Place Recognition Using Image Retrieval with Covariance Descriptors

F. Dornaika [a,b], *A. Assoum* [c], *and A. Moujahid* [a]
[a] *University of the Basque Country UPV/EHU, San Sebastian, Spain*
[b] *IKERBASQUE, Basque Foundation for Science, Bilbao, Spain*
[c] *LaMA Laboratory, Lebanese University, Tripoli, Lebanon*

Abstract

Visual place recognition is an interesting technology that can be used in many domains such as localizing historical photos, autonomous navigation and augmented reality. The main stream of research in that domain was based on the use of local invariant features like SIFT. Little attention was given to region descriptors which can encompass local and global visual appearances. In this paper, we provide an empirical study on a particular visual descriptor: covariance matrices. In order to enhance the discriminative power of the descriptor, multi-block based descriptors are designed and compared. We show further experimental results on matching test images with reference images acquired in dense urban scenes in the streets of the city of Paris. Experiments show that the multi-block based matching algorithms can lead to both high accuracy and scalability.

Introduction

Overview

Image-based localization is the task of determining the location from which a query photo was taken [1, 2, 3, 4, 5, 6, 7]. The problem is commonly formalized as identification of reference images depicting the same scene as the query which can be followed by viewpoint estimation. Researchers start to focus on image-based localization as it enables many interesting applications such as real-time camera pose tracking and robot navigation [8].

Unmanned navigation and more generally Advanced Driver Assistance Systems (ADAS) require localization of the vehicle. For vehicle localization it is shown that GPS can be the most popular tool that performs global localization [9]. However, despite its popularity, GPS has some limitations that are very often observed in dense urban environments. Indeed, in such environments satellite signals might be intercepted or affected by buildings and other urban structures. This limitation in visibility and bad geometry of satellites or the multi-path problem can then decrease the accuracy of the GPS based localization. To alleviate these shortcomings, other sensors are deployed and added to the localization system to complete the GPS information. These perception sensors can be a camera or a Lidar.

Simultaneous Localization and Mapping (SLAM) was considered one of the vision based approach for localization. It was mainly developed for mobile robots. Good progress has been achieved in SLAM, but is still far from being an established and reliable technology due to the lack of robustness. Most place recognition techniques typically utilize an image retrieval approach. As can be seen, whether the location is a geo-referenced 3D pose or simply 2D coordinate on a 2D map, image retrieval is an essential step in any image-based localization system. Image retrieval can benefit from many advances regarding descriptors and their matching [10, 11, 10, 12]. It consists in determining the most similar reference image to a query image, and hence the most likely location from where the query image was captured.

Image retrieval is based on a trivial fact: a visited place should give rise to an image similar to the one in the database for the same location. However, image appearance can drastically affected by many perturbing factors. This technique requires two main processing steps: (i) extracting image descriptors on which the similarity is based, and (ii) computing a similarity score that should be used in the decision phase.

In [13], the author proposes a fast appearance-Based Mapping (FAB-MAP) technique, which employs Bag-of-Words (BOW) image retrieval systems and a Bayesian framework. FAB-MAP utilizes Chow-Liu trees algorithm in order to compute a score for the co-occurring visual words. FAB-MAP enables to recover the vehicle current position from reference key images. The hybrid RatSLAM/FAB-MAP system [14] has shown that mapping can be performed even in difficult outdoor conditions when the environment appearance varies due to changes in illumination and structure. However, the results clearly show that the map diverges when there are long sections of path where no matches occur.

In [15], the authors propose an image retrieval framework for image-based localization. They introduce the concept of co-occurrence matrix that encodes both visual words (dictionary atoms) and spatial words. The image matching is then scored according to the matches found between a request image and each image in the reference database. The similarity score between a query image and a database location is equal to the number of correspondences in the spatially-consistent group.

Paper contribution

Most of image retrieval techniques for image-based localization use feature points such as SIFT and SURF [16] and many frameworks have been built upon the use of such feature points. However, much less attention was given to block descriptor. These visual features as such are convenient under small variations in lighting and orientation. However, under complex and similar environments, image mismatch can occur. In other words, a query image, taken from the current location might be matched to the content of one or more reference images in a large database. This would eventually lead to the fail of the mapping result.

In this paper, we provide an empirical study on a particular

image descriptor: Covariance matrix. The paper has two main contributions. Firstly, applying the descriptor on image-based localization. Secondly, we provide a quantitative evaluation using several classifiers.

Global Descriptors

Covariance descriptors

The region covariance descriptor was firstly proposed by Tuzel *et al.* in [17]. The idea is to represent a feature distribution using its sample covariance matrix. Let I be a $W \times H$ one-dimensional intensity or three-dimensional color image, and F be the $W \times H \times d$ dimensional feature image extracted from I. $F(x;y) = \psi(I;x;y)$ where ψ is a function extracting image features such as intensity, color, gradients, and filter responses, etc. For a given region $R \in I$, let $\{\mathbf{f}_i\}_{i=1\ldots N}$ denote the d-dimensional feature points obtained by ψ within R. N denotes the number of pixels in R. The region R is then represented by a $d \times d$ covariance matrix:

$$\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{f}_i - \mu)(\mathbf{f}_i - \mu)^T$$

where μ is the mean vector of $\{\mathbf{f}_i\}_{i=1\ldots N}$.

Whenever the region R is rectangular, fast calculation of covariance matrices can be provided by the intermediate representation called integral image. With this representation, covariance descriptor of any rectangular region can be computed within constant time.

Covariance matrices do not lie on the Euclidean space. Therefore, an arithmetic subtraction of two matrices would not measure the distance of the corresponding regions. In fact, nonsingular covariance matrices are Symmetric Positive Definite (SPD) and lie on a connected Riemannian manifold. Accordingly, Riemannian metrics should be used for computing distance and mean of covariance matrices.

Under the Log-Euclidean Riemannian metric, distance measure between covariance matrices preserves much of the natural properties of the affine-invariant metric while being computationally straightforward: the distance between two covariance matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ is given by,

$$d(\mathbf{C}_1,\mathbf{C}_2) = ||\log(\mathbf{C}_1) - \log(\mathbf{C}_2)||$$

where $||.||$ is the vector norm operator (ℓ_1 or ℓ_2) and $\log(\mathbf{C})$ is the matrix logarithm of the square matrix $\mathbf{C}$.

Thus, every image region, R, can be characterized by $\log(\mathbf{C}_R)$. Since this is a symmetric matrix, then the feature vector can be described by a $d \times (d+1)/2$ where d is the number of channels used to build the covariance matrix.

Another way used to compute the distance between two covariance matrices under the affine-invariant Riemannian metric is

$$\rho(\mathbf{C}_1,\mathbf{C}_2) = \sqrt{\sum_{i=1}^{d}\log^2\lambda_i(\mathbf{C}_1,\mathbf{C}_2)}$$

where $\{\lambda_i(\mathbf{C}_1,\mathbf{C}_2)\}_{i=1\ldots d}$ are the generalized eigenvalues of $\mathbf{C}_1$ and $\mathbf{C}_2$ computed from

$$\lambda_i\mathbf{C}_1 x_i - \mathbf{C}_2 x_i = 0,\ i = 1\ldots d$$

and $x_i \neq 0$ are the corresponding generalized eigenvectors.

Block-based descriptors

In order to increase the contribution of a bigger part of their regions in the resulting descriptor, input images are broken up into blocks. Then, for each block, the features extraction process is applied. The final descriptor of the image is obtained by concatenating the features vectors of all the constituent blocks. In our work, we tested three blocks decomposition methods that we denote by MB1, MB2 and MB3. These methods were proposed in [17], [18] and [19] respectively. Examples of blocks decomposition are depicted in Figure 1.

(a) Multi-blocks decomposition MB1: overlapping regions case

(b) Multi-blocks decomposition MB2: multiple-patches case

(c) Multi-blocks decomposition MB3: contiguous regions case

Figure 1. *Illustration of the multi-blocks image representation. The sought features correspond to the concatenation of those of the blocks resulting from the decomposition of the initial image into sub-regions.*

Experimental Setup

Generating the image database

The retrieval performance depends on the specificity or distinctiveness of the scene. Naturally, not all scenes in real world scenario satisfy this requirement. Fortunately, an autonomous system is usually collecting a stream of images at a fixed rate either in time domain or in space domain. Thus, to build the set of

reference images, we adopt the Google Street View service [20] that allows having panoramic views from positions along many streets in the world. A virtual tour was made in the city of Paris using Street View and a set of 400 reference color images that depict building facades and other structures was collected with an average of 4 images per faade. The size of each image is 512×1024 pixels. Figures 2 and 3 show some examples of the reference dataset for different and same faade respectively.

Figure 2. *Examples of different facades in the reference dataset.*

Figure 3. *Examples of the same facade in the reference dataset.*

For testing purpose, we use the reference images in order to generate four test sets in each of them the similarity with the reference set is decreased. The test images correspond to zoomed, cropped and resized views of the reference images. An example of a reference image and the corresponding test images is shown in Figure 4.

Descriptor extraction details

The Covariance matrix corresponding to a given image was generated using $d = 14$ channels extracted from it. These channels are respectively x-abscissas, y-ordinates, RGB components, HSV components, image gradient with respect to x, y and xy. The size of the descriptor vector is given by $d(d+1)/2$ i.e. 105 features.

Certainly when dealing with a block-based descriptors case, the size of the features vector will be multiplied by the number of blocks into which the image is divided.

Image matching

In order to match the test images with the set of reference images, two different classifiers were used: the classic k Nearest Neighbor (kNN) [21] and the Partial Least Square (PLS) [22].

Figure 4. *Example of a reference image and the ad-hoc test images corresponding to a zoom factor of 5, 10, 15 and 20% respectively.*

The Partial Least Squares (PLS) classifier or regressor is a statistical method that retrieves relations between groups of observed variables X and Y through the use of latent variables. It is a powerful statistical tool which can simultaneously perform dimensionality reduction and classification/regression. It estimates new predictor variables, known as components, as linear combinations of the original variables, with consideration of the observed output values. PLS is also extended to deal with non-linear cases [23]. In our work, the number of latent components is fixed to 50. Note that each test image is matched against all images in the reference image set.

Experiments performed

The localization framework used to perform the experiments is built on two main processing modules which are the descriptor computation and the classification ones. For comparison purposes, several combinations/scenarios are considered for the four test sets. Table 1 summarizes the set of combinations tested during the experiments (each row corresponds to a combination).

In another group of experiments, in order to test the effects of eventual occlusions that may take place during localization, e.g. due to the passage of a car near the faade being captured, a rectangle of a size close to the one of a car has been randomly added to the test images before running the recognition process (Figure 5).

Experimental Results

Table 2 shows, for each tested combination, and for the four test sets, the rate of successful matching corresponding to covariance descriptors. The best performances are shown in bold. The test sets correspond to increasing zoom levels ranging from 5% to 20%. As expected, the recognition rate decreases as the zoom level of the images increases

Table 1. The combinations distance-block layouts for covariance descriptors.

Distance	Blocks	Classifier
Log difference - L1	Mono	NN PLSR
Log difference - L1	Multi MB2	NN PLSR
Log difference - L2	Mono	NN PLSR
Log difference - L2	Multi MB2	NN PLSR
Log difference - L2	Multi MB3	NN PLSR
Eigenvalue-based	Mono	NN PLSR
Eigenvalue-based	Multi MB1	NN PLSR

Figure 5. *Simulating occlusions by adding a rectangle to the test images.*

At low zoom level, both classifiers give maximum recognition rates with no significant differences. However, for a given classifier, the distance measures and blocks decomposition seem to have a great impact on the performance. Indeed, Table 2 shows that the KNN classifier based on multiblock (MB2) combination achieves a maximum accuracy of about 95%. While, for the monoblock case the maximum performance was only 89%.

When the zoom level increases, the difference in maximum performances of classifiers also increases. At 20% zoom level (the most challenging case), the PLSR classifier reaches a maximum accuracy of about 95% which is 2% less than the accuracy achieved at 5% zoom level. For this classifier, the log-difference L2 distance and the multi-block MB3 decomposition give the best performance. This tends to confirm that the MB3 layout was better than MB1 and MB2 (two known decompositions for covariance descriptors). It seems that the PLSR is less affected by the zoom increase when compared with the KNN classifier. In fact, at 20% zoom level, the KNN classifier experiences a 27.5% decrease of the maximum accuracy achieved at 5% zoom level. We can also observe that the eigenvalue based distance gave better performance than the Log difference distance for the mono-block case.

Tables 3 shows the results of the same tested combinations while simulating occlusion in the test image sets. Similarly to the discussion made above, the classification rates given by the two classifiers, for a given combination, are close to each other and decrease with the increasing of the zoom level of the test images while being smaller than the ones obtained in the absence of occlusions. The difference in values obtained with and without simulated occlusions depends on the classifier used in the tested combination. This difference is negligible for PLSR (less than 1% on average) and varies between 3.7% and 8.2% in average for NN. The combinations that led to the best performance when adding occlusion were the same as in the case without occlusion.

On the other hand, in order to investigate the influence of the number of the ranking of reference images on the rate during classification, several ranks are used by our framework. The cumulative score obtained for the main combinations (mono and multi-blocks, L1 and Chi-Squared distances) are depicted on Figures 6 and 7.

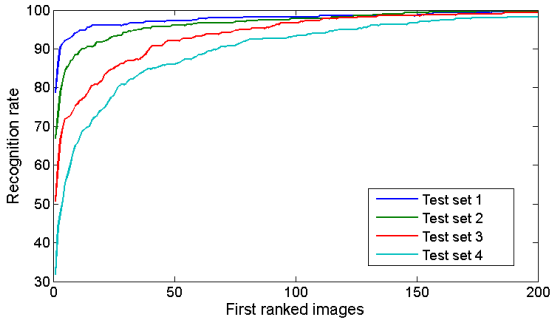

Figure 6. *Cumulative score (NN) for covariance monoblock log difference.*

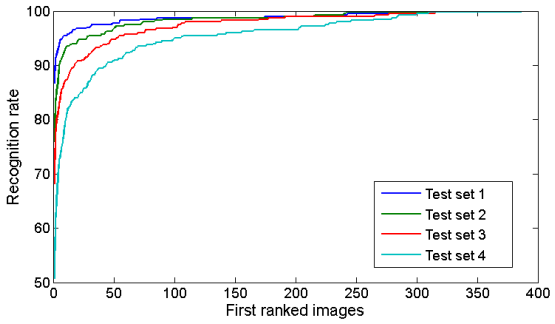

Figure 7. *Cumulative score (NN) for covariance multiblocks log difference.*

Conclusion

In this paper, we have presented an empirical study on a particular image descriptors and its use for visual place recognition in urban scenes. The study showed that the use of multi-block based features can enhance the discrimination of the obtained final descriptor. The study was limited to a relatively small reference dataset. This is justified by the fact that in real application the system can have partial information on the neighborhood so the automatic image retrieval does not need to do a search among millions of images. The conducted study showed that the use of multi-block based covariance descriptor with the PLS classifier can lead to good and robust results. Future work can investigate the use of deep learning paradigms in order to simultaneously solve the feature extraction and the matching process.

Table 2. Correct classification rate (%) obtained with covariance descriptors for different tested distances and different kinds of image decomposition (mono-block and multi-block cases).

	Test set 1 (5% zoom)		Test set 2 (10% zoom)		Test set 3 (15% zoom)		Test set 4 (20% zoom)	
Distance and decomposition	**NN**	**PLSR**	**NN**	**PLSR**	**NN**	**PLSR**	**NN**	**PLSR**
Log difference - L2 (monoblock)	78.8	75.8	67.0	70.0	50.8	62.8	31.8	53.8
Log difference - L1 (monoblock)	89.0		80.3		73.3		54.8	
Eigenvalue-based (monoblock)	91.5		84.8		**87.3**		65.8	
Log difference - L2 (multiblock MB2)	86.8	89.5	76.0	85.3	68.2	76.5	50.7	53.8
Log difference - L1 (multiblock MB2)	**95.3**		**92.0**		83.0		**67.8**	
Eigenvalue-based (multiblock MB1)	87.0	96.8	80.3	94.5	68.5	91.3	51.3	87.8
Log difference - L2 (multiblock MB3)	92.8	**97.5**	87.0	**97.5**	81.3	**96.5**	65.8	**95.3**

Table 3. Correct classification rate (%) obtained with covariance descriptors, in the presence of simulated occlusions, for different tested distances and different kinds of image decomposition (mono-block and multi-block cases).

	Test set 1 (5% zoom)		Test set 2 (10% zoom)		Test set 3 (15% zoom)		Test set 4 (20% zoom)	
Distance and decomposition	**NN**	**PLSR**	**NN**	**PLSR**	**NN**	**PLSR**	**NN**	**PLSR**
Log difference - L2 (monoblock)	72.3	73.3	61.0	68.5	43.2	59.5	27.8	50.8
Log difference - L1 (monoblock)	86.8		76.0		68.3		50.8	
Eigenvalue-based (monoblock)	85.8		77.3		76.8		59.0	
Log difference - L2 (multiblock MB2)	78.8	89.3	67.0	82.5	50.8	73.3	31.8	62.3
Log difference - L1 (multiblock MB2)	**94.8**		**89.8**		72.8		**61.5**	
Eigenvalue-based (multiblock MB1)	84.0	95.5	75.5	94.5	63.3	90.3	47.8	86.0
Log difference - L2 (multiblock MB3)	92.6	**97.5**	85.8	**97.5**	**79.8**	**95.8**	60.5	**94.3**

References

[1] G. Baatz, K. Koser, D. Chen, R. Grzeszczuk, and M. Pollefeys. Handling urban location recognition as a 2d homothetic problem. In *European Conference on Computer Vision*, 2010.

[2] M. Brown G. Schindler and R. Szeliski. City-scale location recognition. In *IEEE Conferene on Computer Vision and Pattern Recognition*, 2007.

[3] Y. Li, N. Snavely, and D. P. Huttenlocher. Location recognition using prioritized feature matching. In *European Conference on Computer Vision*, 2010.

[4] D. Robertson and R. Cipolla. An image-based system for urban navigation. In *British Machine Vision Conference*, 2006.

[5] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *International Conference on Computer Vision*, 2011.

[6] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *European Conference on Computer Vision*, 2010.

[7] W. Zhang and J. Kosecka. Image based localization in urban environments. In *3DPVT*, 2006.

[8] L. Meier, P. Tanskanen, F. Fraundorfer, and M. Pollefeys. Pixhawk: A system for autonomous flight using onboard computer vision. In *ICRA*, 2011.

[9] E. Johns and G.Z. Yang. Localization independent of location based on place recognition and GPS observations. In *IEEE/SICE International Symposium on System Integration*, pages 43–48, 2012.

[10] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *European Conference on Computer Vision*, 2010.

[11] M. Perdoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *IEEE Conferene on Computer Vision and Pattern Recognition*, 2009.

[12] R. Raguram, C. Wu, J.-M. Frahm, and S. Lazebnik. Modeling and recognition of landmark image collections using iconic scene graphs. *Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs*, 95(3), 2011.

[13] M. Cummins and P. Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *International Journal of Robotics Research*, 27(6):647–665, 2008.

[14] A.J. Glover, W.P. Maddern, M.J. Milford, and G.F. Wyeth. FAB-MAP + RatSLAM: Appearance-based slam for multiple times of day. In *IEEE International Conference on Robotics and Automation*, pages 3507–3512, May 2010.

[15] E. Johns and G.Z. Yang. Feature co-occurrence maps: Appearance-based localisation throughout the day. In *IEEE International Conference on Robotics and Automation*, pages 3212–3218, 2013.

[16] C. Valgren and A. Lilienthal. Sift, surf, and seasons: Long-term outdoor localization using local features. In *Proc. European Conference on Mobile Robots*, 2007.

[17] F. Porikli, O. Tuzel, and P. Meer. A fast descriptor for detection and classification. In *European Conf. on Computer Vision*, pages 589–600, 2006.

[18] L. Qin. *Online machine learning methods for visual tracking*. PhD thesis, Universit de Technologie de Troyes - France, 2014.

[19] V. Takala, T. Ahonen, and M. Pietikiinen. Block-based methods for image retrieval using local binary patterns. In *Image Analysis, SCIA*, volume LNCS, 3540, 2005.

[20] A. Charles. Where the streets all have google's name. *The Guardian.com (US ed.)*, March 2009.

[21] E. Fix and J.L. Hodges. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[22] R. Rosipal and N. Kramer. *Subspace, Latent Structure and Feature Selection Techniques*, chapter Overview and recent advances in partial least squares, pages 34–51. Springer, 2006.

[23] Nicole Krämer and Mikio L. Braun. Kernelizing pls, degrees of freedom, and efficient model selection. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 441–448, New York, NY, USA, 2007. ACM.