Extraction, matching and pose recovery based on dominant rectangular structures.

Jana Košecká and Wei Zhang

Department of Computer Science
George Mason University
Fairfax, VA 22030

Abstract

Man-made environments possess many regularities which can be efficiently exploited for image based rendering as well as robotic navigation and localization tasks. In this paper we present an approach for automatic extraction of dominant rectangular structures from a single view and show how they facilitate the recovery of camera pose, planar structure and matching across widely separated views. In the presented approach the rectangular hypothesis formation is based on a higher level information encoded by the presence of orthogonal vanishing directions, the dominant rectangular structures can be detected and matched despite the presence of multiple repetitive structures often encountered in a variety of buildings. Different stages of the approach are demonstrated on various examples of images of indoors and outdoors structured environments.

Key words: Model based reconstruction, pose recovery, wide-baseline matching

1 Introduction and Related work

Previous approaches to acquisition of 3D models from multiple views differ in the type of chosen geometric primitives, estimation algorithms as well as level of human interaction. There exist several systems for completely automated recovery of camera motion and 3D structure of the scene [4]. In many instances these general methods lack robustness, are well conditioned only in restricted scenarios and rely on successful solution to feature correspondence, which becomes difficult when the views are widely separated. The techniques that have enjoyed success in limited domains typically employ structural information of the environment. Examples of such systems are PhotoModeler [8] and Facade [9]. These systems were used for building 3D models of architectural environments, which are naturally parameterized by cubes, tetrahedrons,

prisms, arches, surfaces of revolutions and their combinations, used partial human interaction to instantiate the model primitives in respective views and yielded quality of the models superior to the fully automated methods.

The past attempts to automate the geometric model selection and matching typically resorted to weaker geometric assumptions, such as presence of linear and planar structures combined with orthogonality and parallelism relationships between them. These weaker modelling assumptions have been successfully incorporated into fully automated system for multi-view reconstruction [10]. Examples of stronger more constrained models (e.g. doorways, different window types, facades) and their automatic instantiations have been explored in automated methods as well [11]. The constraints of parallelism and orthogonality between planes and lines were used for reconstruction of 3D models in case of uncalibrated camera [12] from single view. Partially calibrated camera and linearly parameterized models have been used for the recovery of 3D structure from a single view [13].

The assumptions and models for the wide-baseline feature matching explored in the past differed in the type of primitives detected in individual views, descriptors associated with their support regions and chosen similarity criteria. In order to account for variation in the appearance due to the change of viewpoint, methods for selecting and matching neighborhoods based on descriptors invariant to rotation, affine transformations and/or scale have been proposed in [14–17]. These local descriptors work well when the individual feature support regions have distinct appearance characterized by either color or texture. In cases where the perspective fore-shortening effects become dominant the affine models are no longer appropriate. The detection and matching of rectangular regions has been previously proposed by [18], in the context of the same problem and by [19] in the context of texture analysis. The approach of [18] proceeded with instantiation of the planar hypothesis in a bottom up manner by linking and grouping detected line segments to form initial rectangular hypothesis. The rectangular regions obtained in such a manner have a small extent and hence are more prone to mismatching in additional views specially in the presence of repetitive structures.

The work presented here focuses on the automatic extraction, detection and matching of rectangular structures detected in individual views. Rectangular planar structure is an image of a 3D rectangle. For example building facades, windows, bulletin boards can in many instances be modelled appropriately by rectangular planar structures. Given the detected rectangular structures we will show how to recover a relative pose of the camera with respect to the 3D world in case of partially calibrated camera and match the detected structures across wide baselines. The presented approach extends the applicability of the automated image based rendering methods to a larger class of man-made environments and is also useful in the context of visual navigation

and localization tasks. The main contributions of the approach are in: (1) the structure extraction stage, which exploits higher level information encoded by the presence of dominant vanishing directions and does not rely on low-level, often brittle, search for geometric structure. In our case we can establish the notion of dominant rectangular structures which make the process of pose recovery better conditioned and simplify the matching stage; (2) we outline a simple method for the camera pose recovery from single view for the case of partially calibrated camera; (3) and demonstrate improvements in the matching stage, which enable us to handle large changes in the viewpoint and slant of the planar structures and establish matches in the presence of large scale repetitive structures.

2 Approach

Our approach is based on the observation that in man-made environments the majority of lines is aligned with three principal directions of the world coordinate frame. The groups of parallel lines belonging to the same direction intersect in the image at the vanishing point. The fact that in man-made environments the sets of parallel lines often come from three mutually orthogonal vanishing direction provides effective constraints for calibrating the camera and recovering the relative orientation of the camera with respect to the scene [20,21]. Rectangular structure is defined by four line segments which come from two different orthogonal line's groups. While these types of structures are easily detected by humans, automatic detection of rectangular structures from images is not straightforward. Simple exhaustive grouping of the initial set of line segments aligned with three principal directions would yield a large number of candidates for rectangular structures, many of them not corresponding to the actual planar structures in the world. In the first part of this paper we describe an approach for merging, pruning and verifying the rectangular structure hypothesis in the image. In the second part we demonstrate how to recover the relative pose of the camera and 3D structure of the rectangular primitives and match them across widely separated views.

2.1 Vanishing point estimation

The starting point of our method is an efficient line detection procedure and vanishing point estimation. The gradient orientation is first quantized into a set of bins containing pixels with similar gradient orientations [21], followed by connected component analysis within each bin and line fitting. The parallel lines in the world intersect in the image plane in vanishing points. The intersection point can be finite or infinite, depending on the relative orientation of

the camera with respect to the scene.

Consider the perspective camera projection model, where 3D coordinates of points $\mathbf{X} = [X, Y, Z, 1]^T$ are related to their image projections $\mathbf{x} = [x, y, 1]^T$ in the following way

$$\lambda \mathbf{x} = KPg\mathbf{X}. \tag{1}$$

$K \in SL(3)$ is the intrinsic camera parameters matrix, $P = [I_{3\times3}, 0] \in \mathbb{R}^{3\times4}$ is the projection matrix, $g = (R, T) \in SE(3)$ is a rigid body transformation represented by 4×4 matrix using homogeneous coordinates and λ is the unknown scale corresponding to the depth Z of the point $\mathbf{X}$. In the above equation both $\mathbf{x}$ and $\mathbf{X}$ are in homogeneous coordinates. Given two image points $\mathbf{x}_1$ and $\mathbf{x}_2$, the line passing through the two endpoints is represented by a normal of a plane going through the center of projection and intersecting the image in a line l, such that $\mathbf{l} = \mathbf{x}_1 \times \mathbf{x}_2 = \widehat{\mathbf{x}}_1\mathbf{x}_2$ [1]. The vanishing direction of two lines which are parallel in 3D world then corresponds to the plane normal where all these lines lie. Given two lines the common normal is determined by $\mathbf{v} = \mathbf{l}_1 \times \mathbf{l}_2 = \widehat{\mathbf{l}}_1\mathbf{l}_2$. Hence given a set of line segments belonging to the lines parallel in 3D, the common vanishing direction $\mathbf{v}$ can be obtained by solving the linear least squares estimation problem $\min_{\mathbf{v}} \sum_{i=1}^{n} (\mathbf{l}_i^T\mathbf{v})^2$. This corresponds to $\min_{\mathbf{v}} \|A\mathbf{v}\|^2$, where the rows of matrix $A \in \mathbb{R}^{n\times3}$ are the lines segments $\mathbf{l}_i$ belonging to the same vanishing direction. Given a set of line segments sharing the same vanishing direction, the above orthogonal least squares solution is applicable regardless of the camera being calibrated. Prior to the vanishing point estimation the detected line segments need to be grouped into the dominant vanishing directions.

Previous techniques for line segment grouping vary in the choice of the accumulator space, where the peaks correspond to the dominant clusters of line segments; most common alternatives are the Gaussian sphere and Hough space [1,3,5,7]. When the camera is calibrated, the image line segments are represented as unit vectors on the Gaussian sphere and several techniques for both grouping and initialization stage on the Gaussian sphere exist [1–3]. The main advantage of the Gaussian sphere representation is the equal treatment of all possible vanishing directions, including those at infinity. The initialization and grouping are the determining factors of the efficiency of the previously proposed methods. An approach for simultaneous grouping and vanishing point estimation using Expectation Maximization algorithm has been suggested previously by [1], assuming calibrated camera and Gaussian Sphere representation. In the absence of calibration, the peaks on the Gaussian sphere are not well separated making the grouping problem poorly conditioned. In our previous work [21], we have demonstrated an efficient approach for simultaneous grouping of lines into dominant vanishing directions and estimation of vanishing points using expectation maximization algorithm

[1] $\hat{\mathbf{x}}$ is a skew symmetric matrix associated with $\mathbf{x} = [x_1, x_2, x_3]^T$.

(EM) in an uncalibrated setting. Namely we have shown that by applying arbitrary non-singular normalizing transformation A can be applied to our measurements $\mathbf{l}_i$ and then transforming the result $\mathbf{v}$ back does not affect the final estimates. Hence we can first transform all the endpoints of lines by A^{-1}, in order to make the line segments and vanishing directions well separated on the unit sphere and consequently similar to the calibrated setting [2]. We can now apply the Expectation Maximization algorithm (EM), which estimates the coordinates of vanishing points as well as the probabilities of individual line segments belonging to particular vanishing directions. The posterior distribution of the vanishing points given line segments can be expressed using Bayes rule in terms of the conditional distribution and prior probability of the vanishing points

$$p(\mathbf{v}_k \mid \mathbf{l}_i) = \frac{p(\mathbf{l}_i \mid \mathbf{v}_k)p(\mathbf{v}_k)}{p(\mathbf{l}_i)} \tag{2}$$

where $p(\mathbf{l}_i \mid \mathbf{v}_k)$ is the likelihood of the line segment belonging to a particular vanishing direction $\mathbf{v}_k$. This posterior probability captures the membership probability of a line $\mathbf{l}_i$ belonging to k-th vanishing direction and will be denoted by w_{ik}. For a particular line segment, $p(\mathbf{l}_i)$ can be expressed using the conditional mixture model representation

$$p(\mathbf{l}_i) = \sum_{k=1}^{m} p(\mathbf{v}_k)p(\mathbf{l}_i \mid \mathbf{v}_k) \tag{3}$$

The EM algorithm then proceeds in a two stage iterative fashion, where during each iteration, the posterior probabilities $p(\mathbf{v}_k \mid \mathbf{l}_i)$ are computed given the currently available vanishing points estimates. In the maximization step, the vanishing points are estimated by minimizing negative log likelihood. This in case of Gaussian likelihood distribution yields the following linear least-squares estimation problem

$$J(\mathbf{v}_k) \;=\; \min_{\mathbf{v}_k} \sum_i w_{ik}(\mathbf{l}_i^T \mathbf{v}_k)^2 \;=\; \min_{\mathbf{v}_k} \|(WA\mathbf{v}_k)\|^2 \tag{4}$$

[2] In our case A is given by

$$\mathbf{x} = A^{-1}\mathbf{x}' = \begin{bmatrix} \frac{1}{f^*} & 0 & -\frac{o_x^*}{f^*} \\ 0 & \frac{1}{f^*} & -\frac{o_y^*}{f^*} \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x}'.$$

Given an image of size $s = [nrows, ncols]$ the choice of the transformation A is determined by the size of the image and captures the assumption that the optical center is in the center of the image and the aspect ratio $k = 1$. The focal length in the pixel units is $f^* = nrows$, $o_x^* = \frac{nrows}{2}$ and $o_y = \frac{ncols}{2}$. Given the assumptions about optical center and aspect ratio, the chosen focal length f^* is related to the actual focal length by a scale factor.

where $\mathbf{v}_k$ is a vanishing point associated with k-th vanishing direction, $W \in \mathbb{R}^{n \times n}$ is a diagonal matrix of weights (membership probabilities) and rows of $A \in \mathbb{R}^{3 \times n}$ are the detected line segments. Figure 2 depicts the iterations of EM algorithm and shows an example of vanishing point estimation. The line segments which are not aligned with principal directions are classified as outliers and are discarded from the matching process. More detailed description of the initialization stage, estimation and adjustment of the number of models can be found in [21].

2.2 Rectangular structure extraction

As we mentioned above one rectangular structure in 3D world is delimited by four lines from two principal directions. One approach would be to extend the existing line segments and search for all possible pairs of lines from two orthogonal directions. We next describe the process of refining the detected line segments, and forming and verifying the initial rectangular structure hypothesis.

Line Segment merging. For efficient and accurate rectangular regions extraction, we want to handle only small number of long line segments. The line segments estimates are first refined by combining vanishing point information and original line orientation. Each image line is modelled as $(\mathbf{x}_c, \theta)$ where $\mathbf{x}_c$ is the centroid of the segment and θ is its direction. In case the segment belongs to k-th (finite) vanishing direction ($k = 1, 2, 3$), the line orientation is refined by weighting θ with vanishing point direction defined by $\theta_v = \text{atan}(d_y, d_x)$, where $\mathbf{d} = [d_x, d_y]^T = \mathbf{v}_k - \mathbf{x}_c$. The new direction then becomes

$$\theta_{new} = \mu \times \theta + (1 - \mu) \times \theta_v \tag{5}$$

where μ is the membership probability of the line belonging to the k-th vanishing direction. After this step, the line segments are more consistent with the vanishing directions. This enables us to merge the shorter line segments detected in the first stage. In Hough space a line candidate $(\mathbf{x}_c, \theta)$ is represented by a point (ρ, θ), such that

$$\rho = x_c \cos\theta + y_c \sin\theta. \tag{6}$$

By transforming the obtained lines to Hough space while keeping the resolution of the space high, we check whether multiple lines fall in the same cell and merge them. The new extended line candidate is obtained by 1) computing the two end points of a new line defined by maximum and minimum of extremal points of the incident lines; 2) the middle point is defined by new centroid of two end points; 3) the mean of contributing line directions is considered to

be the new line direction. In the second stage, the resolution of the Hough space is decreased and only single dominant line segment is kept for each cell. This step substantially improves the line segments used for initial hypothesis formation and also eliminates dramatically the number of line segment candidates. Figure 1 shows the originally detected lines and refined lines. We can see now the structure information is much more evident.

Rectangle hypotheses initialization. Given only small number of extended line segments, we exhaustively choose two line candidates from each group, and compute their intersection points. In case the selected lines indeed delineate a rectangular planar patch, there should exist real corners points within a small neighborhood of the predicted corner position. If all four points satisfy the requirement we initiate a rectangular structure hypothesis. This enables us to reject hypotheses as the one depicted in Figure 3.

The image patch represented by the four hypothesis corners and corresponding line segments undergoes an additional verification stage.

Hypothesis verification. In this stage, given the rectangular structure hypothesis, we choose to keep or discard it by checking whether the whole patch delimited by four line segments indeed comes from the same plane. Recall that any planar mapping between the 3D world plane and the image plane can be characterized by a homography $H \in \mathbb{R}^{3\times3}$ which relates the coordinates of points from two respective planes. Without loss of generality we can assume that the points in the 3D world plane are specified by homogeneous coordinates $\mathbf{X} = [X, Y, 1]$ and point coordinates in the image plane are denoted by $\mathbf{x} = [x, y, 1]^T$. The relationship between the points is then

$$\mathbf{x} \sim H\mathbf{X} \tag{7}$$

where $\sim$ denotes an equality up to scale and H is the homography matrix. Consider the coordinate frame associated with the plane with one of the rectangle points (e.g. upper left) being the origin and the axis aligned with the sides of the rectangle. The coordinates of four extremal points of a rectangular structure $\mathbf{S}$ expressed in this frame are then

$$\mathbf{S} = \begin{bmatrix} 0 & a & 0 & a \\ 0 & 0 & b & b \\ 1 & 1 & 1 & 1 \end{bmatrix}, \tag{8}$$

where a and b the dimensions of 3D rectangle. Given at least four corresponding points and knowledge of $\mathbf{S}$, H can be recovered linearly from Equation 7.

However H can be recovered only up to scale and in general a and b are unknown. For the purpose of hypothesis verification we can assume $a = b$, which will only introduce a different scaling of two principal directions but won't affect the verification. The recovered homography enables us to warp the hypothesized image patch to a *normalized* fronto-parallel view. Since the choice of the scale a essentially controls the resolution of the warped image we adjust its value depending on the size of the image patch.

The verification step is based on our previous assumption of the presence of dominant orientations used in the vanishing points detection stage. The gradient orientation histogram of warped image should also contain two dominant horizontal and vertical directions. In case additional peaks in the histogram are detected the hypothesis is discarded. Figure 4 shows two rectangular hypotheses and their corresponding warped images. We can easily identify the true one by either of the proposed methods.

In certain instances checking only the corner areas of the warped image is sufficient for verification, since corners are the most likely areas where the planar hypothesis is violated. By discarding the center part, the verification becomes more robust, in scenarios where for example there is a tree or clutter in front of the building. Currently we consider only larger structures in order reduce the number of initial hypotheses. The detected structures can be alternatively organized in hierarchical manner. Figure 5 shows the final set of the verified rectangular structures. Note that they are naturally divided into two groups, coming from the composition of vertical directions with two horizontal vanishing directions respectively and cover most area of the two facades. In this case, the intersection line of the two facades is well defined. This information will be used later for merging the individual single reconstructions to obtain consistent relative pose between the views.

Figure 5 depicts the examples of detected rectangular structures. Some of the rectangles in the figure are artifacts of the visualization method, since they due to intersections of more dominant rectangular structures and do not correspond to structures detected from and verified by images.

3 Camera pose recovery and partial scene reconstruction

In this section we describe a method for recovery of the relative pose of the camera with respect to the world plane from single view. This problem is a variation of techniques used previously for camera pose recovery from a single view [13] and can be solved very efficiently. Detailed analysis of the existing constraints provided by rectangular structures in multi-view uncalibrated setting can be found in [22].

Recall the image coordinate $\mathbf{x}$ is related to its 3D counterpart $\mathbf{X}$ via projection equation (1). In case sufficient number of 3D coordinates is available, the entire projection matrix $\Pi = KPg \in \mathbb{R}^{3\times4}$ can be recovered and factored into intrinsic and extrinsic parameters of the camera. For uncalibrated camera, its intrinsic calibration matrix K and its simplified form K_f have following form

$$K = \begin{bmatrix} f & \alpha_\theta & o_x \\ 0 & kf & o_y \\ 0 & 0 & 1 \end{bmatrix} \qquad K_f = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where f is the focal length of the camera in pixel units, k is the aspect ratio, α_θ is skew factor and $[o_x, o_y]^T$ is the principal point of the camera. We assume zero image skew $\alpha_\theta = 0$, the aspect ratio is $k = 1$ and principal point $[o_x, o_y]^T$ lies in the center of image (or is known) and the calibration matrix thus assumes simple form of K_f above. The basic projection equation can be simplified in the special case, when the partially calibrated camera is viewing a planar scene. Without loss of generality we assume that 3D planar points $\mathbf{X} = [X, Y, 0, 1]^T$ lie on the plane which goes though the origin in the world frame and has a normal vector $\nu = [0, 0, 1]^T$. In such case we have

$$\lambda\mathbf{x} = \begin{bmatrix} \pi_1^T \\ \pi_2^T \\ \pi_3^T \end{bmatrix} \begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix}, \tag{9}$$

where $\pi_1^T, \pi_2^T, \pi_3^T$ are the rows on the projection matrix Π. Since the third coordinate of $\mathbf{X}$ is zero and the intrinsic parameter matrix is K_f, the projection equations can be written explicitly in the following form

$$\lambda\mathbf{x} = \begin{bmatrix} fr_{11} & fr_{12} & ft_x \\ fr_{21} & fr_{22} & ft_y \\ r_{31} & r_{32} & t_z \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = H \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}. \tag{10}$$

$H \in \mathbb{R}^{3\times3}$ here gives an explicit form of homography between the world plane and image plane in case only the focal length f of the camera is unknown. In order to estimate the homography we need to know at least four correspondences between the world and the image plane. Note that despite the fact that we do not know the actual world coordinates of the points $\mathbf{X}$, assuming that we are viewing a rectangular structure, we can parameterize the unknown shape $\mathbf{S}$ in the following way. The four corner points of the rectangular structure $\mathbf{S}$

are

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & \alpha b & \alpha b \\ 0 & b & b & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \tag{11}$$

where b is the height of the rectangle in 3D world and α is ratio between the height and width of the rectangular structure. Factoring S into scaling matrix and the structure part

$$\mathbf{S} = \mathbf{S}_\alpha \mathbf{S}_\mathbf{s} = \begin{bmatrix} \alpha b & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \tag{12}$$

Substituting it into equation (10) we obtain

$$\lambda \mathbf{x} = H \mathbf{S}_\alpha \mathbf{S}_\mathbf{s}. \tag{13}$$

Denote $H_\alpha = H\mathbf{S}_\alpha$ which has the following form

$$H_\alpha = \begin{bmatrix} \alpha b f r_{11} & b f r_{12} & f t_x \\ \alpha b f r_{21} & b f r_{22} & f t_y \\ \alpha b r_{31} & b r_{32} & t_z \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}.$$

Note that H_α is the homography between the rectangular structure and a unit square, which is well defined by the four corner points. Now we are in the position that given the structural information $\mathbf{S}$ the unknown homography H_α can be recovered up to scale as $H = \gamma H_\alpha$ from the constraint (13). In case the homography estimation is poorly conditioned, we choose an alternative dominant rectangular structure from the detected set for the purpose of pose recovery. Due to the special structure of H_α it is now possible to recover the unknown camera pose as well as dimensions of the rectangular structure. Note now that the columns of the rotation matrix can be expressed in terms of the homography matrix and unknown scales. Exploiting the constraints that the columns of the rotation matrix have to be orthogonal ($r_1^T r_2 = 0$) and of unit norm ($\|r_1\| = \|r_2\| = 1$), we can solve for unknown scaling factors by expressing these constraints in terms of entries of the homography matrix. From the orthogonality constraint we obtain

$$\frac{1}{\gamma^2}\frac{1}{\alpha}\frac{1}{b^2}\left(\frac{h_{11}h_{12} + h_{21}h_{22}}{f^2} + h_{31}h_{32}\right) = 0. \tag{14}$$

From the above equation we can estimate f in the following way

$$\hat{f} = \sqrt{\frac{h_{11}h_{12} + h_{21}h_{22}}{-h_{31}h_{32}}} \tag{15}$$

Note that the recovered $\hat{f}$ is independent of b. Dividing the first two rows of H by $\hat{f}$ we obtain

$$H' = \gamma \begin{bmatrix} \alpha b r_{11} & b r_{12} & t_x \\ \alpha b r_{21} & b r_{22} & t_y \\ \alpha b r_{31} & b r_{32} & t_z \end{bmatrix} = \gamma[h'_1, h'_2, h'_3]. \tag{16}$$

Imposing the unit norm constraint on the rotation matrix columns the unknown ratio of dimensions can be calculated as $\hat{\alpha} = \frac{\|h'_1\|}{\|h'_2\|}$ where h'_1, h'_2 are the column vectors of H'. Denoting $\gamma_b = \gamma b = \|h'_2\|$ as a scale factor and eliminating the unknown scales γ_b and α, the unknown camera pose can be extracted from the above equation as

$$g = \begin{bmatrix} r_{11} & r_{12} & \frac{t_x}{b} \\ r_{21} & r_{22} & \frac{t_y}{b} \\ r_{31} & r_{32} & \frac{t_z}{b} \end{bmatrix}. \tag{17}$$

The final column of the rotation matrix can be obtained as $r_3 = r_1 \times r_2$. So finally we recover the focal length and the complete camera pose (R, T) as well as the dimensions of the rectangle up to universal scale. Note the recovered T is inversely proportional to true rectangle dimension b, because we are using a unit square instead of true dimension to compute the homography.

Sensitivity and Degeneracy

In the general configuration the recovery of pose and structure dimensions is well conditioned. Although the orthogonality constraints for the recovery of the focal length has been used previously in the past [20], there were instantiated in terms of estimated vanishing point coordinates as opposed to homography entries. In practical experiments we have found our method to yield more reliable estimates. Since it is difficult to compare the two approaches on a equal footing we demonstrate the sensitivity of our estimates as a function of errors in image coordinates on several simulations with the synthetic data. Assuming focal length $f = 1000$, the length ratio $\alpha = 2$ and the size of rectangular structure was set to 220×400, which was a typical size of the largest rectangular structure detected in our experiments. The four corner coordinates were perturbed by Gaussian noises with different σ. Figure 13 shows the

median of estimated error obtained from 1000 trials. Note that for $\sigma = 2$ the median error of estimated length ratio is around 0.06 and the relative error is only $0.06/2 = 3\%$. The effect of the size of rectangular structure and the relative orientation of the camera on the final estimates of focal length and length ratio parameters is in Figure 14. Note that reducing the size of the structure affects the final estimates noticably.

There are two degenerate configuration of the above method. When there is no rotation $R = I$, h_{21}, h_{12}, h_{31} and h_{32} are all 0 and no unique solution for the focal length can be obtained. The length ratio α can still be computed as $\frac{h_{11}}{h_{12}}$. Second, when the rotation axis coincides with horizontal (R_x) or vertical coordinate axis (R_y), h_{31} or h_{32} are equal to 0 and the focal length and the length ratio cannot be recovered.

Partial reconstruction of facade

The two rectangular structures belonging to different dominant planes enable us to recover their dimensions and camera pose with regard to the reference frames they define, say (R_l, T_l) and (R_r, T_r), up to "different" universal scales. We can reconcile this by assigning the same origin to the two frames. Any point in the intersection line between of the two planes can be used, with the end corner point they share being the most convenient one. The most upper vertex of the building is chosen in Figure 5. The relative scale is then $\eta = \frac{T_r}{T_l}$ and scale T_l can then be adjusted accordingly to ηT_l, because the two translations should be the same. The two recovered rotations show the perpendicular relationship between two facades with only 3^o error. To visualize the result, we still use the world coordinate frame defined by two facades, with the camera pose expressed in this frame as $g' = (R^T, -R^T T)$. Figure 6 shows the recovered structure and pose based on two facades recovered in Figure 5.

Additional examples of the structure detection results are in Figure 7 and Figure 8 applied to indoors environment.

Given the detected rectangular structures in two views we now demonstrate how to establish their correspondence, and use it for the recovery of the relative camera pose between the views.

4 Matching rectangular structures

Similarly as in the verification stage we warp the rectangular structures detected in individual views into canonical fronto-parallel view. The matching uses both pictorial and geometric information and proceeds in three stages:

(1) comparison of ratios of rectangle sizes; (2) normalized cross correlation of normalized warped views; (3) consistency coplanarity check based on homography between the two views. We next describe these three steps. Since we already have the size ratio $\alpha = \frac{a}{b}$ of the height and the width of each rectangular structure, in order for two structures to match, they must in ideal case have the same size ratio α. In practice, we allow for a small variance of α. In case the structure to be matched has ratio α_t, candidates with α between $[\frac{\alpha_t}{1+err}, (1+err) \times \alpha_t]$ will pass the pre-selection, err is set to be 20% in our experiments. In the second stage the remaining candidates are compared based on pictorial cues. Given the normalized warped views of rectangular structures, we simply choose Normalized Cross Correlation (NCC) measure to asses the similarity between the structures. The corresponding pair is kept if their correlation score is larger than some specified threshold t_{ncc}. After these two stages there are still remaining ambiguities, due to the repetitive nature of the rectangular structures in man-made environments; *i.e.* for one candidate in the first view, there still may be several structures in the second view matching both geometrically and pictorially. As the top of Figure 9 illustrates, multiple structures pass both the geometric and the pictorial test. Note that in the second view (Figure 9 top-left), there are valid matches on the left side of the building, demonstrating that the selected matching criteria and our structure detection method can handle very large distortion.

These remaining ambiguous matches are resolved by using a geometric consistency criterion. The basic assumption behind this criterion is the fact that the dominant rectangular structures detected in the individual views come from the same 3-D plane. In such case, we can exploit the two view relationship between matched structures characterized by a homography matrix H which relates coordinates of two sets of planar points between two views; $\mathbf{x}_2 \sim H\mathbf{x}_1$. Hence, the two view homography can be estimated by selecting a pair of rectangular structures in respective views. For the remaining structure candidates it can be then verified whether they are consistent with the detected homography, by looking at the residual error between warped and actual corner points coordinates

$$|\mathbf{x}_2^j - H\mathbf{x}_1^i| < \epsilon \tag{18}$$

where $\mathbf{x}_1^i$ are coordinates of i-th rectangular structure in the first view and $\mathbf{x}_2^j$ are coordinates of j-th rectangular structure in the second view. Within some tolerance characterized by value ϵ, two structures which are not exactly in the same 3D plane can be matched, as long as the distance of the plane from the camera coordinate system differs by a small amount. This process of estimation of the dominant homography is carried out in spirit similar to RANSAC. First a pair of corresponding structures is picked randomly and it's support is computed. In the final stage the homography with the largest support is chosen. The final estimate is then obtained using all correspondences which comply with the homography. This process enables us to eliminate the remaining mismatches and establish a small number of corresponding rectan-

gular structures in two views. Given the estimated homography H the relative displacement between the views can be obtained by standard decomposition of H into motion $(R, T) \in SE(3)$ and structure parameters. Once the correspondence between rectangular structures has been established the relative displacement can be alternatively computed from the two absolute displacements between the camera and planar structure. Figure 11 shows structure extraction results of two views of a library and their matching results. The motion estimates obtained from the homography decomposition are: the rotation axis $\omega = [0.98408, 0.15094, -0.093846]^T$ and rotation angle $\theta = 12°$, and translation is $T = [-0.080845, 0.34867, -0.36336]^T$, where x-axis is aligned with the vertical direction.

Additional examples of matched structures and recovered relative pose are shown in Figure 12. Note that even though our matching algorithm uses more global information captured by dominant rectangular structures, occlusions caused by trees does not affect the matching results. In the example in Figure 12 the actual camera poses were in reality far apart, but the focal length in the right view was much larger, yielding almost the same apparent size of the building.

5 Summary and discussion

In this paper we described an approach for extraction and matching of dominant rectangular structures. The approach was motivated by our previous work on vanishing points detection and used the assumption that the majority of detected line segments comes from principal vanishing directions associated with the world coordinate frame. We have also demonstrated a simple method for the recovery of planar structure and camera pose from a single view in the absence of focal length. This enabled us to develop a three stage wide baseline matching strategy, which utilized both pictorial and geometric cues. We have demonstrated successful matching and relative pose recovery from widely separated views in the presence of multiple repetitive structures.

We are currently exploring applicability of the proposed method in the context of robotic visual navigation and image based rendering. The presented approach demonstrates, that the use of descriptors with a larger spatial extent, which utilize higher level structural constraints, simplifies certain difficult matching tasks. We are currently investigating alternative choices of matching primitives and representations of their spatial relationships which would enable the applicability of both geometric and appearance based matching in the context of wide-baseline pose and structure recovery as well as recognition tasks.

References

[1] M. Antone and S. Teller. Automatic recovery of relative camera rotations for urban scenes. In *IEEE Proceedings of CVPR*, 2000.

[2] B. Brillaut-O'Mahony. New method for vanishing point detection. *CVGIP: Image Understanding*, 54(2):289–300, September 1991.

[3] R. Collins. Vanishing point calculation as statistical inference on the unit sphere. In *Proceedings of International Conference on Computer Vision*, pages 400–403, 1990.

[4] R. K. M. Pollefeys, L. V. Gool, Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters, in: Proc. of the IEEE ICCV, 1998.

[5] L. Quan and R. Mohr. Determining perspective structures using hierarchical hough transforms. *P.R. Letters*, 9:279–286, 1989.

[6] C. Rother. A new approach for vanishing point detection in architectural environments. In *Proceedings of the British Machine Vision Conference*, 2000.

[7] T. Tuytelaars, M. Proesmans, and L. Van Gool. The cascaded Hough transform. In *Proceedings of ICIP*, pages 736–739, 1998.

[8] R. Cipolla, D. Robertson, E. Boye, Photobuilder – 3D models of architectural scenes from uncalibrated images, in: Proc. IEEE International Conference on Multimedia Computing and Systems, Firenze, 1999.

[9] P. Debevec, C. Taylor, J. Malik, Modelling and rendering architecture from photographs, ACM Computer Graphics, SIGRAPH (1996) 11–20.

[10] T. Werner, A. Zisserman, New techniques for automated reconstruction from photographs, in: ECCV, 2002, pp. 541 – 555.

[11] A. Dick, P.Torr, S. Ruffle, R. Cipolla, Combining single view recognition and multiple view stereo for architectural scenes, in: Proc. 8th IEEE International Conference on Computer Vision (ICCV'01), 2001.

[12] A. Criminisi, I. Reid, A. Zissernman, Single view metrology, International Journal of Computer Vision.

[13] D. Jelinek, C. Taylor, Reconstruction of linearly parametrized models from single images with camera of unknown focal lenght, IEEE Transactions of PAMI (2001) 767–774.

[14] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. Journal on Computer Vision.

[15] K. Mikolajcsyk, C. Schmidt, Affinely invariant feature detector, in: European Conference on Computer Vision, 2002, pp. 128–142.

[16] T. Tuytelaars, L. Gool, Wide baseline stereo matching based on local, affinely invariant regions, in: Proceedings BMVC, 2000, pp. 412–425.

[17] F. Schaffalitzky, A. Zisserman, View-point invariant texture mapping and wide-base line stereo, in: Proceedings of ICCV, 2001.

[18] P. Pritchett, A. Zisserman, Wide baseline stereo matching, in: Proceedings of ICCV, 1998, pp. 767–774.

[19] Y. Liu, R. Collins, Y. Tsin, A computational model for periodic pattern perception based on frieze and wallpaper groups, in: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 26 (3) (2004) 354–372.

[20] D. Liebowitz, Combining scene and auto-calibration constraints, in: Proceedings of ICCV, 1999.

[21] J. Košecka, W. Zhang, Video compass, in: European Conference on Computer Vision, 2002, pp. 657 – 673.

[22] M. Wilczkowiak, E. Boyer and P. Sturm, 3D Modeling Using Geometric Constraints: A Parallelepiped Based Approach. in: European Conference on Computer Vision 2002, Vol IV, pp. 221–237, June 2002.

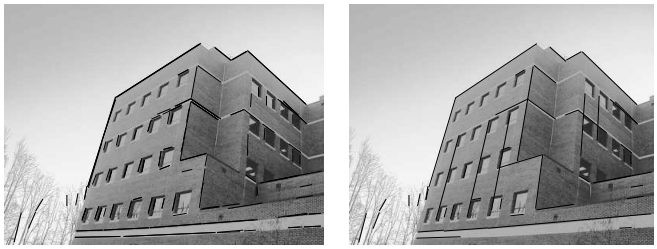

Fig. 1. Initial line segments (left) and refined line segments (right).

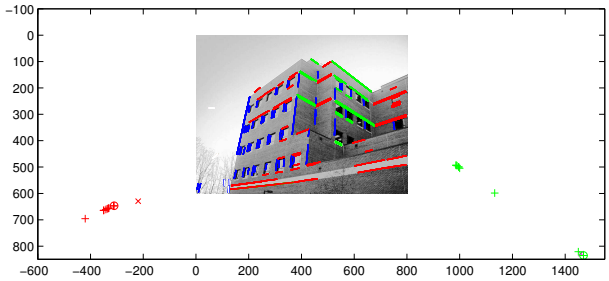

Fig. 2. Iterations of the EM algorithm, detected vanishing points (vertical vanishing point not shown here) and lines belonging to different vanishing directions.

Fig. 3. An example where the intersection of extended lines is outside the rectangular structure.

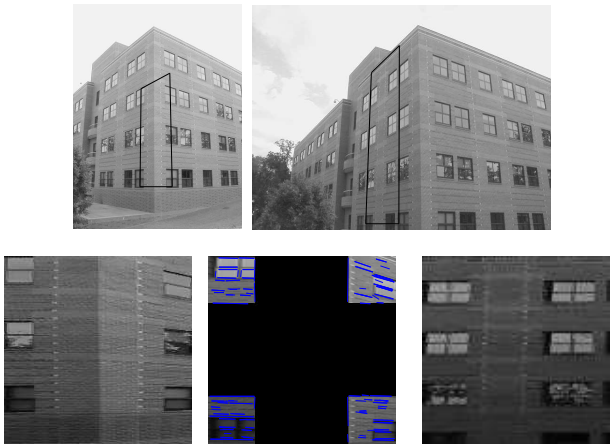

Fig. 4. Hypothesis rectangles and their corresponding warped images, the left is invalid hypothesis and the right one is correct. Verification based on corner areas only, is demonstrated on the warped image patch in the center.

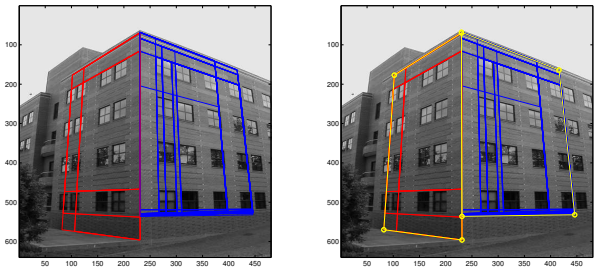

Fig. 5. Rectangle structure extraction result and two initialized building facades.

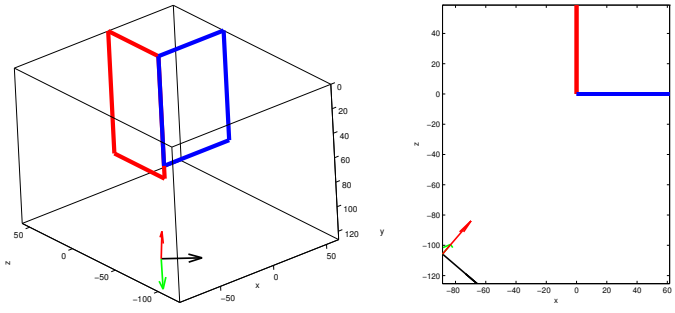

Fig. 6. Frontal view (left) and top view (right) of recovered structures and camera pose based on the two initiated facades of a building in Figure 5.

Fig. 7. Rectangle structure recovered for another building, crosses mark the corners of structures which failed the verification stage.

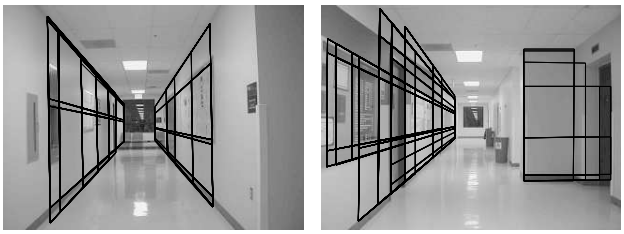

Fig. 8. Rectangle structures detected in indoor environment.

Fig. 9. Structure in first view have multiple matches pictorially (top). The repeated pattern causes mismatch (bottom).

Fig. 10. Library image pair: structure detected in the first view (upper left); structure detected in the second view (upper right); successfully matched structures (bottom).

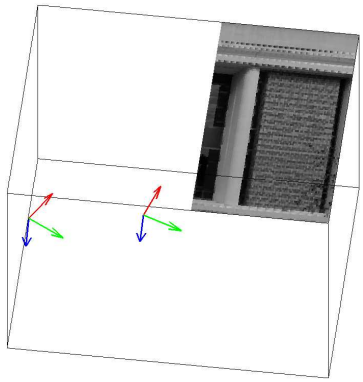

Fig. 11. Pose recovery results for the library image pair.

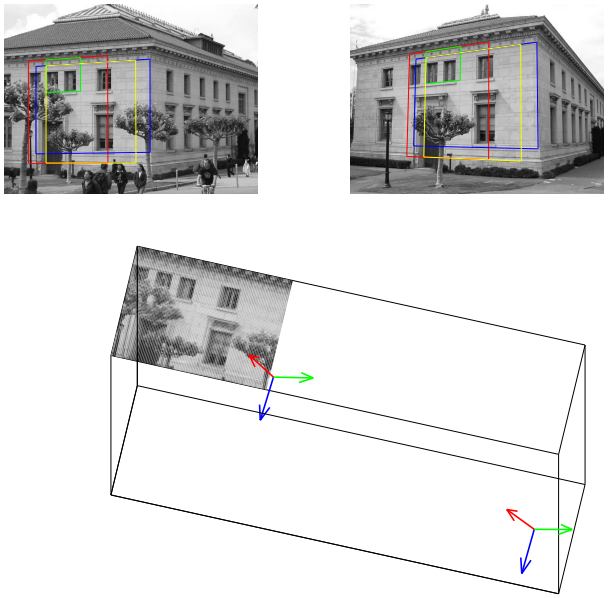

Fig. 12. Matching and relative pose recovery result for California Hall.

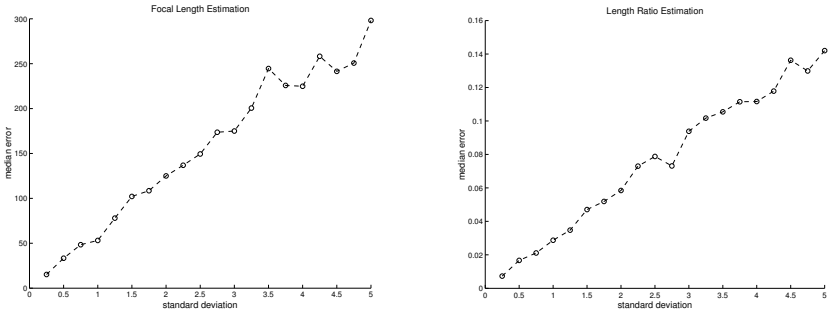

Fig. 13. Sensitivity of the focal length (left) and length ratio (right) estimates as a function of errors in image coordinates.

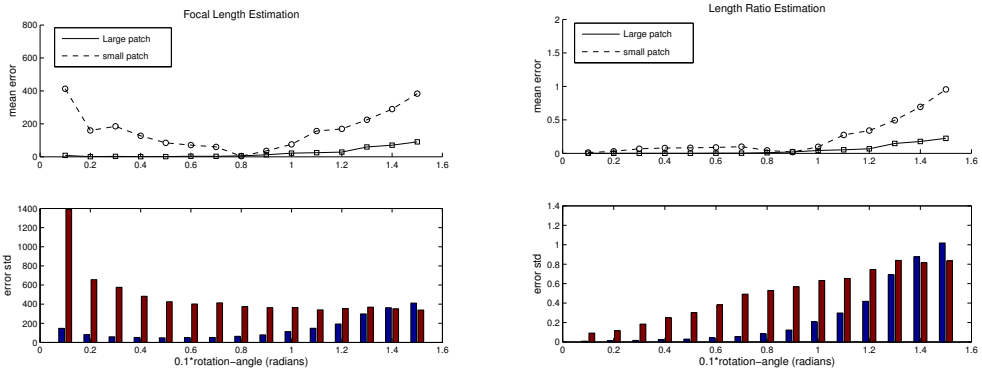

Fig. 14. Sensitivity of the focal length and length ratio estimates as a function of relative rotation (R_y) with respect to the plane and size of the rectangular structure. The error in image coordinates was set to $\sigma = 2$.