Pattern Recognition 84 (2018) 197–210

Contents lists available at ScienceDirect

Pattern Recognition

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

Learning spatial relations and shapes for structural object description and scene recognition

Michaël Clément[a,b,*], Camille Kurtz[a], Laurent Wendling[a]

[a] LIPADE–SIP (EA 2517), Université Paris Descartes, Sorbonne Paris Cité, Paris, France
[b] Centre for Vision Research, York University, Toronto, Canada

ARTICLE INFO

Article history:
Received 25 April 2017
Revised 17 April 2018
Accepted 25 June 2018

Keywords:
Spatial relations
Relative position descriptors
Bags of relations
Structural object description
Hierarchical representation
Force histograms

ABSTRACT

Being able to describe the content of an image, adapted to a particular application, is essential in various domains related to image analysis and pattern recognition. In this context, taking into account the spatial organization of objects is fundamental to increase both the understanding and the accuracy of the perceived similarity between images. In this article, we first present the Force Histogram Decomposition (FHD), a graph-based hierarchical descriptor that allows to characterize the spatial relations and shape information between the pairwise structural subparts of objects. Then, we propose a novel bags-of-features framework based on such descriptors, in order to produce discriminative structural features that are tailored for particular object classification tasks. An advantage of this learning procedure is its compatibility with traditional bags-of-features frameworks, allowing for hybrid representations gathering structural and local features. Experimental results obtained both on the recognition of structured objects from color images and on a parts-based scene recognition task highlight the interest of this approach.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

In the domains of image analysis and pattern recognition, the development of efficient image representations constitutes a challenging problem. Such image representations must be discriminative, robust, and with the increase of the amount of visual data, they must allow for fast processing. Classical methods dedicated to the recognition of images usually rely on a structural or statistical description of their content, summarizing different features such as outer contour, geometry, or texture and color effects. A limit of these methods is that these different types of features (and combinations) are sometimes not discriminant enough to successfully describe images composed of complex objects. This issue is particularly raised when these images are highly spatially and semantically structured.

In recent years, the spatial organization of objects in a scene (or between object subparts) has received much attention in the domain of image analysis. Indeed, the structural relations between image components are fundamental in the human perception of image similarity. Therefore, the spatial relations between the regions composing an object can be considered as important features to recognize the nature of the object itself. However, such features usually suffer from strong structural constraint issues, making them difficult to exploit in different generic object recognition tasks.

In parallel, bags-of-features strategies have been proposed by the computer vision community to efficiently learn and then exploit the discriminative aspects of local features in images. Such popular strategies have led to encouraging results in many image classification tasks. But it is also recognized that one of their inherent downside is their difficulty to take into account spatial information, because images are represented as orderless collections of local features.

Our purpose in this article is to investigate how spatial relations and shape features can be efficiently learned for structural object description and scene recognition. We seek to obtain image representations that are able to capture salient and characteristic spatial configurations occurring across image subparts. To this end, we present a novel bags-of-features framework, based on spatial relations and shapes for the recognition of complex objects in images. The contributions of this article, which is an extended version of preliminary works [1], are the following:

- We propose a generic formalization of the Force Histograms Decomposition (FHD), a graph-based descriptor allowing to characterize the pairwise spatial relations and shapes between arbitrary regions in images;

* Corresponding author at: Centre for Vision Research, York University, Toronto, Canada.
E-mail address: mclement@yorku.ca (M. Clément).

https://doi.org/10.1016/j.patcog.2018.06.017
0031-3203/© 2018 Elsevier Ltd. All rights reserved.

- We present an original strategy to learn FHD descriptors that are tailored for particular image recognition tasks, inspired by bags-of-features strategies. Notably, we show that improvements such as soft assignment methods can also be employed with our structural representations;
- Extensive experimental results are reported on three datasets of structured objects contained in color images and on a large dataset of natural scenes, highlighting the interest of considering the learned structural features in such image classification tasks. In particular, we illustrate that hybrid representations combining both structural and local features can significantly improve classification results.

The remainder of this article is organized as follows. Section 2 reviews some related works in the fields of spatial relations and bags-of-features. Section 3 introduces the proposed generic, graph-based and multi-level descriptor called Force Histograms Decomposition (FHD). Section 4 presents the proposed learning strategy for FHD descriptors resulting in shape and spatial visual words. Experimental validations are reported in Sections 5 and 6, respectively for structured objects and scene recognition. Finally, conclusions and perspectives will be found in Section 7.

2. Related work

2.1. Spatial relations

Many studies have been conducted for the analysis of spatial relations in different application domains of pattern recognition and computer vision, with the common objective of describing the relative position of objects in images [2]. We can distinguish in the literature two main research axes based on strong dual concepts [3]: the concept of spatial relation, and the relative position of an object with regards to another.

In the first axis, a spatial relation such as "*to the left of*" is considered, and a fuzzy evaluation of this relation is obtained for two given objects. For instance, the fuzzy landscape framework [4] is oriented towards this type of evaluations. This approach is based on a fuzzy modeling of spatial relations directly in the image space, using morphological operations. Typical applications include for example graph-based face recognition [5], brain segmentation from MRI [6], or handwritten text recognition [7].

In the second axis, the relative position of an object with regards to another one can have a representation of its own, from which it is then possible to derive evaluations of spatial relations. Different spatial relations can be assessed from this representation and the associated descriptors can be integrated in pattern recognition processes to match similar spatial configurations. A typical relative position descriptor is the Force Histograms [8], which is a generalization of Angle Histograms [9]. Notably, Force Histograms are isotropic and less sensible to discretization issues, while also allowing to explicitly take into account the distance between objects, depending on the application needs. Force Histograms are involved in several application domains such as linguistic descriptions [10], scene matching [11] or content-based image retrieval [12–14]. Note that many other approaches were also proposed for modeling more specific spatial relations such as the "*surrounded by*" relation [15], the "*between*" relation [16,17], the "*enlaced by*" relation [18]. Recent works introduced the ϕ-descriptor [19,20], which provides a generic framework to assess any spatial relation from a set of specific operators, based partially on Allen intervals [21]. This descriptor provides an important advancement, while requiring an extraction of a set of suitable operators dedicated to each usual spatial relation.

Going back to the model of Force Histograms, the authors of [13,14] introduced a structural object descriptor called Force Histogram Decomposition (FHD) in the particular context of image retrieval. The key idea of this descriptor is to encode the pairwise spatial relations between disjoint layers of pixels composing an object, obtained with an image segmentation strategy, using an homogeneous set of F-Histograms. These preliminary works showed the interest of this structural representation based on spatial relations for object description.

However, this approach suffered from different problems. First, it is required to *a priori* determine and fix the number of object subparts (or image regions) to run the decomposition step of this descriptor, which is a major problem when dealing with automatic image recognition tasks. Secondly, the comparison of two objects described with FHD descriptors is viewed as a costly graph matching problem – where nodes represent object subparts and edges represent pairwise spatial relations – implying strong structural constraints on the representation and a high sensitivity to image segmentation issues.

In this context, we propose to consider a multi-level FHD descriptor that allows to characterize objects with different numbers of subparts and at different scales of decomposition. Such a representation is naturally more adjustable to deal with images in databases containing a highly heterogeneous semantic content. Besides, to overcome the previously mentioned structural constraints, we propose to adapt this descriptor to a bags-of-features representation model.

2.2. Towards bags-of-relations

Bags-of-features strategies, originally proposed for text retrieval problems, have recently attracted numerous research attentions for object recognition and image classification tasks [22,23]. Due to their computational simplicity, they have achieved several scalability breakthroughs in image classification tasks. Typical bags-of-features approaches use local features (e.g., SIFT [24] or HOG [25] descriptors), either based on sparse interest points or in a dense grid, to learn with a clustering technique a vocabulary of visual words adapted to a specific application. An image is then represented by a composition histogram of such visual words, and these feature vectors form the basic visual entity for image classification, using a supervised machine learning procedure (e.g., SVM [26] or Random Forests [27]).

An inherent downside of classical bags-of-features strategies is the lack of spatial information, because these models generally represent images as orderless collections of local features. To deal with this issue, different works have been proposed to incorporate spatial information into these models. For example, at the descriptor level, the correlogram features take into account the spatial co-occurrences of color features [28]. It has been proposed in [29] to incorporate the relative positions of visual words in generative bags-of-features models. The authors of [30] presented a hierarchical shape and appearance model for human action recognition. It relies on a new part layer between the mixture proportion, and the extracted features capturing the spatial relationships among parts in the layer. For discriminative bags-of-features models, it has been proposed in [31] to perform pyramid matching by partitioning the image into fine regions, and to compute histograms of local features inside each image region iteratively. Another strategy is to enrich local features by injecting their spatial coordinates into the descriptor, normalized by the image dimensions [32,33]. Other works in this direction propose to also include shape information into the classical bags-of-features framework [34,35]. In [36], the authors recently proposed the Bag of Graphs (BoG), a Bag-of-Words model that encodes as graphs the local structures of an object contained in an image. These works have presented different attempts

Download English Version:

https://daneshyari.com/en/article/6938659

Download Persian Version:

https://daneshyari.com/article/6938659

Daneshyari.com