Adding discriminative power to a generative hierarchical compositional model using histograms of compositions

Domen Tabernik [a,*], Aleš Leonardis [a,b], Marko Boben [a], Danijel Skočaj [a], Matej Kristan [a]

[a] *Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia*
[b] *CN-CR Centre, School of Computer Science, University of Birmingham, Birmingham, UK*

ABSTRACT

In this paper we identify two types of problems with excessive feature sharing and the lack of discriminative learning in hierarchical compositional models: (a) similar category misclassifications and (b) phantom detections in background objects. We propose to overcome those issues by fully utilizing a discriminative features already present in the generative models of hierarchical compositions. We introduce descriptor called histogram of compositions to capture the information important for improving discriminative power and use it with a classifier to learn distinctive features important for successful discrimination. The generative model of hierarchical compositions is combined with the discriminative descriptor by performing hypothesis verification of detections produced by the hierarchical compositional model. We evaluate proposed descriptor on five datasets and show to improve the misclassification rate between similar categories as well as the misclassification rate of phantom detections on backgrounds. Additionally, we compare our approach against a state-of-the-art convolutional neural network and show to outperform it under significant occlusions.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

The problem of visual object categorization and detection has been extensively researched in the last decade with different approaches developed. Many of them include different kind of bag-of-words models [32] or dense features [7,13] combined with a powerful classifiers, however, a lot of promises have also shown biologically inspired hierarchical approaches [1,3,17,19,25,28,36]. Many hierarchical methods [17,36] use compositions to form simple visual features in an increasingly more complex mid-level features throughout the different levels of the hierarchies. Such representation enables a description of higher-level concepts using complex visual features in higher layers of the hierarchy. Hierarchical compositional approaches enjoy a range of advantages for object detection: compact and scalable representations [15], robustness to occlusions [28] and faster and efficient inference [17].

The important aspect of hierarchical methods that enable those benefits is feature sharing. As an inherent characteristic of hierarchical model it allows complex mid-level features to be composed from simple features of lower layers, but those same simple features can also be shared with other complex features in higher layers. For

instance, a shape description of a wheel can easily be used for categories of vehicles, motorbikes or bicycles. Another example is a category of a cow sharing many similar features with a horse. They share similarity in description of a leg, of a head and to certain degree in description of their overall shape (see Fig. 1). The nature of compositions allows hierarchical approaches to efficiently encode such similarities and achieve a compact and scalable representation. Additionally, some hierarchical models [17] also utilize the same power of shareability to lower the computational complexity of detection by introducing fast indexing of features. They implement an efficient bottom-up inference by initiating detection process from simpler lower layer features and gradually inferring more complex features in higher layers. Initial features limit the search space of possible objects that can be found at specific location and can be used as indexes in higher layers to avoid exhaustive search of all objects.

While shareability is important in hierarchical models it also introduces potential problems. When hierarchies rely only on generative learning without explicitly optimizing for discrimination the shareability limits the performance and makes the model prone to two types of problems: (a) misclassifications among similar categories and (b) phantom detections on background objects (see Fig. 1). Both issues are directly related to feature sharing and to the lack of discriminative learning as similar objects are allowed to be composed from the same set of features and no distinctive features are explicitly searched for during the learning. Category misclassifications therefore often occur with objects that have similar compositions, such as

* Corresponding author.
  E-mail addresses: domen.tabernik@fri.uni-lj.si (D. Tabernik), ales.leonardis@fri.uni-lj.si (A. Leonardis), marko.boben@fri.uni-lj.si (M. Boben), danijel.skocaj@fri.uni-lj.si (D. Skočaj), matej.kristan@fri.uni-lj.si (M. Kristan).
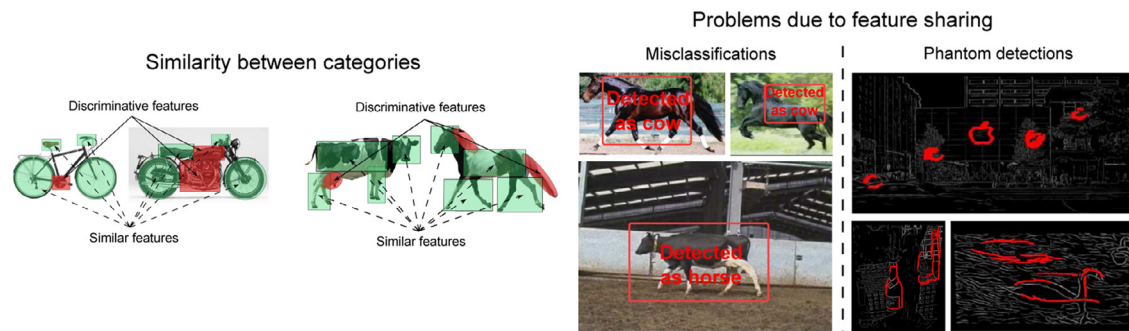
Fig. 1. When two categories have similar features many of them can be shared in a more compact representation (two sets of examples on the left). However two types of problems occur if discriminative features are not considered: (a) misclassifications (example between horses and cows in the middle) and (b) phantom detections (examples of false detections of apple logo, bottle and swan in the right with correct detections highlighted in green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in the case of a cow detector firing on horses. The second problem, phantom detections, is frequently observed in textured and cluttered backgrounds because the presence of different variety of low-level shapes expends the search space to higher layer compositions that are composed from those shared low-level features. Without having an ability to learn and focus on a distinctive features shareability prevents proper classification and severely hinders the overall detection rate in such hierarchies.

In this paper we address the poor discriminative power of hierarchical compositions that perform only generative learning by proposing a descriptor called histogram of compositions (HoC) and show that in combination with a classifier, such as support vector machine (SVM), we are able to reduce the misclassification rate of similar categories as well as eliminate many false detections on backgrounds. We propose to utilize existing generative models produced by the hierarchical method to obtain initial promising detections of objects and then further verify each detection with discriminative features encoded in the HoC descriptor. We refer to the last step as hypothesis verification. With existing detections we fully utilize the category specific information from the highest layers of the hierarchy, while we design HoC descriptor, used to verify detections, to include discriminative features not always present in the highest layers of the hierarchy. We achieve this by re-introducing lower layer features with certain distinctive information capable of discriminating between similar categories and between foreground and backgrounds, and further utilize a classifier to learn discriminative features.

The proposed approach bears several advantages stemming from reuse of hierarchical structure inferred during detection: additional computational time for calculating other types of shapes or compositions can be avoided, detections can be verified only for the category detected and an exhaustive search with sliding windows can be avoided. The latter also enables a computationally more expensive classifiers to be used, while still supporting the scalability for large-scale category detection. HoC descriptor is also applicable to other hierarchical compositional models that allow top-down reasoning and explicitly model compositions.

Our main contribution is HoC descriptor, however, an additional contribution can be found in the provided analysis behind the reasons for low discriminative power in hierarchical compositional method with generative learning where we identify feature sharing in higher layers, reliance on only root feature responses and reliance on only positive features as important factors in poor discriminative power. A preliminary results of our method have been published in two conference papers [24,37].

This paper is structured as follows. Section 2 overviews the related work, Section 3 provides a basic notation for hierarchical compositional model followed by the analysis of the problem of the low discriminative power in Section 4. A histogram of compositions is introduced in Section 5 with the evaluation in Section 6 with final conclusions in Section 7.

2. Related work

Many methods following the hierarchical approach have been developed, however, one method that includes all of the previously presented positive benefits of shareability was introduced by [17] as learned-hierarchy-of-parts (LHOP) model. It is designed as hierarchical compositional model with Gabor filters as first layer features which are gradually combined in specific geometrical configurations into higher layer compositions. In contrast to [3], the features are not hand-coded, but are trained in a generative way based on a statistics of feature co-occurrences found in training images. Additionally, LHOP incorporates fast matching algorithm by using efficient indexing to infer compositions in bottom-up manner as opposed to matching of all features such as in the hierarchical clustering [1] or the hierarchical convolutional network [22,25]. The LHOP method also reduces the exhaustive matching to only lower layer features, while search in higher layers is focused only on features that are supported by lower layers. Additionally, compared to other hierarchical methods [1,19,22], the LHOP model incorporates incremental learning which becomes crucial for large-scale category learning. However, the learning does not explicitly focus on potentially useful discriminative features. A discriminative features can be extracted from the lower layers with a linear classifier when performing classification of a whole image, but such classifier is not used when performing detection and localization with higher layer compositions. Higher layer compositions are only validated on a separate set of images and removed if they produce too many false positives.

Many other generative models have been shown to improve performance when focusing on discriminative features. While [18] do not follow hierarchical approach, they used a generative bag-of-words model and showed to achieve better discriminative performance between similar categories when adding ability to discriminate using an SVM classifier. Similarly, the work of Enzweiler et al. [9] showed the same trend but they approached the problem from different direction. They started with a discriminative model and added a generative one to show on the problem of pedestrian detection to obtain better results when generative and discriminative models are combined.

In shallow hierarchies, such as deformable parts model [13], a state-of-the-art performance is achieved by also incorporating discriminative training of deformable parts. A discrimination was incorporated as latent SVM optimization problem based on a dense histogram of oriented gradients [7] descriptor. The same approach with the latent SVM optimization was used in [30] where they extended the deformable parts model with an additional third layer. While they achieved excellent results their model is still not