# 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**.