Grouping of co-planar local features

Katarina Mele and Jasna Maver
University of Ljubljana, Faculty of Computer and Information Science
Tržaška 25, 1001 Ljubljana
e-mail: katarina.mele@fri.uni-lj.si, jasna.maver@ff.uni-lj.si

Abstract

In this work an adaptive method for accurate and robust grouping of local features belonging to planes of interior scenes and object planar surfaces is presented. For arbitrary set of images acquired from different views, the method organizes a huge number of local SIFT features to fill the gap between low-level vision (front end) and high level vision, i.e., domain specific reasoning about geometric structures. The proposed method consists of three steps: exploration, selection, and merging with verification. The exploration is a data driven technique that proposes a set of hypothesis clusters. To select the final hypotheses a matrix of preferences is introduced. It evaluates each of the hypothesis in terms of number of features, error of transformation, and feature duplications and is applied in quadratic form in the process of maximization. Then, merging process combines the information from multiple views to reduce the redundancy and to enrich the selected representations. As demonstrated by experimental results, the proposed method is an example of unsupervised learning of planar parts of the scene and objects with planar surfaces.

1 Introduction

The use of local features is becoming increasingly popular for solving different vision tasks. Recently SIFT descriptor has been proposed for describing distinctive scale-invariant features in images [4]. SIFT features can be used to perform reliable matching between different images of an object or scene. The invariance to image translation, scaling, and rotation makes them appropriate for stereo matching, tracking applications and also suitable for mobile robot localization. SIFT features are good natural visual landmarks appropriate for tracking over a long period of time from different views, e.g., in [5] authors propose to use SIFT features for building 3D maps. It has also been demonstrated that SIFT features are appropriate for recognizing general object classes [3].

In this work we present a method for accurate and robust grouping of local features belonging to planes of interior scenes such as walls, floor, and planar surfaces of objects. For arbitrary set of images acquired from different views, the method organizes a huge number of local SIFT features to fill the gap between low-level vision (front end), i.e. outputs of various filtering operations and high level vision, i.e., domain specific reasoning about geometric

structures. The proposed method consists of three steps: exploration, selection, and merging with verification. The exploration is a data driven technique that proposes a set of hypothesis clusters. To select the final hypotheses a matrix of preferences is introduced. It evaluates each of the hypothesis in terms of number of features, error-of-transformation, and feature duplications and is applied in quadratic form in the process of maximization. Since the set of local features vary from view to view, the goal of the merging process is to combine the information from multiple views to reduce the redundancy and to enrich the selected representations. As demonstrated by experimental results, the proposed method is an example of unsupervised learning of planar parts of the scene and objects with planar surfaces.

2 Step 1: Exploration

Given a set of descriptors of local patches of interior scene the goal is to group them in clusters in accordance with some geometric property or a model. Here we examine the planar surfaces.

Let us assume that we have a set of images $\mathcal{I} = \{I_1, I_2, ...I_N\}$ of particular interior scene. The first step of our approach is the detection of DoG points and the computation of SIFT descriptor for each local region [4] (Figure 1). Next, for each pair of images, $\{(I_i, I_j)|i < j, i = 1, \ldots, N-1,\ j = 2, \ldots, N\}$, a set of matching features is determined. The matches are obtained on the basis of Euclidean distance between SIFT descriptors. Each SIFT feature in image I_i is compared to all SIFT features in image I_j. The feature has a match, if the Euclidean distance to the closest SIFT feature is at least 4 times shorter than the Euclidean distance to the next closest SIFT feature. Let $\mathcal{S}_{ij}$ denote a set of SIFT features of I_i having a match in I_j (Figure 2).

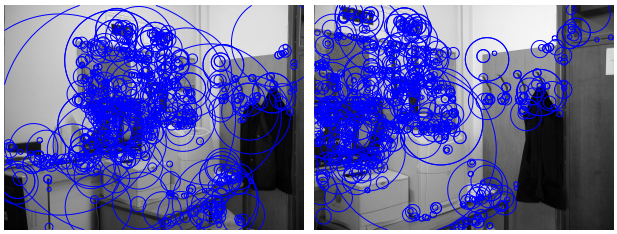


Figure 1: Illustration of feature extraction. Each circle corresponds to one DoG point. The circle defines the size of local region described by SIFT descriptor.

Figure 2: The best matches between two images.

Now, the task is to find in $\mathcal{S}_{ij}$ the features that belong to planar parts of the scene and to group them in accordance with the plane they belong to. For this purpose we apply a plane to plane homography [2]. The computation of the plane to plane homography requires at least four features in two images of the same plane. For a larger set of points the system is over determined and the plane to plane homography is estimated by homogeneous estimation method. A reliable solution, imposes to start the process of plane searching with a large set of small SIFT feature clusters, i.e., initial hypotheses. The features of $\mathcal{S}_{ij}$, here represented by their coordinates, $\{f_i^t; f_i^t = (x_i^t, y_i^t), t = 1, 2, \ldots, |\mathcal{S}_{ij}|\}$, are clustered by the k-mean clustering algorithm. The algorithm is performed several times, each time starting with different arbitrarily initial set of cluster centers. The value k denotes the number of clusters obtained by one iteration and depends on the number of features $|\mathcal{S}_{ij}|$. It is as large as $max\{round(|\mathcal{S}_{ij}|/30), 3\}$.

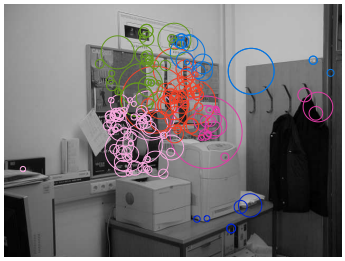


Figure 3: Clusters obtained by one call of k-mean clustering algorithm.

Obtained clusters of features define a set of initial hypotheses $\mathcal{H}_{ij} = \{H_{ij}^1, H_{ij}^2, \ldots, H_{ij}^n\}$. For each hypothesis H_{ij}^l a plane to plane homography P_{ij}^l from I_i to I_j is computed by applying the RANSAC algorithm (Algorithm 1). If the algorithm fails to find a solution the portions of features denoted by D and K are decreased by a factor 0.95 and the RANSAC is proceeded again.

Next, the coordinates of all matching features of $\mathcal{S}_{ij}$ are transformed to image I_j in accordance with transformation P_{ij}^l. Displacement errors $d(f_j^t, f_i^t P_{ij}^l); t = 1, 2, \ldots, |\mathcal{S}_{ij}|$ are computed as Euclidean distances. All features with displacement error below a pre-specified tolerance are included in the hypothesis (Figure 4). Note that features of the initial hypothesis can also be excluded from the hypothesis. Then, a plane to plane homography is recomputed and new features are included in the hypothesis. The process is repeated until there exist features that can be added to the hypothesis. This is demonstrated by Figure 5.

3 Step 2: Selection

Redundant set of clusters results in many 'overlapping' hypotheses. To reduce the redundancy and to keep the hypotheses that efficiently group the data a matrix of preference Q is introduced. It is preferred to have hypothesis with large number of features and small error-of-transformation. Duplication of features in hypotheses also has to be penalized. The selection of hypothesis is performed by maximization of an objective function of quadratic form $\mathbf{hQh}^T$ [6]. $\mathbf{h}$ is a binary vector of length n and denotes a set of selected hypotheses. A value 1 at position i

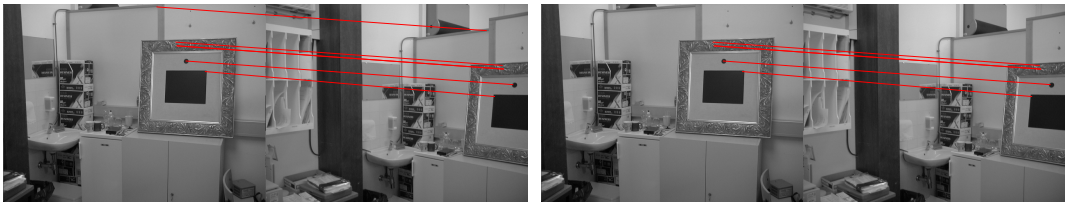
Algorithm 1 Random Sample Consensus Algorithm.

Assume:
The parameters are estimated from D data items.
There are T data items in total. (In our experiments $D = 0.7 \times T$.)
Tolerance t corresponds to the distance of maximal allowable displacement between features in a matching pair when transformed to the same image plane and is set to 1 pixel.

1. Select D data items at random.
2. Estimate parameters $\mathbf{p}$.
3. Find how many data items of T fit the model with parameters $\mathbf{p}$ within a tolerance t. Call this K.
4. If K is big enough exit with success. (In our experiments $K = 0.8 \times T$.)
5. Repeat steps from 1 to 4 L times. (In our experiments L=100.)
6. Fail if you get here.

(a) (b)

Figure 4: (a) Features of one cluster (left) and their matches (right). (b) The hypothesis is enlarged by adding all the features that satisfy the prespecified tolerance of plane to plane homography P.



(a) (b)

Figure 5: (a) One of the initial hypotheses. (b) The top most feature pair do not satisfy the tolerance criterium of plane to plane homography P, therefore, it is removed from the initial hypothesis.

indicates the presence of the i-th hypothesis and a 0 its absence. $\mathbf{Q}$ is a $n \times n$ symmetric matrix. The elements of $\mathbf{Q}$ are defined as $q_{cc} = K_1|Z_c| - K_2\xi_{c,c}$; and $q_{cr} = \frac{-K_1|Z_c \cap Z_r| + K_2\xi_{c,r}}{2}$; $c \neq r$. $|Z_c|$ is the number of features in the c-th hypothesis H^c_{ij}, i.e., $|Z_c| = sum(H^c_{ij})$. $\xi_{c,r}$, so called

the error-of-transformation, is defined as $\max(\sum_{f\in|Z_c\cap Z_r|} d(f, fP_{ij}^c)^2, \sum_{f\in|Z_c\cap Z_r|} d(f, fP_{ij}^r)^2)$. The constants K_1 and K_2 are the weights determined experimentally. (In our experiments $K_1 = 4$ and $K_2 = 1$.)

To maximize the objective function $\mathbf{hQh}^T$ we use the tabu search [1]. $\mathbf{h}$ that maximizes the objective function represents the final selection. Figure 6 depicts the hypotheses selected by maximization process. Note that each of them describes one plane.

(a) Hypothesis 1 (front site of the printer)

(b) Hypothesis 2 (wall newspaper)

(c) Hypothesis 3 (coat hanger)

Figure 6: The final set of hypotheses. Each of the selected hypothesis describes one plane.

3.1 Hypothesis rejection

Due to a small difference in camera locations for some acquired image pairs, (I_i, I_j), the computed plane to plane homography lacks the sensitivity and therefore groups together SIFT features which do not lie on the same plane. See for example Figure 7. To refuse such hypotheses the rejection process is applied to the set of final hypotheses. For each hypothesis H_{ij}^k we find all image pairs that contain matches determined by the hypothesis. The plane to plane homography is determined for each such image pair. If for at least one image pair the plane to plane homography does not satisfy most of the matches, the hypothesis H_{ij}^k is removed from further consideration.

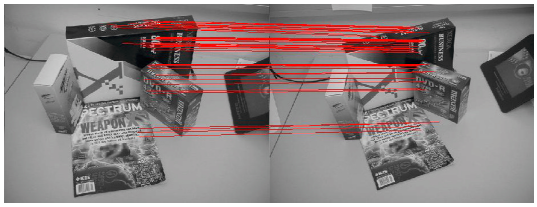


Figure 7: Refused hypothesis. If in $\mathcal{I}$ there exists an image pair for which the features of selected H_k do not lie on the same plane, H_k is removed from further consideration.

4 Step 3: Merging

Selections on pairs of images $\{(I_i, I_j)|i < j, i = 1, \ldots, N-1, j = 2, \ldots, N\}$ end up with a set of final hypotheses $\mathcal{H} = \{H_1, \ldots, H_m\}$. Each hypothesis determines a cluster of SIFT features. A SIFT feature is represented as a structure of feature coordinates (x, y), a SIFT vector, and a weight which determines the importance of the feature. At the beginning all weights are set to 1.

In $\mathcal{I}$, there are images representing the same parts of the scene acquired from different locations and viewing directions. Hence, many hypotheses determine the same parts of the scene. To reduce the redundancy and to enrich the final representation we apply to $\mathcal{H}$ a merging process.

SIFT descriptors are highly distinctive local parts of the scene therefore, even a small number of SIFT features uniquely determines the particular part of the scene. If in H_i and H_j there exists a subset of common matching features the hypotheses are candidates for merging. It is still possible that H_i and H_j describe two different planar parts or different parts of slightly bending surface. To refuse such cases features in both hypotheses are examined in the following way. First, we divide the features of H_i and H_j in three subsets: $A = H_i \cap H_j$, $B = H_i \setminus H_j$, and $C = H_j \setminus H_i$. Next, we find all image pairs that contain matches from all three above determined subsets. We require at least one such image pair to do the merging. By applying a plane to plane homography to each such image pair we test, if the matching features from subsets A, B, and C lie on the same plane. If for all such image pairs the test is positive, we merge H_i and H_j. Features of both hypotheses are transformed to the same image, for features in H_i and H_j the weights are summed, and all SIFT descriptors are kept. The process of merging is repeated (also on newly generated hypotheses) until there is no pair of hypotheses with sufficient number of matching features. The weights of features give us information about feature stability. Features with high weights are more stable while feature with low weights are very likely outliers.

The reader has to keep in mind that the merged hypotheses are still only hypotheses. By acquiring new images of the scene new information is obtained and the rejection of a hypothesis is still possible.

5 Experiments

he results are presented for two experiments. In the first experiment the scene is fixed. In the second the configuration of objects in the scene is different for the acquired set of images. In both experiments we deal with gray images of resolution 640×480.

In the first experiment the feature clustering was generated from 15 images leading to 86 final hypotheses. After the process of merging we end up with 8 different planes (Fig. 8).

In the second experiment 10 different images were acquired. The process ends up with 54 hypotheses (Fig. 9). Some hypotheses of feature clusters, belonging to the some plane, were not merged due to the sparse nature of SIFT features and insufficient number of acquired images. Since the scene is altered from image to image, the results can show the location of features belonging to one cluster, even though the planar part is occluded by other objects.

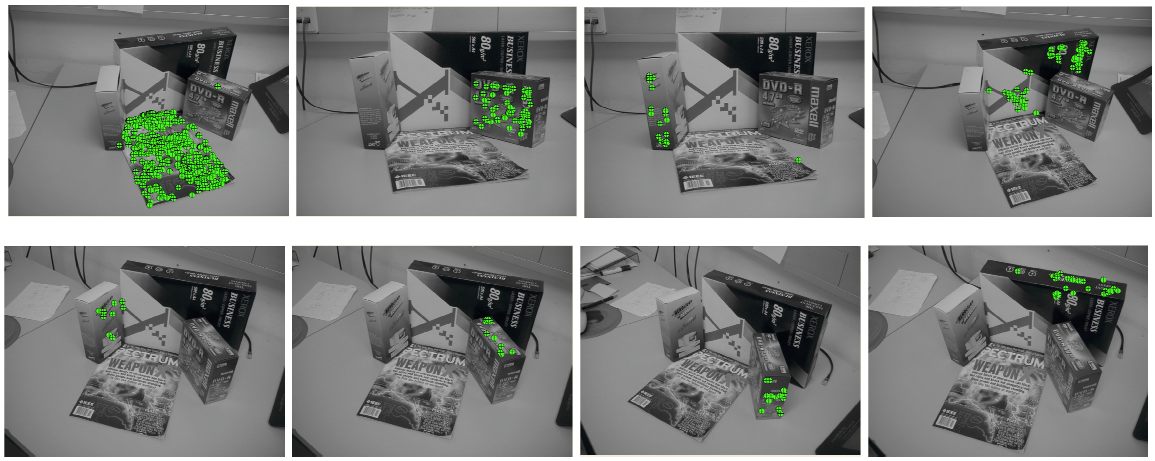


Figure 8: Experiment with a fixed scene. The method finds seven clusters of SIFT features belonging to seven different planar parts of the scene.

6 Conclusion

In the work we represent a method for clustering the SIFT features belonging to planar surfaces. The clusters obtained through the phases of exploration, selection and merging can be used as initial structures for building higher level scene representations. The proposed method can also be understood as unsupervised learning of objects with planar parts, what is demonstrated by the second experiment. The attached weights to the SIFT descriptors can also be exploit to detect changes in the interior scene, e.g., changes on wall newspaper, a coat hanger, and would together with time parameter allow continuous long time learning.

Figure 9: Scene is altered from view to view. Eleven different clusters are found belonging to five different planar parts of the scene. Only six clusters are displayed.

Acknowledge

This work was supported in part by the EU project *CogVis (IST-2000-29375)*, the grants funded by the Ministry for Education, Science and Sport: Research Program *Computer Vision-1539-506* and *SLO-A/07* and by the Federal Ministry for Education, Science and Culture of Austria under the CONEX program.

References

[1] D. de Werra A. Hertz, E. Taillard. A tutorial on tabu search. In *Proc. of AIRO'95*, pages 13–24, Italy, 1995.

[2] A. Criminisi, I. Reid, and A. Zisserman. A plane measuring device. In *In Proc. BMVC*, September, 1997.

[3] Gy. Dorkó and C. Schmid. Selection of scale invariant neighborhoods for object class recognition. In *Proceedings of the 9th ICCV, Nice, France*, pages 634–640, 2003.

[4] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the 7th ICCV, Corfu*, pages 1150–1157, 1999.

[5] S. Se, D. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings of the IEEE ICRA*, pages 2051–2058, Seoul, Korea, May 2001.

[6] Markus Stricker and Ales Leonardis. Exsel++: A general framework to extract parametric models. In *Computer Analysis of Images and Patterns*, pages 90–97, 1995.