurring offline on a computer. This work is based on the same idea, but our aim is to show that image deblurring is feasible on modern smartphones and not requiring any other devices.

The main contribution of this work is to illustrate that blur estimation with built-in inertial sensors is possible and to implement image deblurring on a smartphone, which works in practical situations and is relatively fast to be acceptable for a general user. The next section shows the relation between the camera pose and the image blur, and discusses simplifications that we make. Section 3 briefly describes implementation on our test device (Samsung smartphone). Section 4 shows results of our experiments and addresses pitfalls that are common for cameras embedded in smartphones.

2 Camera Motion Blur Analysis

We start the discussion with a general camera motion. Since our primary goal is a handy implementation for mobile devices, we then introduce simplification of the problem that allows a fast and memory-conserving solution with promising results.

2.1 *Model*

The image degradation model is represented by relation

$$g = H(u) + n, \tag{1}$$

where H is a linear degradation operator and n is additive noise. Image coordinate indices are omitted here for simplicity. Our goal is to find an estimate of the original image u from the observed blurred image g.

To track the effect of camera motion on the output image, we first assume a standard perspective projection $\Pi:\mathbb{R}^3 \rightarrow \mathbb{R}^2$ that transforms a three-dimensional (3-D) point $[x, y, z]$ in the observed scene to a two-dimensional (2-D) location $[x', y']$ in the image plane:

$$\Pi([x, y, z]^T) = \left[\frac{xf}{z}, \frac{yf}{z}\right]^T. \tag{2}$$

For the sake of brevity, we assume here only the focal length f in the intrinsic camera matrix. The optical axis is identical with the z axis. During camera motion, projection of a point $p = [x, y, z]^T$ at time τ within the exposure period is given by

$$C(\tau) = \Pi\left(R(\tau)\begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} t_x(\tau) \\ t_y(\tau) \\ t_z(\tau) \end{bmatrix}\right) = \Pi[R(\tau)p + t(\tau)], \tag{3}$$

where R and t are the 3-D rotation matrix and translation vector, respectively, that define the camera pose at time τ. The rotation matrix $R(\tau)$ is given by three rotation angles $\phi_x(\tau)$, $\phi_y(\tau)$ and $\phi_z(\tau)$.

The resulting curve C makes up a trajectory of a trace that is left on the sensor by a point light source. Assuming a constant illuminance over the exposure period, the light energy emitted from the point is distributed evenly (with respect to time) over the curve C. This effectively gives us a time parametrization of a PSF for a given point, which forms the blur operator H. The operator H can be written in a form naturally generalizing standard convolution as

$$H(u)[x, y] = \int u(x - s, y - t)\tilde{h}(s, t, x - s, y - t)\mathrm{d}s\mathrm{d}t, \tag{4}$$

where $\tilde{h}$ depends on the position (third and fourth variable) and can be regarded as a space-variant PSF.

Now we can draw the relation between $\tilde{h}$ in Eq. (4) and the curve C. For any given 3-D point at position p rendered on the image plane to $[x', y'] = \Pi(p)$ the point-spread blur function $\tilde{h}(s, t, x', y')$ is a 2-D function of $[s, t]$, which can be interpreted as a blurred image of an ideal light point displayed at $[x', y']$. It can be thus obtained by rendering the curve C on a plane with the total integral of $\tilde{h}$ (which has to be equal to 1 to conserve distribution of energy) distributed along the path evenly in respect to the time parameter.

In the next section, we will show how to simplify this model and assume the space-invariant case, i.e., $\tilde{h}(s, t, x, y) = h(s, t)$.

2.2 *Space-Invariant Simplification*

We will consider a situation when the operator H is spatially invariant, so Eq. (1) becomes

$$g = h{*}u + n, \tag{5}$$

where "$*$" denotes convolution and h is a space-invariant PSF.

The PSF Eq. (3) is spatially variant in general, so it will be modified for our purposes. First of all, the translation t affects the projection differently depending on the object distance from the camera. The relation is inversely proportional, as shown in Fig. 1(a). In the case of our test device, if the camera shifts by 1 mm, objects at distance of 2 m or more move by less than 1 pixel in the image. We can thus effectively ignore translation as a cause of blur in many practical situations.

Rotation about the optical z axis (yaw) intuitively interferes with the space-invariant blur assumption. This type of rotation applied on a point light source placed in the center of the picture (on the optical axis) leaves the projection unchanged, but points outside the center form arc-shaped traces that grow toward the image borders. Provided that the camera is rotated with an equal amount around all three axes, which is a fair assumption under normal circumstances, a yaw has the least effect on the resulting blur, especially in the center of the sensor. Cellphone cameras typically

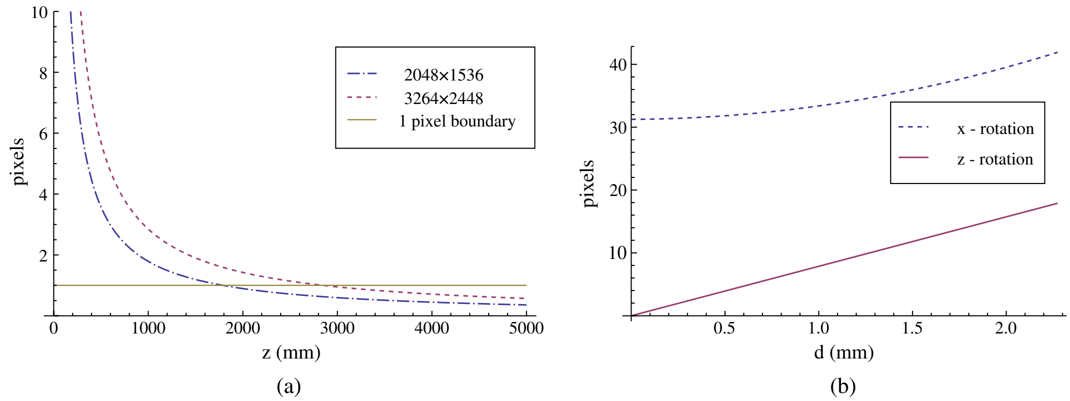


Fig. 1 Dependence of projection shift on translation and z rotation for a test device. (a) Influence of 1 mm x or y translation depending on object distance. Angle of view is 60 deg; two curves represent different image sensor resolution; (b) influence of 1 deg rotation about x and z axis depending on a distance d from the image sensor center. The full sensor extent corresponds to $d = 2.3$ mm; image resolution is 2048×1536.

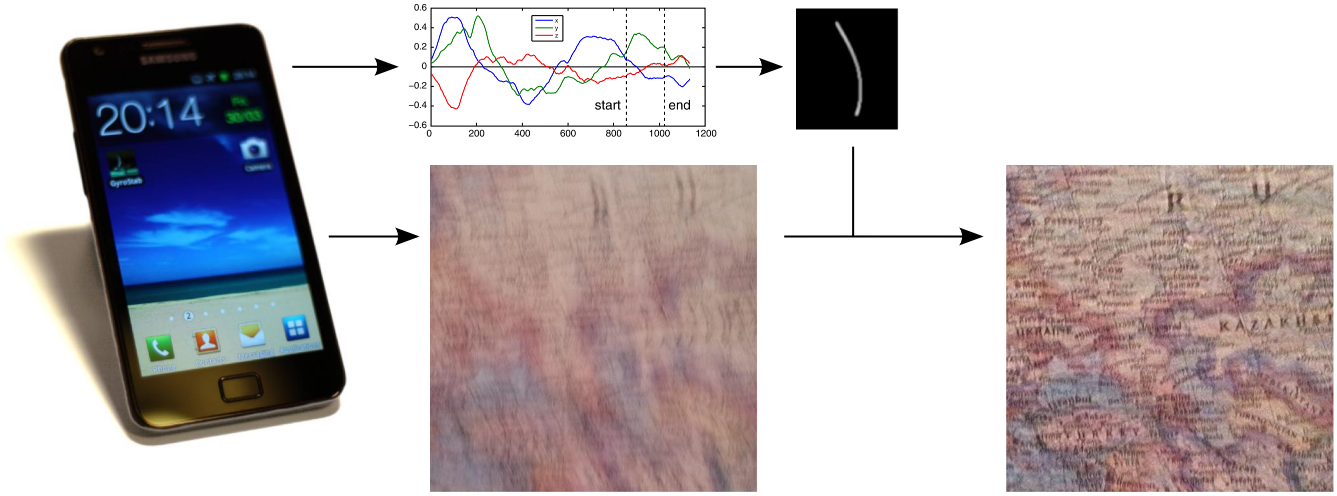


Fig. 2 Basic application workflow. Together with a taken photograph gyroscope data are recorded, which is a base for blur kernel estimation. A deconvolution is then performed to remove blur from the image.

have the focal length close to the sensor size, which means that only close to the image borders the blur size produced by yaw is approaching the blur size produced by rotation about x or y; see Fig. 1(b).

The last obstacle towards the space-invariant PSF is the perspective projection itself. Length of a trace caused by x and y rotations are projected slightly differently depending on the distance from the optical center, because the rectilinear projection in Eq. (2) casts a point at an angle α from the optical axis to a point at a distance of $f \cdot \tan(\alpha)$ from the image center. The tangent function is close to linear for small angels, so both x and y rotations by a small angle α shift a point in the sensor center approximately $f \cdot \alpha$ away in the direction of the given axis. Using the same rule for all points on the sensor gives us the space-invariant simplification of Eq. (3):

$$C(\tau) \approx \begin{bmatrix} x' \\ y' \end{bmatrix} + f \begin{bmatrix} \phi_x(\tau) \\ \phi_y(\tau) \end{bmatrix}, \qquad (6)$$

where $[x', y']$ is the location of a point in the image. This approximation holds if z is large, and $x'\phi_x \ll f$ and $y'\phi_y \ll f$, which is true at least in the central part of the image.

3 Implementation

As a testing platform, we have chosen a *Samsung Galaxy S II* smartphone with Android OS. It is equipped with all the apparatus needed for our experiments; namely a relatively high-quality camera, motion sensors, a fast CPU, and enough RAM to perform computations.

3.1 *PSF Estimation*

During the photo acquisition, samples of angular velocity are recorded using the embedded gyroscopes, which are afterward trimmed to fit the exposure period. An estimation of the PSF is rendered by integrating the curve position from the recorded data using Eq. (6).

3.2 *Deconvolution*

State-of-the-art nonblind deconvolution methods use sparse image priors, and the solution is usually found by some iterative minimization algorithms, such as in Ref. 4. However, the limited computational power of the smartphone prevents us from implementing these sophisticated deconvolution methods. We thus use a simple but fast Wiener filter in the form

$$\hat{U} = G \frac{H^*}{|H|^2 + \Phi}, \tag{7}$$

where Φ is an estimation of the inverse signal-to-noise ratio, and G, H, and $\hat{U}$ are discrete Fourier transforms of the observed image g, PSF h, and the estimated latent image $\hat{u}$, respectively.

Filtering in the frequency domain treats the image as a periodic function, which causes ringing artifacts around image borders. To overcome this problem, several less or more sophisticated techniques were proposed in the literature.[18,19] We have found sufficient to preprocess the input image g by blending the opposite image borders at the width of the PSF, which creates a smooth transition and eliminates the artifacts.

The intensity values of the output image $\hat{u}$ sometimes lie outside the 8-bit range (0 to 255), therefore we added optional normalization with clipping of outliers. The normalization is especially useful in the case of larger blurs and scene with high illumination.

For conversions of the images to frequency domain and back, we use fast Fourier transform (FFT) algorithm implemented in the fastest Fourier transform in the West (FFTW) library. Utilizing a fast ARM Cortex-A9 CPU with two cores and support for a single instruction, multiple data instruction set (NEON), FFTW proved to be remarkably fast on the tested smartphone (see Table 1).

The acquired images with native camera resolution of 3264 × 2448 is by default scaled down to 2048 × 1536 to take the advantage of better performance of FFTW when the image size is a factor of small primes. Image downsampling has a negligible effect on the image quality, because native camera resolution is unnecessarily high. The optical system of the camera has a very small aperture, which, because of diffraction and optical aberrations, limits the number of pixels that can be effectively captured by the image sensor.

To perform Wiener filtering, FFT must be applied several times: once for the PSF and twice (forward and backward-inverse) for each color channel. That yields a total of seven FFT operations. With some overhead of bitmap transfers, the deconvolution phase for the image resolution 2048 × 1536 takes about 2.6 s. The whole process starting from the camera shutter is done in a little over 6 s. This includes image resizing, PSF estimation, compressing, and saving the original and deblurred image files. The main application workflow is summarized on a schematic diagram in Fig. 2.

Table 1 Speed (in milliseconds) of FFT transform of grayscale images with different sizes and different CPU settings.

Resolution	No NEON, No hardware FPU	NEON, 1 core	NEON, 2 cores
1536 × 1152	2900	185	110
2048 × 1536	5300	330	195
2050 × 1538	—	1000	540
3264 × 2448	21200	1450	800

4 Results

In this section we display several of our results together with estimated PSFs; see Figs. 3, 4, and 5. All results were computed with the signal-to-noise parameter Φ set to 0.01. This value was determined experimentally to provide the best looking results. The original intention was to set Φ proportionally to the film speed (ISO value) extracted from the exchangeable image file format data of a photo, which should determine the amount of noise present in the image. However, we found the dependency of Φ on ISO very negligible. We explain this behavior by the denoising step that the mobile phone internally performs on the captured photos.

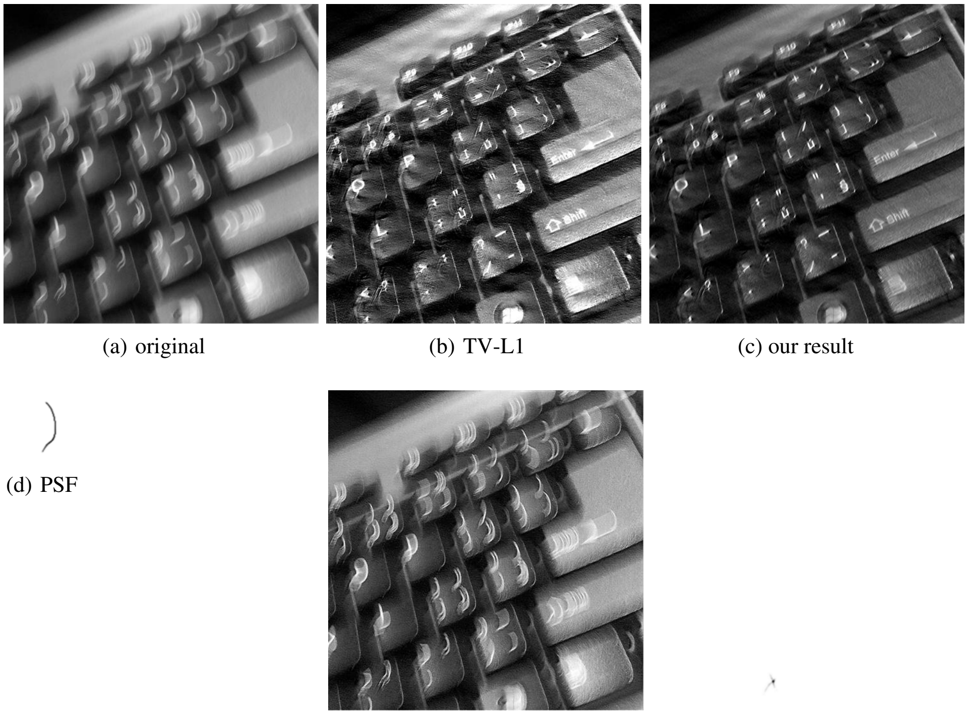
For comparison, we show an advanced nonblind iterative method (TV-L1) by Xu and Jia (Ref. 5)*, which minimizes image total variation and data term in the L_1-norm. We also tested blind deconvolution proposed in the same, which is probably currently the best blind deconvolution method. However, the result of the first test image shown in Fig. 3(e) illustrates a total failure of this method when applied to images taken by our test device. The PSF [Fig. 3(f)] estimated by the blind deconvolution method is close to a delta function and the estimated image [Fig. 3(e)] is thus a slightly sharpened image. We suspect that small PSF variations in space and/or the image postprocessing done by the smartphone prevents a successful estimation of the correct motion blur. The same unsatisfactory behavior was observed in all our tests. However, our results in Figs 3(c), 4c, and 5c illustrate that in spite of a relatively simple approach, which incorporates the Wiener filter with the space-invariant PSF estimated by inertial sensors, the proposed method is capable of producing convincing images exposing many details that were hidden in the original. The nonblind algorithm of Xu and Jia, which is using the same PSF estimated by inertial sensors, tends to amplify the signal, which rather emphasizes noise and false edges than gains signal improvement. Conversely, high-frequency details are more suppressed, probably due to being treated as noise, despite of careful attempts to tune the parameters of the method. Within our testing environment, the simplified Wiener filter is more advantageous as it filters all frequencies evenly which apparently matches the spectrum characteristics of most of the tested images.

Our results seem to lack contrast, which is largely because of the normalization. On the other hand, it helps retaining the full dynamic range without saturation as clearly seen in the comparison Fig. 3.

Our deconvolution process admittedly has downsides, as well. Focusing in a dark environment may be unsuccessful, and then the deconvolved result cannot be sharp even if the PSF estimation is correct, since we lack any means to estimate the out-of-focus blur.

The subjective quality of the deconvolution output is not entirely consistent. Images presented in this section are the best-looking results. Outputs of the similar quality are frequently achieved by our method, but sometimes the result is impaired by visual anomalies worsening its appearance. Most often it is manifested as ringing artifacts surrounding sharp edges in the picture, as demonstrated in Fig. 6.

*An executable is available for download at ~http://appsrv.cse.cuhk.edu.hk/~xuli/deconv.zip



(a) original

(b) TV-L1

(c) our result

(d) PSF

(e) Xu, Jia — blind

(f) PSF est. – Xu, Jia

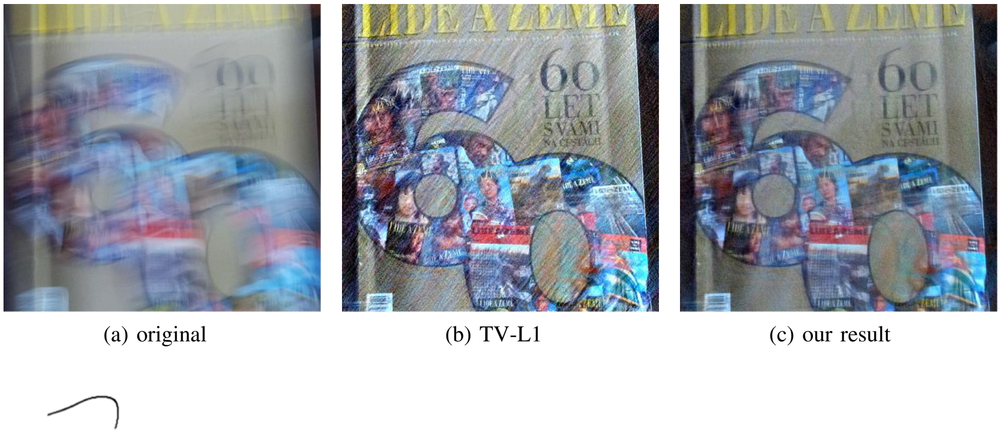
Fig. 3 Test 1: 1/7 s exposure, 16×59 estimated PSF.

The lack of control over camera hardware in the phone (no manual exposure settings, no access to raw data from the image sensor) and inaccurate timing of exposure events prevents us to systematically evaluate our method and find sources of malfunctioning.

The main problem is most likely the space-variant nature of the PSF as discussed in Sec. 2, which is particularly noticeable when a rotation about the z axis is significant or a translation movement is present and the scene depth is small. The example in Fig. 6 is influenced by a combination of both of these factors. The space-invariant approximation of camera projection is often apparent in parts close to image borders, because of a relatively wide camera field of view (60 deg).

However, another cause is the shutter mechanism. Contrary to systems with a mechanical shutter, values of illuminated pixels are here read successively line by line. The readout from the CMOS sensor takes several tens of milliseconds as shown in Fig. 7, which results in a picture not taken at a single moment, but with a slight time delay between the first and last pixel row. This process, called rolling shutter, is therefore another cause of the blur variance as the PSF depends on the vertical position in the image. The correct approach to PSF estimation is thus shifting inertial sensor data in time according to the vertical position in the image.

The application programming interface (API) of the tested device does not allow accurate synchronization between camera and gyroscope samples. Therefore we have



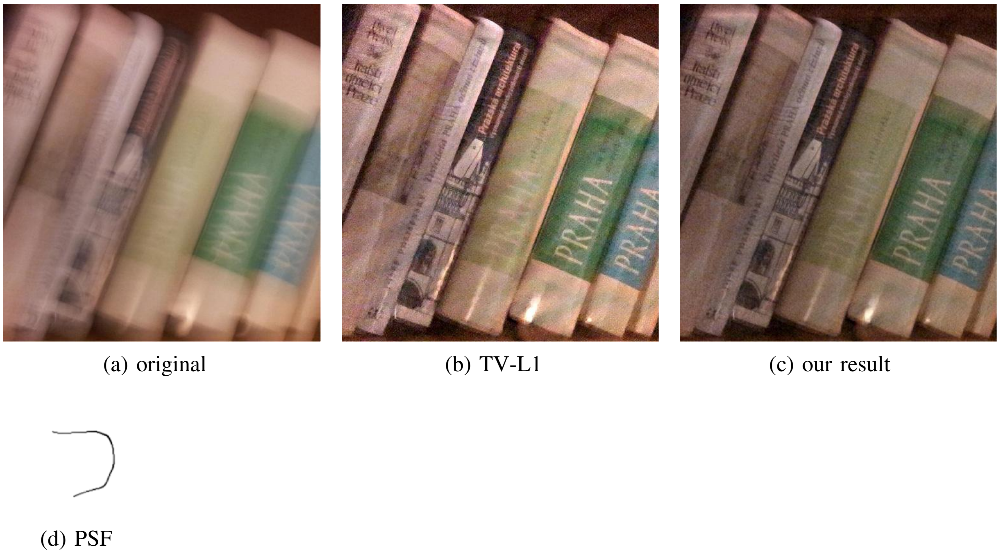
(a) original

(b) TV-L1

(c) our result

(d) PSF

Fig. 4 Test 2: 1/9 s exposure, 21×28 estimated PSF.


(a) original (b) TV-L1 (c) our result

(d) PSF

Fig. 5 Test 3: 1/2 s exposure, 72 × 76 estimated PSF.

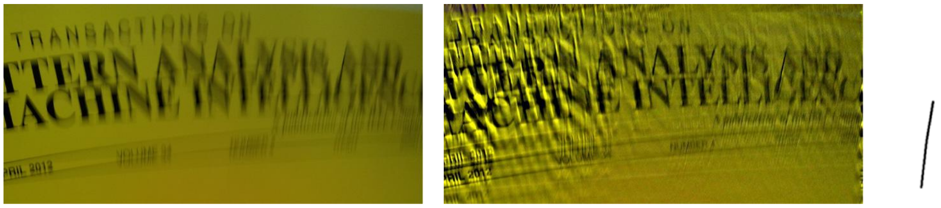
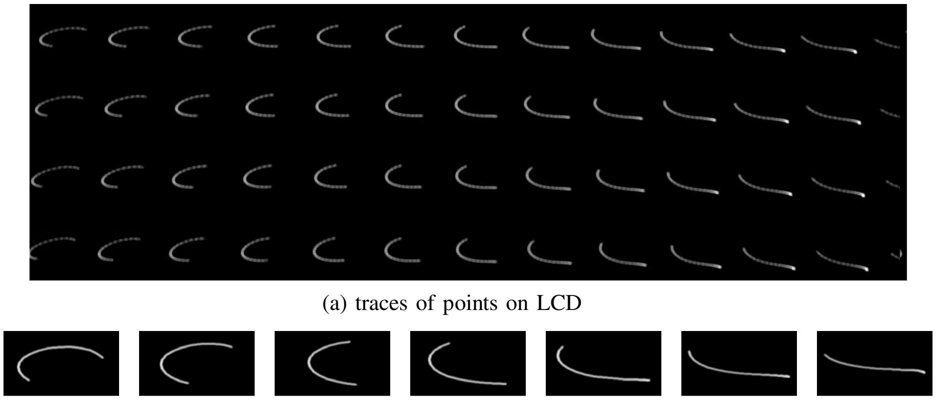

(a) original (b) result (c) PSF

Fig. 6 An example of an unsatisfactory result.


(a) traces of points on LCD

(b) 40 ms (c) 50 ms (d) 60 ms (e) 70 ms (f) 80 ms (g) 90 ms (h) 100 ms

Fig. 7 A snapshot of point grid displayed on a liquid crystal display screen showing the rolling shutter effect. The bottom row shows a series of blur kernels rendered using data from the gyroscope sensor shifted in time. Exposure 1/14 s, PSF images were created from sensor data starting 40 to 100 ms after a synchronization timestamp.

implemented a deconvolution preview, where the user picks the best option from a set of results created with time-shifted PSFs. The preview also partly solves the rolling shutter problem, since the selected time shift corresponds to a horizontal image band of a certain height that can be considered as acquired at one moment, thus eliminating the rolling shutter effect for that image part.

Image post-processing might also present a serious problem for the deconvolution. Since the original raw data from the image sensor are not available, we are forced to work with the JPEG-compressed image, which is most likely processed by a denoising, contrast-enhancement algorithm, or lens-distortion compensation. These adjustments are undesirable for our purposes, as they were not taken into account in our model.

Noise present in gyroscope measurement data can also be a problem, as displayed in Figs. 8 and 9. This has been examined in a following synthetic experiment. A test image was

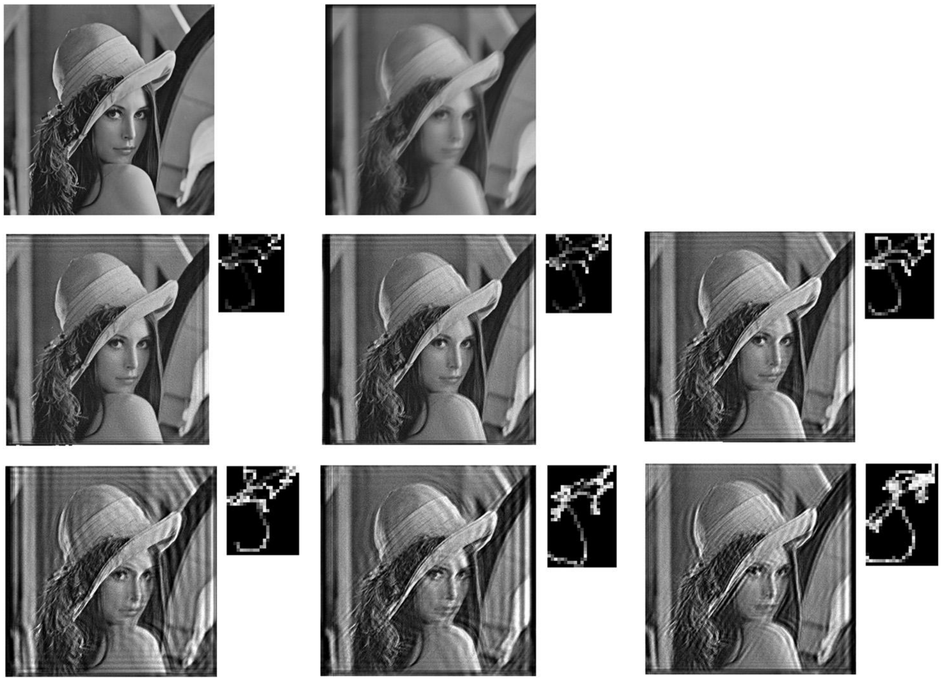


Fig. 8 Noise in gyroscope data. Synthetically blurred Lena image using PSF from recorded gyroscope samples and afterward deblurred using PSF from measurements with variable amount of noise. Images are from left to right, top to bottom: original, blurred, and six deblurred images using original gyroscope data altered by random Gaussian noise with variance from 0 to 0.05 (gyroscope measurements are in rad/s).

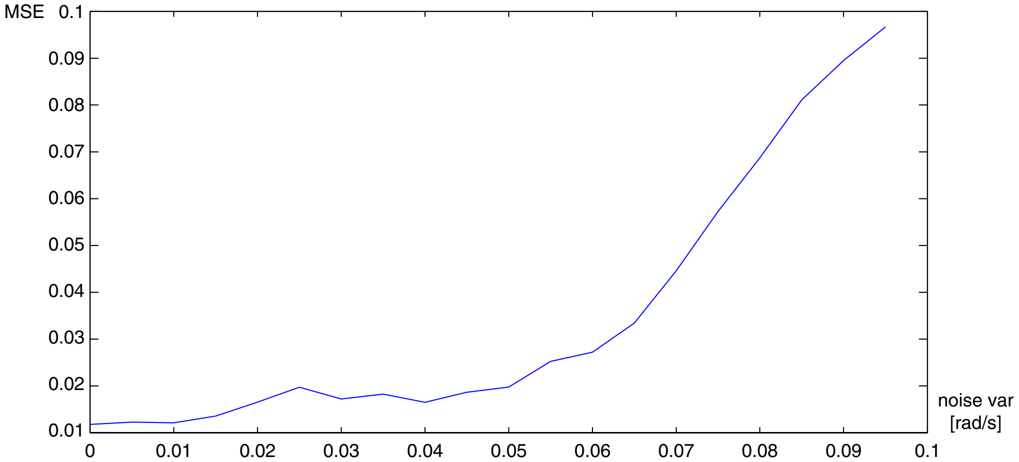


Fig. 9 Mean squared error (MSE) of difference between the original and deblurred image in relation to amount of added sensor noise. Gaussian noise of variance 0 to 0.1 was added to gyroscope measurements (angular velocity in rad/s). Deconvolution algorithm was then performed using computed blur kernels based on these altered measurements. MSE of difference to the original image is plotted in the graph (pixel value was normalized to $\langle 0, 1 \rangle$ range). The graph shows mean of 10 iterations for each of the variance values. Lena image was used for the test.

first blurred using convolution with a PSF counted from one set of gyroscope samples recorded in our mobile application. An additive noise was added to the image in accordance with the model 1 (40-dB Gaussian noise was used). Gaussian noise was also added to the gyroscope samples to simulate errors in sensor measurement. Corrupted image was then repaired using our deblurring algorithm from the altered motion data. Results for different amounts of noise in gyroscope samples are shown in Fig. 8. The mean square error of the result as a function of the gyroscope noise level (variance) is in Fig. 9. We can see that the performance starts to drop for noise levels above 0.05 rad/s. The gyroscope noise level typically encountered in the motion sensors inside mobile devices (in our case Samsung Galaxy S II) is 0.007 rad/s for our sampling rate, and it is therefore way below the critical level.

5 Conclusion

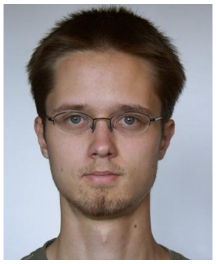
We have presented an image deblurring method that can effectively remove blur caused by camera motion using information from inertial sensors. The proposed method is fully implemented on a smartphone device, which is to our knowledge the first attempt in this direction and renders the method particularly appealing for end users. We have justified the space-invariant simplification for certain camera motions, but simultaneously we have uncovered intrinsic sources of space-variant blur, such as rolling shutter. The space-variant implementation of the deblurring algorithm, which would solve some of the current issues, is in theory possible, but the computational cost on the smartphone may be too high. It will be a topic of our future research to find out whether this is viable.

Acknowledgments

This work was supported in part by the Academy of Sciences of the Czech Republic under Grant M100751201 and by the Grant Agency of the Czech Republic under Project P103/11/1552.

References

1. R. Fergus et al., "Removing camera shake from a single photograph," in *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*, pp. 787–794, ACM, New York, NY (2006).
2. J. Jia, "Single image motion deblurring using transparency," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit., CVPR '07*, pp. 1–8, IEEE Computer Society Press, Silver Spring, MD (2007).
3. N. Joshi, R. Szeliski, and D. J. Kriegman, "PSF estimation using sharp edge prediction," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit., CVPR 2008*, pp. 1–8, IEEE Computer Society Press, Silver Spring, MD (2008).
4. Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," in *SIGGRAPH '08: ACM SIGGRAPH 2008 papers*, pp. 1–10, ACM, New York, NY (2008).
5. L. Xu and J. Jia, "Two-phase kernel estimation for robust motion deblurring," in *Proc. 11th European Conf. on Comput. Vis: Part I, ECCV'10*, pp. 157–170, Springer-Verlag, Berlin, Heidelberg (2010).
6. A. Levin et al., "Understanding blind deconvolution algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(12), 2354–2367 (2011).
7. A. Levin et al., "Understanding and evaluating blind deconvolution algorithms," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit., CVPR '09*, IEEE Computer Society Press, Silver Spring, MD (2009).
8. J. Miskin and D. J. C. MacKay, "Ensemble learning for blind image separation and deconvolution," in *Advances in Independent Component Analysis*, M. Girolani, Ed., Springer-Verlag, Berlin (2000).
9. N. P. Galatsanos et al., "Hierarchical bayesian image restoration from partially known blurs," *IEEE Trans. Image Process.* **9**(10), 1784–1797 (2000).
10. O. Whyte et al., "Non-uniform deblurring for shaken images," in *Proc. IEEE Conf. on Comput. Vis. and Pattern Recognit. CVPR '10* , pp. 491–498, IEEE Computer Society Press, Silver Spring, MD (2010).
11. A. Gupta et al., "Single image deblurring using motion density functions," in *Proc. 11th European Conf. on Comput. Vis.: Part I, ECCV'10*, pp. 171–184, Springer-Verlag, Berlin, Heidelberg (2010).
12. M. Tico, M. Trimeche, and M. Vehvilainen, "Motion blur identification based on differently exposed images," in *Proc. IEEE Int. Conf. Image Process.*, pp. 2021–2024, IEEE Computer Society Press, Los Alamitos, CA (2006).
13. L. Yuan et al., "Image deblurring with blurred/noisy image pairs," *in SIGGRAPH '07: ACM SIGGRAPH 2007 papers*, p. 1, ACM, New York, NY (2007).
14. M. Šorel and F. Šroubek, "Space-variant deblurring using one blurred and one underexposed image," in *Proc. IEEE 16th Int. Conf. on Image Process., ICIP 2009*, IEEE Computer Society Press, Los Alamitos, CA (2009).
15. M. Ben-Ezra and S. K. Nayar, "Motion-based motion deblurring," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(6), 689–698 (2004).
16. Y.-W. Tai et al., "Correction of spatially varying image and video motion blur using a hybrid camera," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(6), 1012–1028 (2010).
17. N. Joshi et al., "Image deblurring using inertial measurement sensors," *ACM Trans. Graph.* **29**(4), 30:1–30:9 (2010).
18. R. Liu and J. Jia, "Reducing boundary artifacts in image deconvolution," in *15th IEEE Int. Conf. on Image Process., ICIP 2008*, pp. 505–508 (2008).
19. M. Sorel, "Removing boundary artifacts for real-time iterated shrinkage deconvolution," *IEEE Trans. Image Process.* **21**(4), 2329–2334 (2012).



Ondřej Šindelář received an MS degree in computer science from Charles University in Prague, Czech Republic, in 2012. He is currently working on improvement of the project described in this article in collaboration with Dr. Šroubek and the Institute of Information Theory and Automation in Prague.

Filip Šroubek received an MS degree in computer science from the Czech Technical University, Prague, Czech Republic, in 1998 and PhD degree in computer science from Charles University, Prague, Czech Republic, in 2003. From 2004 to 2006, he was on a postdoctoral position in the Instituto de Optica, CSIC, Madrid, Spain. In 2010 and 2011, he was the Fulbright Visiting Scholar at the University of California, Santa Cruz. He is currently with the Institute of Information Theory and Automation and partially also with the Institute of Photonics and Electronics, where both institutes are part of the Academy of Sciences of the Czech Republic. He is an author of seven book chapters and over 80 journal and conference papers on image fusion, blind deconvolution, super-resolution, and related topics.

Compensating specular highlights for non-Lambertian projection surfaces

Chen-Tai Kao
Tai-Hsiang Huang
National Taiwan University
Graduate Institute of Communication Engineering
Taipei, Taiwan

Hua Lee
University of California
Department of Electrical and Computer Engineering
Santa Barbara, California

Homer H. Chen
National Taiwan University
Graduate Institute of Communication Engineering
Taipei, Taiwan
E-mail: homer@cc.ee.ntu.edu.tw

Abstract. *This paper concerns the compensation of specular highlight for handheld image projectors. By employing a projector-camera configuration, where the camera is aligned with the viewer, the distortion caused by nonideal (e.g., colored, reflective) projection surfaces can be estimated from the captured image and compensated for accordingly to improve the projection quality. This works fine when the viewing direction relative to the system is fixed. However, the compensation becomes inaccurate when this condition changes, because the position of the specular highlight changes as well. We propose a novel method that, without moving the camera, can estimate the specular highlight seen from any position and integrate it with Grossberg's radiometric compensation framework to demonstrate how view-dependent compensation can be achieved. Extensive results, both objective and subjective, are provided to demonstrate the performance of the proposed algorithm. © 2013 SPIE and IS&T [DOI: 10.1117/1.JEI.22.1.011004]*

1 Introduction

Using an image projector in a mobile phone or digital camera greatly overcomes the screen-size limitation of the handheld device and allows the image to be conveniently projected onto a bigger area on any nearby surface, such as a wall. Ideally, we would like the handheld projector to be able to project a clear image regardless of the physical characteristics of the projection surface. In practice, however, the projection surface available in the surroundings is often far from ideal and causes distortions to the projected image. As a result, geometric warping and radiometric compensation must be applied to the image before projection to counteract the nonideal characteristics of the projection surface. This compensation operation is especially important for immersive displays[1–3] and other related applications, where it is practically difficult to acquire an ideal screen. In general, algorithms for such visual computing utilize a projector-camera (procam) system, in which a number of calibration images are projected in advance, and the camera's feedback is analyzed to rectify the geometric[4–10] and photometric[11–24] properties of the projection surface. This paper concerns the radiometric compensation of a procam system.

A highly desirable technique to combat the effect of color distortion is tone mapping. It shifts the original image's color space such that, when projected on a colored screen, the compensated image is perceived undistorted. Grossberg et al.[21] proposed a procam model with a simple tone mapping. Only six calibration images are required to recover the screen's spectral response, prior to online compensation of each input image. While the chroma of the projected image is corrected, the contrast is inevitably reduced since the light absorbed by the screen is unrecoverable. Huang et al.[11] proposed to optimize the offset between chroma correctness and contrast such that radiometric compensation can be applied to deeply colored screen. Some other algorithms consider content-dependent compensation[25–27] or optimal tone mapping that is adaptive to the projector's gamut.[28,29] A detailed review of existing radiometric compensation techniques can be found in Ref. 30.

It is worth noting that, while these kinds of radiometric compensation algorithms have been successful in many cases, three major limitations remain. First, their use for non-Lambertian screens has been somewhat limited. The reason lies in the fact that, for reflective screens, the visual quality tends to be ruined by the specular highlight generated by the projection itself. The importance of this issue should not be neglected because it is hard to find an ideally diffusive surface to project whenever needed. To our knowledge, the

Paper 12343SS received Sep. 5, 2012; revised manuscript received Nov. 26, 2012; accepted for publication Dec. 21, 2012; published online Jan. 31, 2013.

problem of specular reflection regarding radiometric compensation has not been addressed before. Park et al. tried to eliminate specular light using multiple projectors,[31–33] but their work only estimated the position of the specular light. Without modeling specular light's intensity, Park's method is not applicable to radiometric compensation. Second, it is hard to characterize the specular light's intensity in a procam system because the bidirectional reflectance distribution function (BRDF) of the projection surface is unknown. For a procam system, the material of the projection surface cannot be determined in advance, so it is impossible to recover the BRDF from existing BRDF databases of various materials. Also, performing full measurement of the BRDF in a procam system is time-consuming and impractical for real-time usages. These obstacles make it hard to model, or even to eliminate, specular light in conventional radiometric compensation techniques. Third, one fundamental assumption of most radiometric compensation techniques is that the camera is placed where the viewer is supposed to be. This assumption is easily violated since the procam device can be placed statically while the viewer is allowed to move freely. In this scenario, the position of the specular light changes depending on the viewer's position, which nullifies the compensation calculated based on the statically placed camera. Thus, for non-Lambertian screens, one should design a more general radiometric compensation algorithm that takes the dynamics of specular highlight with respect to the viewing direction into consideration.

In this paper, we propose an algorithm for view-dependent radiometric compensation of non-Lambertian surfaces, with multifold contributions. First, it is a simple scheme that does not require additional projectors or cameras to reconstruct the BRDF of the surface—only one camera and one projector suffice. Second, it is the first that predicts the calibration images for different viewing angles from those captured at a single viewing angle, which greatly extends the capability of a procam system. Third, it introduces a feedback to estimate the specular light iteratively, which avoids over-compensation. Since the proposed method predicts the calibration images for an arbitrary viewing angle, it can be treated as a preprocessing module for existing radiometric compensation techniques. We show the effectiveness of the proposed method by integrating it with Grossberg's radiometric compensation framework,[21] which is accurate only at the viewing angle where the calibration images are captured. By predicting calibration images at various viewing angles using our method, automatic photometric compensation for an arbitrary viewing angle is made possible.

This paper is organized as follows. We introduce specular light in Sec. 2, and then we describe the proposed radiometric compensation algorithm in Sec. 3 and the experimental results in Sec. 4. Then the subjective and objective evaluations are provided in Sec. 5. Finally, the conclusions are drawn in Sec. 6.

2 Specular Highlight Modeling

Consider projection onto a non-Lambertian (or reflective) screen. While most of the light is evenly scattered, a small portion of light rays directly reflect as if the surface is a mirror. This mirror-like reflection of light is commonly known as specular highlight and has been considered important in three-dimensional (3-D) computer graphics.[34–38] There exist different models for predicting the distribution of the specular highlight, such as the Phong model,[38] the Blinn-Phong model,[37] the Gaussian distribution model,[39] and the Cook-Torrance model.[36]

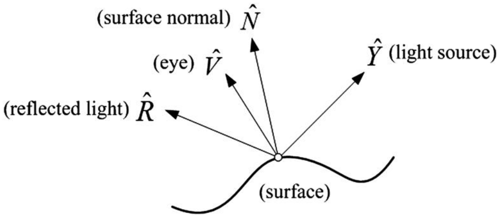


Fig. 1 Notations used in the Phong model. It characterizes how a viewer (at $\hat{V}$) perceives the reflected light (at $\hat{R}$) of an incident light (at $\hat{Y}$) on a point with surface normal $\hat{N}$.

We model the specular highlight using the Phong model and the notations shown in Fig. 1 in the following derivation. For each point on the surface, the specular light I_s is given by

$$I_s = \sum_{m \in \mathcal{L}} k_s (\hat{R}_m \cdot \hat{V})^{\gamma} i_{m,s}, \tag{1}$$

where $\mathcal{L}$ is the set of all light sources, k_s is the specular reflection constant, $\hat{R}_m$ is the direction of the perfectly reflected light, $\hat{V}$ is the direction toward the viewer, γ is the shininess constant for the screen material, and $i_{m,s}$ is the intensity of the light source. Note that $\hat{R}_m$ and $\hat{V}$ are unit vectors.

3 Proposed Approach

Figure 2 shows the overall architecture of the proposed method, which generates view-dependent compensated images for a non-Lambertian surface. The method first projects a uniform image onto the surface. Then it utilizes the projected image to fit the specular highlight model, estimate the calibration images viewed at different positions, and perform radiometric compensation using the estimated calibration images. The radiometric compensation framework proposed in Ref. 21 is adopted in our work.

To perform geometric calibration, a chessboard pattern is projected on the surface and captured by the camera. We then detect the chessboard corners in the captured image and apply Zhang's method[40] to find the geometric transformation that calibrates perspective distortion and radial distortion. Subsequently, all captured images are geometrically calibrated by the same geometric transformation.

Since the perceived specular light varies with respect to the viewing angle of a viewer, we need a precise description of the viewing angle. Assuming the viewer can only move along the xz-plane in Fig. 3, which depicts the viewing geometry, we define the viewing angle θ, measured in degree, to be 90 deg when the viewer stands right in front of the screen. In its simplest form, we assume the viewer and the procam system are at the same height. That is, the change of the viewing angle only has one degree-of-freedom. When precise spatial information of the viewer is available, the proposed method can be extended to account for 2-D variation of the viewing angle. In this section, the details of the proposed method are described in accordance with the processing flow shown in Fig. 2.

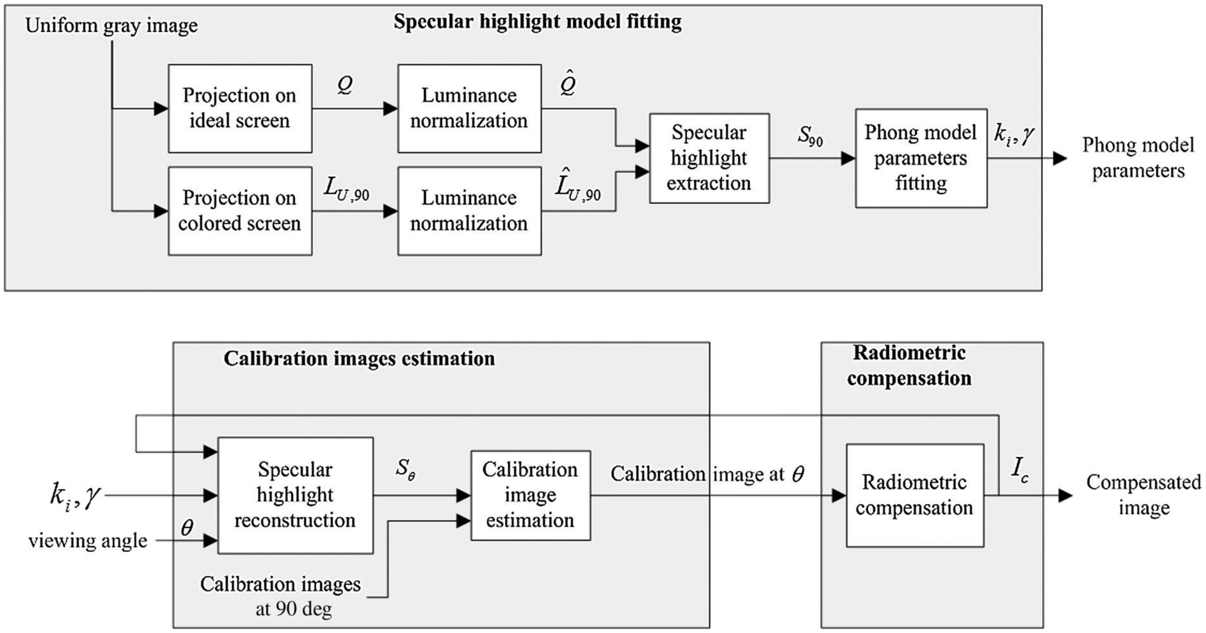

Uniform gray image
Specular highlight model fitting
Projection on ideal screen
Q
Luminance normalization
$\hat{Q}$
Projection on colored screen
$L_{U,90}$
Luminance normalization
$\hat{L}_{U,90}$
Specular highlight extraction
S_{90}
Phong model parameters fitting
k_i, γ
Phong model parameters
Calibration images estimation
Radiometric compensation
k_i, γ
viewing angle
θ
Specular highlight reconstruction
S_θ
Calibration image estimation
Calibration images at 90 deg
Calibration image at θ
Radiometric compensation
I_c
Compensated image

Fig. 2 The architecture of the proposed method, which consists of three major components: specular highlight model fitting, calibration images estimation, and radiometric compensation.

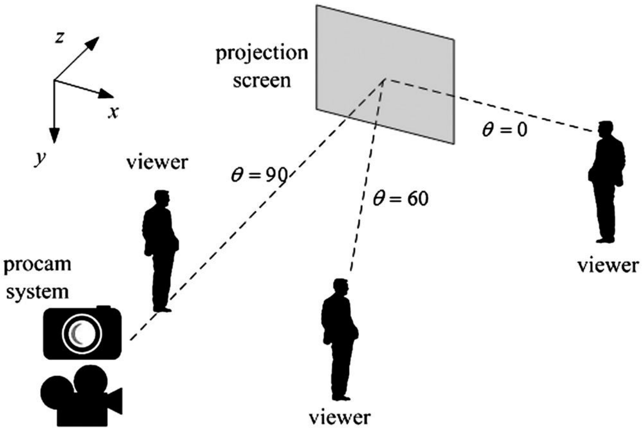


Fig. 3 The definition of viewing angle θ (measured in degree). $\theta = 90$ deg when the viewer stands right in front of the screen.

3.1 *Specular Highlight Model Fitting*

We estimate the unknown parameters, k_s, γ, and $i_{m,s}$, in the Phong model from the luminance variation of a projected image. We project a uniform gray image onto the screen and capture it by a camera placed at 90 deg. The captured image $L_{U,90}$ is normalized to $\hat{L}_{U,90}$ by

$$\hat{L}_{U,90} = \frac{L_{U,90}}{n}, \tag{2}$$

where

$$n = \max_{x,y} L_{U,90}(x,y) \tag{3}$$

is the normalizing constant. $\hat{L}_{U,90}$, with value ranging from 0 to 1, denotes the spatial variation of the luminance. Note that we cannot estimate the Phong model directly from $\hat{L}_{U,90}$, because the luminance variation is caused by the combination of the following two factors: (1) vignetting and (2) specular highlight. Vignetting, introduced by the imperfection of lens, often results in luminance reduction at the periphery of a photo. Therefore we need to estimate and exclude the vignetting factor before reconstructing the Phong model.

The vignetting effect can be calculated by projecting the same uniform gray image onto an ideal projection screen, which is assumed to be highly, if not perfectly, diffusive. More specifically, no specular light should appear on the ideal projection screen. We normalize the captured image Q to obtain $\hat{Q}$, which is the luminance variation caused by pure vignetting. Since the vignetting effect remains identical under the same procam configuration, we can extract the specular highlight S_{90} by

$$S_{90}(x,y) = \frac{\hat{L}_{U,90}(x,y)}{\hat{Q}(x,y)}. \tag{4}$$

It should be noted that the division is performed pixelwise. The two-dimensional (2-D) specular highlight S_{90} is then sampled along then sampled along the x-axis [see Fig. 4(a)] to obtain the one-dimensional (1-D) curve of the specular highlight, as shown in Fig. 4(b). The 1-D curve, denoted by s, fully characterizes the change of the specular light along the x-axis of the captured image. Data formed by s are used as samples of I_s in Eq. (1) to estimate the unknown parameters k_s, γ, and $i_{m,s}$. In our scenario, the screen is assumed to be homogeneous, i.e., made of the same material, so one set of k_s and γ suffices. Since there is only one light source (the projector), Eq. (1) can be rewritten as

$$s(x) = k_s[\hat{R}(x,y) \cdot \hat{V}(x,y)]^\gamma i_s|_{y=h}, \tag{5}$$

where h is the height of the camera. Note that the value of $\hat{R}$ and $\hat{V}$ can be obtained by considering the relative position of

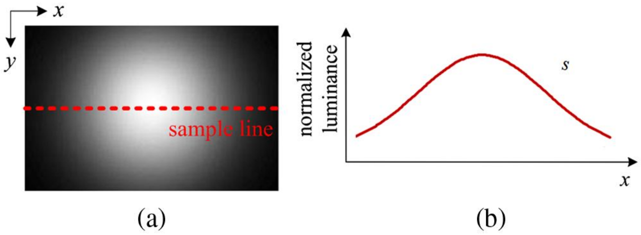


Fig. 4 (a) Specular highlight of a reflective projection surface when seen at 90 deg; (b) 1-D sample of the specular highlight.

the projector, the camera, and each point $p(x,y)$ on the projection surface. For each point $p(x,y)$ on the projection surface, $\hat{V}$ points from $p(x,y)$ toward the camera, $\hat{Y}$ points from $p(x,y)$ toward the projector, and the direct reflected light $\hat{R}$ can be computed as

$$\hat{R} = 2(\hat{Y}\cdot\hat{N})\hat{N} - \hat{Y}, \tag{6}$$

where $\hat{N}$ is the surface normal. The geometric interpretation of Eq. (6) is shown in Fig. 5(c). Note that $\hat{V}$, $\hat{R}$, $\hat{N}$, and $\hat{Y}$, are all unit vectors in 3-D space. Figure 5(a) demonstrates the camera direction for planar projection surface. With variables rearranged, and logarithm taken on both sides of Eq. (5), the equation becomes

$$\log[s(x)] = \log(k_s i_s) + \gamma \log[\hat{R}(x,y)\cdot\hat{V}(x,y)]|_{y=h}. \tag{7}$$

Based on the scatter plot of $\log(s)$ and $\log(\hat{R}\cdot\hat{V})$, we use linear regression to obtain $k_s i_s$ and γ.

The method can also be extended to use all samples in $S_{90}(x,y)$ to fit the Phong model, instead of using only the 1-D samples of s. In this case, Eq. (1) is rewritten as

$$S_{90}(x,y) = k_s[\hat{R}(x,y)\cdot\hat{V}(x,y)]^{\gamma} i_s, \tag{8}$$

and $k_s i_s$ and γ can be fitted by linear regressing all data points in $S_{90}(x,y)$. Since more samples are used, the fitted parameters may be more accurate.

3.2 *Calibration Images Estimation*

Given the specular light response of the screen at $\theta = 90$, we now move the virtual camera to an arbitrary viewing angle θ and reconstruct the 2-D specular highlight S_θ observed there. It follows from Eq. (1) that we can predict S_θ by

$$S_\theta(x,y) = k_s[\hat{R}(x,y)\cdot\hat{V}_\theta(x,y)]^{\gamma} i_i, \tag{9}$$

where $\hat{V}_\theta(x,y)$ denotes the direction pointing from the pixel $p(x,y)$ to the virtual camera at viewing angle θ. Note that k_i, i_i, and γ are as computed in Sec. 3.1, and $\hat{R}$ remains unchanged because the projector stays at the same place. Figure 5(b) demonstrates how the values of $\hat{V}_\theta(x,y)$ and $\hat{R}(x,y)$ can be obtained.

Five calibration images should be captured ($L_{R,\theta}$, $L_{G,\theta}$, $L_{B,\theta}$, $L_{U,\theta}$, and $L_{S,\theta}$) for the radiometric compensation framework (see Sec. 3.3). To generate radiometric compensation for viewing angle θ, the calibration images at that viewing angle should be estimated. Calibration images at θ($L_{R,\theta}$, $L_{G,\theta}$, $L_{B,\theta}$, $L_{U,\theta}$, and $L_{S,\theta}$) are estimated by adding the change of specular light on the calibration images at $\theta = 90$. That is,

$$L_{M,\theta} = L_{M,90} + n(S_\theta - S_{90}), \qquad M \in \{R,G,B,U,S\}, \tag{10}$$

where n is the normalizing constant defined in Eq. (3) and θ denotes the viewing angle (measured in degree) of the calibration images.

3.3 *Radiometric Compensation*

We adopt the framework proposed in Ref. 11 for radiometric compensation, which is based on the procam model first introduced by Grossberg et al.[21] Grossberg's model has great advantage in that only six calibration images are needed for the photometric compensation. Because Grossberg's model deals with screens with spatially variant color, an "invariant value" is computed pixel-wise in order to reflect the color distortion of each pixel.[21] In our scenario, where the screen color is assumed to be spatially uniform, the computation of that "invariant value" is not needed. This saves one calibration image, and thus only five calibration images are required. These calibration images consist of

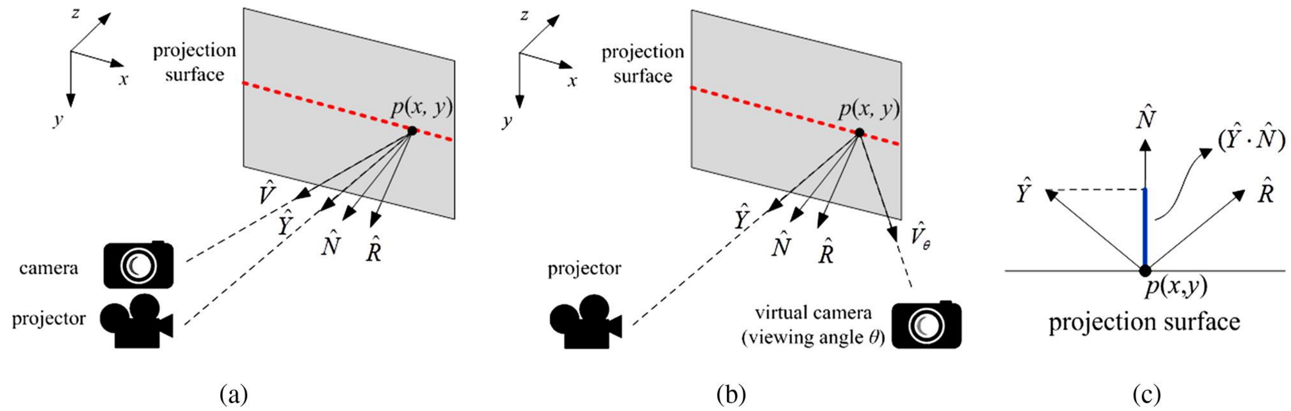


Fig. 5 Illustration of the camera direction in (a) Phong model fitting for planar projection surface and (b) specular light estimation. In the model fitting, the camera is placed right above the projector. In the specular light estimation, the virtual camera is placed at a different angle θ. In both conditions, the direct reflected light $\hat{R}$ is calculated by Eq. (6), whose geometric interpretation is shown in (c).

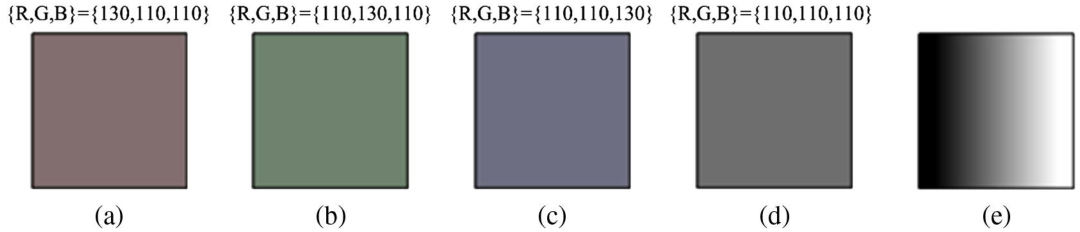


Fig. 6 The five calibration images: (a) red, (b) green, (c) blue, (d) gray, and (e) slope—a ramp with pixel values ranging from 0 to 255.

Table 1 Configuration of the five calibration images.

Color	Pixel value
Red	$\{R,G,B\} = \{130,110,110\}$
Green	$\{R,G,B\} = \{110,130,110\}$
Blue	$\{R,G,B\} = \{110,110,130\}$
Gray	$\{R,G,B\} = \{110,110,110\}$
Slope (ramp)	$\{R,G,B\} = \{0,0,0\} \sim \{255,255,255\}$

four uniform-colored images (red, green, blue, and gray) and one color ramp consisting of pixels ranging from gray-level 0 to gray-level 255 (Slope). We denote the gray image as U. These images are shown in Fig. 6(a) to 6(e), whose pixel values are summarized in Table 1. The pixel values of the four uniform-colored images should not be set too low; otherwise, the captured image may be subject to severe noise generated by the camera sensor. They should not be set too high, either, since the projector's response curve is nearly flat when displaying very bright content. Therefore, we set the pixel values to be around 110 to 130, a much safer choice near the middle of 0 to 255. We project these calibration images and denote the captured images as $L_{M,\theta}$, $M \in \{R,G,B,U,S\}$, where the camera is placed at viewing angle θ.

Here we want to emphasize a crucial contribution of the proposed method: no extra image is needed for the reconstruction of the specular highlight response. We use the gray calibration image $L_{U,90}$ as the source image to estimate the specular highlight. The brightness of U is carefully chosen such that specular light appears, while the resulting luminance variation is recordable by the camera. Empirically, the pixel value is set to 110.

The procam model is shown in Fig. 7, with each part illustrated as follows. First, the original image I, after being mapped by the projector's response curve f_p, becomes the projected luminance P,

$$P = f_p(I), \tag{11}$$

where f_p is typically a nonlinear transfer function. Note that $I = [I_r, I_g, I_b]^T$ is the input image with pixel value ranging from 0 to 255, while $P = [P_r, P_g, P_b]^T$ is in the luminance domain. Both I and P are 3-D vectors (red, green, and blue channel).

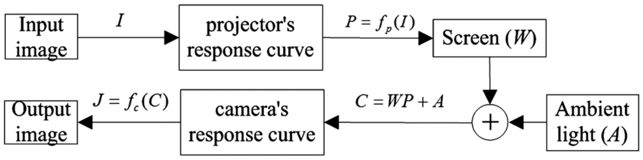


Fig. 7 Grossberg's procam model. The input image I is first mapped by the projector's response curve f_p. The projected luminance P is then modulated by the screen's color (modeled as a 3×3 color mixing matrix W) and the ambient light A before being captured by the camera. The captured luminance is then mapped by the camera's response curve f_c to be the output image J.

The overall modulation of the screen is modeled by a 3×3 matrix W that characterizes all possible interactions between the incident light and the reflected light:

$$W = \begin{bmatrix} W_{RR} & W_{RG} & W_{RB} \\ W_{GR} & W_{GG} & W_{GB} \\ W_{BR} & W_{BG} & W_{BB} \end{bmatrix}. \tag{12}$$

Generally, W is called the color mixing matrix, which captures the coupling between each color channel of the projector and the camera. W_{GR}, for example, denotes the portion of red channel of the projector that contributes to the green channel of the camera. It should be noted that W is spatially variant for each pixel in the projected screen.

By adding the ambient light A, the captured image C is modeled by

$$C = WP + A, \tag{13}$$

where C, P, and A are 3-D vectors (red, green, and blue channel), and W is as defined in Eq. (12). The captured image C is then transformed back by the camera response curve f_c to result in the output image J,

$$J = f_c(C). \tag{14}$$

Grossberg et al.,[21] proposed an efficient method to recover all parameters of the procam model. First, the color-mixing matrix W is decomposed by

$$W = \tilde{W}D = \begin{bmatrix} 1 & \tilde{W}_{RG} & \tilde{W}_{RB} \\ \tilde{W}_{GR} & 1 & \tilde{W}_{GB} \\ \tilde{W}_{BR} & \tilde{W}_{BG} & 1 \end{bmatrix} \begin{bmatrix} W_{RR} & 0 & 0 \\ 0 & W_{GG} & 0 \\ 0 & 0 & W_{BB} \end{bmatrix}, \tag{15}$$

where $\tilde{W}$ and D encodes the inter-channel and the intra-channel interaction, respectively. $\tilde{W}$ can be determined using just four calibration images ($L_{R,\theta}$, $L_{G,\theta}$, $L_{B,\theta}$, and $L_{U,\theta}$). Let

$\bar{L}_{R,\theta}$, $\bar{L}_{G,\theta}$, $\bar{L}_{B,\theta}$, $\bar{L}_{U,\theta}$ denote the mean pixel value of $L_{R,\theta}$, $L_{G,\theta}$, $L_{B,\theta}$, $L_{U,\theta}$, respectively, then $\tilde{W}$ can be computed as

$$\tilde{W} = \begin{bmatrix} 1 & \frac{\{\bar{L}_{G,\theta}-\bar{L}_{U,\theta}\}_r}{\{\bar{L}_{G,\theta}-\bar{L}_{U,\theta}\}_g} & \frac{\{\bar{L}_{B,\theta}-\bar{L}_{U,\theta}\}_r}{\{\bar{L}_{B,\theta}-\bar{L}_{U,\theta}\}_b} \\ \frac{\{\bar{L}_{R,\theta}-\bar{L}_{U,\theta}\}_g}{\{\bar{L}_{R,\theta}-\bar{L}_{U,\theta}\}_r} & 1 & \frac{\{\bar{L}_{B,\theta}-\bar{L}_{U,\theta}\}_g}{\{\bar{L}_{B,\theta}-\bar{L}_{U,\theta}\}_b} \\ \frac{\{\bar{L}_{R,\theta}-\bar{L}_{U,\theta}\}_b}{\{\bar{L}_{R,\theta}-\bar{L}_{U,\theta}\}_r} & \frac{\{\bar{L}_{G,\theta}-\bar{L}_{U,\theta}\}_b}{\{\bar{L}_{G,\theta}-\bar{L}_{U,\theta}\}_g} & 1 \end{bmatrix}, \tag{16}$$

where indices r, g and b denote the red, green and blue channels, respectively. Multiplying $\tilde{W}^{-1}$ on both sides of Eq. (13) yields

$$\tilde{C} = DP + \tilde{A}, \tag{17}$$

where

$$\tilde{C} = \tilde{W}^{-1}C, \tag{18}$$

and

$$\tilde{A} = \tilde{W}^{-1}A. \tag{19}$$

We say that Eq. (17) decouples the color response of the screen, since each color channel in $\tilde{C}$ is now affected only by the same channel of the projector, the screen and the ambient light. The monotonic mapping ρ between the pixel value (0 to 255) and the corresponding decoupled luminance is defined as

$$\tilde{C} = \rho(I). \tag{20}$$

We reconstruct ρ by the regression of $\tilde{L}_{S,90}(x,y)$, where $\tilde{L}_{S,90}(x,y) = \tilde{W}^{-1}L_{S,90}(x,y)$.

We reconstruct $\tilde{W}$ and ρ for the white screen and the colored screen, denoted by $\{\tilde{W}_w, \rho_w\}$ and $\{\tilde{W}_c, \rho_c\}$, respectively, by projecting the calibration images onto both screens. Once the parameters are reconstructed, we gain full information about how the projected image is perceived on both screens. Therefore it becomes possible to compensate the radiometric distortion and to make the projection onto a colored screen looks as if it is projected onto a white screen.

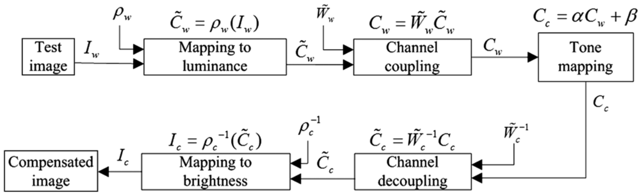


Fig. 8 The block diagram of radiometric compensation. The test image I_w is first converted to the decoupled luminance $\tilde{C}_w$, then the color-mixing-matrix $\tilde{W}_w$ is used to compute the coupled luminance C_w. The tone mapping function in Eq. (21) is used to map the luminance C_w seen on white-screen to the luminance C_c on colored-screen. Finally, the luminance is decoupled by $\tilde{W}_c^{-1}$ to be the decoupled luminance $\tilde{C}_c$, which is then converted back to the compensated image I_c.

Figure 8 illustrates the radiometric compensation framework for a test image I_w. Under this framework, we generate a compensated image I_c that, when projected on the colored screen, is perceived almost the same as projection of I_w on the white screen. The test image is converted first to the decoupled luminance ($\tilde{C}_w$) by ρ_w, then to the desired luminance (C_w) by multiplying the color-mixing matrix $\tilde{W}_w$. Here C_w serves as a simulation of the perceived luminance supposing the test image is projected on the white screen.

Projection onto a colored screen often leads to loss of the dynamic range because the upper bound of displayable luminance is modulated by the screen color, while the lower bound is determined by the ambient light that brings additional luminance to black pixels (see Fig. 9). Therefore compensation toward photometric correctness requires that the dynamic range be compressed within the recoverable range. This can be achieved by tone mapping, but the image contrast is inevitably reduced. A technique was developed in Eq. (11) to optimize the tradeoff between the photometric correctness and the contrast. The tone mapping function is defined as follows:

$$C_c = \alpha C_w + \beta. \tag{21}$$

Here, α and β are determined by considering two kinds of error: over-upper-bound error (E_u) and below-lower-bound error (E_l). E_u and E_l account for pixels that, after this tone mapping, lie outside the recoverable dynamic range. In practice, E_u is computed as the sum of pixels whose luminance lies over the upper bound; E_l for pixels with luminance below the lower bound. α and β are determined by the following optimization:

$$(\alpha, \beta) = \underset{\alpha,\beta}{\operatorname{argmin}}(E_u + E_l) + \lambda E_b, \tag{22}$$

where $E_b = (1-\alpha)^2$ is the penalty term for brightness and λ is a weighting factor. The optimization attempts to maximize the number of pixels that lie within the recoverable dynamic range, while preserving as much contrast as possible.

We compute the compensated image I_c from the desired luminance C_c by decoupling the color channels, followed by mapping the decoupled luminance to an 8-bit pixel value. The following equations give the compensated image I_c:

$$\tilde{C}_c = \tilde{W}_c^{-1}C_c, \tag{23}$$

$$I_c = \rho_c^{-1}(\tilde{C}_c). \tag{24}$$

The quality of the compensated image is determined by the reconstructed parameters of the procam model, which in turn depends fully on the captured calibration images. Since the

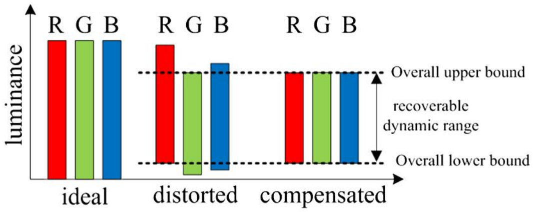


Fig. 9 When the brightest white is projected onto a colored screen, each color channel suffers from a different amount of distortion. The upper bound of each channel is independently modulated by the screen's color, while the lower bound is determined by the reflected ambient light that adds additional luminance to black pixels. Radiometric compensation recovers the color, with the recoverable dynamic range determined by the minimum dynamic range of the three channels. The contrast is therefore decreased.

calibration images are estimated by a virtual camera colocated with the viewer, the compensation is view-dependent. For simplicity, we refer to the image compensated for viewing angle θ as "θ compensated image." For example, if the viewer is at 60 deg, the compensated image is denoted by 60 deg compensated image.

3.4 *Specular Highlight Estimation Feedback*

The luminance of the projected light determines the chroma and the intensity of the specular highlight. In particular, when a compensated image is projected, the specular light slightly differs from that estimated in the initial condition, under which the calibration image is projected. This often leads to over-compensation, which is explained as follows. Suppose a green compensation image is projected onto a magenta screen, which is expected to recover a purely gray image. Due to the compensation, the projection is now much greener, making the specular light greener as well. Extra hue distortion is thus added to the originally well-compensated gray image, resulting in overcompensation.

We propose to incorporate the specular highlight estimation feedback that estimates the specular light based on the content of the compensation image. Since the new compensation image only affects the intensity of specular light (i_s) in the Phong model, we change the estimated parameter $k_s i_s$ by replacing i_s with i_c according to the following equation:

$$i_c = i_s \frac{\bar{I}_c}{\bar{U}}, \tag{25}$$

where $\bar{U}$ and $\bar{I}_c$ are the mean pixel intensity of U and I_c, respectively. Note that Eq. (25) is computed respectively for each color channel. In practice, we iterate the feedback loop three times.

Here the light from the projector is modeled as a point source. This is definitely a simplified, yet usable, model. More complex modeling can be designed to compute the projector's light field, which considers the specular light caused by each single beam projected onto each pixel on the surface. Nevertheless, this is beyond the scope of this paper and deserves a separate treatment.

3.5 *Possible Extension*

The proposed method is designed for, but not limited to, procam systems consisting of a projector and a camera bound together. This is the most applicable configuration that can be nicely packaged as one unit and is well suited for mobile projectors. For a procam system with multiple projectors and cameras, some extra work is needed for the proposed technique to be adopted. In the case of multiple projectors, geometric registration (e.g., using Zhang's method)[40] is needed for each projector to seamlessly merge all projections. Then our technique can be used to estimate the specular light contributed by each projector. Specifically, we can divide the image to be projected into a number of patches, each of which is then projected onto the projection surface by the specific projector that produces the least specular light for that patch. In this way, the system is able to support multiple viewers at different viewing angles simultaneously with minimum specular light perceived by each viewer. With multiple cameras, on the other hand, the system is able to estimate the specular light distribution more accurately, because images captured at various viewing angles now provide more samples to fit the specular light model.

4 Experimental Results

The proposed algorithm is applied to the radiometric compensation framework described in Ref. 11 and shown in Fig. 2 as the radiometric compensation module. Specifically, we integrate specular highlight model fitting and calibration images estimation with the radiometric compensation module.

Figure 10 shows the experimental setup of our procam system, which consists of a projector (SanyoPLC-XW56) and a camera (Canon40D). The screen is an A4 paper, one half of which is printed in color by a color laser printer. It should be noted that the printed ink is itself reflective, making the colored side a non-Lambertian surface. Also note that both the projector and the camera are placed at 90 deg viewing angle.

We report results for two test images, namely Waterfall and Motorbikes. The resolution of both images is 640 × 320. The projection of both images on the white screen is shown in Fig. 11, which serves as ideal images for other projections to compare with. We project Waterfall on a magenta screen, where the photometric distortion is shown in Fig. 12(a). To compensate for the distortion, 90 deg compensated image is projected and the result is captured by a camera placed at 90 deg [see Fig. 12(b)]. Although Fig. 12(b) recovers most visual quality, the compensation becomes inaccurate when seen at 60 deg, as shown in Fig. 12(c). In fact, the specular light present at 90 deg almost vanishes when seen from 60 deg, thus the image's chroma is subject more to the screen color. Figure 12(d) gives the projection of 60 deg compensation, in which the color is corrected by the proposed algorithm.

The effectiveness of the proposed method can be identified by examining the green color that is recovered. As a matter of fact, green is the most absorbed color for a magenta screen. Therefore, when the viewing angle changes from 90 to 60 deg, it is the green channel that is the most severely affected. By comparing the green channel that is recovered in Fig. 11(c) and 11(d), it becomes clear that the proposed method offers a satisfactory solution for retaining consistent visual quality across different viewing angles.

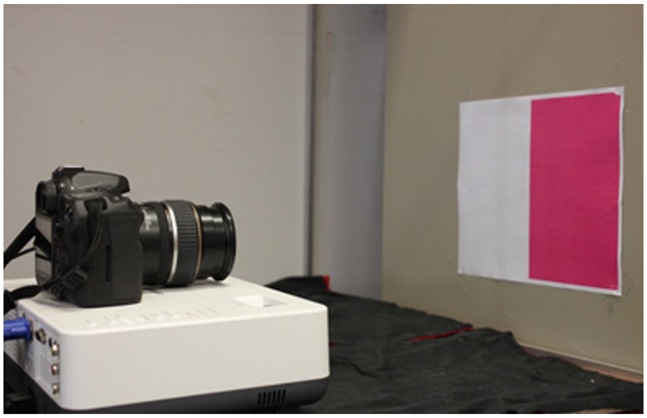


Fig. 10 Experimental setup. The procam consists of a projector (Sanyo PLC-XW56) and a camera (Canon 40D). The screen is an A4 white paper, one half of which is printed in color by a color laser printer.

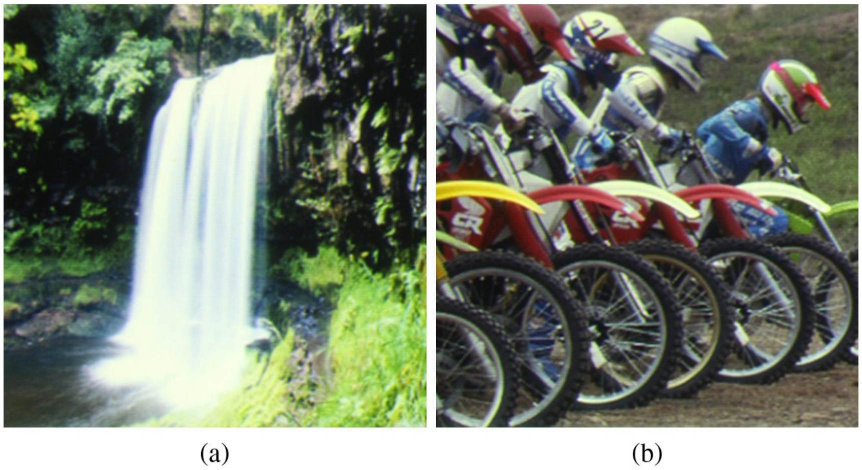


Fig. 11 Two test images projected on a white surface: (a) Waterfall and (b) Motorbikes.

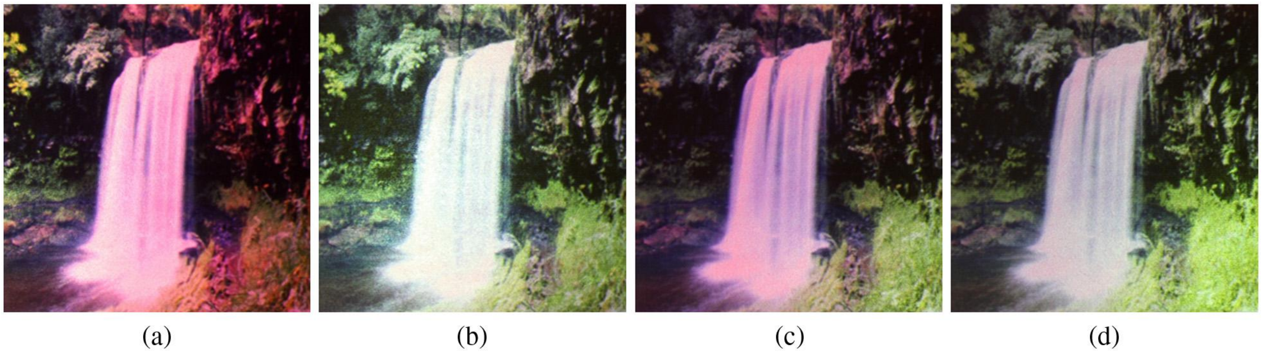


Fig. 12 Waterfall projected on a magenta surface and seen at different viewing angles:(a) original image seen at 90 deg; (b) 90 deg compensated image seen at 90 deg; (c) 90 deg compensated image seen at 60 deg; (d) 60 deg compensated image seen at 60 deg.

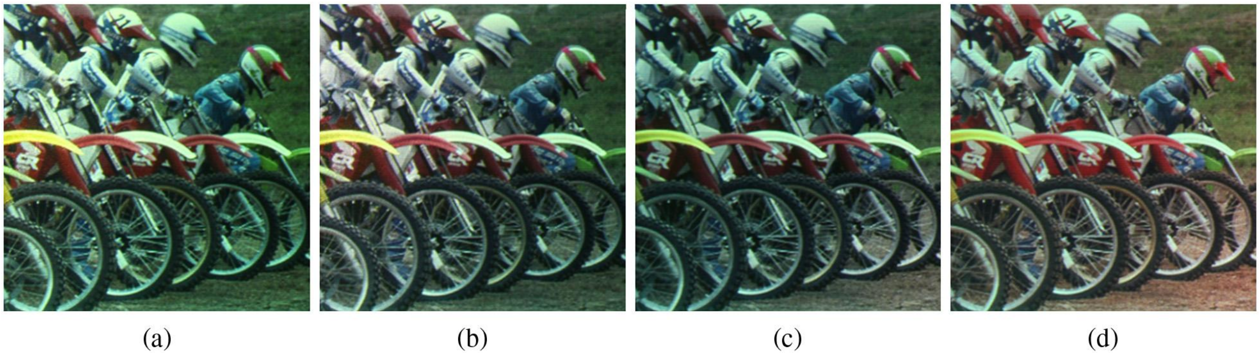


Fig. 13 Motorbikes projected on a green surface and seen at different viewing angles: (a) original image seen at 90 deg; (b) 90 deg compensated image seen at 90 deg; (c) 90 deg compensated image seen at 60 deg; (d) 60 deg compensated image seen at 60 deg.

Figure 13 shows the result for another test image Motor bikes. In this case, a green screen is used. The implication of the result is similar to that of Fig. 12.

5 Evaluation

To validate the proposed method, an evaluation is conducted to test the model under a representative condition. The same set of 90 deg compensated and 60 deg compensated images are projected onto a reflective color surface, and the projection results are evaluated both objectively and subjectively.

5.1 *Comparison with the Ideal Image*

This section presents a quantitative analysis of the proposed method in terms of its ability of projection quality enhancement. To evaluate the method's generality, a test image is first projected on an ideal screen as the ground truth T_i. We then project the following onto the colored screen: (1) original image T_o; (2) 90 deg compensated image T_{90}; and (3) 60 deg compensated image T_{60}. To compare the visual quality seen at different viewing angle, we capture the images of T_o, T_{90}, and T_{60} at 60 deg. The quality of T_o,

Table 2 Comparison of the sum of squared error (unit: 10^7).

Screen color	Pixel value of the test image [R G B]	Seen at 90 deg			Seen at 60 deg		
		T_o	T_{90}	T_{60}	T_o	T_{90}	T_{60}
Red	[100 50 50]	8.48	**1.70**	4.23	12.4	2.74	**1.87**
	[50 100 50]	16.7	**3.73**	7.34	27.7	5.64	**3.57**
	[50 50 100]	17.8	**5.41**	7.32	27.6	9.98	**8.16**
Green	[100 50 50]	14.1	**3.82**	3.98	18.1	5.72	**1.70**
	[50 100 50]	14.1	**7.55**	7.73	18.4	9.95	**7.42**
	[50 50 100]	12.0	**6.12**	6.71	14.3	7.55	**4.76**
Blue	[100 50 50]	17.8	**2.39**	4.92	29.6	8.38	**4.88**
	[50 100 50]	21.8	**6.52**	7.66	35.6	15.4	**13.6**
	[50 50 100]	8.88	**4.67**	8.92	12.8	3.59	**2.51**

T_{90}, and T_{60} can be represented by summing the pixel-wise difference deviated from T_i. That is, the lower the sum of squared error, the better the quality. A general representation of the error is as follows:

$$E = \sum\nolimits_{x,y} |T_t(x,y) - T_i(x,y)|^2, \qquad T_t \in \{T_o, T_{90}, T_{60}\}. \tag{26}$$

We choose a representative set of test images and the screen colors and adopt red, green, and blue as the primary colors. We use three uniform-colored images as test images and set their pixel value [R G B] to [100 50 50], [50 100 50] and [50 50 100], respectively. They are paired with red, green and blue screens to form nine combinations, covering the basis of all possible image-screen pairing in this setup.

Table 2 lists the results in sum of squared error. There are two key observations. First, it is evident that, though 90 deg compensation performs better when seen at 90 deg, it is outperformed by 60 deg compensation when the viewing angle is changed to 60 deg. The result strongly indicates that the radiometric compensation should be adaptive to the viewer's position. More importantly, our algorithm is able to further lower the error. Second, the quality of the original image is further worsened when seen from 60 deg. This can be explained by considering the following two factors that affect the projection: specular light and the screen color. The perceived quality at 90 deg tends to incorporate large portion of specular light, which is of the same color of the original image. However, when seen from 60 deg, the specular light dwindles, and the influence from the screen color increases, leading to greater photometric distortion. The worsened quality implies the necessity for radiometric compensation being used under such condition.

5.2 *Subjective Evaluation*

We took a total of 26 images in the Kodak lossless true color image suite[41] as the test images. We recruited 20 volunteers to judge the quality of the projections. Among them, 13 were male and seven were female, with their age ranging from 22 to 27. All subjects reported no severe visual abnormalities other than myopia.

Each test image was projected onto a magenta screen with three variations: (1) the original image T_o; (2) the 90 deg compensated image T_{90}; and (3) the 60 deg compensated image T_{60}. For each test image, the projection process consisted of three rounds, and in each round two of the three candidates (T_o, T_{90}, and T_{60}) were projected. Actually, it was a round-robin contest for T_o, T_{90}, and T_{60} in a random order. The subjects were asked to select from the two candidates the one with better quality. The criteria for judging the quality included chroma correctness, brightness, and contrast. The voting system was implemented using MATLAB, where the subject used keyboard to cast the vote. The same process was repeated twice, where the subject viewed the projection at 60 deg in the first time and 90 deg in the second.

Figure 14 summarizes the result of each match. Figure 14(a) to 14(c) are for subjects being at 90 deg viewing angle, while Fig. 14(d) through 14(f) are for 60 deg. The overall statistics of each match is shown in the bottom-right box. Here, we use the symbol "$>$" to denote quality superiority. For instance, by "$x > y$" we mean "candidate x has better quality than candidate y." We now discuss the result of each match: $\{T_o, T_{90}\}$, $\{T_o, T_{60}\}$, and $\{T_{90}, T_{60}\}$ as follows.

$\{T_o, T_{90}\}$: This match evaluates the effectiveness of the conventional radiometric compensation technique. When seen at 90 deg [Fig. 14(a)], T_{90} is highly preferred (90.19%) over T_o, indicating that the radiometric compensation is successful. It seems counterintuitive that, when seen at 60 deg, T_{90} performs even better as it received 98.85% of the votes. This can be explained by noting that the quality of T_o severely worsens when the viewing angle changes from 90 to 60 deg, making it even less competitive. As the specular light abounding at 90 deg become out of sight when seen at 60 deg, the screen color gradually dominates, and hence ruins, the quality of T_o. For this reason, one should not mistakenly consider that the conventional technique also performs well for T_{90} when seen at 60 deg, although T_{90} received more votes in 60 deg (98.85%) than in 90 deg (90.19%).

$\{T_o, T_{60}\}$: This match shows that T_{60} received votes as high as 97.69% against T_o, indicating that T_{60} has great quality when seen at 60 deg [see Fig. 14(e)]. The match also shows that, though highly preferred at 60 deg, T_{60} was considered poor in quality (only 62.88% votes) when seen at 90 deg. This is in accordance with our expectation: The amount of compensation in T_{60} exceeds what is needed for 90 deg viewing angle so that the specular highlight of the compensated color reflects and, as a result, spoils the visual quality. When viewing a magenta screen at 90 deg, for example, strong green color in T_{60} would mostly be reflected as specular light. Consequently, the resulting image is perceived much greener than the original image. In this case, subjects showed no strong preference over either T_o or T_{60}.

$\{T_{90}, T_{60}\}$: This is the key part of the whole evaluation, which proves the validity of the proposed algorithm. From Fig. 14(f) we can see that T_{60}, generated by the proposed method, outperforms T_{90} when seen at 60 deg. Recall that the traditional radiometric compensation schemes do not

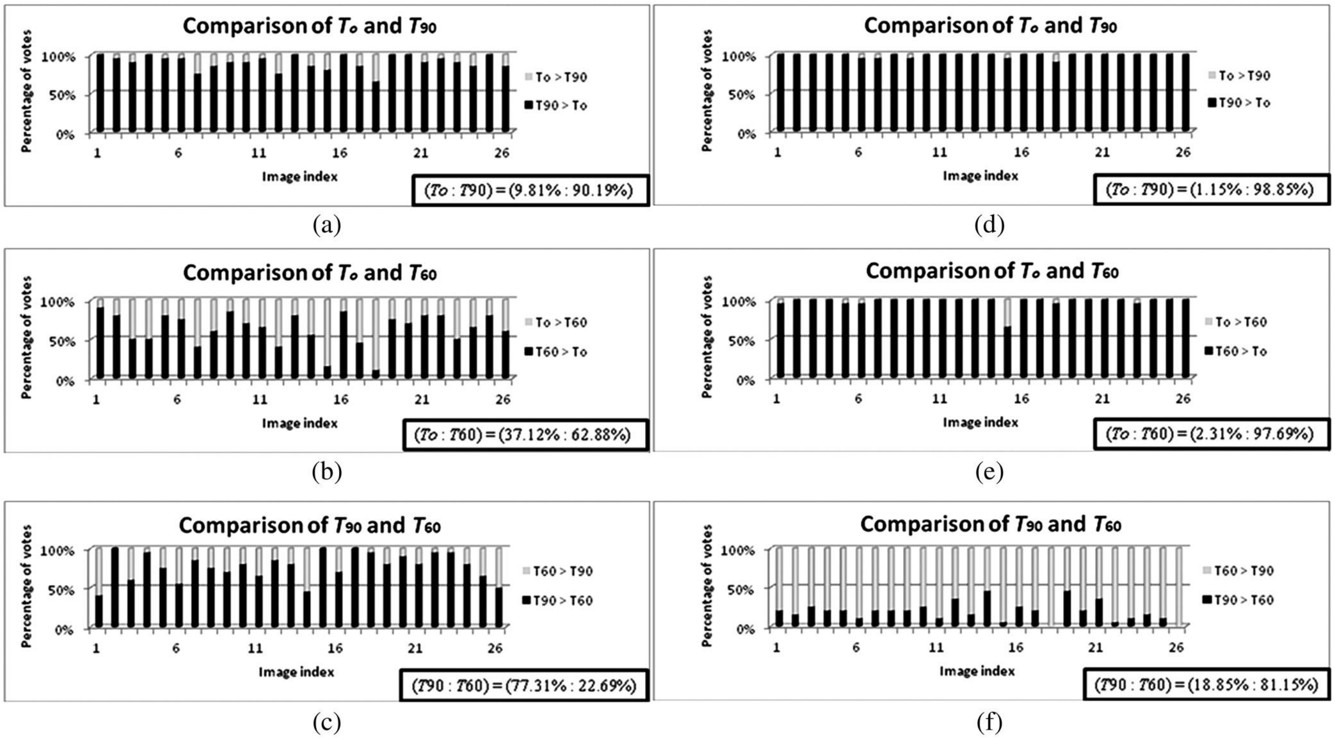


Fig. 14 Statistics of the subjective evaluation. By "$x > y$" we mean that "x has better quality than y." For (a), (b), and (c), the subject views the projection at 90 deg viewing angle. For (d), (e), and (f), the viewing angle is 60 deg. The statistics of total votes are shown in the bottom-right box of each subfigure.

take specular light and viewing angle into account. In contrast, our technique is able to provide better visual quality. Figure 13(c) and 13(f) together strongly suggest that the radiometric compensation should be adaptive to the viewing angle. In other words, it is not the best practice to apply a one-for-all compensation and ignore the viewing direction. A view-dependent compensation works much better.

6 Conclusion

In this paper, we have addressed the problem of recovering the projection quality for non-Lambertian surfaces. We have proposed a novel specular highlight estimation algorithm for radiometric compensation and discussed how calibration images for different viewing angles can be predicted from the reconstructed specular light model.

The proposed algorithm has been rigorously tested. Both objective and subjective evaluations have been carried out on a number of test images, and it is shown that the proposed algorithm consistently outperforms conventional ones. The proposed technique provides good visual quality for the projection. In addition, due to its low complexity, it can be easily incorporated into existing procam systems.

Acknowledgments

This project is supported in part by a grant from the National Science Council of Taiwan under Contract NSC 100-2221-E-002-197-MY3, a grant from National Taiwan University under Contract 102R7609-2, and a grant from the Himax Technologies, Inc.

References

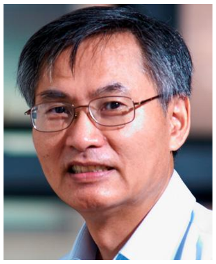
1. S. Zollmann, T. Langlotz, and O. Bimber, "Passive-active geometric calibration for view-dependent projections onto arbitrary surfaces," *J. Virtual Reality Broadcast.* **4**(6), 1–10 (2007).
2. J. Oh et al., "Portable projection-based AR system," in *Proc. 3rd Int. Conf. on Adv. in Vis. Comput.*, pp. 742–750, Springer-Verlag, Berlin, Heidelberg (2007).
3. H. Park et al., "Surface-independent direct-projected augmented reality," in *Proc. 7th Asian Conf. Comput. Vision*, pp. 892–901, Springer-Verlag, Berlin, Heidelberg (2006).
4. T. Johnson and H. Fuchs, "Real-time projector tracking on complex geometry using ordinary imagery," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 1–8, IEEE Computer Society, Washington, DC (2007).
5. E. S. Bhasker, R. Juang, and A. Majumder, "Registration techniques for using imperfect and partially calibrated devices in planar multi-projector displays," *IEEE Trans. Visual. Comput. Graph.* **13**(6), 1368–1375 (2007).
6. M. Brown, A. Majumder, and R. Yang, "Camera based calibration techniques for seamless multi-projector displays," *IEEE Trans. Visual. Comput. Graph.* **11**(2), 193–206 (2005).
7. R. Raskar et al., "RFIG lamps: interacting with a self-describing world via photosensing wireless tags and projectors," in *Proc. ACM SIGGRAPH*, pp. 406–415, ACM, New York, NY (2004).
8. J. Salvi, J. Pages, and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern Recognit.* **37**(4), 827–849 (2004).
9. R. Yang and G. Welch, "Automatic and continuous projector display surface calibration using every-day imagery," in *Proc. Int. Conf. in Central Europe on Comput. Graph., Visual. and Comput. Vis.*, Union Agency, Plzen, Czech Republic (2001).
10. R. Raskar, "Oblique projector rendering on planar surfaces for a tracked user," in *ACM SIGGRAPH Conf. Abstr. Appl.*, pp. 260, ACM, New York, NY (1999).
11. T. H. Huang, C. T. Kao, and H. H. Chen, "Quality enhancement of a procam system by radiometric compensation," to appear in *Proc. 14th IEEE Int. Workshop on Multimedia Signal Process.*, pp. 192–197, IEEE Computer Society, Washington, DC (2012).
12. D. C. Kim et al., "Color correction for projected image on colored screen based on a camera," *Proc. SPIE* **7866**, 786606 (2011), .
13. M. H. Lee, H. Park, and J. I. Park, "Fast radiometric compensation accomplished by eliminating color mixing between projector and camera," *IEEE Trans. Consumer Electron.* **54**(3), 987–991 (2008).

14. S. Zollmann and O. Bimber, "Imperceptible calibration for radiometric compensation," in *Proce. Eurograph.*, pp. 61–64, EUROGRAPHICS Association, Goslar, Germany (2007).
15. G. Wetzstein and O. Bimber, "Radiometric compensation through inverse light transport," in *Proc. Pacific Conf. Comput. Graph. and Appl.*, pp. 391–399, IEEE Computer Society, Washington, DC (2007).
16. A. Grundhofer et al., "Dynamic adaptation of projected imperceptible codes," in *Proc. IEEE/ACM Int. Sym. on Mixed Augmented Reality*, pp. 1–10, IEEE Computer Society, Washington, DC (2007).
17. O. Bimber, A. Emmerling, and T. Klemmer, "Embedded entertainment with smart projectors," *J. Comps.* **38**(1), 48–55 (2005).
18. K. Fujii, M. D. Grossberg, and S. K. Nayar, "A projector-camera system with real-time photometric adaptation for dynamic environments," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, Vol. 1, pp. 814–821, IEEE Computer Society, Washington, DC (2005).
19. T. P. Koninckx et al., "Scene-adapted structured light," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 611–618, IEEE Computer Society, Washington, DC (2005).
20. O. Bimber et al., "Enabling view-dependent stereoscopic projection in real environments," in *Proc. IEEE/ACM Int. Sym. on Mixed Augmented Reality*, pp. 14–23, IEEE Computer Society, Washington, DC (2005).
21. M. D. Grossberg et al., "Making one object look like another: controlling appearance using a projector-camera system," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, Vol. 1, pp. I-452–I-459, IEEE Computer Society, Washington, DC (2004).
22. S. K. Nayar et al., "A projection system with radiometric compensation for screen imperfections," in *IEEE Int. Workshop on Projector-Camera Systems*, IEEE Computer Society, Washington, DC (2003).
23. T. Yoshida, C. Horii, and K. Sato, "A virtual colour reconstruction system for real heritage with light projection," in *Proc. Int. Conf. Virtual Systems and Multimedia*, pp. 161–168 (2003).
24. D. Caspi, N. Kiryati, and J. Shamir, "Range imaging with adaptive colour structured light," *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(5), 470–480 (1998).
25. A. Grundhoefer and O. Bimber, "Real-time adaptive radiometric compensation," *Trans. Visual. Comput. Graph.* **14**(1), 97–108 (2008).
26. H. Park et al., "Contrast enhancement in direct-projected augmented reality," in *IEEE Int. Conf. Multimedia and Expo*, pp. 1313–1316, IEEE Computer Society, Washington, DC (2006).
27. D. Wang et al., "Radiometric compensation in a projector-camera system based on the properties of human vision system," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 100, IEEE Computer Society, Washington, DC (2005).
28. M. Ashdown et al., "Perceptual photometric compensation for projected images," *IEICE Trans. Inf. Sys.* **J90-D**(8), 2115–2125 (2007).
29. M. Ashdown et al., "Robust content-dependent photometric projector compensation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. Workshop*, pp. 6, IEEE Computer Society, Washington, DC (2006).
30. O. Bimber et al., "The visual computing of projector-camera systems," in *ACM SIGGRAPH*, pp. 84-1–84-25, ACM, New York, NY (2008).
31. H. Park et al., "Radiometrically-compensated projection onto non-Lambertian surface using multiple overlapping projectors," in *Proc. Pacific-Rim Conf. Adv. in Image and Video Technol.*, pp. 534–544, Springer-Verlag, Berlin, Heidelberg (2006).
32. H. Park et al., "Specularity-free projection on nonplanar surface," in *Proc. Pacific-Rim Conf. Multimedia*, pp. 606–616, Springer-Verlag, Berlin, Heidelberg (2005).
33. H. Park et al., "Specular reflection elimination for projection-based augmented reality," in *Proc. IEEE/ACM Int. Sym. Mixed Augmented Reality*, pp. 194–195, IEEE Computer Society, Washington, DC (2005).
34. E. A. Merritt and M. E. P. Murphy, "A program for photorealistic molecular graphics," *Acta Crystallograph. D* **50**(6), 869–873 (1994).
35. C. M. Goral et al., "Modeling the interaction of light between diffuse surfaces," in *Proc. Conf. Comput. Graph. and Interactive Tech.*, pp. 213–222, ACM, New York, NY (1984).
36. R. L. Cook and K. E. Torrance, "A reflectance model for Comp. graphics," in *Proc. Conf. Comput. Graph. and Interactive Tech.*, pp. 307–316, ACM, New York, NY (1981).
37. J. F. Blinn, "Models of light reflection for Comp. synthesized pictures," *SIGGRAPH Comp. Graph.* **11**(2), 192–198 (1977).
38. B. T. Phong, "Illumination for Comp. generated pictures," *Commun. ACM* **18**(6), 311–317 (1975).
39. A. S. Glassner, *An Introduction to Ray Tracing*, Academic Press, Waltham, MA (1989).
40. R. Franzen, "Kodak Lossless True Color Image Suite," (30 October 2012), http://r0k.us/graphics/kodak/ (06 January 2013).
41. Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proc. 7th IEEE Int. Conf. on Comput. Vis.*, Vol. 1, pp. 666–673, IEEE Computer Society, Washington, DC (1999).

Chen-Tai Kao received a BS degree in electrical engineering from National Taiwan University in 2012. He is currently working toward an MS degree at the Graduate Institute of Communication Engineering, National Taiwan University. His research interests are in the area of perceptual-based image processing.

Tai-Hsiang Huang received his BS degree in electrical engineering from National Taiwan University in 2006. He is currently working toward a PhD degree at the Graduate Institute of Communication Engineering, National Taiwan University. His research interests are in the area of perceptual-based image and video processing.

Hua Lee prior to his return to UCSB in 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. His research interests cover the areas of imaging system optimization, high-performance image formation algorithms, synthetic-aperture radar and sonar systems, acoustic microscopy, microwave nondestructive evaluation, and dynamic vision systems. His research laboratory was the first to produce the holographic and tomographic reconstructions from a scanning laser acoustic microscope, and his research team is also known as the leader in pulse-echo microwave nondestructive evaluation of civil structures and materials.



Homer H. Chen received a PhD degree in electrical and computer engineering from University of Illinois at Urbana-Champaign. Since August 2003, he has been with the College of Electrical Engineering and Computer Science, National Taiwan University, where he is Irving T. Ho chair professor. Prior to that, he held various R&D management and engineering positions with U.S. companies over a period of 17 years, including AT&T Bell Labs, Rockwell Science Center, iVast, and Digital Island (acquired by Cable & Wireless). He was a US delegate for ISO and ITU standards committees and contributed to the development of many new interactive multimedia technologies that are now part of the MPEG-4 and JPEG-2000 standards. His professional interests lie in the broad area of multimedia signal processing and communications. He is an IEEE Fellow. He was an associate editor of *IEEE Transactions on Circuits and Systems for Video Technology* from 2004 to 2010, *IEEE Transactions on Image Processing* from 1992 to 1994, and *Pattern Recognition* from 1989 to 1999. He served as a guest editor for *IEEE Transactions on Circuits and Systems for Video Technology* in 1999 and *IEEE Transactions on Multimedia* in 2011.