Learning Lightness from Human Judgement on Relative Reflectance

Takuya Narihira[1,2], Michael Maire[3], Stella X. Yu[1]
[1]University of California at Berkeley / International Computer Science Institute. [2]Sony Corporation. [3]Toyota Technological Institute at Chicago.

We develop a new approach to inferring lightness, the perceived reflectance of surfaces, from a single image. Classic methods view this problem from the perspective of the intrinsic image model, which assumes that the image intensity I is the product of reflectance image R and shading image S. Lightness L is simply the solution of R attempted by the visual system given its knowledge about the regularity of reflectance and shading:

$$I = R \cdot S, \tag{1}$$

$$L = R^*, \text{s.t. priors}(R,S) \tag{2}$$

Such a decomposition is ill-defined without priors, and the classic strategy is to exploit strong priors in order to constrain the search space for the solution (R^*, S^*) that satisfies the per-pixel factorization.

Our approach involves a complete change in intuition and strategy. We learn a lightness model directly from data, leveraging a training set of many relative reflectance comparisons made by human subjects. That is, we focus on learning the relative ordering of L, *i.e.* $L_i - L_j$, directly from contextual cues present in two local image patches, x_i and x_j, without resorting to an absolute pixel-wise decomposition of I into plausible R and S. Our model is built upon two key elements of recent work:

1. The Intrinsic Images in the Wild (IIW) dataset [1] provides a large collection of ground-truth in the form of human judgements of relative reflectance: 5230 indoor images with a total of 872,161 pairs of comparisons, about 106 ± 45 comparisons per image. These pairwise comparisons take three values: *same, lighter, or darker*, and they are noisy across human subjects.

2. Rich contextual features computed through either hierarchical sparse coding (HSC) [2, 6] or deep convolutional neural networks (CNN) [4, 5] provide an informative fine-to-coarse, small-to-large context feature representation of every patch in the input image, enabling a simpler and direct local classification approach without heavy reliance on any hand-designed global priors or expensive inference algorithms.

Figure 1 illustrates the ground-truth labeling from which we train as well as some example results of our learned models. Figure 2 diagrams the overall architecture within which we utilize HSC or CNN algorithms as feature extractors. We extract features z_i, z_j of patches x_i, x_j and learn weights w for a linear classifier f:

$$L_i - L_j = f(z_i, z_j) = w^T (z_i - z_j) \tag{3}$$

When using HSC as a feature extractor, we learn the sparse representation generatively as in [6] and then train w by ridge ranking regression on the human ground-truth data for reflectance:

$$\min \varepsilon(w) = \sum_{i,j} \log\left(1 + \exp(-J_{ij} w^T (z_i - z_j))\right) + \gamma w^T w \tag{4}$$

where:

$$J_{ij} = \begin{cases} 1, & R_i^h > R_j^h \\ -1, & R_i^h < R_j^h \end{cases} \tag{5}$$

Here R^h refers to human ratings of relative reflectance (lightness) on the IIW dataset. γ calibrates regularization. For each example where humans judge equal reflectance ($R_i^h = R_j^h$), we create two virtual examples with both $R_{ij} = 1$ and $R_{ij} = -1$ in order to force prediction $f(z_i, z_j)$ toward zero.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

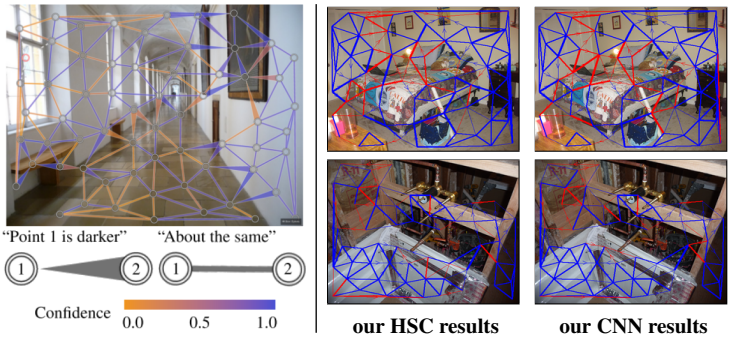

Figure 1: *Left:* Human lightness annotations [1]. *Right:* Our model predictions compared to human ground-truth (blue for correct, red for incorrect).

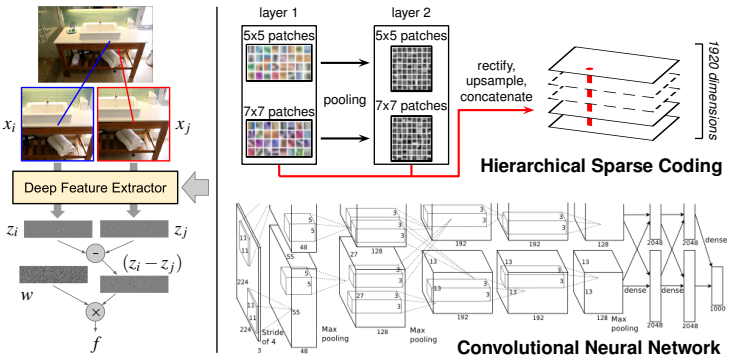

Figure 2: Direct learning of pairwise lightness relationships.

For our CNN-based feature extractor, we use the Caffe [4] implementation of the 7-layer convolutional neural network of [5] and take the 4096-dimensional activations of the final fully connected layer as a feature descriptor for a patch presented as input to the network. We use the standard cross entropy classification loss to train the CNN from both randomly initialized weights and weights initialized by pre-training on ImageNet [3].

As measured by the standard benchmark of confidence-weighted disagreement with human ground-truth (WHDR) on the IIW dataset, our HSC and CNN models offer state-of-the-art performance, achieving 20.9% and 18.1% WHDR, respectively. This matches that of more complicated algorithms such as the CRF method of Bell *et al.* [1]. Use of rich patch representations, obtained via hierarchical sparse coding or convolutional neural networks, and a large amount of training data, enables our purely local model to compete with global inference approaches. Our work opens up new areas of exploration within the classic problem of intrinsic image decomposition.

[1] Sean Bell, Kavita Bala, and Noah Sanvely. Intrinsic images in the wild. In *ACM Trans. on Graphics*, 2014.

[2] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Multipath sparse coding using hierarchical matching pursuit. *CVPR*, 2013.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *arXiv preprint arXiv:1408.5093*, 2014.

[5] A. Krizhevsky, S.Ilya, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[6] Michael Maire, Stella X. Yu, and Pietro Perona. Reconstructive sparse code transfer for contour detection and semantic labeling. *ACCV*, 2014.