Bifocal search: embedding context in local descriptors

Fabrizio Falchi

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo",
via G. Moruzzi 1, Pisa 56124, Italy
fabrizio.falchi@isti.cnr.it

Abstract. Local descriptors are state-of-the-art of representing low-level visual information in object recognition. Because of their effectiveness, they are also largely used in content-based image retrieval whenever the query visually express a specific object to be retrieved between the images in the archive. Given that searching for the local descriptors can be very costly, many recent works have proposed to encode the local descriptors in a compact representation. In this paper, we propose to embed the aggregated information in the local descriptors in order to achieve higher effectiveness. The experimental results, obtained on a largely used public dataset, reveal the potential of the approach. Even if we only tested our approach in a content-based image retrieval scenario, the idea of combining aggregated and local information is general and could be applied in other similarity search tasks. We call the proposed approach bifocal searching because of the similarity with bifocal eyeglasses which have two parts with different focal lengths.

Keywords: permutation-based indexing, similarity search, content based image retrieval

1 Introduction

Content-based image search has received increasing interest in recent years. While in the last century it was only possible to search for global features (e.g., colors, texture, etc...), the technologies developed in the Computer Vision field during the last few years, allow searching for particular objects. Starting with the pioneering paper from David Lowe [14] defining the Scale Invariant Features Transform (SIFT), local features extracted from stable regions are at the base of almost all of the proposed approaches. The high effectiveness of object recognition obtained by matching local features (e.g., SIFT) and geometric verification typically performed using Randon Sample Consensus (RANSAC), has the drawback of an high computation cost and very low scalability. In fact, for each local descriptors in the query image (typically 1,000), a similarity search has to be performed in the whole dataset.

In 2003, particular object searching on a large scale became possible thanks to the bag-of-features or bag-of-(visual)words (BoF) approach presented by Sivic

and Zisserman [19]. In terms of effectiveness, BoF largely outperformed the other methods. The relevant lost in effectiveness has been considered acceptable in many practical cases. Moreover, many extension to the basic BoF approach have been proposed [11]. Very recently, novel aggregation approaches [13] have been proposed that significantly outperform BoF both in terms of efficiency and effectiveness. The most famous are Vector of Locally Aggregated Descriptors (VLAD) [9] and Fisher Vectors (FV) [16]. The effectiveness obtained by these novel aggregations techniques is so high that they even outperform the more costly approaches based on local features matching combined with sophisticated geometric verification.

The impressive results obtained by the most recent aggregations approaches suggest that local features similarity based searching should be revised. Our intuition is that the information about the context in which the local features appear is more relevant than the actual similarity between any two local descriptors. However, we believe that the local features searching can be improved by adding the information about the context to the local descriptor.

In this paper, we propose to focus at the same time on both local and aggregated information when searching for local features. We call this approach bifocal searching because of the similarity between this approach and the bifocal eyeglasses. We propose to compare any two local descriptors considering not only the local information but also the aggregated information of all the local descriptors in the whole image. The intuition is that the probability of two descriptors to be a real match is related not only to the similarity between two descriptors but also to the overall similarity of the images. We tested our proposal adding the VLAD aggregation information to the SIFT local descriptors obtaining larger local descriptors that comparable using the Euclidean distance. Thus, the resulting extended local descriptors can be indexed by both general metric data structures and specific vector based indexes. We call this new feature obtained combining local and aggregated information the bifocal feature. In fact, we don't need the data structure to have the knowledge that the bifocal descriptor is actually composed of two parts because it can be compared as a whole resulting in a combination of local and global similarity.

In the experimental section, we show that effectiveness of the proposed approach outperform both traditional local features matching approaches and state-of-the-art aggregation techniques even when reordering of the results with geometric verification is considered.

2 Backgrounds and Related Work

2.1 Local Features Matching

Local descriptors are fixed size vector automatically extracted from relevant regions automatically detected in an image typically in the order of thousands. Various detection techniques [21] have been proposed in order to obtain regions invariant to projective transformations and variou local features have been defined to describe the visual content of each region [15].

We define I_i as the set of local descriptors $l_{i,j}$ extracted from image i, i.e., $I_i = l_{i,1} \dots l_{i,n}$, where $n = |I_i|$ is the number of descriptors which largely depend on the complexity of the image. A function of dissimilarity $d_l(l_{i,j}, l_{i',j'})$ between any two local descriptors is necessary in order to perform any kind of match. For SIFT [14] and for many other local features as Speeded Up Robust Features (SURF) [5], the Euclidean distance is typically used.

For each local features $l_{q,j}$ in the query image q, we define $m(l_{q,j}, I_o)$, if it exists, as the matching local descriptors in a generic image o in the dataset. Almost all the approaches proposed in the literature select the most similar local descriptor in o with respect to $l_{q,j}$ as the candidate match [14]. We indicate this nearest neighbor as $NN_1(l_{q_j}, I_o)$. The distance between the query descriptor and its nearest neighbor in o is usually considered to filter out wrong matches defining a maximum distance [7]. We define this threshold as t. Thus,

$$m(l_{q,j}, I_o) = \begin{cases} NN_1(l_{q_j}, I_o), & \text{if } d_l(l_{q,j}, NN_1(l_{q_j}, I_o)) \leq t \\ \text{null}, & \text{otherwise} \end{cases} \quad (1)$$

In [14], it was suggested to consider the ration between the distances of $n_1(l_{q_j}, I_o)$ and $NN_2(l_{q_j}, I_o)$ as a confidence on the match. However, this is intended to rely on reliable matches in order to perform geometric verification.

In the following, for each local descriptor l_{q_i} in the query image we perform a range search with range t in the whole datasets and we consider the first nearest neighbor in any image o, if it exists, as a matching descriptor. A laregely used measure of similarity between any two images q and o is given by the percentage of local descriptors $l_{q,j}$ in the query image q that do have a match in o.

This set of candidate match can be refined considering the information related to the region from which the local descriptor was extracted as explained in the following section.

2.2 Geometric Verification

Many local descriptors (e.g., SIFT) report information about the original position, orientation and size of the region from which they were extracted. This information is necessary in order to perform geometric verification in object recognition and for augmented reality application.

In [14], Lowe proposed *Hough Transform* to group local features matches between two images that have an agreement in terms of relative rotation and scale. While the *Hough Transform* is usually followed by an affine or homogrpahy transformation estimation, the size of the largest cluster built can be used as a measure of similarity.

The Random Sample Consensus (RANSAC) [6] approach is largely used in order to estimate a projective trasformation able to map the matching regions of a given image on top of the other. Two are the transformation that are typically considered: *homography* and *affine*. While the homography trasformation is more genera,l the affine trasformation is typically more reliable. In fact, very unlikely projective transformation can be expressed as homographies. Filtering out these

noisy results can be difficult [3]. In the experiments we report the results obtained by both.

Please note that geometric verification is not scalable because it can not be indexed. Thus, geometric verification is typically considered when reranking top-k results obtained with a scalable approach. When searching for matching local features, for instance, the geometric verification is only performed between the query image q and the dataset image o for which there are the largest amount of candidate matches. Thus, the images o in the dataset are typically ordered according to the number of matches as defined in Equation 1.

2.3 Aggregation Techniques

The first aggregation technique proposed for local features was inspired by traditional text retrieval. The bag-of-features (BoF) representation [19] groups local descriptors by defining a codebook of n_w visual words $C = c, 1 \dots c_{n_w}$ usually obtained by clustering (e.g., with k-means) the local descriptors of the whole dataset. The size of the codebook is typically in the order of hundred thousands. Each local descriptor is then assigned to the closest centroid in the vocabulary. The BoF representation is defined as the histogram of the assignment of all the descriptors $l_{i,j}$ in image i to the visual words. Similarity between two BoF representation is typically evaluated using the *cosine* function [19]. Moreover, TF-IDF weighting scheme is usually adopted as for the traditional bag-of-words approach used in text retrieval [20].

After many years in which BoF remain the state-of-the-art approach for large scale searching, in 2010 Vector of Locally Aggregated Descriptors (VLAD) [9] and Fisher Vectors (FV) [16] were proposed. The Vector of Locally Aggregated Descriptors (VLAD) aggregates descriptors on a locality criterion in the feature space. A small codebook (i.e., hundreds) of visual words selected with k-means is used as in BoF. For each image i, VLAD accumulate the differences between the local descriptors $l_{i,j}$ and the nearest centroids $NN_1(l_{i,j}, C)$. The VLAD representation is then obtained by concatenating the sum of the residuals v_y and subsequently L_2-normalizing. Formally, the aggregation function $a_{_{VLAD}}(I_i)$ of the local descriptors I_i extracted from image i is defined as:

$$a_{_{VLAD}}(I_i) = \frac{(v_1 \dots v_{n_w})}{||(v_1 \dots v_{n_w})||}, \qquad v_y = \sum_{NN_1(l_{i,j},C)=c_y} l_{i,j} - c_y$$

Please note that the dimension of the VLAD representation is n_w times the dimension of the local descriptors (e.g., 128 for the SIFT). Thus, Principal Component Anlysis [1] is typically used in order to improve efficiency. Experiments [4] showed that PCA also improve effectiveness.

The most common approach for selecting words for the aggregation techniques vocabulary is k-Means. However, recent pivots selection approaches [2] presented in the context of similarity search in metric spaces could be tested.

2.4 Bifocal Searching

In this section, we present our proposal to combine the approaches presented in Section 2.1 and Section 2.3. State-of-the-art aggregations techniques typically outperform local features matching techniques. In other terms, a similarity measure between two image based on the number of matches found is less informative than the distance between the aggregated information. In order to improve effectiveness, we propose to revise the similarity function d_l used to combine local features and aggregated information. To this goal we define the similarity between two local descriptors as the weighted sum of the standard local descriptors distance d_l and the distance d_a between the aggregated information. Formally, we define the bifocal distance between two local descriptors $l_{q,j}$ and $l_{o,y}$ extracted from images q and o as:

$$d_b(l_{q,j}, l_{o,y}) = w_l * d_l(l_{q,j}, l_{o,y}) + w_a * d_a(a(I_q), a(I_o))$$

where a_q and o_q are the vector resulting from the aggregation of the local descriptors (e.g., VLAD).

It is worth to note that whenever a threshold on the dissimilarity between a query local descriptors and its nearest neighbor in another image is used to filter out false matches (see Equation 1), the bifocal distance results in varying the threshold according to the similarity between the aggregated information. In fact,

$$d_b(l_{q,j}, l_{o,y}) \leq t \iff d_l(l_{q,j}, l_{o,y}) \leq t - \frac{w_a}{w_l} * d_a(a(I_q), a(I_o)) \tag{2}$$

The intuition is that, given a distance between two local descriptors, the probability for them to match is higher if the aggregated information is similar. Thus, by using Equation 2, we can obtain a varying threshold for descriptor matching. Moreover, from Equation 2, we have

$$d_b(l_{q,j}, l_{o,y}) \leq t \Rightarrow d_l(l_{q,j}, l_{o,y}) \leq \frac{t}{w_l}, \quad d_a(a(I_q), a(I_o)) \leq \frac{t}{w_a} \tag{3}$$

where t/w_l and t/w_a can be interpreted as excluding distances. In fact, whenever the local descriptor distance d_l is above t/w_l or the aggregated distance d_a is above t/w_a, the two descriptors don't match. Given that the t, w_l and w_a parameters actually control two levels of freedom, in the experiments we fixed $t = 1$ without loss of generality.

In our experiments we used SIFT local features and VLAD aggregation in combination with Principal Component Analysis to reduce the dimensionality of the aggregated vector. In this case, both d_l and d_a are the Euclidean distance (L_2). Thus, we can combine the local and global information in one vector before applying the distance, i.e.,

$$d_b(l_{q,j}, l_{o,y}) = L_2(b_{q,j}, b_{o,y}), \qquad b_{i,j} = (w_l * l_{i,j}, w_a * a(I_i))$$

Table 1. mAP obtained using the BoW approach for various number of visual words.

#VW	cos	TF-IDF
100	38.1	38.4
200	39.4	39.8
500	40.4	40.5
1,000	40.3	39.8
2,000	39.6	39.9
5,000	38.6	40.8
10,000	40.2	41.7
20,000	43.8	44.2
50,000	46.7	48.0
100,000	51.9	53.4
200,000	54.2	55.7

where $b_{i,j}$ is the weighted concatenation of the local descriptor $l_{i,j}$ and the aggregation $a(I_i)$ of all the descriptors in the same image i. We call b_{i_j} the bifocal descriptor.

Please note, that the definition of b_{i_j} allows indexing with any metric data structure [22], vector based indexes and even with specific approach for L_2 distance [18]. A range search of query bifocal descriptors over the whole datase results in a variable threshold over local descriptors distance d_l as expressed in Equation 3. The price to pay is the replication of the aggregated information $a(I_i)$ of an image i in all the bifocal descriptor.

3 Experiments

In the experiments we focused on effectiveness. Thus, even if all the approaches can be used in combination with efficient and scalable data structures, we only performed sequential scan.

We performed experiments on the INRIA Holidays [12, 13] collection of 1,491 public available images largely used by the Computer Vision community. For this collection a ground-truth consisting of 500 queries and expected results is also public available together with SIFT descriptors and set of visual words (vocabularies) obtained using k-Means [1]. The quality of the retrieved images on this is typically evaluated [17, 8, 16, 13] by using the mean Average Precision (mAP), which represents the area below the precision and recall curve.

In Table 1, we report the mAP obtained by the BoW approach discussed in Section 2.3 using the *cosine* similarity alone and in combination with the TF-IDF weighting for various visual words vocabulary size. The vocabularies, public available together with the INRIA Holidays dataset images, have been created performing k-means on a distinct largest dataset. As expected, the quality of the results is largely affected by the size of the vocabulary. Please note, that the overall number of SIFT extracted from the dataset is about 1 million. Thus, a

[1] `http://lear.inrialpes.fr/~jegou/data.php`

200,000 words dictionary is already very large. The TF-IDF weighting improve the quality of the results only marginally.

Table 2. mAP obtained using the VLAD approach for various number of centroids and distinct number of principal components selected.

		#principal components					
#centroids	**full**	16	32	64	128	256	512
64	54.9	45.6	51.7	57.2	61.6	62.0	58.0
128	56.6	45.8	51.4	58.2	64.1	65.8	60.9
256	59.2	44.6	52.0	58.8	65.0	**67.5**	61.7

In Table 2, we report the mAP obtained by the VLAD approach (Section 2.3) for various number of centroids and varying the number of principal components selected performing PCA. As reported in other papers [10], PCA not only reduce the size of the VLAD vector, but also helps in improving effectiveness especially for number of principal components near to 128. The comparison of the results obtained by the VLAD approach wit the ones obtained by the BoW approach (Table 1, reveals that VLAD better describe the overall distribution of the local descriptors in the image. In Figure 1, we report the recall of good and bad results varying the range of a search considering when using the VLAD aggregation with 128 centroids and 128 principal components. The figure shows that there are no good results above distance 1.0 while bad results start appearing at 0.6 where more than half of the good results have already been found. With a range of 0.8 about 95% of the good results are retrieved but also 40% bad results are present.

For testing the bifocal approach, we decided to use the VLAD with 128 centroids and 128 principal components given the results reported in Table 2. With 128 principal components we obtained a bifocal descriptors composed of 256 dimensions (128 for SIFT and 128 for VLAD). We used 128 centroids instead of 256 that obtained the best results because for 128 principal components the difference in effectiveness do not probably justify the extra cost of using 256 centroids. Please note that in the following, when comparing our approach with the standard VLAD, we will used the 256 centroids and 256 principal components settings in order to compare with the best VLAD settings. In order to perform bifocal search, we have to define w_g and w_a. As reported in Section 2.4, we use threshold $t = 1.0$ for filtering good local matches. In this case, $1/w_g$ and $1/w_a$ are the excluding distance for the local and aggregated distance respectively. Given the results reported in Figure 1, we decide to test values of $1/w_a$ between 1 and 1.4. To reduce the degree of freedom of the problem we fixed $1/w_g$ to 0.06. In Figure 2, we report the actual local descriptor matching threshold $1 - \frac{w_a}{w_l} * d_a(a(I_q), a(I_o))$ (see Section 2.4) as a function of the aggregated distance d_a. Following the previous discussion related to Figure 1, the most problematic range is between 0.6 and 0.8. Above 0.8, we almost have only bad results. Below 0.6, we almost have only good results. Thus, the goal of the bifocal approach is

varying the resulting local descriptor threshold allowing to better discriminate results in this aggregated distance range.

Table 3. Results obtained using the bifocal approach

$1/w_l$	$1/w_a$	**mAP**	**avg matches**
0.06	1.0	**70.2**	2.1
0.06	1.2	69.3	6.6
0.06	1.4	68.1	19.6

In Table 3, we report the mAP obtained by the bifocal approach by using fixed threshold of 1.0 for the bifocal distance d_b and varying w_a (we kept w_l fixed). The measure of similarity between two image is the percentage of local descriptors $l_{q,j}$ that have matches in a given image o in the dataset, i.e., it exists at least one local descriptor in o at distance $d_b(l_{q,j}, l_{o,y}) \leq 1.0$. Given that the number of matches per query local descriptors in the whole dataset is significantly affected by the weights, we also reported this information in the table.

Table 4. Results obtained using the local descriptors matching approach

t	**mAP**	**avg matches**
0.03	59.6	45.5
0.02	56.8	5.4

In Table 4 we report the same information for the standard local descriptor matching approach (Section 2.1. The results show that the bifocal approach is much better.

Table 5 show the results obtained reordering top results by using the geometric consistency verification approaches discussed in Section 2.2. Please note that while for the local features and bifocal approaches the matches to be filtered are available after the range search phase, for VLAD we actually had to compare all the local descriptors in the query image with all the local descriptors in image to be reordered to get candidate matches that the aggregated information can not given. The VLAD results were obtained with 256 centroids and 256 principal components.

Before writing this paper, it was a surprise to notice that even when reordering, the results obtained by the VLAD approach outperform the local features based one. We expected the local features matching to be more appropriated in identifying images with relevant matches that can survive the geometric verification. We believe that the VLAD plus geometric verification approaches combine a global view with the use of local descriptors in the reordering phase. This is not possible with tradition local features based approach. Our intuition was that combining the aggregated and local view (bifocal search) could outperform both

approaches in terms of effectiveness. The results confirm our intuition. In fact, for all the setting we tested the bifocal approaches obtained the best results.

Table 5. mAP obtained by the various approach performing reordering of top results with various type of geometric matching approaches.

	#results reordered		
method	8	16	32
LF+Hough	60.7	62.7	64.2
LF+Affine	61.0	63.6	65.1
LF+Homography	59.5	62.1	63.1
VLAD+Hough	67.0	69.8	70.3
VLAD+Affine	68.3	69.8	70.7
VLAD+Homography	68.4	68.0	68.7
Bifocal (0.06,1.0) +Hough	**71.2**	72.9	72.8
Bifocal (0.06,1.0) +Affine	70.9	72.9	72.7
Bifocal (0.06,1.0) +Homography	69.5	70.9	70.2
Bifocal (0.06,1.2) +Hough	70.4	71.8	73.1
Bifocal (0.06,1.2) +Affine	71.1	**73.1**	**75.2**
Bifocal (0.06,1.2) +Homography	70.0	71.9	72.4
Bifocal (0.06,1.4) +Hough	69.5	71.5	72.4
Bifocal (0.06,1.4) +Affine	71.0	73.2	74.4
Bifocal (0.06,1.4) +Homography	68.5	70.1	71.3

4 Conclusions and Future Work

In this paper we propose to combine local features matching techniques with state-of-the-art aggregation approaches in order to achieve higher effectiveness. We defined a bifocal search approach that combine local and aggregated information in the same vector (the bifocal feature) in order to obtain a variable threshold for matching local features using a fixed bifocal threshold. This allows to index bifocal descriptors in order to efficiently and effectively match local descriptors. The results show that our approach outperform both local features and aggregated approaches in terms of effectiveness. Because of the high efficiency, the aggregation techniques during the search execution they are still preferable on large scale scenarios. However, whenever high accuracy is needed we prove that bifocal searching can outperform traditional local features matching. The bifocal approach could be adopted in any similarity search scenario in which for a given object a set of descriptor of the same features are given. Aggregations approaches are still under investigation and even better techniques than VLAD already exist (e.g., Fisher Vectors) and probably will come in the future. Our bifocal approach will benefit from these research results adding even better aggregated information to the local descriptions.

References

1. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2(4), 433–459 (2010)
2. Amato, G., Esuli, A., Falchi, F.: A comparison of pivot selection techniques for permutation-based indexing. Information Systems 52, 176 – 188 (2015), http://www.sciencedirect.com/science/article/pii/S0306437915000204, special Issue on Selected Papers from SISAP 2013
3. Amato, G., Falchi, F., Gennaro, C.: Geometric consistency checks for knn based image classification relying on local features. In: SISAP '11: Fourth International Conference on Similarity Search and Applications, SISAP 2011, Lipari Island, Italy, June 30 - July 01, 2011. pp. 81–88. ACM (2011)
4. Arandjelović, R., Zisserman, A.: All about VLAD. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
5. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: In ECCV. pp. 404–417 (2006)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), 381–395 (1981)
7. Homola, T., Dohnal, V., Zezula, P.: Sub-image searching through intersection of local descriptors. In: Proceedings of the Third International Conference on SImilarity Search and APplications. pp. 127–128. SISAP '10, ACM, New York, NY, USA (2010)
8. Jegou, H., Douze, M., Schmid, C.: Packing bag-of-features. In: Computer Vision, 2009 IEEE 12th International Conference on. pp. 2357 –2364 (29 2009-oct 2 2009)
9. Jegou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 3304–3311 (June 2010)
10. Jégou, H., Douze, M., Sánchez, J., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 3304 –3311 (june 2010)
11. Jégou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. Int. J. Comput. Vision 87, 316–336 (May 2010)
12. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: IEEE Conference on Computer Vision & Pattern Recognition. pp. 3304–3311 (jun 2010)
13. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence (Sep 2012), qUAERO
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
15. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27(10), 1615–1630 (2005)
16. Perronnin, F., Liu, Y., Sanchez, J., Poirier, H.: Large-scale image retrieval with compressed fisher vectors. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 3384 –3391 (june 2010)
17. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2007)

18. Samet, H.: Foundations of Multidimensional and Metric Data Structures. Computer Graphics and Geometric Modeling, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)
19. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2. pp. 1470–. ICCV '03, IEEE Computer Society, Washington, DC, USA (2003), http://portal.acm.org/citation.cfm?id=946247.946751
20. Tirilly, P., Claveau, V., Gros, P.: Distances and weighting schemes for bag of visual words image retrieval. In: Proceedings of the international conference on Multimedia information retrieval. pp. 323–332. MIR '10, ACM, New York, NY, USA (2010)
21. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: a survey. Found. Trends. Comput. Graph. Vis. 3(3), 177–280 (2008)
22. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search - The Metric Space Approach, Advances in Database Systems, vol. 32. Kluwer (2006)

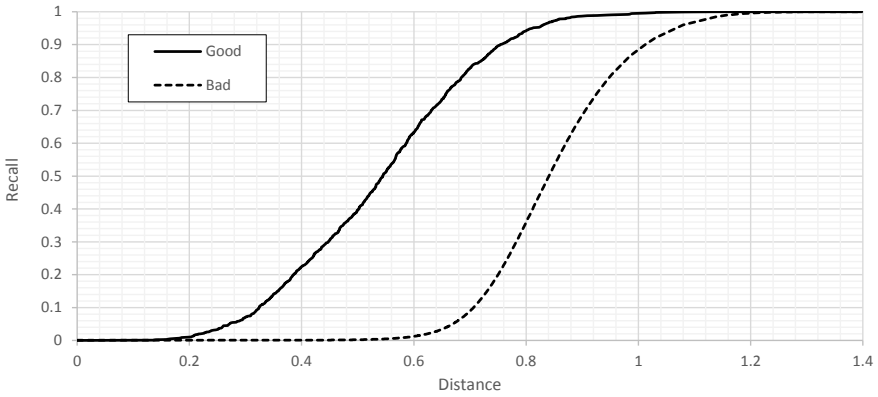

Fig. 1. Recall of bad and good results performing range searches using VLAD with 128 centroids and 128 principal components.

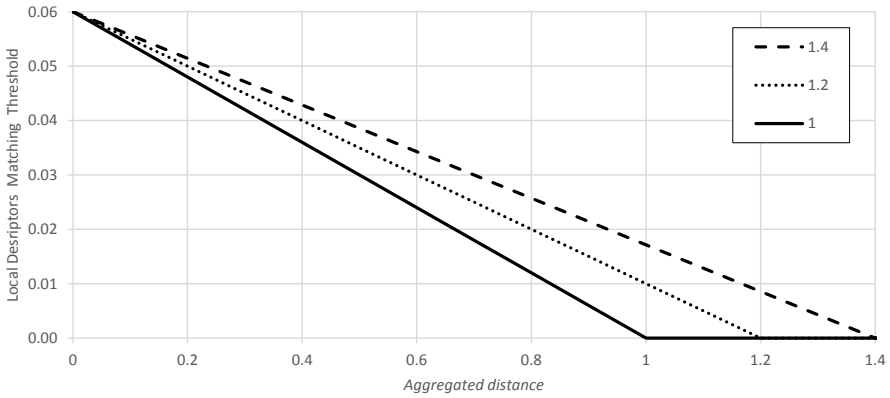

Fig. 2. Resulting local features matching threshold by using the bifocal approach as a function of the aggregated distance for $1/w_a$ and fixed $1/w_l = 0.06$.