Simultaneous Cast Shadows, Illumination & Geometry Inference Using Hypergraphs

Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras and Nikos Paragios, *Fellow, IEEE*

Abstract—The cast shadows in an image provide important information about illumination and geometry. In this paper, we utilize this information in a novel framework, in order to jointly recover the illumination environment, a set of geometry parameters and an estimate of the cast shadows in the scene, given a single image and coarse initial 3D geometry. We model the interaction of illumination and geometry in the scene and associate it with image evidence for cast shadows using a higher-order Markov Random Field (MRF) illumination model, while we also introduce a method to obtain approximate image evidence for cast shadows. Capturing the interaction between light sources and geometry in the proposed graphical model necessitates higher-order cliques and continuous-valued variables, which make inference challenging. Taking advantage of domain knowledge, we provide a two-stage minimization technique for the MRF energy of our model. We evaluate our method in different datasets, both synthetic and real. Our model is robust to rough knowledge of geometry and inaccurate initial shadow estimates, allowing a generic coarse 3D model to represent a whole class of objects for the task of illumination estimation, or the estimation of geometry parameters to refine our initial knowledge of scene geometry, simultaneously with illumination estimation.

Index Terms—Markov random fields, Photometry, Shading, Image models

1 Introduction

The appearance of a scene depends considerably on the illumination conditions, which therefore influence a large number of computer vision tasks. Illumination is one of the three components of the image formation process, along with the 3D geometry of the scene and the reflectance properties of the surfaces in it. The interaction among these three components means that estimation of one or two of them requires knowledge or strong assumptions about the rest ([18], [23], [26], [28], [29]). Previous work in illumination estimation usually assumes known scene geometry and makes strong assumptions about reflectance. Our goal in this work is, through a statistically robust inference approach, to diminish the effect violations of such assumptions have on the final illumination estimate, based on the information contained in cast shadows. Compared to illumination cues such as shading or specularities, cast shadows are relatively stable in the presence of large inaccuracies in the knowledge of geometry and reflectance.

Estimating illumination from cast shadows implies obtaining an estimate of the cast shadows in the image, which can be challenging in complex images. Shadow detection, in the absence of illumination estimation or knowledge of 3D geometry is a well studied problem. Salvador et al. [25] use invariant color features to segment cast shadows in still or moving images. Finlayson et al. [5], [4] propose illumination invariant features to detect and remove shadows from a single image. Their method makes several assumptions about the lights and the camera and its performance reduces in lower-quality consumer-grade photographs. Recently, Zhu et al. [31] combine a number of different features in a probabilistic framework to recognize shadows in monochromatic images, while in [16], Lalonde et al. propose a learning approach to detect shadows in consumer-grade photographs, focusing on shadows cast on the ground. Guo et al. [22] consider image regions, combining classifiers for individual regions as well as region pairs in a graph in order to label the shadow regions.

Much research in the computer vision community has dealt with extracting illumination from shading, specular reflections or cast shadows. Yang and Yuille [29] estimate multiple light sources from the intensity along occluding boundaries and critical points; Wang et al. [28] estimate multiple directional illuminants utilizing both shading and shadows, assuming known scene geometry. Sato et al. [26] estimate illumination from cast shadows, assuming known geometry illuminated by a set of infinitely distant light sources, casting shadows onto a planar lambertian surface. Their method uses non-negative least squares optimization to obtain an illumination estimate. Hara et al. [9] remove the distant illumination assumption, while estimating simultaneously illumination and reflectance. In [30], Zhou et al. propose a unified framework to estimate both distant and point light sources.

The prior art on illumination estimation from shadows cast on textured surfaces is limited. Sato et al. [26] require an extra image to deal with texture. Li et al. [18] propose a method that integrates multiple

- *A.Panagopoulos and D.Samaras are with the Dept. of Computer Science, Stony Brook University, Stony Brook, NY 11794, USA. E-mail: {apanagop, samaras}@cs.stonybrook.edu*
- *C.Wang and N.Paragios are with Center for Visual Computing, Ecole Centrale Paris, Châtenay-Malabry, France, and Equipe GALEN, INRIA Saclay - Île-de-France, Orsay, France. E-mail: {chaohui.wang, nikos.paragios}@ecp.fr*

cues from shading, shadow, and specularities, utilizing physical consistencies between lighting cues to handle textured surfaces. Kim et al. [12] use regularization by correlation to estimate illumination from shadows in the presence of texture, but require extra user-specified information and assume lambertian reflectance and known geometry. Panagopoulos et al. [19] propose a method able to deal with inaccurate geometry and texture, but the shadow detection results on textured surfaces are limited. In the more special case of daytime outdoor scenes, Lalonde et al. [15] propose an approach that combines cues from the sky, cast shadows on the ground and surface brightness to estimate illumination where the sun is the single light source, which does not require known 3D geometry but is not applicable to general scenes.

In general, illumination estimation necessitates strong assumptions about geometry. The main goal of this paper is to relax such assumptions, so that simplistic geometry approximations, like the ones that can be extracted automatically or provided from limited user input, will suffice to estimate the illuminants in a scene. To this end, we propose a novel framework to recover the illumination environment of a scene, a rough cast shadow estimate and a set of geometry parameters from a single observed image, given coarse initial 3D geometry. With our method, very coarse approximations of geometry, such as bounding boxes, are enough to estimate illumination, while the geometry of the occluders can be refined as part of the illumination estimation process. The initial approximate geometric information we require could be derived as part of more general scene understanding techniques, while enabling illumination estimation to be incorporated in the scene understanding loop; the obtained illumination and geometry information could be a crucial contextual prior in addressing various other scene understanding questions.

Graphical models can efficiently incorporate different cues within a unified framework [27]. In order to deal with the complex illumination/geometry/shadows estimation problem robustly in a flexible and extensible framework, we jointly model the geometry, light sources, and shadow values within an MRF model. All the latent variables can then be simultaneously inferred through the minimization of the energy of the MRF. This work was originally reported in [21].

The MRF model we propose captures the interaction between geometry and light sources and combines it with image evidence of cast shadows. Cast shadow detection is well-posed in terms of graph topology, since it can be expressed using a graph in the form of a 2-dimensional 4-connected lattice, where each image pixel corresponds to a graph node. Modeling in the MRF model the creation of cast shadows from the interaction of light sources and geometry, on the other hand, implies a potential dependence between each pixel and all nodes representing the light sources and the occluder geometry. This generally results in higher-order cliques in the graph representing our MRF model. Further complications arise by the fact that the number of light sources is unknown, resulting in unknown MRF topology, and that the search space is continuous. We are able to reduce the search space and identify the MRF topology through an initial illumination estimate obtained using a voting algorithm. Inference in higher-order MRF models has received a lot of attention recently [11], [13]; here we take advantage of domain knowledge to describe a two-stage minimization approach that can effectively minimize the MRF energy. Our approach is based on a decomposition of the energy and requires solving only pairwise MRF problems. We make the following *assumptions* (common in illumination modeling): an initial coarse 3D geometry is known, the illumination environment can be approximated by a set of distant light sources, the reflectance of surfaces is roughly lambertian and there are no interreflections. Furthermore, if estimation of occluder geometry parameters is desired, these occluders have to be identified in the original image by providing a 2D bounding box, and one or more candidate geometric models. The shadow detection method of Sec. 5 assumes that shadows are cast on flat surfaces. Our illumination MRF model does not rely on this assumption, however.

To obtain the initial shadow estimate required by our method, we describe a method based on the observation that illumination affects the whole image in a consistent way. Therefore, features such as hue or brightness changes across shadow edges are consistent across the whole image, a fact that we exploit to detect shadows. Our approach is also aided by a simple measure of image brightness, the *bright channel* [20]. It should be noted, however, that the proposed MRF model can incorporate other shadow cues.

We provide qualitative evaluation of our method on different datasets, including images captured in a controlled environment, car images collected from Flickr and images from the Motorbikes class of Caltech 101 [17]. We also provide quantitative results on a synthetic dataset. The experimental evaluation shows that our method is robust enough to be able to use geometry consisting of bounding boxes or a common rough 3D model for a whole class of objects, while it can also be applied to scenes where some of our assumptions are violated. Results on geometry parameter estimation show that through our model we can extract useful information about object geometry and pose from the cast shadows.

This paper is organized as follows: Sec. 2 presents related fundamentals; Sec. 3 describes the MRF model to jointly estimate the shadows, illumination and geometry parameters. In Sec. 4 we discuss the inference process. Section 5 presents the shadow cue we used, the bright channel cue. Experimental evaluation is provided in Sec. 6, and Sec. 7 concludes the paper.

2 Problem Description

The input required by our method is a single color image $\mathbf{I}$, an approximate 3D model of the geometry $\mathcal{G}$ of the scene and approximate camera parameters. If geometry parameter estimation is desired, a 2D bounding box and a set of candidate geometric models should be provided for each identified occluder, as described in Sec. 2.1.

We adopt a commonly used set of assumptions: the surfaces in the scene exhibit lambertian reflectance, and the scene is illuminated by N point light sources at infinity, as well as some constant ambient illumination term. Under these assumptions, the outgoing radiance at a pixel i can be expressed as:

$$L_o(\mathbf{p}) = \rho_{\mathbf{p}} \left(\alpha_0 + \sum_{i=1}^{N} V_{\mathbf{p}}(\mathbf{d}_i)\alpha_i \max\{-\mathbf{d}_i \cdot \mathbf{n}_{\mathbf{p}}, 0\} \right), \tag{1}$$

where $\rho_{\mathbf{p}}$ is the albedo at point $\mathbf{p}$ with normal $\mathbf{n}_{\mathbf{p}}$, α_0 is the ambient intensity, $\alpha_i, i \in \{1, ..., N\}$ is the intensity of the i-th light source, $\mathbf{d}_i$ is the illumination direction of the i-th light source, and $V_{\mathbf{p}}(\mathbf{d}_i)$ is a visibility term for direction $\mathbf{d}_i$ at point $\mathbf{p}$:

$$V_{\mathbf{p}}(\mathbf{d}_i) = \begin{cases} 1, & \text{if ray to } \mathbf{p} \text{ along } \mathbf{d}_i \text{ intersects } \mathcal{G} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Therefore, illumination information is fully captured by parameters $\theta_{\mathcal{L}} = \{\alpha_0, \alpha_1, ..., \alpha_N, \mathbf{d}_1, ..., \mathbf{d}_N\}$, where the set of lights $\mathcal{L}$ includes the ambient light.

If we assume a simplified linear model for the camera sensors, the observed value at pixel (x, y) is:

$$I(x, y) = cL_o(\mathbf{p}) + \epsilon, \tag{3}$$

where c is an exposure parameter and ϵ is noise. Since we can only estimate light source intensities up to scale, we can safely assume that $c = 1$.

In our method, we first obtain an initial cast shadow estimate from the input image $\mathbf{I}$ (see Sec. 5). This estimate should contain the shading intensity at each pixel in shadow, without any variations due to albedo ρ, and the non-shadow pixels of $\mathbf{I}$ should be masked out. Ideally, therefore, the value of each shadow pixel (x, y) in such a shadow image $\mathbf{I}_s$ would be the shading at that point due to the non-occluded light sources:

$$I_s(x, y) = \alpha_0 + \sum_{i=1}^{N} V_{\mathbf{p}}(\mathbf{d}_i)\alpha_i \max\{\mathbf{d}_i \cdot \mathbf{n}_{\mathbf{p}}, 0\}, \tag{4}$$

where $\mathbf{p}$ is the 3D point where (x, y) projects to. In practice we can obtain a cast shadow cue $\hat{\mathbf{I}}_s$ which is a rough approximation of $\mathbf{I}_s$.

2.1 Geometry modeling

One of the goals of this work is to provide a model that allows reasoning about illumination to be incorporated in more complex scene understanding tasks. Towards this goal, we describe here how we can incorporate objects with unknown parameters to be estimated to our model. Estimation of these parameters happens jointly with the estimation of illumination and cast shadows. Different parametrizations of the scene geometry could be handled by our model without significant changes, as long as the total number of geometry parameters remains small.

As mentioned, $\mathcal{G}$ is the known, approximate 3D geometry which is provided as input. We assume that there may also exist a (small) set of objects $\mathcal{O}$, which are the parametric objects to be estimated. The information we assume as known about the objects $\mathcal{O}$ is restricted, for each object i, to a 2D bounding box that bounds the object in the image, and a set ${\mathcal{G}_{\mathcal{O}}}^{(i)}$ of potential approximate 3D models for this object. The potential 3D models can be thought as the geometric models representing common instances of the class to which object i belongs (e.g. if the object is a car, we could assume a small number of 3D models representing common car shapes). Our goal is to recover, concurrently with illumination estimation, the most probable geometry and the pose (orientation/translation/scale) for each of these objects, in order to best approximate the real scene geometry.

In the following sections we will present a model to jointly estimate the shadows, the illumination parameters $\theta_{\mathcal{L}}$ and a set of geometry parameters from the approximate shadow cue $\hat{\mathbf{I}}_s$. In section 5 we present the shadow cue which we used to obtain our results.

3 Global MRF for Cast Shadow Formation

We associate the image-level evidence for cast shadows with high-level information about geometry and the lights through the MRF model described below.

3.1 Markov Random Field Formulation

The proposed MRF consists of one node for each image pixel $i \in \mathcal{P}$, one node for each light source $l \in \mathcal{L}$, one node for the ambient intensity α_0 and one node for the geometry of each object k in the set of objects $\mathcal{O}$. Each pixel node, all the light nodes and all the object nodes compose a high-order clique $c \in \mathcal{C}$. The 4-neighborhood system [1] composes the edge set $\mathcal{E}$ between pixels. The energy of our MRF model has the following form:

$$\begin{aligned} E(\mathbf{x}) \;\; = \;\; & \sum_{i \in \mathcal{P}} \phi_p(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_p(x_i, x_j) + \sum_{k \in \mathcal{O}} \phi_k(x_k) \\ & + \sum_{l \in \mathcal{L}} \phi_l(x_l, \mathbf{x}_{\mathcal{O}}) + \sum_{i \in \mathcal{P}} \psi_c(x_i, \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}), \end{aligned} \tag{5}$$

where $\phi_p(x_i)$ and $\phi_k(x_k)$ are the singleton potentials for pixel nodes and object nodes respectively, $\psi_p(x_i, x_j)$ is the pairwise potential defined on a pair of neighboring pixels, $\phi_l(x_l, \mathbf{x}_{\mathcal{O}})$ is the clique potential expressing a shadow shape-matching prior, and $\psi_c(x_i, \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}})$ is the high-order potential associating all lights in $\mathcal{L}$, all objects in $\mathcal{O}$ and a pixel x_i.

The latent variable x_i for pixel node $i \in \mathcal{P}$ represents the intensity value for that pixel. We uniformly discretize the real intensity value $[0, 1]$ into N bins to get the candidate set $\mathcal{X}_i$ for x_i. The latent variable x_l for light node $l \in \mathcal{L}$ is composed of the intensity and the direction of the light. We sample the space in the vicinity of the light configuration obtained by the voting approach explained later, to initialize the candidate set $\mathcal{X}_l$ for x_l (see details later in this section).

By $\mathbf{x}_{\mathcal{O}}$ we signify the labels corresponding to the objects in $\mathcal{O}$. The label x_k^O of object node k determines a set of parameters $(g_k, \phi_k, t_x, t_y, t_z, s_x, s_y, s_z)$, where g_k is an index into $\mathcal{G}_{\mathcal{O}}{}^{(k)}$ that determines which of the potential object geometries is selected for label x_k^O, ϕ_k is the azimuth orientation of the object, (t_x, t_y, t_z) is the translation and (s_x, s_y, s_z) is the scale of the object.

3.1.1 Singleton Potentials for Pixel Nodes

This term encodes the similarity between the estimated intensity value at pixel i and the shadow cue value $\hat{I}_s(i)$ and is defined as:

$$\phi_p(x_i) = w_s \min\{\left|x_i - \hat{I}_s(i)\right|, t_p\}. \tag{6}$$

where an upper bound t_p for this cost term is used to avoid over-penalizing outliers and w_s is a positive weight coefficient (same for w_l, w_p and w_c below).

3.1.2 Singleton Potentials for Geometry

In our attempt to extract information about the geometry of object k, in the model of Eq.5 we obviously take into account the information in the shadow cast by object k. However, the cast shadow provides only one projection of the object, which is often insufficient to extract useful information about the object shape. We can, however, obtain a second projection of the object, the one onto the image plane, which will provide us with extra information to make reasoning about the object pose and shape possible.

To obtain the shape of the object on the image plane, we use GrabCut [24] with the user-provided 2D bounding box for the object as input. GrabCut gives us a foreground/background segmentation, where pixels in the foreground $\mathcal{F}$ are the pixels most likely to belong to the object contained in the initial 2D bounding box.

The singleton potentials $\phi_k(x_k)$ penalize geometry labels x_k that are inconsistent with the extracted shape $\mathcal{F}$ of the object $k \in \mathcal{O}$ in the image. This potential also penalizes geometry labels x_k that correspond to a scale that significantly deforms the initial geometry. The form of the potential is:

$$\phi_k(x_k) = \sum_{i \in \mathcal{P}} (\mathcal{F}(i) - \mathcal{M}_{x_k}(i))^2 + w_s \left\|\mathbf{x}_k^{(scale)} - [1, 1, 1]\right\|_2, \tag{7}$$

where $\mathbf{x}_k^{(scale)}$ is a vector (s_x, s_y, s_z) determining the object scale corresponding to label x_k, $\mathcal{F}$ is the object mask obtained by GrabCut:

$$\mathcal{F}(i) = \begin{cases} -1 & \text{if } i \in \text{background} \\ +1 & \text{if } i \in \text{foreground} \end{cases} \tag{8}$$

and $\mathcal{M}$ is the mask corresponding to the projection $I_k^{\mathcal{O}}$ of the geometry assigned to object k from label x_k, at the corresponding rotation, translation and scale:

$$\mathcal{M}(i) = \begin{cases} -1 & \text{if } i \notin I_k^{\mathcal{O}} \\ +1 & \text{if } i \in I_k^{\mathcal{O}} \end{cases}. \tag{9}$$

As demonstrated in our experiments (Fig.10), the obtained mask $\mathcal{M}$ is not by itself adequate for determining the geometry parameters. The combination of $\mathcal{M}$ with the information contained in shadow regions in our MRF model, however, allows us to obtain a good estimate of the geometry parameters.

3.1.3 Pairwise Potentials

We adopt the well-known *Ising* prior [7] to define the pairwise potential between neighboring pixels $(i, j) \in \mathcal{E}$ to favor neighboring pixels having the same value:

$$\psi_p(x_i, x_j) = \begin{cases} w_p & \text{if } x_i \neq x_j \\ 0 & \text{if } x_i = x_j \end{cases} \tag{10}$$

3.1.4 Shadow Shape-matching Prior

Terms $\phi_l(\mathbf{x}_l, \mathbf{x}_{\mathcal{O}})$ incorporate into the MRF model a shadow shape-matching prior for light l, in order to favor illumination/geometry configurations generating shadows that match observed shadow outlines.

We apply gaussian smoothing and the Sobel edge detector [8] to detect edges in the shadow cue image. Let $\tau(i) \in [0, 2\pi)$ be the angle of the gradient at pixel i with the x-axis, and $\hat{\tau}(i) \in \{0, K-1\}$ a quantization of $\tau(i)$. For each possible direction $d \in \{0, K-1\}$, we compute a distance map υ_d so that, for pixel i, $\upsilon_d(i)$ is the distance from pixel i to the closest edge pixel of orientation d.

For pixel i with gradient angle $\tau(i)$, the distance function is computed by interpolating between the distance map values for the two closest quantized orientations:

$$dist_{\tau(i)}(i) = (1 - \lambda) \cdot \upsilon_{\hat{\tau}(i)}(i) + \lambda \cdot \upsilon_{\hat{\tau}(i)+1}(i), \tag{11}$$

$$\lambda = \left\{\frac{K \cdot \tau(i)}{2\pi}\right\}, \tag{12}$$

where $\{.\}$ indicates the fractional part. In our experiments, we chose $K = 4$.

The shape-matching prior expresses the quality of the match between the observed edges in the shadow cue image and the edges of the synthetic shadow $\mathcal{S}_l$ associated with $\mathbf{x}_l$ and geometry configuration $\mathbf{x}_{\mathcal{O}}$:

$$\phi_l(\mathbf{x}_l, \mathbf{x}_{\mathcal{O}}) = w_l \frac{1}{|\mathcal{E}_{\mathcal{S}_l}(\mathbf{x}_l, \mathbf{x}_{\mathcal{O}})|} \sum_{i \in \mathcal{E}_{\mathcal{S}_l}(\mathbf{x}_l, \mathbf{x}_{\mathcal{O}})} dist_{\tau_{\mathcal{S}_l}(i)}(i), \tag{13}$$

where $\mathcal{E}_{\mathcal{S}_l}(\mathbf{x}_l, \mathbf{x}_{\mathcal{O}})$ is the set of all pixels that lie on edges of the shadow $\mathcal{S}_l$ generated by light label $\mathbf{x}_l$ and

$\tau_{\mathcal{S}_l}(i)$ is the gradient angle of the synthetic shadow edge generated by x_l at pixel i. To determine the set of shadow edge pixels $\mathcal{E}_{\mathcal{S}_l}(\mathbf{x}_l, \mathbf{x}_{\mathcal{O}})$, we generate the shadow $\mathcal{S}_l$ created by light label $\mathbf{x}_l$ and the geometry $\mathbf{x}_{\mathcal{O}}$ and then apply gaussian smoothing and the Sobel edge detector. The set $\mathcal{E}_{\mathcal{S}_l}(\mathbf{x}_l, \mathbf{x}_{\mathcal{O}})$ contains all pixels whose gradient magnitude is above θ_e.

3.1.5 Higher-order Potentials

The higher-order terms $\psi_c(x_i, \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}})$ impose consistency between the light source labels $\mathbf{x}_{\mathcal{L}}$, the geometry labels $\mathbf{x}_{\mathcal{O}}$ and the pixel intensity values.

Let $\mathcal{S}$ be the synthetic shadow, generated by light configuration $\mathbf{x}_{\mathcal{L}}$ and geometry configuration $\mathbf{x}_{\mathcal{O}}$. The intensity at pixel $i \in \mathcal{S}$ is:

$$s'_i(\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}) = \mathbf{x}^{\alpha_0} + \sum_{l \in \mathcal{L}} x_l^{\alpha} V_i(\mathbf{x}_l^{dir} | \mathbf{x}_{\mathcal{O}}) \max\{-\mathbf{x}_l^{dir} \cdot \mathbf{n}(i), 0\}, \tag{14}$$

where $\mathbf{x}^{\alpha_0}$ corresponds to the ambient intensity, x_l^{α} is the light intensity component of x_l, $\mathbf{x}_l^{dir}$ is the light direction component, $\mathbf{n}(i)$ is the normal at 3D point $\mathbf{p}$ imaged at pixel i and $V_i(\mathbf{x}_l^{dir}) \in \{0, 1\}$ is the visibility term for light direction $\mathbf{x}_l^{dir}$ at 3D point $\mathbf{p}$ (cf. Eq.2). For pixels $i \notin \mathcal{S}$, we set $s'_i(\mathbf{x}_{\mathcal{L}}) = 1$, according to the definition of our shadow cue $\mathbf{I}_s$. The clique potential is defined as:

$$\psi_c^{(1)}(x_i, \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}) = w_c \min\{(s'_i(\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}) - x_i)^2, t_c\}, \tag{15}$$

where t_c is also an upper bound to avoid over-penalizing outliers.

In cases where the geometry $\mathcal{G}$ is far from the real scene geometry, a light configuration that does not generate any visible shadows in the image might result to a lower MRF energy than the true light source. Similarly, if there are falsely identified shadows covering a large portion of the image, a configuration where the whole image is in shadow (light source under the ground plane) might correspond to a lower energy. To avoid these degenerate cases, we introduce the term $\psi_c^{(2)}(\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}})$, which assigns a very high penalty to light configurations that do not generate any visible shadows or that generate shadows at every pixel. The final form of the clique potential is:

$$\psi_c(x_i, \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}) = \psi_c^{(1)}(x_i, \mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}) + \psi_c^{(2)}(\mathbf{x}_{\mathcal{L}}, \mathbf{x}_{\mathcal{O}}). \tag{16}$$

3.2 Initializing the MRF Model

As mentioned earlier, the continuous search space complicates inference in our MRF model. Furthermore, in our discussion of the model so far, we assumed that the number of light sources $|\mathcal{L}|$ is known. In practice, however, $|\mathcal{L}|$ may be unknown, which results in unknown MRF topology. To deal with these two issues, we use a rough initial illumination estimate both to determine $|\mathcal{L}|$, if it is unknown, and to set the initial values of the light source variables, before inference begins.

To obtain this rough illumination estimate, we use the greedy approach described in Algorithm 1, based on the shadow cue $\hat{\mathbf{I}}_s$ and geometry $\mathcal{G}$. We examine a fixed set of possible illumination directions, corresponding to the nodes of a geodesic sphere [26]. In each iteration of this algorithm, the pixels in shadow, which are not explained by already discovered light sources, vote for all occluded illumination directions. Pixels not in shadow vote for all directions that are not occluded. After all pixels cast a vote, the most popular direction is chosen as the direction of the new light source. Having the light source direction, we estimate the light source intensity using the median of local intensity estimates from each pixel in the shadow of this light source, and the new light source is added to the set of discovered light sources. The algorithm stops when the estimated intensity of the new light source is near zero, meaning that it has no significant contribution to the observed shadows.

Algorithm 1 Voting for initial illumination estimate

```
Lights Set: $\mathcal{L} \leftarrow \varnothing$
Direction Set: $\mathcal{D} \leftarrow$ all the nodes of a unit geodesic sphere
Pixel Set: $\mathcal{P} \leftarrow$ all the pixels in the observed image
loop
    votes[$\mathbf{d}$] $\leftarrow 0, \forall \mathbf{d} \in \mathcal{D}$
    for all pixel $i \in \mathcal{P}$ do
        for all direction $\mathbf{d} \in \mathcal{D} \setminus \mathcal{L}$ do
            if $I_s(i) < \theta_S$ and $\forall \mathbf{d}' \in \mathcal{L}, V_i(\mathbf{d}') = 0$ then
                if $V_i(\mathbf{d}) = 1$ then votes[$\mathbf{d}$] $\leftarrow$ votes[$\mathbf{d}$] $+ 1$
            else
                if $V_i(\mathbf{d}) = 0$ then votes[$\mathbf{d}$] $\leftarrow$ votes[$\mathbf{d}$] $+ 1$
    $\mathbf{d}^* \leftarrow \arg\max_{\mathbf{d}}(votes[\mathbf{d}])$
    $\mathcal{P}_{\mathbf{d}^*} \leftarrow \{i | c_i(\mathbf{d}^*) = 1 \text{ and } \forall \mathbf{d} \neq \mathbf{d}^*, c_i(\mathbf{d}) = 0\}$
    $\alpha_{\mathbf{d}^*} \leftarrow median\left\{\frac{1 - I_s(i)}{\max\{-\mathbf{n}(\mathbf{p}(i)) \cdot \mathbf{d}^*, 0\}}\right\}_{i \in \mathcal{P}_{\mathbf{d}^*}}$
    if $\alpha_{\mathbf{d}^*} < \epsilon_\alpha$ then
        stop the loop
    $\mathcal{L} \leftarrow \mathcal{L} \cup \{(\mathbf{d}^*, \alpha_{\mathbf{d}^*})\}$
```

4 Inference

We simultaneously estimate the cast shadows, illumination and geometry parameters by minimizing the MRF's energy defined in Eq. 5:

$$\mathbf{x}^{opt} = \arg\min_{\mathbf{x}} E(\mathbf{x}) \tag{17}$$

Minimizing this energy, however, is challenging, because our MRF model contains high-order cliques of size up to $|\mathcal{L}| + |\mathcal{O}| + 1$.

To efficiently perform inference, we can split the minimization of the energy in Eq.5 in two stages [2]. In the light and geometry parameter selection stage, we choose a candidate set of light and geometry parameters for which we will compute the MRF energy, and if this energy is lower than the current minimum, we accept them. In the pixel label selection stage, assuming fixed light and geometry parameters, we compute the MRF energy solving a pairwise MRF.

If we assume that the light parameters are fixed, the high-order clique potentials $\psi_c^{(1)}$ in Eq.15, which are part of ψ_c, become singleton potentials of the form:

$$\psi_c^{(1)}(x_i|\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O}) = w_c \min\{(s_i'(\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O}) - x_i)^2, t_c\}. \quad (18)$$

This way, for a fixed light configuration $\mathbf{x}_\mathcal{L}$ and a fixed geometry configuration $\mathbf{x}_\mathcal{O}$, after we split ψ_c in $\psi_c^{(1)}$ and $\psi_c^{(2)}$ as in Eq.16, we can rewrite the energy of the MRF model in Eq.5 as:

$$E(\mathbf{x}) = E_I(\mathbf{x}|\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O}) + E_L(\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O}) + E_G(\mathbf{x}_\mathcal{O}), \quad (19)$$

where

$$\begin{aligned} E_I(\mathbf{x}|\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O}) &= \sum_{i\in\mathcal{P}} \Big(\phi_p(x_i) + \psi_c^{(1)}(x_i|\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O})\Big) \\ &\quad + \sum_{(i,j)\in\mathcal{E}} \psi_p(x_i,x_j) \quad (20) \end{aligned}$$

$$E_L(\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O}) = \sum_{l\in\mathcal{L}} \Big(\phi_l(x_l) + |\mathcal{P}|\psi_c^{(2)}(\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O})\Big), \quad (21)$$

$$E_G(\mathbf{x}_\mathcal{O}) = \sum_{k\in\mathcal{O}} \phi_k(x_k) \quad (22)$$

are the energy terms associated with the (fixed) light configuration $\mathbf{x}_\mathcal{L}$ and the (fixed) geometry configuration $\mathbf{x}_\mathcal{O}$ but independent of the per-pixel variables.

For a given light configuration $\mathbf{x}_\mathcal{L}$ and geometry configuration $\mathbf{x}_\mathcal{O}$, the energy $E_I(\mathbf{x}|\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O})$ can be minimized using any inference algorithm for pairwise MRFs. The speed of the chosen algorithm is, however, important, because the energy $E_I(\mathbf{x}|\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O})$ is minimized many times (for different light and geometry configurations). The FastPD algorithm [14] is a fitting choice and was adopted in our experiments.

The energy minimum $\min_x\{E_I(\mathbf{x}|\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O})\}$ changes with different light configurations and different geometry configurations. To minimize $E(\mathbf{x})$, a (blocked) coordinate descent approach in the light and geometry parameter domain is used.

Let $\hat{\mathbf{x}}_\mathcal{L}^{(s-1)}$, $\hat{\mathbf{x}}_\mathcal{O}^{(s-1)}$ be the set of light and geometry parameters that correspond to the minimum energy encountered up to iteration $s-1$. At iteration s, we generate proposed light labels $\mathbf{x}_\mathcal{L}^{(s)}$ and geometry labels $\mathbf{x}_\mathcal{O}^{(s)}$ by sampling the light parameter space around the current light estimate $\hat{\mathbf{x}}_\mathcal{L}^{(s-1)}$ and the geometry parameter space around the current geometry configuration estimate $\hat{\mathbf{x}}_\mathcal{O}^{(s-1)}$. We then compute the total MRF energy as

$$\begin{aligned} E^{(s)}(\mathbf{x}) &= \min_x\{E_I(\mathbf{x}|\mathbf{x}_\mathcal{L}^{(s)},\mathbf{x}_\mathcal{O}^{(s)})\} \\ &\quad + E_L(\mathbf{x}_\mathcal{L}^{(s)},\mathbf{x}_\mathcal{O}^{(s)}) + E_G(\mathbf{x}_\mathcal{O}^{(s)}), \quad (23) \end{aligned}$$

which includes minimizing the pairwise energy $E_I(\mathbf{x}|\mathbf{x}_\mathcal{L}^{(s)},\mathbf{x}_\mathcal{O}^{(s)})$. If the new energy $E^{(s)}(\mathbf{x})$ is lower than the previous lowest energy, we keep the proposed illumination and geometry labels $\mathbf{x}_\mathcal{L}^{(s)}$ and $\mathbf{x}_\mathcal{O}^{(s)}$, otherwise they are discarded.

As the number of geometry and illumination parameters is increasing, the choice of which dimensions of the illumination-geometry parameter domain to re-sample in order to generate proposals $\mathbf{x}_\mathcal{L}^{(s)}$ and $\mathbf{x}_\mathcal{O}^{(s)}$ becomes crucial for the effectiveness of the minimization. In our experiments we used the following proposal schedule: At some iteration s, a single light source l is chosen, and new values are generated only for the parameters of light source l and the ambient intensity to produce $\mathbf{x}_\mathcal{L}^{(s)}$. At iteration $s+1$, new values for the azimuth rotation and geometry class label of a single object k are generated to produce $\mathbf{x}_\mathcal{O}^{(s+1)}$. At iteration $s+2$ new values are generated for the 6 scalar parameters defining the 3D translation and 3D scale of a single object k to produce $\mathbf{x}_\mathcal{O}^{(s+2)}$. This proposal schedule is repeated every 3 iterations.

The proposed labels at each iteration are generated in the following way:

Light directions: Proposed light source direction $\mathbf{x}_l^{dir}$ is generated by drawing a sample from a von Mises-Fisher distribution [6] with mean direction $\hat{\mathbf{x}}_l^{dir}$ and concentration parameter κ_{sample}, where $\hat{\mathbf{x}}_l^{dir}$ is the current light direction estimate. The estimate from the voting algorithm is used for the first iteration. In our experiments, $\kappa_{sample} = 200$ was chosen and samples were drawn using the accept-reject algorithm.

Light intensities: The proposed intensity for light source l is computed from the current light source intensity estimate adding a random offset, drawn from a normal distribution. The same method is used for ambient intensity proposals x^{α_0}.

Geometry parameters: The parameters used to define the geometry of an object are azimuth rotation, 3D translation, 3D scale and a geometry class label. This means that geometry for an object is defined by 7 scalars and 1 discrete value. The scalar values are drawn from normal distributions with the current value of the respective parameters used as the distribution mean. The geometry class label is drawn from a uniform distribution for each proposal.

The final solution corresponds to the light parameter sample s that generated the labeling with the lowest energy:

$$\mathbf{x}^{opt} = \arg\min_s E^{(s)}(\mathbf{x}). \quad (24)$$

This method is more tolerant to local minima in the model energy (which appear often in practice) and it requires a limited number of the costly evaluations of energy $E_I(\mathbf{x}|\mathbf{x}_\mathcal{L},\mathbf{x}_\mathcal{O})$.

5 Shadow Cues

In our discussion so far, we have assumed that some per-pixel estimate $\hat{\mathbf{I}}_s$ of the shadow image $\mathbf{I}_s$ is available to be used as input in our MRF model. In this section we explain how we obtain this initial estimate of shadow intensity.

We detect shadows by examining the change of image features across the borders of potential shadow regions. We start from the observation that light sources affect the whole image in a consistent way; therefore, edges due to cast shadows will generally exhibit characteristics that are consistent across the whole image, while edges due to other effects, such as albedo variations, will exhibit a more random behavior. To aid in the detection of shadows, we also utilize an appropriate measure of brightness, the bright channel [20]. In this section, we explain our approach to initial shadow detection in detail.

5.1 Bright Channel

We first extract a measure of brightness from the image, the *bright channel* cue [20] (similar to [10]):

$$I_{bright}(i) = max_{c \in \{r,g,b\}} \left(max_{j \in \Omega(i)} (I^c(j)) \right) \tag{25}$$

where $I^c(j)$ is the value of color channel c for pixel j and $\Omega(i)$ is a rectangular patch of size $m \times m$ pixels, centered at pixel i (in our experiments, $m = 6$).

The bright channel cue is based on the following intuition: The image values in patch $\Omega(i)$ are bounded by the incident radiance and modulated by the albedo at each pixel. However, in natural images, often a patch will contain some pixels with albedo that has high values in at least one color channel. By maximizing over color channels over all pixels in the patch, we reduce the effect of local variations of albedo within the image patch, getting a measure of brightness which is closer to the incident radiance at pixel i than the brightness at that pixel only.

We post-process the bright channel by choosing a white point I^{β}_{bright}, such that at least β % of the pixels are fully illuminated, corresponding to bright channel values of 1.0 (in our experiments, $\beta = 20\%$). Then the adjusted bright channel values $\dot{I}_{bright}$ are:

$$\dot{I}_{bright}(i) = min \left\{ I_{bright}(i)/I^{\beta}_{bright}, 1.0 \right\} \tag{26}$$

Furthermore, the max operator in Eq. 25 implies a dilation operation, meaning that the dark regions in the bright channel image appear shrunk by $m/2$ pixels ($m \times m$ is the size of patches $\Omega(i)$). We correct this by expanding the dark regions in the bright channel image by $m/2$ pixels, using an *erosion* morphological operator [8]. An example of the bright channel is shown in Fig. 1.b.

5.2 Shadow detection

As mentioned above, we take advantage of the global nature of the effects of illumination to detect cast shadows. For example, if we examine features like the brightness ratio or the hue difference across the two sides of shadow edges, in a scene with a single light source we will notice that the values we observe are concentrated around a clearly defined center. Intuitively, the shadows are similarly dark and exhibit a similar color change everywhere when they are caused by the same light source. On the other hand, the same features across the sides of non-shadow edges are distributed in a much more random way in most images, because they are caused by albedo variations and other effects that are local in the image. The distribution of such features exhibits peaks that correspond to shadow borders in the image. Our goal is to detect such peaks.

All our computations to obtain confidence values for shadows are based on comparing image features on the two sides of potential shadow borders. To improve the robustness of such computations, when examining values on the two sides of pixel i lying on the border of segment S_j, we compare the *average* of values on two semi-circular patches P^i_{in} and P^i_{out} centered at pixel i, and oriented so that P^i_{in} is inside segment S_j and P^i_{out} is outside, as seen in Fig.1.d. We examine only border pixels where the ratio of the average bright channel value between the two patches P^i_{in} and P^i_{out} is larger than θ_e or smaller than $1/\theta_e$, to ignore pixels that do not correspond to image edges (in our experiments, $\theta_e = 1.2$).

We first obtain a segmentation $\mathcal{S}$ of the bright channel image $\dot{I}_{bright}$ [3]. From the set of segments in $\mathcal{S}$, we choose a subset of segments that are ”good candidates” to correspond to shadow regions. We define a ”good candidate” for shadow as a segment where all three RGB color channels reduce in value across most of its edges, as we move from outside the segment towards the inside. We compute the confidence $q_{cand}(S_j)$ that a segment S_j is a ”good candidate” to be a shadow as:

$$q_{cand}(S_j) = 1/|S_j| \sum_{i \in S_j} q(i; S_j), \tag{27}$$

where $q(i; S_j) = 1$ if the average of r, g and b color channels in P^i_{in} is darker than P^i_{out}, and 0 otherwise.

Let f be the chosen feature across segment borders (bright channel ratio or hue difference in our experiments) that depends on illumination. We create a histogram h^{all}_f of the values of feature f at all segment border pixels. We also create a histogram h^{good}_f of the values of feature f at each border pixel i of each segment S_j, where each border pixel i contributes to the histogram proportionally to the confidence $q_{cand}(S_j)$. These two histograms represent the distribution of the values of feature f over all segment borders and over only segment borders that may be shadows. Normalizing them and taking their difference gives us a third histogram h^{diff}_f which corresponds to peaks in the distribution of feature f at borders in the set of ”good candidates” that are not prominent in the distribution of f in the set of all segment borders. We expect that these peaks will correspond to the characteristics of the shadows: for example, if f is the bright channel ratio, then the peaks in h^{diff}_f will indicate how dark the shadows in the image are.

Based on the extracted histograms, we compute a confidence for each segment to correspond to a shadow. We approximate the distribution of feature f in h^{diff}_f by a mixture of normal distributions. Each

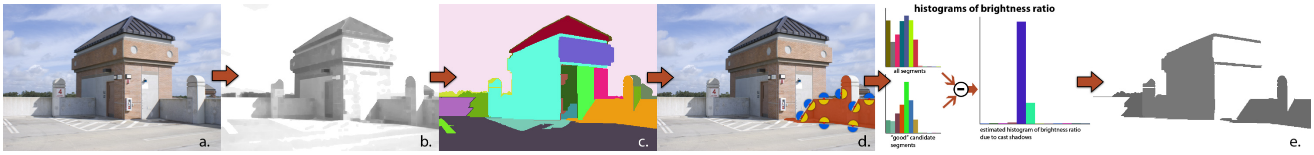
a.
b.
c.
d.
histograms of brightness ratio
all segments
"good" candidate segments
estimated histogram of brightness ratio due to cast shadows
e.

Fig. 1. Shadow detection: a. original image (from [31]); b. bright channel; c. segmentation; d. for each segment border pixel, feature values are compared between two patches inside (yellow) and outside(blue) the segment; then we form histograms of the features observed for all segments, and for segments that are good candidates to correspond to shadows, and compute the difference of the two distributions; e. the final shadow estimate.

component k of this mixture model is characterized by mean μ_k^f, variance σ_k^f and mixing factor π_k^f. We estimate these parameters through an Expectation-Maximization algorithm. To choose the number of distributions in the mixture we minimize a quasi-Akaike Information Criterion (QAIC). The confidence, based on a feature f, for segment $S_j \in \mathcal{S}$ is then defined as:

$$p^f(S_j) = \frac{1}{|\mathcal{B}_j|} \max_k \sum_{i \in \mathcal{B}_j} P_k\left(\Delta f(\mathcal{P}_1^i, \mathcal{P}_2^i)\right), \qquad (28)$$

where $\mathcal{B}_j$ is the set of all border pixels of segment S_j, k identifies the mixture components, and, for patches $\mathcal{P}_1^i$ and $\mathcal{P}_2^i$ on the two sides of border pixel i, $P_k\left(\Delta f(\mathcal{P}_1^i, \mathcal{P}_2^i)\right)$ is the probability of observing the difference $\Delta f(\mathcal{P}_1^i, \mathcal{P}_2^i)$ in the average value of feature f between the two patches $\mathcal{P}_1^i$ and $\mathcal{P}_2^i$, according to mixture component k (and weighed by the mixture factor π_k).

If we know that there is only a single light source, as in the case of outdoor scenes, we can improve performance further by fitting a single normal distribution centered at the highest peak of h_f^{diff}.

The features used in our work are the bright channel value ratio and hue difference across patches $\mathcal{P}_1^i$ and $\mathcal{P}_2^i$. We compute the final confidence $p(S_j)$ that segment S_j is a shadow as:

$$p(S_j) = q_{cand}(S_j)\left(p^{bright}(S_j) + p^{hue}(S_j)\right)/2. \qquad (29)$$

The shadow intensity for a segment S_j is computed as the median of the bright channel value ratio of patch pairs inside and outside the segment (Fig.1.e), assuming shadows are cast on a roughly flat surface.

This process is based on a segmentation of the image. In order to reduce our method's dependency on the quality of segmentations, we compute confidence values for different initial segmentations of the image. The final confidence value at pixel i is the mean of confidence values computed from each segmentation. Shadow detection can then be performed by thresholding the confidence value at each image pixel. In our experiments, we chose the threshold for shadow detection to maximize the classification rate on 100 training images from the UCF dataset [31].

6 Experimental Validation

In this section we present results with our approach. We first evaluate our shadow detection approach used to obtain an initial shadow estimate. We then evaluate illumination estimation with the proposed MRF model, both quantitatively in a synthetic dataset, as well as qualitatively in real datasets. Finally, we present results when geometry parameters are estimated jointly with shadows and illumination.

6.1 Shadow Cue Evaluation

We evaluated our shadow detection approach quantitatively on the UCF dataset [31], which consists of 356 images and manually annotated ground truth for the cast shadows, using the same set of 123 test images as [31]. We also evaluated our approach on the 135 image dataset of [16]. In Fig.2 we show ROC curves with our method on both datasets and compare with [31], [4] and [16].

Our method performs similarly to [31] and significantly better than [4], which is affected by the low image quality and unknown camera sensors. One reason for the difference in performance to [16] is that the annotation of the ground truth in the dataset of [16] generally includes edges of cast but not attached shadows, whereas our method does not differentiate between the two. When the shadow is partially cast and partially attached, the ground truth in [16] contains only the partial boundary that corresponds to the cast shadow and thus cannot be matched correctly by our method that produces always closed shadow borders. In Tab. 1 we show pixel classification rates on the 123 test images from UCF dataset. To obtain these classification rates, we chose the decision threshold (see Sec.5.2) as the optimal threshold for a different set of 100 training images from the UCF dataset. The results show that our method is comparable to much more complex approaches. The average running time of our method for the test images in the UCF dataset is 2.7 sec which compares very favorably to the other methods.

The results in Fig.2 also justify our choice of the bright channel compared to simple image brightness (from the HSV color model), by examining the performance of each in shadow detection when used with simple thresholding.

6.2 Illumination Estimation

We used three different datesets to evaluate the performance of illumination estimation: images collected

method	our method (bright channel ratio)	our method (hue)	our method (combined)	[31]
classification rate	87.7%	86.7%	**89.1%**	88.7%

TABLE 1

Pixel classification results with our method using different features, and with [31], on the UCF dataset ([31]).

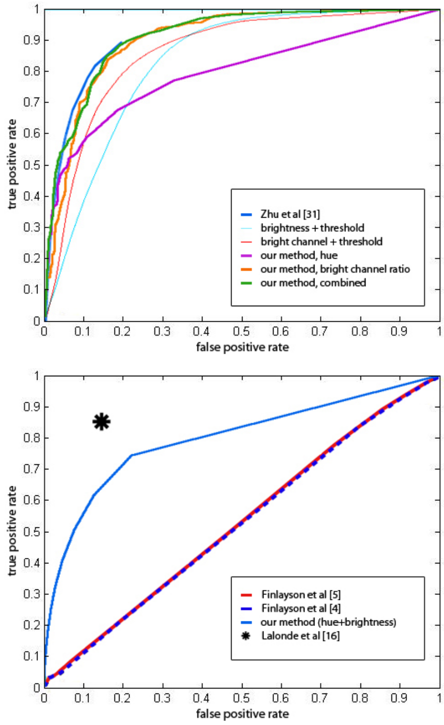

Fig. 2. Comparison of our shadow detection method with different features and methods ([31], [16], [4]). ROC curves computed on the datasets of [31] (top) and of [16] (bottom).

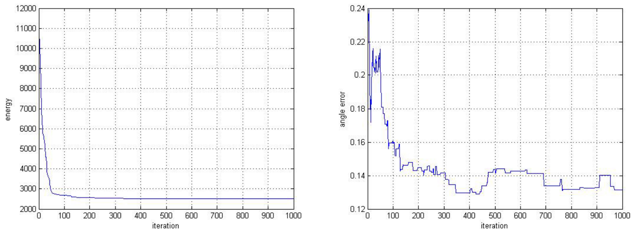

Fig. 3. Convergence of our algorithm. Left: The energy $E(x)$ for each iteration, averaged over a set of synthetic test images (using approximate geometry and added noise to the initial shadow estimate); right: the angular error per iteration, averaged over the same test set.

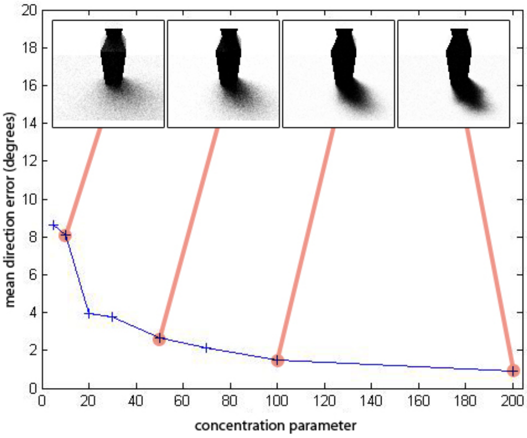

Fig. 4. Behavior for soft shadows. Illumination was modeled by a vMF distribution of varying concentration κ to produce sets of images with shadows of varying "softness". Even for very "soft" shadows, the error (in degrees) in the light source direction estimate is relatively small. Examples of the images produced for sample κ values are shown on the top.

under controlled illumination conditions, real-world images of cars collected from Flickr, and the Motorbike images from Caltech 101 [17]. We overlayed a synthetic vertical pole (sun dial) onto the original images, rendered under the illumination estimated by our method, in order to visualize the results.

The weights used in our experiments were: $(w_s, w_l, w_p, w_c) = (8, 1, 1, 4)$. The upper bounds for the truncated potentials were $(t_p, t_c) = (0.5, 0.5)$. Pixel node labels were quantized to 8 values and 1000 iterations of our algorithm were performed. Illumination estimation takes 5 to 30 minutes per image for the images in this paper, depending on image size. However, 60% to 70% of the running time is spent performing raytracing, which can be sped up significantly with a faster raytracer implementation.

6.2.1 Synthetic Dataset

To evaluate our method quantitatively we used a set of synthetic images, rendered using a set of known area light sources. The number of light sources was randomly chosen from 1 to 3. The direction and intensity of the light sources was also chosen randomly. We examined three different cases:

Exact geometry: We used the same 3D model to render the image and to estimate illumination.

Approximate geometry: We used a bounding box and a ground plane that coarsely approximated the original geometry to estimate illumination.

Approximate geometry and noisy shadow input: We estimated illumination parameters using a coarse 3D model, as above, and a noisy initial shadow estimate. The latter was obtained by adding random dark patches to the rendered shadow (Tab. 2.c). We used such noise because, on one hand our methods are relatively insensitive to spatially-uniform random noise, and on the other hand, in real data the errors generally affect large image regions which get mislabeled, which is emulated by this patch-based noise.

We computed the difference between the estimated light source parameters and the parameters of the true light source that was closest in direction to the estimated one. The average light source direction errors are presented in Tab. 2. We compare the results from the voting method used to obtain the

	a. Exact geometry			b. Approx. geometry			c. Approx. geometry + noisy shadow input		
#lights:	1	2	3	1	2	3	1	2	3
Voting	7.06	6.94	8.23	5.83	11.51	13.31	20.78	28.61	29.30
NNLS [26]	3.84	6.20	6.35	13.95	15.21	14.15	33.69	32.10	33.96
MRF(HOCR [11])	3.29	5.41	8.13	5.14	14.67	13.99	14.35	20.60	22.83
MRF(2-stage minim.)	**0.44**	**1.31**	**2.36**	**2.53**	**9.06**	**8.57**	**6.97**	**12.36**	**17.77**
MRF(2-stage minim.) - w/o shadow shape prior	1.27	3.82	5.40	3.11	11.12	11.95	10.81	12.24	17.91
Number of light sources: mean error (error %)	0 (0%)	0.047 (4.7%)	0.143 (14.2%)	0 (0%)	0.309 (17.6%)	0.32 (23.8%)	0 (0%)	0.285 (26.7%)	0.33 (38%)

TABLE 2

Synthetic results. From left to right, we show the mean error (in degree) for the estimated light directions on a synthetic dataset: a) using the exact geometry; b) using geometry approximated by bounding boxes (blue) and a ground plane; c) using approximate geometry and a noisy initial shadow estimate. For each case, we show results for scenes rendered with 1, 2 or 3 light sources. We show results obtained with the voting algorithm used for the initialization; with NNLS [26]; with our MRF model, when the MRF energy is minimized using [11]; and when the MRF energy minimized using our 2-stage approach, which achieves the best results. We also include results with our MRF model and 2-stage approach without the shadow shape-matching prior, which shows the benefits of this term. In the bottom we show the mean error in the estimated number of light sources and in what portion of images that the number was estimated inaccurately.

initial estimate, and our MRF model. We compare the proposed inference method with a state-of-the-art method to perform inference on higher-order MRF models, the higher-order clique reduction (HOCR) technique of [11]. The results show that our method, taking advantage of the topology of this particular MRF model to efficiently perform inference, is able to achieve significantly better results, compared to our initialization method, HOCR inference on our model, as well as the non-negative least squares optimization approach of [26] (NNLS).

Furthermore, Tab. 2 shows that the shadow shape-matching prior significantly improves illumination estimates. This is more pronounced in the case of inaccurate input data, where a large number of pixels may be different between the noisy observed shadow and the one produced by the coarse geometry and true illumination. However, when there are multiple light sources, leading to a large number of potential shadow edges, the benefits of the shadow shape-matching prior are reduced.

We also evaluated the estimation of the number of light sources through our voting procedure on our synthetic dataset. Tab. 2 shows the mean error in the estimated number of light sources in that dataset. We are generally able to get a good estimate of the number of light sources. The accuracy of that estimate is reduced when the true number of light sources and the errors in the initial shadow estimate increase. We further evaluated our light source number estimation on the motorbike images of Caltech 101. The images we selected contained a single light source (the sun) and the average estimated number of light sources was 1.17, with the number of light sources correctly estimated 91% of the time. We should also note that any extraneous light sources identified by our voting algorithm are generally assigned low intensities during MRF inference, resulting in small errors in the synthesized cast shadows.

We further quantitatively evaluated the behavior of our method in the case of soft shadows. We rendered the set of synthetic scenes under illumination produced by a single light source modeled by a vMF distribution of varying concentration parameter κ. Lower values of κ mean a more spread-out light distribution and softer shadows. Fig.4 shows the error in the estimated light source direction as the concentration parameter of the light source changes. Even in the case of very soft shadows, our method is able to estimate the direction of illumination with good accuracy.

6.3 Geometry Reasoning

6.3.1 Real Datasets

To evaluate our approach in real images, we used the class "Motorbikes" of the Caltech 101 dataset [17] and images of cars we collected from Flickr.

In the case of "Motorbikes", we used *the same* coarse 3D model (Fig.11) corresponding to an average motorbike and *the same* average camera parameters for every image. In this dataset there are significant variations in geometry, pose and camera position in each individual image, deviating from our average 3D model and camera parameters. Despite these variations, our results in Fig.5 show that our algorithm is robust enough to effectively estimate illumination using the same generic 3D model for all instances of a class of objects.

In the case of car images collected from Flickr (Fig.6), the geometry was limited even further to a

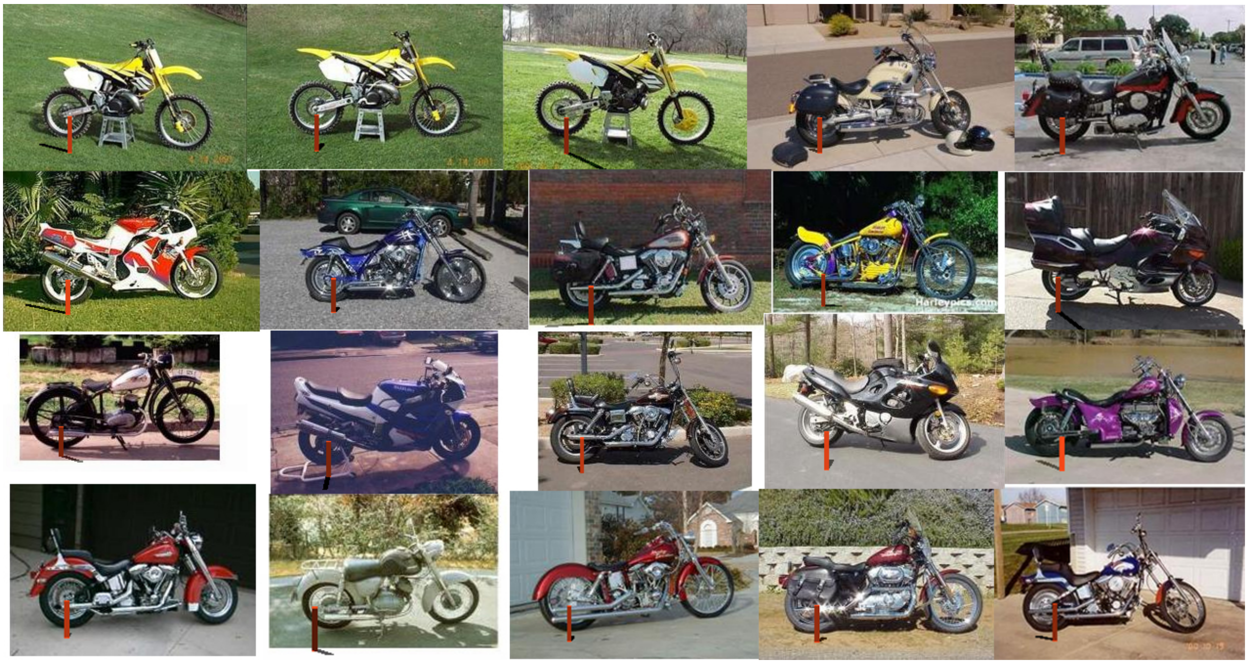

Fig. 5. Results for the Motorbikes class of the Caltech101 dataset. We rendered a synthetic sun dial (orange) under the estimated illumination and overlayed it on each original image. The geometry used for all instances was the same 3D model capturing an average motorbike, with the same common camera parameters.

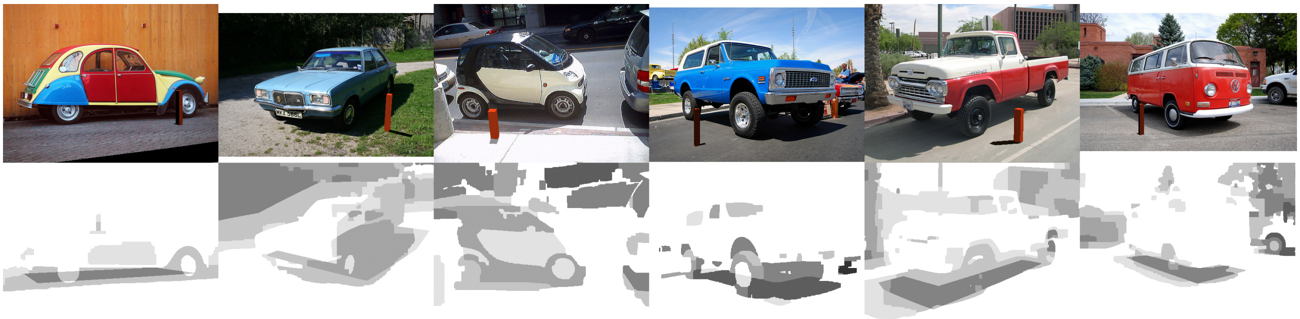

Fig. 6. Results with car images from Flickr. Top row: the original image and a synthetic sun dial rendered with the estimated illumination; Bottom row: the final shadow values. The geometry consists of the ground plane and a bounding box for the car.

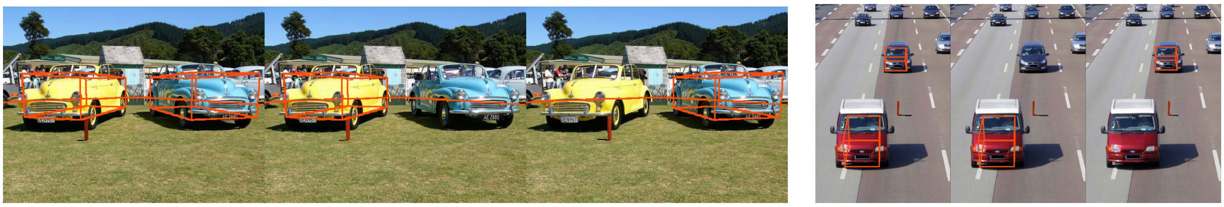

Fig. 7. Examples of scenes with more objects: the orange bounding boxes show the geometry provided as input to our method, and the synthetic orange sundial rendered using the estimated illumination shows our light source estimate. The illumination estimate is very stable regardless of which part of the scene we choose to model.

bounding box approximating to the car body and a ground plane (Fig.11). Camera parameters were matched manually. For both Fig.6 and Fig.8 we assumed known number of light sources. Despite our initial assumption of Lambertian reflectance, the results show that our algorithm can cope with the abundance of non-lambertian surfaces in these images.

We further evaluated our algorithm in a set of images captured under controlled illumination conditions in the lab. This set includes shadows cast on a variety of textured surfaces, under 1, 2 or 3 light sources. Results on images from this dataset can be found in Fig.8. To estimate the illumination in this images we used rough approximate geometry, which can be seen in Fig.11. In Fig.8 we also show two synthetic examples of illumination estimation where shadows are cast on arbitrary geometry, demonstrating that we do not make any assumptions about scene geometry.

Fig.9 shows common cases where our algorithm fails. One general reason is challenges in shadow detection. While the shadow shape-matching prior helps our method differentiate between adjacent shadows

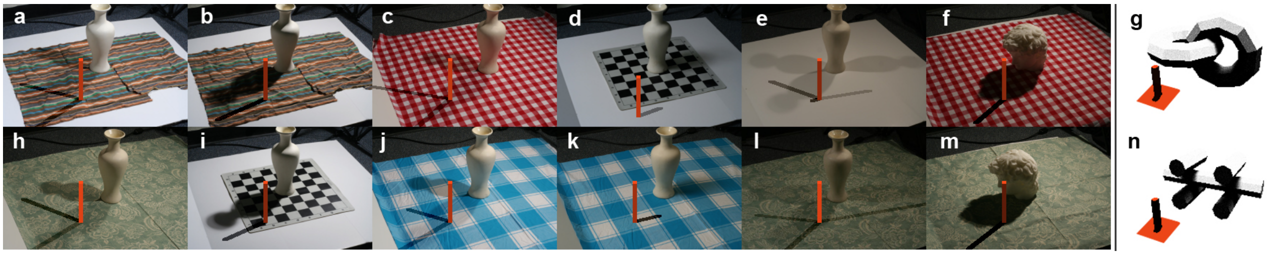

Fig. 8. Results on images captured using different background textures and number of lights. The vertical image pairs (a,h),(b,i),(c,j),(d,k),(e,l),(f,m) are captured under the same illumination. An orange synthetic sundial has been rendered under the estimated illumination and inserted into the original image. We also show two results on synthetic images (g,n) where the input image was used as the initial shadow estimate, without using the shadow detection method of Sec. 5. These images show that our MRF illumination model can be applied to arbitrary scene geometry, where shadows are not cast on a flat ground (mean light direction error for g,n is 2.27 degrees).

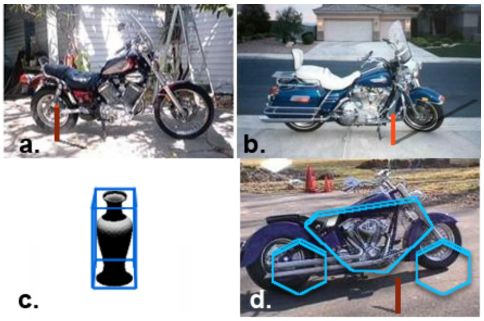

Fig. 9. Common failure modes. Errors due to shadows (top): (a) shadows of other objects not modeled may overlap the shadows of the objects of interest, or (b) very dim shadows may not be detected, in which case our algorithm tries to use other dark image regions. Errors due to geometry (bottom): (c) approximate geometry (in blue) cannot explain the observed shadows for any choice of illumination direction, since the approximate geometry (blue) fully covers the cast shadow. (d) Large errors in the positioning of geometry in the scene (when geometry parameters are not estimated) affect the relative position of shadows in the image to the object geometry.

from different occluders, it can still be challenging to correctly estimate illumination when shadows from objects that are not modeled by the geometry are very close to or overlap shadows of interest. Furthermore, very dim shadows, as in the case of cloudy outdoor scenes, can be hard to detect, therefore not allowing us to obtain a good solution. On the other hand, coarse geometry knowledge can sometimes lead to observed shadows that cannot be explained under any illumination configuration given the coarse geometry (as in Fig.9.c). Inaccuracies in the placement of 3d models in the scene (e.g. with the Caltech 101 ”Motorbike” images) or in the camera parameters can also lead to inaccurate illumination estimates (Fig.9.d). Light sources close to the horizon also cause inaccuracies, because they generate long shadows which reside in large part out of the image, offering ambiguous image evidence about the illumination direction.

We evaluate joint illumination and geometry/pose estimation qualitatively on the car images we collected from Flickr, as seen in Fig.10. The input to our algorithm in this case was the original image, a 2D bounding box around the object of interest (in this case, the car), a common ground plane, the camera parameters and a common set of 4 candidate geometric models for cars (shown in Fig.10). The geometric models represent 4 common car shapes. The 2D bounding box can be provided by a car detector. The camera parameters are very similar across these images, probably because of the common subject, and could approximated automatically using the information in the image EXIF tag, along with horizon line estimation (and assuming the camera is at eye level of an average human). In our experiments shown in Fig.10 however, we set camera parameters manually.

For experiments with geometry parameter estimation we did not use our voting initialization method, because the random initial geometry reduces the benefits of such an initialization. We assumed a single light source and used a random initialization of the other parameters. A larger number of iterations (4000) was performed to obtain a solution, with larger variance for the generation of light parameter proposals. Despite the random initialization, our MRF model is able to obtain a satisfactory solution.

Our results show that we can approximate the orientation of the object with good accuracy (around 10 degrees), and get visually convincing estimates of scale and orientation. The object geometry is identified correctly in 3 of the 4 images of Fig.10. Notice that although we could fit an infinite number of very different (and mostly incorrect) combinations of geometry/rotation/translation/scale values to the object outline obtained by GrabCut, as shown in Fig.10.b, the combination of the object outline and the shadow leads our algorithm to select parameter combinations close to the truth (Fig.10.c), while estimating the illumination at the same time. In some cases the pose estimate further improves when when combined with geometry class estimation.

An important observation is that, as the number of free parameters that define geometry grows, local minima in the energy become a bigger issue. An example of this problem is the fourth image in Fig.10.d, where the geometry class used for the pick-up truck corresponds to ”jeep”, and at the same time the size

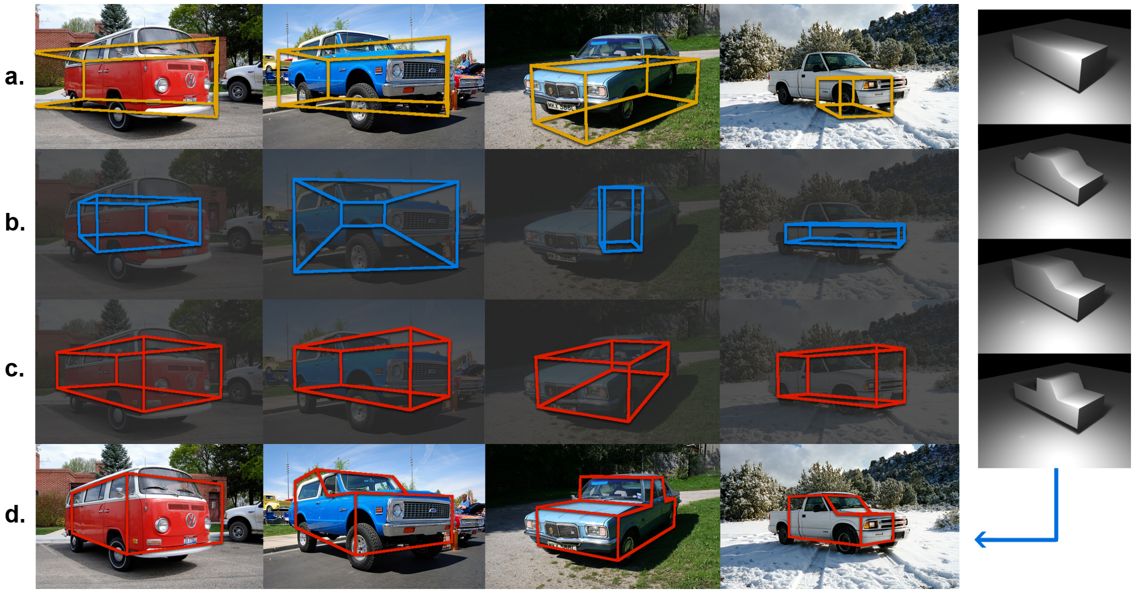

Fig. 10. Results of joint estimation of shadows, illumination and geometry parameters. The geometry used in this case consists of a ground plane and a bounding box for the object. The geometry parameters estimated are the azimuth rotation, 3D translation and 3D scale of the object's bounding box. a) input: the original image and the initial configuration of the geometry; b) the estimated geometry when only fitting the object to the mask obtained by GrabCut; c) the geometry estimated by our method. While the object silhouette is not enough to estimate the geometry parameters, the combination of the object silhouette with information in the shadows allows us to obtain a good geometry estimate. d) Here we also allow our model to select the most probable of 4 candidate geometry classes. The estimated geometry class for each image is, from left to right: *box*, *jeep*, *sedan*, *jeep*. The 4 geometry classes are shown on the right.

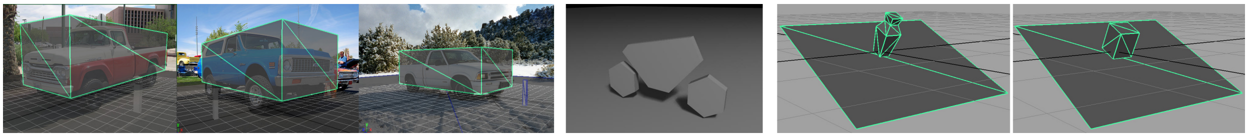

Fig. 11. The 3D models for the experiments on cars (Fig.6), motorcycles (Fig.6), and the images of Fig.8.

chosen for the model omits the rear part of the pick-up truck. In this case our algorithm has found a local minimum of the energy; to continue to the global minimum, a large change in scale and translation along with the change in the selected geometry class is needed. A clever selection of the dimensions which change to produce the new step on each iteration can help as the number of geometry parameters grow - for example, the geometry class could be locked to the simple bounding box for a number of iterations, expecting that the bounding box will be positioned properly over the object before we begin examining more specific geometry classes. Random initializations of geometry very far from the true geometry can also affect the final result, but constraining the initial pose within the GrabCut mask is often sufficient.

7 Conclusions

In this paper, we introduced a higher-order MRF model of illumination, which allows us to jointly estimate the illumination parameters, cast shadows and a set of geometry parameters for the occluders in a scene, given a single image. Our model incorporates both high-level knowledge about the scene, such as illumination and geometry, and low-level image evidence. Although this leads to a complex formulation that makes inference challenging, we demonstrate that inference can be performed effectively. Our results in various datasets demonstrate the potential of the proposed model. We are able to estimate the illumination parameters using the same geometry, pose and camera parameters for a large number of scenes which belong to the same class, as shown by the results on Caltech101. Bounding boxes can be sufficient approximations of occluders for our method, as is the case with our experiments with car images from Flickr. Geometry reasoning is also incorporated in our model to allow estimation of the object pose in the 3D scene, as well as reasoning about the 3D geometry that best represents the object. Our experiments show that the proposed approach is more general and more applicable in real-world images where other methods fail. In the future, we are interested in incorporating our method in more general scene understanding applications. Geometry parameter estimation, as presented here, is the first step towards this direction.

Acknowledgements: This work was partially supported by NIH grants 5R01EB7530-2, 1R01DA020949-01 and NSF grants CNS-0627645, IIS-0916286, CNS-0721701, IIS-1111047. Nikos Paragios's work was partially supported by the European Research Council Starting Grant DIOCLES (ERC-STG-259112).

References

[1] Y. Boykov and G.F. Lea. Graph cuts and efficient n-d image segmentation. *IJCV*, 70(2):109–131, November 2006.

[2] M. Bray, P. Kohli, and P. H. S. Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *In ECCV*, pages 642–655, 2006.

[3] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[4] Drew M.S. Finlayson, G.D and C.Lu. Entropy minimization for shadow removal. *IJCV*, 79(1):13–30, 2009.

[5] G.D. Finlayson, S.D. Hordley, and M.S. Drew. Removing shadows from images. In *ECCV*, 2002.

[6] R.A. Fisher. Dispersion on a sphere. *Proc. Royal Soc. London*, 217:295–305, 1953.

[7] S. Geman and D. Geman. *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, pages 452–472. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[8] R.C. Gonzalez and R.E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., 2006.

[9] K. Hara, K. Nishino, and K. Ikeuchi. Light source position and reflectance estimation from a single view without the distant illumination assumption. *PAMI*, 27(4):493–505, 2005.

[10] K.M. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. In *CVPR*, 2009.

[11] H. Ishikawa. Higher-order clique reduction in binary graph cut. In *CVPR*, 2009.

[12] T. Kim and K.S. Hong. A practical approach for estimating illumination distribution from shadows using a single image. *IJIST*, 15(2):143–154, 2005.

[13] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. In *CVPR*, 2009.

[14] N. Komodakis, G. Tziritas, and N. Paragios. Performance vs computational efficiency for optimizing single and dynamic mrfs: Setting the state of the art with primal-dual strategies. *CVIU*, 112(1):14–29, 2008.

[15] J.F. Lalonde, A.A. Efros, and S.G. Narasimhan. Estimating natural illumination from a single outdoor image. In *ICCV*, 2009.

[16] J.F. Lalonde, A.A. Efros, and S.G. Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *ECCV*, 2010.

[17] F.F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, April 2007.

[18] Y. Li, S. Lin, H. Lu, and H.Y. Shum. Multiple-cue illumination estimation in textured scenes. In *ICCV*, 2003.

[19] A. Panagopoulos, D. Samaras, and N. Paragios. Robust shadow and illumination estimation using a mixture model. In *CVPR*, 2009.

[20] A. Panagopoulos, C. Wang, D. Samaras, and N. Paragios. Estimating shadows with the bright channel cue. In *CRICV 2010 (in conjuction with ECCV'10)*, 2010.

[21] A. Panagopoulos, C. Wang, D. Samaras, and N. Paragios. Illumination estimation and cast shadow detection through a higher-order graphical model. In *CVPR*, 2011.

[22] Q. Dai R. Guo and D. Hoiem. Single-image shadow detection and removal using paired regions. In *CVPR*, 2011.

[23] R. Ramamoorthi, M. Koudelka, and P. Belhumeur. A fourier theory for cast shadows. *PAMI*, 27(2):288–295, 2005.

[24] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23:309–314, 2004.

[25] E. Salvador, A. Cavallaro, and T. Ebrahimi. Cast shadow segmentation using invariant color features. *CVIU*, 95(2):238–259, 2004.

[26] I. Sato, Y. Sato, and K. Ikeuchi. Illumination from shadows. *PAMI*, 25(3):290–300, 2003.

[27] C. Wang, M. De la Gorce, and N. Paragios. Segmentation, ordering and multi-object tracking using graphical models. In *ICCV*, 2009.

[28] Y. Wang and D. Samaras. Estimation of multiple directional light sources for synthesis of augmented reality images. *Graphical Models*, 65(4):185–205, 2003.

[29] Y. Yang and A. Yuille. Sources from shading. In *CVPR*, 1991.

[30] W. Zhou and C. Kambhamettu. A unified framework for scene illuminant estimation. *IVC*, 26(3):415–429, 2008.

[31] J. Zhu, K. G. G. Samuel, S. Masood, and M. F. Tappen. Learning to recognize shadows in monochromatic natural images. In *CVPR*, 2010.

Alexandros Panagopoulos received the PhD degree from the Computer Science Dept. at Stony Brook University in 2011 and the B.Sc. degree from the Computer Engineering and Informatics Dept. in the University of Patras in 2004. His research interests lie in computer vision and machine learning and more specifically in inverse rendering problems, shadow detection and the application of graphical models in computer vision.

Chaohui Wang received his PhD degree in applied mathematics and computer vision from Ecole Centrale Paris in 2011. He is currently a postdoctoral fellow at the Vision Lab of University of California, Los Angeles. His research interests include computer vision, machine learning, image processing and medical image analysis. More detail about his research work can be found at: http://vision.ucla.edu/~cwang/index.php.

Dimitris Samaras received the diploma degree in computer science and engineering from the University of Patras in 1992, the MSc degree in computer science from Northeastern University in 1994, and the PhD degree from the University of Pennsylvania in 2001. He is currently an associate professor in the Department of Computer Science at Stony Brook University, where he has been teaching since 2000. He is a Digiteo Chair at Ecole Centrale de Paris. He has been a visiting professor in Ecole Centrale de Paris, Ecole Central de Lyon and Universitat Autonoma de Barcelona. He specializes in deformable model techniques for 3D shape estimation and motion analysis, illumination modeling and estimation for recognition and graphics, and biomedical image analysis with emphasis on human behavioral data. He is a member of the ACM and the IEEE.

Nikos Paragios received the PhD (highest honors) and DSc (Habilitation a Diriger de Recherches) degrees in electrical and computer engineering from the University of Nice/Sophia Antipolis, France, in 2000 and 2005, respectively. Currently, he is professor of Applied Mathematics and Computer Science, director of the Center for Visual Computing of Ecole Centrale de Paris & Ecole des Ponts - ParisTech, member of the Laboratoire d'informatique Gaspard-Monge and scientific leader of GALEN group of Ecole Centrale de Paris/INRIA Saclay, Ile-de-France. He has coedited four books, published more than 200 papers in the most prestigious journals and conferences of computer vision and medical imaging, and has 17 US-issued patents. He is a fellow of the IEEE, an associate editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence, an area editor for the Computer Vision and Image Understanding Journal, and a member of the Editorial Board of the International Journal of Computer Vision, the Medical Image Analysis Journal, the Journal of Mathematical Imaging and Vision,the SIAM Journal of Imaging Sciences and the Imaging and Vision Computing Journal. He is one of the program chairs of the 11th European Conference in Computer Vision (ECCV'10). He is also member of the scientific council of SAFRAN conglomerate & the Intrasene group.