Combining Textural and Geometrical Descriptors for Scene Recognition

Neslihan Bayramoğlu, Janne Heikkilä, and Matti Pietikäinen

Center for Machine Vision Research, University of Oulu, Finland
{nyalcinb,janne.heikkila,matti.pietikainen}@ee.oulu.fi

Abstract. Local description of images is a common technique in many computer vision related research. Due to recent improvements in RGB-D cameras, local description of 3D data also becomes practical. The number of studies that make use of this extra information is increasing. However, their applicabilities are limited due to the need for generic combination methods. In this paper, we propose combining textural and geometrical descriptors for scene recognition of RGB-D data. The methods together with the normalization stages proposed in this paper can be applied to combine any descriptors obtained from 2D and 3D domains. This study represents and evaluates different ways of combining multi-modal descriptors within the BoW approach in the context of indoor scene localization. Query's rough location is determined from the pre-recorded images and depth maps in an unsupervised image matching manner.

Keywords: 2D/3D description, feature fusion, localization.

1 Introduction

The scene recognition problem is widely studied in many research areas such as robot localization, path planning, similarity retrieval, matching and classification. In this study we focus on the indoor image matching problem in which the scene information is gathered from multi-modal (2D/3D) sensors. Today, extracting such information from real environments is rather effortless with the help of the new generation depth cameras and range scanners such as Kinect [1].

These RGB-D cameras, can acquire both color data (RGB) and the depth data in real dimensions. They have significant advantages compared to laser scanning devices: *i)* they can operate in real time (up to 30 Hz), *ii)* they are affordable, and *iii)* depth data is synchronized with the color information.

Unlike the Simultaneous Localization and Mapping (SLAM) [2] methods where the exact pose of the robot (position and orientation) is required and unlike the scene classification problems [3] in which semantics and categories are extracted, in this paper, scene recognition is regarded as determining a rough location (topological place) from the pre-recorded images and depth-maps associated with labels in an unsupervised image matching manner (Fig. 1). In this context, our main aim is to investigate the effects of local geometrical properties in scene recognition research. This study also represents different ways of combining textural and geometrical descriptors within the Bag-of-words approach.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 32–41, 2012.

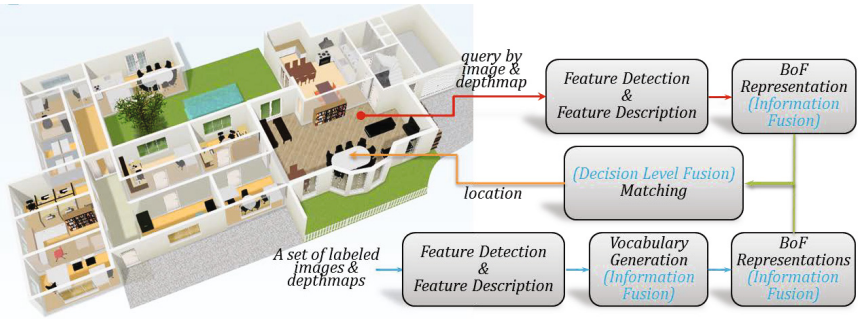

Fig. 1. Flowchart of the proposed method. Information fusion can be performed in different stages depending on the combination method.

1.1 Related Work

In literature, indoor scene matching is considered to be more challenging than the outdoor scene matching problem [4, 5] because outdoor scenes contain more discriminative and unique features leading to comparatively easy recognition. Besides, indoor scenes comprise many similar structures such as doors, windows, chairs, etc. Consequently, such images resemble each other in the global sense but can still contain local distinguishing patterns. The accuracy of this inspiration is testified by the state-of-the-art scene matching methods. Sivic and Zisserman [6] demonstrated efficiency of the region based descriptors within the Bag-of-words (BoW), or Bag-of-Features (BoF), approach for scene retrieval by making analogy between the text retrieval. After that, substantial amount of work adopted this BoF model for scene matching [2, 4, 5].

Until recently, vision based scene recognition methods were utilizing texture based features (2D) [4–7]. After the release of RGB-D cameras, 3D features have also started to be used for scene recognition purposes [8–11]. Ren et al. [8] utilized local 3D features together with texture based descriptors for supervised object classification. Similarly, Janoch et al. [9] applied histogram of oriented gradients (HOG) on depth maps for supervised object classification. However, they did not integrate depth features with the texture based ones. Browatzki et al. [11] also combined 2D features with 3D ones for supervised object categorization in which the objects are isolated and viewed from several angles.

The closest work to our study is described in [10]. While the main emphasis in their work is on the scene classification and segmentation with object labelling, they applied texture based descriptors onto the depth maps for scene classification. However, in this study, we are extracting several 3D features from the 3D point cloud representations instead of mapped 3D information (depth maps). We also demonstrate three different ways of combining 2D and 3D information for scene matching.

2 Theoretical Background

2.1 Bag-of-Words/Features

In the BoF approach there are three main steps: *i)* feature detection and description, *ii)* construction of visual vocabulary(dictionary), *iii)* matching. The main goal in the feature detection is to find keypoints holding significant information that are also robust across transformed versions of the image. A comparison of some of the interest point detection algorithms can be found in [12]. Feature descriptors which are usually represented as vectors carry local information in the neighbourhood of each keypoint. Scale Invariant Feature Transform (SIFT)[13] and Speeded Up Robust Feature (SURF) [14] are popular descriptors because of their accomplished performances. Visual vocabulary is built by clustering all the extracted features from a dataset of images. The selection of the number of clusters (k) is empirical, although it is critical. Obtaining BoF representation of database images is the next stage. First, features are extracted from an image. After that features are assigned to the closest cluster (word) in the vocabulary. Then, the count of each word that appears in the image is used to form the BoF representation of the image. When a query is placed, firstly the BoF representation is constructed. After that, the BoF representation of query image and the BoF representation of the database images are compared and matched.

2.2 Local 3D Features

Since the 3D object description has become popular during the last decade the number of research efforts is less than the 2D counterparts. In the shape analysis research, global description of 3D mesh models is popular. Still there exist local 3D feature detectors and descriptors. We refer the reader to Tangelder and Veltkamp [15] for a detailed survey.

Methods that rely on the global descriptors are usually utilized in similarity retrieval of 3D mesh models among the database of the same type, whereas local ones are employed in partial matching and point correspondences. There are two challenges regarding the 3D local descriptors. First one is the time complexity. Most of the 3D (or point cloud) descriptors rely on the normal vector information. The normal vector estimation step usually requires analysis of a covariance matrix. Besides, the nearest neighbourhood search or a similar strategy is utilized in finding the local neighbourhood of a keypoint which, if repeated many times, lead to high computation cost. The second and the critical challenge is the following: 3D local descriptors are considered to be less discriminative and far from being robust, since 3D shapes (surfaces) have insufficient features and keypoint repeatability is usually not satisfied [16].

In this study, we employed *spin images* descriptor which is one of the well-known local surface descriptors [17]. It is a two-dimensional histogram of the spatial distribution of neighbouring points around a keypoint (Fig. 2). Utilizing the keypoint's normal vector makes the descriptor rotation and translation invariant. However, *spin images* descriptor does not ensure the scale invariance.

We also utilized *D2 distributions* descriptor which is a simple yet an efficient descriptor. It is originally proposed as a global descriptor for 3D mesh model retrieval [18] and corresponds to the distribution of the Euclidean distances between object points that are selected randomly. The *D2 distributions*, which is also rotation and translation invariant, does not rely on the normal vector information, it is faster than the *spin images* descriptor. In addition to these descriptors, one could also use more sophisticated methods such as *Fast Point Feature Histograms* [19] for describing local geometry.

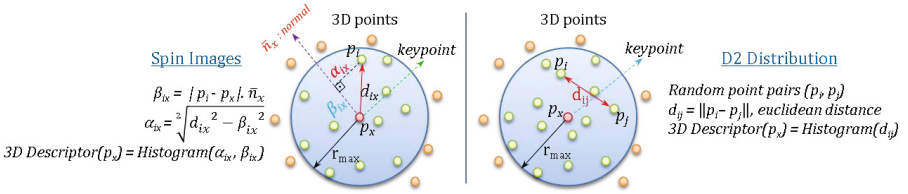

Fig. 2. Local 3D descriptors, *left*: spin images, *right*: D2 Distributions.

3 Combining 2D and 3D Features

Since RGB-D data is composed of two different modalities (texture and 3D geometry), there are several ways for extracting and also for combining this information. Table 1 shows straightforward ways of feature detection and feature description for RGB-D data. Feature detectors in 3D domain generally depend on the curvature information. Mapped data can be considered as 2D image in which the pixels represent a function value at the corresponding 3D location for the given geometry such as depth maps, shape index mapped images, etc.

Table 1. Feature Detection and Description for RGB-D Data

Feature Detection	Feature Description
2D Detector	2D Descriptor
3D Detector	3D Descriptor
Detection on mapped data	

The most critical point of feature detection is keypoint repeatability. We believe that keypoint repeatability is higher in 2D images than in the 3D domain. Therefore, in our tests feature detection is performed only on 2D images. We employed SURF as a feature detector. Also, in this study, we utilized only SURF features as the texture based descriptor, since the objective is not to optimize the performance of the 2D descriptors but to find the best way to combine the two types of descriptors.

After extracting information from the 2D and 3D representations, gathering them to obtain a single descriptor is the next stage. We employed following strategies to combine these descriptors: *i)* Point Description Fusion, *ii)* Scene Description Fusion, and *iii)* Decision Level Fusion.

3.1 Point Description Fusion

A keypoint can be described locally by its textural and geometrical properties if both the 2D image and the 3D information are available. These descriptors can be concatenated to form a single vector as is done in [10] and in [20]. In this type of fusion, keypoints should be selected from the single modality. Dense feature detectors can also be utilized; however, both (2D and 3D) descriptors should exist for each keypoint. Contrary to the previous studies, we carried out a normalization step before concatenating descriptors. Normalization is necessary if the 2D and 3D descriptor vectors are having different lengths and/or their order-of-magnitudes differs a lot. We propose the following normalization on the descriptor vectors f_k:

$$\mu[i] = \frac{1}{N}\sum_{k=1}^{N} f_k[i], \;\; \tilde{f}_k[i] = \frac{f_k[i]}{\sqrt{\frac{1}{N}\sum_{j=1}^{N}\sum_{l=1}^{n}\left(f_j[l]-\mu[l]\right)^2}}, \; i \in (1,2,\ldots,n) \quad (1)$$

where N is the total number of descriptors (2D/3D) obtained from the training set, n is the size of the descriptor vectors and $\tilde{f}_k$ is the normalized descriptor of point k. The final descriptor is then formed by scaling these normalized descriptor vectors by some constants φ and γ as in Fig. 3. These constants can be selected as $\varphi = 1$ and $\gamma = 1$ for equal contribution and also be adjusted if one of the modalities (2D/3D) is desired to have more influence on the description than the other. This normalization enables to concatenate point descriptors which are extracted from any means of modality.

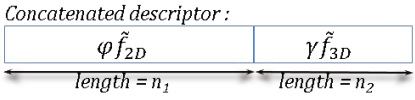

Fig. 3. Final descriptor is a concatenation of normalized and scaled feature vectors which are obtained from texture and shape

3.2 Scene Description Fusion

Information fusion can also be done in the stage of scene representation (Fig 4). In this case, BoF descriptors that are separately obtained from textures and point clouds are concatenated. This configuration introduces the flexibility of choosing different number of clusters for 2D and 3D descriptors. Also keypoint detection can be performed in different modalities.

The algorithm makes two passes over the scenes: one-pass to construct the dictionary of 2D descriptors and second to obtain the dictionary of 3D descriptors. After that, each scene is described by two histograms obtained from the

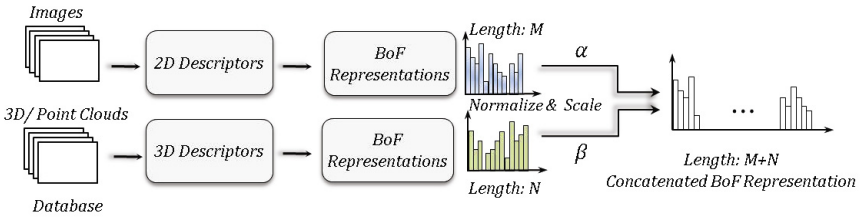

Fig. 4. Scene representations are combined after Bof Representations are extracted from 2D and 3D features seperately

BoF representations. We propose normalizing and scaling these individual BoF representations before concatenating. Histograms are normalized such that their sums are equal to one. Scaling histograms with some constants α and β is again necessary to compensate the effects of different vocabulary sizes and also for assigning weights to the description type (2D/3D) because individual description performances of 2D and 3D descriptors are different.

3.3 Decision Level Fusion

Decision level fusion is a common approach for combining classifiers [21]. A similar strategy can be employed in image matching methods. For combining the 2D and 3D information, we propose evaluating the BoF based indexing approach separately for 2D images and 3D point clouds. After the query is presented, images in the database are indexed (rank list) according to their similarity measures for both 2D and 3D. Then, using the ranked results of each modality re-indexing is utilized. The new criterion puts emphasis on the joint position on the ranked lists. That is, if the similarity score of an image (location) is higher in both the 2D similarity list and the 3D similarity list, then the probability of being the true location increases. We propose the new score of an image as a weighted combination of the individual scores:

$$\begin{gathered} score(image_k)_{2D/3D} = 1 - \frac{index_{k(2D/3D)}}{\#ofimages} \\ score(image_k)_{combined} = w_1 \times score(image_k)_{2D} + w_2 \times score(image_k)_{3D} \end{gathered} \tag{2}$$

where $index_k$ denotes the position of the $image_k$ in the similarity list and w_1 and w_2 are the weights. The ratio of the weights w_1 and w_2 can be tuned such that it can reflect the ratio of the matching performances of 2D and 3D descriptors. The highest scored image's label is returned as the location of the query.

4 Dataset and Experimental Results

Our dataset sampled from the publicly available dataset [10] consists of 1626 unique Kinect frames, spread over 64 different indoor environments. In the original dataset, *"Bookstore"* scenes have higher number of samples. We reduced

this number since this may affect the fair evaluation of the methods. Fig. 5 represents a sample image from the dataset, corresponding raw depth map and the registered point cloud respectively. Table 2 gives a list of location types and the number of scenes belonging to that type. In our experiments, instead of using raw depths we utilized processed depth maps provided by [10].

Fig. 5. Sample image from the dataset, associated raw depth map and the registered point cloud respectively

Given a query we search for the closest image from the dataset depending on the distance measure. To evaluate the accuracy of descriptors all images are included in both test and training sets. Therefore the top match is always the query itself, we report recognition rates for the second match. *OpenCV*'s SURF implementation is employed in our experiments. The implementation of BoF stage is standard, built dictionaries using k-means clustering and hard assignment is utilized in feature mapping. We use the Fast Library for Approximate Nearest Neighbors *(FLANN)* also from the *OpenCV* library for ranking the image similarities. BoF representations are compared using L_2 norm. SURF generates feature descriptors of size 64, and in our implementation sizes for *D2 Distribution* and *spin images* are 16 and 100 ($\alpha \times \beta$) respectively. Originally *spin images* are 2D histograms, but we vectorized them for combination purposes.

Firstly, individual recognition rates are evaluated for 2D and 3D descriptors. The optimal cluster number k is decided by evaluating the recognition rates for different k values. Fig. 6 represents the recognition rate as a function of dictionary size. The SURF descriptor achieves the highest recognition rate at 87.57%, whereas *D2 Distribution* has a 72.44% and *spin images* has a 68.17% recognition rate at best. The highest recognition rate is achieved with 2D descriptor at dictionary size of 100, whereas it is 50 with the 3D descriptors. However, k is not the the only parameter which needs to be tuned. The local region size is another important parameter which affects the discriminative power of the 3D descriptors. For very small regions, 3D descriptors cannot convey significant information, since surfaces tend to change smoothly. On the other hand, if evaluation is performed on bigger regions, due to clutter and occlusion, descriptor may contain unrepeatable representation of a combination of different objects. Fig. 7 shows the relation between the recognition rate and the local region size. Both of the 3D descriptors achieved highest rates at 0.7 m with recognition rates of 72.44% with *D2 Distribution* and 75.46% with *spin images* descriptor.

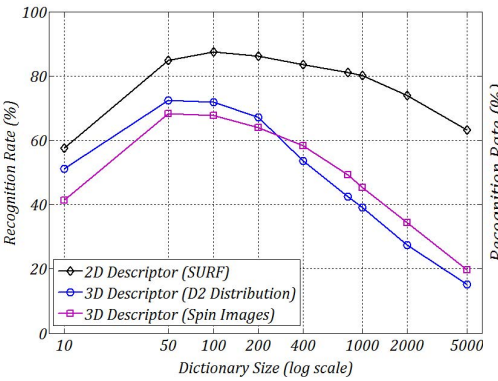

Fig. 6. Individual performances

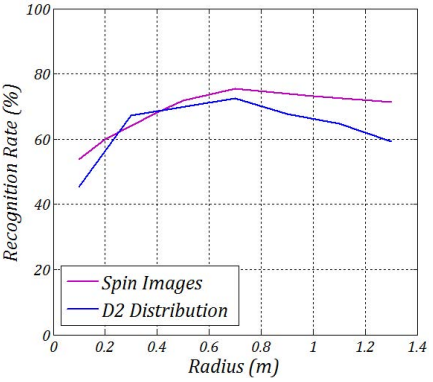

Fig. 7. Effet of local region size

Table 2. Dataset Statistics

Place Type	Scenes
Living Room	13
Office	14
Kitchen	10
Bedroom	17
Bathroom	6
Book store	3
Cafe	1
Total	64

Table 3. Scene Description Fusion

	Recognition Rate (%)	
Parameters	SURF-D2	SURF-Spin
$(\alpha = 1, \beta = 1)$	85.17	87.27
$(\alpha = 1, \beta = 2)$	86.41	88.49
$(\alpha = 1, \beta = 4)$	**89.85**	**89.98**
$(\alpha = 1, \beta = 8)$	88.42	88.93

Table 4 gives the overall evaluation of the methods presented here. In combining the point descriptors, normalization constants are tuned as $\varphi = 2, \gamma = 1$ for *D2 Distribution* and $\varphi = 4, \gamma = 1$ for *spin images* descriptor with the dictionary size of 100. Dictionary sizes *(M/N)* for scene description fusion are 100 and 50 for 2D and 3D respectively. The normalization constants are tuned as $\alpha = 1, \beta = 4$ for both *D2 Distribution* and *spin images* descriptor (Table 3). In decision-level fusion, highest recognition rate is achieved with parameters $w_1 = 20$ and $w_2 = 1$.

Individual performances of 3D descriptors are lower than that of the 2D descriptor, since local 3D features are not as rich as 2D ones. After combining 2D and 3D features we obtained about a 6% improvement with the decision-level method. Other combination methods did not improve the recognition rates significantly. At first glance, one can expect that concatenated feature vector should describe a keypoint more precisely. However, some keypoints which own different texture descriptors may have same 3D descriptors (similar surface properties) and some other keypoints having similar textures may reside on different surfaces. Therefore, introducing 3D features decreases the discriminative power of the texture based descriptors for the former case, similarly 3D features will

Table 4. Overall Evaluation

Method	Recognition rates
2D only Description (SURF)	**87.57%**
3D only Description (D2-Distributions)	72.44%
3D only Description (Spin Images)	75.46%
Point Fusion, SURF/D2-Distributions	87.69%
Point Fusion, SURF/Spin-Images	86.28%
Scene Fusion, SURF/D2-Distributions	89.85%
Scene Fusion, SURF/Spin-Images	89.98%
Decision-Level, SURF/D2-Distributions	**93.54%**
Decision-Level, SURF/Spin-Images	93.41%

become less effective in the latter case. Besides, using single dictionary size also decreases the performance. Therefore, fusing scene descriptions that enables utilizing individual dictionaries in different sizes performs better.

5 Discussions and Conclusions

In this study, we proposed and evaluated integrating local descriptors of images and 3D data in the context of indoor scene localization. We proposed normalization methods which are necessary and critical in information fusion. Since the RGB-D data contains complementary information with depth maps and RGB, their combination always introduces improvements in recognition. In our dataset, images belonging to same location is captured at the same time with similar illumination conditions. As a result, 2D descriptors achieved a considerably high recognition rate by themselves. We believe that 3D information will be more effective for datasets containing high illumination variations among images. In that case, 2D descriptors will not be as successful as in our case. Since 3D descriptors are not affected from the illumination they would have better impact on the combined result. We can conclude that local 3D features enhance the performances of local 2D descriptors with proper combinations. However, depending on the application, the increased complexity introduced by 3D descriptor extraction and also by file I/O operations of 3D point clouds should be considered.

Acknowledgments. This work is supported by Infotech Oulu.

References

1. Microsoft: Introducing kinect for xbox 360, `http://www.xbox.com/en-US/Kinect/`
2. Cummins, M., Newman, P.: Fab-map: Probabilistic localization and mapping in the space of appearance. Int. J. Rob. Res. 27, 647–665 (2008)

3. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE CVPR, pp. 2169–2178 (2006)
4. Kang, H., Efros, A.A., Hebert, M., Kanade, T.: Image matching in large scale indoor environment. In: IEEE CVPR Workshop on Egocentric Vision (2009)
5. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: IEEE CVPR, pp. 413–420 (2009)
6. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: IEEE ICCV, pp. 1470–1477 (2003)
7. Grauman, K., Darrell, T.: Efficient image matching with distributions of local invariant features. In: IEEE CVPR, pp. 627–634 (2005)
8. Ren, X., Bo, L., Fox, D.: Rgb-(d) scene labeling: Features and algorithms. In: IEEE CVPR (2012)
9. Janoch, A., Karayev, S., Jia, Y., Barron, J., Fritz, M., Saenko, K., Darrell, T.: A category-level 3-D object dataset: Putting the kinect to work. In: IEEE ICCV Workshops, pp. 1168–1174 (2011)
10. Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. In: IEEE ICCV Workshop on 3DRR (2011)
11. Browatzki, B., Fischer, J., Graf, B., Bulthoff, H., Wallraven, C.: Going into depth: Evaluating 2D and 3D cues for object classification on a new, large-scale object dataset. In: IEEE ICCV Workshops, pp. 1189–1195 (2011)
12. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. Int. J. Computer Vision 65, 43–72 (2005)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Computer Vision 60, 91–110 (2004)
14. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer Vision Image Underst. 110, 346–359 (2008)
15. Tangelder, J.W.H., Veltkamp, R.C.: A survey of content based 3D shape retrieval methods. Multimedia Tools Appl. 39, 441–471 (2008)
16. Bronstein, A.M., Bronstein, M.M., Guibas, L.J., Ovsjanikov, M.: Shape google: Geometric words and expressions for invariant shape retrieval. ACM Trans. Graph. 30, 1–20 (2011)
17. Johnson, A.E., Hebert, M.: Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Transactions on PAMI 21, 433–449 (1999)
18. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Matching 3D models with shape distributions. In: IEEE Int. Conf. on Shape Mod. & App. (2001)
19. Rusu, R.B., Blodow, N., Beetz, M.: Fast Point Feature Histograms (FPFH) for 3D Registration. In: IEEE ICRA, pp. 3212–3217 (2009)
20. Tombari, F., Salti, S., Di Stefano, L.: A combined texture-shape descriptor for enhanced 3D feature matching. In: IEEE ICIP, pp. 809–812 (2011)
21. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Transactions on PAMI 20, 226–239 (1998)