Mosaics Construction from a Sparse Set of Views

J. Košecká, W. Zhang and F. Li*

Abstract

In this paper we describe a flexible approach for constructing mosaics of architectural environments from a sparse set of uncalibrated views. The main contribution this paper is the use of environment constraints in order increase the efficiency and level of automation of the mosaic construction process. The observation that in architectural environments, the majority of lines is aligned with the principal orthogonal directions of the world coordinate frame, will be exploited in different stages of the mosaic construction pipeline. The automated detection of vanishing directions will enable us to partially calibrate the camera an estimate the relative orientation of the camera with respect to the scene from a single view. These initial estimates will facilitate efficient feature matching, computation of displacements between the views as well as alignment of multiple views.

While the approach described here will be presented in the context of rotational mosaics, the alignment and matching techniques are applicable for general displacements, where the constraints of man-made environments are present and the displacement between the views is large.

Key words: panoramic mosaic construction, vanishing point estimation, relative orientation, partial calibration using vanishing points.

1 Introduction

The problem of construction of visually realistic models of the surrounding environments is actively being pursued by various researchers in computer vision, computer graphics and photogrammetry. One of the widely used approaches for acquisition of such models are so called image based rendering techniques. The philosophy of image based rendering techniques is to create realistic models of the environments from a set of photographs or a video stream, since these media capture both the geometry and the appearance of the environment. The popularity of image based rendering techniques partly lies in a broad dissemination of digital (still and video) cameras, ease of data acquisition as well as advances in the understanding of the geometric relationship between multiple views in an uncalibrated setting and associated algorithms.

Given multiple views of the scene the approaches typically vary in the type of models they attempt to capture, the amount of 3D information represented in the model, the level of automation in the model acquisition process, model complexity and capability to create new views. The techniques which strive for the full 3D model acquisition enable creation of arbitrary novel views [12]. While it is possible to recover the relative camera displacements and camera calibration information from multiple views, in order to obtain photorealistic models some level of assistance it is often required. This is either at the level of choice of the model (planar surfaces and associated textures maps) or during the image matching and model instantiation process [11, 3]. Alternative representations, which sample the space of all possible appearances of the scene often require storage of large amounts of data [13, 8] and are more applicable for object level models.

Alternative image-based representations of environments are so called image mosaics. In case of mosaics the capabilities of generating novel views of the scene are limited, the mosaics are however superior when it comes to ease of their acquisition. Mosaics have been used in a variety of applications including collection of photographs from aerial and satellite imagery, for video indexing, virtual reality and were also proposed as means of representing visual scenes. Comprehensive review as well as motivation and

associated algorithms can be found in [15].

Most commonly used are so called panoramic or spherical mosaics, which capture the complete 360° view or a full spherical view of the surrounding environment. In such case the images are obtained from cameras mounted on tripods, where individual views are related by pure rotation. Some recent recent efforts focus on enriching the type mosaic representation to the case of general displacements [16] as well as improving the level of automation and efficiency of the existing techniques.

The efficiency and level of automation of the mosaic construction process is mostly determined by the algorithms used for matching and alignment of neighboring views as well as availability of the knowledge of the camera intrinsic parameters. It is often assumed that the sequence of closely separated images (e.g. as acquired by video camera) is available [15] and that intrinsic parameters of the camera are partially known. In such case the individual views have relatively large overlap and are typically aligned using differential motion models (e.g. pure translation, affine models, pure rotation), which are estimated by iterative techniques [16]. In case of large displacements multi-scale representations are adopted and the iterative alignment and warping are interleaved across different levels of the pyramid. Once the views are aligned, and the camera internal parameters are known, the rotations between individual views can be calculated and are followed by the the mapping of each pixel into spherical or cylindrical coordinates. As the displacement between individual views becomes larger, the alignment techniques based of differential motion models often exhibit difficulties of converging to local minima or require large number of iterations.

The flexibility of automatic mosaic construction can be greatly enhanced by developing techniques which enable mosaicing from a sparse set of uncalibrated views. In the sequel we will describe such approach by combining environment constraints and single view analysis in order to yield efficient matching and multi-view alignment of a sparse set of views. Alternative techniques for matching across widely separated views using affine invariants associated with textured regions was proposed in [14].

Paper Outline The goal of this paper is to demonstrate how the qualitative knowledge of the environment can increase the flexibility and efficiency of the image alignment process and mosaic acquisition. We will demonstrate how can the constraints of man-made environments be utilized for partial camera calibration and image matching across widely separated views. The main premise of the approach is the presence of sets of parallel and orthogonal lines and planes aligned with principal orthogonal directions of the world coordinate frame. We briefly review an efficient technique for simultaneous estimation and grouping of the de-

tected lines into dominant vanishing directions and demonstrate how to use this information for partial calibration and estimation of camera orientation with respect to the environment from a single view. The partial calibration and orientation information is then used for guided image matching between widely separated views and estimation and refinement of planar homographies between adjacent views. In this paper we will demonstrate applicability of these techniques for construction of panoramic mosaics. The proposed approach can be applied for the alignment of views related by general displacements.

There is a large body of work related to individual steps of our approach. We will point out the differences and commonalities with our approach along the way.

2 Single View Analysis

Recent efforts in building large city models as well as basic surveillance and monitoring applications often encounter the alignment problem of registering current view to the model or registering widely separate views. Single view analysis can be very instrumental in providing some information about position of the camera and camera intrinsic parameters.

The structural regularities of man-made environments, such as presence of sets of parallel and orthogonal lines and planes can be exploited towards determining the relative orientation of the camera with respect to the scene using the information about vanishing points and vanishing lines. The problem of vanishing point detection and estimation have been addressed numerous times in the past and comprehensive review can be found in more recent publications on the topic [1]. The geometric constraints imposed by vanishing directions on the camera intrinsic parameters and camera rotation as well as associated estimation techniques are well understood and have been used previously in the context of structure and motion recovery problems in the uncalibrated case [4]. Once the detected line segments are grouped into common vanishing directions, the MAP estimates of vanishing points can be obtained by minimizing the distance of the line end points from the estimated line segments leading to a nonlinear optimization problem [4]. An alternative to the nonlinear minimization is a covariance weighted linear least squares formulation suggested first in [6], which tries to minimize the algebraic errors.

Vanishing points detection

The line segments parallel in the 3D world intersect in the image in the vanishing point. Depending on the line orientation the vanishing point can be finite or infinite. The grouping of the line segments into vanishing directions has been often considered separately from the geometric estimation

problems, or it has been studied in the case of calibrated camera [1]. 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 a following way:

$$\lambda \mathbf{x} = Pg\mathbf{X}$$

where $P = [I_{3\times3}, 0] \in \mathbb{R}^{3\times4}$ 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 $\mathbf{X}$. Given two image points $\mathbf{x}_1$ and $\mathbf{x}_2$, the line segment passing through the two endpoints is represented by a plane normal of a plane passing through the center of projection and intersecting the retinal plane in a line l, such that $\mathbf{l} = \mathbf{x}_1 \times \mathbf{x}_2 = \widehat{\mathbf{x}_1}\mathbf{x}_2$ [1]. The unit vectors corresponding to the plane normals $\mathbf{l}_i$ can viewed as points on a unit sphere. The vectors $\mathbf{l}_i$ corresponding to the parallel lines in 3D world all lie in some plane, whose intersection with the Gaussian sphere forms a great circle. The vanishing direction $\mathbf{v}$ then corresponds to the plane normal where all these lines lie and in the noise free case $\mathbf{l}_i^T \mathbf{v} = 0$. 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 following 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 $\mathbf{l}_i$. This particular least squares estimation problem has be studied in [2] assuming the unit vectors on the sphere are distributed according to Binghman distribution. The optimal solution to this type of orthogonal least squares problems is also described in [6].

Uncalibrated camera

In order to be able to determine and adjust along the way the number of groups present in the image some notion of a distance between the line and vanishing direction or two vanishing directions is necessary. In the calibrated setting the angle between two directions is represented by the inner product between two vectors $u^T v$ with $u, v \in \mathbb{R}^3$. In the case of uncalibrated camera the image coordinates undergo an additional transformation K which depends on the

---

$^1\widehat{\mathbf{x}} = \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix}$ is a skew symmetric matrix associated with $\mathbf{x} = [x_1, x_2, x_3]^T$.

internal camera parameters:

$$\mathbf{x}' = K\mathbf{x} \text{ with } K = \begin{bmatrix} \alpha_x & \alpha_\theta & o_x \\ 0 & \alpha_y & o_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} f & \alpha_\theta & o_x \\ 0 & kf & o_y \\ 0 & 0 & 1 \end{bmatrix}.$$

where f is the focal length of the camera in pixel units, k is the aspect ratio and $[o_x, o_y, 1]^T$ is the principal point of the camera. In the uncalibrated setting the vectors undergo additional transformation K and we have $u' = Ku$ and $v' = Kv$ and the inner product $u^T v$ becomes:

$$u^T v = u'^T K^{-T} K^{-1} v'$$

where the unknown matrix $S = K^{-T}K^{-1}$ can be interpreted as a metric of the uncalibrated space. In the following we will demonstrate that by transforming the measurements by an arbitrary nonsingular transformation A has no effect on the computation of the vanishing points. In order to proceed note that the following fact holds: If $v \in \mathbb{R}^3$ and $A \in SL(3)$, then $A^T \widehat{v} A = \widehat{A^{-1}v}$.

Suppose that the endpoints of two lines are $\mathbf{x}'_1, \mathbf{x}'_2$ and $\mathbf{x}'_3, \mathbf{x}'_4$, such that $\mathbf{l}'_1 = \mathbf{x}'_1 \times \mathbf{x}'_2$ and $\mathbf{l}'_2 = \mathbf{x}'_3 \times \mathbf{x}'_4$, where the measurements $\mathbf{x}'_i = A\mathbf{x}_i$ are related to the calibrated image coordinates by some unknown nonsingular transformation A. We can to show that the vanishing point $\mathbf{v}'$ corresponding to the plane normal spanned by vectors $\mathbf{l}'_1$ and $\mathbf{l}'_2$; $\mathbf{v}' = \mathbf{l}'_1 \times \mathbf{l}'_2$ is related to the actual vanishing direction in the original space by the unknown transformation A, namely $\mathbf{v}' = A\mathbf{v}$. Hence we have:

$$\begin{aligned} \mathbf{v}' &= \mathbf{l}'_1 \times \mathbf{l}'_2 = (\widehat{A\mathbf{x}_1}A\mathbf{x}_2) \times (\widehat{A\mathbf{x}_3}A\mathbf{x}_4) \\ &= (A^{-T}\widehat{\mathbf{x}_1}\mathbf{x}_2) \times (A^{-T}\widehat{\mathbf{x}_3}\mathbf{x}_4) = \quad\quad (1) \\ &= (A^{-T}\mathbf{l}_1) \times (A^{-T}\mathbf{l}_2) = A\widehat{\mathbf{l}_1}\mathbf{l}_2 = A\mathbf{v} \quad (2) \end{aligned}$$

The above fact demonstrates that in the context of vanishing point estimation, transforming the image measurements by an arbitrary nonsingular transformation A and then transforming the result back, does not affect the final estimate. We will use this fact in the normalization step of the least squares estimation in the context of EM algorithm.

In the case of man-made environments we will exploit the fact that the dominant vanishing directions are aligned with the principal orthogonal axes e_i, e_j, e_k of the world reference frame. Similarly as in [1] we address the grouping stage and vanishing point estimation stage simultaneously as a problem of probabilistic inference with an unknown model. We assume however that the camera is not calibrated. In such instances the algorithm of choice is the Expectation Maximization algorithm (EM), which estimates the coordinates of vanishing points as well as the probabilities of individual line segments belonging to a particular vanishing directions. We will demonstrate that with proper

normalization, the simultaneous grouping and estimation of the vanishing points using EM can be accomplished in the case of an uncalibrated camera.

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:

$$p(\mathbf{v}_k \mid \mathbf{l}_i) = \frac{p(\mathbf{l}_i \mid \mathbf{v}_k)p(\mathbf{v}_k)}{p(\mathbf{l}_i)} \tag{3}$$

where $p(\mathbf{l}_i \mid \mathbf{v}_k)$ is the likelihood of the line segment belonging to a particular vanishing direction $\mathbf{v}_k$. Hence for a particular line segment, $p(\mathbf{l}_i)$ can be expressed using the conditional mixture model representation:

$$p(\mathbf{l}_i) = \sum_{k=1}^{m} p(\mathbf{v}_k)p(\mathbf{l}_i \mid \mathbf{v}_k) \tag{4}$$

The number of possible vanishing directions m, will vary depending on the image. We assume that there are at most four significant models, three corresponding to the dominant vanishing directions and an additional one modeling the outlier process. The choice of the likelihood term $p(\mathbf{l}_i \mid \mathbf{v}_k)$ depends on the form of the objective being minimized as well as the error model. In the noise free case we have $\mathbf{l}_i^T \mathbf{v}_k = 0$. In the case of noisy measurements we assume that the error ξ_i in $\mathbf{l}_i^T \mathbf{v}_k = \xi_i$ is a normally distributed random variable with $N(0, \sigma_1^2)$. Then the likelihood term is given as:

$$p(\mathbf{l}_i \mid \mathbf{v}_k) \propto exp\left(\frac{-(\mathbf{l}_i^T \mathbf{v}_k)^2}{2\sigma_1^2}\right) \tag{5}$$

Given initial estimates of the vanishing points $\mathbf{v}_k$, the membership probabilities of a line segment $\mathbf{l}_i$ belonging to the k-th vanishing direction are computed in the following way:

$$p(\mathbf{v}_k \mid \mathbf{l}_i) = \frac{p(\mathbf{l}_i \mid \mathbf{v}_k)p(\mathbf{v}_k)}{\sum_{k=1}^{m} p(\mathbf{l}_i \mid \mathbf{v}_k)p(\mathbf{v}_k)} \tag{6}$$

The posterior probability terms $p(\mathbf{v}_k \mid \mathbf{l}_i)$ represent so called membership probabilities, denoted by w_{ik} and capture the probability of a line segment $\mathbf{l}_i$ belonging to k-th vanishing direction $\mathbf{v}_k$. Initially we assume that the prior probabilities of all vanishing directions are equally likely and hence do not affect the posterior conditional probability. The prior probabilities of the vanishing directions can be estimated from the likelihoods and can affect favorably the convergence process as demonstrated in [1]. In the following paragraph we describe the main ingredients of the EM algorithm for simultaneous grouping and estimation of the vanishing directions. Prior to the estimation of vanishing points and grouping of the line segments into common vanishing directions, we first transform all the endpoint measurements by $\tilde{K}^{-1}$; $\mathbf{x} = \tilde{K}^{-1}\mathbf{x}'$. The transformation $\tilde{K}^{-1}$

will make the line segments and the vanishing directions well separated on the unit sphere (as in the calibrated setting). Given an image of size $[dx, dy]^T$ the transformation $\tilde{K}$ is simply related to the size of the image where $\tilde{f} = dx$, $\tilde{o}_x = dx/2$ and $\tilde{o}_y = dy/2$ and has the following form:

$$\tilde{K} = \left[\begin{array}{ccc} \tilde{f} & 0 & \tilde{o}_x \\ 0 & \tilde{f} & \tilde{o}_y \\ 0 & 0 & 1 \end{array}\right]$$

The E-step of the EM algorithm amounts to computation of posterior probabilities $p(\mathbf{v}_k \mid \mathbf{l}_i)$ given the currently available vanishing points estimates. The M-step of the algorithm involves maximization of the expected complete log likelihood with respect to the unknown parameters $\mathbf{v}_k$ [10]. This step yields a maximization of the following objective function:

$$\max_{\mathbf{v}_k} \prod_{i=1}^{n} p(\mathbf{l}_i) = \sum_{i=1}^{n} \log p(\mathbf{l}_i) \tag{7}$$

where $p(\mathbf{l}_i \mid \mathbf{v}_k)$ is the likelihood term defined in equation (5). The above objective function in the case of linear log likelihood model yields a solution to a weighted least squares problem; one for each model. Each line has an associated weight w_{ik} determined by posterior probabilities computed in the E step. In such case the vanishing points are estimated by solving the following linear least-squares problem:

$$J(\mathbf{v}_k) = \min_{\mathbf{v}_k} \sum_i w_{ik}(\mathbf{l}_i^T \mathbf{v}_k)^2 = \min_{\mathbf{v}_k} \|WA\mathbf{v}_k)\|^2 \tag{8}$$

Where $W \in \mathbb{R}^{n \times n}$ is a diagonal matrix associated with the weights and rows of $A \in \mathbb{R}^{3 \times n}$ are the detected line segments. Closed form solution corresponds to the eigenvector associated with the smallest eigenvalue of $A^T W^T W A$ and yields the new estimate of $\mathbf{v}_k$. EM algorithm is an iterative technique guaranteed to increase the likelihood of the available measurements. The iterations of the EM algorithm are depicted in Figure 1. The initially large number of vanishing point estimates, is reduced through the merging process to three dominant directions and the process usually converges in 3-5 iterations. The EM process is efficiently initialized from orientation histograms of the detected line segments, with no need for costly computation of the vanishing points using Hough Transforms. For more technical details of the algorithm see [7].

2.1 Partial Calibration from Single View

The constraints among detected vanishing directions can be used for partial self-calibration. In the uncalibrated case

4
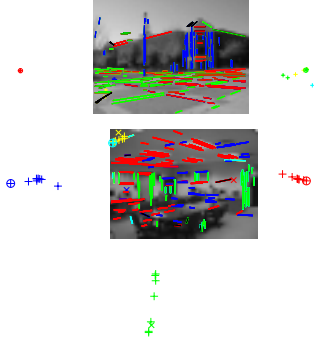
Figure 1. The iterations of the EM algorithm and the detected vanishing points. The lines aligned with the principal orthogonal directions are color coded and the final finite estimates of vanishing points are denoted by 'o'.

the relationship between image coordinates of a point and its 3D counterpart is as follows:

$$\lambda \mathbf{x} = R\mathbf{X} + T$$

Multiplying both sides by calibration matrix K we have:

$$\lambda \mathbf{x}' = KR\mathbf{X} + KT \qquad (9)$$

where $\mathbf{x}'$ denotes a pixel coordinate of $\mathbf{X}$. Let's denote the unit vectors associated with the world coordinate frame to be: $e_i = [1,0,0]^T$, $e_j = [0,1,0]^T$, $e_k = [0,1,0]^T$. The vanishing points corresponding to 3D lines parallel to either of these directions are:

$$\mathbf{v}_i = KRe_i \quad \mathbf{v}_j = KRe_j \quad \mathbf{v}_k = KRe_k$$

and note that the coordinates of vanishing points depend only on rotation and internal parameters of the camera. The orthogonality relations between e_i, e_j, e_k readily provide constraints on the calibration matrix K. In particular we have:

$$
\begin{aligned}
e_i^T e_j &= \mathbf{v}_i^T K^{-T} R R^T K^{-1} \mathbf{v}_j \qquad (10) \\
&= \mathbf{v}_i^T K^{-T} K^{-1} \mathbf{v}_j = \mathbf{v}_i^T S \mathbf{v}_j \qquad (11)
\end{aligned}
$$

where S is the metric associated with the uncalibrated camera introduced earlier:

$$S = K^{-T}K^{-1} = \begin{bmatrix} s_1 & s_2 & s_3 \\ s_2 & s_4 & s_5 \\ s_3 & s_5 & s_6 \end{bmatrix}$$

When three finite vanishing points are detected, they provide three independent constraints on matrix S:

$$
\begin{aligned}
\mathbf{v}_i^T S \mathbf{v}_j &= 0 \\
\mathbf{v}_i^T S \mathbf{v}_k &= 0 \\
\mathbf{v}_j^T S \mathbf{v}_k &= 0 \qquad (12)
\end{aligned}
$$

In general symmetric matrix $S_{3\times3}$ has six degrees of freedom and can be recovered up to a scale, so without additional constraints we can recover the S only up to two parameter family. Other commonly assumed assumption of zero skew and known aspect ratio can also be expressed in terms of constraints on the metric S as proposed earlier in [9]. The zero skew constraint expresses the fact that the image axes are orthogonal can be written as:

$$[1,0,0]S[0,1,0]^T = 0$$

In the presence of zero skew assumption, the known aspect ratio constraint can be expressed as $s_1 = s_4$. With these two additional constraints we have a sufficient number of constraints and the solution for $\mathbf{s} = [s_1, s_2, s_3, s_4, s_5, s_6]^T$ can be obtained by solving a linear least squares estimation problem. Writing the individual constraints as $\mathbf{b}_j^T \mathbf{s} = 0$ and stacking them into matrix B, $\mathbf{s}$ can be obtained by minimizing $\|B\mathbf{s}\|^2$ and corresponds to the eigenvector associated with the smallest eigenvalue of $B^T B$. The calibration matrix K^{-1} can be obtained from S by Cholesky decomposition. In the case one of the vanishing directions lies close to infinity one of the constraints becomes degenerate and recovered S fails to be positive definite. This situation can be also noticed by checking the condition number of B. In such case we assume that the principal point lies in the center of the image and hence S is parameterized by the focal length only as:

$$S = \begin{bmatrix} 1/f^2 & 0 & 0 \\ 0 & 1/f^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In such case the focal length can be recovered in a closed form [5] from the remaining constraint $\mathbf{v}_j^T S \mathbf{v}_k = 0$. The recovered calibration matrices for the examples outlined in Figure 1 are below:

$$K_{building} = \begin{bmatrix} 409.33 & -0.0000 & 177.46 \\ 0 & 409.33 & 165.75 \\ 0 & 0 & 1 \end{bmatrix} \qquad (13)$$

$$K_{room} = \begin{bmatrix} 361.133 & -0.0000 & 263.99 \\ 0 & 361.133 & 129.038 \\ 0 & 0 & 1 \end{bmatrix} \qquad (14)$$

Note that in the above examples the difference in the focal length is due to the difference in the image size. While the

subsampling affects also the position of the principal point, the above statement assumes that the focal length of the sub-sampled images is related to the original focal length by the subsampling factor. The quality of the estimates depends on the accuracy of the estimated vanishing points. As the vanishing points approach infinity their estimates become less accurate. This affects in particular the estimate of the principal point, which in case one of the vanishing points is at infinity cannot be uniquely determined unless additional constraints are introduced [9]. In such case we assume that the principal point of the camera lies in the center of the image and estimate the focal length in the closed form, using a single orthogonality constraint between vanishing directions. The estimate of the focal length obtained in this manner is less accurate then if all the constraints are used simultaneously (if available) and the principal point is estimated as well.

Relative orientation

Once the vanishing points have been detected and the unknown camera parameters determined by the above procedure, the relative orientation of the camera with respect to the scene can be computed. Note first that since the vanishing directions are projections of the vectors associated with three orthogonal directions i, j, k and depend on rotation only. In particular we can write that:

$$K^{-1}\mathbf{v}_i = Re_i \quad K^{-1}\mathbf{v}_j = Re_j \quad K^{-1}\mathbf{v}_k = Re_k$$

with each vanishing direction being proportional to the column of the rotation matrix $R = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3]$. Choosing the two best vanishing directions and properly normalizing them, the third row can be obtained by enforcing the orthogonality constraints as $\mathbf{r}_3 = \widehat{\mathbf{r}}_1 \mathbf{r}_2$. There is a four way ambiguity in R due to the sign ambiguity in $\mathbf{r}_1$ and $\mathbf{r}_2$. Spurious solutions can be eliminated by considering relative orientation or structure constraints.

2.2 Image alignment

The techniques described in the previous paragraph were limited to the single view analysis and enabled us to recover partial calibration of the camera as well as orientation of the camera with respect to the scene. In the case of two uncalibrated views related by rotation only, the image coordinates of corresponding points in two views satisfy:

$$\mathbf{x}_j \propto H\mathbf{x}_i \qquad (15)$$

where $H = KRK^{-1} \in \mathbb{R}^{3\times 3}$ is so called homography with the plane at infinity sometimes denoted as H_∞ and $\mathbf{x}_i = [x_i, y_i, 1]^T$ and $\mathbf{x}_j = [x_j, y_j, 1]^T$ are the image coordinates of the corresponding points. When the orientation
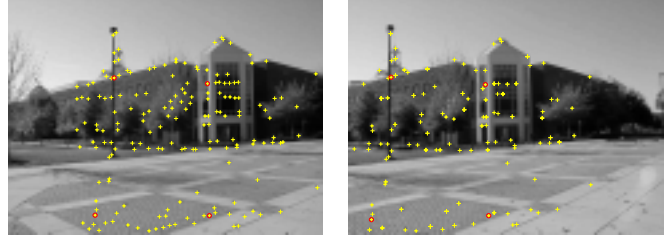
Figure 2. Original set of correspondences, with putative matches found by from homography computed from relative orientations of two single views $\hat{H}_{ij}$.
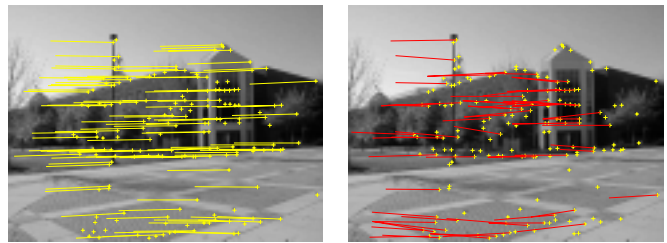
Figure 3. The matched correspondences (left) and detected outliers (right) after RANSAC iterations. Outliers are denoted in red.

and partial calibration information are estimated from a single view, as described in previous section, the homography between two views can be computed and becomes:

$$\hat{H}_{ij} = K R_i R_j^T K^{-1}$$

Together with the knowledge of the camera calibration matrix $K_{building}$ from equation 13 the rotation angle and the axis can be recovered from the rotation matrix $R_i R_j^T$, yielding in this case a rotation of $\phi_{ij} = 14.7^o$ around y axis. The quality of H_{ij} estimate is often not satisfactory for accurate alignment of the views. However the estimate of H_{ij} can serve as a good starting point for guided feature matching and re-estimation of the homography. Note that in order to estimate H in equation 15 at least 4 corresponding points in two views are needed each providing two constraints on H. In the first step of the refinement we detect additional features from each view using Harris corner detector with the results in Figure 2. Traditional robust techniques for estimation of H require at least 4 corresponding points, in order to bootstrap the robust matching algorithm, such as RANSAC. In the case of large displacements the initialization step however can be quite expensive. Without any prior knowledge about the motion potentially large regions in the image have to be searched for possible matches. Given that

Figure 4. Image alignment using the refined homography estimates.

Figure 5. Final mosaic obtained with the refined estimate of the focal length yield correct mapping to the cylindrical coordinates.

we can approximately estimate the homography based on single view analysis, we can select candidate feature points in one view and use $\hat{H}_{ij}$ to predict their location in another view. The predicted location of the feature then limits the neighborhood search to few pixels around the predicted location. The displacement was over 300 pixels in the horizontal direction which would make the search for putative correspondences very consuming and the differential techniques would require a large number of iterations in this case. Examples of the initial and refined estimates of the homographies between two views are:

$$
\begin{aligned}
H_{ij} &= \begin{bmatrix} 1.1799 & 0.0945 & -43.1874 \\ 0 & 1.252 & -356.39 \\ 0 & 0.004 & 1 \end{bmatrix} \\
H &= \begin{bmatrix} 1.1425 & 0.0881 & -34.1227 \\ 0.0071 & 1.2142 & -265.0491 \\ 0 & 0.003 & 1 \end{bmatrix}
\end{aligned} \quad (16)
$$

If we assume H is the reference homography then the error between the two estimates $\frac{\|H - H_{ij}\|}{\|H\|} \times 100$ is 34%.

Applying the homography refinement and matching techniques to all views we can obtain refined homography estimates, which will enable us the to align all the views to the common reference plane 4. The above alignment can be also achieved without any knowledge of the intrinsic parameters of the camera, using computationally more expensive image matching techniques. Once the homographies between the reference view and additional views have been computed the intrinsic parameters of the camera can be reestimated. In the case of panoramic mosaics, where the views are related by the rotation around single independent axis the matrix S can be determined only up to two parameter family and hence all the intrinsic parameters of the camera cannot be estimated. In the case of panoramic mosaics we assume that the skew is zero, the aspect ratio of the camera is one (known) and the principal point is in the image center. In such case the only unknown is the focal

length f, which can be computed in closed form exploiting properties of rotation matrices.

$$
H = \begin{bmatrix} h_0 & h_1 & h_2 \\ h_3 & h_4 & h_5 \\ h_6 & h_7 & 1 \end{bmatrix} \propto \begin{bmatrix} r_{11} & r_{12} & r_{13}/f \\ r_{21} & r_{22} & r_{23}/f \\ r_{31} & r_{32} & r_{33}/f \end{bmatrix} \quad (17)
$$

The fact the any two rows (or columns) of the scaled rotation matrix must have the same norm and be orthogonal gives us constraints for computing f. In order to obtain accurate estimates of rotation and camera intrinsic parameters, spherical mosaics are favorable, since they are composed from multiple views related by rotations of around more then two independent axis. In such case the complete intrinsic parameters can be determined and associated rotation matrices computed [17]. For the mosaic construction accurate focal length parameters are necessary in order to establish the mapping between image coordinates $[x, y, 1]^T$ and cylindrical coordinates:

$$
\theta = \tan^{-1}(\frac{X}{Z}) = \tan^{-1}(\frac{x}{f}) \quad (18)
$$

$$
v = \frac{Y}{\sqrt{X^2 + Z^2}} = \frac{y}{\sqrt{x^2 + f^2}} \quad (19)
$$

where the radius of the cylinder is the focal lenght. The final mosaic obtained from the estimates of homographies and global refinement of the focal length is depicted in figure 5.

3 Conclusions

We presented an efficient, completely automated approach for construction of rotational mosaics from an uncalibrated sparse set of views. Along the way the assumptions about the structure of man-made environments, were used towards efficient initialization and grouping of line segments into dominant vanishing directions aligned with the axes of the world coordinate frame. The estimation and grouping problems for vanishing point estimation were addressed simultaneously using the Expectation Maximization algorithm. The single view analysis was instrumental

for obtaining initial estimates of homographies as well as intrinsic parameters of the camera. These initial estimates enabled us to do efficient matching and final homography computation and view alignment for the case of largely separated views. The presented approach extends currently available techniques for mosaic construction and demonstrates how the constraints of architectural environments can be used efficiently towards this task.

References

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

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

[3] Paul E. Debevec, C.J.Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs:a hybrid geometry-and image-based approach. In *SIGRAPH*, 1996.

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

[5] D. Jelinek and C.J. Taylor. Reconstruction of linearly parametrized models from single images with camera of unknown focal lenght. *IEEE Transactions of PAMI*, pages 767–774, July 2001.

[6] K. Kanatani. *Geometric Computation for Machine Vision*. Oxford Science Publications, 1993.

[7] J. Košecka and W. Zhang. Video compass. Technical report, George Mason University, 2001.

[8] McMillan L. and Bishop G. Plenoptic modellng. an image based rendering system. In *In Computer Graphics Annual Conference Series*, 1995.

[9] D. Liebowitz. Combining scene and auto-calibration constraints. In *Proceedings of ICCV*, 1999.

[10] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications*. Marcel Dekker Inc., N.Y., 1988.

[11] D.P. Robertson R. Cipolla and E.G. Boye. Photobuilder – 3d models of architectural scenes from uncalibrated images. In *Proc. IEEE International Conference on Multimedia Computing and Systems, Firenze*, 1999.

[12] M. Pollefeys R. Koch and L. Van Gool. Multi viewpoint stereo from uncalibrated video sequences. In Springer Verlag, editor, *European Conference on Computer Vision*, 1998.

[13] R. Szeliski S. Gotler, R. Grzeszczuk and M. Cohen. The lumingraph. In *SIGRAPH*, 1996.

[14] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *International Conference on Computer Vision, Vancouver, Canada*, 2001.

[15] R. Szeliski. Video mosaics for virtual environments. In *IEEE Computer Graphics and Applications*, 1996.

[16] R. Szeliski and H. Shum. Creating full view panoramic image mosaics and environment maps. In *Microsoft Research Technical Report*, 1996.

[17] J. Košecka Y. Ma, S. Soatto and S. Sastry. Reconstruction and reprojection up to subgroups. *International Journal of Computer Vision*, 38(3):218:29, 2000.