USING ASSOCIATION RULES AND SPATIAL WEIGHTING FOR AN EFFECTIVE CONTENT BASED-IMAGE RETRIEVAL

Ismail Elsayad, Jean Martinet, Thierry Urruty, Taner Danisman, Haidar Sharif and Chabane Djeraba
LIFL-CNRS, Lille 1 University, Villeneuve d'ascq, France

Keywords: SURF, Bag-of-visual-words, Visual phrases, Gaussian mixture model, Spatial weighting.

Abstract: Nowadays, having effective methods for accessing the desired images is essential with the huge amount of digital images. The aim of this paper is to build a meaningful mid-level representation of visual documents to be used later for matching between the query image and other images in the desired database. The approach is based firstly on constructing different visual words using local patch extraction and fusion of descriptors. Then, we represent the spatial constitution of an image as a mixture of n Gaussians in the feature space. Finally, we extract different association rules between frequent visual words in the local context of the image to construct visual phrases. Experimental results show that our approach outperforms the results of traditional image retrieval techniques.

1 INTRODUCTION

In typical Content-Based Image Retrieval (CBIR) systems, it is always important to select an appropriate representation for documents (Baeza-Yates and Ribeiro-Neto, 1999). Indeed, the quality of the retrieval depends on the quality of the internal representation for the content of the documents.

A popular approach (bag-of-visual-words) that appeared recently is to consider images as a collection of quantized local patches. This approach achieves good results in representing variable object appearances caused by changes in pose, scale and translation, etc. Despite the success of the bag-of-visual-words approach in recent studies (Sivic and Zisserman, 2003; Willamowski et al., 2004; Jurie and Triggs, 2005), there are still three important drawbacks and this paper aims to resolve them.

Firstly, most of the local descriptors are based on the intensity or gradient information of images, so neither shape nor color information is used. In the proposed approach, in addition to the SURF descriptor (Bay et al., 2008), we introduce a novel descriptor (edge context) that is based on the distribution of edge points.

Secondly, since the bag-of-visual-words approach represents an image as a collection of local descriptors, ignoring their order within the image, the resulting model provides a rare amount of information about the spatial structure of the image. In this paper we propose a new spatial weighting scheme that consists of weighting visual words according to the probability of each visual word belonging to one of the n Gaussians in the 5 dimensional color-spatial feature space.

Thirdly, the low discrimination power of visual words leads to low correlations between the image features and their semantics. In our work, we build a higher-level representation, namely: **visual phrase** from groups of adjacent words using **association rules** extracted with the *Apriori* algorithm (Agrawal et al., 1993). Having a higher-level representation, from mining the occurrence of groups of lower-level features (visual words), enhances the image representation with more discriminative power since structural information will be added.

The remainder of the article is structured as follows: in Section 2, we describe the method for constructing visual words from images and mining visual phrases from visual words to obtain the final image presentation. In Section 3, we present an image similarity method based on visual words and visual phrases. We report on the experimental results in Section 4, and we give a conclusion to this paper in Section 5.

Elsayad I., Martinet J., Urruty T., Danisman T., Sharif H. and Djeraba C. (2010).
USING ASSOCIATION RULES AND SPATIAL WEIGHTING FOR AN EFFECTIVE CONTENT BASED-IMAGE RETRIEVAL.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 112-117
DOI: 10.5220/0002836101120117

2 IMAGE REPRESENTATION

In this section, we describe different components of the chain of processes in constructing the image representation (see Figure 1).

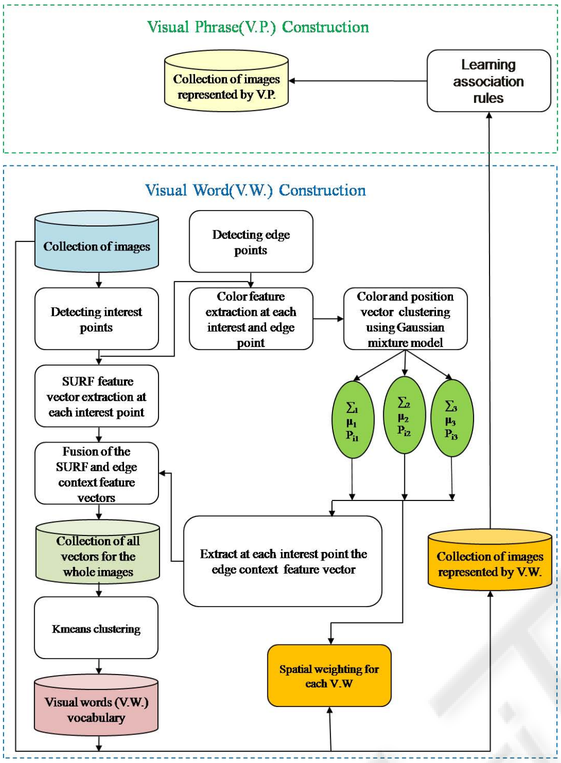
Visual Phrase(V.P.) Construction
Collection of images represented by V.P.
Learning association rules
Visual Word(V.W.) Construction
Collection of images
Detecting edge points
Detecting interest points
Color feature extraction at each interest and edge point
Color and position vector clustering using Gaussian mixture model
SURF feature vector extraction at each interest point
Σ1 μ1 Pi1
Σ2 μ2 Pi2
Σ3 μ3 Pi3
Fusion of the SURF and edge context feature vectors
Collection of all vectors for the whole images
Extract at each interest point the edge context feature vector
Collection of images represented by V.W.
Kmeans clustering
Spatial weighting for each V.W
Visual words (V.W.) vocabulary

Figure 1: Flow of information in the visual document representation model.

2.1 Visual Word Construction

We use the **fast Hessian detector** (Bay et al., 2008) to extract interest points. In addition , the **Canny edge detector** (Canny, 1986) is used to detect edge points. From both sets of interest and edge points, we use a clustering algorithm to group these points into different clusters in the 5 Dimensional color-spatial feature space (see the visual construction part in Figure 1). The clustering result is necessary to extract our edge context descriptor (to be discussed later in this paper) and to estimate the spatial weighting scheme for the visual words.

2.1.1 Gaussian Mixture Model

In our approach, based on **Gaussian mixture model** (GMM) (Bilmes, 1997), we model our color and position feature space to cluster the set of interest and edge points in different clusters. The objective is to cluster the salient structure of the image on some information that can be extracted from the image, rather than intensity and the appearance information that is used in the description process. In addition, by using the GMM, we present a novel spatial weighting scheme for visual words as follows:

Firstly, a 5 Dimensional color-spatial feature vector, built from the 3 Dimensional RGB color features plus 2 Dimensional (x,y) spatial position, is created to represent each interest and edge point. In an image with m interest/edge points, a total of m feature vectors: $Z_1,...,Z_m$ can be extracted.

The set of points is assumed to be a mixture of n Gaussians in the 5 Dimensional color-spatial feature space and the Expectation-Maximization (EM) algorithm is used to iteratively estimate the parameter set of the Gaussians. The parameter set of Gaussian mixture is: $\theta = \{\mu_i, \Sigma_i, P_i\}, i = 1, ..., n$ where μ_i is the mean of the i$^{\text{th}}$ Gaussian cluster, Σ_i is the covariance of the i$^{\text{th}}$ Gaussian cluster and P_i is the prior probability of the i$^{\text{th}}$ Gaussian cluster.

By applying Bayes theorem at each E-step, we can estimate the expected value of the log likelihood function, with respect to the conditional distribution of β_i (denotes the Gaussian which Z_j come from under the current estimate of the parameters $\theta(t)$).

$$P(\beta_i/Z_j,\theta(t)) = \frac{P(Z_j/\theta(t))P(\beta_i/\theta(t))}{P(Z_j)} \quad (1)$$

$$P(Z_j) = \sum_{k=1}^{n} P(Z_j/\beta_k,\theta(t))P(\beta_k/\theta(t)) \quad (2)$$

At each M-step, the parameter set θ of the n Gaussians is updated to maximize the log-likelihood

$$Q(\theta) = \sum_{j=1}^{m}\sum_{i=1}^{n} P(\beta_i/Z,\theta(t)) lnP(Z_j/\beta_i,\theta(t))P(\beta_i/\theta(t)) \quad (3)$$

At the final step of the EM algorithm, we obtain all the parameters needed to construct our set of Gaussians. Then each point is assigned to one of the Gaussians.

2.1.2 Extracting and Describing Local Features

In our approach, we use the SURF low-level feature descriptor that describes how the pixel intensities are distributed within a scale-dependent neighborhood of each interest point detected by the Fast-Hessian. This approach is similar to the SIFT one (Lowe, 2004), but integral images (Viola and Jones, 2001) are used in conjunction with filters known as Haar wavelets in order to increase robustness and decrease the computation time. Haar wavelets are simple filters which can be used to find gradients in the x and y directions. The extraction of the descriptor can be divided into two distinct tasks. Each interest point is assigned to a reproducible orientation. Then a scale-dependent window is constructed in which a 64 Dimensional vector is extracted. In order to achieve scale-invariant results, it is important that all calculations for the descriptor are based on measurements relative to the detected scale. In addition to the **SURF** descriptor, we introduce a novel **Edge context descriptor** at each interest point detected by the Fast-Hessian, based on the distribution of the **edge** points in the same Gaussian (by returning to the 5 Dimensional color-spatial feature space).

Our descriptor is inspired by the **shape context** descriptor proposed by (Belongie et al., 2002) with regard to extracting information from edge point's distribution. Describing the distribution of these points enriches our descriptor with more information, rather than the intensity that is described by SURF. Moreover, the distribution over relative positions is a robust, compact, and highly discriminative descriptor. As shown in Figure 2, vectors from each interest point in the 2D spatial image space are drawn point to all other edge points (that are within the same cluster in 5 Dimensional color-spatial feature space). Then the edge local descriptor for each interest point is represented as a **histogram** of 6 bins for R (*magnitude* of the drawn vector from the interest point to the edge points) and 4 bins for θ (*orientation angle*). For this novel descriptor many invariance can be applied:

Firstly, invariance to translation is intrinsic to the edge context definition since the distribution of the edge points is measured with respect to fixed interest points.

Secondly, invariance for scale can be achieved by normalizing the radial distance by a mean distance between the whole set of points within the same Gaussian in the 5 Dimension color-spatial feature space.

Thirdly, invariance for rotation is achieved by measuring all angles relative to the tangent angle of each interest point.

Following the visual construction part in Figure 1, after the extraction of the edge context feature, fusion between this descriptor and the SURF descriptor is performed. This mixed feature vector is composed of 88 dimensions (64 from SURF + 24 from the edge context descriptor). The new feature vector describes information on the distribution of the intensity and the edge points of the image at the same time. This enriches our image representation with more local information.

Visual words are created by clustering the mixed feature vectors (SURF + edge context feature vector) in order to form a visual vocabulary. We quantize the 88 Dimensional feature vector space by assigning to each observed feature the closest visual word.

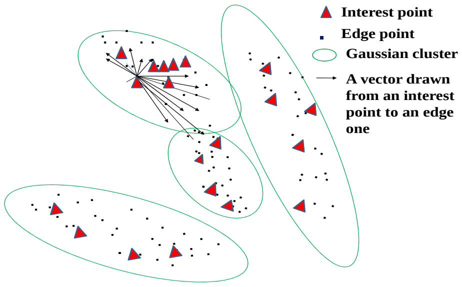

Figure 2: Extraction of the edge descriptor in the 2D spatial space where the points are already clustered before in the 5 Dimensional color-spatial Gaussian space.

2.1.3 Spatial Weighting for the Visual Words

For the ***spatial weighting***, we used an adapted scheme from the one used by (Chen et al., 2009) which differs from ***tf-idf*** weighting scheme. Supposing that in an image there exist local descriptors obtained from the interest point set $\{1, 2, ..., n_i\}$ that belong to the same Gaussian and are assigned to a visual word w_l where $1<l<k$ and k is the number of visual words in the visual vocabulary. Then the summation of the probabilities of the occurrence of the salient points will indicate the contribution of the visual word w_l to the Gaussian β_i. Therefore, the weighted term frequency ($Tf_{w_l\beta_i}$) of a visual word w_l with respect to Gaussian β_i is defined as follow:

$$Tf_{w_l\beta_i} = \sum_{m=1}^{n_i} P(\beta_i/Z_m) \tag{4}$$

The average weighted term frequency (Tf_{w_l}) of w_l with respect to an image I where w_l occurs in the n_{w_l} Gaussians is defined as follow:

$$Tf_{w_l} = \sum_{i=1}^{n_{w_l}} (Tf_{w_l\beta_i})/n_{w_l} \tag{5}$$

The weighted inverse Gaussian frequency of w_l with respect to an image I with n Gaussians is defined as follow:

$$If_{w_l} = ln\frac{n}{n_{w_l}} \tag{6}$$

The final spatial weight of the visual word w_l is defined by the following formula:

$$Sw_{w_l} = Tf_{w_l} \times If_{w_l} \tag{7}$$

2.2 Visual Phrase Construction

Before proceeding to the construction phase of visual phrases for the set of images, let us examine phrases in text. A phrase can be defined as a group of words functioning as a single unit in the syntax of a sentence and sharing a common meaning. For example, from the sentence ”*James Gordon Brown is the Prime Minister of the United Kingdom and leader of the Labor Party*”, we can extract a shorter phrase ”*Prime Minister*”. The meaning shared by these two phrases is the governmental career of James Gordon Brown.

Images are particular arrangements of patches in a 2D space. Such patches in an image are not independent but are likely to belong to the same physical object with each other and, consequently, they are likely to have the same conceptual interpretation. The inter-relationships among patches encode important information for our perception. Applying association rules, we used both the patches themselves and their inter-relationships to obtain a higher-level representation of the data known as **visual phrase**.

We are not alone in applying the association rules to images. (Martinet and Satoh, 2007) adapted the definition of association rules to the context of perceptual objects in order to merge strongly associated features and get a more compact representation of the data. We apply an adapted version of this to the frequent, consecutive visual words that share the strong association rules and are located within the same local context. All local patches are within the same context whenever the distance between their centers are less than or equal to a given threshold. Considering that the set of all visual words (visual vocabulary) $W = \{w_1, w_2, ..., w_k\}$, D is a database (set of images I), $T = \{t_1, t_2, ..., t_n\}$ is the set of all different sets of visual words located in the same context (see Figure 3).

An association rule is a relation of an expression $X \Rightarrow Y$, where X and Y are sets of items. The properties that characterize association rules are:

- The rule $X \Rightarrow Y$ holds in the transaction set T with support s if s % of transactions in T contains X and Y.
- The rule $X \Rightarrow Y$ holds in the transaction set T with confidence c if c % of transactions in T that contains X also contains Y.

Given a set of documents D, the problem of mining association rules is to discover all strong rules, which have a support and confidence greater than the pre-defined minimum support (*minsupport*) and minimum confidence (*minconfidence*). Although a number of algorithms have been proposed to improve various aspects of association rule mining, Apriori (Agrawal et al., 1993) remains the most commonly used.

Since our aim is to discover the inter-relationships between different visual words, we consider the following:

- W denotes the set of items.
- T denotes the set of transactions.
- X and Y can be sets of one or more of frequent visual words that are within the same context.

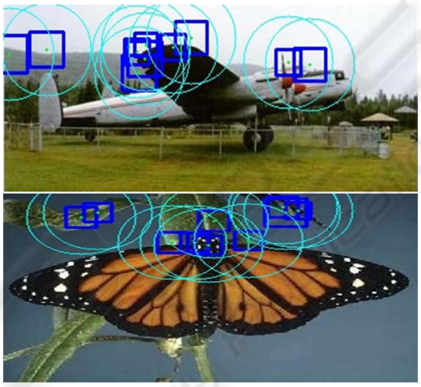

Figure 3: Two sample images where 10 randomly chosen visual words are represented in the local context for each one. The square resembles a local patch, which denotes one of the visual words, and the circle around the center of the patch denotes the local context for this visual word.

After mining the whole transactions and finding the association rules, all visual words located in the same context and involved in at least one strong association rule will form the visual phrases.

3 IMAGE SIMILARITY MATCHING AND RETRIEVAL

Given the proposed image representation discussed in Section 2, we describe here how documents are matched, by estimating a similarity value from the 2-faceted representation. The traditional Vector Space Model of Information Retrieval (Salton et al., 1975) is adapted to our representation, and used for similarity matching and retrieval of images. The doublet represents each image in the model:

$$d = \begin{cases} \vec{W}_d \\ \vec{P}_d \end{cases} \tag{8}$$

Where $\vec{W}_d$ and $\vec{P}_d$ are the vectors for the word and phrase representations of a document respectively:

$$\vec{W}_d = (w_{1,d}, \ ..., \ w_{n_w,d}) \ , \ \vec{P}_d = (p_{1,d}, \ ..., \ p_{n_p,d}) \tag{9}$$

Note that the vectors for each level of representation lie in a separate space. In the above vectors, each component represents the weight of the corresponding dimension. We used the ***spatial weight scheme*** defined in Section 2.1, for the words and the standard ***td.idf-weighting scheme*** for the phrases. We have designed a simple measure that allows evaluating the contribution of words and phrases. The similarity measure between a query q and a document d is estimated with:

$$sim(q,d) = (1-\alpha)RSV(\vec{W}_d, \vec{W}_q) + (\alpha)RSV(\vec{P}_d, \vec{P}_q) \quad (10)$$

The Retrieval Status Value (RSV) of 2 vectors is estimated with the cosine distance. The non-negative parameter α is to be set according the experiment runs in order to evaluate the representation independently, and a combination of the two representations.

4 EXPERIMENTS

This section describes the set of experiments we performed to explore the performance of the proposed methodology.

4.1 Data Set and Experimental Setup

The image data set used for these experiments is the Caltech101 Dataset1 (Fei-Fei et al., 2007). It contains 8707 images, which include objects belonging to 101 classes. For the various experiments, we construct our test data set by randomly selecting 10 images from each class (1010 images). We randomly select 30 images (different from the test dataset) from each class to build the visual vocabulary (3030 images). We manually choose to set the size of the vocabulary at k=2750.

Firstly, we run experiments with a similarity matching parameter α=0 in order to compare our spatial weighting scheme with other approaches. Then, we evaluate the contribution between words and phrases by running the experiments several times with different values of α. All our experiments have been run on a 3GHz Intel Xeon machine with 3GB memory running under Microsoft Windows XP.

4.2 Evaluation for the Spatial Weighting Performance

We compare the proposed spatial weighting scheme to 2 other approaches, the '***blobworld***' approach (Belongie et al., 1998) and the ***'bag of visual words'*** approach (Sivic and Zisserman, 2003). The *'blobworld'* approach is a well known image representation method, which simply represents images by the parameter sets of Gaussian mixture models. The bag-of-visual-words approach is based on local image patch extraction using SIFT-like region descriptors. For each category, we compute the *average precision* for the top 20 retrieved images. It is clear from the results displayed in Figure 4 that the spatial weighting scheme generally outperforms the other two approaches. In categories in which the image content is highly heterogeneous, exhibiting a lot of textures, and thus being more complicated (such as brain or watch images), our scheme outperforms the others. In categories in which image content is relatively homogeneous (like human face or dolphin images), the bag-of-visual-words approach performs as well as our approach. We noticed that the blobworld approach shows similar results to other approaches only when the image colors are uniform.

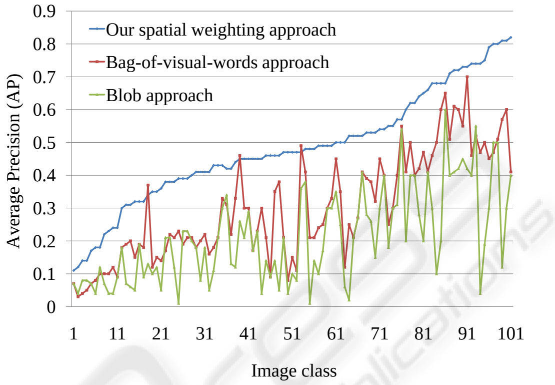

Figure 4: Comparison of image retrieval effectiveness between our spatial weight scheme approach and the bag of visual words and blob approaches. For a clear presentation, we arrange the 101 classes from left to right in the figure with respect to the ascending order of their average retrieval precision.

4.3 Evaluation of the Contribution of Words and Phrases

In the previous section, we demonstrated the good performance of the visual phrase approach. We are now interested in combining visual phrase and visual word approaches by varying the parameter α used in the similarity matching approach. Figure 5 plots the average precision for different values of α over all 101 classes.

When considering only visual phrases in the similarity matching ($\alpha = 1$), the *mean average precision (MAP)* is better than the scenario in which only visual

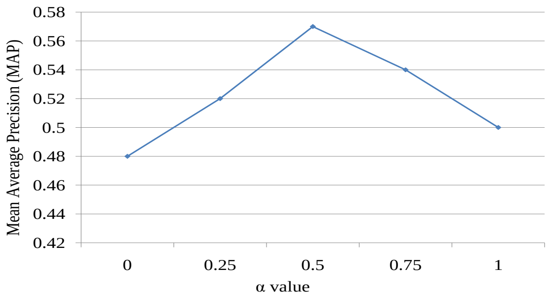

Figure 5: Contribution of words and phrases.

words in the similarity matching ($\alpha = 0$) are taken into consideration. However, the combination of both yields better results than using words or phrases separately.

The explanation is that there are some images, which are not texture-rich like human face, stop sign or umbrella pictures, which leads to detect a small number of interest points. From this study, we conclude that visual phrase alone can not capture all the similarity information between images, the visual word similarity is still required.

5 CONCLUSIONS

A new spatial weighting technique has been developed which enhances the basic bag-of-visual-words approach by using spatial relations. We also devised methods to construct visual phrases based on the association rule technique. Our experimental studies showed that a combined use of words and phrases could perform better than using them separately. It also showed good performance when compared to similar recent approaches.

In our future work, we will perform more studies about the interrelationship between different visual words in order to further investigate the higher representation level. This will improve the discrimination power of the visual words.

REFERENCES

Agrawal, R., Imielinski, T., and Swami, A. N. (1993). Proceedings of the 1993 acm sigmod international conference on management of data, washington, d.c., may 26-28, 1993. In Buneman, P. and Jajodia, S., editors, *Mining Association Rules between Sets of Items in Large Databases*. ACM Press.

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Bay, H., Ess, A., Tuytelaars, T., and Gool, L. J. V. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359.

Belongie, S., Carson, C., Greenspan, H., and Malik, J. (1998). Color- and texture-based image segmentation using the expectation-maximization algorithm and its application to content-based image retrieval. In *ICCV*, pages 675–682.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522.

Bilmes, J. A. (1997). A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698.

Chen, X., Hu, X., and Shen, X. (2009). Spatial weighting for bag-of-visual-words and its application in content-based image retrieval. In *PAKDD '09*, pages 867–874.

Fei-Fei, L., Fergus, R., and Perona, P. (2007). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70.

Jurie, F. and Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *ICCV*, pages 604–610.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Martinet, J. and Satoh, S. (2007). A study of intra-modal association rules for visual modality representation. In *CBMI '07*.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477. IEEE Computer Society.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR 2001*, volume 1, pages I–511–I–518 vol.1.

Willamowski, J., Arregui, D., Csurka, G., Dance, C. R., and Fan, L. (2004). Categorizing nine visual classes using local appearance descriptors. In *In ICPR Workshop on Learning for Adaptable Visual Systems*.