In: Proc ICCV (2001)

Deriving intrinsic images from image sequences

Yair Weiss
Computer Science Division
UC Berkeley
Berkeley, CA 94720-1776
yweiss@cs.berkeley.edu

Abstract

Intrinsic images are a useful midlevel description of scenes proposed by Barrow and Tenenbaum [1]. An image is decomposed into two images: a reflectance image and an illumination image. Finding such a decomposition remains a difficult problem in computer vision. Here we focus on a slightly easier problem: given a sequence of T images where the reflectance is constant and the illumination changes, can we recover T illumination images and a single reflectance image? We show that this problem is still ill-posed and suggest approaching it as a maximum-likelihood estimation problem. Following recent work on the statistics of natural images, we use a prior that assumes that illumination images will give rise to sparse filter outputs. We show that this leads to a simple, novel algorithm for recovering reflectance images. We illustrate the algorithm's performance on real and synthetic image sequences.

1 Introduction

Barrow and Tenenbaum (1978) introduced the term "intrinsic images" to refer to a midlevel decomposition of the sort depicted in figure 1. The observed image is a product of two images: an illumination image and a reflectance image. We call this a midlevel description because it falls short of a full, 3D description of the scene:the intrinsic images are viewpoint dependent and the physical causes of changes in illumination at different points are not made explicit (e.g. the cast shadow versus the attached shadows in figure 1c).

Barrow and Tenenbaum argued that such a midlevel description, depsite not making explicit all the physical causes of image features, can be extremely useful for supporting a range of visual inferences. For example, the task of segmentation may be poorly defined on the input image and many segmentation algorithms make use of arbitrary thresholds in order to avoid being fooled by illumination changes. On the intrinsic, reflectance image, on the other hand, even primitive segmentation algorithms would correctly segment the cylinder as a single segment in figure 1b. Similarly, view-based template matching and shape-from-shading would be significantly less brittle if they could work on the intrinsic image representation rather than on the input image.

Recovering two intrinsic images from a single input image remains a difficult problem for computer vision systems. This is a classic ill-posed problem: the number of unknowns is twice the number of equations. Denoting by $I(x, y)$ the input image and by $R(x, y)$ the reflectance image and $L(x, y)$ the illumination image, the three images are related by:

$$I(x, y) = L(x, y)R(x, y) \tag{1}$$

Obviously, one can always set $L(x, y) = 1$ and satisfy the equations by setting $R(x, y) = I(x, y)$. Despite this difficulty, some progress has been made towards achieving this decomposition. Land and McCann's Retinex algorithm [7] could successfully decompose scenes in which the reflectance image was piecewise constant. This algorithm has been continuously extended over the years (e.g. [4]). More recently, Freeman and Viola [3] have used a wavelet prior to classify images into one of two classes: all reflectance or all illumination.

In this paper we focus on a slightly easier version of the problem. Given a sequence of T images $\{I(x, y, t)\}_{t=1}^{T}$ in which the reflectance is constant over time and only the illumination changes, can we then solve for a single reflectance image $R(x, y)$ and T illumination images $\{L(x, y, t)\}_{t=1}^{T}$? Our work was motivated by the ubiquity of image sequences of this form on the world wide web: "webcam" pictures from outdoor scenes as in figure 2. The camera is stationary and the scene is mostly stationary: the predominant change in the sequence are changes in illumination.

While this problem seems easier, it is still completely ill-posed: at every pixel there are T equations and $T + 1$ unknowns. Again, one can simply set $L(x, y) = 1$ and $R(x, y, t) = I(x, y, t)$ and fully satisfy the equations. Obviously, some additional constraints are needed.

One version of this problem that has received much attention is the case when $L(x, y, t)$ are attached shadows of a single, convex, lambertian surface: $L(x, y, t) = \vec{N}(x, y) \cdot$

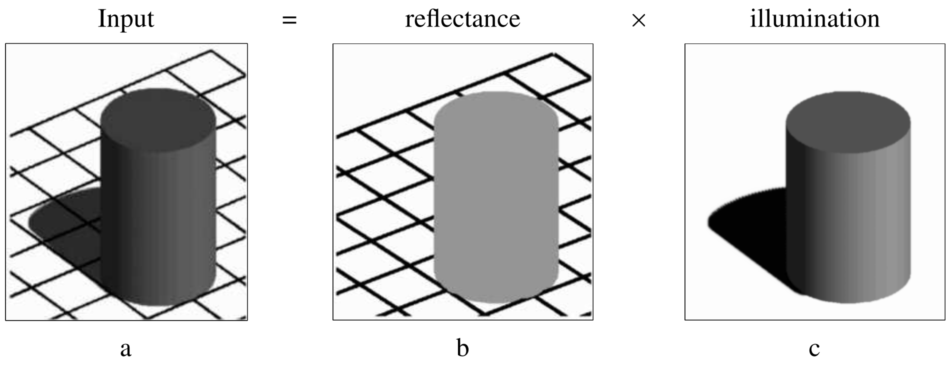


Figure 1: The intrinsic image decomposition [1].

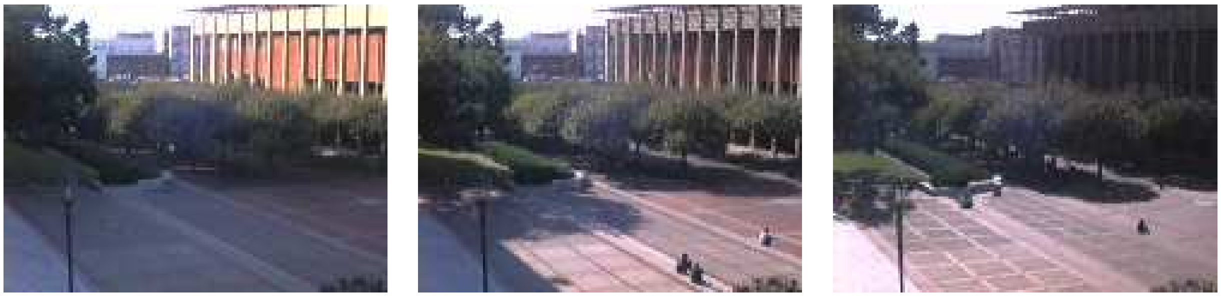


Figure 2: Images from a "webcam" at www.berkeley.edu/webcams/sproul.html. Most of the changes are changes in illumination. Can we use such image sequences to derive intrinsic images?

$\vec{S}(t)$, with $N(x, y)$ the surface normal at x, y and $\vec{S}$ a vector in the direction of the light source. This is the photometric stereo problem with unknown light-source, and can be solved using SVD techniques up to a Generalized Bas Relief ambiguity [12, 6].

Farid and Adelson addressed a special case of this problem [2]. They assumed that $L(x, y, t) = \alpha(t)L(x, y)$, i.e. that all illumination images are related by a scalar. They used independent component analysis to solve for $L(x, y)$ and $R(x, y)$.

A similar problem has recently been addressed by Szeliski, Avidan and Anandan [11]. They dealt with additive transparency sequences so that $I(x, y, t) = R(x, y) + L(x - tv_x, y - tv_y, t)$. Of course, if we exponentiate both sides we get equation 1 with an additional constraint that the illumination images are warped images of a single illumination image. They showed that since $L(x, y, t)$ is bounded below (i.e. that $L(x, y, t)$ is positive), setting $\hat{R}(x, y) = \min_t I(x, y, t)$ can give a good estimate for the reflectance. They used the min filter as an initialization for a second estimation procedure that estimated the motion of $L(x, y)$ and improved the estimate.

In this paper we take a different approach. We formulate the problem as a maximum-likelihood estimation problem based on the assumption that derivative-like filter outputs applied to L will tend to be sparse. We derive the ML estimator under this assumption and show that it gives a simple, novel algorithm for recovering reflectance. We illustrate the algorithm's performance on real and synthetic image sequences.

2 ML estimator assuming sparseness

For convenience, we work here in the log domain. Denote by $i(x, y), r(x, y), l(x, y)$ the logarithms of the input, reflectance and illumination images. We are given: $i(x, y, t) = r(x, y) + l(x, y, t)$ and wish to recover $r(x, y)$ and $l(x, y, t)$.

To make the problem solvable, we want to assume a distribution over $l(x, y, t)$. Our first thought was to make a similar assumption to that made in the Retinex work: that illumination images are lower contrast than relfectance images. We found, however, that while this may hold true in the Mondrian world studied by Land and McCann, it is rarely true for the outdoor scenes of the type shown in figure 2. Edges due to illumination often have as high a contrast as those due to reflectance changes.

We use a weaker, more generic prior that is motivated by recent work on the statistics of natural images. A remarkably robust property of natural images that has received

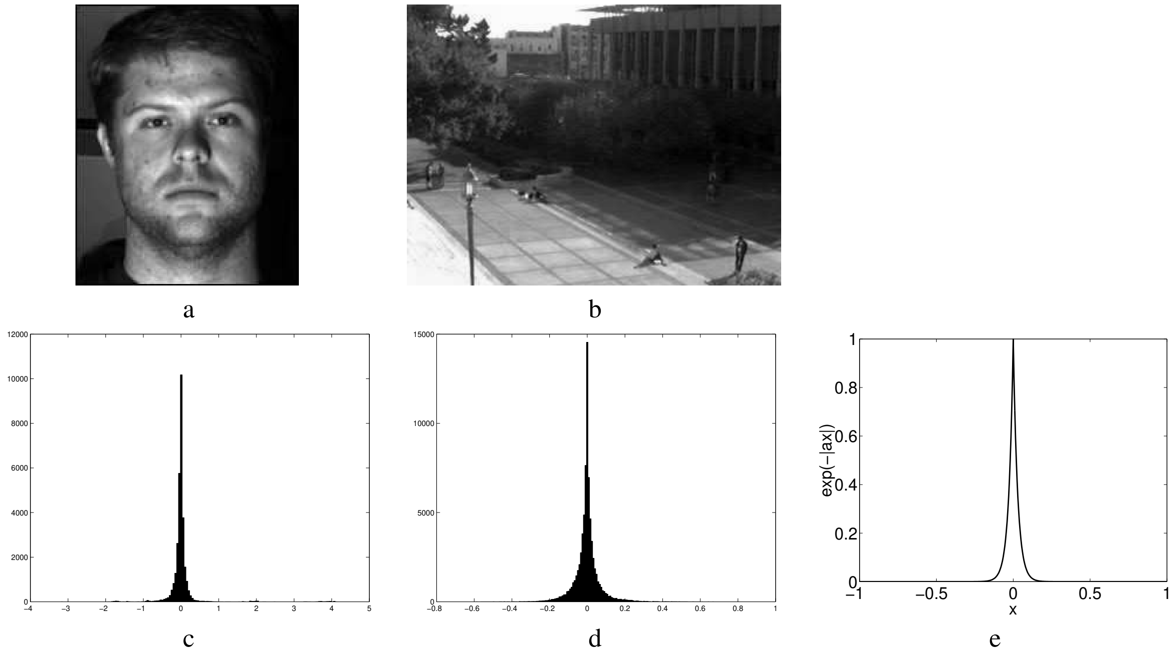


Figure 3: We use a prior motivated by recent work on the statistics of natural scenes. Derivative filter outputs tend to be sparse for a wide range of images. **a-b** images **:** **c-d** histograms of horizontal derivative filter outputs. **e** A Laplcian distribution. Note the similar shape to the observed histograms.

much attention lately is the fact that when derivative filters are applied to natural images, the filter outputs tend to be sparse [8, 10]. Figure 3 illustrates this fact: the image of the face and the outdoor scene have similar histograms that are peaked at zero and fall off much faster than a Gaussian. This property is robust enough that it continues to hold if we apply a pixelwise log function to each image (the histograms shown are actually for the log images). These prototypical histograms can be well fit by a Laplacian distribution $P(x) = \frac{1}{Z}e^{-\alpha|x|}$ (although better fits are obtained with richer models). Figure 3e shows a Laplacian distribution.

We will therefore assume that when derivative filters are applied to $l(x, y, t)$ the resulting filter outputs are sparse: more exactly, we will assume the filter outputs are independent over space and time and have a Laplacian density. Assume we have N filters $\{f_n\}$ we denote the filter outputs by $o_n(x, y, t) = i \star f_n$. We use r_n to denote the reflectance image filtered by the nth filter $r_n = r \star f_n$.

Claim 1: Assume filter outputs applied to $l(x, y, t)$ are Laplacian distributed and indpendent over space and time. Then the ML estimate of the filtered reflectance image $\hat{r}_n$ are given by:

$$\hat{r}_n(x, y) = \mathsf{median}_t o_n(x, y, t) \qquad (2)$$

Proof: Assuming Laplacian densities and independence yields the likelihood:

$$P(o_n|r_n) = \frac{1}{Z}\prod_{x,y,t} e^{-\beta|o_n(x,y,t)-r_n(x,y)|} \qquad (3)$$

$$= \frac{1}{Z}e^{-\beta\sum_{x,y,t}|o_n(x,y,t)-r_n(x,y)|} \qquad (4)$$

Maximizing the likelihood is equivalent to minimizing the sum of absolute deviations from $o_n(x, y, t)$. The sum of absolute values (or l_1 norm) is minimized by the median. $\square$.

Claim 1 gives us the ML estimate for the filtered reflectance images $\hat{r}_n$. To recover r, the estimated reflectance function, we solve the overconstrained systems of linear equations:

$$f_n \star \hat{r} = \hat{r}_n \qquad (5)$$

It can be shown that the psuedo-inverse solution is given by:

$$\hat{r} = g \star \left(\sum_n f_n^r \star \hat{r}_n\right) \qquad (6)$$

with f_n^r the reversed filter of f_n: $f_n(x, y) = f_n^r(-x, -y)$ and g a solution to:

$$g \star \left(\sum_n f_n^r \star f_n\right) = \delta \qquad (7)$$

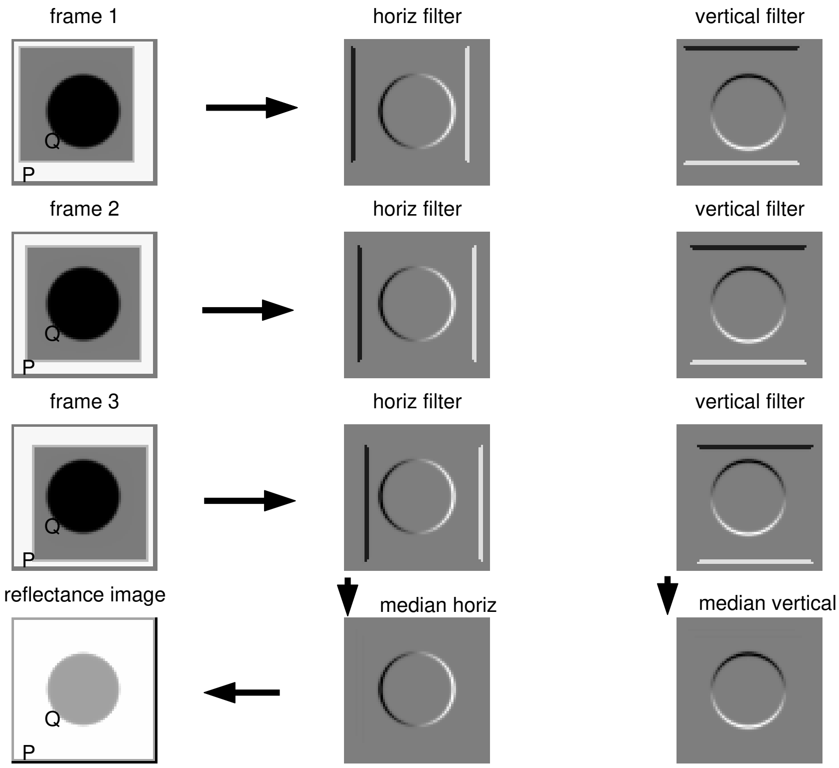


Figure 4: An illustration of the ML estimation algorithm

Note that g is indpendent of the image sequence being considered and can be computed in advance.

Figure 4 illustrates the algorithm for calculating the ML reflectance function. The three frames in the leftmost column show a circle illuminated with a square cast shadow. The shadow is moving over time. The middle and right columns show the horizontal and vertical filter outputs applied to this sequence. Taking a pixelwise median over time gives the estimated filtered reflectance images in the bottom row. Finally, applying equation 6 gives the ML estimated reflectance function. Once we have an estimate for $r(x, y)$ we can estimate $l(x, y, t) = i(x, y, t) - r(x, y)$.

The ML estimate has some similarities to the temporal median filter that is often used in accumulating mosaics from image sequences (e.g. [9]) but has very different performance characteristics. In figure 4 taking the temporal median of the three frames in the left column would *not* give the right reflectance function. A pixel whose intensity is bright in all frames (e.g. pixel P in figure 4a) and a pixel whose intensity is dark in all frames (e.g. pixel Q in figure 4a) must have different medians. Thus, a pixel whose intensity is always dark must be estimated as having a different reflectance that a pixel whose intensity is always light.

In the ML estimate, on the other hand, this is not the case. Note that pixels P and Q are estimated as having the same reflectance even though one was always lighter than the other. This is because the ML estimate performs a temporal median on the *filtered* images, not the original images. In terms of probabilistic modeling, a temporal median filter on images is the ML estimate if we assumed that $l(x, y, t)$ is sparse, i.e. that most pixels in $l(x, y, t)$ are close to zero. This is rarely true for natural images. In contrast, here we are assuming that the *filter outputs* applied to $l(x, y, t)$ are sparse, an assumption that often holds for natural images.

What if $f_n \star l(x, y, t)$ does not have exactly a Laplacian distribution? The following claim shows that the exact form of the distribution is not important as long as the filter outputs are sparse.

Claim 2: Let $p_\epsilon = P(|f_i \star l(x, y, t)| < \epsilon)$. Then the estimated filtered reflectances are within ϵ of the true filtered reflectances with probability at least:

$$\sum_{k=1}^{T/2} \left(\begin{array}{c} T \\ k \end{array} \right) (1 - p_\epsilon)^{T-k} p_\epsilon^k$$

Proof: If more than 50% of the samples of $f_n \star l(x, y, t)$ are within ϵ of some value, then by the definition of the median, the median must be within ϵ of that value. The claim follows from the binomial formula for the sum of T independent events. $\square$

Claim 2 does not require that the illumination images have a Laplacian distribution in their filter outputs, rather the more sparse the filter outputs are the quicker the median estimate will converge to the true filtered reflectance function. For example, the lighting image in figure 1c, does not have a Laplacian distribution but is very sparse: 85% of the filter outputs have magnitude less than $\epsilon = 1\%$ of the maximal magnitude. Claim 2 guarantees the $|\hat{r}_n - r_n| < \epsilon$ with probability at least 0.93 given only three frames and with probability at least 0.97 given five frames.

3 Results - synthetic sequences

All the results shown in this paper used just two filters: horizontal and vertical derivative filters.

Figure 5 show the first and last frames from a sequence in which a square cast shadow is moving over a circle and ellipse. Note that the circle is always in shadow and that the ellipse is always half in shadow. Despite this, the ML reflectance estimate correctly gets rid of the shadows on both. For comparison, figure 5 also shows the temporal min filter and temporal mean filter. Both of these approaches suffer from the fact that if a pixel is always in shadow, the estimated reflectance is also darker.

Figure 6 shows the first and last frames from a synthetic additive transparency sequence. The image of Reagan is constant for all frames and we added an image of Einstein that is moving diagonally (speed 4 pixels per frame). Figure 6 also shows the recovered Reagan and Einstein images. They are indistinguishable from the original image. For comparison, figures 6 also shows the min and median filters. Again, the min filter assumes that all pixels at some time see a black pixel from the Einstein image. Since that assumption does not hold, the estimate is quite bad.

4 Results - real sequences

Figure 7 shows two frames from a sequence that is part of the Yale face database B [5]. A strobe light at 64 different locations illuminated a person's face, giving 64 images with changing illumination. The images were taken with a camera with linear response function. Figure 7 also shows the estimated reflectance and illumination images. Note that nearly all the specular highlights are gone although there are still some highlights at the tip of the nose. Although it is hard to measure performance in this task, observers describe the illumination images as "looking like marble statues", as would be expected from an illumination image of a face.

Figure 8 shows two frames from a sequence taken from the "WebCam" at UC Berkeley. We used 35 images taken at different times. Figure 8 also shows the estimated reflectance and illumination images. Note that the cast shadows by the trees and buildings are mostly gone. Note also that we did not have to do anything special to get rid of the people in the images: since we use a generic prior for the lighting images we can easily accomodate changes that are not purely due to lighting. The results are not as good as the Yale sequence, and we believe this is partially due to the automatic gain control on the web camera so that the response function is far from linear.

Even these results can be good enough for some applications. Figure 9a shows a color scene of Sproul plaza in Berkeley. Suppose a malicius Stanford hacker wanted to insert the Stanford logo on the plaza. Figure 9b shows what happens when α blending is used to composite the logo and the image. The result is noticeably fake. Figure 9c shows the result when α blending is used on the estimated reflectance image, and the image is rerendered with the estimated illumination image. The logo appears to be part of the scene.

5 Discussion

Deriving intrinsic images from a single image remains a difficult problem for computer vision systems. Here we have focused on a slightly easier problem: recovering intrinsic images from an image sequence in which the illumination varies but the reflectance is constant. We showed that the problem is still ill-posed and suggested adding a statistical assumption based on recent work in statistics of natural scenes: that derivative filter outputs on the illumination image will tend to be sparse. We showed that this assumption leads to a novel, simple algorithm for reflectance recovery and showed encouraging results on a number of image sequences.

Both the camera model and the statistical assumptions we have used can be extended. We assumed a linear response in the camera model and one can derive the ML estimator when the camera is nonlinear. We have also assumed that filter outputs are independent across space and time. It would be interesting to derive ML estimators when the dependency is taken into account. We hope that progress in statistical modeling of illumination images will enable us to tackle the original problem posed by Barrow and Tenenbaum: recovering intrinsic images from a single image.

Acknowledgements

I thank A. Efros for pointing out the utility of webcams for computer vision. Supported by MURI-ARO-DAAH04-96-1-0341.

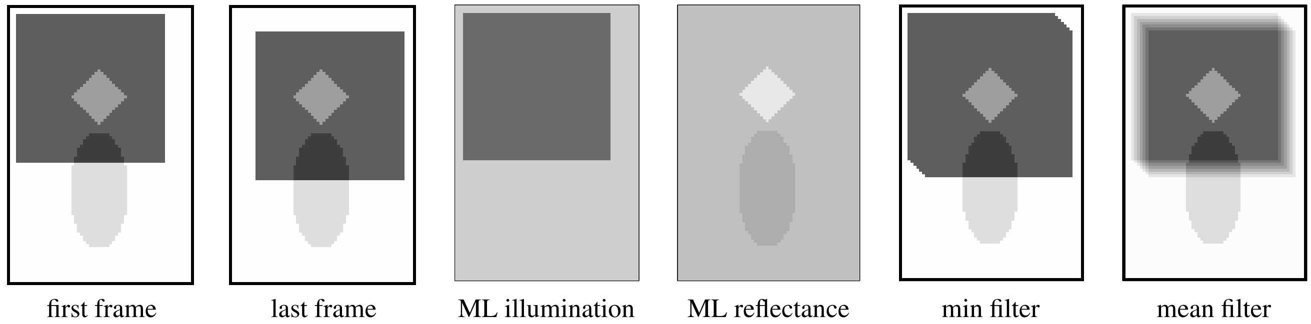


Figure 5: A synthetic sequence in which a square cast shadow translates diagonally. Note that the pixels surrounding the diamond are always in shadow, yet their estimated reflectance is the same as that of pixels that were always in light. In the min and mean filters, this is not the case and the estimated reflectances are quite wrong.

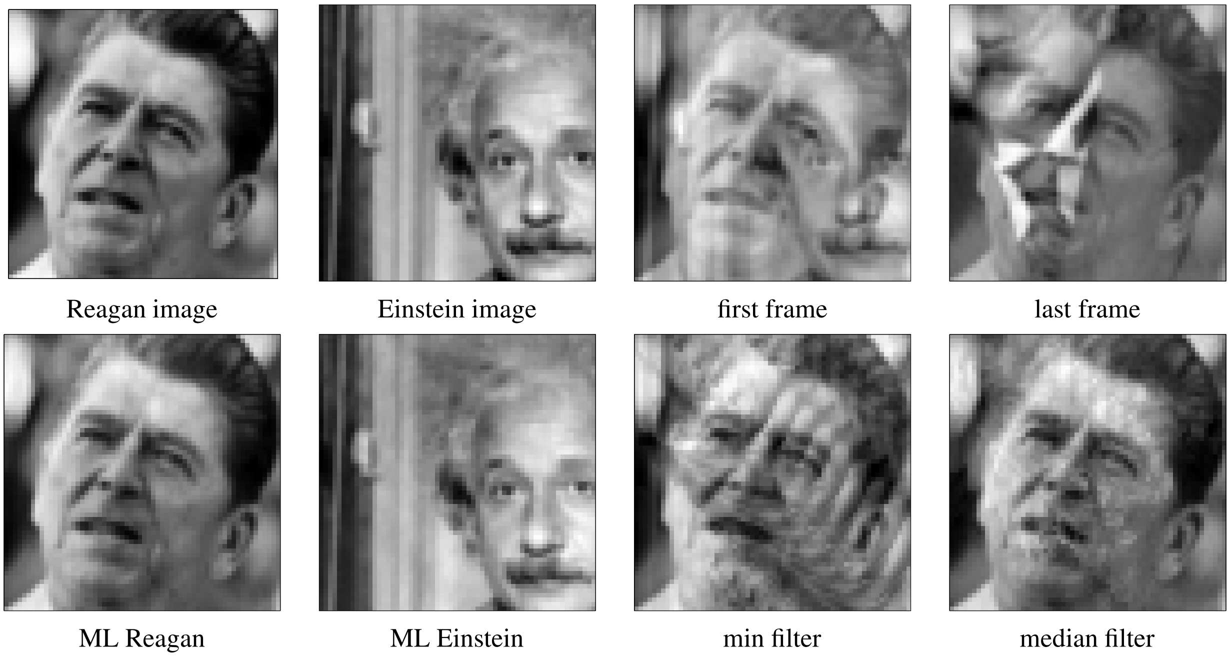


Figure 6: Results on a synthetic additive transparency sequence. The Einstein image is translated diagonally with speed 4 pixels per frame and added to the Reagan image. The ML estimates are nearly exact while the min and median filters are not.

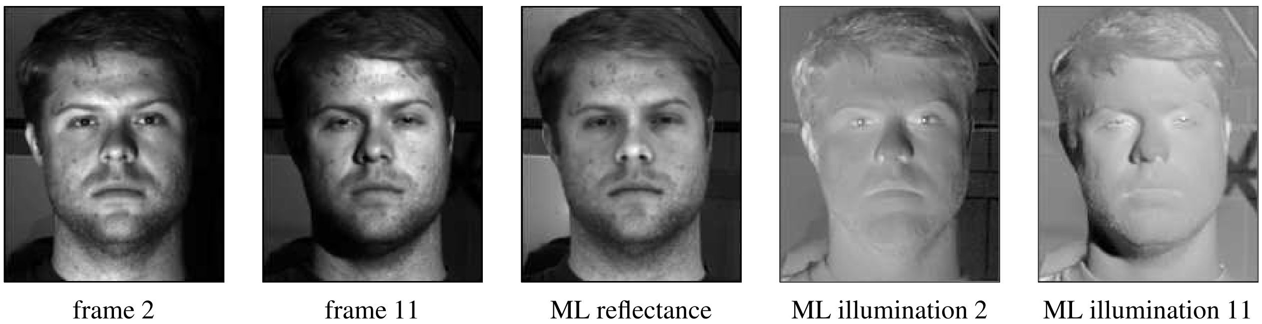
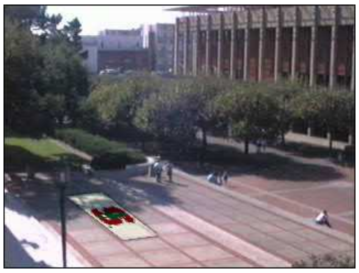


Figure 7: Results on one face from the Yale Face Database B [5]. There were 64 images taken with variable lighting. Note that the recovered reflectance image is almost free of specularities and is free of cast shadows. The ML illumination images are shown with a logarithmic nonlinearity to increase dynamic range.

frame 1 frame 11 ML reflectance

ML illumination 1 ML illumination 2

Figure 8: Results on a webcam sequence from: www.berkeley.edu/webcams/sproul.html. There were 35 images that varied mostly in illumination. Note that the ML reflectance image is free of cast shadows. The illumination images are shown with a logarithmic nonlinearity to increase dynamic range.



a b c

Figure 9: Intrinsic images are useful for image manipulation. **a.** The original image of Sproul plaza in Berkeley. **b.** The Stanford logo is α blended with the image: the result is noticeable fake. **c.** The Stanford logo is α blended in the reflectance image and then rendered with the derived illumination image.

References

[1] H.G. Barrow and J.M. Tenenbaum. Recovering intrinsic scene characteristics from images. In A. Hanson and E. Riseman, editors, *Computer Vision Systems*. Academic Press, 1978.

[2] H. Farid and E.H. Adelson. Separating reflections from images by use of independent components analysis. *Journal of the optical society of america*, 16(9):2136–2145, 1999.

[3] W.T. Freeman and P.A. Viola. Bayesian model of surface perception. In *Adv. Neural Information Processing Systems 10*. MIT Press, 1998.

[4] B. V. Funt, M. S. Drew, and M. Brockington. Recovering shading from color images. In G. Sandini, editor, *Proceedings: Second European Conference on Computer Vision*, pages 124–132. Springer-Verlag, 1992.

[5] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Generative models for recognition under variable pose and illumination. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 277–284, 2000.

[6] K. Hayakawa. Photometric stereo under a light source with arbitrary motions. *Journal of the optical society of America*, 11(11), 1994.

[7] E. H. Land and J. J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 1971.

[8] B.A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–608, 1996.

[9] H.Y. Shum and R. Szeliski. Construction and refinement of panoramic mosaics with global and local alignment. In *Proceedings IEEE CVPR*, pages 953–958, 1998.

[10] E.P. Simoncelli. Statistical models for images:compression restoration and synthesis. In *Proc Asilomar Conference on Signals, Systems and Computers*, pages 673–678, 1997.

[11] R. Szeliksi, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *Proceedings IEEE CVPR*, 2000.

[12] A.L. Yuille, D. Snow, R. Epstein, and P.N. Belhumeur. Determining generative models of objects under varying illumination: shape and albedo from multiple images using SVD and integrability. *International Journal of Computer Vision*, 35(3):203–222, 1999.