Pattern Recognition Letters 84 (2016) 134–141

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Semantic parts based top-down pyramid for action recognition☆

Zhichen Zhao, Huimin Ma*, Xiaozhi Chen

Department of Electrical Engineering, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 24 November 2015
Available online 8 September 2016

Keywords:
Semantic part learning
Spatial arrangement
Action recognition
Convolutional neural network features

ABSTRACT

We focus on the problem of recognizing actions in still images, and this paper provides an approach which arranges features of different semantic parts in spatial order. Our approach includes three components: (1) a semantic learning algorithm that collects a set of part detectors, (2) an efficient detection method that divides multiple images by the same grid and evaluates parallelly, and (3) a top-down spatial arrangement that increases the inter-class variance. The proposed semantic parts learning algorithm captures both interactive objects and discriminative poses. Our spatial arrangement can be seen as a kind of adaptive pyramid, which highlights spatial distribution of body parts in different actions, and provides more discriminative representations. Experimental results show that our approach outperforms the state-of-the-art significantly on two challenging benchmarks: (1) PASCAL VOC 2012 and (2) Stanford-40 (by 2.6% mAP and 5.2% mAP, respectively).

1. Introduction

Action recognition in still images is one of the core topics in computer vision. The task is to identify, from a single image, the action a human is performing. Pose variations make it extremely challenging even when bounding boxes are provided at both the training and test stages.

For action recognition, two main cues have been studied: interactive objects and discriminative pose parts. Some methods are devoted to modeling human-object interactions [3,8,18]. Compared with human body parts, objects are much easier to detect, and can provide contextual information. Most of the other methods capture discriminative poses [1,7,16]. Such methods attempt to learn and distinguish typical poses for specific actions, which also helps a lot. Besides those two cues, spatial description is also critical. Conventional method [14] concatenates region features from a fixed grid in spatial order. Since spatial description highlights part distribution of different actions, it has been widely used as an effective technique.

There are two problems when applying the above approaches. (1) Finding interactive objects and detecting pose parts both provide promising results for action recognition. However, using only objects or poses may not be sufficient to cover all categories of actions. For example, in categories such as “jumping” and “applauding”, interactive objects can hardly be found, thus recognizing these actions relies on pose parts. In categories such as “blowing bubbles” and “smoking” where persons perform very similar poses, interactive objects become main cues. Such observations suggest that using only a single kind of cue is not enough to provide comprehensive and discriminative features. (2) For spatial description techniques, artificially designed grids are not flexible enough when there are many spatially variable image instances, which always causes wide intra-class variance.

In this paper, we tackle the two problems by (1) detecting and employing multiple “semantic parts” to extract discriminative features, (2) arranging features of detected regions in a top-down spatial order. A “semantic part” is defined as any region that provides great contribution to the right recognition, given specific constraints (see Section 4). Such parts are learned to capture both objects and poses. To learn which parts are semantically meaningful and detect them, we employ a coarse-to-fine learning algorithm. For each image, we detect multiple semantic parts and combine them to obtain more comprehensive and discriminative representations. For spatial description, we only choose features of detected parts, and arrange them in a top-down spatial order to increase inter-class variance (see Section 5). We also demonstrate how to extract part features efficiently. Our approach spends 6.5 ms on detecting semantic parts, which accelerates both learning and test processes significantly.

To evaluate our approach, we perform experiments on two challenging action recognition datasets: (1) PASCAL VOC 2012 [4] and (2) Stanford-40 [22]. We show that multiple parts and spatial description help a lot in action recognition. The learned detectors

☆ This paper has been recommended for acceptance by Prof. M. Couprie.

* Corresponding author. Fax: +86 10 62770317.
E-mail address: mhmpub@tsinghua.edu.cn (H. Ma).

http://dx.doi.org/10.1016/j.patrec.2016.08.020
0167-8655/

flexibly find semantic parts in different categories and images (see Fig. 7). Our approach outperforms the state-of-the-art significantly on the two datasets by 2.6% and 5.2% (mean average precision, mAP), respectively.

2. Related work

Action recognition. For action recognition, two main cues have been studied. One is presence of interactive objects. [18] provide a human-centric approach that first localizes person and object, and then models their relationship. [8] employ generic object proposals [21] to find proper interactive objects. The other cue is the presence of discriminative poses. [1] learn a set of pose parts, poselets, which are compound parts consisting of multiple anatomical parts, highly clustered in 3D configuration space. Recently, [7] employ a "deep" version of poselets on head, torso, and legs to extract discriminative features. In this paper we combine both interactive objects and poses to form more comprehensive and discriminative representations.

Spatial description. Spatial description has been widely used to improve performance for extensive tasks. One of the most common spatial description is Spatial Pyramid (SP, [14]). A spatial pyramid divides an image into fixed grids of multiple scales, and concatenates all features extracted in all cells to form the final representation. Semantic pyramid [11] improves conventional spatial pyramid by employing pre-trained detectors to detect head and upper body regions (Fig. 5b), and then concatenates features of these regions. Our approach employs detectors that position parts precisely to reduce intra-class variance, and concatenates features in a top-down spatial order to increase inter-class variance.

CNN features. Compared with handcrafted features [2,15], convolutional neural networks (CNNs) have shown remarkable results on many computer vision tasks, such as image classification [13,20] and detection [5,6]. While being more discriminative, CNN feature also have a very different property: convolution and pooling only operate locally, so the convolutional feature maps keep the relative position relationship of original image patches. With pooling operation, we can obtain all features of all probable patches by once feedforward operation. Our approach utilizes the same architecture to extract part features efficiently.

3. Overview

Fig. 1 outlines our framework. In the training stage, our goal is to find which parts are semantically meaningful and learn the corresponding detectors. Considering that bounding boxes are provided at both the training and test stages, in this paper, all learning and detection operations are implemented within the bounding boxes. Since there is no additional supervision, we start with weak and general detectors, and boost them to be targeting and specific. The learning algorithm follows a coarse-to-fine process (Fig. 1a and b). First, we learn initial detectors (which are essentially SVM weights, see Section 4.2), and search the most discriminative part for each image, obtaining a set of coarse parts. Then we learn intermediate detectors on these parts. Next, we detect the most discriminative parts again using the intermediate detectors, and update them by new detected parts. We repeat the two steps (1) detect semantic parts and (2) update detectors over and over, obtaining fine semantic parts finally. In principle, Step 1 and Step 2 are exchangeable, however, we learn initial weak detectors at first. So from Fig. 1a to Fig. 1b, we begin with weak detectors and end up with a set of fine semantic parts.

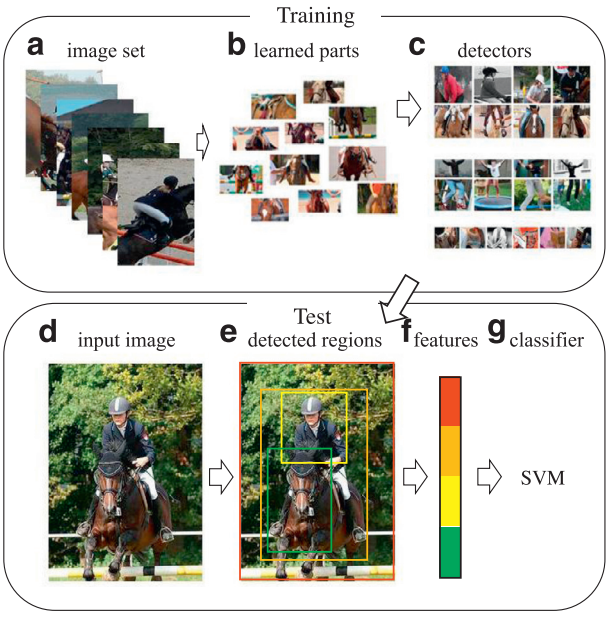

Fig. 1. Overview of our method. First, we employ our semantic learning algorithm to collect a set of parts. Each one is the most discriminative part in the corresponding image (a) and (b). Then, the parts belonging the same category are clustered into M subsets, according to their appearance and locations (b) and (c), here $M = 2$). We learn detectors for each subset and use all detectors of all categories to detect the most discriminative parts in test stage (c) and (e). Finally, features extracted from the whole image (red), the bounding box (orange) and M semantic parts (yellow and green) are concatenated to form the final representation (f). For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

There may be multiple probable semantic parts in the same category. For instance, both raised arms and moving legs are semantic parts for the category "running". In different image instances, sometimes semantic parts are arms, and sometimes semantic parts are legs. We found that the set of fine semantic parts obtained above always contains various kinds of semantics. To generate more targeting and specific detectors, we cluster learned parts into M subsets and learn M final detectors for each category (Fig. 1b–c). That is also why we end up with semantic parts in the previous step.

In the test stage, we detect M semantic parts in each image using the final detectors (Fig. 1e). The chosen parts are the M parts that obtain the maximum detection scores. Features extracted from the whole image, the bounding box and these M parts (from top to down) are concatenated to form the final representation (Fig. 1f). At last, the final representation is fed into a classifier (Fig. 1g).

In the next two sections we describe each step of our approach in details. First we introduce how to obtain part features efficiently based on the structure of CNNs in Section 4.1, which accelerates both the training and test stages. In Section 4.2 we define what a semantic part is, and describe how to find such parts and learn corresponding detectors in the training stage. We demonstrate the detection process in Section 4.3, and its implementation in matrix form in Section 4.4. How to arrange features is discussed in Section 5.

4. Semantic part detection

In this section, we introduce how to calculate part features efficiently, and propose the semantic learning algorithm in the training stage. Furthermore, an accelerated detection process is provided in matrix form.