USING A REGION AND VISUAL WORD APPROACH TOWARDS SEMANTIC IMAGE RETRIEVAL

Yannis Kalantidis, Evaggelos Spyrou, Phivos Mylonas and Stefanos Kollias

Image, Video and Multimedia Laboratory
School of Electrical and Electronics Engineering
National Technical University of Athens
{ykalant,espyrou,fmylonas,stefanos}@image.ntua.gr

ABSTRACT

This paper presents a region-based approach towards semantic image retrieval. Combining segmentation and the popular Bag-of-Words model, a visual vocabulary of the most common “region types” is first constructed using the database images. The visual words are consistent image regions, extracted through a k-means clustering process. The regions are described with color and texture features, and a ”model vector” is then formed to capture the association of a given image to the visual words. Opposite to other methods, we do not form the model vector based on all region types, but rather to a smaller subset. We show that the presented approach can be efficiently applied to image retrieval when the goal is to retrieve semantically similar rather than visually similar images. We show that our method outperforms the commonly used Bag-of-Words model based on local SIFT descriptors.

1. INTRODUCTION

It is still rather difficult to grasp the actual semantic description of a given image and this is regarded as the main obstacle from an efficient and successful semantic image retrieval scheme. Several research approaches range from text-based to content-based ones. The former tend to apply text-based retrieval algorithms to a set of usually (pre-)annotated images including keywords, tags, or image titles, as well as filenames. The latter typically apply low-level image processing and analysis techniques to extract visual features from images, whereas their scalability is questionable. Most of them are limited by the existing state-of-the-art in image understanding, in the sense that they usually take a relatively low-level approach and fall short of higher-level interpretation and knowledge.

In this paper, we shall provide our research view on modelling and exploiting visual information towards efficient retrieval of multimedia content. Our goal is to create a meaningful representation of visual features of images by constructing a visual vocabulary. This vocabulary contains the most common region types encountered within a large-scale image database. A model vector is then formed to capture the association of a given image to the visual dictionary. The goal of our work is to retrieve semantically similar images. This means that given a query image, depicting a semantic concept, only the returned images that contain the same semantic concepts will be considered as relevant. Thus, images that appear visually similar, without containing the semantic concept of the query image will be considered irrelevant.

The idea of using a visual dictionary in order to quantize image features has been used widely in both image retrieval and high-level concept detection. In [1] images are segmented into regions and regions correspond to visual words based on their low level features. Moreover, in [2] the bag-of-words model is modified in order to include features which are typically lost within the quantization process. In [3], fuzziness is introduced in the process of the mapping to the visual dictionary. This way the model does not suffer from the ”curse of dimensionality”. In [4] images are divided into regions and a joint probabilistic model is created to associate regions with concepts.

In [5] a novel image representation is proposed (bag of visual synset), defined as a probabilistic relevance-consistent cluster of visual words, in which the member visual words induce similar semantic inference towards the image class. The work presented in [6] aims at generating a less ambiguous visual phrase lexicon, where a visual phrase is a meaningful spatially co-occurrent pattern of visual words.

The rest of this paper is structured as follows: Section 2 discusses the idea of using a visual vocabulary in order to quantize image features and presents the approach we adopt. Section 3 presents the algorithm we propose in order to create a model vector that will describe the visual properties of images. Experiments will be presented in Section 4 and conclusions will be drawn in Section 5.

This research was partially supported by the European Commission under contract FP7-215453 - WeKnowIt.

2. BUILDING A VISUAL VOCABULARY

As it has already been mentioned, the idea of using a visual vocabulary to quantize image features has been used in many multimedia problems. In this Section we discuss the role of the visual vocabulary and we present the approach used in this work for its construction.

Given the entire set of images of a given database and their extracted low-level features, it may easily be observed that regions that correspond to the same concept have similar low-level descriptions. Also, images that contain the same high-level concepts are typically consisted of similar regions. For example, regions that contain the concept *sky* are generally visually similar, i.e. the color of most of them should be some tone of "blue". On the other hand, images that contain *sky*, often are consisted of similar regions.

The aforementioned observations indicate that similar regions often co-exist with some high-level concepts. This means that region co-existences should be able to provide visual descriptions which can discriminate between the existence or not of certain high-level concepts. As indicated in the bibliography, by appropriately quantizing the regions of an image dataset, we can create efficient descriptions. Thus, this work begins with the description of the approach we follow in order to create a visual vocabulary of the most common region types encountered within the data set. Afterwards, each image will be described based on a set of region types.

In every given image I_i we first apply a segmentation algorithm, which results to a set of regions R_i. The segmentation algorithm we use is a variation of the well-known RSST [7], tuned to produce a small number of regions. From each region r_{ij} of I_i we extract visual descriptors, which are then fused into a single feature vector f_i as in [8]. We choose to extract two MPEG-7 descriptors [9], namely the Scalable Color Descriptor and the Homogeneous Texture Descriptor, which have been commonly used in the bibliography in similar problems and have been proved to successfully capture color and texture features, respectively.

After the segmentation of all images of the given image dataset, a large set $\mathcal{F}$ of the feature vectors of all image regions is formed. In order to select the most common region types we apply the well-known K-means clustering algorithm on $\mathcal{F}$. The number of clusters which is obviously the number of region types N_T is selected experimentally.

We define the visual vocabulary, formed by a set of the region types as

$$T = \Big\{w_i\Big\},\ i = 1, 2, \dots N_T,\ w_i \subset \mathcal{F}\ , \tag{1}$$

where w_i denotes the i-th region type. We should note here that after clustering the image regions in the feature space, we chose those that lie nearest to the centroid of each cluster.

We should emphasize that although a region type does not contain conceptual semantic information, it appears to carry a higher description than a low-level descriptor; i.e. one could intuitively describe a region type as "green region with a coarse texture", but would not be necessarily able to link it to a specific concept such as *vegetation*, which neither is necessary a straightforward process, nor falls within the scope of the presented approach.

3. CONSTRUCTION OF MODEL VECTORS

In this Section we will use and extend the ideas presented in [10] and [11], in order to describe the visual content of a given image I_i using a model vector m_i. This vector will capture the relation of a given image with the region types of the visual vocabulary. For the construction of a model vector we will not use the exact algorithm as in [10]. Instead and for reasons that will be clarified later we will modify it, in order to fit in the problem of retrieval.

Let R_i denote the set of the regions of a given image I_i after the application of the aforementioned segmentation algorithm. Moreover, let N_i denote its cardinality and r_{ij} denote its j-th region. Let us also assume that a visual vocabulary $= \{w_i\}$ consisting of N_T region types has been constructed following the approach discussed in Section 2.

In previous work we constructed a model vector by comparing all regions R_i of an image to all region types. For each region type, we chose to describe its association to the given image by the smallest distance to all image regions. Let

$$m_i = \{m_i(1)\ m_i(2)\ \dots\ m_i(N_T)\}\ , \tag{2}$$

denote the model vector that describes the visual content of image I_i in terms of the visual dictionary. We calculated each coordinate as

$$m_i(j) = min_{r_{ij} \in R_i}\{d(f(w_j), f(r_{ij})\}, j = 1, 2, \dots, N_T\ . \tag{3}$$

In this work, instead of m_i we calculate a modified version of the model vector which will be referred to as $\hat{m}_i$. After calculating the distances among each region r_{ij} and all the region types, let $\mathcal{W}_{ij}$ denote an ordered set that contains all the region types with an ascending order, based on their distances d_{ij} to r_{ij}, as

$$\mathcal{W}_i = \{w_{ij} \mid \forall k, l \leq N_T, k \leq l : w_{ik} \leq w_{il}\}\ . \tag{4}$$

For each region r_{ij} we select its closest region types, which obviously are the first K elements of $\mathcal{W}_i$. This way and for each region we define the set of its K closest region types as

$$\mathcal{W}_i^K = \{w_{ij} : j \leq K\}\ . \tag{5}$$

To construct a model vector $\hat{m}_i$, instead of using the whole visual vocabulary, we choose to use an appropriate subset. This will be the union of all ordered sets $\mathcal{W}_i^K$

$$W^K = \bigcup_i \mathcal{W}_i^K\ . \tag{6}$$

This way, the set W^K consists of the closest region types of the visual dictionary to all image regions. We will construct the model vector using this set, instead of the set of all region types. Again,

$$\hat{m}_i = \{\hat{m}_i(1)\ \hat{m}_i(2)\ \ldots\ \hat{m}_i(N_T)\}\ . \tag{7}$$

We define as $\hat{m}_i(j)$ the minimum distance of a region type to all image regions, thus it is calculated as

$$\hat{m}_i(j) = \begin{cases} min\{d(f(w_{ij}), f(r_{ij}))\} & \text{if } w_{ij} \in W^K \\ 0 & \text{else} \end{cases} . \tag{8}$$

If we compare Eq.8 with Eq.3 we can easily observe that the resulting model vector $\hat{m}_i$, it becomes obvious that it is not constructed based on the full visual vocabulary. Instead, our method selects an appropriate subset.

The method we followed in order to construct $\hat{m}_i$ contains an intermediate step when compared to the one for the construction m_i. The latter has been used successfully in a high-level concept detection problem. The use of a neural network classifier practically assigned weights to each region type. Thus, those that were not useful for the detection of a certain concept had been ignored. However, in the case of the retrieval we do not assign any weights to the region types. This means that if the model vector consisted from all region types, those with a small distance to the image regions would act as noise. In this case, retrieval would fail, as many images would have similar descriptions despite being significantly different in terms of their visual content.

To further explain the aforementioned statement, we also give a semantic explanation on why the choice of K instead of one region types for each image region is meaningful and crucial. From a simple observation of a given data set, but also intuitively, it is obvious that many high-level concepts are visually similar to more than one region types. For example, let us assume that the concept *sand* appears "brown" in an image of the database and "light brown" in another. Let us now consider a query image containing the concept *sand*. If the given visual vocabulary contains both a "brown" and a "light brown" region types, in order to retrieve both the aforementioned images of the database, their description should contain both region types and not the most similar. Thus, this way we tackle the problem of quantization.

An artificial example of the K most similar region types to each image region is depicted in Fig.1, for the case of $K = 2$.

4. EXPERIMENTAL RESULTS

In order to test the efficiency of the proposed approach, we selected a dataset[1] created by Oliva and Torralba. This collection was used in a scene recognition problem and is annotated both globally and at a region level. A sample of the dataset is depicted in Fig.2. We used only the global annotations for all 2688 images. For each concept, we calculated the mean Average Precision (mAP).

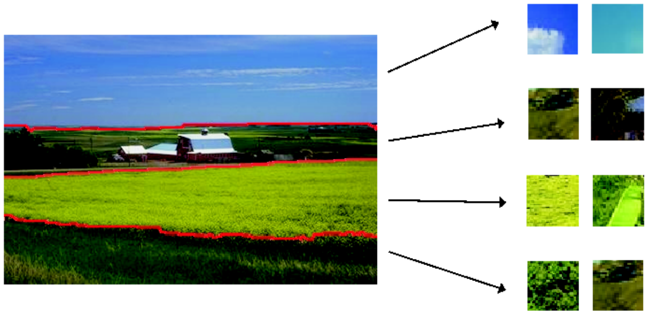

Fig. 1. A segmented image and the 2 most similar region types to each region.

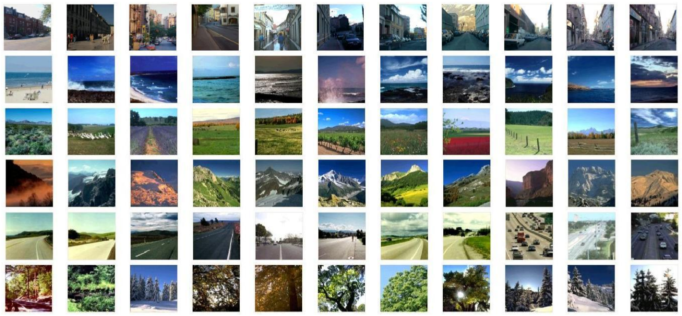

Fig. 2. A subset of the Torralba dataset.

We should remind here that given a query image belonging to a certain semantic category, only those images within the results that belong to the same category were considered as relevant.

For comparison, we chose to evaluate the proposed method against the two most common techniques used in image retrieval: A *global* visual descriptor approach, where the images are represented by a color and texture feature vector extracted from the whole image, and a *local* descriptor approach, the K=1 version of [2].

We investigated the effect on performance of two parameters: the size of the region types vocabulary, and the number of closest region types we keep per region.

Table 1 summarizes the results of our method on selected categories of the Torralba dataset. We may observe that the proposed retrieval algorithm achieved better performance in concepts *coast* and *forest*. On the other hand, the mAPs for concepts *highway* and *street* were not as high. This result can be explained if we consider the visual properties of these concepts. In the case of *coast* and *forest*, the segmentation algorithm created regions which can easily discriminate those

[1] `http://people.csail.mit.edu/torralba/code/spatialenvelope/`

	Nt=150, K=1	Nt=150, K=2	Nt=150, K=4	Nt=270, K=1	Nt=270, K=2	Nt=270, K=5
coast	0.336	0.351	0.374	0.472	**0.674**	0.465
mountain	0.332	0.300	0.321	0.335	0.447	**0.469**
forest	**0.291**	0.162	0.190	0.285	0.248	0.193
open country	0.150	0.120	0.152	0.135	0.129	**0.172**
street	0.080	0.116	0.146	0.074	0.106	**0.159**
inside city	0.108	0.118	0.141	0.155	0.142	**0.215**
tall buildings	0.097	0.098	0.122	0.140	0.142	**0.167**
highways	0.078	0.082	0.107	0.071	0.112	**0.159**

Table 1. The mAP calculated for six different visual vocabularies, whose size is denoted as N_T and for six cases of closest region types K for the **Torralba dataset**.

concepts, while in the images depicting *street* and *highway*, the segmented regions are similar and also similar to those of the concept *inside city* which was used as a distractor.

We also investigated the effect of the number K of region types which are considered to be similar to the image regions, to the mAP that is achieved. We can see in Table 1 that the mAP of *street* increases for higher values of K, while we observe the opposite for *coast*. The same observations stand for *highway* and *forest*. This leads to the conclusion that the concepts that may be considered as intuitively ”simpler”, can be efficiently described and retrieved by a smaller value K of their closest region types.

Table 2 depicts the performance of the proposed method against the global and local descriptor methods, when it comes to semantic image retrieval. The otherwise popular SIFT features, fail to capture the semantic–visual connection between the concepts of this dataset, and even perform sometimes worse than the naive global descriptors. Finally, Figure 3 depicts the actual variation of retrieval results based on the herein proposed approach in terms of mAP values for all eight concepts, given the visual vocabulary of $NT = 270$ region types.

5. CONCLUSIONS

In this paper we presented an approach for semantic image retrieval using a region-based visual vocabulary. We proposed a Bag-of-Words model based on image regions for capturing the visual properties of regions of the images. Instead of using the entire vocabulary, we selected a subset consisting of the closest region types to the image regions. This led to a simple and effective representation of the image features that is close to the notion of soft visual word assignment. We showed that the presented approach can be efficiently applied to image retrieval when the goal is to retrieve semantically similar rather than visually similar images. Experimentally, our method outperforms commonly used Bag-of-Words models based on either local or global descriptors.

6. REFERENCES

[1] P. Duygulu, K. Barnard, JFG De Freitas, and D.A Forsyth, “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary,” *Lecture Notes in Computer science*, 2002.

[2] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Lost in quantization: Improving particular object retrieval in large scale image databases,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[3] J. van Gemert, J. Geusebroek, C. Veenman, and A. Smeulders, “Kernel codebooks for scene categorization,” in *European Conference on Computer Vision (ECCV)*. Springer, 2008.

[4] V. Lavrenko, R. Manmatha, and J. Jeon, “A model for learning the semantics of pictures,” in *in NIPS*. 2003, MIT Press.

[5] Y. Zheng, S. Neo, T. Chua, and Q. Tian, “Object-Based Image Retrieval Beyond Visual Appearances,” *Lecture Notes in Computer Science*, vol. 4903, pp. 13, 2008.

[6] J. Yuan, Y. Wu, and M. Yang, “Discovery of collocation patterns: from visual words to visual phrases,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. Citeseer, 2007, vol. 1.

[7] Y. Avrithis, A. Doulamis, N. Doulamis, and S. Kollias, “A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases,” *Computer Vision and Image Understanding*, vol. 75, no. 1, pp. 3–24, 1999.

[8] E. Spyrou, H. Le Borgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O Connor, “Fusing mpeg-7 visual descriptors for image classification,” in *International Conference on Artificial Neural Networks (ICANN)*, 2005.

	Region-based, K=5	SIFT descriptors	Global descriptors
coast	**0.465**	0.130	0.394
mountain	**0.469**	0.122	0.340
forest	**0.193**	0.100	0.173
open country	0.172	0.096	**0.227**
street	**0.159**	0.112	0.134
inside city	**0.215**	0.209	0.186
tall buildings	0.167	**0.354**	0.154
highways	0.159	**0.217**	0.108

Table 2. Comparison of the proposed approach (for $K = 5$) against two local and global descriptor-based retrieval schemes.

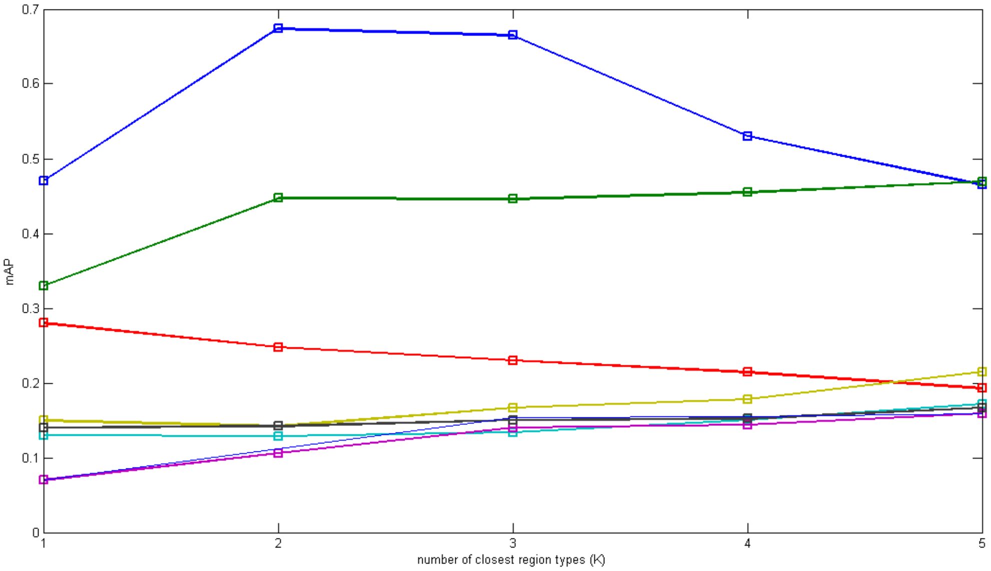

Fig. 3. Change of mAP according to the number K of region types. The mAP is calculated for a visual vocabulary of size $NT = 270$.

[9] S.F. Chang, T. Sikora, and A. Purl, "Overview of the MPEG-7 Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 688–695, 2001.

[10] E. Spyrou, G. Tolias, P. Mylonas, and Y. Avrithis, "Concept detection and keyframe extraction using a visual thesaurus," *Multimedia Tools and Applications*, vol. 41, no. 3, pp. 337–373, 2009.

[11] P. Mylonas, E. Spyrou, Y. Avrithis, and S. Kollias, "Using Visual Context and Region Semantics for High-Level Concept Detection," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 229, 2009.