Journal of Visual Communication and Image Representation 49 (2017) 315–331

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Exploring visual dictionaries: A model driven perspective☆

Sinem Aslan[a,*], Ceyhun Burak Akgül[b], Bülent Sankur[b], E. Turhan Tunalı[c]

[a] International Computer Institute, Ege University, İzmir, Turkey
[b] Electrical and Electronics Engineering Department, Boğaziçi University, İstanbul, Turkey
[c] Department of Computer Engineering, İzmir University of Economics, İzmir, Turkey

ARTICLE INFO

Keywords:
Model-driven
Visual dictionary
Bag of Visual Words
Shape models
Primitive image structures
Image understanding
Object recognition
Scene classification

ABSTRACT

Good representative dictionaries is the most critical part of the BoVW: Bag of Visual Words scheme, used for such tasks as category identification. The paradigm of learning dictionaries from datasets is by far the most widely used approach and there exists a plethora of methods to this effect. Dictionary learning methods demand abundant data, and when the amount of training data is limited, the quality of dictionaries and consequently the performance of BoVW methods suffer. A much less explored path for creating visual dictionaries starts from the knowledge of primitives in appearance models and creates families of parametric shape models. In this work, we develop shape models starting from a small number of primitives and develop a visual dictionary using various nonlinear operations and nonlinear combinations. Compared with the existing model-driven schemes, our method is able to represent and characterize images in various image understanding applications with competitive, and often better performance.

1. Introduction

The Bag-of-Visual Words (BoVW) paradigm provides state-of-the-art performance for tasks of object recognition, image category determination, and in general scene understanding. BoVW methods use visual words, extracted for instance, from Scale-Invariant Feature Transform (SIFT) vectors as mid-level representations of image patches. It is important in the BoVW framework to obtain good representative dictionaries. A plethora of visual dictionaries have been generated in the literature according to the following three paradigms:

1. *Dictionaries built from mixtures of the column set of known transform matrices*, such as DCT [1], DWT [2], Gabor filter [2], curvelet [3], edgelet [4], ridgelet [5], contourlet [6], bandelet [7], and steerable filters [8]. The main advantage of these dictionaries is their realization by means of their fast implementation. However these dictionaries have limitations, i.e., they can only be successful as their underlying model. For example DCT is good at representing images with homogeneous components, DWT is good at representing point singularities and, edgelets, curvelets, ridgelets, contourlets, and bandelets, are good at representing line singularities in images [9].
2. *Dictionaries learned from data.* In a number of methods, one obtains dictionaries directly from pixel data, based on matrix factorization principles under sparsity constraints such K-Singular Value Decomposition (K-SVD) [10] and Online Dictionary Learning (ODL) [11]. Another set of approaches follow the steps of: dense sampling of images, obtaining local features such as HOG [12] or SIFT [13], and building a dictionary via clustering [14]15. These can be grouped under the name of *unsupervised dictionary learning techniques*. Recent studies have introduced supervised dictionary learning [16–22] for better classification performance where a class-specific discrimination term is added to the learning algorithm. The main advantage of both unsupervised and supervised dictionary-learning techniques is that dictionaries can be fine-tuned to the underlying dataset as compared to the transform-based approaches. Furthermore, results in the literature indicate that better performance can be achieved. Their main disadvantage is that unsupervised techniques result in an unstructured dictionary, and they are computationally costlier to apply compared to the transform-based ones. Supervised techniques are more discriminative than unsupervised ones and better in classification tasks, yet they still have some drawbacks, e.g., very large sized dictionaries may be encountered in [16,17]. Supervised pruning of the dictionary, initially learned without supervision, does not necessarily improve the performance [18,19]. The related optimization problem is non-convex and can become quite complex as in [20–22].

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: sinem.aslan@ege.edu.tr, siinem@gmail.com (S. Aslan), cb.akgul@gmail.com (C.B. Akgül), bulent.sankur@boun.edu.tr (B. Sankur), turhan.tunali@ieu.edu.tr (E. Turhan Tunalı).

http://dx.doi.org/10.1016/j.jvcir.2017.09.009
Received 3 October 2016; Received in revised form 27 July 2017; Accepted 19 September 2017
Available online 25 September 2017
1047-3203/

3. *Dictionaries that are crafted on models of local image appearances.* These are typically models of gray-level image topological features, such as *ramps*, *corners*, *wedges*, *bars*, *crosses*, *saddles*, *mesas*, *valleys*, *potholes*, *valleys*, *depressions*, *gorges*, *ridges*, and *flat zones*, etc. This technique has been a much less explored path for creating visual dictionaries. Marr's studies in 1980s [23,24] can be accepted as the beginning of describing natural images in terms of a geometrical structures set. Inspired from the findings in physiology [25–27], Marr claimed that in order to achieve visual perception for machine vision systems, some primitive shape structures such as edge, bar and blob should be detected on the images firstly. Recently, Griffin et al. [28–30] have introduced a dictionary construction method, where images are described in terms of a pre-determined dictionary of merely 7 basic qualitative structures, that are *flat*, *dark* and *light bar*, *dark* and *light blob* and *saddle*, called as Basic Image Features (BIFs). The shape models are defined by a parametric mapping from a jet space to a partitioned orbifold. These authors have subsequently enriched their coarse dictionary by replicating BIFs in different orientations, though its performance in object categorization tasks was far from being competitive [29].

We believe that model-based dictionary methods have further room for exploration and improvement. The potential for improvement lies in a more detailed quantization of the parameter space of the shape models as well as exploring new representative shape types. This paper presents our work in this direction.

Fig. 1 shows the three main operations in the pipeline of visual dictionary construction. These operations are (i) Feature extraction, (ii) Descriptor computation, and (iii) Signature extraction.

Feature extraction. In the first stage, characteristics of the local patches around selected image points can be used. The simplest image feature can be the vector of pixel values or their histogram within a patch. However, raw pixel values are sensitive to position, illumination, and noise variations, or geometrical transform effects. Thus, image features have been developed in the literature [31,32], that, if not totally invariant, mitigate spatial and/or photometric transformations. One can use HOG [12], SIFT [13], GLOH [31], SURF [33], etc. features, on sparse points of interest or densely sampled points on a regular grid [34]. Other possibilities consist of the family of filter kernels, e.g., steerable filters [8], and Gabor filters [35,36]. Derivative-based features investigated by Koenderink and van Doorn [37] are some other examples. These have been used successfully in many applications such as image coding [38], foreground/background segmentation [39,40], moving object detection [41], pose estimation [42] or image registration [43]. Recently, binary features, i.e., BRIEF [44], ORB [45], BRISK [46], FREAK [47], have attracted some attention, due to their computational simplicity, memory-efficiency and their inherent robustness against image variability. In the training stage, the local features are processed to extract a visual dictionary (a.k.a. codebook), consisting of code words.

Descriptor computation. The extracted local features are encoded in a descriptor, regarding to their association to the elements (a.k.a. *code words*) of a predetermined visual dictionary (a.k.a. *codebook*). Principal encoding methods that have been used in the literature [48] can be grouped under categories of (i) voting-based methods such as hard-voting [14] and soft-voting [49], (ii) reconstruction-based methods such as sparse coding [50], Local Coordinate Coding (LCC) [51], and Local-constraint Linear Coding (LLC) [52], and (iii) Fisher coding methods [53,54]. Fisher coding and reconstruction-based methods outperform voting-based methods [48]. Among all, Fisher coding is reported as the best performing one, as the Gaussian mixture model (GMM) provides richer information, and it is more robust to unusual, noisy features. However, Fisher coding gives rise to very high dimensional descriptor vectors. Reconstruction-based methods yield a more exact representation of features than voting-based methods, but computational complexity is higher and they are the least robust ones among all as reported in [48].

Signature extraction. Signature vector is a unique representation of the image to enable its similarity comparison with other images. One way to accomplish this is to combine the descriptors occurrences (hits) into a "bag of features" vector. Essentially this is a *spatial pooling* operation. Spatial pooling provides not only compactness of representation, but also, invariance to transformations such as changes in position, and robustness to lighting conditions, noise and clutter [55]. Sum (or average) and max pooling are the two common ways used for this purpose [55,56]. Sum pooling can reduce discriminability since it is influenced strongly by the most frequent features, which may not however be informative as in the stop words case in text retrieval [56]. Max pooling balances can have better discrimination as it focuses on the most strongly expressed features. However, it is not necessarily the best method for every coding scheme. For example it does not perform well with Fisher coding, but works quite well with soft-voting and sparse coding [56]. Furthermore, pooling spatially close descriptors as in *Spatial Pyramid Matching* (SPM) [34] and *macro-features* [57] has been shown to bring substantial improvements.

In this paper, we propose a novel dictionary generation method adopting the model-driven perspective. The proposed dictionary based scheme, that we call **Sym**bolic **Pa**tch **D**ictionary (**SymPaD**), follows the steps of BoVW paradigm in that, pixels are visited on a dense grid, local image characteristics are extracted in terms of shape similarity scores to the dictionary atoms, the scores are pooled, and finally an image signature is obtained. We differ from BoVW schemes in the literature in the generation of our shape dictionary. These shape patterns are generated by mathematical formulae encoding qualitative image characteristics [23,24,58–60,28,29,61].

Our contributions can be summarized in two items. First, our scheme can incorporate any shape primitive in the visual dictionary thanks to its parametric generative function. More importantly, the parametric representation allows a more thorough sampling of the

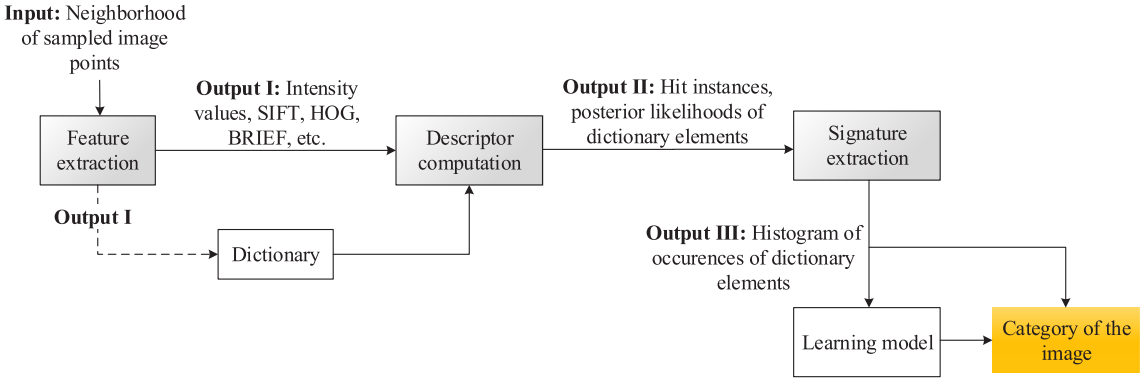

Fig. 1. Processing steps in the pipeline of a dictionary-based computer vision task (a dashed line indicates dictionary learning stage).