MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Learning local evidence for shading and reflectance

Matt Bell, William T. Freeman

TR2001-04 December 2001

Abstract

A fundamental, unsolved vision problem is to distinguish image intensity variations caused by surface normal variations from those caused by reflectance changes–ie, to tell shading from paint. A solution to this problem is necessary for machines to interpret images as people do and could have many applications. The labelling allows us to reconstruct bandpassed images containing only those parts of the input image caused by shading effects, and a separate image containing only those parts caused by reflectance changes. The resulting classifications compare well with human psychophysical performance on a test set of images, and show good results for test photographs.

MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

Learning local evidence for shading and reflectance

Matt Bell* and William T. Freeman
Mitsubishi Electric Research Labs (MERL)
201 Broadway
Cambridge, MA 02139

TR-2001-04 January 2001

Abstract

A fundamental, unsolved vision problem is to distinguish image intensity variations caused by surface normal variations from those caused by reflectance changes–ie, to tell shading from paint. A solution to this problem is necessary for machines to interpret images as people do and could have many applications.

We take a learning-based approach. We generate a training set of synthetic images containing both shading and reflectance variations. We label the interpretations by indicating which coefficients in a steerable pyramid representation of the image were caused by shading and which by paint.

To analyze local image evidence for shading or reflectance, we study the outputs of two layers of filters, each followed by rectification. We fit a probability density model to the filter outputs using a mixture of factor analyzers. The resulting model indicates the probability, based on local image evidence, that a pyramid coefficient at any orientation and scale was caused by shading or by reflectance variations. We take the lighting direction to be that which generates the most shape-like labelling.

The labelling allows us to reconstruct bandpassed images containing only those parts of the input image caused by shading effects, and a separate image containing only those parts caused by reflectance changes. The resulting classifications compare well with human psychophysical performance on a test set of images, and show good results for test photographs.

1. First printing, TR2001-04, January, 2001

* Matt Bell's present address: Psychology Dept., Stanford University, Stanford, CA 94305

Learning local evidence for shading and reflectance

Matt Bell* and William T. Freeman
Mitsubishi Electric Research Labs (MERL)
201 Broadway
Cambridge, MA 02139
MERL-TR2001-04, January, 2001

Abstract

A fundamental, unsolved vision problem is to distinguish image intensity variations caused by surface normal variations from those caused by reflectance changes–ie, to tell shading from paint. A solution to this problem is necessary for machines to interpret images as people do and could have many applications.

We take a learning-based approach. We generate a training set of synthetic images containing both shading and reflectance variations. We label the interpretations by indicating which coefficients in a steerable pyramid representation of the image were caused by shading and which by paint.

To analyze local image evidence for shading or reflectance, we study the outputs of two layers of filters, each followed by rectification. We fit a probability density model to the filter outputs using a mixture of factor analyzers. The resulting model indicates the probability, based on local image evidence, that a pyramid coefficient at any orientation and scale was caused by shading or by reflectance variations. We take the lighting direction to be that which generates the most shape-like labelling.

The labelling allows us to reconstruct bandpassed images containing only those parts of the input image caused by shading effects, and a separate image containing only those parts caused by reflectance changes. The resulting classifications compare well with human psychophysical performance on a test set of images, and show good results for test photographs.

1. Introduction

A fundamental problem in image interpretation is to identify the cause of intensity variations in the image. Humans can easily look at a photograph and identify which parts of the image are due to shading, which are due to reflectance variations, occluding contours, etc. These assignments are crucial for proper interpretation of the image. For example, shape-from-shading algorithms typically assume all intensity changes are due to surface normal changes, and reconstruct spurious shapes when confronted with reflectance changes. Here, we restrict ourselves to distinguishing shading from paint.

Figure 1 (a) illustrates the problem. Some of the image intensity changes are caused by the graffiti paint; others of the intensity variations are caused by the shape of the rock on which the paint was sprayed. Some locations show both effects. (b) shows the same location a few months later, after an attempt was made to enforce a uniform reflectance over the rock. It is simple to see the underlying shape of (b) in the image (a); we want to develop a computer program to do the same thing.

This problem has not yet been solved for real images. Sinha and Adelson [11] solved the problem in a blocks world domain, based on heuristic rules over a set of junctions and contours, which were pre-identified by hand. Like other blocks world vision solutions, this hasn't led to an analogous solution for real images.

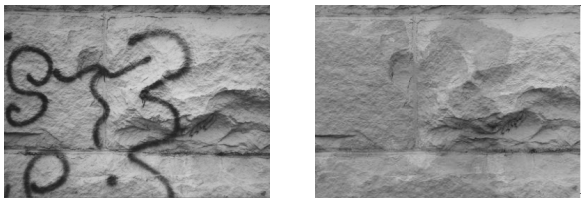
Freeman and Viola [4] proposed a prior probability for shapes which penalized the elaborate shapes that were required to explain images made by reflectance changes. Their method assumed each image was either all shading or all paint and couldn't process an image containing both shading and reflectance changes.

Freeman, Pasztor, and Carmichael [3] used a Markov network to solve for the shape and reflectance combinations which best explain the input image data, using a training set of labelled images. However, their method required prestoring all possible shape and reflecance interpretations for any patch. This caused the conjectured scene to be a poor fit to the observed image, limiting the applicability of the method.

To date, there is no adequate solution which identifies which components of an image are caused by shading variations, and which are caused by reflectance variations.

Our approach is training-based, like that of [3], but we use representations that make good solutions feasible. We represent the image data using multi-layer filter energy models, which allows better generalization over inputs than

*Present address: Psychology Dept., Stanford University, Stanford, CA 94305



(a) (b)

Figure 1: (a) Example image of the sort that we would like to separate shading and reflectance information. People can trivially make the separation, yet it is a difficult task for the computer. (b) Graffiti removal by the local town reveals the underlying shape on which the paint was applied.

a pixel representation. We identify shading by assigning labels to image pyramid coefficients, which allows the proper interpretation in image regions where both shading and reflectance events occur.

Because the interpretation in some regions, such as isolated contours, will be ambiguous, a full solution will follow the local analysis with a propagation stage. However, we show here that even the initial local analysis can go far in disambiguating shading from paint.

2. Training set

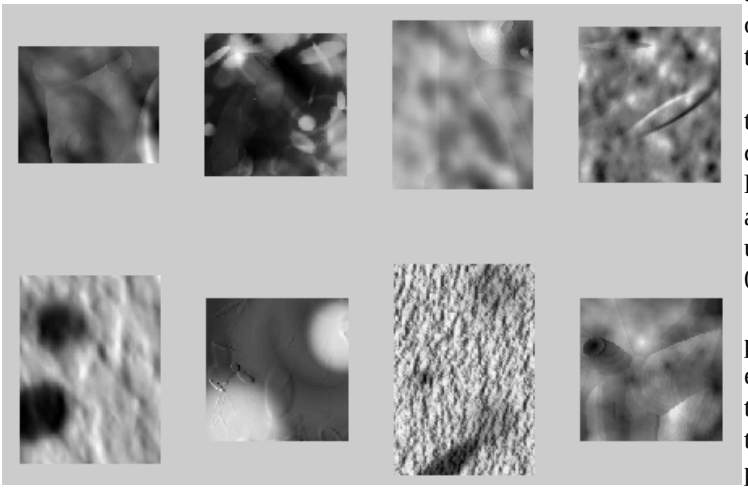


Figure 2: Example of training images for mixed shape and reflectance classification.

We generate test images in the following way. We first generated test images of all shading or all reflectance. Some of these were made by a fractal process, using the mid-point displacement procedure. Others were generated by summing randomly placed ellipses of randomized position, orientation, size, and eccentricity in an image. Fig. 2 shows a sampling of the all-shading and all-reflectance images. We shaded the shapes by assuming Lambertian surfaces, with the light direction fixed, from the left, with a mid-gray constant added to make all the light intensities positive. (We will rotate the input image to generalize to other assumed light directions).

Not all products of shading and reflectance images are easily interpretable by people. We used visual interpretability as a guide in synthesizing the labelled training set. We found that a paint image could be multiplied by a rendered shape image to yield an interpretable result provided that at least one of the two images was very sparse, for example, was generated from a small number of ellipses.

3. Representation and labelling

How should we label the shading and reflectance components of an image? In many images, such as Fig. 4, a given spatial localtion can have intensity changes caused by both shading and reflectance effects in the image. A labelling of shading or reflectance at each pixel would not be adequate.

Instead, we provide a label for each possible orientation, scale, and position. We represent the input image as a steerable pyramid [10] and provide one label for each coefficient of the steerable pyramid. (We chose the steerable pyramid representation because it is self-inverting, and the image subbands are not aliased.) This can be thought of as the generalization, for a signal processing approach, of the assumption in [11] that each intensity edge can have only one cause. (No paint and shape changes were allowed at the same contour).

To label these images, we made pyramids from each of the multiplicand images, and set the label of each coefficient in the product image to be that of modality with the larger coefficient of the multiplicand images. We omitted ambiguous classification points: if the ratio of absolute values of the pyramid coefficients for each class was between 0.2 and 0.8, we omitted that point from the training set.

Multiplication of the source images causes the steerable pyramid coefficients to interact in a non-linear way. However, we found that for our sparse images, this approximation was adequate. Figure 4 (a) shows a test image and (d) the corresponding shading/reflectance labelling of steerable pyramid coefficients.

4. Local evidence

Influenced by the success of [2, 9, 6], we use a cascade of local filters to represent the local image area in a way that allows generalization from the training data. It is thought [9] that such representations generalize better over image intensities and over slight variations in relative positions than a pixel representation does. There is evidence from psychophysics research [12] that such cascades may exist in the human visual system.

We apply the filters of Fig. 3 in a cascade, also used by [13] for image retrieval applications. These filters are simple combinations of first and second derivatives. We first form three copies of the original image, downsampled by 0, 1, and 2 factors of two in each dimension. Then, to each of those images, we apply each of the filters, and take the absolute value of the output. Then we subsample and apply each filter again to the magnitude of the filter responses. Three starting images, times 25 initial filters, times 25 second filterings gives a 1875 dimensional feature vector, with each dimension corresponding to a different combination of initial subsamplings and first and second level filterings.

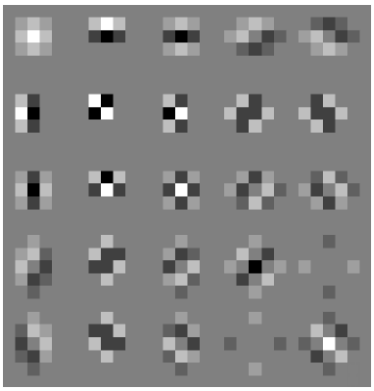


Figure 3: Filters used in cascaded energy model of local image region.

Given the cascaded filter responses, we want to assign a probability that each different image pyramid coefficient corresponds to the shading or reflectance label.

(Note there are two different sets of filters–the cascaded analysis filters and the steerable pyramid filters. The first are small and fast to compute. The pyramid filters are larger and designed to allow exact reconstruction of the input image. While we could do all the processing using only the pyramid filters, we use the smaller and separable second set of filters for speed).

Once these dimensions are sampled, two normalization procedures are performed. First, the value of each dimension, is scaled so that its average value over all training points, is 1. The second normalization scales the feature vector for each training point so that the average value is 1. The purpose of the second normalization is to normalize for contrast differences; thus, the training done in an area of a certain contrast can be applied to classifying areas with a wide variety of contrasts. The first normalization is designed to equalize the importance of each dimension so that the second normalization does not favor certain dimensions. When new images are being classified, the same normalizations are applied.

Many of the dimensions of our 1875-d feature vector are redundant, although we do not know ahead of time which dimensions those will be. We prune them to a more manageable set in the following way: First, we compute the correlation between each of the dimensions in the training data. Then, we delete one of the pair of dimensions with the highest absolute correlation. This deletion process is iterated until we have pared down the dataset to 150 dimensional feature vectors $\vec{d}$. This must be done so that the mixture of factor analyzers algorithm can run in a reasonable amount of time.

We modelled the probability density for each class label (shading and reflectance) by a mixture of gaussians [1], using the mixture of factor analyzers approach of [5]. By evaluating the resulting categorization accuracy using cross-validation with the labelled training data, we chose to use 10 gaussians in the mixture, and 4 dimensions in each gaussian. We fit one such mixture for each class label, l, and for each orientation, o, and scale, s of steerable pyramid coefficient:

$$P_{l,o,s}(\vec{d}) = \frac{1}{Z}\sum_{i=1}^{10} \alpha_i N(\vec{d}; \vec{d}_i, \Lambda_i), \qquad (1)$$

$N(\vec{d}; \vec{d}_i, \Lambda_i)$ is a gaussian in $\vec{d}$ of mean $\vec{d}_i$ and covariance Λ_i, α_i is the weight of the ith mixture gaussian, and Z is a normalization constant,

Given an image to analyze, we apply the layered filters, rectifying the output at each level. We select the 150 outputs determined in the training set to be most uncorrelated with each other, and evaluate the probability density for each class labelling (shading or paint) at each orientation and scale of pyramid coefficients, using the learned densities, $P_{l,o,s}(\vec{d})$. We compare the shading and paint probabilities, and assign the label of the larger density. (We do not assign confidence measures, or indicate ties).

We illustrate an initial test on our labelling method by applying it to an image generated by the same program as generated the training set images, although this image itself was not in the training set. This is thus a typical image of the set described by the training data, yet a novel image. Figure 4 (a) is the test image. (d) is the labels of the steerable pyramid coefficients (dark means paint, light means shape). Using these labels, one can reconstruct the ideal result bandpass paint (b) and shading (c) images.

The learning-based algorithm's performance is not perfect, as above, but is good. (e) through (g) show the class labels on the 4 orientations of the steerable pyramid at the three spatial scales (black is paint; white is shape). We use the class labels to mask out only those pyramid coefficients estimated to be due to paint, then reconstruct the image from those pyramid coefficients to give (h). Doing the same for the coefficients labeled as shape gives (i). Because we are only labelling bandpass pyramid coefficients, these

images have no DC component. The reconstructed images compare well with the best possible bandpassed reconstructions, (b) and (c), showing that we can separate shading and reflectance effects in images similar to the training set images.

4.1 Unknown lighting direction

To a first approximation, the lighting direction that humans perceive can be summarized by the azimuthal angle, in the plane of the image, that the light arrives from (see discussion in [7]). We make that assumption and allow for unknown lighting direction by rotating the image through different angles before applying the labeling algorithm, which is only trained on images with the light arriving from the left. While the majority of reflectance changes are likely to be classified as reflectance changes no matter what direction they are facing, the algorithm has difficulty properly recognizing shape changes when they are produced by a lighting direction that is not from the left or right. As a result, the rotated image that produces the most shapelike classification is presumed to have the lighting from the left or right, relative to the orientation of the rotated image. It is this image's classification that is chosen as the final classification.

We determine an index of shapeness by computing the ratio of total variance of the extracted shape image to the total variance of the extracted reflectance image.

5. Results

5.1 Application to psychophysics database

We applied the labelling to a random subset of the images of the psychophysics test. The Spearman rank ordering correlation [8] between the mean of the subjects' rankings and that of the algorithm was 0.46. at a significance level of 0.0027. The Spearman between differrent human subjects ranged between 0.32 and 0.9; the algorithm's agreement with the subjects' rankings was within that range. Figure 5 shows test images, in decreasing order of shapeness as determined by the algorithm. The ordering looks very plausible. (The optimal azimuthal light direction was computed separately for each image. The images are shown at the orientation determined by the algorithm to give maximal shapeness, assuming the light comes from the left.)

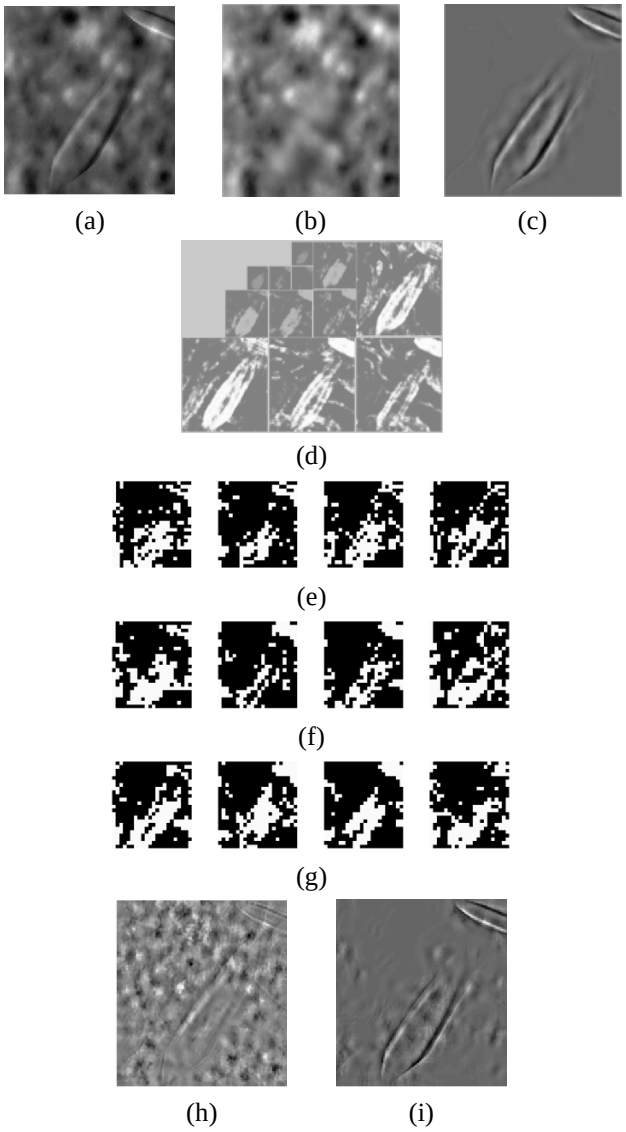


Figure 4: (a) Input image, from the algorithm that generated the training images (assuming light coming from the left, which we also assume here for the analysis), but not in the training set. (b) reflectance component, (c) shape component. (d) are the true labels for each pyramid band. The origin labelling algorithm of the paper was applied to each coefficient of a steerable representation of the input image, resulting in a label (black = paint, white = shape) for pyramid coefficients at each of 4 orientations at each of 3 scales, (e), (f), and (g). The low-pass band of the pyramid was not labelled. Reconstructing the pyramid for a label category, using only the coefficients corresponding to that category, yields an image showing only those features which correspond to (h) reflectance and (i) shape. Note that the algorithm has correctly separated the components due to shading from those due to shape.

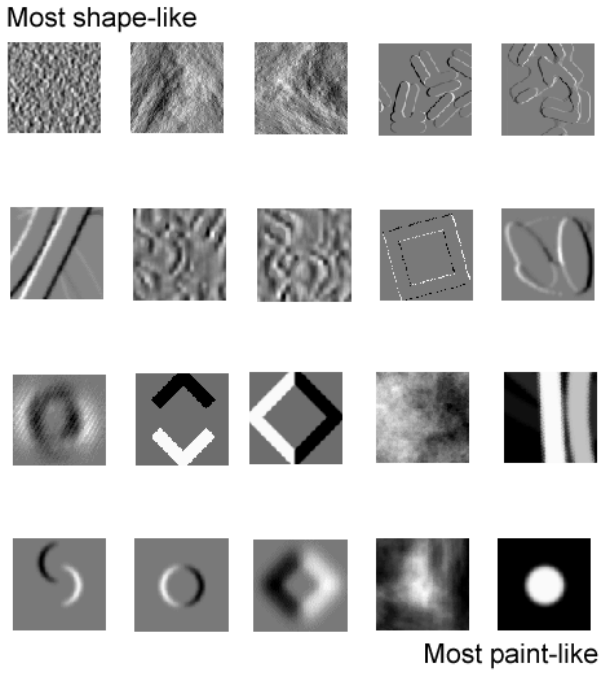


Figure 5: Test images, from [4], listed in decreasing order of rated "shapeness", as determined by our algorithm. Apart from the bump in the bottom row, the ordering looks very reasonable. The extent of the bump was larger than the support region of the local evidence filters, so the image was never viewed in its entirety by any single filter. The Spearman rank correlation with the mean subjects' score was 0.45.

5.2 Graffiti image

We apply our labelling method to the graffiti image of Fig. 1. We first sought the proper light direction. Over a range of assumed light directions (or, equivalently, image rotations assuming a fixed lighting orientation), we computed the variance of the reconstructed shape and reflectance bandpassed images. The orientation corresponding to the maximum ratio (the fourth from left in Fig. 6 (b)) was assumed to be the true lighting direction.

Using that lighting direction, we can then label each of the steerable pyramid coefficients as being due to shading or paint, based on our mixture of gaussians probability density model for the responses of the layered analyzing filters. From that labelling, we can then reconstruct bandpassed versions of images corresponding to each cause–shading and reflectance. To our knowledge, this is the first analysis of separate shading and reflectance causes in natural, grayscale images.

6. Summary and Conclusions

We have developed a learning-based method to separate shading and reflectance in images.

We assume that each local filter in a steerable pyramid has only one cause in the image, either shading or paint, and we seek a labelling for each coefficient describing an image. This labelling allows a bandpassed reconstruction of the image components due to shading, and those due to paint.

We use a training-based approach, first creating a synthetic visual world showing typical examples of images combining shading and reflectance variations.

We analyze the input images using a cascaded energy model: we apply spatial filters, rectify their outputs, then apply them again and rectify again. The pruned outputs of these operations are input to a probability density for each image event class (shading or paint), learning from the training data. The local classification is taken to be the class of the higher probability density.

This simple method works well. The result for an image typical of the training set (but not in it) agreed well with the best possible bandpassed results. The output for a set of psychophysics test images agree well with the judgements of humans. The bandpassed separation for a photograph of mixed shape and paint looks very reasonable.

Because local evidence alone does not always determine the image interpretation, a complete solution would need to propagate the local evidence to uncertain regions. However, it is encouraging to note how much progress can be made from a local analysis alone. To handle more general images, categories of image events other than just shading or paint, such as occluding contours, would have to be accounted for, as well.

This work illustrates an important learning-based approach to low-level vision problems. We constructed a labelled training set, and an appropriate set of image and scene representations which allowed a straightforward machine learning algorithm to solve an important problem in computational vision.

Acknowledgments

We thank E. Adelson and J. Tenenbaum for helpful discussions.

References

[1] C. M. Bishop. *Neural networks for pattern recognition*. Oxford, 1995.

[2] J. S. DeBonet and P. Viola. Texture recognition using a non-parametric multi-scale statistical model. In *Proc. IEEE Computer Vision and Pattern Recognition*, 1998.

[3] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *Intl. J. Computer Vision*, 40(1):25–47, 2000. See also http://www.merl.com/reports/TR2000-05/.

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

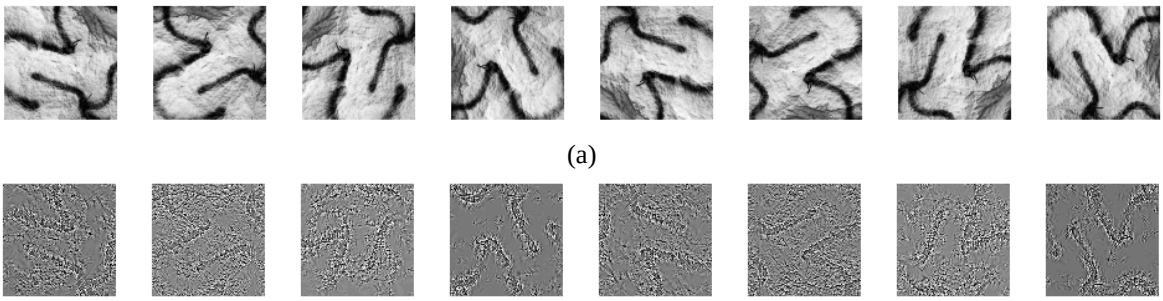
[5] Z. Ghahramani. Factorial learning and the EM algorithm. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Adv. in Neural Information Processing Systems*, volume 7, pages 617–624, Cambridge, MA, 1995. MIT Press.

[6] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time-series. In M. Arbib, editor, *The handbook of brain theory and neural networks*. MIT Press, Cambridge, MA, 1995.

[7] A. P. Pentland. Linear shape from shading. *Intl. J. Comp. Vis.*, 1(4):153–162, 1990.

[8] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge Univ. Press, 1992.

[9] M. Ptzsch, N. Kruger, and C. v.d.Malsburg. Improving object recognition by transforming gabor filter responses. *Network: Computation in Neural Systems*, 7(2), 1996.



(a)

(b)

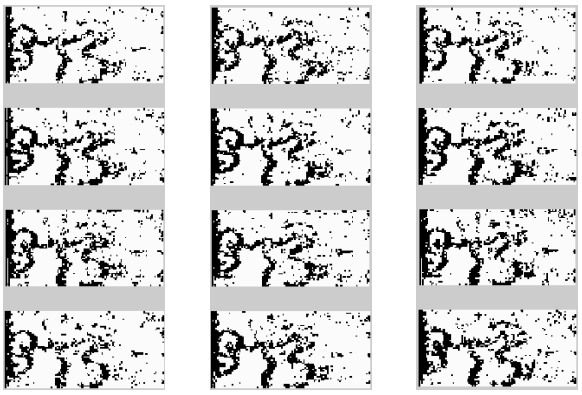
Figure 6: (a) The same input image, rotated to different orientations for analysis with model trained on light always coming from the left. (b) Bandpass image reconstruction from pyramid coefficients labelled as "reflectance". Note that for the two orientations of the input image where the light was indeed coming from the left, only the graffiti gets attributed to paint, while for all the other orientations, both the shading and the paint is attributed to paint. The bandpassed image contrasts were boosted for clarity. The assumed light angle which maximized the total image variance in the shape reconstruction was assumed to be the correct light orientation relative to the input image.

[10] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *2nd Annual Intl. Conf. on Image Processing*, Washington, DC, 1995. IEEE.

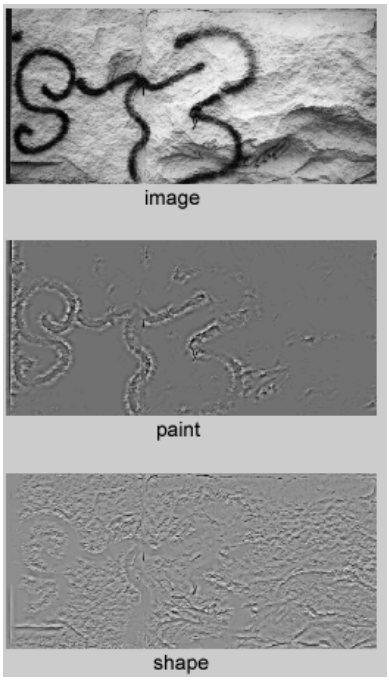
[11] P. Sinha and E. H. Adelson. Recovering reflectance and illumination in a world of painted polyhedra. In *Proc. 4th Intl. Conf. Comp. Vis.*, pages 156–163. IEEE, 1993.

[12] A. Sutter and N. Graham. Investigating simple and complex mechanisms in texture segregation using the speed-accuracy tradeoff method. *Vision Research*, 35(20):2825–2843, 1995.

[13] K. Tieu and P. Viola. Boosting image retrieval. In *Proc. IEEE Computer Vision and Pattern Recognition*, 2000.



(a) (b) (c)



(d)

Figure 7: Result of applying learning-based image labelling method to image of graffiti on a rock. Using the best lighting direction, found as shown in Fig. 6, we apply our learning-based method to the graffiti image of Fig. 1. (a), (b), and (c): The labelling of the steerable pyramid coefficients, determined by our training-based method, for the 3 resolution levels of the image pyramid. Black indicates a "paint" classification; white indicates reflectance. (d): Since we only label steerable pyramid coefficients, the best we can expect is bandpassed reconstructions of the parts of the image caused by shading and the parts caused by paint. The pictured decomposition is largly correct, as compared with the image of the rock with the graffiti painted over, Fig. 1 (b).