MatchBox: Indoor Image Matching via Box-like Scene Estimation

Filip Srajer[1] Alexander G. Schwing[2] Marc Pollefeys[3] Tomas Pajdla[1]

[1]CTU in Prague
{srajefil, pajdla}@fel.cvut.cz

[2]University of Toronto
aschwing@cs.toronto.edu

[3]ETH Zurich
pollefeys@inf.ethz.ch

Abstract

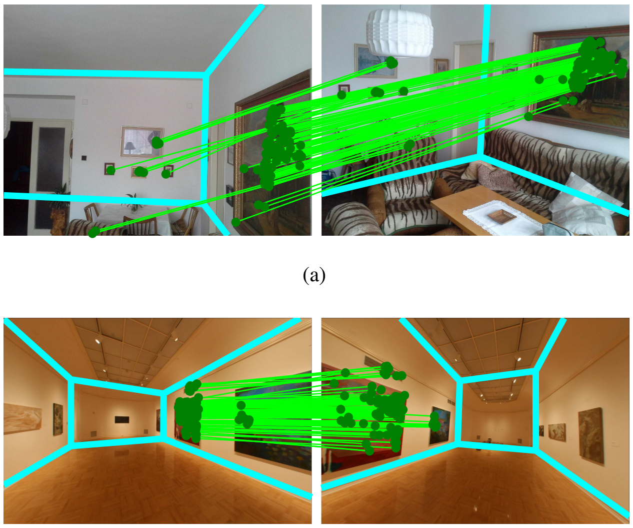
Keypoint matching in images of indoor scenes traditionally employs features like SIFT, GIST and HOG. While those features work very well for two images related to each other by small camera transformations, we commonly observe a drop in performance for patches representing scene elements visualized from a very different perspective. Since increasing the space of considered local transformations for feature matching decreases their discriminative abilities, we propose a more global approach inspired by the recent success of monocular scene understanding. In particular we propose to reconstruct a box-like model of the scene from every single image and use it to rectify images before matching. We show that a monocular scene model reconstruction and rectification preceding standard feature matching significantly improves keypoint matching and dramatically improves reconstruction of difficult indoor scenes.

1. Introduction

Image matching is an important component of 3D scene reconstruction [1, 18], image retrieval [28] and scene and video completion [34, 13, 35]. Considering the variability of the applications, a multitude of approaches have been considered and evaluated in the past.

In general, all the approaches use some feature representation for image keypoints and a distance metric to find visually similar local patches. While the similarity measure is often chosen as the standard Euclidean distance, the employed image features vary largely from variants of the scale-invariant feature transform (SIFT) [22] and histograms of oriented gradients (HOG) [8] to GIST descriptors [25]. Importantly, all those image and keypoint representations are very helpful in capturing *local* transformations such as rotation, illumination or scaling.

Hence, if we assume the data to be reasonably homogeneous, *e.g.*, for small-baseline disparity map estimation, a simple and efficient sum-of-squared-differences approach leveraging the pixel intensity as a feature works very well in practice. For many other situations, however, matching is often surprisingly difficult. In particular, if the observed scene is similar only on a higher level; consider the translation and rotation of the camera as illustrated in the left and right column of Fig. 1 for example. Due to the large transformation, local pixel neighborhoods appear differently in the image space and standard matching approaches often fail in finding any corresponding keypoints.


(a)

(b)

Figure 1. Image pairs that could not be matched by standard techniques were successfully matched by *MatchBox*. The figure visualizes geometrically verified keypoint matches. *MatchBox* found 100 matches in (a) and 99 matches in (b). The original images in (b) are from the dataset of Furukawa *et al*. [11].

The aforementioned issue is well known, as standard keypoint detection typically finds an insufficient number of keypoints in indoor scenarios, *e.g.*, because of low-textured walls. In addition, we note that standard techniques are challenged by large camera movements typically present in indoor scenarios. Therefore, we argue to differentiate between indoor and outdoor scenarios and subsequently suggest a solution specifically tailored for the indoor setting. Restricting ourselves to indoor environments enables us to employ the *Manhattan world assumption*, *i.e*., we model a scene to be aligned according to three dominant and orthogonal directions defined by vanishing points [14, 33].

Hence our approach titled *MatchBox* retrieves keypoint matches after a global scene estimation. We first predict a coarse global scene representation for the observed scenario using a textured 3D cuboid, estimated from each individual image. In the second step we rectify local constituents instead of sacrificing discriminative power by only increasing the space of local transformations. Using MatchBox on the image pairs illustrated in Fig. 1 enables us to find about 100 matching keypoints that undergo a large transformation while standard feature matching does not result in any correspondences.

We evaluate *MatchBox* image matching on ten challenging image datasets and illustrate results that significantly outperform standard approaches employed in frequently utilized tools like Bundler [30].

2. Related work

In computer vision we often aim at designing general approaches to tackle a specific task. This is particularly true for object detection where classification approaches are almost invariably based on SIFT [22], GIST [25] and HOG [8] features or more recently deep learning methods. According to our opinion, image matching, a task frequently employed in early stages of an application, is no exception to this principle.

But designing a visual metric to ignore small, unimportant details while being capable of focussing on the important structures that render two patches similar, remains a challenge. This work presents an attempt to follow the intuition that humans observe a novel scene by first establishing a *global correspondence* before nailing small details.

Such an approach contrasts common image matching methods which often directly focus on small local transformations from the very beginning. The two central elements for finding corresponding keypoints are the feature space and the similarity metric. Common keypoint representations are the aforementioned cues, like SIFT, GIST and HOG as well as various wavelet and gradient decompositions and combinations such as spatial pyramids [19] or bag-of-words representations [29] which capture salient structures that are however purely local.

To relax the aforementioned locality property of the considered transformations, image matching techniques were predominantly extended in two directions. Either the space of considered local transformations is modified, which influences computational efficiency and discriminative properties, or the distance metrics are adapted [23, 4, 3, 6, 10, 5]. Another line of work formalizes image matching from a data driven perspective to learn a better visual similarity. Tieu and Viola [31] use boosting to learn image specific features and Hoiem *et al*. [17] employ a Bayesian framework to find close matches. Contrasting the aforementioned work which is based on multiple training examples, Shrivastava *et al*. [28] showed how to achieve cross-domain matching using structured support vector machines learned from a single example. They illustrate impressive results across multiple domains but their approach can still deal with only minor viewpoint changes.

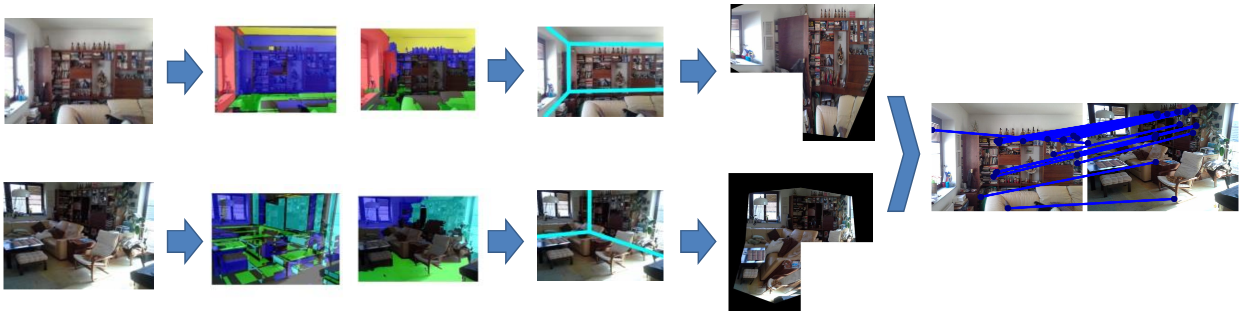
Instead of extending a standard image matching approach to deal with a larger number of local transformations we subsequently follow the physiologic intuition by first investigating an observed scene from a more global perspective. To this end, we specifically consider image matching for indoor scenes by leveraging the *Manhattan world* assumption, the restriction that scenes are commonly aligned to three dominant and orthogonal vanishing points. This assumption was already utilized for 3D indoor scene reconstruction by Furukawa *et al*. [11] for depth-map estimation. Although both [11] and MatchBox employ the Manhattan world assumption, both approaches differ in that [11] assumes a piecewise planar scene and we make use of cuboids.

Common to the most successful methods for monocular layout retrieval [14, 33, 20, 9, 26, 27] is the use of the Manhattan world property. As a consequence, a simple parameterization of the 3D layout based on four variables exists [14, 33, 20]. By exploiting the inherent decomposition of the additive energy functions with an integral geometry technique [26], globally optimal inference of frequently utilized cost functions was shown to be possible [27]. Given high quality image cues known as geometric context [16] and orientation maps [21], accuracies exceeding 85% are achieved [27] on standard datasets [14, 15].

While image matching has been extended to take into account predominantly local transformations, we present the first work for indoor structure from motion to use global scene information for rectification of local patches as discussed next. Note that rectification based on global image properties has been done for outdoor façades [7, 2].

3. Approach

Encouraged and inspired by the quality of the results obtained from monocular scene understanding algorithms, we aim at using global scene interpretation to improve keypoint matching. In the following we first present an overview of our approach and then describe the individual components (scene estimation, image rectification and image matching) in more detail.



(1) original image (2) orientation maps and geometric context (3) scene layout (4) rectification (5) tentative matches

Figure 2. The proposed *MatchBox* procedure. Given input images (1) we extract orientation maps and geometric context (2) which enable optimization to find a scene interpretation (3). The scene estimate enables rectification of the detected faces (floor, ceiling, left, right and front wall) (4) used for keypoint matching (5). This particular example could not be matched by standard approach whereas we obtained 20 tentative matches.

3.1. Overview

Consider Fig. 2 for an overview of the proposed approach. We are given a pair of images and first detect three corresponding vanishing points for each image using the algorithm from Hedau *et al.* [14]. Subsequently we minimize an energy function to retrieve a 3D parametric box that best describes the observed room layout. The energy function is based on image cues which were proven valuable for indoor scene understanding. Those are orientation maps [21] and geometric context [16] which are described in greater detail below and visualized in step 2 of Fig. 2. To minimize the cost function, we employ a variant of a branch-and-bound algorithm following [27] which is briefly discussed for completeness in Sec. 3.2. The resulting 3D scene reconstruction is visualized in step 3 of Fig. 2. Given the reconstructed scene, we rectify the prediction of the walls individually as detailed in Sec. 3.3 and depicted in step 4.

We extract standard SIFT features from the rectified floor and ceiling as well as upright SIFT from the rectified walls. Next, we match the features, transform them back to the original image and combine them with the result from standard feature matching. Afterwards, the correspondences are geometrically verified using epipolar geometry.

3.2. Scene Estimation

For the scene estimation task, let the 3D layout be referred to via variable y. Optimizing for the best 3D interpretation of the observed scene, *i.e.*, finding y^*, is commonly phrased as the general energy minimization problem

$$y^* = \arg\min_{y\in\mathcal{Y}} E(y). \qquad (1)$$

To make this general framework more concrete we subsequently first consider how to parameterize a 3D scene. Hence we answer how to describe the space of all layouts $y \in \mathcal{Y}$. In a second part we detail the involved energy function E before we afterwards discuss its optimization. Note that our exposition follows the approach described in [27] and we refer the interested reader to this work for additional details.

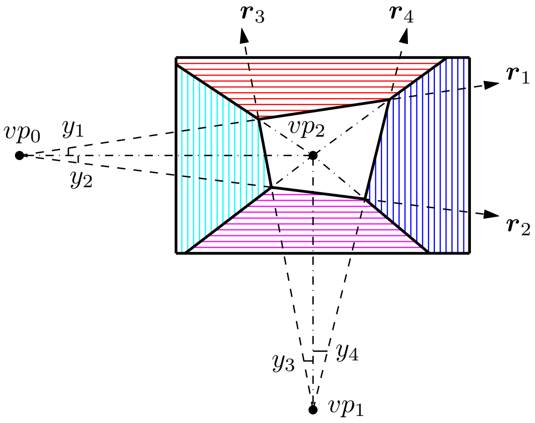


Figure 3. Parameterization of 3D layout estimation

Scene Parameterization Following standard monocular scene understanding literature [14, 33, 20, 9, 26, 27], we use the Manhattan world assumption, *i.e.*, we assume that the observed scene is described by three mutually orthogonal plane directions. Taking a single image as input we therefore detect the three dominant and orthogonal vanishing points $\{vp_0, vp_1, vp_2\}$ using the algorithm also employed by Hedau *et al.* [14]. Given those vanishing points, we parameterize the observed room as a single 3D parametric box. With at most three walls as well as floor and ceiling being observable, we follow the standard approach [14, 33] and parameterize such a box by four parameters y_i, $i \in \{1, \dots, 4\}$ each corresponding to an angle describing a ray r_i, $i \in \{1, \dots, 4\}$ as visualized in Fig. 3. Note that the rays r_1, r_2 are limited to lie either above or below the horizon which is the line that connects vp_0 and vp_2. To only parameterize valid layouts a similar constraint is employed for rays r_3 and r_4. For efficient computation we discretize the possible angles $y_i \in \mathcal{Y}_i = \{1, \ldots, |\mathcal{Y}_i|\}$ such that the space of all valid layouts $\mathcal{Y} = \prod_{i=1}^{4} \mathcal{Y}_i$ is a product space describing a countable amount of possibilities. To ensure a sufficiently dense discretization we let the number of discrete states $|\mathcal{Y}_i|$ depend on the location of the vanishing points while making sure that the area within the image domain covered by successive rays is smaller than 3000 pixel.

Energy function Having a parameterization of possible layouts at hand, we score a given layout hypothesis y using an energy function $E(y)$. We subsequently investigate the structure of the employed energy function. Let the five layout faces be subsumed within the set $\mathcal{F} = \{\text{left-wall}, \text{right-wall}, \text{front-wall}, \text{floor}, \text{ceiling}\}$. We design an energy function which decomposes into a sum of terms each depending on a single layout face, *i.e.*,

$$E(y) = \sum_{\alpha \in \mathcal{F}} E_\alpha(y_{g(\alpha)}), \quad g : \mathcal{F} \to \mathcal{P}(\{1, \ldots, 4\}) \quad (2)$$

Note that the set of variables involved in computing a face energy E_α is a subset of all variables, *i.e.*, g denotes a restriction of $(y_1, \ldots, y_4)$ to $(y_i)_{i \in g(\alpha)}$ and hence maps from a face $\alpha \in \mathcal{F}$ to a member of the powerset $\mathcal{P}(\{1, \ldots, 4\})$. Importantly it was shown in independent work by many authors [20, 9, 26] that the most promising image cues for 3D layout estimation are geometric context (GC) [16] and orientation maps (OM) [21]. We therefore let the face energy decouple according to $E_\alpha = E_{\alpha,\text{GC}} + E_{\alpha,\text{OM}}$. An example for both image cues is given in step 2 of Fig. 2. Note that their construction is different. While OMs are based on sweeping lines to obtain one of five possible wall orientations, GCs are computed using classifiers trained on a dataset provided in meticulous work by Hedau *et al.* [14].

Scene model construction Using "integral geometry," it was shown by Schwing *et al.* [26] that the energy functions $E_{\alpha,\cdot}$ decouple for every wall face α into a sum of terms with each summand depending on at most two variables. This enables efficient storage. More importantly it was further shown [27] that the geometric properties of the parameterization can be leveraged to design an efficient branch-and-bound approach which retrieves the globally optimal solution $y^* = \arg\min_{y \in \mathcal{Y}} E(y)$ of the initially given optimization problem stated in Eq. (1).

The approach proceeds by successively dividing a set of layouts $\hat{\mathcal{Y}} = \hat{\mathcal{Y}}^1 \cup \hat{\mathcal{Y}}^2$ into two disjoint sets $\hat{\mathcal{Y}}^1$, $\hat{\mathcal{Y}}^2$, *i.e.*, $\hat{\mathcal{Y}}^1 \cap \hat{\mathcal{Y}}^2 = \emptyset$. A lower bound on the energy function is computed for each set such that we know that any layout member of the set scores equally well or worse. The sets are inserted into a priority queue according to their score. We iterate by retrieving the lowest scoring set to be considered for further partitioning until such a set contains only a single element y^*. Despite having to evaluate all layout hypothesis in a worst case, it was shown in [27] that such an approach yields an efficient algorithm that partitions only a few layouts in practice.

3.3. Image rectification

Finding the optimal 3D layout by solving the problem given in Eq. (1) yields a result similar to the one visualized in step 3 of Fig. 2, which enables a 3D reconstruction of the observed scene up to scale.

To unfold the estimated and textured 3D parametric box into a 2D image, we apply a homography to each wall as well as floor and ceiling separately. Given the three vanishing points and the inference result being the four angles y_i, the four corners of the front wall are completely specified by intersecting the rays r_i. Since we only observe three walls as well as floor and ceiling, but not the closing wall of the box, the other corners are not specified uniquely but are computed such that no image region is cropped.

Hence, every wall is given by four points and we compute a projective transformation to warp each quadrilateral into a square-shaped image. The length of a side of the resulting image is given by $\min\left(\lfloor \frac{h+w}{2} \rfloor, 1600\right)$ where h and w are the height and width of the original image.

To give an example, let the four corners of the quadrilateral describing the front wall be referred to via $x_1, \ldots, x_4 \in \mathbb{R}^2$. We solve a linear system of equations to obtain the transformation matrix $T_{\text{front-wall}}$ which projects $x_1, \ldots, x_4$ to the corners of a square-shaped image. We subsequently warp the texture of the front-wall to the square-shaped image by applying a bi-linear transformation. The resulting rectification upon processing all walls, ceiling and floor is illustrated in step 4 of Fig. 2.

3.4. Image matching

Next we describe the two types of keypoint matchings that we employ before we discuss their verification.

Tentative standard matches What we subsequently refer to as "tentative standard matches" is computed following the procedure employed in Bundler [30], except that we replace the original SIFT features [22] with a publicly available SIFT feature computation and we utilize randomized kd-trees for matching instead of a single kd-tree.

First, we find keypoints and corresponding 128 dimensional SIFT descriptors on the given *original* pair of images using the library by Vedaldi and Fulkerson [32]. To construct possible matches, we find two nearest neighbors in the 128 dimensional feature space using the fast approximate nearest neighbor procedure [24]. We establish a tentative match between the feature in the first image and its closest feature in the second image if the ratio of the distance to the closest over the distance to the second closest

Set	# images	# cameras		# points		Error	
		Std	Ours	Std	Ours	Std	Ours
1	101	**100**	**100**	16118	**34567**	**0.446**	0.475
2	75	33	**37**	8620	**12020**	**0.462**	0.475
3	116	**110**	101	31358	**38923**	**0.618**	0.818
4	129	106	**116**	36628	**50266**	0.677	**0.636**
5	79	74	**78**	21949	**28305**	**0.595**	0.642
6	79	14	**55**	1209	**11988**	**0.687**	0.929
7	98	**96**	**96**	**36860**	34563	**0.748**	1.793
8	70	**70**	**70**	15170	**27183**	**0.767**	0.856
9	57	9	**23**	1369	**3142**	**0.411**	0.826
10	492	294	**329**	40688	**72336**	**0.614**	0.663

Table 1. The number of recovered cameras, the number of reconstructed 3D points and the average reprojection error obtained with the standard matches only and with our proposed approach for ten different datasets.

feature point is smaller than 0.6, *i.e.*, we only regard two keypoints as tentative matches if there is a sufficient distance between a possibly disambiguating point that might arise from repetitive structures such as windows or picture frames.

Tentative scene model matches To obtain what we refer to as "tentative scene model matches" we detect keypoints and corresponding feature vectors using [32] on the rectified images illustrated in step 4 of Fig. 2. We compute standard 128 dimensional SIFT descriptors on the floor and ceiling, and employ the more discriminative upright SIFT feature computation on the three possible walls. We use upright SIFT only on the walls since we assume their structure to be aligned with gravity.

Similar to the aforementioned standard matches and in order to filter out keypoints arising from repetitive structures, we find two nearest neighbors [24] and accept a tentative match only if the distance ratio does not exceed a threshold of 0.6. To gain computational savings, we match only ceilings with ceilings, floors with floors and walls with walls. In a last step, we transform all the tentative keypoints from the matching domain back into the original image domain using inverse mappings for every wall, *e.g.*, $T^{-1}_{\text{front-wall}}$. The resulting tentative scene model matches refer to the set of all matches, *i.e.*, matches of ceilings, floors and walls.

Verified matches In our experiments we compare the standard matching with our proposed approach combining both, the matches obtained with the standard approach augmented by the matches obtained with the scene model approach. Even if a scene model is estimated incorrectly, we have at least as many matches as the standard procedure.

To filter outliers from both sets, we let Bundler [30] verify them. First, the epipolar geometry is estimated using an eight-point algorithm [12] inside a "vanilla RANSAC." A tentative match is defined as an outlier if the residual for the generated fundamental matrix fails to lie within a predefined threshold of 9 pixels. A total of 2048 epipolar geometries are investigated. If the best epipolar geometry is supported by 16 or more matches then the inlier matches are retained. Otherwise, the hypothesis is rejected as unreliable and the set of verified matches remains empty.

4. Experiments

In the following, we evaluate our proposed approach on 10 challenging datasets, each visualizing a particular indoor scene. These datasets contain 101, 75, 116, 129, 79, 79, 98, 70, 57 and 492 images each, as summarized in Tab. 1. Every dataset visualizes an indoor scene showing living rooms in set 1 and 5, a bathroom in 2, kitchens in 3 and 8, general rooms in 4 and 6, a library in 7, an empty room with white walls in 9, and a whole floor of a gallery in 10. The dataset referred to by 10 is obtained from Furukawa *et al.* [11].

4.1. Quantitative Evaluation

We show results of our approach combining the standard matches and the scene model matches (denoted by *Ours*) in comparison to results of the standard method (denoted by *Std*) which utilizes only the standard matches.

In Tab. 1 we compare the 3D reconstruction results obtained with the standard approach and with our *MatchBox* algorithm. Reconstruction is carried out by Bundler [30]. We show the number of cameras that were recovered, the number of 3D points that could be reconstructed and additionally we provide the average reprojection error. *MatchBox* is able to reconstruct more cameras on six out of ten datasets while we are on par with the standard approach for three sets of images. Considering the number of reconstructed 3D points as our score we want to maximize we are able to improve over Bundler on nine image sets. On the other hand, the average reprojection error is better only for one of the ten datasets. This is connected to the fact that the final optimization contains more cameras and/or more 3D points. Note that worse reprojection error does not mean

Set	# matches				# pairs				Graph diameter	
	Tentative		Verified		Tentative		Verified			
	Std	Ours	Std	Ours	Std	Ours	Std	Ours	Std	Ours
1	102	**166**	108	**178**	890	**1120**	805	**981**	7	**5**
2	121	**130**	134	**161**	405	**510**	317	**339**	12	**10**
3	**144**	139	160	**227**	972	**1709**	776	**849**	7	**6**
4	**178**	**178**	234	**280**	1238	**1641**	841	**890**	**8**	**8**
5	197	**199**	213	**249**	550	**760**	469	**527**	6	**5**
6	94	**121**	100	**129**	279	**313**	231	**246**	9, **2**	**6, 2**
7	272	**361**	349	**537**	1846	**2276**	1286	**1320**	**6**	**6**
8	188	**284**	189	**288**	1183	**1278**	1022	**1070**	5	**4**
9	78	**121**	84	**120**	257	**271**	160	**172**	(4 + 2), (4 + 5), **1**	**6, 9, 1**
10	95	**137**	97	**138**	6415	**6968**	5264	**5835**	21	**19**

Table 2. The average number of tentative and verified matches and the total number of matched image pairs before and after verification. In addition and separated by ',' we show the graph diameter for all connected components of size larger than two obtained with our approach. If such a component is disconnected when utilizing the standard, we show diameters of all the sub-components that got connected in parenthesis, separated by '+'.

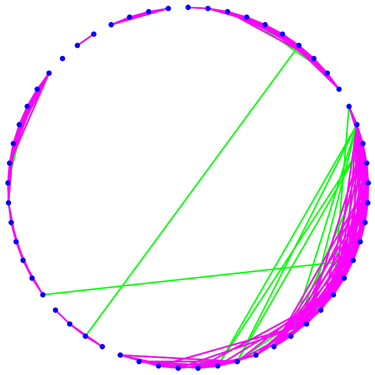


Figure 4. Nodes represent images of the dataset 9. An image pair is connected by an edge if it is matched and verified by the standard approach (magenta) or our proposed method (green+magenta). We observe how *MatchBox* is able to reduce the number of connected components obtained with standard keypoint matching.

worse true reconstruction error.

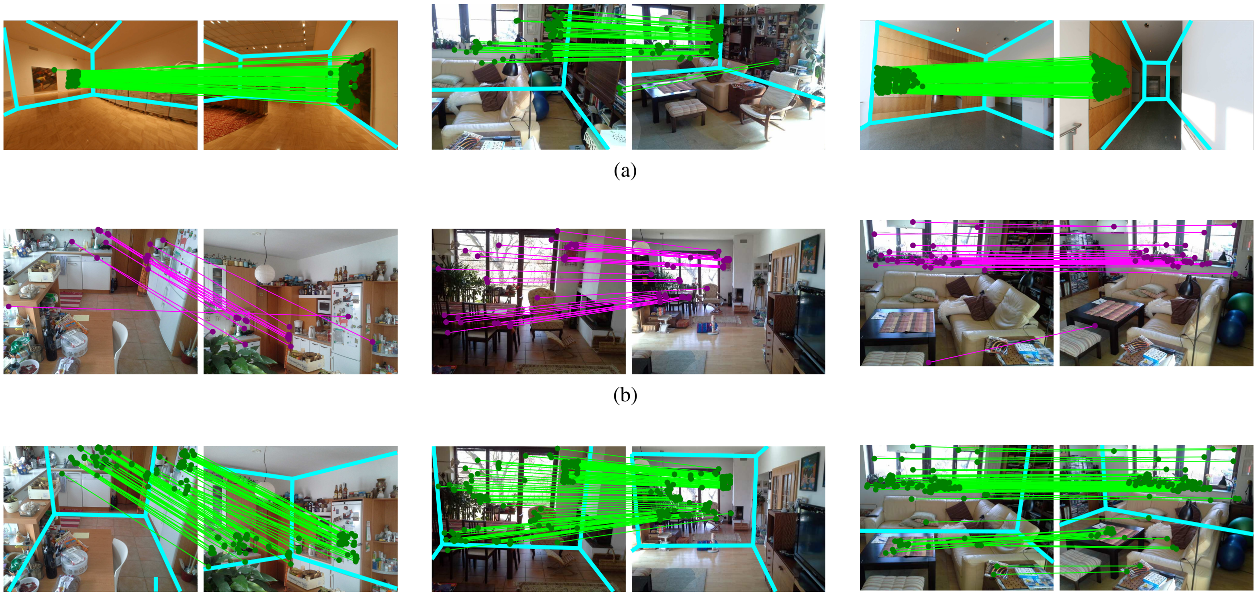
For two datasets the additional features surprisingly result in worse reconstruction. Less cameras were recovered for the image set 3 and the reprojection error of the reconstructed dataset 7 is too high. This is due to the fact that our approach matches similar structures just like the standard matching. Therefore, if a scene contains repetitive structures, which can be falsely matched, our method can increase also the number of false matches and not just correct ones. Since the eight-point algorithm [12] is employed for verification, false matches might not get rejected if they lie on a plane which is often the case for indoor scenes. Higher contamination with false matches can therefore cause worse reconstruction. To fully exploit the benefit of our contribution in structure from motion pipelines, this problem has to be addressed.

Next we aim to evaluate the merit of incorporating a more global scene representation. Therefore we consider a graph where vertices represent images and two vertices are connected via an edge if we found matching keypoints on both image pairs after verification. See Fig. 4 for a visualization of such a graph where nodes are ordered according to the connected components. If a graph consists of more than one connected component, 3D reconstruction will omit some images. Hence we want to minimize the total number of connected components. In Fig. 4, green edges illustrate the *additional* pairs matched by our approach. Note that our method takes advantage of both, standard and scene matches. Therefore, we observe how the additional matches successfully connect components for dataset 9 visualized in Fig. 4.

Not only are we interested in minimizing the number of connected components but we are also aiming for many connections between images. In Tab. 2, we compare standard matching to our approach using the average number of matches between image pairs as well as the total number of matched image pairs. For an additional error metric capturing the connectedness of the matched image pairs, we propose to utilize the graph diameter of a matching, *i.e.*, the maximum length of all the shortest paths. Hence the graph diameter measures how tightly we are able to connect the different images. We provide the diameter for all datasets in Tab. 2 using the connected components of size larger than two found with our approach. We observe that *MatchBox* performs very well in all the metrics.

4.2. Qualitative Evaluation

Having shown that our proposed approach outperforms a standard matching algorithm on a set of ten indoor scene datasets for various error metrics we next provide some typ-



(a)

(b)

(c)

Figure 5. (a) visualizes verified matches of our approach. The standard method would not retrieve any matches while we match 105, 65 and 221 keypoints respectively. Cyan color depicts detected scene interpretation. (b) and (c) visualize verified matches found on the same image pairs except that (b) illustrates results obtained by standard matching and (c) provides our *MatchBox* performance. We observe that our approach exploits knowledge of the scene and matches more keypoints (standard approach was able to match 20, 30 and 28 keypoints and we obtained 105, 121 and 98). The original images on the left and on the right in (a) are from Furukawa *et al*. [11].

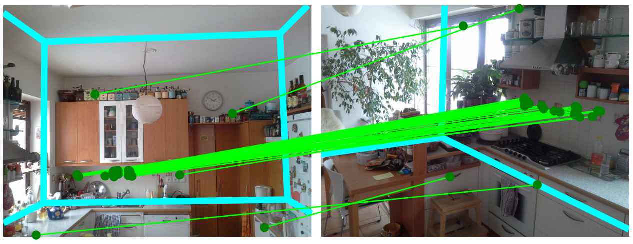


Figure 6. Visualization of an incorrectly matched image pair. The color green visualizes verified matches obtained by our approach and the color cyan shows detected scene layout.

ical examples.

In Fig. 5(a) we visualize verified matches obtained with our proposed approach. The standard matching, did not retrieve any matching keypoints. For completeness we overlay the image with the predicted room layout.

In Fig. 5(b) we show verified matches obtained with the standard keypoint matching and compare them to our proposed approach with results visualized in Fig. 5(c). Again, we observe significantly more matches.

In Fig. 6 we show a wrongly matched image pair. The image pair was incorrectly matched since tiles had the same pattern but they were situated on different walls.

Additionally, we visualize a comparison of sparse 3D reconstructions in Fig. 7.

5. Conclusion

In this work we proposed to employ monocular scene understanding to improve image matching for indoor environments. Importantly, we suggested to follow the physiological concept of first reconstructing a global scene representation before matching salient details via commonly utilized image features that focus on local transformations. We showed that our proposed approach outperforms standard keypoint matching on challenging datasets illustrating various indoor scenes.

Acknowledgments The authors were supported by The under Project SGS12/191/OHK3/3T/13 and by The Technology Agency of the Czech Republic under Project TA02011275. A. G. Schwing was supported by ONR Grant N00014-14-1-0232.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski. Building rome in a day. In *Proc. ICCV*, 2009. 1

[2] G. Baatz, K. Köser, D. M. Chen, R. Grzeszczuk, and M. Pollefeys. Handling urban location recognition as a 2d homothetic problem. In *Proc. ECCV*, 2010. 2

[3] A. Baumberg. Reliable Feature Matching across widely separated views. In *Proc. CVPR*, 2000. 2

[4] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool. SURF: Speeded Up Robust Features. *CVIU*, 2008. 2

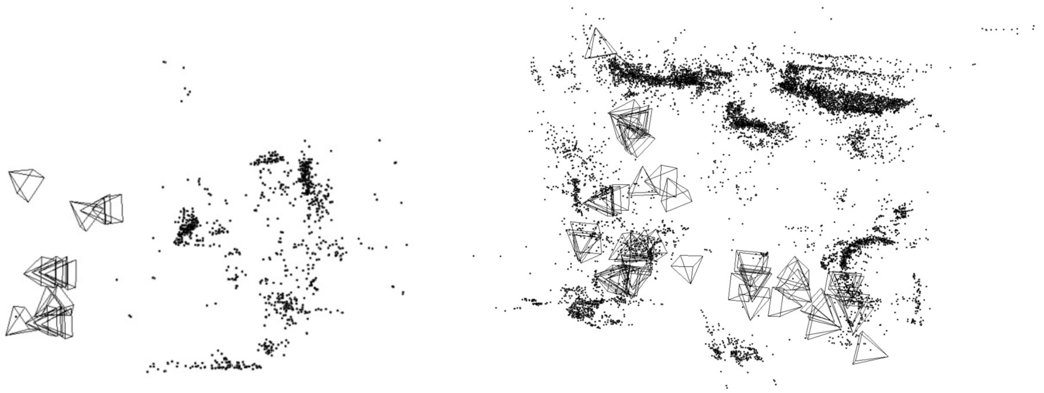


Figure 7. The top view of reconstructions of the dataset 6. On the left, we see a result of a reconstruction procedure employing standard matches and on the right *MatchBox* achievement. The square pyramids represent cameras. A camera center is in a top of its pyramid and an orientation of the camera is given by a base of its pyramid.

[5] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proc. CVPR*, 2005. 2

[6] G. Carneiro and A. D. Jepson. Multi-scale Phase-based local Features. In *Proc. CVPR*, 2003. 2

[7] R. Cipolla, R. Duncan, and B. Tordoff. Image-based localisation. In *Proc. Virt. Sys. and Mult.*, 2004. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005. 1, 2

[9] L. del Pero, J. Guan, E. Brau, J. Schlecht, and K. Barnard. Sampling Bedrooms. In *Proc. CVPR*, 2011. 2, 3, 4

[10] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *PAMI*, 1991. 2

[11] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing building interiors from images. In *Proc. ICCV*, 2009. http://grail.cs.washington.edu/projects/interior/. 1, 2, 5, 7

[12] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge, 2nd edition, 2004. 5, 6

[13] J. Hays and A. A. Efros. Scene Completion using Millions of Photographs. *Siggraph*, 2007. 1

[14] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the Spatial Layout of Cluttered Rooms. In *Proc. ICCV*, 2009. 2, 3, 4

[15] V. Hedau, D. Hoiem, and D. Forsyth. Thinking Inside the Box: Using Appearance Models and Context Based on Room Geometry. In *Proc. ECCV*, 2010. 2

[16] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 2007. 2, 3, 4

[17] D. Hoiem, R. Sukthankar, H. Schneiderman, and L. Huston. Object-based image retrieval using the statistical structure of images. In *Proc. CVPR*, 2004. 2

[18] B. Klingner, D. Martin, and J. Roseborough. Street view motion-from-structure-from-motion. In *Proc. ICCV*, 2013. 1

[19] S. Lazebnik, C. Schmid, and J. Ponce. Spatial pyramid matching. *Object Categorization: Computer and Human Vision Perspectives*, 2009. 2

[20] D. C. Lee, A. Gupta, M. Hebert, and T. Kanade. Estimating Spatial Layout of Rooms using Volumetric Reasoning about Objects and Surfaces. In *Proc. NIPS*, 2010. 2, 3, 4

[21] D. C. Lee, M. Hebert, and T. Kanade. Geometric Reasoning for Single Image Structure Recovery. In *Proc. CVPR*, 2009. 2, 3, 4

[22] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 1, 2, 4

[23] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. *PAMI*, 2005. 2

[24] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, 2009. 4, 5

[25] A. Olivia and A. Torralba. Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, 2006. 1, 2

[26] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient Structured Prediction for 3D Indoor Scene Understanding. In *Proc. CVPR*, 2012. 2, 3, 4

[27] A. G. Schwing and R. Urtasun. Efficient Exact Inference for 3D Indoor Scene Understanding. In *Proc. ECCV*, 2012. 2, 3, 4

[28] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. *Siggraph Asia*, 2011. 1, 2

[29] J. Sivic and A. Zisserman. Video google: A thext retrieval approach to object matching in videos. In *Proc. ICCV*, 2003. 2

[30] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. In *SIGGRAPH*, 2006. 2, 4, 5

[31] K. Tieu and P. Viola. Boosting image retrieval. *IJCV*, 2004. 2

[32] A. Vedaldi and B. Fulkerson. VLFeat: An Open and Portable Library of Computer Vision Algorithms. http://www.vlfeat.org/, 2008. 4, 5

[33] H. Wang, S. Gould, and D. Koller. Discriminative Learning with Latent Variables for Cluttered Indoor Scene Understanding. In *Proc. ECCV*, 2010. 2, 3

[34] Y. Wexler, E. Shechtman, and M. Irani. Space-Time Completion of Video. *PAMI*, 2007. 1

[35] O. Whyte, J. Sivic, and A. Zisserman. Get out of my picture! Internet-based Inpainting. In *Proc. BMVC*, 2009. 1