Visual Sentence-Phrase-Based Document Representation for Effective and Efficient Content-Based Image Retrieval

Ismail Elsayad, Jean Martinet, Thierry Urruty, Chabane Djeraba

LIFL/CNRS-UMR 8022-University of Lille 1, France
{ismail.elsayad, jean.martinet, thierry.urruty, chabane.djeraba}@lifl.fr

Abstract. Having effective and efficient methods to get access to desired images is essential nowadays with the huge amount of digital images. This paper presents an analogy between content-based image retrieval and text retrieval. We make this analogy from pixels to letters, patches to words, sets of patches to phrases, and groups of sets of patches to sentences. To achieve a more accurate document matching, more informative features including phrases and sentences are needed to improve these scenarios. The proposed approach is based first on constructing different visual words using local patch extraction and description. After that, we study different association rules between frequent visual words in the context of local regions in the image to construct visual phrases, which will be grouped to different sentences.

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 of the content of documents. A popular technique that was used recently consists in considering images as bag-of-words (Grauman et al., 2005, Jurie et al., 2005, Djeraba 2003, and Sivic et al., 2005). Similarly to document representation in terms of words in text domain, the bag-of-words approach models an image as a bag of visual words, which is formed by a vector quantization of local region descriptors. On one hand, the bag-of-words approach achieves good results in representing variable object appearances caused by changes in pose, scale, translation, etc. On the other hand, the low discrimination power of visual words leads to low correlations between image features and its semantics.

We develop a rich and full-bodied global structure representation of visual documents by considering not only visual words, but also introducing two higher-level representations namely: **visual phrases** and **sentences.** In our approach, we extract scale and orientation invariant local image patches from each image using SURF (Bay et al., 2008). Patches are clustered into different groups to form a visual vocabulary. Images are divided into vertical and horizontal stripes that define local regions where association rule learners are used to discover patterns of visual words that co-occur frequently within these regions. From different visual words that have strong ***association rules*** within these regions, **visual phrases** are constructed. Finally, neighbor phrases that are within the same stripe are grouped into **sentences**.

Compared to the state of the art techniques (Lew et al., 2006), we propose an approach based on visual words, phrases and sentences that maintains the different structural information between local patches and within a set of local patches that are located in image regions. This enriches the presentation with more information and gives a better global structural representation for the whole image.

The remainder of the article is structured as follows. In Section 2, we describe our method for constructing visual words from images and mining visual phrases from visual words that will leads us to build the visual sentences. In Section 3, we present our image similarity method based on visual words, visual phrases and visual sentences. Section 4 concludes the paper.

2 Image representation

In this section, we describe three components of the chain of processes in constructing the visual sentence-phrase representation (see Figure 1).

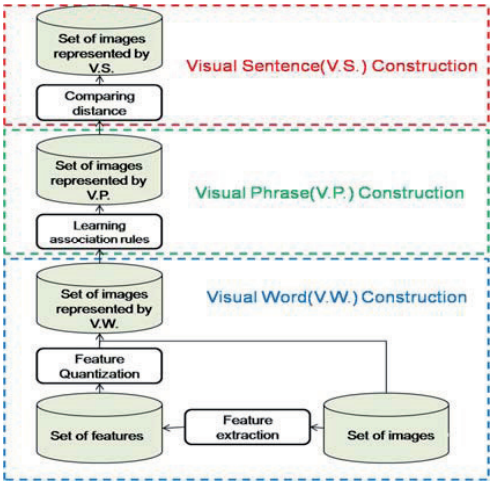

FIG. 1 – Flow of information in the visual document representation

2.1 Visual word construction

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 SIFT (Lowe, 2004), but integral images (Viola and Jones, 2001) used in conjunction with filters known as Haar wavelets are used in order to increase the 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 (see Figure 2). First, each interest point is assigned a reproducible orientation. Secondly, a scale dependent win-

dow is constructed, in which a 64-dimensional vector is extracted. It is important that all calculations for the descriptor are based on measurements relative to the detected scale in order to achieve scale invariant results. Visual words are created by clustering the observed features in order to form a visual vocabulary. We quantize the feature vector space by assigning each observed feature to the closest visual word.

2.2 Visual phrase construction

By returning to text documents, a phrase can be defined as a group of words functioning as a single unit in the syntax of a sentence and having a different meaning taken together or separately. This is also applicable in an image but in a 2D space. In our approach, we segment the image into different local stripes through columns and rows covering the whole image (see the red and green lines in Figure 2). Having an image represented by *visual words*, we examine association rules (Simovic and Djeraba 2008) between different frequent visual words that occur in the same local stripes. Considering that the set of the all visual words (visual vocabulary) is $W= \{w_1, w_2 \ldots w_k\}$, D is the database (set of images I), and $T= \{t_1, t_2 \ldots t_n\}$ is the set of all different sets of visual words located in a same stripe. By returning to the definition of association rules, W denotes the set of items and T denotes the set of transactions. 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 contain X also contain Y.

After mining the whole transaction set and finding the association rules, the association rules are called **strong** if they have support and confidence above ***minsupport*** and ***minconfidence*** respectively. Finally, all ***visual words*** that are within the same stripe and involved in strong association rules will form a ***visual phrase***.

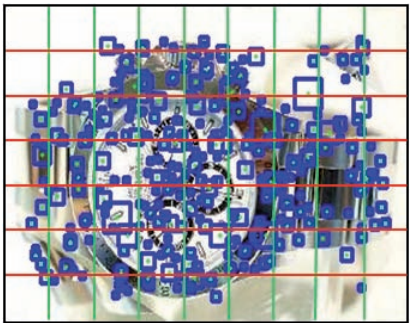

FIG. 2 – *An example of an image after local patches (blue squares) extraction by SURF then it is segmented to different vertical and horizontal stripes (green and red lines).*

2.3 Visual sentence construction

Once visual phrases have been constructed, we can process to the next step for constructing the visual sentences by grouping ***neighbor phrases*** that are within the same stripe. Constructing visual sentences has an intrinsic advantage since a visual sentence can be shared by different objects within an image, and this gives a good representation for the structural relations between different objects, which are not represented by the visual word or phrase.

3 Similarity matching and retrieval

Given the proposed image representation in Section 2, this section describes how images are matched, by estimating a similarity value from the 3-faceted representation. The traditional Vector Space Model (Salton et al., 1975) of Information Retrieval is adapted to our representation, and used for the similarity matching and retrieval of images. The triplet represents each image in the model:

$$d = \begin{cases} \vec{W}_d \\ \vec{P}_d \\ \vec{S}_d \end{cases} \qquad \vec{P}_d = (p_{1,d}, p_{2,d}, \ldots, p_{n_p,d});\ \vec{W}_d = (w_{1,d}, w_{2,d}, \ldots, w_{n_w,d});\ \vec{S}_d = (s_{1,d}, s_{2,d}, \ldots, s_{s_w,d})$$

Where $\vec{W}_d$, $\vec{P}_d$, and $\vec{S}_d$ are the vectors for the word, phrase, and sentence facet in the representations of a document respectively. 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 have used the standard *tf.idf* weighting scheme, where the *tf* and *idf* values are estimated independently for words, phrases, and sentences. We have designed a simple similarity measure that allows evaluating the contribution of words, phrases, and sentences. The similarity measure between a query q and a document d is estimated with

$$similarity \quad (q, d) = \alpha \cdot RSV\ (\vec{W}_q, \vec{W}_d) + \beta \cdot RSV\ (\vec{P}_q, \vec{P}_d) + \gamma \cdot RSV\ (\vec{S}_q, \vec{S}_d)$$

$$\alpha + \beta + \gamma = 1$$

The Retrieval Status Value ***(RSV)*** of 2 vectors is estimated with the cosine. The 3 non negative parameters α, β, and γ are to be set according the experiment runs in order to evaluate the contribution of each representation level independently, and a combination of all representations levels.

4 Experiments

In this Section, we describe a set of experiments dedicated to test the proposed approach. The image dataset used for these experiments is a subset of 1000 images from Caltech101 Dataset1 (Fei-Fei et al., 2004) equally distributed in 10 categories of various objects. We have randomly chosen 10 images from each of these categories and used them as query im-

ages. All experiments have been done on a 3GHz Intel Xeon machine with 3GB memory running Microsoft Windows XP. The algorithms are implemented in C++ using OpenCv library ver1.0. (Bradski and Kaehler, 2008).

To measure the efficiency of our approach, we compute the time used to retrieve the 10 nearest neighbors from each query image. This query processing time includes the time used to retrieve candidate images from the database and the time to rank them. The average query processing time varies. It ranges from less than 10 milliseconds to about 160 milliseconds, depending on the number of visual words, phrases and sentences in the query images. In average overall set of query images, it takes about 48 milliseconds for each category to get the retrieval results.

To evaluate the effectiveness of our visual sentence-phrase-based representation, we try to retrieve images using (i) visual words only (visual word-based), (ii) visual phrases only (visual phrase-based) and (iii) visual words, phrase and sentences at the same time (visual sentence-phrase-based). The effectiveness of each setting is judged by the average precision, which is percentage of the relevant images from the 10 retrieved images in each category (see Figure 3).

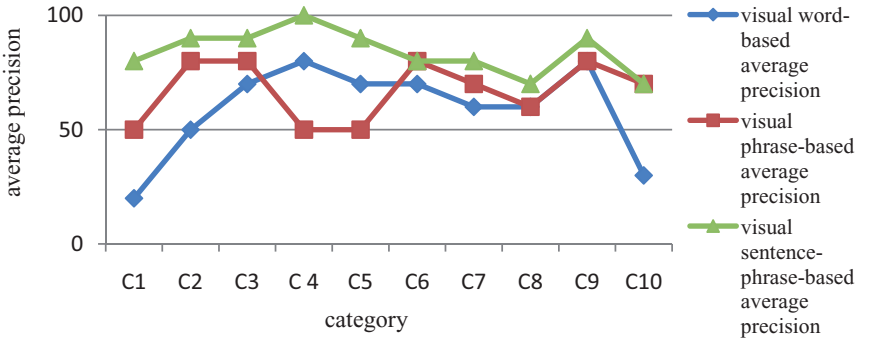

FIG. 3 – *Comparison of the image retrieval effectiveness between visual word-based, phrase-based and visual sentence-phrase-based algorithms*

If we compare the visual word-based and phrase-based approaches, the phrase approach doesn't not outperforms the visual word based in some category because some images do not have enough local patches to create a set of phrases that could stand alone as a good representation for the image. On the other hand, it is obvious that the sentence-phrase based approach outperforms others because the word and phrase approaches are integrated in this approach.

5 Conclusion

We presented a new approach for content-based image retrieval that proposed a chain of processes for constructing visual words, phrases, and sentences. We also presented the retrieval methodology that uses a similarity measure based on visual words, phrases and sen-

tences. Finally, our experimental results demonstrated that the proposed approach could retrieve images efficiently and effectively.

In our future work, we will investigate how to measure image similarity by applying different techniques via different representation levels (words, phrase, and sentences). Moreover, more work has to be done in selecting and combining low-level feature descriptor.

References

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

Bay, H., A. Ess, T. Tuytelaars, and L. V. Gool (2008).SURF: Speed up feature, *Computer Vision and Image Understanding (CVIU),* Vol. 110, No. 3, pp. 346—359.

Bradski, G., and A. Kaehler (2008). Learning OpenCV. *Computer Vision with the OpenCV Library.*

Djeraba, C. (2003). Association and Content-Based Retrieval. *IEEE Trans. Knowl. Data Eng. 15(1): 118-135*.

Fei-Fei, L., R. Fergus, and P. Perona (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. *CVPR 2004, Workshop on Generative-Model Based Vision*. IEEE Computer Society.

Grauman, K., and T. Darrel (2005). The pyramid match kernel: discriminative classification with sets of image features. *Proceedings of International Conference on Computer Vision, pp.1458–1465.* IEEE Computer Society, ICCV 239. USA.

Jurie, F., and B. Triggs (2005). Creating efficient codebooks for visual recognition. Proceedings *of International Conference on Computer Vision.* Washington, DC, USA.

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

Lew, M. S., N. Sebe, C. Djeraba, and R. Jain (2006). Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP 2(1): 1-19.* Burrus, C.S., R.A.

Sivic, J., B. Russell, A. Efros, A. Zisserman, and W. Freeman (2005). Discovering objects and their location in images. *ICCV*, volume 1, pages 370–377.

Simovic, D., C. Djeraba (2008). Mathematical tools for data mining. Set Theory, Partial Orders, Combinatorics, Series Advanced Information and Knowledge Processing. *Springer, ISBN 978-1-84800-200-5.*

Salton, G., A. Wong, and C. S. Yang (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, vol. 18, nr. 11, pages 613–620.

Viola, P., and M. Jones (2001). Rapid object detection using a boosted cascade of simple features. *International conference in computer vision and pattern recognition (CVPR 2001).*