Exploiting a region-based visual vocabulary towards efficient concept retrieval

Evaggelos Spyrou, Yannis Kalantidis, and Phivos Mylonas

Image, Video and Multimedia Systems Laboratory,
School of Electrical and Computer Engineering
National Technical University of Athens
9 Iroon Polytechniou Str., 157 80 Athens, Greece,
espyrou@image.ece.ntua.gr,
WWW home page: http://www.image.ece.ntua.gr/∼espyrou/

Abstract. This paper presents our approach for semantic concept retrieval based on visual characteristics of multimedia content. The former forms a crucial initial step towards efficient event detection, resulting into meaningful interpretation of available data. In the process, a visual vocabulary is constructed in order to create a representation of the visual features of still image content. This vocabulary contains the most common visual features that are encountered within each still image database and are referred to as "region types". Based on this vocabulary, a description is then formed to capture the association of a given image to all of its region types. Opposite to other methods, we do not describe an image based on all region types, but rather to a smaller representative subset. We show that the presented approach can be efficiently applied to still image retrieval when the goal is to retrieve semantically similar rather than visually similar image concepts by applying and evaluating our method to two well-known datasets.

1 Introduction

It is true that the main obstacle in order to successfully implement the task of (semantic) concept retrieval in multimedia is that the actual semantic description of image objects or even of entire image scenes, is rather difficult to grasp. Several research approaches exist in the literature and they 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 semantic interpretation and retrieval of multimedia content. Our goal is to create a meaningful representation of visual features of images by constructing a visual vocabulary, which will be used at a later stage for efficient concept detection. The proposed 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 through the detection of semantic similar concepts within them. 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 plain 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 well-known "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. However, as it will be showed in the following sections, all above references lack significantly in comparison to the herein proposed work, both in terms of representation modelling and scalability/expressiveness.

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 are presented in Section 4 and brief conclusions are finally drawn in Section 5.

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 for concepts that can be characterized as "scenes" or "materials" 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

$$T = \Big\{ w_k \Big\},\ k = 1, 2, \ldots N_T,\ w_k \subset \mathcal{F}\ , \tag{1}$$

where w_k denotes the k-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 $T = \{w_i\}$ 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

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

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

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

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

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

For each region r_{ij} we select its K closest region types, which obviously are the first K elements of $\mathcal{W}_{ij}$. 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 $K = 2$.

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

4 Experimental Results

In order to test the efficiency of the proposed approach, we selected two descriptive image collections, one dataset[1] created by Oliva and Torralba and one comprised by images derived from the Corel image collection [12]. The first collection was used in a scene recognition problem and is annotated both globally and at a region level. A sample of the first dataset is depicted in Fig.2. We used only the global annotations for 2688 images, as well as all 8 categories of the dataset to evaluate our approach, namely: *coast*, *mountain*, *forest*, *open country*, *street*, *inside city*, *tall buildings* and *highways*.

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

A similar procedure was followed for the second dataset, containing 750 images and 6 concepts, namely: *vegetation*, *road*, *sand*, *sea*, *sky* and *wave*.

In order to meaningfully evaluate our work, we calculated the mean Average Precision (mAP) measure for each concept. At this point we should remind the reader 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 to be relevant. In addition, the well-known Euclidean distance was applied in order to compare the visual features between regions and region types. The mAP that has been achieved for several visual vocabularies and for several cases of the region types that were considered to be similar to the image regions is depicted in the following Tables.

	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	0.317	0.336	0.360	0.460	**0.660**	0.450
mountain	0.320	0.287	0.311	0.317	0.428	**0.459**
forest	**0.275**	0.146	0.170	0.270	0.230	0.180
open country	0.134	0.109	0.133	0.121	0.111	**0.158**
street	0.063	0.106	0.130	0.060	0.090	**0.140**
inside city	0.094	0.098	0.121	0.145	0.130	**0.204**
tall buildings	0.084	0.081	0.105	0.124	0.131	**0.152**
highways	0.067	0.066	0.090	0.060	0.100	**0.140**

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 **Oliva/Torralba dataset**.

Table 1 summarizes the results of the application of our method to the first aforementioned dataset. We may observe that the proposed retrieval algorithm achieved its best performance in concepts *coast* and *mountain*. Concept *forest* appears to be somewhere in the middle range, whereas mAPs for concepts *open country*, *street*, *inside city*, *tall buildings* and *highways* were not as high, with all of them ranging equal or below value 0.20. This result can be explained if we consider the visual properties of these concepts. In the case of *coast* and *mountain*, the segmentation algorithm created regions which can easily discriminate those concepts, while in the images depicting the rest of the concepts, segmented regions are more similar to each other and thus not discriminated thoroughly.

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. Fig. 3 depicts the evolution of mAP vs. K and N_T for all sets of concepts utilized. It is obvious that in the case of low mAPs (e.g. for all 5 concepts mentioned above), mAP values increase for higher values of K, while we observe an intermediate behavior for concepts with significantly better mAPs like *coast* or *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 presents the corresponding results of the application of our method to the second dataset. In this case the proposed retrieval algorithm worked more efficiently,

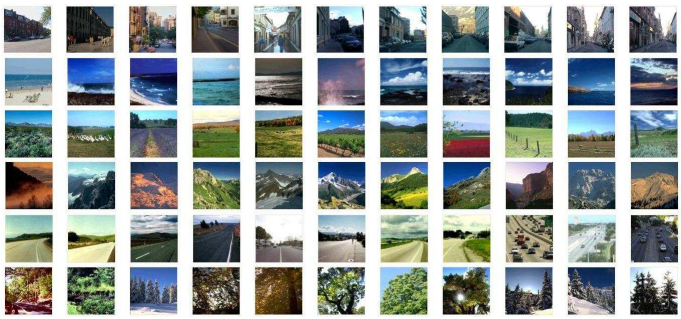

Fig. 2. A subset of the Torralba dataset.

Fig. 3. A subset of the Corel dataset.

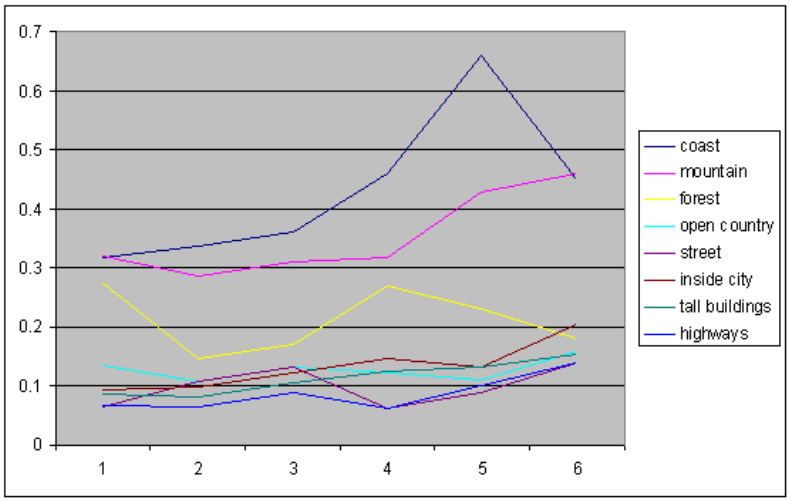

Fig. 4. The achieved mAP for all Torralba concepts, while increasing the number of the closest region types K and the size of visual vocabularies N_T.

especially with respect to concept *sea*, which is to be explained due to the actual nature of the concepts themselves. More specifically, concepts *vegetation* and *sky* performed also very well (i.e. mAP above 0.70), whereas mAP values obtained for concepts *road* and *sand* were average. On the other hand, mAPs for concept *wave* was not as high. This result can again be explained based on the actual visual properties of the particular

	Nt=150, K=1	Nt=150, K=2	Nt=150, K=4	Nt=270, K=1	Nt=270, K=2	Nt=270, K=5
vegetation	0.641	0.664	0.688	**0.721**	0.688	0.670
road	0.581	0.605	0.628	0.607	0.629	**0.657**
sand	0.497	0.520	0.544	**0.574**	0.544	0.552
sea	0.815	0.838	0.862	**0.891**	0.863	0.786
sky	0.618	0.641	0.665	0.673	0.667	**0.734**
wave	0.405	0.429	0.453	**0.457**	0.451	0.401

Table 2. 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 **Corel dataset**.

concept, i.e. a *wave* is difficult to segment and discriminate in a visual manner. Fig. 5 depicts again the evolution of mAP vs. K and N_T for all Corel concepts. In this case, we observe a more unified distribution of mAPs for higher values of K, which results to rather small variations to the actual values, e.g. ranging from a low of 0.401 up to 0.457 for concept *wave* or a low of 0.786 up to 0.891 for concept *sea*.

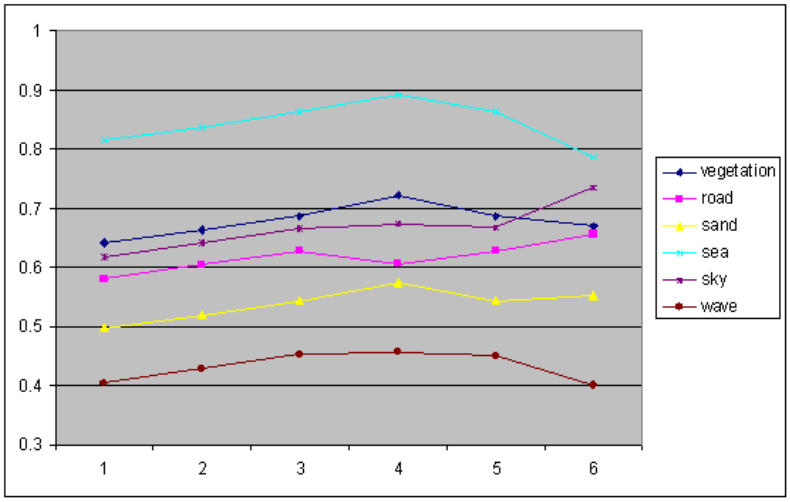

Fig. 5. The achieved mAP for all Corel concepts, while increasing the number of the closest region types K and the size of visual vocabularies N_T.

5 Conclusions

In this paper we presented an approach for semantic image retrieval by exploiting a region-based visual vocabulary. More specifically, we introduced an enhanced bag-of-

words model for capturing the visual properties of images and instead of using the entire vocabulary, we selected a meaningful subset consisting of the closest region types to the image regions. This led to a simple yet effective representation of the image features that allow for efficient retrieval of semantic concepts. Early experimental results on two well-known still image datasets are promising.

References

1. Duygulu, P., Barnard, K., De Freitas, J., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. Lecture Notes in Computer science (2002)
2. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2008)
3. van Gemert, J., Geusebroek, J., Veenman, C., Smeulders, A.: Kernel codebooks for scene categorization. In: European Conference on Computer Vision (ECCV), Springer (2008)
4. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: in NIPS, MIT Press (2003)
5. Zheng, Y., Neo, S., Chua, T., Tian, Q.: Object-Based Image Retrieval Beyond Visual Appearances. Lecture Notes in Computer Science **4903** (2008) 13
6. Yuan, J., Wu, Y., Yang, M.: Discovery of collocation patterns: from visual words to visual phrases. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Volume 1., Citeseer (2007)
7. Avrithis, Y., Doulamis, A., Doulamis, N., Kollias, S.: A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases. Computer Vision and Image Understanding **75**(1) (1999) 3–24
8. Spyrou, E., Le Borgne, H., Mailis, T., Cooke, E., Avrithis, Y., O Connor, N.: Fusing mpeg-7 visual descriptors for image classification. In: International Conference on Artificial Neural Networks (ICANN). (2005)
9. Chang, S., Sikora, T., Purl, A.: Overview of the MPEG-7 Standard. IEEE Transactions on Circuits and Systems for Video Technology **11**(6) (2001) 688–695
10. Spyrou, E., Tolias, G., Mylonas, P., Avrithis, Y.: Concept detection and keyframe extraction using a visual thesaurus. Multimedia Tools and Applications **41**(3) (2009) 337–373
11. Mylonas, P., Spyrou, E., Avrithis, Y., Kollias, S.: Using Visual Context and Region Semantics for High-Level Concept Detection. IEEE Transactions on Multimedia **11**(2) (2009) 229
12. J.Z. Wang, J. Li, G.W.: SIMPLIcity: Semanticssensitive Integrated Matching for Picture LIbraries. In: IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 23, No.9, IEEE (2001) 947–963