REGION-BASED EPIPOLAR AND PLANAR GEOMETRY ESTIMATION IN LOW-TEXTURED ENVIRONMENTS

Vincent Gaudillière, Gilles Simon, Marie-Odile Berger

Inria Nancy - Grand Est / Loria, Nancy, France

ABSTRACT

Given two views of the same scene, usual correspondence geometry estimation techniques exploit the well-established effectiveness of keypoint descriptors. However, such features have a hard time in poorly textured man-made environments, possibly containing repetitive patterns and/or specularities, such as industrial places. In that paper, we propose a novel method for two-view epipolar and planar geometry estimation that first aims at detecting and matching physical vertical planes frequently present in these environments, before estimating corresponding homographies. Inferred local correspondences are finally used to improve fundamental matrix estimation. The gain in precision is demonstrated on industrial and urban environments.

***Index Terms*—** Epipolar geometry estimation, homography estimation, plane detection, low-textured industrial environments.

1. INTRODUCTION

The problem of two-view correspondence geometry estimation has been extensively studied over the past decades. Most commonly used techniques consist in matching previously extracted keypoint descriptors [1]. Keypoints extraction and description has known its golden age as soon as SIFT was introduced [2]. Since then, a great deal of research has focused on designing faster methods [3, 4, 5], or more invariant ones [6]. However, those methods work well with highly textured scenes, but fail with poorly textured ones. Moreover, the Lowe's ratio most often used to get potential matches [2] leads to ignoring repetitive features in the initial matches set.

Therefore, line segments may appear as suitable features for complex environments such as industrial ones, even if line matching remains a very tough issue. Several types of matching strategies have been investigated in the past, and most of them consist in iteratively estimating a geometric model among a set of putative matches (RANSAC-like algorithms). Potential segment matches can be determined from the appearance similarity of their neighborhoods [7, 8, 9]. However, line segments in industrial environments often correspond to 3D object edges, i.e. their neighborhoods consist of an information-poor content (e.g. solid color) on one side, and a viewpoint-dependant content on the other side. The same argument can be opposed to segments described by keypoints in their neighborhoods [10, 11]. Conversely, one can create the initial set of matches thanks to geometric criteria [12, 13, 14], yet such invariant-based methods are highly sensitive to noise, and often lead to very combinatorial problems. Finally, matching groups of lines can compensate for the lack of information, but here again at the cost of a high combinatorics [15, 16].

In our work, we take advantage of global properties of the environment to guide matching:

- First, industrial environments often contain a substantial number of vertical planes. Starting from the observation that information about existing planes is somehow encapsulated into vanishing points, the way our method takes advantage of them is two-fold. First, we associate each line segment to its vanishing point, and that information is then used to constrain segment matching. Second, vanishing points are used as primitives for local homography estimation, thereby reducing the combinatorial complexity and avoiding to place excessive reliance on sparse visual keypoints.
- Second, it appears in [17] that image regions are robust features to match, even in the presence of severe viewpoint and condition changes. Indeed, appropriately-designed descriptors of subimages can offer the advantage of being robust to both viewpoint changes (as local descriptors) and condition changes (as global descriptors) [18]. In our method, subimage correspondences are used as prior for vertical planes detection and matching. By doing that, line matching is more constrained, thereby circumventing inherent difficulties of industrial environments, without compromising the processing time.

Given two different views of the scene, our processing pipeline is finally broken down as follows: (1) **Region correspondences** are computed between images (Section 2). (2) **Local homographies** are detected between matched regions, then merged to identify physical planes (Section 3). (3) Segment correspondences defined by local homographies are used to improve **epipolar geometry** estimation (Section 4).

This work was funded by the DGA/DGE RAPID EVORA project.

2. REGION PROPOSAL AND MATCHING

Our method for generating a set of hypotheses of matched regions is inspired from the work of Sünderhauf *et al.* [17]. In their method, subimage correspondences are computed to robustly recognize places. To do that, object proposals are extracted, before mid-level ConvNet features are computed for each of them. Thus, an image is described by a collection of CNN descriptors corresponding to different subimages. After a dimensionality reduction step, matching of descriptors is performed using a nearest neighbor search based on the cosine distance between them, followed by a cross-check ensuring that only mutual matches are accepted.

In our method, we broadly follow the same approach, except for the fact that we directly consider high-level ConvNet features (*e.g.* the last layer before fully-connected ones) from [19] as region descriptors, enabling us to do without the dimensionality reduction step. *Edge Boxes* [20] is used as object proposal, since it appears to be particularly suitable for scenes made of contours, such as in industrial environments.

3. LOCAL PLANAR HOMOGRAPHIES ESTIMATION

Once subimage correspondences have been computed, our algorithm aims at detecting local homographies between matched regions. To ensure efficient estimations, we have developed a dedicated RANSAC framework in which model hypotheses are first generated based on vanishing point and visual keypoint correspondences, and then validated on keypoints and line segments. That scheme enables us to circumvent problems encountered in poorly-textured images (sparsity of visual keypoints and difficulties to match segments) while taking advantage of the abundance of segments and vanishing points characteristic of industrial environments.

3.1. Features extraction and matching

Vanishing points (*VPs*) are detected in both images using [21]. This method detects the zenith and all existing horizontal VPs, therefore associated to vertical planes. Zeniths are easily identified, and thus matched, by selecting in each image the VP with the greatest absolute coordinate along the vertical axis.

Line segments are extracted with LSD [22]. Segments associated to previously computed VPs (according to a threshold on the angular distance between the line the segment belongs to and the line connecting the VP to the middle of the segment) are retained, while the others are discarded.

Visual keypoints (*KPs*) are finally detected with a learned variant of SIFT, called LIFT [6]. Correspondences are obtained by applying the Lowe's ratio on both directions, then keeping only mutual matches in order to maximize the inlier rate within the initial set of putative matches S_0.

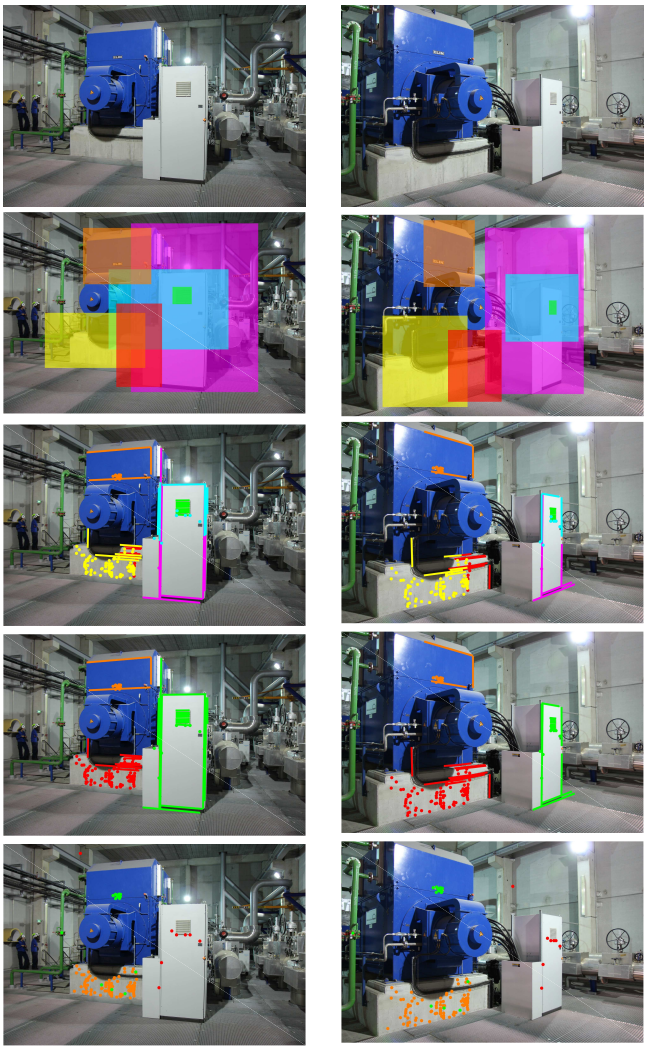

Fig. 1. Overview of our homographies estimation method. Row 1: Two views of an industrial environment. Row 2: Examples of matched regions (one color per match). Row 3: Inliers (points and segments) from homographies estimated between matched regions (one color per homography). Row 4: Inliers after merging homographies. Row 5: Inliers of the first three planes obtained by 4-points multi-RANSAC.

3.2. Homographies estimation

Selection of matched boxes Due to the specificities of our environments, regions matched during the first stage can be considered as vertical plane proposals. Since those correspondences have an important overlap, testing all pairs of boxes to find homographies would unnecessarily increase the processing time. To handle that issue, pairs of boxes are roughly ranked from smallest ones to biggest ones, according to the mean area of boxes. The key idea is to process pairs of boxes in that order, while updating two maps of already visited pixels. These maps have the same resolution as the original im-

ages, and each pixel takes the value 1 if a box containing that pixel has already been tested, 0 otherwise. A pair of boxes is tested if less than 50% of the pixels of each box have already been visited.

Feature selection The processing of one pair of boxes first consists in selecting features it contains (KPs, segments, VPs associated to the segments), and then in applying a merging procedure on line segments in order to offset over-segmentation induced by LSD. Among segments that share the same VP, those to merge are determined using the same kind of tangential and normal constraints as those presented in [14]. If segments are sufficiently aligned, they are thus merged. Finally, only the longest segments associated to each VP are retained, in order to keep the most significant ones and reduce the combinatorial complexity of matching.

Robust homography estimation Due to the difficulties to generate correct segment correspondences, RANSAC hypotheses are only generated from VPs and KPs. At each iteration, we randomly pick one pair of horizontal VPs, two matched pairs of KPs, and the pair of zeniths. These four pairs of points thus define an homography induced by a vertical plane. Unlike [9], VPs are directly used here as primitives to define the models. That choice enables us to drastically reduce the combinatorics and the reliance on visual KPs, by limiting to just two the number of KP matches needed to generate an homography model.

To determine the number of RANSAC iterations, we use a customized version of the adaptive algorithm described in [1] (Section 4.7.1). Given N_{total}^{hvp} the total number of possible pairs of horizontal VPs, the probability to pick the correct pair at any iteration is $w_{hvp} = 1/N_{total}^{hvp}$. Given N_{total}^{p} the number of putative pairs of KPs, the probability to pick a correct pair is $w_p = N_{inliers}^{p}/N_{total}^{p}$, where $N_{inliers}^{p}$ is the number of point inliers in the current iteration. The number N_{iter} of iterations required to ensure, with a probability p, that at least one set of primitives is free from outliers is thus given by the formula:

$$N_{iter} = \frac{log(1-p)}{log(1 - w_{hvp}.w_p^2)} \quad (1)$$

At each iteration, N_{iter} is updated based on the current value of w_p, and the algorithm stops if the current iteration is greater than N_{iter}.

Note that when only one type of features is used, N_{iter} is computed as: $N_{iter} = \frac{log(1-p)}{log(1-w^4)}$. If we assume a configuration with 2 horizontal VPs per image, 40% of point inliers, and 1% of segment inliers (non-matched segments), our method requires 113 iterations, while the 4-point algorithm requires 178 iterations, and the 4-segment method requires 4.6×10^8 iterations. If the considered box pair is not correct, the gain in performance between our method and the 4-point algorithm is much more important.

Model validation (RANSAC) Validation is based on KPs and segments. KP inliers are determined in the classic way, whereas all possible pairs of segments (between segments associated to the zeniths on the one hand, and between segments associated to the picked horizontal VPs on the other hand) are tested. The pairs that satisfy both tangential and normal constraints presented in [14] are retained. To prevent from degenerate cases, one segment in the original image can be matched with only one segment in the target image (the closest in the sense of the normal constraint). From there, the number of target segments involved in a pair that satisfy the homography ($N_{inliers}^{s}$) is added to the number of keypoint inliers, thus defining a score for the model H: $Score(H) = N_{inliers}^{p} + N_{inliers}^{s}$. Since only the most significant segments have been retained, segment matches are likely to be of equal importance. Moreover, adding both KPs and segments contribution is justified by the fact that it allows us to rely on any kind of features when the other is scarce.

Final decision Subimages matched during the first stage may not be linked by an homography, either because they do not contain a planar object, or because the match is incorrect. Thus, to distinguish between real and fake homographies, we only consider homographies producing more than 10 inliers.

3.3. Homographies merging

At this point, several local homographies can identify the same vertical plane (see Fig. 1: Row 3). We thus apply a merging step, which consists, for each homography, in testing the inliers of other homographies. If more than 50% of the KP inliers of an homography A are also inliers of an homography B, and if the same performance is achieved with the segment inliers, then A and B are merged.

4. EPIPOLAR GEOMETRY ESTIMATION

As the consensus set of homographies contain pairs of segments, we now use them to improve epipolar geometry estimation. The main idea, inspired by [8], is to add segment intersections to the set of putative KP matches S_0, and then to give the new set of matches as input to a state-of-the-art epipolar geometry estimation algorithm, called ORSA [23].

In our method, segments are first converted into lines in both images. Then, for each consensus set, intersections between vertical lines (those generated from segments associated with the zenith) and horizontal ones (those generated from segments associated with the horizontal VP) are computed. Since several segments can generate approximatively the same line (due to the merging step), and since near-identical duplicate matches can disrupt the epipolar geometry estimation algorithm, we have decided to divide both images according to a regular grid, and to keep at most one intersection point per square of the grid. Finally, these new intersection point matches are added to the initial set of KP matches S_0.

	Avg. nb of inliers		Avg. inlier rate (%)	
	indus.	urban	indus.	urban
SIFT	198.6	102.8	30.02	52.3
LIFT	207.4	100.3	42.24	58.4
Ours	**260.7**	**126.0**	**45.84**	**62.44**

Table 1. Average performance on test datasets, in terms of both number of inliers and inlier rate.

5. RESULTS AND CONCLUSION

Experiments have been conducted on two datasets. The first one consists of 46 pairs of images taken in an electricity plant, and the second one of 14 pairs of urban images extracted from [24]. For each pair, the ground truth consists of approximately 20 pairs of points that have been placed by hand. These points have been selected with the aim of covering the images as widely as possible, while ensuring an homogeneous distribution. Industrial images have a resolution of 1280×1920 pixels, whereas urban ones are 640×640. It must be noted that [9] fails to detect reliable segment correspondences in the considered examples.

In the following, our method is compared to the standard method composed of two steps (i) extraction and matching of keypoints on the whole image by applying the Lowe's ratio on both directions then keeping only mutual matches (ii) computation of the epipolar geometry from this set of matched points using ORSA. SIFT and LIFT are used in our comparison. Estimated epipolar geometries are finally compared in terms of inliers (numbers and rates), and precision (error w.r.t ground truth).

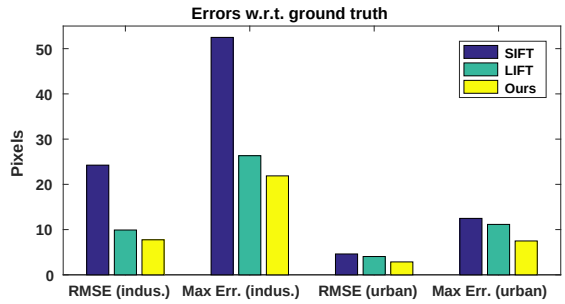

Fig. 2. Average errors on both industrial and urban datasets (referred as *indus.* and *urban*), w.r.t. ground truth.

The precision of estimated fundamental matrices is assessed in Fig. 2. For each image pair, the difference between ground truth points and estimated epipolar lines have been measured through RMSE and maximum error. These criteria have then been averaged over the datasets. In average, our method presents the highest precision, whichever error criterion is used.

Table 1 presents the quality of estimated models in terms of inliers (here again, results are averaged over the datasets). Our method has the highest number of inliers and the best inlier rate, which could substantially improve the quality of further reconstruction or pose computation steps. Fig. 3 illustrates the higher number of inliers obtained with our method. In addition, it is worth noting that there are 6 pairs of industrial images on which ORSA fails at estimating a model based on classical SIFT correspondences.

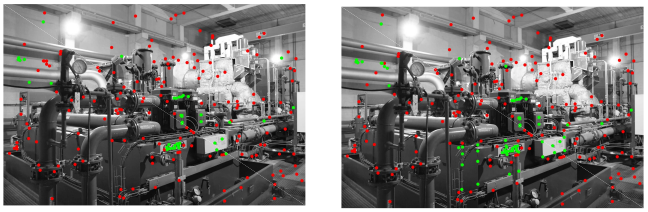

Fig. 3. Example of test images on which ORSA inliers (in green) are more numerous among our matches (right) rather than among LIFT ones (left).

Fig. 1 shows the ability of our method to detect physical vertical planes (Row 4), while a classical multi-RANSAC based on keypoint matches (Row 5) can merge points from different planes (see green and orange inliers), or find inliers out of any physical plane (see red and green dots). Fig. 4 shows other examples of planes detected by our method (convex hulls of point and segment inliers, after homographies merging step), in both industrial and urban environments.

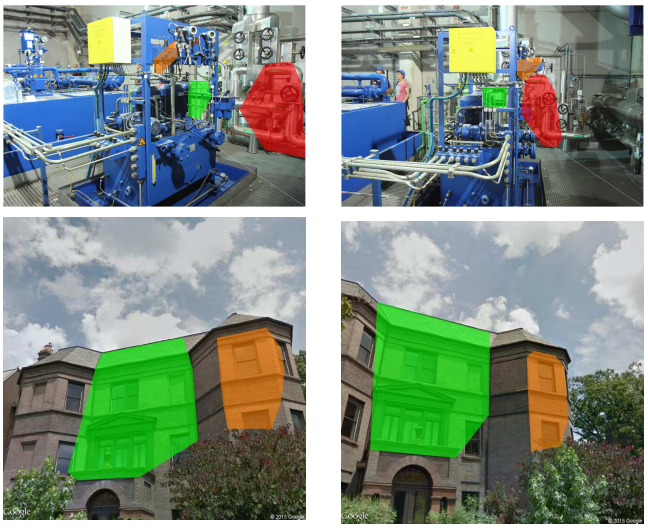

Fig. 4. Convex hulls of homographies inliers (points and segments), after merging step, on two pairs of test images (one color per homography).

These experiments demonstrate that our method can improve correspondence geometry estimation in complex multiplanar environments, even in the absence of rich visual content.

6. REFERENCES

[1] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, June 2008.

[4] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European Conference on Computer Vision*. Springer, 2006, pp. 430–443.

[5] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a Local Binary Descriptor Very Fast," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2012.

[6] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned Invariant Feature Transform," in *Proceedings of the European Conference on Computer Vision*, 2016.

[7] L. Zhang and R. Koch, "An efficient and robust line segment matching approach based on {LBD} descriptor and pairwise geometric consistency," *Journal of Visual Communication and Image Representation*, vol. 24, no. 7, pp. 794 – 805, 2013.

[8] H. Bay, V. Ferrari, and L. Van Gool, "Wide-baseline stereo matching with line segments," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2005, vol. 1, pp. 329 – 336.

[9] Y. Salaün, R. Marlet, and P. Monasse, "Robust and Accurate Line- and/or Point-Based Pose Estimation without Manhattan Assumptions," in *European Conference on Computer Vision*. Springer, 2016, pp. 801 – 818.

[10] B. Fan, F. Wu, and Z. Hu, "Line matching leveraged by point correspondences," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 390–397, 2010.

[11] H. Li, D. Song, Y. Lu, and J. Liu, "A two-view based multilayer feature graph for robot navigation," in *2012 IEEE International Conference on Robotics and Automation*, May 2012, pp. 3580–3587.

[12] T. Kim, J. Woo, and I.-S. Kweon, "Probabilistic matching of lines for their homography," *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 3489–3492, 2009.

[13] B. Fan, F. Wu, and Z. Hu, "Robust line matching through linepoint invariants," *Pattern Recognition*, vol. 45, no. 2, pp. 794 – 805, 2012.

[14] Q. Jia, X. Gao, X. Fan, Z. Luo, H. Li, and Z. Chen, "Novel coplanar line-points invariants for robust line matching across views," in *European Conference on Computer Vision*. Springer, 2016, pp. 599–611.

[15] L. Wang, U. Neumann, and S. You, "Wide-baseline image matching using line signatures.," in *ICCV*. 2009, pp. 1311–1318, IEEE Computer Society.

[16] J. López, R. Santos, X. R. Fdez-Vidal, and X. M. Pardo, "Two-view line matching algorithm based on context and appearance in low-textured images," *Pattern Recogn.*, vol. 48, no. 7, pp. 2164–2184, July 2015.

[17] N. Sunderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Robotics: Science and Systems*, Auditorium Antonianum, Rome, July 2015.

[18] S. M. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. D. Cox, P. I. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, 2012, NIPS'12, pp. 1097–1105.

[20] L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *ECCV*. September 2014, European Conference on Computer Vision.

[21] G. Simon, A. Fond, and M.-O. Berger, "A Simple and Effective Method to Detect Orthogonal Vanishing Points in Uncalibrated Images of Man-Made Environments," in *Eurographics 2016*, Lisbon, Portugal, May 2016.

[22] R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "LSD: a Line Segment Detector," *Image Processing On Line*, vol. 2, pp. 35–55, 2012.

[23] L. Moisan, P. Moulon, and P. Monasse, "Fundamental Matrix of a Stereo Pair, with A Contrario Elimination of Outliers," *Image Processing On Line*, vol. 6, pp. 89–113, 2016.

[24] A. R. Zamir, T. Wekel, P. Agrawal, C. Wei, J. Malik, and S. Savarese, "Generic 3d representation via pose estimation and matching," in *European Conference on Computer Vision*. Springer, 2016, pp. 535–553.