Elektrotehniški vestnik 74(5): 297–302, 2007
Electrotechnical Review, Ljubljana, Slovenija

Unsupervised learning of scene and object planar parts

Katarina Mele, Jasna Maver

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: 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. The proposed method is an example of unsupervised learning of planar parts of the scene and objects with planar surfaces.

Key words: unsupervised learning, visual learning, local descriptors, SIFT descriptor, feature grouping

Nenadzorovano učenje prizora in planarnih objektov v njem

Povzetek. Metoda, predstavljena v članku, je namenjena nenadzorovanemu učenju prizora oziroma planarnih delov objektov, ki ga sestavlajo. Učenje je izvedeno s poljubnim naborom slik, zajetih iz različnih zornih kotov. Predlagani postopek je prilagojen za natančno in robustno razvrščanje velikanskega števila lokalnih deskriptorjev SIFT v skupine, ki določajo posamezne planarme dele v prizoru. Geometrijske enote, ki jih dobimo s takim urejanjem nizkonivojskih značilk, so most med nizkim oz. zaznavnim in visokim oz. vsebinskim nivojem razumevanja vizualne informacije.

Metoda je sestavljena iz treh korakov: raziskovanja prostora, izbire hipotez in združevanja hipotez. Prvi korak, raziskovanje vizualne informacije, je podatkovno voden postopek, ki zgradi širši nabor hipotez. Izbor hipotez je izveden s kvadratno formo preferečne matrike, ki ovrednoti hipoteze glede na število značilnic in transformacijsko napako, pri tem se podvajanje značilnic penalizira. V zadnjem koraku združimo hipoteze, ki so isti planarni del, izračunan iz slik, zajetih iz različnih zornih kotov. Tako se izognemo podvajanju hipotez in hkrati obogatimo predstavitev posameznega dela prizora. Eksperimentalni rezultati potrjujejo uspešnost metode za nenadzorovanno učenje planarnih delov prizora in objektov.

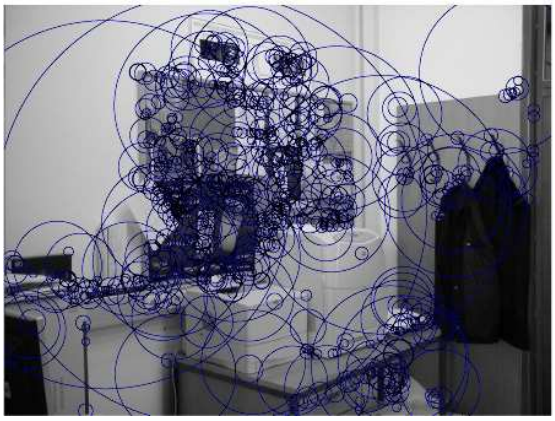
Ključne besede: nenadzorovano učenje, vizualno učenje, lokalni deskriptorji, deskriptor SIFT, grupiranje značilnic.

1 Introduction

The use of local features is becoming increasingly popular for solving different vision tasks. Recently, the SIFT descriptor has been proposed for describing distinctive scale-invariant features in images [7]. 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 [10] the authors propose to use SIFT features for building 3D maps. Local descriptors have previously been used for scene description [4]. In [11, 9] local descriptors are used to extract objects from video clips but no 3D information about the object is generated. On the other hand in work of [6] 3D geometrical information is built about object surfaces. 3D geometrical presentation is model from range images.

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 the planar surfaces of objects. In for example [2, 5, 8] features such as line segments and junctions are selected for plane description. RANSAC algorithm is used to estimate transformation between images [13]. Here we experiment with SIFT descriptors as they uniquely describe a particular part of the scene. For an arbitrary set of images acquired from different views, the method organizes a huge number of local SIFT features to fill the gap between the low-level vision (front end), i.e. outputs of various filtering operations and high-level vision, i.e., domain-specific rea-

Received 3 April 2005
Accepted 28 October 2005



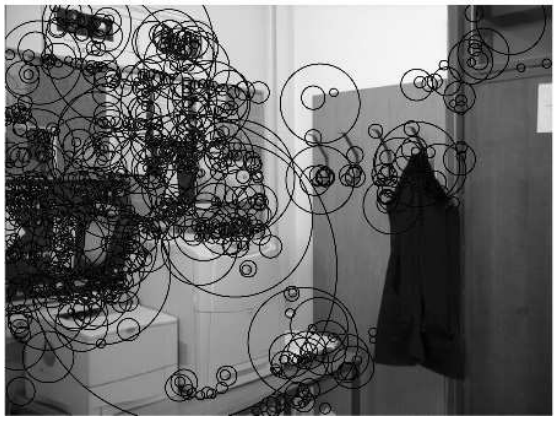


Figure 1. Illustration of feature extraction. Each circle corresponds to a region described by the SIFT descriptor.

soning about geometric structures. The proposed method consists of three steps: exploration, selection, and merging with verification. The exploration step is a data-driven technique that proposes a set of hypothesis models from which the selection step chooses the ones that explain the data in accordance with a matrix of preferences. Since the set of local features varies 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 an interior scene, the goal is to group them into 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 a particular interior scene. The first step of our approach is detection of DoG points and computation of SIFT descriptor for each local region [7] (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 the 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 four 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).

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 [3]. The

Figure 2. The best matches from $\mathcal{S}_{ij}$.

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 a homogeneous estimation method. A reliable solution requires to start the process of plane searching with a large set of small SIFT feature clusters, i.e. the 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-means clustering algorithm. The algorithm is performed several times, each time starting with different arbitrarily initial sets of cluster centers. The value k denotes the number of clusters obtained by one iteration and depends on the number of features $|\mathcal{S}_{ij}|$. In the experiments the k was set to $k = max\{round(|\mathcal{S}_{ij}|/30), 3\}$.

The 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 proportions 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 a displacement error below a pre-specified tolerance are included in the hypothesis (Figure 3). 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 stopped when there are no features that can be added to the hypothesis.

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.

---

3 Step 2: Selection

A 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 a hypothesis with a large number of features and small error-of-transformation encoded in diagonal elements of Q. The off-diagonal terms encode the interaction between the hypotheses. Duplication of features in different hypotheses is penalized. We consider only pairwise overlaps of the hypotheses. Selection of the hypotheses is performed by maximization of an objective function of quadratic form $\mathbf{h}Q\mathbf{h}^T$ [12]. $\mathbf{h}$ is a binary vector of length n and denotes a set of selected hypotheses. A value 1 at position i indicates the presence of the i-th hypothesis and 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 $\max(\sum_{f \in |Z_c \cap Z_r|} d(f, fP^c_{ij})^2, \sum_{f \in |Z_c \cap Z_r|} d(f, fP^r_{ij})^2)$. 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 $\mathbf{h}\mathbf{Q}\mathbf{h}^T$, we use the tabu search [1]. Vector $\mathbf{h}$ that maximizes the objective function represents the final selection. Figure 4 depicts the hypotheses selected by the proposed approach. Note that each of them describes one plane.

3.1 Hypothesis rejection

Due to small differences 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 5. To refuse such hypotheses, the rejection process is applied to give the final set of hypotheses. For each hypothesis H^k_{ij} we find all image pairs that contain matches relevant to 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^k_{ij} 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

(a)

(b)

Figure 3. (a) Initial hypothesis. (b) The hypothesis is enlarged by adding all the features that satisfy the prespecified tolerance of plane-to-plane homography P.

vector, and a weight which determines the importance of the feature. At the beginning all the 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, multiple hypotheses can determine the same parts of the scene. To reduce the redundancy and to enrich the final representation, we apply a merging process to $\mathcal{H}$.

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 a slightly bending surface. To filter out 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 the three above determined subsets. We require at least one match from each subset 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, the weights of features H_i and H_j weights are summed and all the SIFT descriptors are kept. The process of merging is repeated (also on newly generated hypotheses) until there is no pair of hypotheses with a sufficient number of matching features. The weights of features give us information about feature stability. Features with high weights are more stable while features with low weights are very likely to be 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

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. 6).

In the second experiment 10 different images were acquired. The process ends up with 54 hypotheses (Fig. 7). Some hypotheses of feature clusters of the same plane were not merged due to the sparse nature of SIFT features and insufficient number of the acquired images. The hypotheses are built from different images, showing the same planar part from angels where some parts are un-

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

(b) Hypothesis 2 (wall newspaper)

(c) Hypothesis 3 (coat hanger)

Figure 4. Final set of hypotheses. Each of the selected hypothesis represents one plane.

Figure 5. Refused hypothesis.

seen or occluded by objects. Consequently also the results can show the location of features belonging to one cluster, even though the planar part is unseen or occluded by other objects.

6 Conclusion

In this work we present 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 represents unsupervised learning of objects with planar parts as demonstrated by the second experiment. The weights attached to the SIFT descriptors can also be exploited to detect changes in the interior scene, e.g. changes on the wall newspaper, a coat hanger, and would together with the time parameter allow for continuous long-time learning.

7 References

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

[2] C. Baillard and A. Zisserman. Automatic Reconstruction of Piecewise Planar Models from Multiple Views. *CVPR*, pages 2559–2565, 1999.

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

[4] G. 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.

[5] Z. Kim and R. Nevatia. Automatic description of complex buildings from multiple images. In *Comput. Vis. Image Underst.*, pages 1077-3142, 2004.

[6] A. Leonardis Gradnja in Modeliranje parametri nih struktur v slikah In *Electrotechnical Review*, volume 3/4(58), pages 133–141, 1991.

[7] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. In *International Journal of Computer Vision*, volume 60(2), pages 91–110, 2004.

[8] I. K. Park, K. M. Lee and S. U. Lee Perceptual grouping of line features in 3-D space: a model-based framework. In *Pattern Recognition*, pages 145-159, 2004.

[9] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. Segmenting, Modeling, and Matching Video Clips Containing Multiple Moving Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 29(3), pages 477–491, March 2007.

Figure 6. Experiment with a fixed scene. Eight clusters of SIFT features belonging to eight different planar parts of the scene are found.

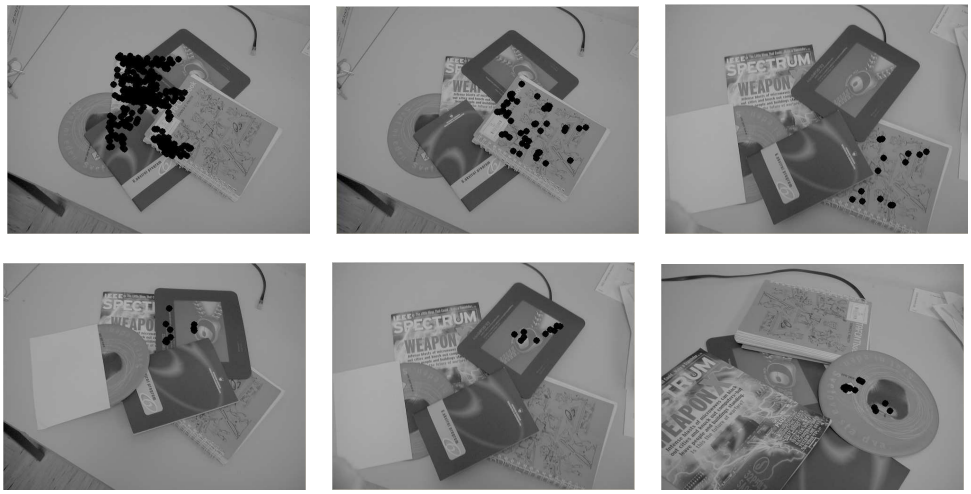


Figure 7. Scene is altered from view to view. Eleven different clusters are found belonging to five different planar parts of the scene. Only six clusters are displayed.

[10] 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.

[11] J. Sivic, F. Schaffalitzky and A. Zisserman, Object Level Grouping for Video Shots. In *Proc. Int. J. Compt. Vision*, volume 67, number 2, pages 189–210, Hingham, MA, USA, 2006.

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

[13] J. Wills, S. Agarwal and S. Belongie, What Went Where. In *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, volume 1, pages 437–54, Madison, WI, June 2003.