Geometric Image Parsing in Man-Made Environments

Olga Barinova[1,*], Victor Lempitsky[2], Elena Tretiak[1], and Pushmeet Kohli[3]

[1] Moscow State University
[2] University of Oxford
[3] Microsoft Research Cambridge

Abstract. We present a new parsing framework for the line-based geometric analysis of a single image coming from a man-made environment. This parsing framework models the scene as a composition of geometric primitives spanning different layers from low level (edges) through mid-level (lines and vanishing points) to high level (the zenith and the horizon). The inference in such a model thus jointly and simultaneously estimates a) the grouping of edges into the straight lines, b) the grouping of lines into parallel families, and c) the positioning of the horizon and the zenith in the image. Such a unified treatment means that the uncertainty information propagates between the layers of the model. This is in contrast to most previous approaches to the same problem, which either ignore the middle levels (lines) all together, or use the bottom-up step-by-step pipeline.

For the evaluation, we consider a publicly available York Urban dataset of “Manhattan” scenes, and also introduce a new, harder dataset of 103 urban outdoor images containing many non-Manhattan scenes. The comparative evaluation for the horizon estimation task demonstrate higher accuracy and robustness attained by our method when compared to the current state-of-the-art approaches.

1 Introduction

Recent years have seen a growing interest in the geometric analysis of a scene based on as little as a single image of this scene. Often the image of interest comes from a man-made environment, e.g. when the image is taken indoors or on a city street. In this case, the image is highly likely to contain a certain number of straight lines, which can be identified in the edgemap of the image, and which often can be further grouped into parallel families. The presence of such lines and their parallelism are known to be valuable cues for the geometric analysis.

When a family of parallel lines is projected on the image, their projections are known to intersect in a single point in the image plane called *vanishing point*. The vanishing point uniquely characterizes the 3D direction of those lines (given

[*] The first three authors are supported by Microsoft Research programs in Russia. Victor Lempitsky is also supported by EU under ERC grant VisRec no. 228180.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part II, LNCS 6312, pp. 57–70, 2010.

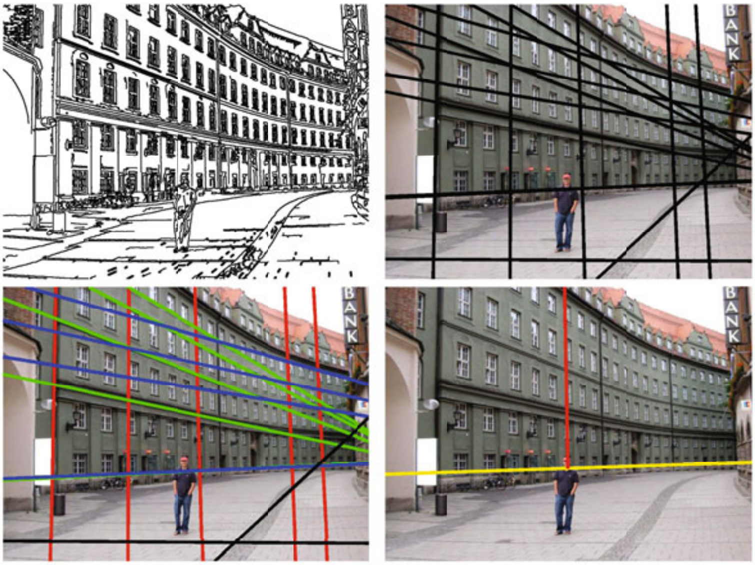


Fig. 1. Geometric primitives of different levels for an example “non-Manhattan” image. *TopLeft* – edge pixels, *TopRight* – straight lines, *BottomLeft* – lines are grouped in parallel families (color indication used), *BottomRight* – the horizon and the zenith (shown with the direction vector in red). Our framework aims at joint estimation of primitives at the latter three levels given the former one (edge pixels).

the camera). When 3D directions of several families are coplanar, the respective vanishing points belong to the same line. Such situation occurs frequently for man-made environments, as there often exist several families with different horizontal directions. In this case, the line containing their vanishing points is called the *horizon*. Most of the remaining lines of the scene are typically vertical. As such, they are parallel to each other and their projections intersect in the vanishing point called the *zenith*[1].

The environments where horizontal lines fall into two orthogonal families, are known as “Manhattan” worlds. A considerable number of previous works investigated the Manhattan case, and the particular simplifications that it brings to the geometric analysis. The parsing framework suggested in this work may be adapted to the Manhattan case, however our work focuses on the non-Manhattan case, assuming the presence of the horizon and the zenith but not the two orthogonal horizontal directions. Surprisingly, very few previous works have paid attention to such scenario (most notably [1]), although we would argue that such assumptions about the scene strike a good balance between the generality and the robustness of the estimation.

[1] Strictly speaking, when this vanishing point lies below the horizon, it should be called the *nadir*. For brevity, we use the term *zenith* in this case as well.

In general, several computer vision and image processing tasks can benefit from the ability to extract the geometric information from a single image. E.g. the knowledge about the location of the horizon may be used to rectify the user photograph with inclined horizon, to facilitate the dense single-view reconstruction and "auto pop-up" [2,3]; this knowledge may also greatly improve semantic segmentation, scene understanding, and object detection [4] as well as video stabilization [5]. The abundance of applications thus motivates the research into better method of geometric analysis of single images leading to more accurate and robust algorithms.

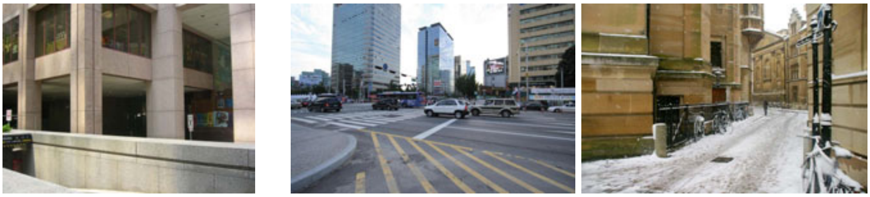
1.1 Related Work

Conceptually, the process of line-based geometric analysis of a single image is well investigated, and typically involves several bottom-up steps. Thus, the process might be initialized with the edge map of an image computed with some edge detector (a standard Canny detector is used in this work). Then, the bottom-up pipeline [6,7,8,9,10,11,12] involves grouping edges into lines, grouping lines into line families and finding the respective vanishing points, and, finally, fitting the horizon and the zenith or the Manhattan directions, depending on the assumptions about the world.

The problem with the step-by-step approach is, however, that neither of the steps can be performed with 100% accuracy and reliability. As the edge maps are always noisy and contaminated with spurious edge pixels not coming from straight lines, the line detection step would miss some of the straight lines and, even worse, detect some spurious lines that do not exist in the scene. Due to these errors, the parallel line grouping step would often group together lines from different families or create groups containing spurious lines (leading to spurious vanishing points) or split actual line families into several (reducing the accuracy of the respective vanishing point estimation). Finally, given an imperfect set of vanishing point, contaminated with outliers, horizon and zenith estimation may lead to gross errors.

Previous works address the challenges associated with each step through several classes of techniques, including robust statistical inference[13], clustering [6,9,14,11], various kinds of Hough transforms [7,9,10], stochastic model fitting [15,12] as well as seeking user supervision [8]. While different approaches possess different strengths and weaknesses, neither results in a perfect accuracy and robustness, leading to the accumulation of errors towards higher stages of the pipeline.

A group of methods [16,17,1,18] goes beyond this pipeline paradigm, as they bypass the line extraction step altogether and directly fit the low-parametric high-level model of the frame (the Manhattan frame [16,17] or a set of Manhattan frames [1]) to the low-level edge map or even to the dense set of image gradients. The joint optimization nature of these methods is similar to our philosophy. However, the simplicity of the model and lack of the edges-to-lines grouping stage limits the accuracy and robustness of their approach as compared to a well-engineered full pipeline approach such as [12].



York Urban dataset [18]

The new “Eurasian cities” dataset

Fig. 2. While the York Urban dataset [18] contains images of “Manhattan” worlds, our framework uses less restrictive scene assumptions that are met by non-Manhattan images in the new dataset that we introduce. Our framework is evaluated on both datasets.

1.2 Overview of Our Method

In this work, we investigate the *geometric parsing* approach to the line-based geometric analysis. By geometric parsing here, we understand the process, when the geometric elements at different levels of complexity (Figure 1), as well as the intra-level grouping relations are explicitly recovered through the joint optimization process. Note, that the term *parsing* is used in a similar meaning in such works as [19], where semantic primitives of different levels (e.g. body parts, individual humans, crowd) as well as the intra-level grouping relations are recovered. In our cases, the primitives at different levels are edge pixels, lines, horizontal vanishing points, the zenith and the horizon.

Our work thus differs from works that employ a single bottom-up pass, as the inference in our case is performed jointly, allowing the information from top levels resolve the ambiguities on the lower levels (and vice versa). Our work also differs from the works that bypass the line detection, as the lines in our method are detected explicitly. To the best of our knowledge, the method presented here is the first that integrates line detection, vanishing point location, and higher-level geometric estimation (the horizon and the zenith in our case) in a single optimization framework. Notably, the optimization in our method does not employ alternations between different levels, and is therefore less prone to getting stuck at poor local minima.

There are several design choices and assumptions in our model that are motivated by the applicability and tractability. Firstly, unlike the majority of previous works, we do not make a Manhattan-world assumption. Instead, we consider a less-restrictive non-Manhattan scenario similar to the “Atlanta world” of [1] that will be detailed below in Section 2. Regarding the camera parameters, we assume that the principal point is known (if unknown we assumed it to be in the center of the frame); we also assume that pixels are square. This assumption holds approximately for the vast majority of cameras in real life, and it makes the inference in our model much easier. We also do not model radial distortion explicitly, which is perhaps a bigger shortcoming of our model, although the robust nature of our algorithm means that considerable distortion might be tolerated without explicit modeling. Finally, we assume the focal length unknown.

Theoretically, locations of the horizon and the zenith allow to estimate the focal length of the camera directly from the results of the parsing, however the accuracy of such estimation is hindered by the degeneracy that occurs when the horizon passes near the principal point, which in practice happens very often.

In a sequel, we detail our energy model in Section 2, and discuss the optimization procedure in Section 3. We then perform the experimental validation on two datasets (Figure 2). The first one is the York Urban dataset presented in [18], where several approaches were benchmarked. This dataset has been recently also used for the evaluation in [12], where improved results have been reported. The second dataset was collected by ourselves and, unlike Urban, contains a lot of more challenging non-Manhattan outdoor scenes. The experimental comparison in Section 4 demonstrates the competitiveness of the parsing approach.

2 The Model for Geometric Parsing

We now explain the energy model of the world within our method. We assume an image to be defined by the set of its edge pixels. The main assumptions about the world are a) that a considerable part of edge pixels may be explained by a set of lines, b) that considerable part of those lines fall into several parallel line families. It is further assumed that c) one of these families is a set of vertical (in 3D) lines converging (in the image plane) to the *zenith* and d) all other families consist of horizontal (in 3D) lines converging (in the image plane) to a set of *horizontal* vanishing points, that all lie close to a single line in the image plane known as the *horizon*. The model thus encompasses the edge pixels, the lines, the zenith, and the horizontal vanishing points, as well as the grouping relations of edge pixels in the lines as well as the lines into the parallel families.

We now introduce the notation and the energy model. The edge pixels are denoted $\mathbf{p} = \{p_i\}_{i=1..P}$. The lines present in the scene are denoted $\mathbf{l} = \{l_i\}_{i=1..L}$. As the model involves grouping of lines into parallel families, we denote with z the vanishing point of the vertical line family (the zenith) and with $\mathbf{h} = \{h_i\}_{i=1..H}$ the set of vanishing points of the horizontal families. The points $h_1, h_2 \ldots h_H$ thus have to lie close to a line in the image plane (we will refer to this fact as the *horizon constraint*).

The energy function in our method includes the individual energy terms corresponding to the (pseudo-)likelihood of each edge and each line. The edge pixel energy term is defined as:

$$E_{edge}(p|\mathbf{l}) = \min\left(\theta_{bg}, \min_{i=1..L} \theta_{dist} \cdot d(p, l_i) + \theta_{grad} * d_{angle}(p; l_i)\right), \quad (1)$$

where $d(p, l_i)$ denotes the Euclidean distance in the image plane between the pixel p and the line l_i, $d_{angle}(p; l_i)$ denotes the angular difference between the local edge direction at pixel p and the direction of the line l_i, θ_{bg} is the constant, corresponding to the likelihood of the background clutter, and θ_{dist} and θ_{grad} are the constants corresponding to the spread of edge pixels generated by a particular line around that line. Thus, the energy term for an edge pixel p is small if this edge pixel is well explained by some line from the set $\mathbf{l}$ and is large otherwise. The largest possible value is θ_{bg}, which corresponds to an edge pixel generated by the background clutter.

The line energy terms are defined as

$$E_{line}(l|\mathbf{h}, z) = \min\left(\eta_{bg}, \min_{i=1..H} \eta_{dist} \cdot \phi(l, h_i)^2, \eta_{dist} \cdot \phi(l, z)^2\right) , \tag{2}$$

where ϕ denotes the distance on the Gaussian sphere [20] between the projection of the line l and projection of the respective vanishing point (h_i or z). η_{bg} is the constant, corresponding to the likelihood of lines that are neither horizontal nor vertical, and η_{dist} is the constant, corresponding to the spread of lines in their families around the respective vanishing points. Thus, the energy term for a line l is small if this line is well explained by (i.e. passes close to) a vanishing point from the set $\mathbf{h} \cup \{z\}$ and is large otherwise. The largest possible value is η_{bg}, which corresponds to a line that is neither vertical nor horizontal.

According to horizontal constraint introduced above all vanishing points except the zenith have to lie close to a line in the image plane. How can we enforce this constraint? Should a separate variable for the position of the horizon be introduced? It turns out [21] that under our assumption about internal camera parameters (square pixels and known principal point) this is not necessary. Under these assumptions, the horizon is perpendicular to the radius vector between the line $L(z)$ connecting the zenith and the principal point, and we enforce this perpendicularity with the following energy term:

$$E_{horizon}(u, h|z) = \kappa_{hor} \cdot \tan \psi(u - h, L(z)) \tag{3}$$

where ψ is the absolute angle between the vector $u - h$ and a perpendicular to $L(z)$, and κ_{hor} is a constant. The tan in (3) was chosen because it imposes significant penalty (upto $+\infty$) on strong non-orthogonality between the horizon and $L(z)$.

The final energy is thus defined as:

$$E_{total}(\mathbf{l}, \mathbf{h}, z|\mathbf{p}) = \sum_{i=1..P} E_{edge}(p_i|\mathbf{l}) + \sum_{i=1..L} E_{line}(l_i|\mathbf{h}, z) + \sum_{1 \le i < j \le H} E_{horizon}(h_i, h_j|z) + E_{prior}(\mathbf{l}, \mathbf{h}) , \tag{4}$$

where $E_{prior}(\mathbf{l}, \mathbf{h}) = \lambda_{line}|\mathbf{l}| + \lambda_{vp}|\mathbf{h}|$, is an MDL prior penalizing the number of lines $|\mathbf{l}| = L$ and the number of horizontal vanishing points $|\mathbf{h}| = H$, thus favouring simpler explanations of the scene (λ_{line} and λ_{vp} are the constants regulating the strength of this prior). The energy (4) thus ties together the different-level components in the image of a non-Manhattan environment, and the line-based parsing of such an image may be performed through the minimization of (4).

Probabilistic interpretation and the model of [22]. Some of the components of our model may be easily formulated with the language of probabilities. In particular, the part of our model related to the edges and their grouping into

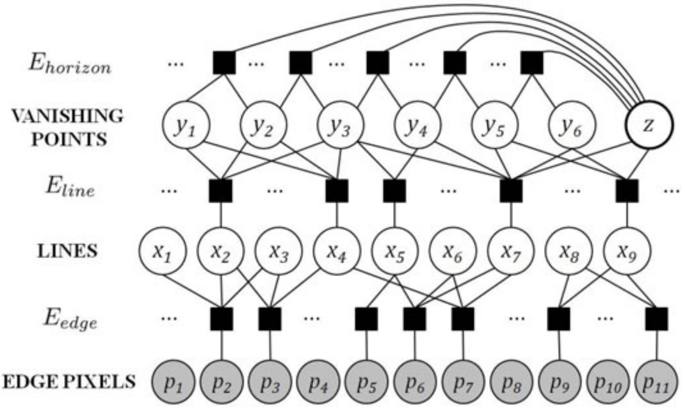


Fig. 3. The graphical model for the discrete approximation of the energy (4). The variables $x_1 \dots x_X$ and $y_1 \dots y_Y$ are binary and correspond to the existence or the absence of candidate lines and horizontal vanishing points. z stands for the location of the zenith and takes the value in a precomputed discrete set of 2D points in the image plane. The unary cliques corresponding to $x_1 \dots x_X$ and $y_1 \dots y_Y$ are omitted for clarity. The shaded nodes (edge pixels) are observed both during training and at test time. *Please, see text for more details.*

lines is in the exact correspondence with the probabilistic model of Hough transform derived in [22]. The next layer of the model concerned with lines and their grouping into families permits an analogous probabilistic treatment. It is unclear, however, if the $E_{horizon}$ term in (3) admits a probabilistic interpretation, as it apparently involves some overcounting of the perpendicularity cues. On practice, this non-probabilistic nature does not present a problem, as we train our model discriminatively by tuning the constants $\theta_{bg}, \theta_{dist}, \theta_{grad}, \eta_{bg}, \eta_{dist}, \kappa_{hor}, \lambda_{line}, \lambda_{vp}$ on the hold-out validation set.

3 Inference

The minimization of (4) is a hard computation problem that necessitates the use of approximations. One possible way would be to minimize it greedily in a layer-by-layer fashion, first choosing the set of lines given the edges, then choosing the set of vanishing points given lines, then fitting the horizon and the zenith into the chosen lines. Such approach would correspond to the traditional bottom-up pipeline from previous methods. Its results might be improved with reiteration of the process through the EM-algorithm, although on practice that suffers from the local minima problem and often gets stuck close to the initial greedy approximation.

A different approach taken in this work is to derive a *discrete approximation* to the original energy that is easier to minimize. To achieve that, we do two steps of the bottom-up pipeline, namely line detection and vanishing point detection, with very low acceptance thresholds, ensuring that an extensive set of X lines $\hat{l}_1..\hat{l}_X$ and an extensive set of Y vanishing points $\hat{h}_1..\hat{h}_Y$ are detected. On practice, one may use any approach that detects lines based on the edgemap and any approach that detect a set of vanishing points based on lines. We detail our choices in the experimental section (see also Figure 4).

The task of the approximate minimization of (4) may then be reduced to the minimization of the energy of discrete variables $\mathbf{x} = \{x_i\}_{i=1..X}$, $\mathbf{y} = \{y_i\}_{i=1..Y}$, and z. Here, each variable x_i is binary and decides whether a candidate line $\hat{l}_i$ is present ($x_i = 1$) or absent ($x_i = 0$) in the image. Similarly, each variable y_i is binary and decides whether a candidate vanishing point $\hat{h}_i$ is a *horizontal* vanishing point that is present ($y_i = 1$) or absent ($y_i = 0$) in the image. Finally, the variable z is, as defined above, a 2D point in the image plane corresponding to the zenith. The set of its possible locations is however restricted to discrete set of candidate vanishing points. For computational efficiency, we may further prune the set of possible locations for z by removing candidate vanishing points that correspond to the horizon inclinations of more than 7.5 degrees. This can be regarded as an additional hard prior on z in our original energy.

The discrete approximation to the energy (4) is then defined by the requirement:

$$E_{discrete}(\mathbf{x}, \mathbf{y}, z|\mathbf{p}) \equiv E_{total}(\{\hat{l}_j\}_{j:x_j=1}, \{\hat{h}_k\}_{k:y_k=1}, z|\mathbf{p}) \,. \quad (5)$$

In other words, the discrete energy is defined as the continuos energy of the appropriate subsets of candidate lines and vanishing points.

In more detail, the discrete energy defined in (5) can be written as:

$$E_{discrete}(\mathbf{x}, \mathbf{y}, z|\mathbf{p}) = \sum_{i=1..P} E_{edge}(p_i|\{\hat{l}_j\}_{j:x_j=1}) + \sum_{i=1..X} x_i \cdot E_{line}(\hat{l}_i|\{\hat{h}_k\}_{k:y_k=1}, z) +$$
$$\sum_{1 \le i < j \le Y} y_i \cdot y_j \cdot E_{horizon}(\hat{h}_i, \hat{h}_j|z) + \sum_{i=1..X} \lambda_{line} \cdot x_i + \sum_{i=1..Y} \lambda_{vp} \cdot y_i . \quad (6)$$

The factor graph for the formula (6) is shown in Figure 3. Note, that due to the truncation effect of the constants θ_{bg} and θ_{dist} in the definition of E_{edge} and E_{line}, the connections between the E_{edge} factors and the line variables as well as between the E_{line} factors and the vanishing points variables are sparse. Each E_{edge} factor is connected only to the lines that pass nearby that edge pixel and, likewise, each E_{line} factor is connected to the vanishing point variables that lie near that line.

Since the values of $\mathbf{p}$ are observed, very big efficiency gains may be easily obtained by merging (summing up) the E_{edge} factors that are connected to the same (or nested) sets of line variables. Since E_{edge} terms constitute the vast majority of terms in (6), this trick dramatically reduces the number of energy terms in the model. It permits us to use quite a simple and brute-force optimization scheme, while still allowing short optimization runtime of several seconds for a typical photograph. In more detail, we exhaustively search through the zenith candidate set (which typically includes less then a dozen of candidates). Given a fixed z, we then perform optimization over the binary variables x and

y through the Iterated Conditional Modes algorithm [23] with the randomized node visiting order.

4 Experiments

Technical details. In our experiments we used the following strategy for choosing candidate lines and candidate vanishing points. For the line detection the probabilistic version of Hough transform [22] was used, where we set the parameters of the method to θ_{bg} and θ_{dist} accordingly. As [22] provides the confidence measure for each detected line, we fixed the number of candidates to 500 and for each image took 500 lines with the highest confidence. Figure 4 gives a typical example of what the candidate set typically looks like.

The candidates for vanishing points were chosen using the J-linkage procedure, described in [12]. This method is based on random sampling, so we ran it several times starting from different random initializations. Usually we got from 50 to 100 candidates for vanishing points. After performing the inference in our model we usually got from 2 to 5 vanishing points and groups of lines supporting each of them.

In the experiments on York Urban dataset we exploited the coordinates of principal point provided, in the experiments on the new dataset we assumed the principal point to lie in the center of the image frame.

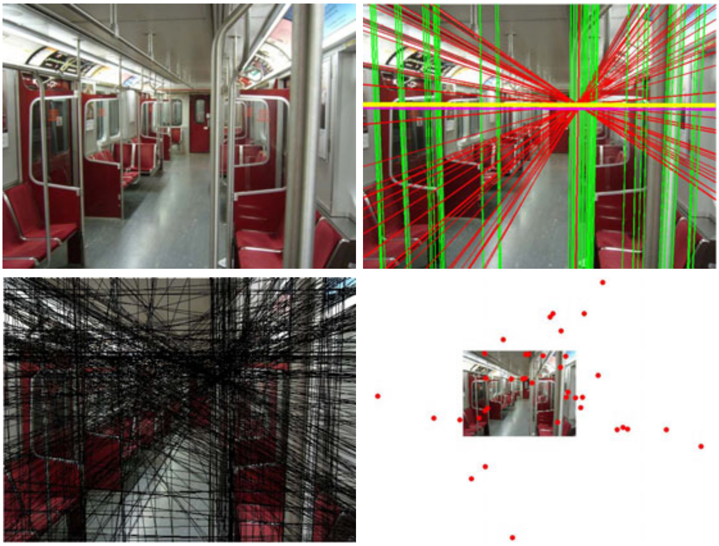


Fig. 4. Sample image from the York Urban dataset: *TopLeft* – the input, *BottomLeft* – all candidate lines superimposed, *BottomRight* – all candidate vanishing points, *TopRight* – the result of the parsing. Coloring reflects grouping into parallel families. Yellow and pink thick lines correspond to the found and the ground truth horizons respectively (the pink line is mostly occluded due to a good fit between the two).

Datasets. Our approach is evaluated on two datasets (Figure 2):

1. The *York Urban* dataset [18] contains 102 images of outdoor and indoor scenes taken within the same location with the same camera. Most of the scenes meet the Manhattan world assumption, as the lines available in the scene mostly fall into the three orthogonal families.

2. The *Eurasian cities* dataset is a new set of 103 outdoor urban images. The images come from the cities of different cultures, hence with different line statistics. They were also taken with different cameras. The main difference of the dataset is the abundance of scenes that fit poorly to the Manhattan assumption. During the annotation, we manually specified several most distinctive lines per each distinctive parallel line family in each image (with the interactive tool similar to that of [18]). This allows to estimate the horizon with good accuracy and we use it as ground truth in the comparative evaluation.

Competing methods. We have compared our approach against the two previously published methods:

1. The method of Tardif [12] is a pipeline approach which reported the top performance on the York Urban dataset. For the experiments on the York Urban dataset we used the author code (with the exception of the EM process that was not published and that we reimplemented by carefully following the text of [12]). For York Urban dataset in cases where more than 3 vanishing points were detected, we chose 3 most orthogonal of them as described in the paper [12]. The coordinates of principal point provided by the authors of the dataset were used during orthogonalization. For the experiments in the Eurasian Cities we did not choose most orthogonal points because the dataset contains non-Manhattan scenes. Parameters of EM were chosen on validation set.

2. The method of Kosecka and Zhang [14] is an approach based on the EM-algorithm, alternating between the two stages: estimation of vanishing point coordinates given distribution of corresponding line segments and re-estimation of distribution of line segments according to positions of vanishing points. The process starts with clustering line segments according to their orientation which results in excessive number of clusters. During EM the clusters with close vanishing points are merged together. Also clusters that have little support are pruned. We took the code from implementation of Automatic Photo Pop Up system [3], which uses that method for vanishing points estimation. Parameters of the method were tuned on the validation sets.

Importantly, to put all the methods on an equal footing, we made sure that all three algorithms are provided with the same Canny edge map (we used the parameters suggested by Tardif in [12]). Both baseline methods use line segments, so we use the line segments detection implementation by [12] for both of them.

After running each method we obtain the zenith, as well as a number of vanishing points corresponding to the parallel families of the line segments (for baseline methods) or lines (for our method). We use this information to estimate the position of the horizon in an image. The horizon is estimated in the same way for all methods. Thus, we restrict it to be perpendicular to the line connecting principal point and zenith. So the slope of horizon is given by zenith and we

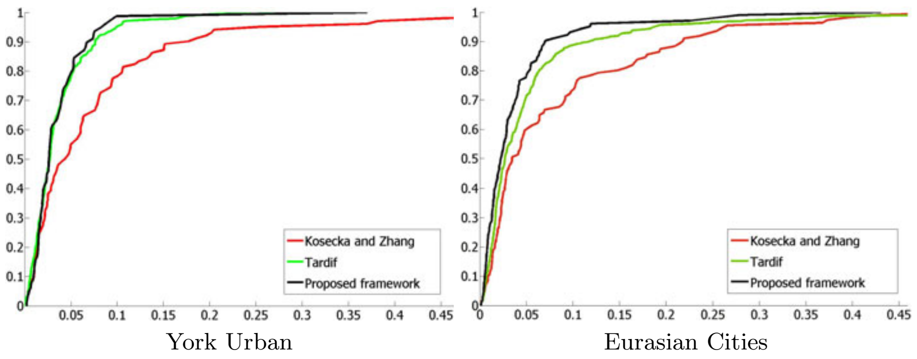


Fig. 5. The results of the comparison of the cumulative statistics for the accuracies of the proposed framework along with the methods of [12] and [14]. The x-axis corresponds to the horizon estimation error measure (see text for more details). The **y**-axis corresponds to the share of the images in the test set that has the error less than the respective x value. In both cases, the proposed framework obtains higher accuracy than the competitors.

estimate only its position along the 1D axis. To do this last step, we perform the weighted least squares fit, where the weight of each detected horizontal vanishing point equals the number of corresponding lines (or line segments).

Accuracy measure. While all the considered approaches essentially output both low-level and high-level primitives, comparing the accuracy of the low-level description of the scenes (e.g. set of lines) is problematic, as the ground truth available for the datasets do not provide full set of lines. Thus, if a line or a vanishing point is present in the output that is missing in the ground truth, it is unclear whether this is due to the error of the algorithm or due to the incompleteness of the ground truth.

We therefore focused on the accuracy of the horizon estimation. Assume that the horizon is given as a (linear) function $H(x)$ of a pixel x-coordinate. Assume that $H_0(x)$ and $H_1(x)$ are the ground truth and the estimated horizon. We then define the estimation error as the maximum of $|H_0(x) - H_1(x)|$ over the image domain ($0 < x <$ image width), divided by the image height. To represent the error over the dataset, we plot the share of the images with the error less then τ for each τ.

Results. Quantitative results are given in Figure 5, while in Figure 4 and Figure 6, we present some qualitative examples from both datasets for our framework. Note that we used the first 25 images of each dataset as a held-out set for the parameter validation for all three competing methods[2]. During the

[2] Through the validation, the parameters for our method for York/Eurasian cities were set to: $\theta_{bg} = 8 \cdot 10^{-5}/7.6 \cdot 10^{-5}, \theta_{dist} = 4 \cdot 10^{-5}/3 \cdot 10^{-5}, \theta_{grad} = 4 \cdot 10^{-5}/2 \cdot 10^{-5}, \eta_{bg} = 0.1/0.1, \eta_{dist} = 1.0/0.8, \lambda_{vp} = 0.015/0.015, \lambda_{line} = 0.003/0.01, \kappa_{hor} = 2.0/5.0$. All angular differences were measured in radians.

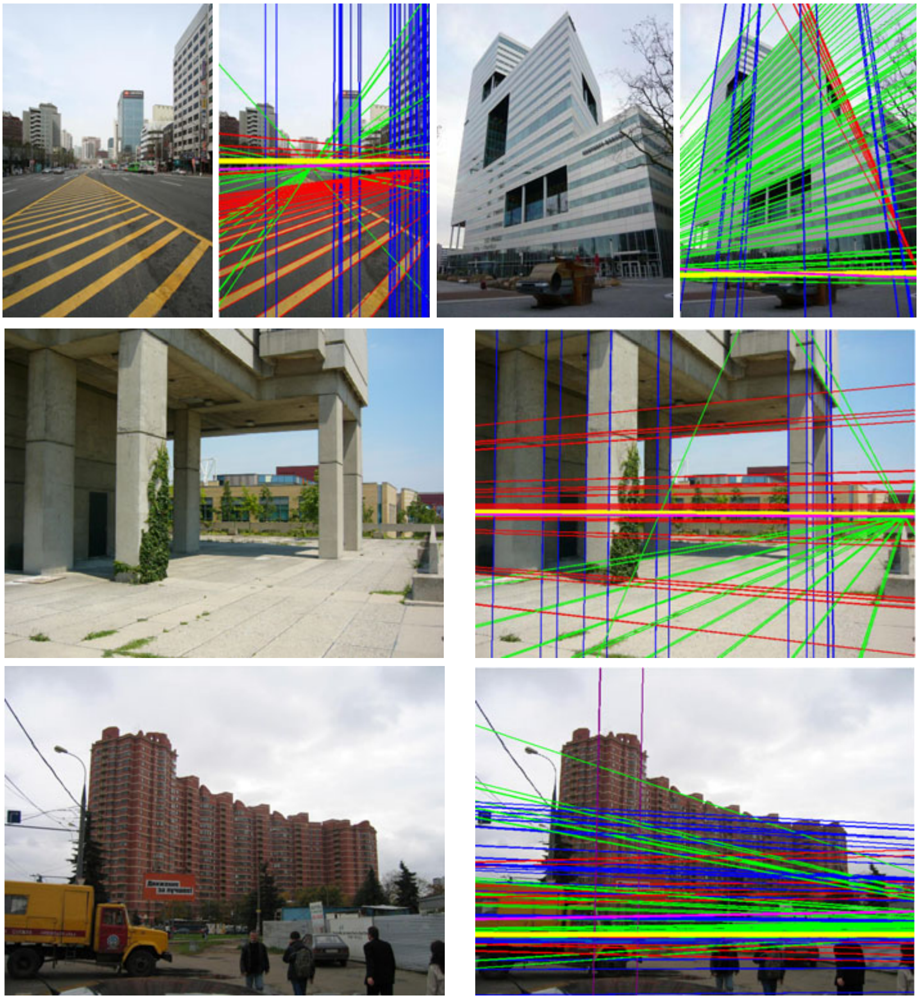


Fig. 6. Sample results of the proposed framework from both datasets. In each pair, we give the input image and the output of the parsing. Coloring reflects grouping into parallel families. Yellow and pink thick lines correspond to the found and the ground truth horizons respectively. Top rows show examples of successful applications, while the bottom one demonstrates one of the worse cases (due to the severely cluttered edge map, the horizon has been estimated significantly below the ground truth).

validation, the area under curve statistics on the hold-out set was optimized. The accuracy measures in the plots in Figure 5 thus reflect the performance on the rest of the images.

As can be seen, the method presented in the paper outperforms both competing methods considerably on the Eurasian cities dataset and performs on a par with [12] and much better than [14] on the York Urban dataset. The latter is all the more important, given the fact the stronger competing method [12] makes explicit use of the Manhattan assumption that is very appropriate for the York dataset, while our method worked with the more general non-Manhattan world model. At the same time, our current implementation is much slower than the competing methods (few minutes per image vs. few seconds per image on a modern PC). The time for our method is dominated by the candidate (lines and VPs) generation and graph construction, and can be reduced significantly is less exhaustive number of candidates would be considered.

In addition to our main error measure (horizon accuracy), we also estimated the error of the zenith estimation on York urban dataset (where ground truth Manhattan geometry allows accurate localization of the zenith). We measured the errors as the angle between directions to the ground truth zenith and the estimated zenith on a Gaussian sphere [20]. The error for our method (0.0118 ± 0.0292) and for the method [14] (0.0133 ± 0.0139) were lower than the error-rate for [12] (0.0402 ± 0.1918).

5 Summary and Discussion

We formulated the problem of geometric analysis of a single image in an optimization framework. Given a set of observed edge pixels, the framework jointly infers groupings of edge pixels into lines, parallel lines, vanishing points and geometric concepts such as the zenith and the horizon. This framework has advantages over previous bottom up methods for inference of such geometric properties; the most significant one being the ability to incorporate a confidence measure about scene elements in a joint framework.

We observed that many failures of the algorithms resulted from the clutter in the edge map (Figure 6 gives an example). As demonstrated by previous works (e.g. [12]), the effect of the clutter may be reduced substantially by local grouping into line segments. In our framework, this can be accomplished by augmenting the graphical model with one more layer situated between the edge pixels layer and the lines layer.

The current framework also ignores appearance information from the scene elements. For instance, parallel lines arising due to a railway track or a road might have similar appearance which may provide additional cues for grouping lines and inferring the location of the zenith and the horizon. This information can produce better results and is a topic for future work. Another interesting direction of work is the incorporation of an uncertainty measure in the presence of edges.

References

1. 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: CVPR, vol. (1), pp. 203–209 (2004)

2. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: ICCV, pp. 654–661 (2005)
3. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. ACM Trans. Graph. 24, 577–584 (2005)
4. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. International Journal of Computer Vision 80, 3–15 (2008)
5. Duric, Z., Rosenfeld, A.: Image sequence stabilization in real time. Real-Time Imaging 2, 271–284 (1996)
6. McLean, G.F., Kotturi, D.: Vanishing point detection by line clustering. IEEE Trans. Pattern Anal. Mach. Intell. 17, 1090–1095 (1995)
7. Tuytelaars, T., Gool, L.J.V., Proesmans, M., Moons, T.: A cascaded hough transform as an aid in aerial image interpretation. In: ICCV, pp. 67–72 (1998)
8. Cipolla, R., Drummond, T., Robertson, D.P.: Camera calibration from vanishing points in image of architectural scenes. In: BMVC (1999)
9. Antone, M.E., Teller, S.J.: Automatic recovery of relative camera rotations for urban scenes. In: CVPR, pp. 2282–2289 (2000)
10. Almansa, A., Desolneux, A., Vamech, S.: Vanishing point detection without any a priori information. IEEE Trans. Pattern Anal. Mach. Intell. 25, 502–507 (2003)
11. Aguilera, D.G., Lahoz, J.G., Codes, J.F.: A new method for vanishing points detection in 3d reconstruction from a single view. In: Proc. of ISPRS Commission V (2005)
12. Tardif, J.P.: Non-iterative approach for fast and accurate vanishing point detection. In: ICCV (2009)
13. Collins, R., Weiss, R.: Vanishing point calculation as a statistical inference on the unit sphere. In: ICCV, pp. 400–403 (1990)
14. Kosecká, J., Zhang, W.: Video compass. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 476–490. Springer, Heidelberg (2002)
15. Rother, C.: A new approach for vanishing point detection in architectural environments. In: BMVC (2000)
16. Coughlan, J.M., Yuille, A.L.: Manhattan world: Compass direction from a single image by bayesian inference. In: ICCV, pp. 941–947 (1999)
17. 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. 175–205. Springer, Heidelberg (2002)
18. Denis, P., Elder, J.H., Estrada, F.J.: Efficient edge-based methods for estimating manhattan frames in urban imagery. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 197–210. Springer, Heidelberg (2002)
19. Tu, Z., Chen, X., Yuille, A.L., Zhu, S.C.: Image parsing: Unifying segmentation, detection, and recognition. International Journal of Computer Vision 63, 113–140 (2005)
20. Barnard, S.: Interpreting perspective images. Artificial Intelligence 21, 435–462 (1983)
21. Beardsley, P., Murray, D.: Camera calibration using vanishing points. In: BMVC, pp. 416–425 (1992)
22. Barinova, O., Lempitsky, V., Kohli, P.: On detection of multiple object instances using hough transforms. In: CVPR (2010)
23. Besag, J.: On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society B-48, 259–302 (1986)