BOOSTING THE WEIGHTS OF POSITIVE WORDS IN IMAGE RETRIEVAL

Emmanouil Giouvanakis, Constantine Kotropoulos

Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece
{egiouvan, costas}@aiia.csd.auth.gr

ABSTRACT

In this paper, an image retrieval system based on the bag-of-words model is developed, which contains a novel query expansion technique. SIFT image features are computed using the Hessian-Affine keypoint detector. All feature descriptors are taken into account for the bag-of-words representation by dividing the full set of descriptors into a number of subsets. For each subset, a partial vocabulary is created and the final vocabulary is obtained by the union of the partial vocabularies. Here, a new discriminative query expansion technique is proposed in which an SVM classifier is trained in order to obtain a decision boundary between the top ranked and the bottom ranked images. Treating this boundary as a new query, words appearing exclusively in top-ranked images are further boosted by rewarding them with larger weights. The images are re-ranked with respect to the their distance from the new boosted query. It is proved that this strategy improves image retrieval performance.

***Index Terms*—** image retrieval, bag-of-words, query-expansion

1. INTRODUCTION

The majority of the object or image retrieval approaches are based on the well-known bag-of-words model (BoW). The BoW model was firstly introduced by Sivic et al. in [1], suggesting that an image is represented by a set of visual words onto which image features are mapped. More specifically, local descriptors are extracted from each image and then clustered to a certain number of code vectors, i.e., the visual words. By doing so, each descriptor is quantized to its nearest visual word and word frequency vectors are finally computed, yielding the BoW representation. Then, a weighting scheme, such as term frequency - inverse document frequency (*tf-idf*) [2] can be used to assign weights to visual words. When an image query is issued, features are extracted from the image query and quantized into weighted words, using the previously computed vocabulary and the *tf-idf* weights.

Many variations of the just described model have appeared in the literature so far, all trying to improve image retrieval performance. Due to the ever growing volume of images, dealing with large-scale datasets is inevitable. Accordingly, efficient clustering algorithms are needed, especially when processing high-dimensional data [3] and fast nearest neighbour search algorithms are required to ensure real-time performance. Memory consumption is another issue that needs to be addressed.

As far as performance is concerned, it has been proved that query expansion (QE) techniques yield better retrieval results. QE is a non-costly process that enriches an initial image query with information contained in the top retrieved images returned by this image query [4], as detailed in Section 2. Thus, QE is considered as a blind relevance feedback, which definitely delivers better results and contributes to performance improvement.

In addition, spatial verification improves the retrieval performance by checking the geometric consistency between the query and the retrieved images [3, 5, 6]. Although visual words between two images may be visually similar, spatial relationships between feature locations are ignored in the BoW representation. During the spatial verification procedure, a transformation is tested between the query feature locations and the ones in the retrieved images, resulting in a number of spatially verified words, called inliers. Then, a re-ranking can be performed based on the number of inliers.

In this paper, an image retrieval system is proposed employing a novel query expansion technique. It is examined how retrieval performance can be improved further by increasing the weights assigned to positive words, i.e., the words appearing only in the top-ranked retrieved images, when using the discriminative query expansion in [7]. The experiments were conducted on an image dataset of tourism interest, containing multiple images from 6 different landmarks around Greece. The proposed technique was tested for different vocabulary sizes and showed a great performance. An improvement of 2% in mean average precision was measured, while for some classes the increase in the

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operation Program "Competitiveness-Cooperation 2011" - Research Funding Program: SYN-10-1730-ATLAS.

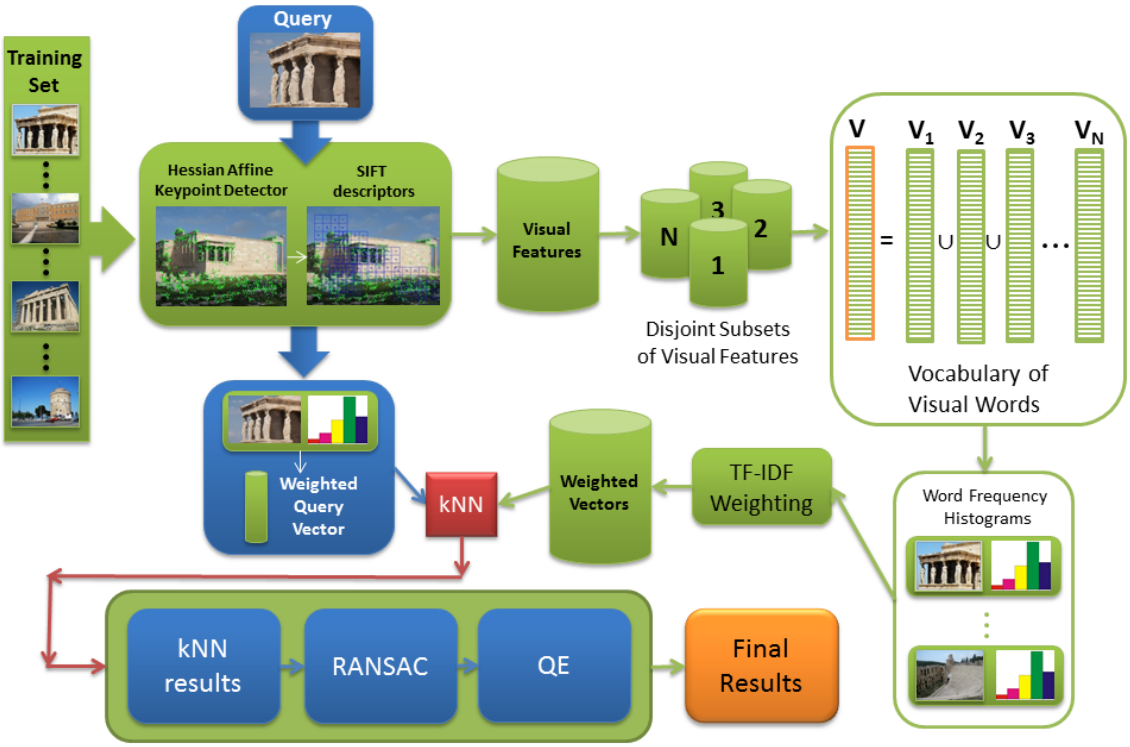

Fig. 1. Overview of the proposed system.

aforementioned figure of merit exceeds 6% An overview of the proposed system is shown in Fig. 1.

The rest of the paper is organized as follows. In Section 2, the QE techniques are reviewed. In Section 3, implementation details of the proposed QE technique are discussed. Experimental results are demonstrated in Section 4, while conclusions are drawn in Section 5.

2. RELATED WORK

Adopted from text retrieval, QE is a standard method of improving retrieval performance, where a number of top ranked images is reissued as a new query. In computer vision, it was introduced in [4]. The new query is computed by averaging the *tf-idf* weights over the top-k retrieved images along with the initial query. There are many variants of QE, which are included in the state-of-the-art approaches [4, 8].

In [3], the user selects a region of a query image containing an object and the system returns a ranked list of images, depicting the same object. The system extracts the SIFT image features and uses the BoW representation to create visual vocabularies. The authors state that flat k-means scales to large collections of image descriptors by using approximate nearest neighbours techniques. This way, the vocabulary is efficiently created and word frequency vectors are computed. Next, the *tf-idf* scheme is used to weigh the visual word frequency vectors. Similarity search is performed by calculating the ℓ_2 distance between a query vector and all the weighted image vectors. Finally, a re-ranking is performed by expanding the query with the top-ranked results, using spatial constraints.

Recently, another approach has been proposed in [9], where locality sensitive hashing is used to solve the large memory consumption problem during visual vocabulary creation requiring less computational time. In particular, exact Euclidean locality sensitive hashing (E2LSH) is used to hash the SIFT features in order to form a group of random visual vocabularies. Next, visual vocabulary histograms are computed and the *tf-idf* weighting scheme is applied to the visual word frequency vectors. Finally, a QE strategy is used to achieve better results.

In Average Query Expansion (AQE), a transformation between the initial query and each of the retrieved images is estimated by means of the RANSAC algorithm [10]. Thus, only images sharing a number of inliers above a predefined threshold are used in the averaged vector.

In [7], a new way of conducting QE is presented. Exploiting the normalized BoW vectors of the spatially verified images retrieved when applying the AQE, an SVM classifier was trained with these vectors as positive data, while *tf-idf* vectors of the bottom ranked images are treated as negative data. This approach is called discriminative query expansion

(DQE). The returned images are ranked with respect to their distance from the decision boundary determined by the SVM.

3. PROPOSED IMAGE RETRIEVAL SYSTEM

3.1. Features extraction and BoW representation

For feature extraction, a modified version of the state-of-the-art Hessian-Affine keypoint detector [11] was used which was proposed in [12]. In particular, using the code used in [12], SIFT descriptors [13] are computed on affine normalized image patches. All SIFT descriptors are then normalized to unit ℓ_2 norm.

Next, a vocabulary of visual words is created by using the VLFeat toolbox [14], and particularly the approximate nearest neighbor k-means algorithm [15–17], which is a variant of Lloyd's algorithm that uses a best-bin-first randomized kd-tree algorithm to approximately find the closest cluster center to each point, quickly. To avoid large memory consumption and increase computational efficiency, a percentage of 20% of total descriptors is usually a good portion for the initial set to extract the visual word vocabulary. However, since vocabulary creation is an off-line process, the full set of descriptors is kept and is divided randomly into N disjoint subsets. For example, if there are D descriptors divided into N subsets, a vocabulary V of size $|V|$ is generated by concatenating V_i, $i = 1, 2, ..., N$ partial vocabularies of size approximately $|V|/N$ words, i.e., $V = \bigcup_{i=1}^{N} V_i$. Using this vocabulary, each descriptor is quantized to its nearest word and visual word frequency histograms (i.e., BoW vectors) are computed for each image.

Then, the *tf-idf* weighting scheme [2] is applied to the BoW vectors, which downweighs the common visual words, holding no discriminating power, while it rewards with higher weights, visual words appearing in a small portion of images. Finally, the weighted *tf-idf* vectors are normalized to unit ℓ_2 norm.

3.2. Spatial verification of the retrieved images

When applying the initial query, a ranked list of images is returned. In order to acquire the closest images to the query, the Euclidean distance between the normalized BoW vectors is used. However, false positives are likely to exist due to the nature of the BoW representation, which ignores the spatial relationships between the words. This means that although two words may be visually similar, they are not geometrically consistent. Thus, expanding the query with this "false" information does not guarantee stable retrieval performance.

In order to test if two images are geometrically consistent an affine homography matrix is estimated by applying the RANSAC algorithm [10] on the matched keypoints between the two images. The RANSAC algorithm is commonly used to find inliers, i.e., spatially verified keypoints between the query image and the retrieved ones. Since spatial verification is a time consuming process, it is not performed on the whole set of retrieved images, but only on the 100 top ranked ones. Employing more images in spatial verification did not show any relative improvement, taking into account computational demands.

To ensure higher retrieval performance, a criterion needs to be set that determines whether two images are considered geometrically consistent or not. Let M_{qi} and S_{qi} denote the number of visual matches and the number of spatially verified keypoints between the query and the ith retrieved image I_i, $i = 1, 2, ..., N$, respectively. Two images are considered spatially verified if an adequate number of visual matches exist and S_{qi}/M_{qi} is above a predefined threshold c. To avoid useless computations, in order to proceed to the spatial verification step, a strict criterion was set: the number of matches has to be above 20, i.e., $M_{qi} \geq 20$, while the threshold is set to $c = 0.2$, meaning that more than 20% of the visual matches need to be inliers.

3.3. Discriminative Positive Query Expansion

As described in Section 2, Arandjelović et al. proposed [7] DQE as a binary classification problem. Top-ranked images are used as positive examples, while images appearing at the bottom of the retrieved images are considered as negative examples. Inspired by DQE, a modified version of the DQE technique is proposed here that boosts the visual words in top retrieved results.

If, for example the bottom-m images were chosen along with the top-m ones to train an SVM classifier, it is expected that among the top-m images there will be many images that are not geometrically consistent with the query. In addition, the negative examples are more than the positive ones, meaning that positive words lying in positive images are less than the negatives ones. Although the positive set of spatially verified images is quite small, it is vital for improving the retrieval performance. So, the motivation behind the proposed DQE is to boost further the visual words contained in positive images.

To achieve this, a straightforward approach is to reward positive words with higher *tf-idf* weights. However, words appearing in the set of positive examples may appear in the set of negative examples. So, it is proposed to boost only the visual words that exclusively appear in the set of positive examples. Let us call them pure positive words.

Let W_P be a binary vector indicating the words appearing in the positive images, while W_N be a binary vector pointing to the words in the set of negative examples. So the vector of pure positive words W_{PP} has elements:

$$W_{PP_i} = \begin{cases} 1, & \text{if } W_{P_i} = 1 \text{ and } W_{N_i} = 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Then, the proposed Discriminative Positive Query Expansion (DPQE) is computed by:

Table 1. Mean average precision using an 80K visual vocabulary.

	1320 images					*1320 images + 2000 distractors*				
Class	**NQE**	**QE**	**AQE**	**DQE**	**DPQE**	**NQE**	**QE**	**AQE**	**DQE**	**DPQE**
Erechtheum	0.50	0.64	0.70	0.71	**0.77**	0.45	0.67	0.74	0.75	**0.79**
Odeon	0.68	0.79	0.82	0.83	**0.84**	0.54	0.68	0.71	0.70	**0.74**
Parliament	0.77	0.87	0.91	0.91	**0.91**	0.80	0.88	0.91	0.91	**0.91**
Parthenon	0.52	0.71	0.74	0.75	**0.76**	0.49	0.70	**0.75**	0.74	0.72
Sounio	0.42	**0.57**	0.52	0.52	0.52	0.34	0.47	**0.47**	0.46	0.46
White Tower	0.61	0.78	0.78	0.78	**0.81**	0.51	0.71	0.73	0.73	**0.75**
Total	0.58	0.73	0.74	0.75	**0.77**	0.52	0.69	0.72	0.71	**0.73**

$$q_{DPQE} = q_{DQE} + b \cdot q_{DQE} * W_{PP} \quad (2)$$

where b is a constant boost factor, q_{DQE} the DQE query vector, and $*$ denotes the Hadamard product between the vectors q_{DPQE} and W_{PP}. Values for the boost factor b varying between 0.1 and 2 were tested and best results were measured for $b = 1.5$.

The classifier used both in DQE and DPQE is a linear SVM trained with LIBSMV [18]. During training, the weighting parameter of the cost function in LIBSVM was set to 5, although no differences in accuracy were noticed for other values.

4. EVALUATION

The image dataset used in this paper contains images of touristic interest, which were crawled from *Flickr* using keywords/tags related to Greece. The total size of the collected image dataset is about 650,000, images containing archaeological monuments, landscapes around Greece, etc. It is stored in an SQL database locally.

The proposed method is tested on two subsets of the original dataset. The first one, consists of 1,320 images, containing landmarks in 6 different classes, while the second one contains the previous one along with 2,000 irrelevant images, which play the role of distractors.

To evaluate the system, 20 test images from each class were given as queries, resulting in an evaluation set of 120 images. The mean Average Precision (mAP) is used as figure of merit between the proposed DPQE and the following QE techniques:

- **NQE**: No query expansion was performed.
- **QE**: The *tf-idf* scores of the top-10 ranked images are averaged and reissued as a new query.
- **AQE**: The top-100 ranked images are tested for spatial verification and only the *tf-idf* scores of the spatial verified images are averaged and reissued as a new query.
- **DQE**: The Discriminative Query Expansionm, as described in [7]. The SVM classifier is trained with spatially verified images among the top-100 ranked images and the bottom-200 images.

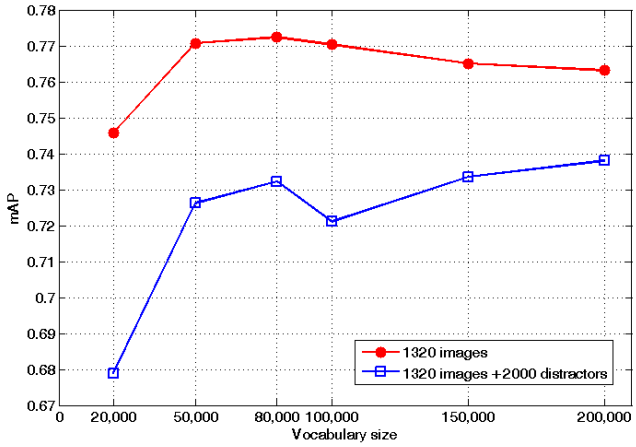

Fig. 2. Mean average precision (mAP) with respect to vocabulary size.

For each dataset, vocabularies of sizes between 20K to 200K words were extracted. During evaluation, the 80K vocabulary showed the best performance, considering both the retrieval performance and the computational efficiency. In Figure 2, the mAP with respect to different vocabulary sizes is plotted for both datasets. The mAP for a vocabulary of 80K visual words across the image classes and on average for the various methods compared are listed in Table 1.

It is noticed that the boosting of pure positive words improves retrieval performance. Especially inside some classes, such as the White Tower class, the improvement exceeds 6%. Finally, distractors do not seem to deteriorate the proposed method, since the mAP of the DPQE exceeds that of the other QE techniques.

To test if the total mAP differences are statistically significant, we assume that the mAPs p_1 and p_2 delivered by two methods are binomially distributed random variables. If $\hat{p_1}$, $\hat{p_2}$ denote the empirical mAPs listed in the last row of Table 1 and $\bar{p} = \frac{\hat{p_1} + \hat{p_2}}{2}$, the hypothesis $H_0 : p_1 = p_2 = \bar{p}$ is tested at 95% level of significance. The variance of the mAP differ-

ence is given by $\beta = 2\,\dfrac{\bar{p}(1-\bar{p})}{T}$, where T is the number of test images (i.e., 120). For $\varphi = 1.65\sqrt{\beta}$, if $|\hat{p_1} - \hat{p_2}| \geq \varphi$, we reject H_0 with risk 5% of being wrong. The aforementioned analysis yields that the performance gain between the NQE and DPQE ($\varphi = 9.98\%$) on the first dataset is statistically significant, while between the DQE and DPQE ($\varphi = 9.09\%$) is not.

5. CONCLUSIONS

An image retrieval system has been proposed. The novelty of the system is a new query technique, which is a modified version of the discriminative query expansion in [7]. It has been shown that rewarding the pure positive words by larger weights the performance of the system is improved. The proposed technique has been compared with 4 query expansion techniques. In most cases, DPQE outperforms these techniques.

In the future, other visual words representations such as those derived from the Probabilistic Latent Semantic Analysis [19] will be tested and words can be assigned into topics, reducing the dimensionality of the representation significantly.

REFERENCES

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. Int. Conf. Computer Vision*, 2003, pp. 1470–1477.

[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search*, Addison Wesley Professional, 2011.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching.," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2007.

[4] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. IEEE Int. Conf. Computer Vision*, 2007, pp. 1–8.

[5] A. Mikulik, O. Chum, and J. Matas, "Image retrieval for online browsing in large image collections," in *Proc. Similarity Search and Applications*, pp. 3–15. Springer, 2013.

[6] L. Dai, X. Sun, F. Wu, and N. Yu, "Large scale image retrieval with visual groups," in *Proc. IEEE Int. Conf. Image Processing*, 2013, pp. 582–2586.

[7] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2012.

[8] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2011, pp. 889–896.

[9] Z. Yongwei, L. Bicheng, and G. Haolin, "Bag-of-visual-words based object retrieval with E2LSH and query expansion," in *Instrumentation, Measurement, Circuits and Systems*, vol. 127 of *Advances in Intelligent and Soft Computing*, pp. 713–725. 2012.

[10] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[11] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[12] M. Perd'och, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2009, pp. 9–16.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. ACM Int. Conf. Multimedia*, New York, NY, USA, 2010, pp. 1469–1472.

[15] C. Silpa-Anan and R. Hartley, "Optimised kd-trees for fast image descriptor matching," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[16] M. Muja and D. G Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration.," in *Proc. VISAPP*, 2009, pp. 331–340.

[17] J. S. Beis and D. G. Lowe, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*. IEEE, 1997, pp. 1000–1006.

[18] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[19] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. ACM Int. Conf. Research and Development in Information Retrieval*, 1999, pp. 50–57.