Ranking by K-Means Voting Algorithm for Similar Image Retrieval

Przemysław Górecki, Krzysztof Sopyła, and Paweł Drozda

Department of Mathematics and Computer Sciences,
University of Warmia and Mazury, Olsztyn, Poland
{pgorecki,pdrozda}@matman.uwm.edu.pl, ksopyla@uwm.edu.pl

Abstract. Recently, the field of CBIR has attracted a lot of attention in the literature. In this paper, the problem of visually similar image retrieval has been investigated. For this task we use the methods derived from the Bag of Visual Words approach, such as Scale Invariant Feature Transform (SIFT) for identifying image keypoints and K-means to build a visual dictionary. To create a ranking of similar images, a novel Ranking by K-means Voting algorithm is proposed. The experimental section shows that our method works well for similar image retrieval. It turned out that our results are more accurate in comparison with a classical similarity measure based on the Euclidean metric in the order of 6% - 15%.

Keywords: Image ranking, CBIR, SIFT, K-means.

1 Introduction

The amount of information available on the Internet is increasing at a tremendous rate in recent times. This necessitates the need to develop methods and algorithms to effectively search in large data collections [14]. The first systems which dealt with this problem have focused on textual information retrieval. The development of image processing and computer vision methods allowed us to search for information encoded in images, giving birth to Content Based Image Retrieval (CBIR).

CBIR systems allow users to query for relevant images either using words describing the content of the image, or using an example image provided by the user. This paper studies the latter approach. In particular, we focus on the problem of retrieving images which contain similar object within a specific category. This issue is of major importance and is examined in many scientific fields. For instance, eCommerce offers the possibility to search for a similar products (like bags, shoes, watches, etc.), which greatly facilitates finding the right product [17]. Medicine is another area that should be mentioned, where CBIR is widely used and is of great importance [1]. In radiology, CBIR techniques assist radiologists in the assessment of medical images and accurate diagnosing by allowing to search for similar images.

L. Rutkowski et al. (Eds.): ICAISC 2012, Part I, LNCS 7267, pp. 509–517, 2012.

Existing CBIR methods are widely based on the well-known descriptors, described in Section 2. These descriptors encode the image features concerning color, texture, shapes and the length of edges, which are then subsequently employed to measure the similarity between images. Recent scientific reports [4] introduced the dictionary methods (Bag of Words), previously applied successfully in Information Retrieval, in the field of image analysis. Bag of Visual Words technique (BoVW) introduces the image representation as a vector containing frequency of similar image patches.

In order to obtain such representation, one should perform the detection of image keypoints. The most frequently used keypoint detectors are Speeded Up Robust Features (SURF) [3] and Scale-Invariant Feature Transform (SIFT) [18]. After the image representation is obtained, it is possible to create the ranking of similar images on the basis of a similarity measure, such as the Euclidean distance.

This paper takes advantage of Bag of Visual Words and SIFT detector to obtain image representations, while for similarity ranking we introduce a novel method, called Ranking by K-means Voting algorithm, where the clustering is repeated multiple times to get the images ranked by similarity.

The paper is organized as follows. In the next section we review the existing image feature extraction algorithms as well as various image representations. Section 3 describes the chosen methods and the main contribution of this work: Ranking by K-means Voting algorithm. Next, section 4 evaluates the proposed approach based on experimental results. We conclude the article and discuss future directions in section 5.

2 State of the Art

In every CBIR system, two main components can be identified. The first one, called feature extractor, is intended to quantitatively express the information encoded in the basic elements of the image, such as color and texture, edges, shapes or spatial layout of objects. The second one, called ranking component, uses the previously extracted features to calculate the similarity between the query image and all other images in the dataset. It can be accomplished using the simple similarity measure, or more complex approach such as machine-learned ranking [11]. In this paper we propose the ranking component based on unsupervised clustering.

Regarding the feature extraction process, we can distinguish methods which capture the global characteristics of an image (global feature extraction) as well as those which indicate locally relevant areas, known as keypoints. Global feature-based algorithms strive to imitate the human way of perception, that is to discern an object in the picture as a whole. The most popular methods are based on color histograms. In particular, in [6] the classic color histogram is proposed, authors of [8] introduce Fuzzy Color Histogram, while in [10] Color Correlogram method is used. Another way of dealing with global feature extraction is related to the study of information encoded in the image texture. Examples of the algorithms that follow this approach are Steerable Pyramid [21], Gabor Wavelet Transform [9], Contourlet Transform [5] or Complex Directional Filter Bank [23].

Global features algorithms are generally considered to be simple and fast, which often results in the lack of invariance to change of perspective or illumination. To overcome these problems local features methods were introduced [16]. For instance, Schmid and Mohr [20] utilize Harris corner detector to identify interest points which is insensitive to change of image orientation.

Lowe [18] introduced Scale Invariant Feature Transform (SIFT), which proved to be robust against variations in rotation, scale and light intensity. The improvement of the SIFT method can be found in Ke & Sukthankar paper [12]. The authors apply Principal Components Analysis (PCA) for relevant keypoints selection, which results in an increased resistance against image deformations. Finally, Bay et al. proposed Speeded Up Robust Feature (SURF) detector [3], which is several times faster then SIFT while retaining similar stability extracted keypoints.

On the basis of the extracted features the image representation can be formed, which is used for the purpose of determining the similarity between any given images. In case of Bag of Visual Words technique [4] image representation consists of histogram of local image features. Such representation does not encode any spatial relationships. In contrast, representations based on graph theory can be applied [1] if interrelation of features is essential eg. their relative spatial distribution. The choice of image representation has a significant impact on the manner in which the similarity is calculated. In case of feature vectors, it is common to use distance functions, i.e. Euclidean or cosine distance. On the other hand, when an image is represented by a graph, the similarity is defined as graph matching [13] or by the effort required to transform one graph into another [2].

Many of the techniques described above were implemented in the existing CBIR systems, from which the following are worth mentioning: ALIPR `alipr.com` automatic photo tagging and visual image search, BRISC a pulmonary nodule image retrieval framework [15], Tineye `tineye.com` commercial online visual search or `like.com` system for visual shopping.

3 Methodology

This section presents the details of our CBIR method. The feature extraction phase involves creation of the visual words dictionary for which SIFT [18] and k-means algorithms are applied. Then, for the given query image, the ranking of similar images is formed using the Ranking by K-means Voting algorithm detailed in section 3.3. Finally, the accuracy of the proposed method is assessed in the experimental session.

3.1 Dictionary of Visual Words

It has been confirmed by numerous research projects [19], [22] that SIFT is one of the most effective and robust keypoint detectors. Therefore, we follow this approach in our work. Keypoint detection by SIFT proceeds as follows: initially, the potential interest points are localized by means of difference-of-Gaussians. Then, unstable keypoints are rejected. In the next phase, for each keypoint, the additional information concerning its relative orientation, scale and location is added. Finally, on the basis of the histograms of local gradients, keypoint descriptors are computed and have the form of numerical vectors. The number of keypoints extracted from an image depends largely on the complexity of the image elements and may oscillate between a hundred and several thousand.

The goal of the next phase is to group the keypoints into k "visual words". This is achieved using k-means clustering, where each cluster contains the keypoints with the smallest distance to the center of a centroid. As a result, a k-visual size dictionary is formed. This allows to unify the number of features for each image and leads to simpler image representation. An important task is the appropriate selection of parameter k, which significantly affects the quality of results as well as the speed of calculation. If the number of clusters is too small, strongly differing key points can be represented by the same visual word and conversely, if the clusters are too many, similar descriptors can be described by different visual words. This may result in a reduced precision of the obtained results. The experiments with different values of the k parameter in Section 4 are presented.

3.2 Image Representation

Given the image keypoints and the visual dictionary, it is possible to assign visual word to each keypoint. From this point, the image can be represented as a histogram of its visual words. In Information Retrieval, such representation is referred to as term frequency (TF). Taking into account only TF histograms can lead to unsatisfactory results, as TF does not include information about its importance among all images. Thus, for the representation of image we also consider TFIDF (TF - term frequency, IDF - inverse document frequency) from the field of Information Retrieval. The idea of TFIDF is to weight each word according to number of its occurrences among entire dataset of images. Visual word which occurs in few images is more informative that the one appearing in many images. Therefore, the weight value for a particular word is calculated by the following formula:

$$(\mathrm{tf\text{-}idf})_{\mathrm{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|D|}{|\{d : t_i \in d\}|}, \quad (1)$$

where $n_{i,j}$ is the number of occurrences of $i-th$ visual word at $j-th$ image, the denominator of the former fraction is equal to the number of all visual words at $j-th$ image, $|D|$ is number of images and $|\{d : t_i \in d\}|$ is a number of images containing $i-th$ visual word.

Moreover, our previous studies [7] proved that normalizing histograms significantly improves classification based on Bag of Visual Words method. Therefore, we apply the Euclidean norm (2), so that each histogram can be interpreted as a unit-length vector.

$$\|\mathbf{x}\|_2 := \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}. \tag{2}$$

3.3 Ranking by K-Means Voting Algorithm

The next step of proposed methodology is the creation of a similarity ranking for the query image on the basis of the image representations. For this task a novel Ranking by K-means Voting procedure is proposed. A series of k-means clustering is executed to divide all images in the dataset (including the query image) into varying number of groups. After each clustering, the pictures from the same centroid receive a vote which is accumulated in the similarity matrix SM. In particular, for each images pair (s, t) in the same cluster, the value of similarity matrix at position (s, t) is incremented. The general outline of the proposed method is presented in Algorithm 1.

Algorithm 1. Ranking by K-means Voting Algorithm

Require: $N > 0$ {number of images}, $M > 0$ {number of iterations}

var SM[N,N] {similarity matrix of size NxN}
for $i = 1 \rightarrow M$ **do**
    for $k = 2, 3, \ldots, \lfloor \lg_2 N \rfloor$ **do**
        Do k-means clustering procedure
        for all $s, t \in 1..N$ in the same cluster **do**
            SM[s,t] = SM[s,t] + 1
        end for
    end for
end for

After the execution of algorithm the similarity matrix SM is obtained, which holds the similarity value between any two given images. This similarity is expressed by the number of votes that particular pair of images has received. Having the similarity matrix SM, the ranking for image s can be easily obtained by sorting $s - th$ row of SM in descending order. It should be noted that the clustering is repeated Mx $\lfloor \lg_2 N \rfloor$ times to minimize the effect of stucking in local minimum solutions.

4 Experimental Results

The goal of the experimental session was to test the effectiveness of our CBIR approach. The dataset, which can be downloaded from `wmii.uwm.edu.pl/~kmmi` consisted of 166 shoe images from four distinctive shoe categories containing 59, 20, 29, and 58 images in each collection (figure 2 presents the exemplary

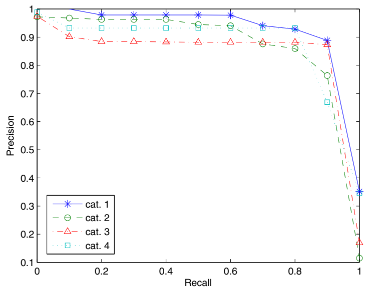

Fig. 1. Precision/Recall graph for Ranking by K-means Voting Algorithm, k=2000 and TF+normalization representation

Fig. 2. Representative images from each category

images from each category). Such structure of the dataset allowed us to preserve the straightforward notion of relevant and irrelevant images for the purpose of retrieval evaluation. In particular, all images from the same category as the query image were considered to be relevant, while images from other category were considered to be irrelevant.

Given the dataset, the visual dictionary was constructed as follows. Initially, SIFT keypoints we extracted from all of the images. Successively, keypoints were clustered into k visual words, using the k-means algorithm, as described in section 3.1. In order to determine the most suitable number of words for the dataset, the image retrieval evaluation was repeated for the following values of k: 50, 100, 250, 500, 1000, 1500, 2000. In addition to different vocabulary sizes, we tested the following term weighting schemes: TF, TF' (normalized TF), TFIDF and TFIDF' (normalized TFIDF), as described in section 3.2. The proposed Ranking by K-means Voting algorithm was compared with the ranking based on the Euclidean distance between the histograms of the query image and all other images in the dataset.

To quantitatively express the quality of the results, we employed standard measures for evaluating retrieval results in unranked datasets, such as precision and recall. Each image in the dataset was used as a query, and its precision and recall values were calculated. This allowed us to obtain the average precision for each query, and the mean average precision (MAP) for each category. Finally, the overall precision among all categories was calculated as the mean of the mean average precisions (MMAP).

The MMAP results obtained for Ranking by K-means Voting are presented in Table 1. In the similar way, the MMAP results for the ranking based on the Euclidean distance are shown in Table 2. It can be noted, that the best results were obtained for theRanking by K-means Voting with the dictionary containing 2000 words and the TF' image representation. In such case MMAP reached over 91 percent. Additionally, it can be observed that Ranking by K-means Voting generally performs better that the ranking based on the Euclidean distance, regardless of the number of words and the weighting scheme used. In particular, when taking into account k values greater than 500, the results were significantly better and the increase of MMAP value oscillated between 6 and 15 percent. Moreover, we examined the effect of the weighting forms applied to visual vocabulary. Considering the TF and TFIDF representations, it can be concluded that the application of the former or the latter slightly affects the results. Figure 1 presents precision/recall graphs for all image categories, that were obtained for Ranking by K-means Voting, $k = 2000$ and TF' weighting.

Table 1. Mean of the mean average precisions (MMAP) for Ranking by K-means Voting Algorithm

	Number of Visual Words						
	50	100	250	500	1000	1500	2000
TF	84.13	83.23	83.18	78.77	67.24	67.72	66.46
TF'	76.37	84.4	87.49	87.52	88.02	88.17	91.85
TFIDF	72.21	74.73	69.63	80.17	75.32	74.27	77.57
TFIDF'	75.01	74.45	81.54	82.85	89.55	90.96	91.38

Table 2. Mean of the mean average precisions (MMAP) for Euclidean distance as similarity measure

	Number of Visual Words						
	50	100	250	500	1000	1500	2000
TF	81,69	80.19	75.8	70.78	64.5	61.98	61.16
TF'	70.34	72.48	74.98	76.03	74.22	74.19	76.01
TFIDF	75.87	77.49	76.69	70.36	64.09	61.89	60.61
TFIDF'	65.66	69.48	74.92	74.89	74.24	74.76	77.27

5 Conclusion and Future Work

The main contribution of this work is the Ranking by K-means Voting algorithm, whose purpose is to create a ranking of similar images. The results obtained in the experimental session show the advantage of the method proposed in this paper over the standard similarity measures, in our case over the Euclidean distance. In particular, we obtained accuracy better by from 6 to 15 percent. In addition, it should be noted that the normalization of image representation has a great impact on the result quality. In most cases, the application of the Euclidean normalization caused a significant increase in accuracy. Finally, studies on the optimal number of "visual words" were undertaken. The results of the experiments show that with the Euclidean normalization the best quality is obtained for $k = 2000$. On the basis of the encouraging results, we plan to test our algorithm on commonly available datasets and compare it with other ranking methods. Other future goals include the verification of the algorithm performance and improving its accuracy.

Acknowledgments. The research described here has been supported by the National Science Center of the Republic of Poland grant N N516 480940. We would like to thank Geox and Salomon companies for providing a set of shoes images for scientific purposes.

References

1. Akgül, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B.: Content-based image retrieval in radiology: Current status and future directions. J. Digital Imaging 24(2), 208–222 (2011)
2. Aschkenasy, S.V., Jansen, C., Osterwalder, R., Linka, A., Unser, M., Marsch, S., Hunziker, P.: Unsupervised image classification of medical ultrasound data by multiresolution elastic registration. Ultrasound Med. Biol. 32(7), 1047–1054 (2006)
3. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
4. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
5. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. IEEE Transactions on Image Processing 14(12), 2091–2106 (2005)
6. Ee, P., Report, P.: Histogram-based color image retrieval. Image Rochester NY, pp. 1–21 (2008)
7. Górecki, P., Artiemjew, P., Drozda, P., Sopyła, K.: Categorization of Similar Objects using Bag of Visual Words and Support Vector Machines. Accepted for Fourth International Conference on Agents and Artificial Intelligence. IEEE (2012)
8. Han, J., Ma, K.K.: Fuzzy color histogram and its use in color image retrieval. IEEE Transactions on Image Processing 11(8), 944–952 (2002)
9. He, C., Zheng, Y.F., Ahalt, S.C.: Object tracking using the gabor wavelet transform and the golden section algorithm. IEEE Trans. Multimedia 4, 528–538 (2002)
10. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., Zabih, R.: Image indexing using color correlograms. In: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR 1997), pp. 762–768. IEEE Computer Society, Washington, DC (1997)
11. Joachims, T.: Optimizing search engines using clickthrough data. In: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 133–142 (2002)
12. Ke, Y., Sukthankar, R.: Pca-sift: a more distinctive representation for local image descriptors. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 506–513 (2004)
13. Keysers, D., Dahmen, J., Ney, H., Wein, B.B., Lehmann, T.M.: Statistical framework for model-based image retrieval in medical applications. Journal of Electronic Imaging 12(1), 59–68 (2003)

14. Korytkowski, M., Scherer, R., Rutkowski, L.: On combining backpropagation with boosting. In: 2006 International Joint Conference on Neural Networks, IEEE World Congress on Computational Intelligence, Vancouver, BC, Canada, pp. 1274–1277 (2006)
15. Lam, M.O., Disney, T., Raicu, D.S., Furst, J., Channin, D.S.: Briscan open source pulmonary nodule image retrieval framework. Journal of Digital Imaging 20(suppl. 1), 63–71 (2007)
16. Lee, Y., Lee, K., Pan, S.: Local and Global Feature Extraction for Face Recognition. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 219–228. Springer, Heidelberg (2005)
17. Li, J.: The application of cbir-based system for the product in electronic retailing. In: 2010 IEEE 11th International Conference on Computer-Aided Industrial Design and Conceptual Design (CAIDCD), pp. 1327–1330 (November 2010)
18. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60, 91–110 (2004)
19. Mikolajczyk, K., Leibe, B., Schiele, B.: Local Features for Object Class Recognition. In: Tenth IEEE International Conference on Computer Vision (ICCV 2005), vol. 1, pp. 1792–1799. IEEE (2005)
20. Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 530–535 (1997)
21. Simoncelli, E.P., Freeman, W.T.: The steerable pyramid: A flexible architecture for multi-scale derivative computation, pp. 444–447 (1995)
22. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: a survey. Found. Trends. Comput. Graph. Vis. 3, 177–280 (2008)
23. Vo, A.P.N., Nguyen, T.T., Oraintara, S.: Texture image retrieval using complex directional filter bank. In: ISCAS (2006)