Computer Vision and Image Understanding 141 (2015) 28–37

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Discriminative part model for visual recognition☆

Ronan Sicre*, Frédéric Jurie

CNRS UMR 6072, University of Caen Basse-Normandie, ENSICAEN, France

ARTICLE INFO

Article history:
Received 17 March 2015
Accepted 3 August 2015
Available online 10 August 2015

Keywords:
Computer vision
Image classification
Visual recognition
Part-based models

ABSTRACT

The recent literature on visual recognition and image classification has been mainly focused on Deep Convolutional Neural Networks (Deep CNN) [A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.] and their variants, which has resulted in a significant progression of the performance of these algorithms.

Building on these recent advances, this paper proposes to explicitly add translation and scale invariance to Deep CNN-based local representations, by introducing a new algorithm for image recognition which is modeling image categories as a collection of automatically discovered distinctive parts. These parts are matched across images while learning their visual model and are finally pooled to provide images signatures.

The appearance model of the parts is learnt from the training images to allow the distinction between the categories to be recognized. A key ingredient of the approach is a *softassign*-like matching algorithm that simultaneously learns the model of each part and automatically assigns image regions to the model's parts. Once the model of the category is trained, it can be used to classify new images by finding image's regions similar to the learned parts and encoding them in a single compact signature.

The experimental validation shows that the performance of the proposed approach is better than those of the latest Deep Convolutional Neural Networks approaches, hence providing state-of-the art results on several publicly available datasets.

1. Introduction

The arrival of effective approaches based on Deep Convolutional Neural Networks (Deep CNN), such as the remarkable work of Krizhevsky et al. [1] has been perceived as a new trend in image classification, relegating the not so distant approaches such as the bag-of-words [2–4] or the even more recent Fisher vectors [5] to what some consider now to be a legacy of previous time.

Since then, the literature on *image classification* – the task consists in predicting whether an image contains an object or, more generally, a visual concept based on the content of the image – has benefited from a revival of interest because of the new perspective Deep CNN provides (*e.g.* [6,7,7–9], to cite only a few recent of them).

However, even if Deep CNN obtain very good performance, most of the recent approaches do not explicitly model objects or scenes as deformable configurations which can potentially result in a lack of robustness to appearance/viewpoint changes. One can see this as a limitation, since scenes (and therefore images) can be seen as spatial arrangements of objects or parts, and a decomposition into distinctive parts can results in more expressive and discriminative models [10–12].

One motivation of this paper is hence to bring together the advantages of Deep CNN and part-based model. The results achieved by Oquab et al. [13] constitute one interesting step toward that end. They indeed show that it is possible to transfer image representations learned with CNNs trained on large datasets to different tasks, even in presence of limited training data. Their method uses ImageNet pre-trained layers of CNN to compute mid-level image signature and can be utilized as an efficient feature encoding system. We use this framework as an alternative to Bag-of-words (BOW) or Fisher vector to encode image regions.

Another key issue raised by the representation of images in the context of image classification, is how to efficiently use geometric information and, as aforementioned, how to decompose images into stable and distinctive regions. While the early works were building on pure bag-of-words *e.g.* [2], which consists of pooling the visual features without using their spatial coordinates in any way, it has been shown later (*e.g.* by Lazebnik et al. [4]) that performance can be significantly improved by encoding separately a set of multiple (possibly overlapping) regions, which constitutes a first step toward the use of geometry. Using fixed regions (usually image quad-trees)

☆ This paper has been recommended for acceptance by Xiaogang Wang.

* Corresponding author. Tel.: +33299847486.
E-mail addresses: ronan.sicre@unicaen.fr, ronan.sicre@gmail.com, ronan.sicre@inria.fr (R. Sicre), frederic.jurie@unicaen.fr (F. Jurie).

http://dx.doi.org/10.1016/j.cviu.2015.08.002
1077-3142/

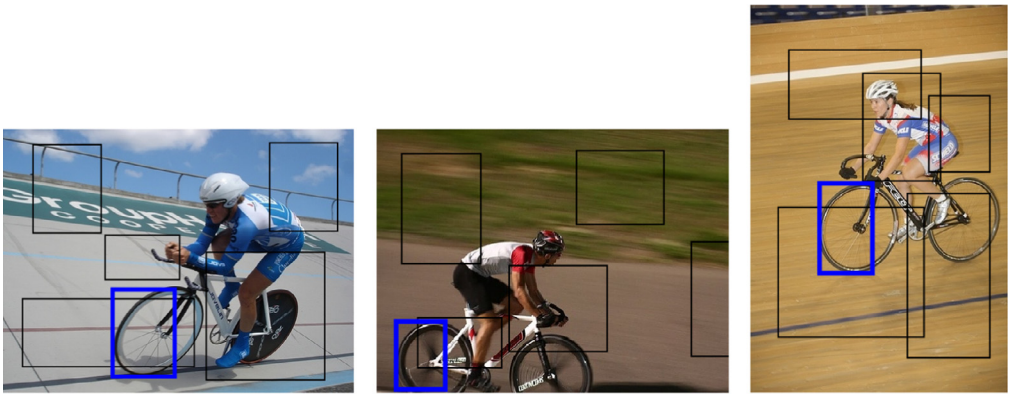

Fig. 1. Our system aims at discovering distinctive parts (blue boxes) from a set of regions (black boxes) randomly extracted from images of a category.

is obviously limited as the corresponding implicit segmentations of the image is not adapted to the image's content. Several more recent works such as [3,14,15] have introduced more flexibility by adapting the shape/position of the regions, but a strong limitation of these works is that the layout of images is still supposed to be fixed, for a given category.

The proposed work starts with the observation that images within a given category can have very different layouts or spatial organization, even if they can be interpreted globally as sharing the same meaning. In line with this observation, several recent works have shown that categories can be efficiently represented by a set of distinctive regions either called *parts* or *fragments* [10–12,16], see Fig. 1. For example, if 'car' images can be recognized because of the joint presence of 'wheel', 'road' or 'window'-like parts, the position of these regions can be any as long as they are in the image. This idea of introducing some invariance (or alignment) with respect to the position of the parts have been successfully utilized in the Deformable Part Model of [17]. However, in the case of image classification the relative position of the parts is much less constrained than in the case of object detection.

In reaction to these observations and concerns, another motivation of our work is precisely to propose a new way to describe images by a set of parts that are aligned across images by construction, without having to use strong geometric constraints between them. This is achieved by proposing a new model for categories, which is based on the fact that (i) a category is defined by a set of K parts (ii) these parts are distinctive in the sense that they occur more frequently in the image of the category than in those from other categories (iii) the presence of regions visually similar to the model's parts is expected in the images of the category. These definitions are implemented into an objective function which is optimized during a learning stage. The objective function relies on a *match* function which automatically discovers and relates model's parts to image regions. Training can be achieved from a set of images describing the category to be recognized, without having to provide any extra annotations. In particular, bounding boxes revealing objects locations are not necessary. During training, a part classifier is learned in conjunction with the alignment of parts to image regions. In a second time, these classifiers can be used to build a global visual descriptor of images, which combines the signatures of the regions discovered in the image. More precisely, the paper proposes three representations: one is obtained by aggregating the Deep CNN signatures of the different image regions, another consists in aggregating the scores of individual part classifiers while the third encodes the distinctive regions of an image with a Fisher vector.

The proposed approach is experimentally validated on three classification datasets. First, Willow [18] aims at classifying seven human actions in still images, while the goal of Boats Datasets is to classify five different categories of boats. Finally the MIT 67 dataset [19] contains images of 67 types of scenes which are to be recognized. These experiments show that not only the proposed method outperform Deep CNN but also that it offers state-of-the-art results on the very competitive MIT 67 dataset.

The rest of the paper is organized as follows. Related work is presented in Section 2, while Section 3 provides details on the proposed system that learns, aligns, and encodes distinctive parts. Finally, the experimental validation is given in Section 4, before concluding the paper.

2. Related work

Image classification. has received a large attention from the computer vision community, *e.g.* see the abundant literature related to the Pascal VOC [20] and ImageNet [21] challenges. A large part of the modern approaches follow the bag-of-word model [2], composed of a four step pipeline: (1) extraction of local image features, (2) encoding of local image descriptors, (3) pooling of encoded descriptors into a global image representation, (4) training and classification of pooled image descriptors for the purpose of object recognition. Several studies evaluated the influence of the first step: the low level features *e.g.* gradient, shape, color, and texture descriptors, such as [22], while other proposed combining different levels (low - mid - high) of information [23]. Regarding the second step: image encoding; Fisher vectors [5] were considered as achieving state-of-the-art performance, in many cases. The third, pooling, step is also shown to provide improvements, and spatial and feature space pooling techniques have been widely investigated [4,24]. Moreover, [3,14] have recently proposed two different strategies for embedding spatial information into the bag-of-words framework. Finally, regarding the last step of the pipeline, discriminative classifiers such as Support Vector Machines (SVM) are widely accepted as the reference in terms of classification performance.

During the last months, the deep CNN approaches have been successfully applied to large-scale image classification datasets, such as ImageNet [21] [1], obtaining state-of-the-art results significantly above Fisher vectors or bag-of-words approaches. These networks have a much deeper structure than standard representations, including several convolutional layers followed by fully connected layers, resulting in a very large number of parameters that have to be learned from training data. By learning these networks parameters on large image datasets, a structured representation can be extracted at an intermediate to a high-level, depending on the extracted layers [25,26]. Deep CNN representation have been recently combined with VLAD descriptors [27] or Fisher vectors [9].

Mid-level features. Several authors have shown the importance of adding intermediate representations [28], also referred as the