Efficient Edge-Based Methods for Estimating Manhattan Frames in Urban Imagery

Patrick Denis[1], James H. Elder[1], and Francisco J. Estrada[2]

[1] York University
{pdenis,jelder}@yorku.ca
[2] University of Toronto
strider@cs.utoronto.ca

Abstract. We address the problem of efficiently estimating the rotation of a camera relative to the canonical 3D Cartesian frame of an urban scene, under the so-called "Manhattan World" assumption [1,2]. While the problem has received considerable attention in recent years, it is unclear how current methods stack up in terms of accuracy and efficiency, and how they might best be improved. It is often argued that it is best to base estimation on all pixels in the image [2]. However, in this paper, we argue that in a sense, less can be more: that basing estimation on sparse, accurately localized edges, rather than dense gradient maps, permits the derivation of more accurate statistical models and leads to more efficient estimation. We also introduce and compare several different search techniques that have advantages over prior approaches. A cornerstone of the paper is the establishment of a new public groundtruth database which we use to derive required statistics and to evaluate and compare algorithms.

1 Introduction

The problem of single-view 3D reconstruction is of enormous theoretical and practical interest. Applications range from urban security to architectural and urban planning systems, to self-localization systems for mobile devices. Given the prevalence of security cameras in urban environments throughout the world, single-view methods could be of great value in the refining of 3D databases such as Google Earth and Microsoft Virtual Earth.

In general, single view reconstruction is an ill-posed problem. The problem becomes more tractable under the so-called "Manhattan-World" assumption [1,2]: that the surfaces of interest are rectangles aligned with a 3D Cartesian frame.

A key step in exploiting the Manhattan World assumption is the estimation of the Manhattan directions, i.e., the 3D rotation of the camera relative to the Manhattan frame [1,2]. This is the problem we address here.

The specific **contributions** we make here fall into four categories. First, we introduce a new public database, with appropriate groundtruthing, that can be used to design and evaluate algorithms. Second, we establish accurate statistics required to ground estimation algorithms. Third, we devise new algorithm

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 197–210, 2008.

variants that have potential advantages in accuracy and speed over previous approaches. Finally, we conduct a comprehensive evaluation of six approaches in terms of both accuracy and efficiency. The results suggest that basing estimates on sparse but accurate edge maps, rather than the dense gradient fields used in other methods [2,3,4], leads to a substantial improvement in both accuracy and computational efficiency.

2 Prior Work

Coughlan & Yuille [1,2] considered the problem of estimating the Manhattan frame using a calibrated camera. They based their estimation on the dense pattern of pixel gradients in the image, using a mixture model in which each pixel gradient has a probability distribution over five possible causes: (1) The pixel is not an edge. (2-4) The pixel is an edge that belongs to one of the three Manhattan-World directions. (5) The pixel is an edge that does not belong to any of the Manhattan-World directions.

Coughlan & Yuille searched over the space of camera rotations for the camera pose that maximized the likelihood of the observed gradient data under their mixture model. Although the gradients were represented in the continuous domain (not Houghed), their coarse-to-fine search technique was discrete.

A number of efforts have been made since to improve on this work, for example, to allow simultaneous calibration of the camera [3,5,4,6], and to extend the Manhattan assumption to more general scene models [3,6]. These methods also replace Coughlan & Yuille's coarse-to-fine discrete search technique, either with an EM method [5,4,6] or stochastic (particle-based) search strategy [3].

There are a number of potential limitations in these methods. Some use heuristics to initialize search, based on dominant orientations [5] or RANSAC methods [6], and it is unclear whether these heuristics will reliably direct the search toward correct solutions. The EM methods [5,4,6] assume that deviations of Manhattan edges and lines from the expected orientations are normally distributed, however this appears to not be the case ([2], and see Figure 3 of this paper). Computing probabilities for dense gradient fields over the entire image [2,4] entails substantial computational cost. This problem is multiplied by complicated iterative search techniques [4]. Subsampling and thresholding the gradient map [4] in order to contain the overall computation time may sacrifice accuracy. While Coughlan & Yuille made an effort to base their method on statistics of scenes, their models were simplified, and subsequent studies have generally been less concerned with accurate statistical models.

In this paper we will address these limitations and evaluate a number of design decisions, using an edge-based method for estimating the Manhattan frame from a calibrated camera. We show how using edges as sparse features allows an accurate statistical model to be derived and employed. We use a new public ground truth database to quantitatively evaluate and compare six algorithms, including the Manhattan World algorithm [2]. Prior quantitative evaluations of

Manhattan frame estimates have been limited, and to our knowledge no quantitative comparisons between algorithms have previously been made.

While there are some advantages to the simultaneous estimation of camera parameters, allowing camera parameters to vary freely also reduces the number of scenes providing sufficient constraints for reliable estimation of the Manhattan frame. In many practical situations camera parameters are known, can be independently estimated, or can be adequately approximated by nominal values. In this paper we assume a calibrated camera, and focus on the issue of how to maximize accuracy and efficiency of Manhattan frame estimation.

3 Approach

There are two main components to our approach: (1) Construction of the ground truth database and camera calibration, and (2) Design of the algorithms to be evaluated.

3.1 Ground Truth Database

We created two databases of 640×480-pixel images of urban Toronto scenes using a standard digital camera (Panasonic Lumix DMC-LC80). The first database, used to calibrate the camera, consisted of 10 images (3 indoor and 7 outdoor) taken with the camera held in a generic attitude with respect to the Manhattan world frame. The second database, used to train and evaluate algorithms, consisted of 102 images (45 indoor and 57 outdoor). The camera was held in a natural attitude. (Posthoc analysis revealed a mean and maximum absolute deviation of the image y-axis from estimated gravitational vertical of 4 deg and 19 deg, respectively.) These 102 images were further randomly divided into training and test sets of 51 images each.

To establish ground truth, we developed an interactive MATLAB program that allows a user to identify lines in the image with sub-pixel precision and to indicate the Manhattan directions with which they are associated (Figure 1).

Vanishing points were first inferred for the calibration database from these ground truth lines using the Gauss Sphere method of Collins & Weiss [7], and nominal values for the internal camera parameters. We then used a standard non-linear search to solve for the focal length and principal point that make the three Manhattan directions for each image mutually orthogonal in 3D space. We assumed a simplified camera model with no skew and unit aspect ratio. The focal length was estimated to be 6.053 ± 0.009mm, and the (x, y) coordinates of the principal point were estimated to be $(-12.9 \pm 4.4, 11.0 \pm 6.4)$ pixels, relative to the image centre.

Vanishing points for the training and test databases were subsequently calculated with the estimated camera parameters [7]. Since each Manhattan direction was estimated independently, the resulting ground truth Manhattan frames were not exactly orthogonal. We therefore fit an orthogonal frame to the three independently-estimated vectors for each image [8].

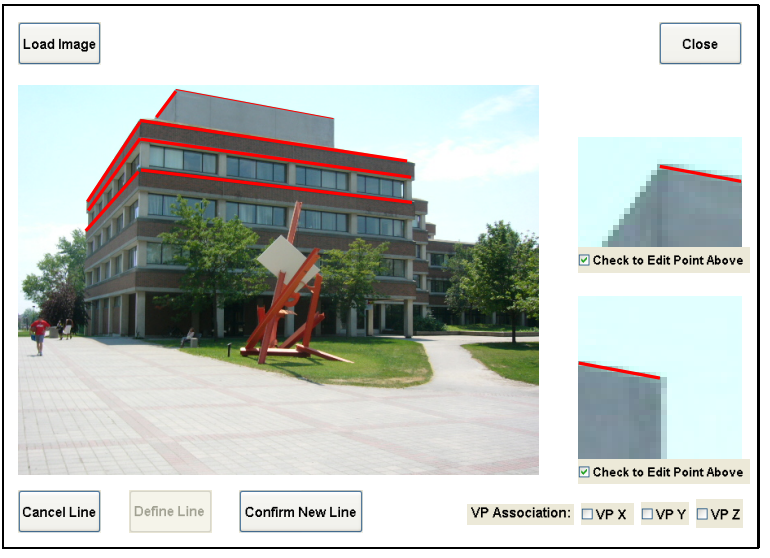

Fig. 1. Ground truth labeling tool

This ground truth database of images, including the estimated Manhattan lines and directions, is available for public use through our website at `http://www.elderlab.yorku.ca/YorkUrbanDB`, in the hope that it will encourage more rapid development as well as quantitative evaluation and comparison of algorithms.

3.2 Mixture Model

We employ a mixture model similar to that used by Coughlan & Yuille ([1], [2]) but applying only to the set of edges E detected and localized to subpixel precision by a standard multiscale edge detector [9]. This reduces the size of the input space by slightly more than a factor of 10.

Each edge $E_{\boldsymbol{u}}$ consists of a pixel location $\boldsymbol{u}$ and orientation $\phi_{\boldsymbol{u}}$, and is assumed to be generated by one of four causes:

$m_{\boldsymbol{u}} = v$	Linear scene structure aligned with the vertical Manhattan direction
$m_{\boldsymbol{u}} = h_1, h_2$	Linear scene structure aligned with one of the two horizontal Manhattan directions
$m_{\boldsymbol{u}} = b$	Background scene structure not aligned with any Manhattan direction

Our goal is to use the edge data to estimate the rotation Ψ' of the coordinate frame of the urban scene (the "Manhattan frame") relative to the camera frame.

As in [2], we assume conditional independence between edges and use a maximum likelihood estimator:

$$\Psi^* = \text{argmax}_\Psi \sum_{\boldsymbol{u}} \log P(E_{\boldsymbol{u}}|\Psi) \tag{1}$$

where

$$P(E_{\boldsymbol{u}}|\Psi) = \sum_{m_{\boldsymbol{u}}} P(E_{\boldsymbol{u}}|m_{\boldsymbol{u}}, \Psi) P(m_{\boldsymbol{u}}). \tag{2}$$

Estimating the Manhattan frame thus requires a reasonable model for the conditional probability $P(E_{\boldsymbol{u}}|m_{\boldsymbol{u}}, \Psi)$ of observing a particular edge $E_{\boldsymbol{u}}$ given that it was generated by a specific cause $m_{\boldsymbol{u}}$, and for the expected proportion $P(m_{\boldsymbol{u}})$ of edges generated by each cause.

3.3 Error Model

In [1], the conditional probability $P(E_{\boldsymbol{u}}|m_{\boldsymbol{u}}, \Psi)$ of observing a particular edge $E_{\boldsymbol{u}}$ given that it was generated by one of the Manhattan causes $m_{\boldsymbol{u}} \in \{v, h_1, h_2\}$ was assumed to be a function only of the angular deviation $\Delta\phi(\boldsymbol{u}, m_{\boldsymbol{u}}, \boldsymbol{\Psi})$ of the observed edge orientation $\phi_{\boldsymbol{u}}$ from the expected direction (Figure 2a). The distribution of $\phi_{\boldsymbol{u}}$ given a background cause was assumed to be uniform. Coughlan & Yuille employed a simple box function to model all three Manhattan conditional distributions. Here we use our ground truth database to determine more accurate models of these distributions.

One of the difficulties in estimating these distributions lies in determining which edges are associated with the hand-labeled lines in the ground truth database. Hand-labelling all of these edges would be prohibitively time-consuming. Instead, consider the density of edges in the image as a function of their distance from the ground truth lines in the database (Figure 2b). The data show a uniform background density of 0.1 edges/pixel, on which is superimposed a peaked distribution of edges generated by the line. We infer from this distribution that the sub-pixel localized edges generated by ground truth lines lie within a distance $x_{\boldsymbol{u}}$ of roughly 1 pixel from the generating line, and thus

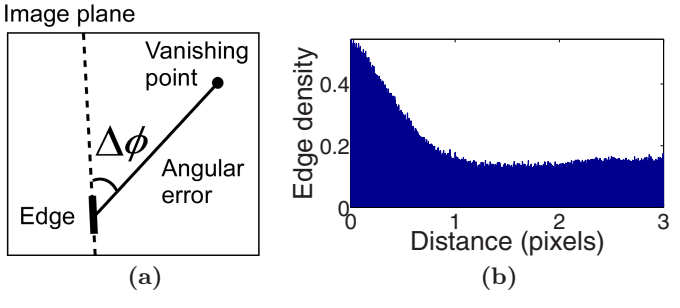

Fig. 2. (a) Angular deviation $\Delta\phi$ of a Manhattan edge from the expected direction, in the image plane. (b) Edge density histogram.

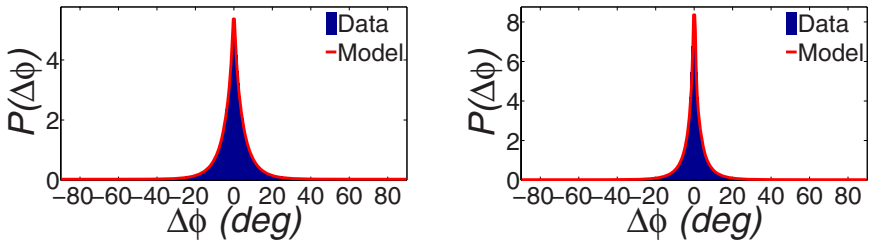

Fig. 3. Empirical densities of image-based angular deviation measure $\Delta\phi$ for edges lying within 1 pixel of horizontal (right) and vertical (left) Manhattan lines. Distributions are fit with a mixture model: see text for details.

confine our analysis to these edges. The conditional distributions of the angular deviation $\Delta\phi_{\boldsymbol{u}}$ for these edges, for both horizontal and vertical Manhattan directions, are shown in Figure 3.

We assume that these distributions arise from a mixture of edges generated by the nearby Manhattan line and edges generated by background structure. Letting $M \in \{v, h_1, h_2\}$ represent the Manhattan direction with which the line is associated, the density $P(\Delta\phi|x_{\boldsymbol{u}} < 1)$ can be represented as

$$P(\Delta\phi|x_{\boldsymbol{u}} < 1) = \lambda P(\Delta\phi|m_{\boldsymbol{u}} = M, \Psi) + (1 - \lambda)P(\Delta\phi|m_{\boldsymbol{u}} = b). \tag{3}$$

We model the former as a Generalized Laplace distribution and the latter by a uniform distribution:

$$P(\Delta\phi|m_{\boldsymbol{u}} = M, \Psi) = \frac{1}{Z} \exp\left(-|\Delta\phi_{\boldsymbol{u}}/b|^{\alpha}\right) \tag{4}$$

and

$$P(\Delta\phi|m_{\boldsymbol{u}} = b) = U(-90, 90) \tag{5}$$

Maximum likelihood parameters for the mixture model are shown in Table 1. Two interesting observations emerge: (1) The Manhattan densities $P(\Delta\phi|m_{\boldsymbol{u}} = M, \Psi)$ are highly leptokurtic ($\alpha < 1$), far from being Normal distributions $\alpha = 2$) as commonly assumed [4] and even further from the boxcar distributions employed by Coughlan & Yuille. (2) The dispersion of the vertical Manhattan density is one third that of the horizontal. This suggests that the orientations of edges generated by vertical structure are more accurate than for horizontal. One possible explanation is that vertical structure may have higher contrast, for example when silhouetted against the sky.

Table 1. Maximum likelihood parameters for mixture model representing deviation of nearby edges from Manhattan lines. See text for details.

	λ	b (deg)	α
Horizontal	0.91	4.0	0.84
Vertical	0.94	1.7	0.65

While Coughlan & Yuille measured error as angular deviation in the image domain, others ([5], [6]) have measured error $\Delta\phi$ as the angle between the vanishing direction and the interpretation plane of the edge in the Gauss Sphere. In the absence of error, this angle should be 0. Using the above method we have also derived estimates of the conditional distributions under this measure of deviation. For lack of space we do not show the distributions here, but we will evaluate and compare the two methods in Section 4.

3.4 Prior Model

To estimate the proportion of edges generated by each causal factor v, h_1, h_2, b, we first compute the density of the angular deviation measure $\Delta\phi$ (in the image) of *all* edges relative to both horizontal and vertical ground truth vanishing points (Figure 4). Note that these densities have peaks at 0 deg due to edges generated by the corresponding Manhattan structure, *and* at 90 deg, due to edges generated by orthogonal Manhattan structure.

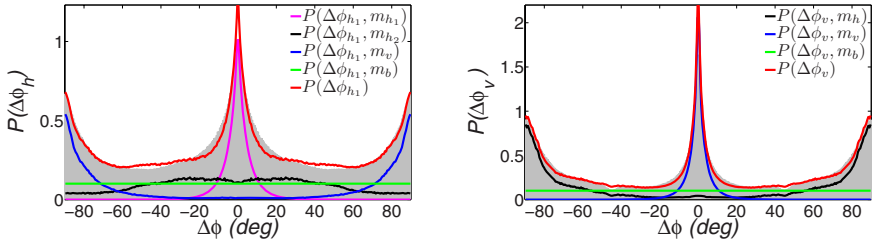

Fig. 4. Density of the angular deviation measure $\Delta\phi$ of all edges relative to both horizontal (right) and vertical (left) ground truth vanishing points. Density is fit with a mixture of densities representing the contributions of edges from all causal factors. See text for details.

To represent this latter contribution, we generalize our notation to let $\Delta\phi_M$, where $M \in \{v, h_1, h_2\}$, denote the deviation measure relative to each of the three Manhattan directions. Now $P(\Delta\phi_M|m_{\boldsymbol{u}})$ represents the conditional distribution of $\Delta\phi$ measured relative to the vanishing point for M, given an edge generated by $m_{\boldsymbol{u}} \in \{v, h_1, h_2, b\}$. We approximate the empirical distributions for $m_{\boldsymbol{u}} \in \{v, h_1, h_2\}, m_{\boldsymbol{u}} \neq M$ by measuring $\Delta\phi$ relative to a Manhattan direction, for edges within $x_{\boldsymbol{u}} < 1$ pixel of a *different* Manhattan direction. Now the total density $P(\Delta\phi_M)$ for all edges can be represented as a mixture from all 4 causes:

$$P(\Delta\phi_M) = \sum_{m_{\boldsymbol{u}} \in \{v, h_1, h_2, b\}} P(\Delta\phi_M|m_{\boldsymbol{u}})P(m_{\boldsymbol{u}}) \tag{6}$$

where the $P(m_{\boldsymbol{u}})$ represent the proportion of edges generated by each causal factor, serving as the mixing parameters in this model. Fitting the mixture to the empirical distribution for all edges thus provides an estimate of these proportions.

Table 2. Comparison between the edge priors estimated here and the gradient priors estimated by Coughlan & Yuille [1]

Model	MW Priors	Our Priors
$P(m_{\boldsymbol{u}} = h_1)$	0.2	0.23
$P(m_{\boldsymbol{u}} = h_2)$	0.2	0.23
$P(m_{\boldsymbol{u}} = v)$	0.2	0.23
$P(m_{\boldsymbol{u}} = b)$	0.4	0.32

Table 2 lists our estimates and compares them to estimates of *gradient* priors made by [1]. The estimates are in reasonable agreement.

3.5 Search Strategies

Our statistical models allow calculation of the likelihood $P(E|\Psi)$ of observing a set of edges E given a camera rotation Ψ relative to the Manhattan frame. Our task now is to estimate the Ψ^* that maximizes this probability.

Coughlan & Yuille [2] employed a coarse-to-fine search strategy over a discretized space representing the three Euler angles composing Ψ. The resolution of the discretization potentially limits the accuracy of the method. Here we explore two alternative search methods in the continuous domain.

Quasi-Newton Method. We first employ a method based upon the BFGS Quasi-Newton gradient descent technique [10]. First, the likelihood $P(E|\Psi)$ is evaluated for some number r of initial guesses $\Psi_i, i \in [1...r]$. Then, for the n most probable guesses, the BFGS Quasi-Newton method is deployed to compute n refined estimates, and the estimate maximizing $P(E|\Psi)$ is selected as the final estimate.

Initial guesses assumed the y-axis of the image plane to be aligned with gravitational vertical, and rotations about the vertical axis were uniformly sampled between 0 and 90 deg. Based on experiments using the training database, values for r and n were selected to achieve maximum speed while remaining within 10% of the achievable accuracy. Suitable values were found to be $r = 5, n = 2$.

EM Method. Schindler & Dellaert [4] proposed a different continuous alternative to Coughlan & Yuille's discrete search strategy, based on Expectation-Maximization (EM). The EM technique had been applied before, for the purpose of estimating individual (not necessarily orthogonal) vanishing directions [5].

As applied to the problem of estimating the Manhattan frame, the EM algorithm alternates between an E-step, in which an estimate is made of the probability $P(m_{\boldsymbol{u}}|E_{\boldsymbol{u}}, \Psi)$ over possible causes $m_{\boldsymbol{u}}$ of each edge $E_{\boldsymbol{u}}$, and an M-step, in which a new estimate of the camera rotation Ψ is made based upon the causal probabilities computed in the E-step. The E-step is straightforward, since, given an estimate Ψ_t of Ψ, $P(m_{\boldsymbol{u}}|E_{\boldsymbol{u}}, \Psi_t) \propto P(E_{\boldsymbol{u}}|m_{\boldsymbol{u}}, \Psi_t)P(m_{\boldsymbol{u}})$ can be calculated directly using the densities and distributions estimated in Section 3. The M-step,

on the other hand, is non-trivial, since no closed-form estimator of Ψ is evident. By assuming the conditional error distributions $P(\Delta\phi|m_{\boldsymbol{u}} = M, \Psi), M \in \{v, h_1, h_2\}$ to be Gaussian, Schindler & Dellaert [4] reduce the M-step to a non-linear least-squares problem, which they solve using iterative nonlinear optimization techniques. However, we know from Section 3 that these distributions are highly *non*-Gaussian.

Here we consider two methods for performing the M-step. In the first, we again use a BFGS Quasi-Newton gradient descent technique [10] to update the estimate of Ψ:

$$\Psi^{t+1} = \arg\max_{\Psi} \sum_{\boldsymbol{u}} \sum_{m} P(m_{\boldsymbol{u}}|E_{\boldsymbol{u}}, \Psi_i) \log(P(E_{\boldsymbol{u}}|m_{\boldsymbol{u}}, \Psi) P(m_{\boldsymbol{u}})) \tag{7}$$

$$= \arg\min_{\Psi} \sum_{\boldsymbol{u}} \sum_{m \in \{h_1, h_2, v\}} P(m_{\boldsymbol{u}}|E_{\boldsymbol{u}}, \Psi_t) \left| \frac{\Delta\phi_{m,\boldsymbol{u}}}{b_m} \right|^{\alpha_m} \tag{8}$$

where $\Delta\phi_{m,\boldsymbol{u}}$ is the angular deviation of the edge at site $\boldsymbol{u}$ from vanishing point m, and b_m and α_m are the Generalized Laplace parameters for that deviation.

As for the Quasi-Newton Method (Section 3.5), the likelihood $P(E|\Psi)$ was first evaluated for some number r of initial guesses. Then, for the n most probable guesses, the above EM method was deployed for t iterations. Based on experiments using the training database, suitable values were found to be $r = 15, n = 3, t = 15$.

As for the Schindler & Dellaert method, the above method nests two iterative procedures, and therefore is unlikely to be efficient. We therefore also considered a quasi-EM approach in which we simplify the M-step by decoupling the estimation of the three Manhattan directions. In particular, we have adapted the fast, closed-form method of Collins & Weiss [7] to estimate each vanishing point direction in the Gauss Sphere independently. To estimate a new vector for a particular Manhattan direction, this adaptation simply requires weighting the interpretation plane normal associated with each edge $E_{\boldsymbol{u}}$ in the image by the probability $P(m_{\boldsymbol{u}}|E_{\boldsymbol{u}}, \Psi)$, computed in the E-step, that the edge was generated by this cause.

Since the vanishing directions are estimated independently, the frame will no longer be orthogonal. At the end of every M-step, we therefore re-orthogonalize the frame by fitting an orthogonal Manhattan frame to the independently-estimated vectors.

As for the other search methods, we empirically estimated the parameters r, n and t that would yield the highest achievable accuracy. We found near-optimal values of $r = 2, n = 1$ and $t = 7$ iterations when measuring error in the image domain, and $r = 1, n = 1$ and $t = 8$ iterations when measuring error in the Gauss Sphere (Section 3.3). However, we also observed erratic convergence behaviour in both cases, suggesting that our approximation of the M-step had destroyed the crucial convergence properties of the EM algorithm.

4 Results

We evaluated six algorithms, summarized in Table 3, on our groundtruth test database. The MW algorithm is the original Manhattan World algorithm of Coughlan & Yuille [2], with code graciously provided by James Coughlan. We list the remaining algorithms in descending order of similarity to the original Manhattan World algorithm. The Edge-Based Newton MW Params algorithm is our edge-based variant of the MW method, using the same statistical distributions as the original algorithm, but using a Quasi-Newton search method instead of the coarse-to-fine search employed by Coughlan & Yuille. The Edge-Based Newton method replaces the Manhattan statistics with our own estimated distributions. The Edge-Based EM algorithm uses the EM method with a nested Quasi-Newton procedure in the M-step (Section 3.5). Finally, the Edge-Based Quasi-EM and Quasi-EM GS algorithms use the quasi-EM method employing the method of Collins & Weiss to update the estimate of the Euler angles in the M-step (Section 3.5).

Empirical accuracy for the six algorithms is shown in Figure 5a. We found that the Edge-Based Newton method performed about 35% better than the original Manhattan World (MW) method, despite using less than 10% of the input data, and even when using the original Manhattan World statistical models (Edge-Based Newton MW Params). We believe this improvement derives from two sources. The first is the higher information content and lower redundancy of the sub-pixel-localized edges [9] underlying the edge-based method, relative to the gradient features used in MW. Given perfect statistical models, noisier gradient features would be downweighted to avoid error. However, in a

Table 3. Properties of the six algorithms evaluated. See text for a description of each algorithm. For each algorithm we list: (1) the image feature used as a cue to estimate the Manhattan frame, (2) the source of the statistical distributions underlying the method, (3) the domain in which the deviation of each feature from the ideal is measured, (4) whether the search space is discretized, and (5) the search method.

Algorithm	**Feature**	**Statistics**	**Domain**	**Search Space**	**Search Method**
MW	Gradients (dense)	Coughlan & Yuille	Image	Discrete	Coarse-to-Fine
Edge-Based Newton MW Params	Edges (sparse)	Coughlan & Yuille	Image	Continuous	Quasi-Newton
Edge-Based Newton	Edges (sparse)	Current study	Image	Continuous	Quasi-Newton
Edge-Based EM	Edges (sparse)	Current study	Image	Continuous	EM
Edge-Based Quasi-EM	Edges (sparse)	Current study	Image	Continuous	Quasi-EM
Edge-Based Quasi-EM GS	Edges (sparse)	Current study	Gauss Sphere	Continuous	Quasi-EM

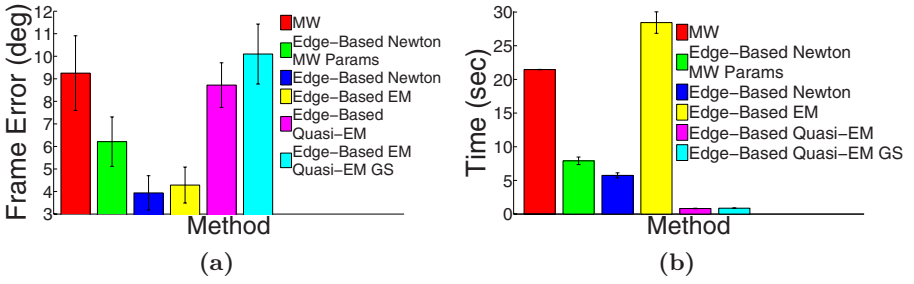

Fig. 5. (a) Average error and (b) average run time for the six algorithms tested. See Table 3 for a definition of each model. Error bars show standard error of the mean.

practical world, models are imperfect, and it is better to exclude noisy features that contain little independent information. The second advantage of the Edge-Based Newton method is the more accurate search algorithm, which is not limited by the discretized search space and discrete search strategy employed in MW.

Using the correct statistics led to a further 35% improvement in accuracy, with the result that the Edge-Based Newton method is more than twice as accurate as the original MW algorithm. Since our estimated priors are fairly close to those of MW, we believe these additional benefits derive mainly from the more accurate likelihood distributions we employ.

We found that replacing the Quasi-Newton search method with the Edge-Based EM method had relatively little influence on accuracy. However, a noticeable and significant loss in accuracy was observed when Manhattan directions were independently estimated in the M-step (Quasi-EM methods), presumably reflecting the loss of convergence properties due to this approximation.

Figure 5b shows the average time taken for each 640 × 480 pixel image in our test database, on a 1.83 GHz Core 2 Duo Intel CPU. All methods have been implemented in MATLAB except for the MW algorithm which was implemented in Python and C++. The Manhattan World method [2] and the edge-based EM method are relatively slow, requiring 20-30 seconds per image. Note that the EM

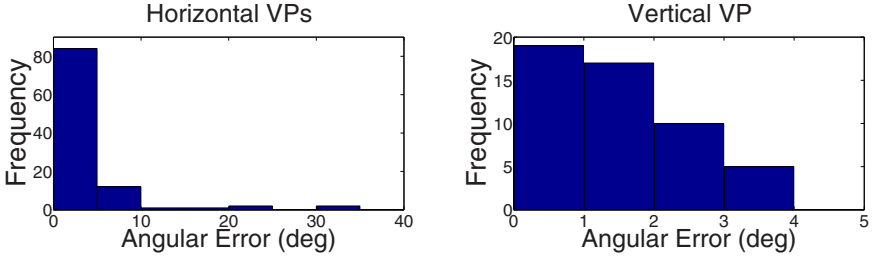

Fig. 6. Distribution of errors for the Edge-Based Newton algorithm over the 26 images in our test database

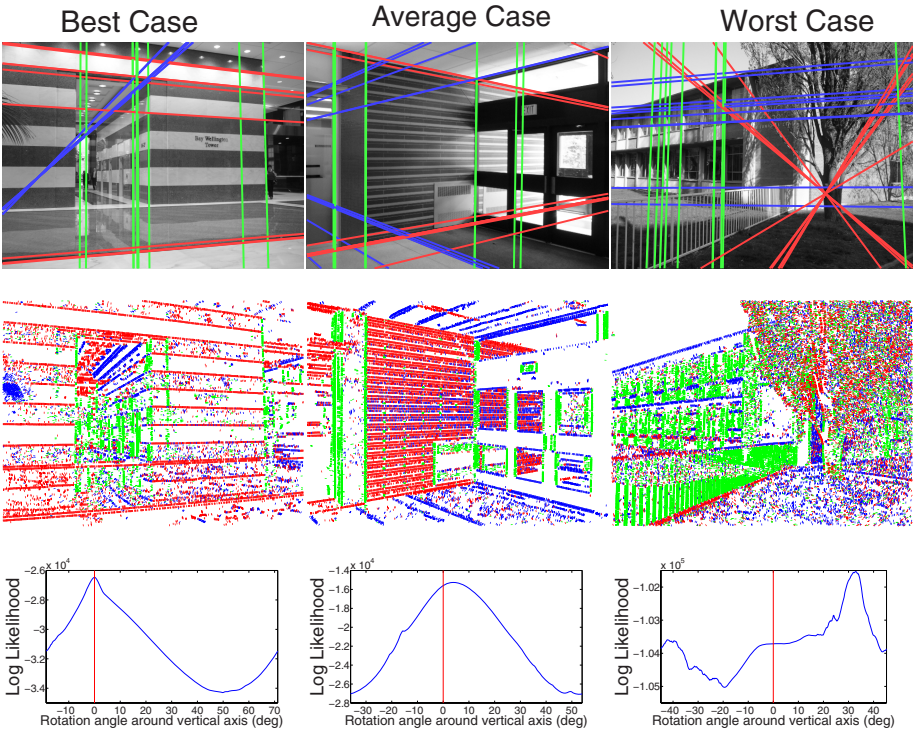

Fig. 7. Example results of our Edge-Based Newton method. Top Row: Lines are drawn through the 10 edges most likely to be generated by each Manhattan direction $m_{\boldsymbol{u}} \in \{v, h_1, h_2\}$, i.e., maximizing $P(\Delta\phi|m_{\boldsymbol{u}}, \Psi^*)$. Middle Row: Edge pixels $\boldsymbol{u}$ for which the most probable model $m^*_{\boldsymbol{u}} \in \{v, h_1, h_2\}$ is a Manhattan direction. Bottom Row: log likelihood distributions $\log P(E|\boldsymbol{\Psi})$ as a function of rotation around the ground truth vertical axis. 0 (red line) indicates ground truth frame.

method run on denser gradient maps, as proposed in the literature (e.g., [4]) would likely run even slower. We find the Edge-Based Newton and Quasi-EM methods run much faster, on the order of 5 seconds or less per image. However, since convergence of the Quasi-EM methods is unreliable, the Edge-Based Newton algorithm is the method of choice, providing the most accurate results of all methods evaluated, and running 4-6 times faster than previous methods from the literature.

Given these findings, we consider further the distribution of errors observed for the Edge-Based Newton algorithm. Figure 6 shows histograms of horizontal and vertical vanishing point errors over the 51 test images. While the great majority of estimates are very accurate

Figure 7 shows results of the top-performing Edge-Based Newton method on three example images (best, average and worst cases) from the test database. The grid-like scene structure in the first (best) image produces a sharp peak in the likelihood very near the ground truth solution. The average case image exhibits a greater variety of orientations, resulting in a broader likelihood with greater deviation from

the ground truth. In our last (worst) image, several factors conspire to produce a horizontal error of 33 deg. Although there is sufficient visible vertical structure, only one of the horizontal directions is adequately represented. A tree not only obscures a large portion of the Manhattan structure in the scene, but also contributes edges forming a central nexus that masquerades as a vanishing point. Finally, an oblique fence contributes another false direction. In sum, the scene does not conform particularly well to the Manhattan assumption. The result is a likelihood function that is peaked very far from the ground truth direction, resulting in a large error.

This failure cannot be avoided by changing the search algorithm, since the likelihood distribution itself is misleading. Allowing inference of more than 3 vanishing points [5,4] could in theory prevent the oblique fence from biasing estimation of the Manhattan frame, however this would not solve the problem of capture by the tree. We find that the maximum log likelihood is negatively correlated with the frame error, and thus could be used as a threshold or a confidence measure for the applicability of the Manhattan model for a particular image [2].

5 Conclusions

We have introduced a new public database that may be used to evaluate and compare methods for estimating Manhattan frames in urban imagery. We have found that basing the estimation of the Manhattan frame on sparse, accurately-localized edges, rather than dense gradient maps, leads to substantially faster and more accurate performance. We believe the gain in accuracy is due mainly to the greater informativeness, accuracy and independence of the sub-pixel-localized edges [9] underlying the edge-based method, relative to the gradient features used in previous methods [2,3,4]. Using edges as input features also reduces the input size by a factor of 10, thereby allowing the deployment of better search methods, leading to further improvements in accuracy.

Several authors have proposed EM-based search methods. We find that a similar EM method based upon edges is reasonably accurate but slow. Methods using denser gradient maps [5,4,6] would likely be even slower. Here we have also explored faster quasi-EM methods in which approximate estimates of the Euler angles are made during the M-step, however convergence of these algorithms is found to be unreliable.

In contrast, we find an edge-based method that uses a Quasi-Newton procedure to directly estimate the Euler angles in the image domain yields the most accurate results, with double the accuracy of the original Manhattan world method, and running 4-6 times faster than methods previously proposed.

References

1. Coughlan, J.M., Yuille, A.L.: Manhattan world: Compass direction from a single image by bayesian inference. In: Seventh International Conference on Computer Vision, vol. 2, pp. 941–947. IEEE, Los Alamitos (1999)
2. Coughlan, J.M., Yuille, A.L.: Manhattan world: Orientation and outlier detection by bayesian inference. Neural Computation 15(5), 1063–1088 (2003)

3. Deutscher, J., Isard, M., MacCormick, J.: Automatic camera calibration from a single manhattan image. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 175–188. Springer, Heidelberg (2002)
4. Schindler, G., Dellaert, F.: Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. I–203 – I–209. IEEE, Los Alamitos (2004)
5. Košecká, J., Zhang, W.: Video compass. In: Seventh European Conference on Computer Vision, pp. 476–490 (2002)
6. Wildenauer, H., Vincze, M.: Vanishing point detection in complex man-made worlds. In: 14th IEEE International Conference on Image Analysis and Processing, pp. 615–622. IEEE, Los Alamitos (2007)
7. Collins, R., Weiss, R.: Vanishing point calculation as a statistical inference on the unit sphere. In: Third International Conference on Computer Vision, pp. 400–403. IEEE, Los Alamitos (1990)
8. Kanatani, K.: Geometric Computation for Machine Vision. Oxford University Press, Inc., New York (1993)
9. Elder, J.H., Zucker, S.W.: Local scale control for edge detection and blur estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(7), 699–716 (1998)
10. Avriel, M.: Nonlinear Programming: Analysis and Methods. Prentice-Hall Inc., Englewood Cliffs (1976)