Atlanta World:
An Expectation Maximization Framework for Simultaneous Low-level Edge Grouping and Camera Calibration in Complex Man-made Environments

Grant Schindler and Frank Dellaert
College of Computing, Georgia Institute of Technology
{grant.schindler, dellaert}@cc.gatech.edu

Abstract

Edges in man-made environments, grouped according to vanishing point directions, provide single-view constraints that have been exploited before as a precursor to both scene understanding and camera calibration. A Bayesian approach to edge grouping was proposed in the "Manhattan World" paper by Coughlan and Yuille, where they assume the existence of three mutually orthogonal vanishing directions in the scene. We extend the thread of work spawned by Coughlan and Yuille in several significant ways. We propose to use the expectation maximization (EM) algorithm to perform the search over all continuous parameters that influence the location of the vanishing points in a scene. Because EM behaves well in high-dimensional spaces, our method can optimize over many more parameters than the exhaustive and stochastic algorithms used previously for this task. Among other things, this lets us optimize over multiple groups of orthogonal vanishing directions, each of which induces one additional degree of freedom. EM is also well suited to recursive estimation of the kind needed for image sequences and/or in mobile robotics. We present experimental results on images of "Atlanta worlds," complex urban scenes with multiple orthogonal edge-groups, that validate our approach. We also show results for continuous relative orientation estimation on a mobile robot.

1. Introduction

We are interested in the problem of 3D scene reconstruction from a single image. This is as part of a larger project that has as its goal the automatic creation of a spatially and temporally registered 3D model of the city of Atlanta, based on historical imagery. The historical nature of the project precludes the use of satellite imagery and/or laser range finders, and in many cases only a single view of a building in a given time period is available. This is why it is important to recover single-view constraints as a precursor to 3D reconstruction.

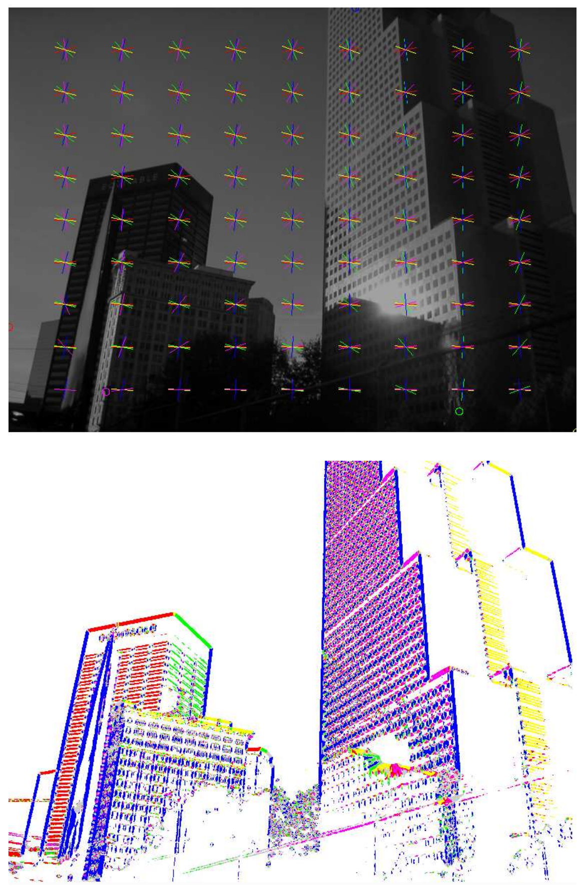

Figure 1: This scene contains two pairs of orthogonal, horizontal vanishing directions in addition to the vertical. In this paper, we introduce an EM-based method to recover vanishing directions (top) while simultaneously grouping edges (bottom). Each colored line segment in the top image is oriented along the vanishing direction of an identically colored edge group in the bottom image.

The extended nature of edges in man-made environments, grouped according to only a few vanishing point directions, provide one kind of single-view constraint. This has been exploited by many authors in the past, both as a precursor to scene understanding as well as extrinsic [2, 1] and intrinsic [6, 10] camera calibration.

A Bayesian approach to edge grouping was proposed in the "Manhattan World" paper by Coughlan and Yuille [2], where they assume three mutually orthogonal vanishing directions in the scene. They perform a one-dimensional exhaustive search over a single camera angle based on a probabilistic classification of each edge, given the gradient image of the scene. In later work [3], they perform a course-to-fine search over 3D camera orientation. This same Bayesian approach has since been applied to the problem of camera calibration using a stochastic search algorithm [6].

In this paper we extend the thread of work spawned by Coughlan and Yuille in several significant new ways. We propose to use the expectation maximization (EM) algorithm to perform the search over the continuous unknowns that influence the classification of the edges along vanishing directions. Because we replace exhaustive or stochastic search with a continuous optimization approach, we are now able to optimize over many more parameters. Both exhaustive search as well as importance sampling suffer drastic performance losses in high-dimensional spaces. The EM algorithm, in contrast, is an iterative optimization method that is guaranteed to converge to a (local) maximum of the log-posterior. EM has been used in the past for estimating and classifying vanishing directions in both calibrated [1] and uncalibrated cases [10]. However, while previous EM-based methods have optimized directly over vanishing directions, we optimize over the unknown parameters that determine the position of these vanishing directions, rather than obtaining those parameters in a separate post-processing step.

A second extension, afforded to us by the use of EM, is the "Atlanta world" assumption, whereby we extend the "Manhattan world" assumption to include multiple groups of orthogonal vanishing directions, each of which induces one additional degree of freedom. This allows our method to correctly group edges for scenes in which not all buildings are aligned along the same Manhattan grid. Finally, the use of EM is well suited to recursive estimation of the kind that needs to be done for image sequences, as the result for one time-step can be used to initialize the EM algorithm in the next time-step, thereby converging much more quickly to the optimally estimated vanishing points.

2. Approach

2.1. Assumptions

Our approach is based on the existence of a set of dominant vanishing directions in the scene, a commonly used scene constraint in man-made environments [2, 4, 9, 6, 10]. We do not require these vanishing directions to be mutually orthogonal. In fact, the *only* assumption we make about the vanishing points (VPs) of the scene is that there is a *finite* set $V = \{v_i | 1 \leq i \leq n\}$ of them, where the number n might or might not be known beforehand.

Given a gradient image G, we wish to estimate a set of unknown parameters Ψ that influence the location of the set of vanishing points V. In general, this can be a probabilistic connection, through some conditional density $P(V|\Psi)$, or a functional relationship $V = f(\Psi)$. Given this, we now wish to obtain the maximum a posteriori (MAP) estimate Ψ^* for the parameters Ψ by maximizing the posterior $P(\Psi|G)$, i.e.,

$$\begin{aligned}\Psi^* &= \underset{\Psi}{\text{argmax}}\ P(\Psi|G) \propto P(G|\Psi)P(\Psi) \\ &= \underset{\Psi}{\text{argmax}}\ P(\Psi)\int_V P(G|V)P(V|\Psi)\end{aligned}$$

In what follows, we always assume a functional relationship $V = f(\Psi)$, in which case the above simplifies to

$$\Psi^* = \underset{\Psi}{\text{argmax}}\ P(\Psi|G) \propto P(G|f(\Psi))P(\Psi) \qquad (1)$$

Note that, in both these cases, if not enough constraints are available from the image, the parameters Ψ might not be determined uniquely from the above formulation. For example, without an orthogonality constraint, we can never hope to recover metric calibration from a single image.

Without loss of generality, this problem formulation holds for a wide variety of situations. Examples include:

- a "Manhattan world" scene assuming three mutually orthogonal vanishing directions and a single unknown camera angle
- a more complex "Atlanta world" scene with multiple pairs of orthogonal vanishing directions where unknowns include a 3D camera rotation and focal length
- a model where we assume no orthogonality constraints whatsoever between the vanishing points
- any of the above, with unknown radial distortion

2.2. An Expectation Maximization Approach

In line with the original "Manhattan World" paper [2], in order to evaluate the MAP criterion (1) we need to integrate over all possible ways of assigning each site p in the gradient image G to one of a finite set of models. In what follows we denote the model at a site p as m_p. Either a site has no edge ($m_p = OFF$), it is assigned to one of the vanishing points in the finite set V ($m_p \in 1..n$), or it is classified as being an edge but not belonging to any of the VPs in V

($m_p = OTHER$). The posterior $P(\Psi|G)$ is then a sum over all possible model configurations M,

$$P(\Psi|G) = P(\Psi) \sum_M P(G|M, f(\Psi))P(M) \qquad (2)$$

where we assume that the prior $P(M)$ does not depend on Ψ. The summation above over all configurations M is intractable for all but the smallest images. If we assume conditional independence and a simple single-site prior $P(M) = \prod_p P(m_p)$ as in [2], the above factors and Equation 2 becomes somewhat more tractable:

$$P(\Psi|G) = P(\Psi) \prod_p \sum_m P(g_p|m, f(\Psi))P(m)$$

Here $P(g_p|m, V)$ is a site-specific density on the gradient measurement g_p at site p, given the model m and the set of vanishing points V. Given this model, in [2] the parameter space was discretized and searched over exhaustively, whereas in [6] a stochastic search was done (iterated importance sampling).

Instead of these search strategies, and inspired by [1, 10], we propose the use of expectation maximization (EM) [5, 11] to locate the MAP estimate Ψ^*. When Ψ is high-dimensional, or worse, the prior on M does not factor (e.g., we could use a Markov Random Field (MRF) prior), both exhaustive search and importance sampling are not efficient enough for practical use. Using EM enables us to perform continuous optimization over Ψ, i.e. to use efficient non-linear optimization methods.

An EM formulation leads to an algorithm where we iterate the following two steps until convergence:

1. In the **E-step**, we keep the current estimate Ψ^{old} for the parameters fixed, and calculate a conditional posterior $P(M|G, \Psi^{old})$ over model configurations M:

   $$P(M|G, \Psi^{old}) \propto P(G|M, V^{old})P(M) \qquad (3)$$

   where $V^{old} = f(\Psi^{old})$ and the vanishing points are constant. A detailed description follows in section 2.4.

2. In the **M-step**, we maximize the expected log-posterior $Q(\Psi; \Psi^{old})$ with respect to Ψ to obtain a new set of parameters Ψ^{new}:

   $$Q(\Psi; \Psi^{old}) \stackrel{\Delta}{=} \langle \log P(G|M, f(\Psi)) \rangle + \log P(\Psi) \quad (4)$$

   $$\Psi^{new} = \underset{\Psi}{\operatorname{argmax}} \; Q(\Psi; \Psi^{old})$$

   where $\langle . \rangle$ denotes expectation, and the expectation is taken with respect to the distribution $P(M|G, \Psi^{old})$ over model configurations M obtained in the E-step. Section 2.5 provides the details for the M-step.

2.3. The Likelihood Model

The likelihood model $P(G|M, V)$ we use is nearly identical to the one used in [2], and is shared between the E-step and the M-step. As was pointed out in [6], conditional independence of the gradient measurements in G does not hold for two adjacent pixels as the gradient operation averages over a window. Thus, we subsample the image at sites p to make the measurements g_p conditionally independent given M and V, and obtain the following likelihood model:

$$P(G|M, V) = \prod_p P(E_p|m_p)P(\phi_p|m_p, V) \qquad (5)$$

where E_p and ϕ_p are respectively the gradient magnitude and perpendicular edge orientation at site p. Magnitude and orientation are modeled independently and hence the likelihood at a given site p is the product of two components:

The gradient magnitude likelihood $P(E|m)$ models the fact that the gradient magnitude is higher on average if an edge is present. Hence, it depends only upon the edge model m, and is defined as

$$P(E|m) = \begin{cases} P_{on}(E) & \text{if } m = 1...n, OTHER \\ P_{off}(E) & \text{if } m = OFF \end{cases}$$

In contrast to [2], we do not use histograms but fitted a Gaussian mixture distribution, $P_{off} = \mathcal{N}(\mu = 1.13, \sigma = 0.77)$ and $P_{on} = \mathcal{N}(\mu = 8.28, \sigma = 6.21)$.

The edge orientation likelihood $P(\phi|m, V)$ models the distribution over orientation given the edge model m and the set of vanishing points V, and is defined as

$$\begin{cases} P_{ang}(\phi - \theta(v_m, u)) & \text{if } m = 1..n \\ 1/2\pi & \text{otherwise} \end{cases}$$

where $u = [u, v, 1]^T$ is the location of the site p in homogeneous coordinates, and v_m is the m^{th} vanishing point in V. The predicted edge orientation $\theta = \theta(v_m, u)$ is computed by taking the cross product

$$v_m \times u = [x, y, 1]^T$$

and computing θ as $\arctan(y/x)$. For the density $P_{ang}()$ we use a zero-mean Gaussian on the difference $\phi - \theta$ between the measured and predicted edge orientations. As standard deviation we use $\sigma = 0.13$ from a fitted Gaussian.

2.4. E-Step

The E-step's only dependence on the current parameter estimate Ψ^{old} is through the corresponding location of the vanishing points V^{old}. *Hence, since both $\Psi^{old}$and V^{old} are kept fixed throughout the E-step, the E-step is completely general and independent of the functional dependence $f(\Psi)$ or the representation of Ψ.*

The goal of the E-step is to obtain the posterior distribution $P(M|G,\Psi^{old})$ over model configurations M defined in Equation 3 on the preceding page,

$$P(M|G,\Psi^{old}) \propto P(M) \prod_p P(E_p|m_p) P(\phi_p|m_p, V^{old})$$

where we use the likelihood model (5). In the results below we assume a simple single-site prior $P(m)$ over the edge model m. Hence, the posterior on M factors as

$$P(M|G,\Psi^{old}) \propto \prod_p P(E_p|m_p) P(\phi_p|m_p, V^{old}) P(m_p)$$

Note that this is without loss of generality: an alternative way to proceed is to use an MRF formulation and approximate the E-step through sampling or belief propagation.

In both cases it is sufficient for the E-step to obtain a set of *weights* w_{pm} for each site p, equal to the marginal posterior probability of the edge model m at site p:

$$w_{pm} \triangleq P(m_p|G,\Psi^{old}) \tag{6}$$

In the simplified non-MRF case the weights w_{pm} can be calculated exactly as (with Z_p a normalization factor):

$$w_{pm} = \frac{1}{Z_p} P(E_p|m_p) P(\phi_p|m_p, V^{old}) P(m_p)$$

2.5. M-Step

In the M-step we re-estimate the parameters Ψ through a non-linear optimization process, but here too the dependence on the chosen parameterization of Ψ is limited: only the function $f(\Psi)$ and its derivative depend upon it.

The objective function to be minimized with respect to Ψ is the expected log-posterior $Q(\Psi;\Psi^{old})$, which simplifies to a non-linear weighted least-squares optimization. Starting from (4) and using the likelihood model (5) we obtain:

$$\begin{aligned} & Q(\Psi;\Psi^{old}) \\ \triangleq\ & \langle \log P(G|M, f(\Psi)) \rangle + \log P(\Psi) \\ =\ & \sum_p \langle \log P(\phi_p|m_p, f(\Psi)) \rangle + \log P(\Psi) \\ =\ & \sum_p \sum_m w_{pm} \log P(\phi_p|m, f(\Psi)) + \log P(\Psi) \end{aligned}$$

where the equality is only up to a constant, and we dropped the terms $P(m_p)$ and $P(E_p|m_p)$ as they do not depend on Ψ. The last equality can be easily verified by plugging in the definition (6) of the weights w_{pm}. A further performance improvement can be obtained by realizing that in the sum above the terms for $m = OFF$ and $m = OTHER$ do not depend on Ψ, as they have no predicted orientation. Thus, the re-estimated parameters Ψ^{new} can be obtained by *minimizing* the weighted least-squares error obtained by summing only over the cases $m = 1...n$, for all sites p:

$$\sum_p \sum_{m\in\{1...n\}} w_{pm} (\phi_p - \theta(f_m(\Psi), u_p))^2 + \log P(\Psi)$$

where $f_m(\Psi)$ is vanishing point v_m, a function of Ψ.

3. Different Worlds

Different world models, i.e., assumptions we make about the scene and the chosen parameterization of the camera, can be fully described in terms of the above problem formulation by specifying three things:

1. the nature of the unknown parameters Ψ
2. the function $f(\Psi)$ that defines the number n of vanishing points and their functional dependence on Ψ
3. the prior $P(m)$ on the edge models

Below we discuss three such models in more detail.

3.1. Manhattan World

As a first application, we consider Coughlan & Yuille's original "Manhattan world" model [2]. In this case,

1. the only unknown parameter is the pan φ of the camera C, hence $\Psi = \{\varphi\}$ is one-dimensional,
2. the vanishing points $V = f(\Psi)$ are assumed to be mutually orthogonal and hence are simply the projections in C of the 3D homogeneous points $[0,0,1,0]^T$, $[1,0,0,0]^T$, and $[0,1,0,0]^T$ lying on the plane at infinity. These vanishing directions are assigned model indices 1, 2, and 3, respectively;
3. the prior over the edge models was taken from [2]:

$$P(m) = \begin{cases} 0.02 & if \quad m = 1,2,3 \\ 0.04 & if \quad m = OTHER \\ 0.90 & if \quad m = OFF \end{cases}$$

3.2. Including Camera Calibration

Next, we consider the case described in [6], where they optimize over a larger set of camera parameters. In this case,

1. the parameters $\Psi = \{f, R\}$, where f is the focal length of the camera C, and the 3D orientation of C is given as a 3×3 rotation matrix $R \in \mathcal{SO}(3)$. Since R has 3 degrees of freedom, the dimensionality of Ψ is 4.
2. the vanishing points $V = f(\Psi)$ are the same as above;
3. the prior over the edge models is the same as above

3.3. Atlanta World

Finally, our EM formulation allows us to introduce much richer sets of constraints. The final model we describe here allows for multiple sets of orthogonal pairs of vanishing directions in the horizontal plane, all orthogonal to the main vertical orientation in the scene (typically defined by gravity). This gives rise to the following modeling assumptions:

1. the parameters $\Psi = \{f, R, \vec{\gamma}\}$ now include one angle $\gamma_i \in \mathcal{SO}(2)$ per additional orthogonal pair of horizontal VPs, relative to the original "Manhattan" triple of vanishing directions;

2. we can easily compute each additional orthogonal pair of vanishing points from their corresponding orthogonal vanishing directions, which are given by the following expressions in the extra parameter γ_i:

   $[\cos(\gamma_i), \sin(\gamma_i), 0, 0]^T,\ [\cos(\gamma_i+\frac{\pi}{2}), \sin(\gamma_i+\frac{\pi}{2}), 0, 0]^T$

   As before, the VPs themselves can be found by projecting these points at infinity in the camera C.

3. we generalize the original model prior by distributing the probability mass for horizontal edges over a possibly larger number n of vanishing directions:

$$P(m) = \begin{cases} 0.02 & if \quad m = 1 \\ \frac{0.04}{n-1} & if \quad m = 2..n \\ 0.04 & if \quad m = OTHER \\ 0.90 & if \quad m = OFF \end{cases} \qquad (7)$$

4. Results

We obtained good results with our method on a number of challenging images, some of which we include here. In all cases, we initialized EM with the result of a quick low-resolution brute-force search over the pan φ of the camera, similar to [2]. However, EM is known to be sensitive to the initial estimate for the parameters, and hence this was not always successful for all images. Hence, strategies for initializing the EM algorithm remain an important component of a fully automatic method.

Images shown below were taken by a hand-held digital camera, except for the robot sequence, which was acquired from a Videre Design firewire camera mounted on an ATRV-mini mobile robot. In addition to sub-sampling each image as described above, we also adaptively threshold on edge magnitude to get a manageable subset of (around 500-1000) image points on which to operate.

The M-step is implemented using Levenberg-Marquardt non-linear optimization in conjunction with a sparse QR solver. To compute the (sparse) Jacobian H, defined as

$$H \triangleq \frac{\partial Q(\Psi; \Psi^{old})}{\partial \Psi}$$

we use an automatic differentiation (AD) framework. AD is neither symbolic nor numerical differentiation, and calculates the Jacobian at any given value exactly, efficiently, and free of numerical instabilities. See [8] for more details.

4.1. Manhattan World Results

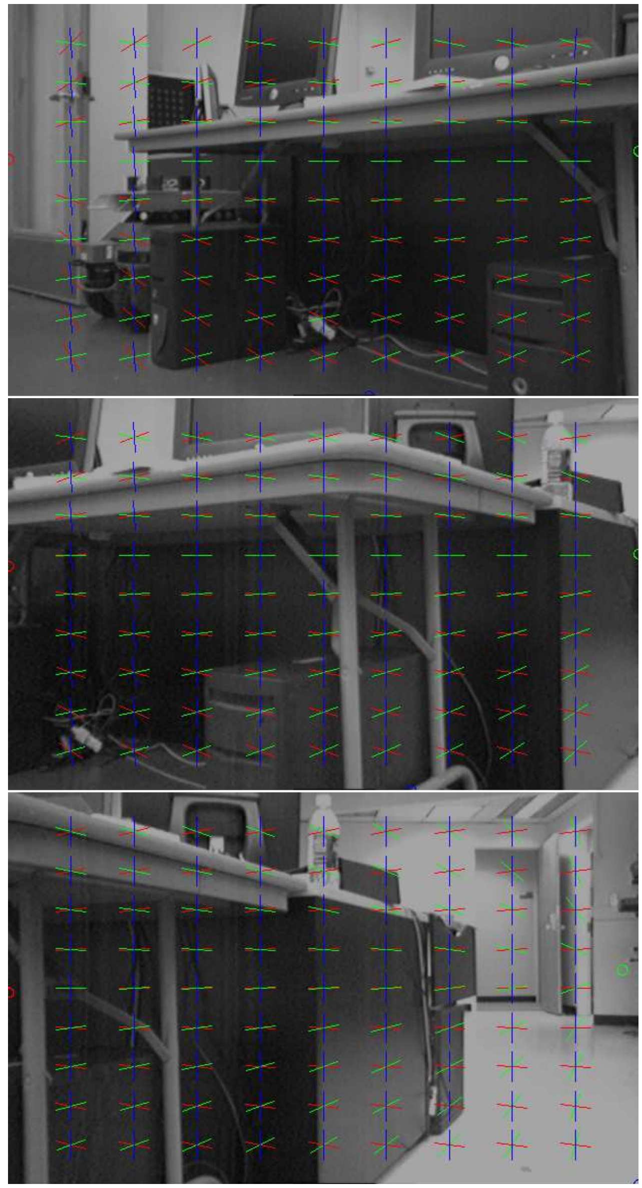

Figure 2: Recursive estimation of relative orientation for a mobile robot in a simple "Manhattan world."

One advantage of using EM is that it is easy to use in a recursive setting, as we can use the MAP estimate Ψ_t^* at time t to initialize the EM search for Ψ_{t+1}^* at time $t+1$. The images in Figure 2 demonstrate this on an image sequence captured by a mobile robot. These results were obtained using the "Manhattan world" modeling assumptions described in Section 3.1. Note that the use of EM enables us to incorporate a motion model for the robot, and/or to perform a batch optimization of the parameters $\{\Psi_t | 1 \leq t \leq T\}$ over a large number of images. In that setting it would be easy to keep some of the parameters constant over the sequence, while others are allowed to vary (within bounds).

4.2. Including Camera Calibration

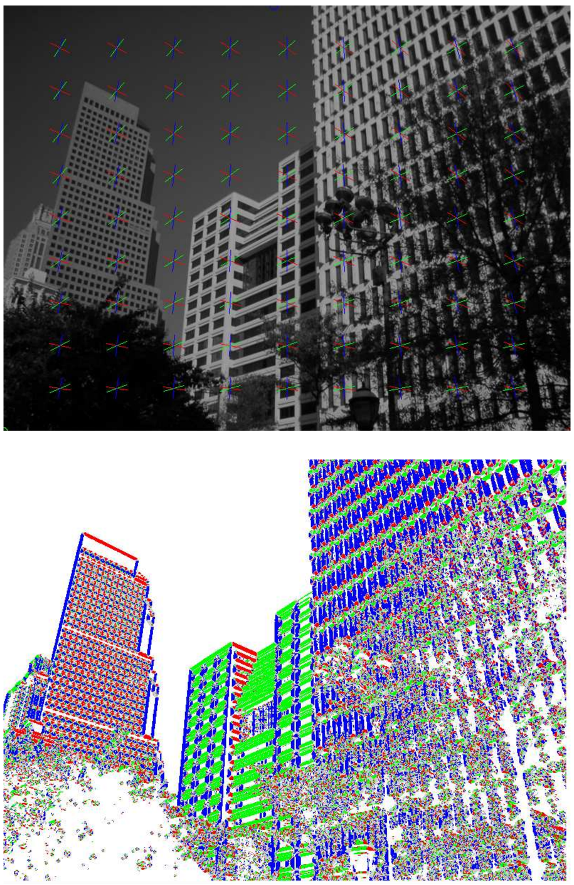

Figure 3: Estimated vanishing directions and edge groupings for a "Manhattan world" scene for which focal length and 3D rotation were unknown.

The main advantage of using EM is that we can include many parameters and simultaneously optimize over all of them, without the need for either discretizing or sampling from the parameter space. To illustrate this, we implemented the model with additional camera parameters (described in Section 3.2). We tested it on a number of outdoor images taken in an urban setting and obtained high-quality vanishing direction estimates and edge groupings within 20-30 iterations of EM. The image in Figure 3 took about 12 seconds for 24 iterations, including initialization.

In contrast to [6], where a quaternion representation is used, we incrementally update the rotation matrix R using Rodrigues' formula [7]. This implements the exponential map from the three-dimensional tangent space, in which our update vector lives, to the space of valid 3D rotation matrices $\mathcal{SO}(3)$.

4.3. Atlanta World Results

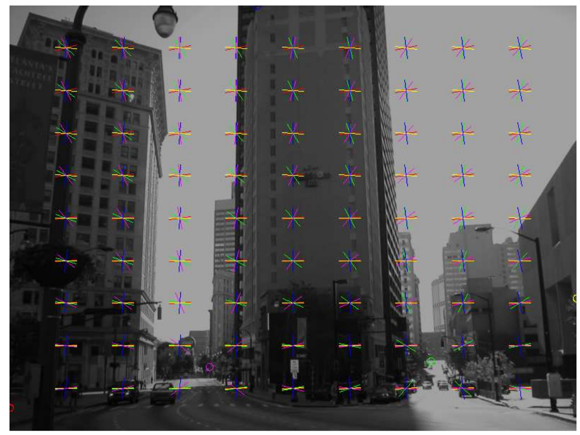

Figure 4: Estimated vanishing directions for an "Atlanta world" scene. Note the correctly identified vanishing points in the image, indicated by small circles where each street vanishes into the distance.

Finally, we take advantage of our method's ability to impose a richer set of constraints on images of complex scenes. In Figure 4 we show a typical city scene where two streets converge, and hence two pairs of orthogonal vanishing directions can be perceived in the scene, which are successfully recovered by the EM algorithm. Figure 1 shows a second example with two dominant VP pairs. As a final example, consider the two images in Figure 5. The left image was recovered using 2 VP pairs, while the right image was recovered assuming 3 VP pairs (in addition to vertical, in both cases). As one can tell from the images, the incorrect assumption of the former case caused some edges to be grouped together although they are clearly not parallel. The image on the right shows the correct grouping and illustrates how our method could potentially be used in a model selection scheme.

5. Conclusions

We have presented a general EM framework for estimating edge groupings, camera parameters, and vanishing directions. In addition, we have demonstrated the success of this framework for a number of different world models. Our method is fast even in continuous parameter spaces of high

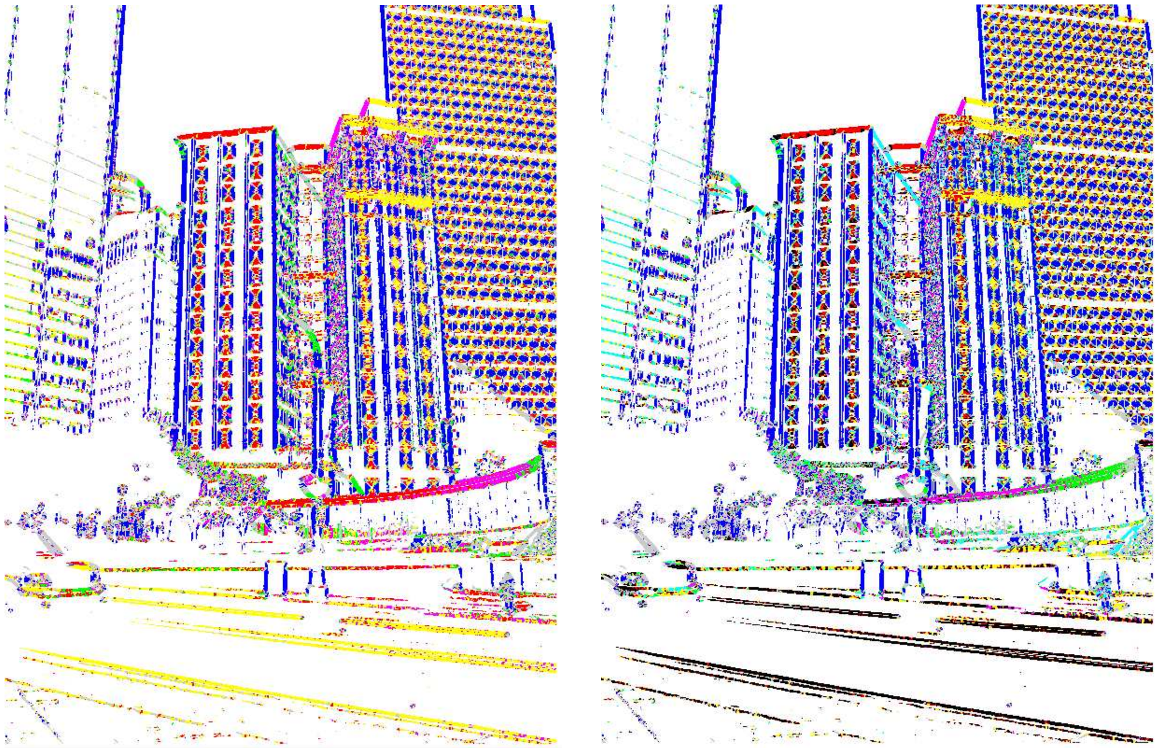

Figure 5: Edge groupings for a complex urban scene under different world models. The world models for the left and right images assume 2 and 3 horizontal vanishing point pairs, respectively, resulting in different edge classifications.

dimensionality. The resulting edge groupings provide an excellent basis for future work on single-view 3D scene reconstruction.

Potential improvements to our method include more robust initialization schemes for camera orientation and horizontal vanishing directions. As illustrated above, it would also be advantageous to include model selection in our framework, enabling us to choose between world models for a scene. Finally, employing an MRF prior in the E-step could reduce the apparent noise in the above edge-grouping images by letting neighboring sites influence each other's edge classification.

References

[1] M. Antone and S. Teller. Automatic recovery of relative camera rotations in urban scenes. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 282–289, June 2000.

[2] J. Coughlan and A. Yuille. Manhattan World: Compass Direction from a Single Image by Bayesian Inference. In *Intl. Conf. on Computer Vision (ICCV)*, pages 941–947, 1999.

[3] J.M. Coughlan and A.L. Yuille. Manhattan world. *Neural Computation*, 2003.

[4] A. Criminisi, I.D. Reid, and A.Zisserman. Single view metrology. *Intl. J. of Computer Vision*, 40(2):123–148, 2000.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[6] J. Deutscher, M. Isard, and J. MacCormick. Automatic camera calibration from a single manhattan image. *Eur. Conf. on Computer Vision (ECCV)*, pages 175–205, 2002.

[7] O.D. Faugeras. *Three-dimensional computer vision: A geometric viewpoint*. The MIT press, Cambridge, MA, 1993.

[8] Andreas Griewank. On Automatic Differentiation. In M. Iri and K. Tanabe, editors, *Mathematical Programming: Recent Developments and Applications*, pages 83–108. Kluwer Academic Publishers, 1989.

[9] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[10] J. Kosecka and W. Zhang. Video Compass. *Eur. Conf. on Computer Vision (ECCV)*, pages 476–491, 2002.

[11] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. John Wiley & Sons, 1997.