3D PRIORS FOR SCENE LEARNING FROM A SINGLE VIEW

By

Diego Rother

Kedar Patwardhan

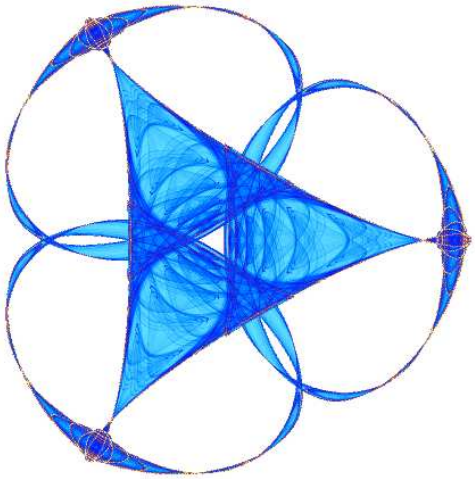
Iman Aganj

and

Guillermo Sapiro

IMA Preprint Series # 2210

(May 2008)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA

400 Lind Hall
207 Church Street S.E.
Minneapolis, Minnesota 55455–0436

Phone: 612-624-6066 Fax: 612-626-7370
URL: http://www.ima.umn.edu

To appear in the Search in 3D IEEE Workshop (held in conjunction with CVPR), 2008.

3D Priors for Scene Learning from a Single View

Diego Rother
University of Minnesota
diroth@umn.edu

Kedar Patwardhan
GE
kedar@umn.edu

Iman Aganj
University of Minnesota
iman@umn.edu

Guillermo Sapiro
University of Minnesota
guille@umn.edu

Abstract

A framework for scene learning from a single still video camera is presented in this work. In particular, the camera transformation and the direction of the shadows are learned using information extracted from pedestrians walking in the scene. The proposed approach poses the scene learning estimation as a likelihood maximization problem, efficiently solved via factorization and dynamic programming, and amenable to an online implementation. We introduce a 3D prior to model the pedestrian's appearance from any viewpoint, and learn it using a standard off-the-shelf consumer video camera and the Radon transform. This 3D prior or "appearance model" is used to quantify the agreement between the tentative parameters and the actual video observations, taking into account not only the pixels occupied by the pedestrian, but also those occupied by the his shadows and/or reflections. The presentation of the framework is complemented with an example of a casual video scene showing the importance of the learned 3D pedestrian prior and the accuracy of the proposed approach.

1. Introduction and motivation

Automated analysis of 3D scenes from 2D videos has attracted a lot of attention in recent years, due in part to the increasing volume of data being generated by surveillance and personal cameras, and fuelled by the computational power of today's computers. Yet, automatic 3D inferences are still challenging, especially when the source of data is a single uncalibrated camera, which is usually the case. Effectively addressing this problem would tremendously benefit numerous computer vision applications, e.g., tracking, detection and occlusion reasoning.

In this paper we present a 3D scene analysis framework that learns the camera transformation and the direction of the shadows using information extracted from pedestrians casually walking in the scene, as seen from a single view. The proposed approach poses scene geometry estimation as a likelihood maximization problem that is efficiently solved via factorization and dynamic programming, and is also amenable to an online implementation. Essentially our approach searches for a set of parameters (camera transformation, shadow direction and people's trajectories) that "*explains*" the input video. For this purpose, we introduce a novel model of the appearance of a pedestrian from a given viewpoint (we call it a *3D prior*), and use it to quantify the "agreement" of a tentative set of parameters with the actual video observations. This "agreement" takes into account not only the pixels occupied by the pedestrian, but also those of his shadows and/or reflections. In addition, we show how to use *a single* consumer video camera and the Radon transform to learn this 3D prior.

1.1. Previous work and key contributions

Knowledge of the camera transformation matrix provides essential information in order to make inferences about a 3D scene from its 2D projection. Numerous methods have been proposed to estimate this matrix. Tracking local features [1] is a successful approach for moving-camera calibration, which has even been integrated into commercial products (e.g., Boujou, from 2d3 [2]). To work, this method requires the tracked features to move in 3D (either because the camera itself is moving or because the objects in the scene move), and it may fail on a scene shot from a single still camera in which featureless objects move (a common case in surveillance). Here we propose to exploit complementary features (such as pedestrians and their shadows), either to address the problem when such systems fail or to improve the accuracy of such systems.

Other camera calibration methods, appropriate for still camera settings, use helpful structures naturally occurring in the scene (see [1, 3] for a review). Though these approaches have been relatively successful, they are not completely automatic and require some level of user intervention and/or modifications in the scene to aid in the calibration (e.g., addition of marks or lines).

In recent years, several techniques, e.g., [4-6], have been proposed to automatically estimate the camera projection matrix from *multiple* images from the *same* viewpoint, by means of *natural experiments.*[1] These techniques exploit the pedestrians in the scene (natural and common ingredients in ubiquitous videos) to compute the camera projection matrix using an offline or batch estimate. Lv *et al.* [4] were, to the best of our knowledge, the first to propose the use of pedestrians in the scene for auto-calibration. Their approach is sensitive to noise and requires motion constraints on the pedestrians in the scene. Krahnstoever and Mendonça, [5], later extended this approach by proposing a bayesian formulation that allowed for the incorporation of measurement uncertainties, outlier models, and prior information about the problem. This

[1] ***Natural experiment*** (from Wikipedia): A naturally occurring instance of observable phenomena which approach or duplicate a scientific experiment. […] a common research tool in fields where artificial experimentation is difficult [...]

approach is non-linear and requires prior knowledge about unknown camera parameters and the location of the people in the scene. Junejo and Foroosh proposed using harmonic homologies [6], and more recently epipolar geometry [7], to improve the scene geometry estimation from the observed pedestrians in the scene. With a similar aim of understanding the 3D placement of foreground objects in a 3D scene, although from a single image, Hoiem *et al.* [8] proposed an elegant approach which uses priors about the appearance and location of scene elements in order to *place* foreground objects in *perspective*.

The aforementioned approaches [4-6], are generally focused on foreground objects in the scene and disregard other important and natural ingredients such as shadows and reflections. Ignoring these elements often results in an incorrect detection of the pedestrians in the scene. In contrast, we consider that shadows (as in [9-11]) and reflections are in fact useful clues that can be exploited to make inferences about a 3D scene, in collaboration with information gleaned from the pedestrians themselves.

In this direction, our previous work (in [9]) provides a simple framework for *online* learning of 3D scenes. Drawbacks of this approach include its use of 2D (and not 3D) templates to model the foreground, limiting its applicability to nearly horizontal camera angles; the closed loop topology of the approach that makes it prone to local optima; the lack of temporal integration among observations that leads to noisy estimates in the feet positions (whose correct estimation is critical); and the independence between the feet/head detection step and the geometry estimation step, that prevents the mutual benefit that can result from their simultaneous estimation.

Regarding pedestrian appearance models or priors, numerous ideas have been proposed in the literature. Notable ones are those where the pedestrian is regarded as a collection of loosely-connected limbs or articulated parts (e.g., [12]), models of the silhouette of the pedestrians (e.g. [13]), models of the interior or 2D priors (e.g. [9, 14]), or distributions of local features (e.g. [15]). For a more detailed review of human models please refer to [16]. The goal of our work is not pedestrian detection/recognition or pose-estimation (as is in the works mentioned above), but to compute the consistency between the pedestrians (objects) as seen in the video and their simulated image from a hypothetical camera view, to test whether the hypothetical camera view explains the pedestrians' observations. Since this consistency between the pedestrians and the camera view is computed in the innermost loop of the framework, it is very important that this is carried out in a computationally in-expensive manner. The object model proposed by Savarese and Li [17] can be used to compute the consistency pedestrian-camera view but is computationally too expensive for our purposes. Hence, in this work we introduce a 3D occupancy prior for pedestrians, which can be efficiently computed and at the same time provides a way of fusing information from the pedestrian, his shadows and/or reflections. In addition, this 3D prior is simple to acquire/construct with a single camera.

Temporal integration among observations is not a new idea in the area. Relevant to our discussion are tracking systems that solve a Hidden Markov Model (HMM) formulation using *dynamic programming* (e.g. [18-20]). Two peculiarities of our proposed HMM implementation are: 1) that it tracks *individuals* using all their observations (processing is carried out once all the observations of an individual are available); and 2) that each HMM state consists of a *pair* of ground locations (i.e. one state encodes the fact that the individual is currently at one location and visited another location in the previous step) representing position *and* velocity of an individual at a given time.

Here, we attempt to address the shortcomings of previously mentioned approaches by: 1) using 3D occupancy-priors of pedestrians to "explain" their 2D appearance in a scene, allowing for a better geometry estimation as detailed in sections 2 and 4; 2) unifying the feet/head detection and geometry estimation into a single step, where the feet/head position contributes to the estimation of the scene geometry and vice-versa, as described in Sec. 3; 3) integrating temporal information from different observations efficiently, to discard outlier detections (Sec. 3.3); 4) efficiently searching the entire parameter space of the scene for the global optimum, thereby avoiding possible local optima; and 5) utilizing the optimal substructure of the proposed likelihood function to derive a low complexity optimization scheme which is amenable to an online implementation.

Note that learning the scene geometry is not the only contribution of this work, since the proposed 3D model and the inclusion of other elements such as shadows and reflections have intrinsic value, and camera calibration is just one possible application for them.

2. Computing 3D priors

The appearance of an object in an image depends on its 3D orientation and position relative to the camera. 3D priors capture the distribution of the object's mass in space and can be efficiently projected to any 2D view to simulate the space occupied by the object in the view. Specifically, a 3D prior is a 3D matrix in which each element contains the probability that the corresponding voxel in a 3D box in space, surrounding the object, is occupied.

In this work we develop and use 3D priors of humans in order to obtain the likelihood of a person's observation from a candidate camera position. This likelihood is a quantitative measure of consistency between the (candidate) scene geometry and the actual observations, and is paramount to estimate the camera transformation in the proposed framework (as explained in Sec. 3).

2.1. Mathematical preliminaries

A 3D prior is a three dimensional collection of Bernoulli random variables, one for each point inside a 3D box in world space[2]. Each Bernoulli variable describes the probability that an object placed inside the box (horizontally centered and lying on the box's floor) will contain the corresponding world point. We consider a discretization of the 3D box into voxels (see Fig. 1), and compute for each voxel the average (occupancy) probability of observing the object (a person walking) in the region enclosed by each voxel.[3] We assume these voxel occupancy probabilities to be independent, an assumption that while only approximately true, considerably simplifies the model, preventing the size of the stored probability tables, and then the memory and computational requirements, from growing excessively with the number of voxels.

Consider p_i as the probability of the object occupying voxel v_i in the 3D box, and P_j as the probability of observing the object at pixel Q_j in the image plane. Let r_j be the ray that originates at a camera center C and passes through Q_j and voxels $v_1, \cdots, v_n$ (Fig. 1). From the independence assumptions and basic rules of probability, it follows that these quantities are related by

$$\log(1 - P_j) = \sum_i R_{ji} \log(1 - p_i)\,, \qquad (1)$$

where R_{ji} is the contribution of voxel v_i to the ray r_j, which is proportional to the length of the intersection between the ray and the voxel. These contributions depend on the camera matrix considered, the position of the box, and the number of voxels in the box along each dimension.

In order to compute the probability of finding the object at a pixel of the image plane (its *foreground 2D prior*) given its tentative 3D location, the 2D projection of a 3D prior is computed for every ray (pixel) that intersects the 3D prior's box, using Eq. (1) and the tentative camera matrix. This process has $O(N_P.\bar{L})$ complexity, where N_P is the number of rays intersecting the 3D prior's box, and $\bar{L}$ is the mean number of voxels intersected by the rays. N_P increases with the square of the video resolution, and therefore so does the computational cost. This computation can be parallelized to run extremely fast on a GPU.

The pixels in the image plane that correspond to the pedestrian's shadow are also related by a projective transformation to the 3D space occupied by the pedestrian. Following [10], we refer to this transformation as the *shadow camera*, since as expected from a camera transformation, it maps points in 3D space to the image plane. This transformation is the composition of a central projection from 3D space to the floor plane, whose center is the light source position, followed by the homography from the floor plane to the image plane. Note that to compute the shadow camera, only the light source position is needed, on top of the camera matrix. To obtain the *shadow 2D prior* (the probability that a pixel in the image belongs to the pedestrian's shadow), the shadow camera is used in the projection.

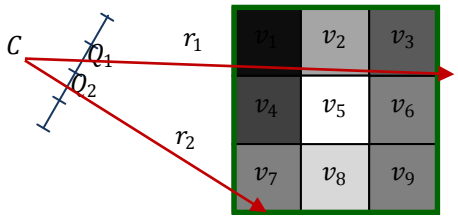


Fig. 1: 2D analog of a 3D prior. Rays r_j, originating at the camera center C, pass through corresponding pixels Q_j in the image plane (in blue) and intersect voxels v_k inside the 2D box (in green).

Analogously, we define the *reflection camera* as the composition of a mirror-plane symmetry and the camera matrix. In scenes having shiny floors or mirrors, this transformation is used to construct the *reflection 2D prior*.

The 2D priors of the foreground, its shadow, and reflections, are all projected from the *same* 3D prior, using the *same* algorithm, simply with different projection matrices. In contrast for example with [9], a single 3D prior is projected to obtain 2D priors from *any* view. This is a critical contribution of this work.

2.2. Learning the 3D prior

Learning a 3D prior amounts to estimate the p_i's, which are the elements of the 3D prior, from the given P_j's, which are the projections of the 3D prior to 2D (we explain below how to obtain them). We use a variation of the *method of moments* [21] for this estimation, matching the unknown values obtained along each ray (rhs of Eq. (1)), to the empirical values (lhs of Eq. (1)). Each pixel (ray) in a video provides an equation in the form of (1)[4] and the estimation of the probabilities p_i can thus be reformulated as a linear system of equations, $y = Rx$, where y is the vector of all the observations, $\log(1 - P_j)$, x is the vector of all unknowns, $\log(1 - p_i)$, and R is the sparse matrix of contributions, R_{ji}, indicating how much each voxel in the 3D prior contributes to each measurement (computed from the camera matrix and the box's size and position). We solved this equation by minimizing the robust L_1 error norm, $\text{argmin}_x \|Rx - y\|_1$. This is a convex function of x, and we used the gradient descent technique to find its global minimum.[5]

[2] In all our experiments we chose the 3D prior's box to be 1×1×2m, large enough to contain a pedestrian.

[3] We actually discretize the function describing the probability of success for each Bernoulli variable, not the collection of variables itself.

[4] The *same* pixel in two *different* frames of the *same* video appears in the *same* equation.

[5] Since a fast convergence was attained using the simple method of gradient descent, we did not apply more advanced optimization techniques.

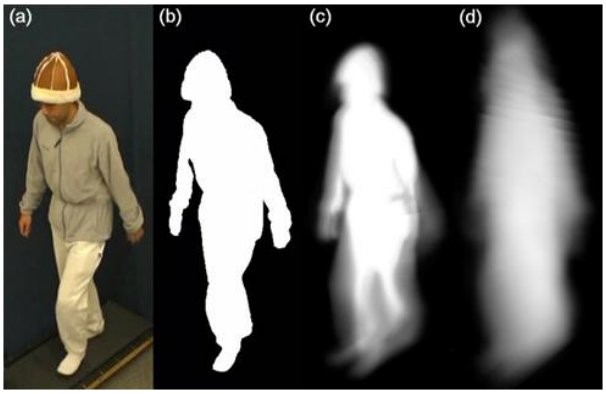


Fig. 2: Learning of the 3D prior: process and results. a) One frame from the original video. b) Segmented person. c) Pixel-wise average of all the segmented frames in the video. d) Projection of the estimated 3D prior to the same view.

To compute the P_j's we followed a number of steps. First, we shot 14 videos (one at a time), of a person walking on a treadmill against a dark background, from different (calibrated) views at diverse heights and angles. Fig. 2a shows a single frame from a single view. Then, we segmented each video frame to obtain masks of the walking person as the light parts of the images (Fig. 2b), and computed the average of all the masks for each video separately, calculating the probability of observing the person in each pixel of the video (Fig. 2c). For each pixel Q_j in the video, this probability is the empirical probability P_j of seeing the object at the pixel.

Since one view does not provide enough information to solve the system of equations mentioned above, we used 14 different views[6] to obtain additional independent equations, which were combined as the rows of the matrix R. This is actually the widely used concept of tomography, and the matrix R is conceptually similar to the fan-beam Radon transform [22]. The 3D prior in this context can then be considered as a semi-transparent object observed from different views.[7] For comparison, Fig. 2d shows the estimated 3D prior projected to the same view.

Fig. 3 compares a 2D prior for the foreground (as in [9]), with a projected 3D prior as described here, for a nearly zenithal angle. Note the increased fidelity of the proposed 3D prior, in particular in the area around the feet.

An alternative approach to learn 3D priors could be to use the very realistic geometric models developed (mostly) by the graphics community (e.g. [23, 24]), together with 3D motion models. We believe that a (simpler) method that can learn directly from the data, as presented here, is valuable, especially since in a multiple camera setup, priors tailored for the particular scene can be learned.

[6] Since we are interested in the pixel occupancy probability as seen from every view, these videos do not need to be shot simultaneously. In fact they were shot sequentially using the same camera, greatly reducing the acquisition complexity and cost.

[7] The code along with the videos and data can be obtained from the authors by request.

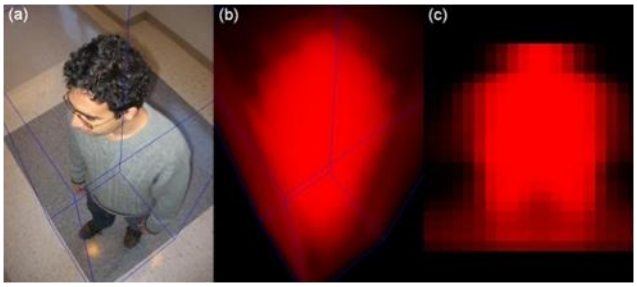


Fig. 3: Comparison of 3D and 2D priors. (a) A view of a walking person. (b) 3D prior for the person as presented here. (c) 2D prior for the person as presented in [9].

Given a person's height (or size) and position in 3D world coordinates, a 3D prior box scaled by the height and placed in the person's position can be projected to estimate the probabilities that the person, its shadow or reflection, will be seen in each video pixel (Fig. 4). For this reason, 3D priors are an essential ingredient of the statistical model discussed in the following section (in particular, Sec. 3.2).

3. Scene learning

In order to learn the desired parameters of the scene (camera matrix and light source 3D location), a model relating them to the observed variables (the input video) is necessary. To this end, we construct a statistical model that explains each pixel in the video as being "produced" by one of three possible classes: background, foreground (people walking in the scene), or shadow (the people's shadow on the floor). Reflections (on shiny floors or mirrors) could also be considered as another class in the same framework, but was not considered in this work.

Each class has an associated color model and class prior. The color models assign probabilities to the colors that each pixel can display if it belongs to a class. The background and shadow color models *at each pixel* consist of a single Gaussian (we explain how to compute them in Sec. 3.1). Not knowing in advance the appearance of people, the foreground color model is a uniform probability in the RGB color space.

Fig. 4: Foreground and reflection priors computed for a person of known height and position in front of a mirror wall. Note that there are four prior projections: the foreground (red); its reflection on the floor (blue); the foreground's reflection in the mirror (red); and its reflection on the floor (blue). *Left*: priors superimposed on the image. *Right*: only the priors.

The class prior defines, for each pixel, the probability of the pixel being background, foreground or shadow. The foreground class 2D prior is computed by placing the learned 3D prior (box) at the pedestrian's (tentative) position, and projecting it (as explained in Sec. 2) using the (tentative) camera matrix. Similarly, the shadow class prior is obtained by projecting the 3D prior using the shadow camera. The background class prior is computed so that the three priors add up to 1.

We now describe the variables used in the model:

M - camera matrix, a 3x4 matrix relating the world and image coordinates. It is characterized in terms of three quantities: the horizon height (*H*), the scaling (S_z) of the *z* axis (vertical), and the scaling (S_y) of the *y* axis (pointing away from camera in the floor plane).
$\vec{L}_P$ - position of the light source in 3D world coordinates.
L_A - attenuation of the background color due to the shadow.
$\boldsymbol{\Theta}$ - set of scene parameters to learn, $\boldsymbol{\Theta} = \{\mathrm{M}, \vec{L}_P\}$. There are other nuisance parameters not included in $\boldsymbol{\Theta}$.
$V_q(i,j,t)$ - color of the pixel (*i*,*j*) in the *t*-th frame for the *q*-th person observed in the scene. This is an input.
$L_q(i,j,t)$ – class label (e.g. background, foreground or shadow), corresponding to the pixel (*i*,*j*) in the *t*-th frame for the *q*-th person.
$\vec{G}_q(t)$ - position (in 3D world coordinates) of the projection of the *q*-th person's center of mass on the floor, in frame *t*.
$\vec{A}_q(t)$ - acceleration (second temporal derivative) of $\vec{G}_q(t)$.
h_q - height (or size in general) of the *q*-th person.

The conditional independencies between these variables are shown in the graphical model of Fig. 5. The only observable variables in the model are the pixel colors (V_q), obtained from the input video. According to the proposed model, they depend only on the pixel classes (L_q), which in turn depend on the position of the person ($\vec{G}_q$), the person's height (h_q), and the scene parameters ($\boldsymbol{\Theta}$). Given these dependencies, the joint probability can be written as

$$p\left(V, L, \vec{G}, h, \boldsymbol{\Theta}\right) = \prod_q \left[\underbrace{p\left(V_q | L_q\right)}_{Color\ Model} \cdot \underbrace{p\left(L_q | \vec{G}_q, h_q, \boldsymbol{\Theta}\right)}_{Label\ Prior} \cdot \underbrace{p\left(\vec{G}_q\right)}_{Traj.Prior} \cdot \underbrace{p\left(h_q\right)}_{Human\ Size\ Prior} \right] \cdot \underbrace{p(\boldsymbol{\Theta})}_{Scene\ Prior} \quad (2)$$

Assuming that all scene configurations are equiprobable, we can safely disregard the scene prior term.

The goal of this work is to estimate the scene parameters $\boldsymbol{\Theta}$, from videos of walking people, using the 3D prior (introduced in Sec. 2). For this purpose, we use the Maximum A Posteriori (MAP) estimator:

$$\Theta = \arg\max_{\Theta} \log p\left(V, L, \vec{G}, h, \boldsymbol{\Theta}\right) \quad (3)$$

This maximization can be carried out efficiently using a nested optimization strategy, where the function is computed in steps (named F_1-F_3) "from the inside out." These steps are carried out simultaneously as suggested by $\max_{x,y} \alpha(x,y) = \max_y(\max_x \alpha(x,y))$. This strategy and each of the terms in Eq. (2) are described in more detail in the following sections.

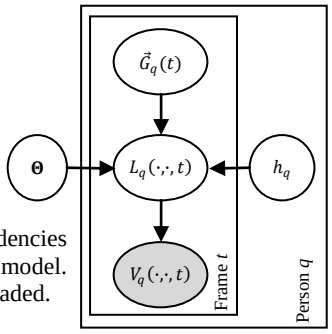


Fig. 5: Conditional independencies between the variables in the model. Observable variables appear shaded.

3.1. Pre-computation

Before starting the actual estimation stage, the proposed pipeline has to compute the background color model, the global attenuation constant, and the rough trajectories corresponding to individuals walking in the scene.

The background model consists of a single Gaussian in the RGB color space for each pixel, with mean and variance estimated following [9]. A global attenuation constant, L_A, which is the factor multiplying the background color model to yield the shadow color model, is computed as the mode (i.e., the most populated histogram bin) of the pixel attenuations[8] for pixels with chroma compatible with the background chroma.

The rough trajectories of individual people walking in the scene are composed of the blobs corresponding to pedestrians, previously detected by background subtraction, which move continuously in the scene while their size changes smoothly.[9] Although better filters based on the appearance of the blobs (e.g. [12, 15]) or advanced people tracking systems could be considered, it is not the goal of this paper, and this simple model is sufficient both to introduce the proposed framework and to present good experimental results.

3.2. Color model and label priors

Given the hypothesized location of a pedestrian in the scene at time *t*, $\vec{G}_q(t)$, and his height (or size in general), h_q, the 3D box containing the 3D prior is scaled by the height and placed at that location in the world. Then, using Eq. (1) and the *postulated* camera matrix, the 3D prior is projected to obtain the foreground class prior probability,

[8] The attenuation of a pixel is defined as the norm of its color (vector) divided by the norm of the background color (vector).

[9] We say that a blob moves *continuously* and *smoothly* if it preserves at least 90% of its pixels from one frame to the next and its size does not change by more than 10%.

$p(L_q = \text{'F'} | \ldots)$ (prior probability that each pixel belongs to the foreground). Similarly, the shadow class prior is obtained using the shadow camera matrix.

Under the simplifying assumptions that the color and label of a pixel ($V_q(i,j,t)$ and $L_q(i,j,t)$ respectively) are independent of the colors and labels of other pixels, the maximum of the first two terms in Eq. (2), for a given frame t, position $\vec{G}_q(t)$, person height h_q, and scene parameters $\boldsymbol{\Theta}$, can be simply computed as

$$\begin{aligned} &F_1\big(t, \vec{G}_q(t), h_q, \boldsymbol{\Theta}\big) \\ &\quad \stackrel{\text{def}}{=} \max_{L_q(t)} \log p\big(V_q(t), L_q(t) | \vec{G}_q(t), h_q, \boldsymbol{\Theta}\big) \\ &\quad = \sum_{(i,j)} \max_{\substack{L_q(i,j,t) \in \\ \{B,F,S\}}} \log p\big(V_q(i,j,t), L_q(i,j,t) | \vec{G}_q(t), h_q, \boldsymbol{\Theta}\big) \end{aligned} \tag{4}$$

This expression is maximized assigning *independently* to each pixel the label which maximizes the (*i*,*j*)-th term. The first step in the maximization therefore only involves finding, for every pixel, the maximum of three numbers.

3.3. Trajectory prior

The *trajectory prior* is a regularization term that penalizes "unlikely" trails. A reasonable assumption is that people tend to walk in relatively straight lines and with roughly constant speed. Therefore, we penalize large acceleration values by modeling the instantaneous accelerations with a zero-mean Gaussian distribution. We also assume that accelerations in different frames are mutually independent, and approximate them at a given time from the current and the two previous positions (second order backward difference). These considerations are summarized in the following expression for the trajectory prior:

$$\log p(\vec{G}_q) = \textstyle\sum_t \log p\left(\vec{A}_q(t)\right) = \sum_{t=3}^{N_F} f\big(\|\vec{A}_q(t)\|\big), \tag{5}$$

where f is the log of the Gaussian distribution of $\vec{A}_q(t)$ and N_F is the total number of frames.

The prior for the entire trajectory can be written as the sum of the priors of its parts. This is helpful in the second step of the maximization of (2), which requires finding the trajectory that maximizes the term in square brackets in

$$F_2\big(h_q, \boldsymbol{\Theta}\big) \stackrel{\text{def}}{=} \max_{\vec{G}_q} \sum_t \begin{bmatrix} F_1\big(t, \vec{G}_q(t), h_q, \boldsymbol{\Theta}\big) \\ +f\big(\|\vec{A}_q(t)\|\big) \end{bmatrix}. \tag{6}$$

For each time *t*, a small number of tentative positions in the ground floor where the *q*-th person may be standing are tested.[10] Let's call the *i*-th of these positions $\vec{G}_q(t,i)$. Then, for an assumed height (h_q) and scene parameters ($\boldsymbol{\Theta}$), $F_2\big(h_q, \boldsymbol{\Theta}\big)$ is the regularized trajectory that best explains the observations.

Consider the three dimensional matrix $Q[t, i_0, i_{-1}]$ that keeps a record of the maximum value that can be achieved in a trajectory that reaches the i_0-th location at time *t*, after visiting the i_{-1}-th location the previous frame. It can be shown that the following recursion computes *Q* efficiently and exactly:

$$\begin{aligned} Q[t, i_0, i_{-1}] = \; & F_1\big(t, \vec{G}_q(t, i_0), h_q, \boldsymbol{\Theta}\big) \\ & + \max_{i_{-2}} \begin{bmatrix} Q[t-1, i_{-1}, i_{-2}] \\ +f\big(\|\vec{A}(\vec{G}_q(t, i_0), \vec{G}_q(t-1, i_{-1}), \vec{G}_q(t-2, i_{-2}))\|\big) \end{bmatrix}. \end{aligned} \tag{7}$$

Then, by definition, $F_2\big(h_q, \boldsymbol{\Theta}\big) = \max_{i_0, i_{-1}} Q[N_F, i_0, i_{-1}]$, can be computed in time $O\big(N_F . {N_G}^3\big)$ where N_G is the number of possible locations per frame. Exact and efficient computation of F_2 is possible since it has *optimal substructure* and therefore dynamic programming can be applied. Note that the state recorded at each instant (in *Q*) has dimension two, in order to include the position as well as the velocity of the person.

3.4. Human size prior

The next nested level of the maximization,

$$F_3(\boldsymbol{\Theta}) \stackrel{\text{def}}{=} \sum_q \max_{h_q} \big[F_2\big(h_q, \boldsymbol{\Theta}\big) + \log p\big(h_q\big)\big] \tag{8}$$

estimates the unknown person's height by searching in a range of values around the expected height value. The term to maximize (in brackets), includes the person's height prior, modeled by a Gaussian centered at 170cm with a standard deviation of 8.5cm (as in [8]). F_3 is computed exactly and efficiently, while still avoiding local maxima in the inner nested optimizations, since only grid search and dynamic programming are used.

3.5. Parameter space search

With the notation introduced in the previous sections, the solution to the problem in (3) can now be stated as,

$$\boldsymbol{\Theta} = \arg\max_{\tilde{\boldsymbol{\Theta}}} F_3\big(\tilde{\boldsymbol{\Theta}}\big) \tag{9}$$

The problem has been reduced to searching the parameter space $\boldsymbol{\Theta}$ (a 4-dimensional space in the example of Sec. 4.1) for the maximum of F_3. The complexity of this search is reduced by a multiscale approach, in the resolution of both the video and the 3D prior.

This whole framework can be implemented in an online fashion keeping track of the value of F_3 at the current solution candidates, and then adding the contributions to Eq. (8) of the newly arrived people.

[10] We estimate these positions using the top location of the head (which in general can be correctly detected) and the tentative height.

4. Experimental results

The framework presented in Sec. 3 was tested on the video in [9] (available at [25]), which includes quantitative results. The goal was to estimate the four parameters that define the scene geometry and illumination: the position of the horizon line, the scale of the *y* axis,[11] and the position of the light (at infinity in the direction given by the spherical coordinates longitude, θ, and latitude, φ). As explained above, the estimate for the parameters is the optimum of F_3.

F_3 was computed in a grid in the parameter space at two video resolutions: half the original resolution and the original resolution. The reason for this is that the parameter values that correspond to the optimum do not change when the video resolution is halved, but the computational cost is one fourth. For the proposed method to work, downsampling can be applied as long as the observed pedestrians are about thirty pixels tall (in this video to half the resolution).

The top part of Fig. 6 shows two orthogonal cuts of F_3 computed in the 4D parameters space surrounding the maximum ($\boldsymbol{\Theta_1}$), computed at half the original video resolution. To locate the maximum more accurately, the neighborhood of $\boldsymbol{\Theta_1}$ was then explored at the original resolution of the video (twice that of the previous one) and the new maximum, $\boldsymbol{\Theta_2}$, was found (bottom of Fig. 6). This is the estimate for the scene parameters.

Fig. 7 shows the axes and shadows that correspond to $\boldsymbol{\Theta_2}$. These parameters define the scene geometry and can be used to perform distance measurements in the scene (see the labeled distances in Fig. 7). Table 1 summarizes the results and compares them to those in [9]. Note that the only input to the system is the assumed average human height.

Table 1: Comparison of estimated measurements.

Measure	Ground truth (m)	From [9] (m)	New model (m)
P1	4.18	3.94	4.27
P2	4.26	4.38	4.25
P3	4.38	4.39	4.36
P4	4.13	3.92	4.29

The new proposed framework leads to significantly better results (1.7% error on average, versus 3.5% in [9]) in the *z* direction. Considering the inherent measurement noise (people's height varies as they walk), the resolution of the video and the size of the people in it, we believe the result obtained is close to the theoretical limit. The measurements in the *x* and *y* directions, being indirectly computed and from less informative clues (i.e. width of the silhouettes, walking speed, etc.), are less exact. It is worth mentioning

[11] The scaling of the *z*-axis is directly computed as the mean height of the persons observed in the scene, and the *x*-axis scaling can then be computed from the scaling of the other two axes.

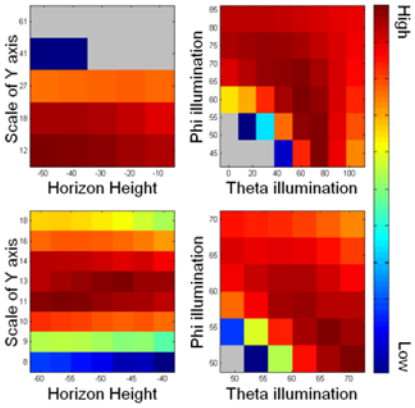


Fig. 6: Two scales of the likelihood in the parameters space, each one including two orthogonal cuts. *Top:* Computed at half the original resolution. *Bottom*: Computed at the original resolution.

that the 3D prior used, via the proposed framework, was learned from a training set containing only empty handed male subjects, while the input video (at [25]) contains mostly female subjects carrying bags, backpacks or purses.

When the camera and illumination parameters are known, they can be exploited through the statistical model presented in Sec. 3 to more accurately localize the pedestrians on the floor plane. The trajectories of the people were computed for the video of Fig. 7, with and without explicitly modeling the shadows. Table 2 contains the mean deviations from the ground truth (hand marked by the user) for each case. Note how using the proposed framework significantly improves the estimation, showing not only the power of the proposed approach but also the importance of explicitly modeling the shadows.

Table 2: Localization error by two different methods.

Video	No Shadows (cm)	Shadows (cm)
Fig. 7	32.3	21.5

5. Conclusions and future work

In this paper, we first introduced a novel representation (3D priors) for non-rigid objects that can be used to reason about the 3D environment, and presented an economic, efficient, and simple method to learn it and use it. We then presented a probabilistic framework to estimate the important camera matrix in situations where other methods have difficulty and to more accurately localize people in the scene. We showed that the combination of 3D priors with the new probabilistic framework outperforms the previous state-of-the-art technique [9] at estimating vertical distances. We also demonstrated that shadows, far from being inconvenient disturbances, could be useful clues to better localize objects in a 3D scene.

An important direction to pursue this work is to remove the independence assumptions built-in into the model and study the effect on the computational cost and results. In particular, it seems worth removing the independence assumption between the colors corresponding to the same part of the body at different times (e.g., the color of the

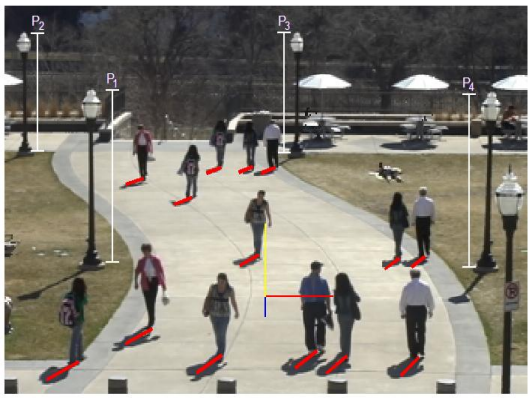


Fig. 7: Results obtained for the video in [9] (to be compared with figures 8 and 11 therein). The estimated horizon line is shown in green, the *x*, *y* and *z* axes are shown in red, blue and yellow, respectively, and the estimated shadow directions are shown in red. Four vertical measurements are marked in white (see the corresponding values in Table 1).

pedestrian's shirt is the same in different frames), the independence assumption between neighboring voxels in the same box (e.g., the voxel below the head's voxel is probably occupied by the neck), and the independence assumption between corresponding voxels at different times (e.g., if an overweight pedestrian occupies many voxels in one frame, he will probably also occupy many voxels in the following frame).

Complementary directions for future work include the learning of multiple light source locations, an important generalization to handle indoor scenes; and the learning and use of multiple 3D priors that include people of both genders and different heights and body shapes, people carrying accessories (e. g., backpacks, purses and bags), and people performing different activities (e.g., running, skating, etc.). (For the current prior only thin males of average height, walking empty handed, were included in the training set.) Results in these directions will be reported elsewhere.

Acknowledgements

This work was carried out while KP was at the University of Minnesota. Support for this work came from NSF, ONR, NGA, ARO, and DARPA.

6. References

[1] Hartley, R. I. and Zisserman, A., *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2003.
[2] 2d3 Ltd., 2d3. Online: http://www.2d3.com.
[3] Faugeras, O. D., *Three-Dimensional Computer Vision: A Geometric Viewpoint.* MIT Press, 1993.
[4] Lv, F., Zhao, T. and Nevatia, R., "Self-calibration of a camera from a video of a walking human." *ICPR*, 2002.
[5] Krahnstoever, N. and Mendoça, P. R. S., "Bayesian autocalibration for surveillance." *ICCV*, 2005.
[6] Junejo, I. N. and Foroosh, H., "Robust auto-calibration from pedestrians." *AVSS*, 2006.
[7] Junejo, I. N. and Foroosh, H. "Trajectory rectification and path modeling for video surveillance." *ICCV*, 2007.
[8] Hoiem, D., Efros, A. A. and Hebert, M., "Putting objects in perspective." *CVPR*, 2006.
[9] Rother, D., Patwardhan, K. A. and Sapiro, G., "What can casual walkers tell us about a 3D scene?" *ICCV*, 2007.
[10] Balan, A. O., Black, M. J., Haussecker, H. and Sigal, L., "Shining a light on human pose: On shadows, shading and the estimation of pose and shape." *ICCV*, 2007.
[11] Cao, X. and Shah, M., "Camera calibration and light source estimation from images with shadows." *CVPR*, 2005.
[12] Sigal, L. and Black, M. J., "Predicting 3D people from 2D pictures." *IV Conference on Articulated Motion and Deformable Objects*, 2006.
[13] Seemann, E., Leibe, B. and Schiele, B., "Multi-aspect detection of articulated objects." *CVPR*, 2006.
[14] Zhao, T., Nevatia, R. and Lv, F., "Segmentation and tracking of multiple humans in complex situations." *CVPR*, 2001.
[15] Leibe, B., Seemann, E. and Schiele, B., "Pedestrian detection in crowded scenes." *CVPR*, 2005.
[16] Wang, J. J. L. and Singh, S., "Video analysis of human dynamics - a survey." *Real Time Imaging*, 2003.
[17] Savarese, S. and Li, F., "3D generic object categorization, localization and pose estimation." *ICCV*, 2007.
[18] Barniv, Y., "Dynamic programming solution for detecting dim moving targets." *IEEE Transactions on Aerospace and Electronic Systems*, 1985.
[19] Han, M., Xu, W., Tao, H. and Gong, Y., "An algorithm for multiple object trajectory tracking." *CVPR*, 2004.
[20] Ardö, H., Berthilsson, R. and Åström, K., "Real time viterbi optimization of hidden Markov models for multi target tracking." *IEEE Workshop on Motion and Video Computing*, 2007.
[21] Kay, S. M., *Fundamentals of Statistical Signal Processing: Estimation Theory.* New Jersey, Prentice Hall, Inc., 1993.
[22] Deans, S. R., *The Radon Transform and some of its Applications.* Wiley, 1983.
[23] Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J. and Davis, J. "SCAPE: Shape completion and animation of people." *SIGGRAPH*, 2005.
[24] de Aguiar, E., Theobalt, C., Stoll, C. and Seidel, H. P., "Rapid animation of laser-scanned humans." *IEEE Virtual Reality Conference*, 2007.
[25] Patwardhan, K. A., What can casual walkers tell us about the 3D scene? Online: http://kedarpatwardhan.org/Research/SceneLearningFromCasualWalkers.htm, 2007.