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

Domen Tabernik[1], Aleš Leonardis[1,2], Marko Boben[1], Danijel Skočaj[1], Matej Kristan[1]

[1] *Faculty of Computer and Information Science, University of Ljubljana, Slovenia*

[2] *CN-CR Centre, School of Computer Science, University of Birmingham*

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.

Keywords: hierarchical compositional model, feature sharing, discriminative features, LHOP, HoC

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 [33] or dense features [7, 13] combined with a powerful classifiers, however, a lot of promise have also shown biologically inspired hierarchical approaches [20, 3, 1, 18, 29, 26, 37]. Many hierarchical methods [17, 37] use compositions to form simple visual features in an increasingly more complex mid-level features throughout the different levels of the hierarchies. Such representation

Email addresses: `domen.tabernik@fri.uni-lj.si` (Domen Tabernik[1]), `ales.leonardis@fri.uni-lj.si` (Aleš Leonardis[1,2]), `marko.boben@fri.uni-lj.si` (Marko Boben[1]), `danijel.skocaj@fri.uni-lj.si` (Danijel Skočaj[1]), `matej.kristan@fri.uni-lj.si` (Matej Kristan[1])

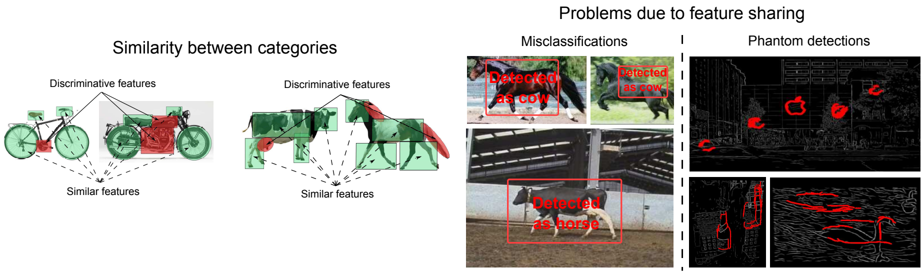


Figure 1: When two categories have similar features many of them can be shared in a more compact representation (two sets of examples on the left). However two types of problems occur if discriminative features are not considered: (a) misclassifications (example between horses and cows in the middle) and (b) phantom detections (examples of false detections of apple logo, bottle and swan in the right with correct detections highlighted in green)

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 [29] and faster and efficient inference [18].

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, Figure 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, Figure 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 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 [38, 25].

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 [18, 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 [23, 26]. 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, 20, 23], 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 [19] 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 [31] where they extended the deformable parts model with an additional third layer. While they achieved excellent results their model is still not hierarchical as features are not build from previous layers as compositions. All benefits of hierarchical principles are missing from such models. The number of layers and the number of parts per layer is fixed and cannot be learnt from natural images, sharing of features is non-existing as each layer is a separate HOG descriptor, and the lack of efficient indexing prevent fast detection, since models of all categories have to be searched and matched. This makes such models difficult to scale to higher number of categories.

A full hierarchical approach was taken by Si and Zhu [37], where they based their model on hybrid image templates (HIT) as initial features, and extended them into hierarchy using AND-OR graphs. They do not model the discriminative features directly, but during the learning process they first create over-fitted AND-OR graph for each training image, and later refine graphs of multiple objects by merging any shared features. With the last step they achieve shareability and compression, while at the same time they are able to indirectly control the degree of discrimination between graphs of different object categories. However, their model lacks incremental learning and does not utilize shareability to implement fast and efficient inference.

The deep neural network hierarchies [22, 23, 30, 26] explicitly address discrimination in the learning process. Their back-propagation learning process relies on a stochastic gradient descent and is tuned to maximize a discriminative objective. Initially, convolutional networks performed poorly in discrimination and a separate layer of Support Vector Machine was needed for a good discrimination [23]. Recently Krizhevsky et al. [26] showed, that excellent results can be achieved without SVM using only fully-connected layers on top of several convolutional layers and using a class specific soft-max output neurons. The main drawback of convolutional networks remains high number of parameters needed to be optimized which can range between 5 and 60 million parameters in certain networks. High number of training images is then needed to converge on the optimal solution thus making the learning computationally expensive. Additionally, connections leading from each feature to every other feature on the previous layer add computational requirements and preclude fast inference since efficient indexing cannot be used to limit the search space of higher layer features. Despite reducing the number of features and restricting to local neighborhood using max-pooling and weights sharing deep approaches still connect each feature map to almost every other feature map on the previous layer, making this approach viable mostly with the specialized hardware such as graphical processing units (GPUs). Alternatively, computational cost could be reduced by approximate convolutional operations [8].

In this work we recognize LHOP as a promising hierarchical compositional method that contains all the benefits of hierarchical compositional approach, however, we further address its poor discriminative power. We approach this problem using a more powerful classifier, a non-linear SVM, and extract important discriminative information using the HoC descriptor created from the hierarchy itself.

3. Learnt Hierarchy of Parts

In this section we first provide basic notation, while we refer the reader to [18, 32] for a more general description of the LHOP model. A quick overview of the model is depicted in Figure 2.

We will denote the vocabulary of hierarchical parts trained for up to L layers as a set of N compositions $\mathcal{L} = \{P_i^l\}_{i=1:N}$, where P_i^l is an identifier of the i-th composition and belongs to the l-th layer of the vocabulary. We will refer to *composition* as a definition of feature that contains a set of sub-compositions and their geometrical constraints stored in the vocabulary, while we will refer to the activation of the composition in the image as *part*. We define connections to other compositions with a set of $Links(P_i^l)$ which holds a list of sub-compositions on layer $l-1$ for each composition P_i^l:

$$Links(P_i^l) = \{(ind_j, P_j^{l-1})\}_{j=1..num\_subparts(P_i^l)}, \tag{1}$$

where P_j^{l-1} is the linked sub-composition on previous layer and ind_j is the local index number for this sub-composition (see, Figure 2).

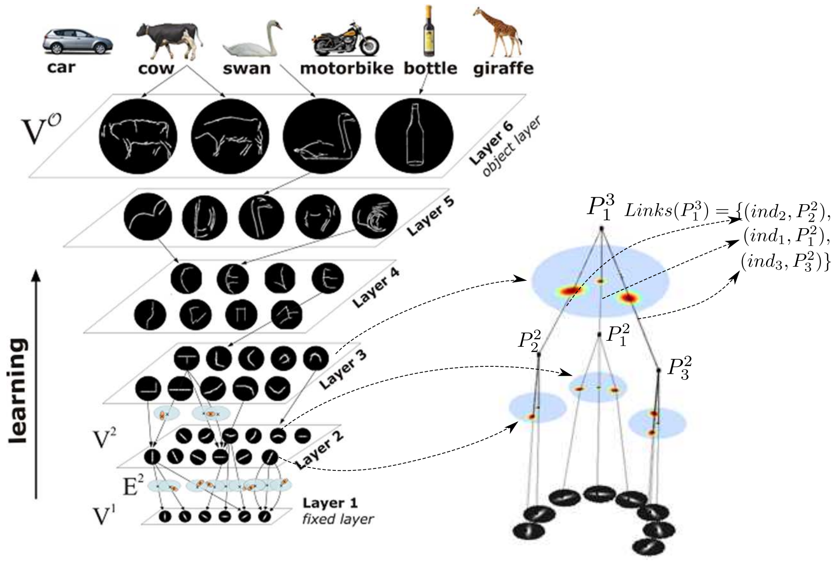


Figure 2: Learnt-hierarchy-of-parts [18] representation with Gabor filters in the first layer and sharing of increasingly more complex features in higher layers with category objects represented in the top layer. The right side of the image depicts one composition from the third layer comprising of three sub-compositions on second layer.

Applying the vocabulary $\mathcal{L}$ on a given image $\mathcal{I}$, the algorithm infers a set of K activations of compositions, i.e. detected parts, $\mathcal{C}(\mathcal{I}, \mathcal{L})$:

$$\mathcal{C}(\mathcal{I}, \mathcal{L}) = \{\pi_k^l\}_{k=1:K}, \tag{2}$$

where k-th detected part on l-th layer $\pi_k^l = [P_k^l, \mathbf{c}_k, \lambda_k]$ is defined by its vocabulary composition P_k^l, its center location $\mathbf{c}_k$ in the image and its response vector $\lambda_k = [\lambda_k^{(S)}, \lambda_k^{(G)}, \lambda_k^{(R)}]$. A response vector contains three separate values where each value is defined recursively from its sub-parts. *A shape's strength value* $\lambda_k^{(S)}$ represents the magnitude of all responses to Gabor features that are connected in the part. *A geometric variability of shape* $\lambda_k^{(G)}$ represents the spatial variability of sub-part relative to the definition of allowed variability in composition P_k^l and *a part realization value* $\lambda_k^{(R)}$ represents a portion of actually detected sub-parts relative to the number of all sub-compositions as defined by the composition P_k^l. We also define a set of links $\Lambda(\pi_k^l)$ as a list of sub-parts pointing to the previous layer $l-1$ parts:

$$\Lambda(\pi_k^l) = \{(\mathit{off}_p, ind_p, \pi_p^{l-1})\}_{p=1..num\_subparts(\pi_k^l)}, \tag{3}$$

where π_p^{l-1} is the linked sub-part on the previous layer, $\mathit{off}_p = (x, y)$ is the offset location relative to part's position $\mathbf{c}_k$ and ind_p is the index number matching to the corresponding sub-composition index in $Links(P_k^l)$. We can use $\Lambda(\pi_k^l)$ to recursively obtain a list of all sub-parts for π_k^l. We term a list of all sub-parts obtained this way *an inferred parse tree*, while using the $Links(P_i^l)$ in the same recursive process would yield a list of all recursively traced vocabulary compositions that we term *a vocabulary parse tree.*

The last layer in the vocabulary is normally category specific and each vocabulary composition P_i^L from layer L directly identifies one trained category. The corresponding $Links(P_i^L)$ at layer L point to compositions that model different views or specific object instances of a category associated with the P_i^L. The inferred parts activated in the image in the last layer L will correspond to detected objects in the image and are defined as a set of detected objects from the given image $\mathcal{I}$:

$$\mathcal{D}(\mathcal{I}, \mathcal{L}) = \{(\pi_j^L, cat_j, r_j)\}_{j=1:J}, \tag{4}$$

where cat_j is a detected category and $r_j = (x, y, w, h)$ is a detection's bounding box. Note that, detection's bounding box r_j is not a simple location of all detected first layer parts, but a more accurate location of the object can be deduced from sub-compositions of corresponding vocabulary composition P_j^L. This can be achieved with a top-down projection by taking the information about the offset from the center of each sub-composition from composition P_j^L and applying those offsets to the correspondingly detected sub-parts in the image to calculate the center and the bounding box of a possible object detection.

4. Understanding low discriminative power in LHOP model

An insight into reasons for low discriminative power in LHOP can be gained by focusing on false detections produced by this method. We can categorize problematic detections into two principal cases as they are depicted in Figure 1:

1. Misclassification of visually similar categories.
2. Phantom detection on background and highly textured object.

4.1. Category misclassifications

Misclassifications occur with objects that have a high degree of visual similarity, such as a similarity between a cow and a horse, or between a bicycle and a motorbike (see, Figure 1). In such cases the hierarchical compositions created during the learning process will naturally share many sub-compositions. When a generative learning process is used the training is focused on finding a general set of features and statistics of spatial feature co-occurrences is used to generate new compositions. This process creates a compact representation of features and has been shown to produce highly compressed set of parts mostly in the lower layers [15]. In lower layers, features have a small receptive field and therefore do not capture complex shapes. In this case, generative learning and sharing of parts is beneficial for the compression as many categories of a different and even visually-dissimilar objects can share as many features as possible. In LHOP, the second and the third layer are therefore trained across all different categories at once and have been shown to learn a stable set of general compositions used across different categories [16]. However, at the higher layers, sharing can introduce ambiguity in discrimination for similar categories. Higher layers are usually trained independently for each category so that the process can focus on a statistics of part co-occurrences for a specific category and produce more complex and category specific parts. Compositions produced this way will be descriptive enough for one category, but very often two categories will share many visual characteristics, such as, for example, in the case of a horse and a cow. In such cases the same compositions produced for one category will also be able to describe another category since the learning process is only generative and not discriminative. This problem can also be viewed from the perspective of granularity at higher layers. For instance, some bicycle wheels and motorbike wheels look sufficiently different and it may be better not to share them, but they are still shaped as circles and at higher layer generative learning will generate similar compositions to describe them. The higher layers therefore lack the granularity to distinguish between such examples, however, this granularity is better at middle or lower layers where there are enough smaller details to make the successful discrimination. Without an ability to focus on such distinctive features, a high degree of sharing prevents the model from discriminating between similar categories, and makes classification highly ambiguous.

4.2. Phantom detections

The second problematic case occurs, similarly as with many other approaches that do not learn with negative samples, when there are too many cluttered objects in the background. Phantom detections may be found in such backgrounds, particularly, when image has many textured objects and they cause first layer features, a responses to Gabor filters, to respond too densely and in too many orientations. In such cases, one location may contain all types of first layer features which brings noise into the inference process. During the inference process a candidate composition is verified if a majority of its sub-compositions satisfy all of the following conditions: (a) matches to one of the allowed sub-composition type, (b) is present around a specific offset from the central sub-composition with allowed degree of variability and (c) has a high enough response vector λ_k values. When too many features are present around a center of a candidate part, all of the above conditions will be met too frequently, which can produce many phantom detections. Additionally, due to the nature of hierarchies false compositions will propagate upwards the hierarchy resulting in exponential increase of false detections. One solution would be to have a more conservative last condition (c) but this also has an effect of eliminating too many correct compositions.

Another reason for phantom detections as well as for category misclassifications can also be attributed to the inference process where only a presence of features is accounted for but the presence of any negative features do not prevent positive compositions from forming. For instance, when a bicycle detector infers an object, it will only be focused on a general features of a bicycle, while the presence of any other features, such as a motor or maybe a petrol tank and exhaust system from a motorbike, will not affect this process. Sometimes this can be beneficial and enables robustness to possible noise around the object. For instance, when a person is riding a bicycle the detector should not consider a person as a negative feature. However, in presence of a motorbike or highly textured background the negative features should be taken into account in the learning process to improve the discrimination.

4.3. Existing approach to discriminative problems in LHOP

In LHOP, a final detection π_j^L found at object layer L is assigned a score from response vector $\lambda_j = [\lambda_j^{(I)}, \lambda_j^{(G)}, \lambda_j^{(R)}]$. However, a response vector is too simplistic to include sufficient information for a successful classification as none of the response values will be fully successful at discriminating. The strength of a shape $\lambda_j^{(S)}$ may be low or high on a correct instance of an object depending on the magnitude of any other objects in the image and is therefore not a good value for discrimination. A geometric variability $\lambda_j^{(G)}$ is also not a good value for discrimination due to a high degree of geometric shape variability within the same category. For instance, a shape of a cow's legs or a head may move in certain directions thus creating variability of a shape that the model must capture for correct description. Additionally, the rotation of an object or a different viewpoint may also introduce certain variability of shape. The third value, a portion of detected sub-compositions $\lambda_j^{(R)}$, performs the best in practice but using this value may also decrease the ability to distinguish between false objects and correct, but slightly occluded, objects.

Using response values of only the top composition in the inferred parse tree, i.e., only the root node of the inferred parse tree, does also reduce the ability for successful discrimination. Although those values are computed recursively from subcompositions of the inferred part tree, any information important for the discrimination may not get fully included. For instance, if a distinction between a cow and a horse is only within a small subset of features, such as features describing an udder, then even if those compositions may have distinctive response values they represent only a small subset of features in the whole inferred parse tree. Based on the definition of response values a small subset of features may not have enough impact on the final response value higher in the hierarchy. To certain degree this approach may mitigate the problem of phantom detections in backgrounds when other positive features present in the inferred parse tree are weak enough, i.e., they have low response values. However, many other phantom detection can not be handled this way, particularly the ones where important negative features are not part of the inferred parse tree.

5. Histogram of Compositions

We address the problem of a low discriminative power by proposing a novel descriptor called Histogram of Compositions (HoC). In the descriptor, we accumulate information of all the compositions from a specific region into a histogram and use a classifier to learn a distinctive features important for the discrimination. Based on observations in Section 4, we focus on accumulating lower layer compositions found within a specific region. This allows us to address three important issues: (i) we can avoid using only root response value which may not contain proper discriminative information present within the features of an inferred parse tree and can focus on compositions that do contain discriminative information, (ii) we address the lack of granularity in higher layer and can capture smaller features that may not have survived all the way up to the higher layers but can still be important for discrimination and (iii) we can capture features that are not part of an inferred parse tree but are located on or in the vicinity of the detected object and can be significant for a proper discrimination.

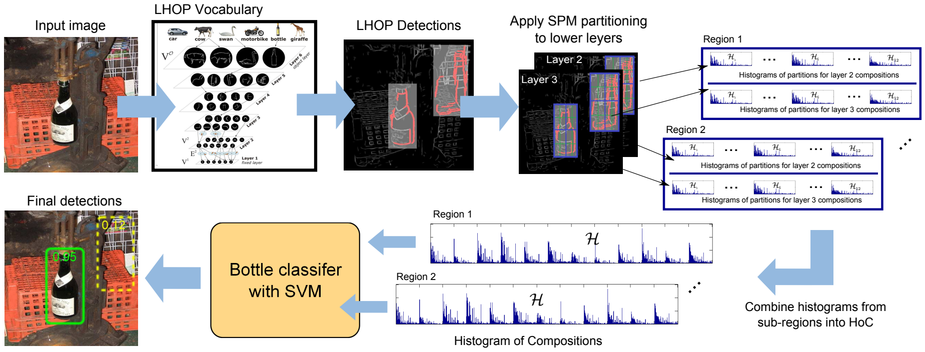

Input image
LHOP Vocabulary
car
cow
swan
motorbike
bottle
giraffe
LHOP Detections
Apply SPM partitioning to lower leyers
Layer 2
Layer 3
Region 1
Histograms of partitions for layer 2 compositions
Histograms of partitions for layer 3 compositions
Region 2
Histograms of partitions for layer 2 compositions
Histograms of partitions for layer 3 compositions
Final detections
0.12
0.95
Bottle classifer with SVM
Region 1
$\mathcal{H}$
Region 2
$\mathcal{H}$
Histogram of Compositions
Combine histograms from sub-regions into HoC

Figure 3: Overview of hypothesis verification with HoC classifier on example of a bottle detection. Possible detection locations are first extracted using LHOP vocabulary. For each region a Histogram of Compositions is composed from the second and the third layer compositions. Additional partitioning in a form of a Spatial Pyramid Match is applied to encode spatial information. Finally, an SVM classifier with HoC features is used to obtain the final probability score of each detection.

5.1. Definition of HoC

In this section we describe the proposed descriptor over a rectangular image region Ω. As a first step, the input image is processed with an LHOP inference process, which activates the set of compositions i.e. detects a set of parts $\mathcal{C}(\mathcal{I}, \mathcal{L}) = \{\pi_k\}_{k=1:K}$ within the region Ω. Histogram of detected parts is then calculated over the library of compositions $\mathcal{L}$ within that region. We provide notation for the parts collected only from the l-th layer. To encode the spatial layouts of the detected parts, the region is split into several subregions, $\Omega = \{\rho_m\}_{m=1:M}$. From each partition ρ_m, a histogram $\mathcal{H}_m$ over the entire library of compositions is extracted. The value of the histogram bin corresponding to the part identifier P_i is defined as

$$\mathcal{H}_m(P_i) = \sum_{\pi_k} w(\pi_k, \rho_m)\delta_{P_i}(P_k), \tag{5}$$

where $\delta_{P_i}(P_k)$ is the Kronnecker delta centered at P_i, and $w(\pi_k, \rho_m)$ is a weighting function that assigns a weight by which a detected part contributes to the histogram bin for the specific composition P_i. The descriptor of a region Ω for l-th layer is defined as a concatenation of all histograms from the subregions, i.e., $\mathcal{H}^{(l)} = [\mathcal{H}_1, \ldots, \mathcal{H}_M]$. The final descriptor $\mathcal{H}$ is formed by a concatenation of histograms from selected layers:

$$\mathcal{H} = [\mathcal{H}^{(l_1)}, \ldots, \mathcal{H}^{(l_N)}], \tag{6}$$

where $[l_1, l_2, ..., l_N]$ is a set of layer numbers used for the final descriptor. In general, an arbitrary set of layers can be used, but as highlighted in the previous section, we focus only on lower layer parts.

5.2. Hypothesis verification using HoC

HoC classifier, an SVM classifier in our case, is used to categorize the region where HoC descriptor was extracted. An overview of *hypothesis verification* is depicted in Figure 3. Initial detections $\mathcal{D}(\mathcal{I},\mathcal{L})$ obtained from the LHOP model are re-used to avoid using sliding windows. HoC descriptor is computed only for LHOP detections and verification is run only for the detected object category. We refer to this step as hypothesis verification. As the number of locations, where HoC descriptor is computed, is limited to only the detected regions, this allows for the use of a more complex radial-based kernel in the SVM. With hypothesis verification we also address phantom detections in background, and can therefore afford to be slightly less conservative in the inference process. This way, potentially more detections are produced, including correct detections, but inference process still has to remain conservative enough to avoid explosion of higher layer parts and consequentially explosion of phantom detections.

We define hypothesis verification for each detected object in $\mathcal{D}(\mathcal{I},\mathcal{L})$ from where category information cat_j and detected bounding boxes r_j is obtained. A detection's bounding box r_j is used as the initial region and HoC descriptor $\mathcal{H}_j$ is computed from the detected parts of lower layers within the region r_j. When extracting compositions for the HoC a vocabulary pre-trained on a general set of images can be used. However, since HoC is based on the same LHOP model that is already used in the detection step, the existing vocabulary can be re-used. An additional step of inference with the second vocabulary is avoided and the time required to perform the inference is reduced. All computed descriptors $\mathcal{H}_j$ are then filtered by the non-linear SVM with a category model c_j. As the final step a non-maximum suppression using a greedy approach is performed.

6. Evaluation

The proposed combination of the HoC descriptor with the LHOP is evaluated on a set of five experiments. First, HoC descriptor is compared against a similar shape-based descriptor, Histogram of Oriented Gradients [7], to asses the performance of descriptor itself. Then, improvement of the discriminative power is evaluated by focusing on the rate of misclassifications between similar categories and on the rate of misclassifications on backgrounds. Experiments are focused on highly shape specific datasets since LHOP is only shape based and does not take into account any information about the texture or color. A comparison to HOG is performed on the *Caltech-101* [12] database while misclassifications are evaluated on the combined database of *Weizmann horses* [4] and *Leeds cows [34]* datasets, on the *MPEG-7 CE-Shape-1* dataset [27] and on the *ETHZ Shape Classes [14]* dataset. As comparison, an additional detection with sliding window and linear HOG classifier is also performed for the latter three datasets. Additionally, we compare our approach to the deep approaches at the end of this section.

6.1. HoC Descriptor

As initial confirmation of our descriptor, a comparison to the Histogram of Oriented Gradients [7] descriptor is performed. HOG descriptor was selected for its popularity

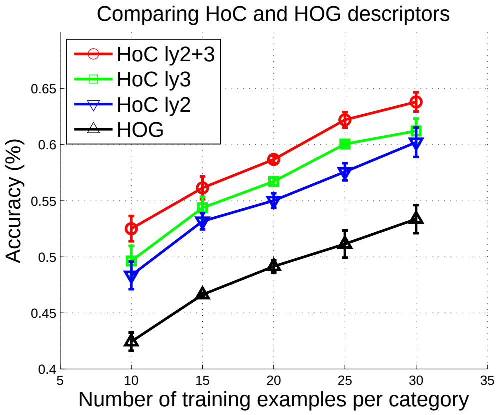


Figure 4: Comparing Histogram of Compositions against a Histogram of Oriented Gradients on *Caltech-101 [12]* dataset. We report results with three versions of HoC: (a) using only second layer parts (blue triangles), (b) using only third layer parts (green squares) and (c) using both second and third layer parts combined (red circles).

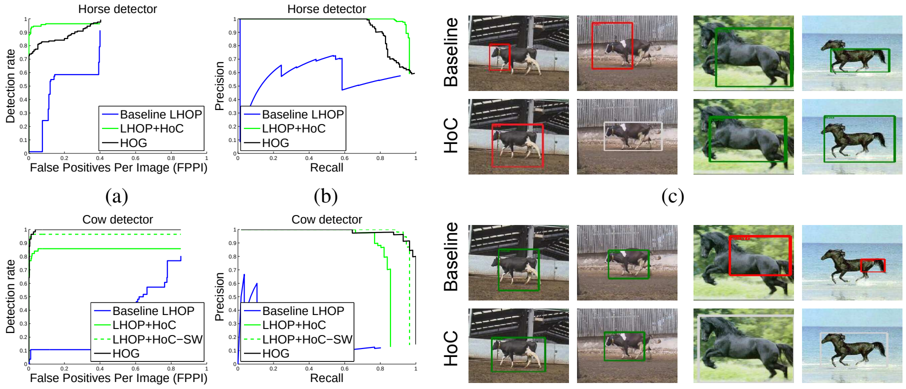
in state-of-the-art detectors [13, 5] as well as due to its similarity to HoC and its shape-specific design. In this evaluation only classification is performed by inferring images only up to the 3th layer and then extracting descriptor on the whole image. We followed the methodology of Vedaldi et al. [41]: different number of examples are used for training, raging from 10 to 30 examples per category, testing is performed on randomly selected 15 examples per category from the rest of the images and finally experiment is repeated 5-times to account for randomness.

We evaluated three versions of HoC descriptors based on different set of LHOP layers: (i) using only the second layer, (ii) using only the third layer and (iii) using both the second and the third layer parts combined. In our preliminary work we also experimented with different types of partitioning for spatial sub-regions and have found Spatial Pyramid Match [28] (SPM) the best performing one. We used SPM with three layers of pyramid, adding to a total of 21 sub-regions. Contribution of each composition π_k in the histogram was weighted by its response vector λ_k, specifically only geometric variability value was used as it performed the best in our preliminary work:

$$w(\pi_k, \rho_m) = \begin{cases} \lambda_k^{(G)} & \mathbf{c}_k \in \rho_m \\ 0 & \mathbf{c}_k \notin \rho_m \end{cases}. \tag{7}$$

To learn libraries of compositions with the second and the third layer, a vocabulary was trained on 250 *general images* which produced the library with 77 compositions on the second layer and 445 compositions on the third layer.

HOG descriptor was computed using the binaries from Dalal and Triggs [7] with 8×8 pixels wide cells and 16×16 pixels wide blocks. For orientation nine binning were used with *L2-Hys* for block normalization schemes. For HOG, all images were resized to 64 × 64 pixels as it proved to perform best in our experiments. This produced a 1764-dimensional HOG descriptor. SVM for both descriptors was implemented using



(a) (b) (c)

(d) (e) (f)

Figure 5: Results of evaluating category misclassification rate between *Weizmann horses* [4] and *Leeds cows [34]* datasets excluding phantom detections on background. The first row shows results for the horse detector with (a) detection-FPPI graph, (b) precision-recall graph and (c) examples of detections with and without HoC classifier. The second row shows similar types of graphs and samples for the cow detector (d-f). Images in (c) and (f) have correct detections with intersection-over-untion intersection over 0.5 shown in green, wrong detections shown in red and ground-truth bounding box when no detection is found shown in gray.

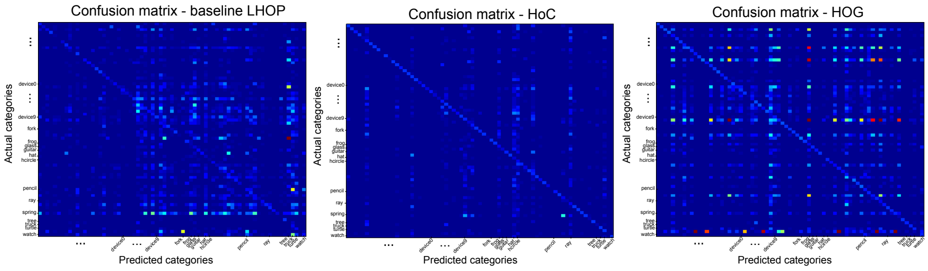
one-vs-all LIBSVM [6] with an RBF kernel using chi-squared distance function (RBF-$\mathcal{X}^2$).

Results of comparison

The results of all three versions of HoC compared to the HOG are shown in Figure 4. All three versions outperform HOG by approximately $5-10$ percentage points, confirming the superior performance of our descriptor. The second layer outperforms HOG by slightly more than 5 percentage points, while the third layer performs even better with additionally $1-2$ percentage points better than the second layer. The best result is obtained with the combined second and third layer which performed by more than 10 percentage points better than the HOG descriptor. Since best performance was achieved by a combination of the second and the third layer, only this version was used in all later experiments.

6.2. Misclassifications between similar categories

The next set of experiments focuses on evaluating discriminative capabilities of HoC descriptor by comparing the performance of only baseline LHOP detections without the HoC and detections verified with the HoC classifier. The LHOP model is first trained for up to 6 layers with additional 7th layer acting as a category identification. During training, all input objects were resized to its diagonal size of 250 pixels, thus creating category models of the same size. In the baseline LHOP a portion of detected



(a) (b) (c)

Figure 6: Confusion matrix for results of *MPEG-7 Shapes* [27] dataset for three methods: (a) LHOP baseline, (b) our method with LHOP and HoC as hypothesis verification and (c) HOG descriptor as additional comparison.

sub-compositions $\lambda_k^{(R)}$ is used as probability score while in HoC classifier a score returned by the SVM is used for its probability score. When learning HoC classifier additional hard-negatives are also mined from the same set of training images. All detections that had an intersection-over-union (PASCAL overlap criteria [10]) factor with the ground truth less than 0.3, were used as negative samples while all detections with the overlap criteria of more than 0.7 were used as additional positive samples. Due to a possible small training sample size, 10% of *Caltech-101* images were sampled for additional hard-negative examples. Detection for hard-negatives, as well as the final evaluation is always performed using the multiscale approach. Firstly, images are resized by a factor of 2.0 and then gradually down-scaled by a factor of $\sqrt{2}$ to produce a maximum of 10 scales per image.

Weizmann horses and Leeds cows dataset

The rate of misclassification between two similar categories is assessed with the performed experiment on *Weizmann horses* [4] and *Leeds cows* [34] dataset both merged together. Each category is evaluated separately by first learning the model for each category on half of the images containing the trained category and testing on all the remaining images of both *Weizmann horses* and *Leeds cows* datasets. For both, the baseline method and the HoC classifier, LHOP learning was performed on masked images provided by both datasets. Ground-truth masks were used for learning as they provided noise-free contour. Note that, masked images were used only during the training process while non-masked images were used for testing.

Since focus of this evaluation is only on a problem of misclassification between similar categories, the detections on background objects that do not intersect with the ground-truth region at all are excluded from the evaluation. We provide results with included background detections in Section 6.3. Results between misclassifications of horses and cows are shown in Figures 5a and 5b for the horse detector and in Figures 5d and 5e for the cow detector. Focusing on horses, an improved ability to discriminate can be seen in both detection-FPPI as well as in precision-recall metrics. While baseline method is able to correctly identify horses with a mean average precision (mAP)

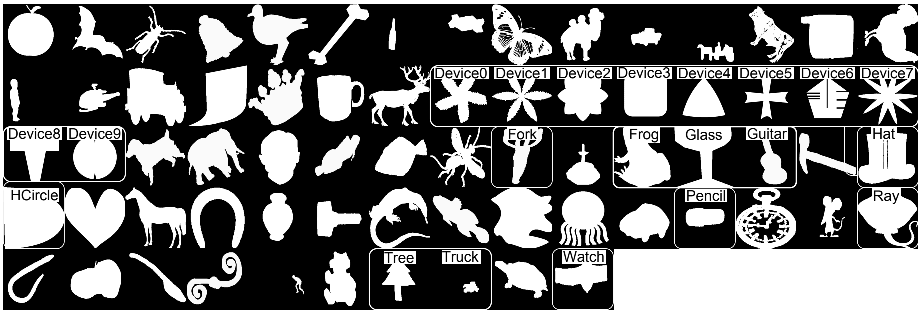


Figure 7: All 70 categories of *MPEG-7 Shapes* [27] dataset. High degree of misclassifications was observed between highlighted classes.

of around 0.61, the inclusion of the HoC classifier is able to improve this to mAP of around 0.96. Similar improvements can be seen in cow detection, where performance in precision-recall, as well as in detection-FPPI metrics are both significantly improved with HoC classifier. In this case, a baseline cow detector is able to detect cows with a mAP of only 0.15 while the HoC classifier increases mAP to 0.84 and therefore significantly boosts the discriminative power. Those results are summarized in the top row of Table 2. Additionally, improved discrimination also increased the overall detection rate as with HoC classifier we are able to detect around 10% more horses and cows. Improved detection rate can be contributed to an improved ability to discriminate using the provided probability score. When the probability score is not discriminative enough, incorrect detections within the object may suppress correct ones during the non-maximum suppression process and reduce the number of correctly identified positive detections.

As additional reference, HOG descriptor with sliding windows was also run with the results reported in the same set of graphs (see, Figure 5). We can notice that HoC outperforms HOG in horse detection but in cow detection HOG preformed better. Worse performance of LHOP method in cow detection can be attributed to missed detections where only 85% of samples were detected. This is confirmed by an experiment when HoC is used with sliding windows instead of LHOP detections. In this case the detection rate increases above 95%. Nevertheless, discriminative power of HoC is comparable to HOG's where sharp leveling of the curve (see, Figure 5d) can be observed in both descriptors.

MPEG7 shape dataset

Further evaluation of category misclassification was performed on *MPEG-7 Shapes* [27] dataset. This dataset was chosen for its large number of shape-specific classes with 70 different categories contained in the dataset as depicted in Figure 7. Many of those categories are fairly similar to each other, which makes them a good candidate for evaluating misclassifications between multiple categories. The dataset was pre-processed by rotating all objects in images to the position of the first sample in class to ensure correct orientation required for the LHOP method, which can only deal with around

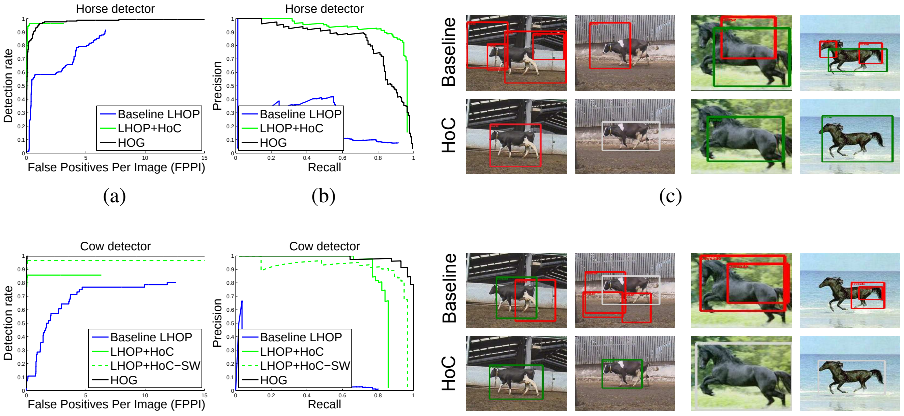
	LHOP	HoC	HOG		LHOP	HoC	HOG
device4	0.46	**0.63**	0.03	hcircle	0.01	0.12	**0.48**
device6	0.46	**0.56**	0.19	ray	0.27	**0.60**	0.02
device8	0.57	**0.77**	0.02	tree	0.01	0.21	**0.68**
frog	0.20	**0.57**	0.11	truck	0.02	**1.00**	0.79
glas	0.17	0.50	**0.89**	watch	0.14	**1.00**	0.18
hat	0.07	**0.14**	0.01	*Average*	*0.38*	***0.72***	*0.49*

Table 1: Results of the selected *MPEG-7 Shape*s [27] classes where high degree of misclassification can be observed. We report mean averaged precision (mAP) of each selected category and averaged mAP over all 70 categories in the last row.

10−15 degree of rotation. To accommodate for the detection of certain smaller-sized objects present in the *MPEG-7 Shapes* database, initial resize factor was increased to 2.2 and a maximum of 15 scales per image were allowed when running multiscale detection. Learning was performed on all categories at once in an incremental order for each category. Half of the images were used for training and another half for testing, therefore having 10 training images and 10 testing images per category. While learning was performed jointly for all categories, the evaluation was performed for each category separately and detections from one class did not influence other classes.

The confusion matrices are shown in Figure 6, where matrices were created from columns of misclassifications, with one column representing detections for a single category. We created the confusion matrix by taking only detections at the ideal point in the ROC curve [11] and performed this for each category individually. As ideal point we took a point on the curve closest to the best detection rate with the minimal number of false positives per image (FPPI), i.e., point closest to the [0,1]. This allowed to gauge the best discriminative performance of the score returned by the detector at the point where most of the positive regions were correctly classified.

Looking at both confusion matrices, we can see significant improvements of the misclassification rate across all categories. The baseline LHOP has many misclassifications among different categories named *device*. This can be expected due to a high degree of similarity between those categories (see, Figure 7), however, misclassifications are also present in other categories. For instance, categories of *frog*, *glass* and *guitar* often misidentify different classes of *devices*, and detections of a *truck* and a *tree* frequently misidentify *forks*, *pencils* and *watches* for *trucks* and *trees*. Poor misclassifications in baseline LHOP can also be seen in the overall detection accuracy with an averaged accuracy over all categories of only 0.49. After applying HoC classifier most of the misclassifications are eliminated, but some new misclassifications of categories *hcircle*, *hat* and a *ray* are now present on certain classes of *devices*. However, even in those cases, the mAP is considerably improved with the overall mAP averaged over all 70 categories increased from 0.38 to 0.72, as can be seen in Table 1. A significantly smaller amount of overall misclassifications also affected the overall detection accuracy where averaged accuracy over all 70 classes has increased by almost 40% from 0.49 to 0.69. Additionally, comparing HoC classifier to the HOG descriptor a significantly better performance of our method can also be observed, with 45% increase



(a) (b) (c)

(d) (e) (f)

Figure 8: Results of evaluating phantom detection rate with included category misclassification on *Weizmann horses* [4] and *Leeds cows [34]* datasets. The first row shows results for the horse detector with (a) detection-FPPI graph, (b) precision-recall graph and (c) examples of detections with and without HoC classifier. The second row shows similar types of graphs and samples for the cow detector (d-f). Images in (c) and (f) have correct detections with intersection-over-untion intersection over 0.5 shown in green, wrong detections shown in red and ground-truth bounding box when no detection is found shown in gray .

in the mAP.

6.3. Misclassifications on background objects

In this section the evaluation of misclassified detections on background objects was performed, following the same experimental setup as in both previous subsections. The LHOP vocabulary was trained with up to 7th layer, portion of realized parts was used as detection's probability score in the baseline LHOP, the SVM classifier score was used for the hypothesis verification with the HoC, and hard-negative detections were mined from the training images and from the additional 10% of the *Caltech-101* images when learning the HoC classifier.

Weizmann horses and Leeds cows dataset

The evaluation of background misclassifications on *Weizmann horses* and *Leeds cows* dataset was performed with the same experiment as in the previous section with the exception that this experiment considered detections on backgrounds as well.

The results are shown in Figures 8a and 8b for the horse detector and in Figures 8d and 8e for the cow detector, with all following the same trends as in our previously reported results. With included background detections the baseline LHOP achieved a mAP for a horse detection of around 0.20 while our improved method achieved mAP of 0.85 for the same category. The same trend appears in a cow detection, where mAP for baseline method is only 0.07, while the HoC classifier is able to increase mAP

	Horse (mAP)			Cow (mAP)			
	LHOP	LHOP/HoC	HOG	LHOP	LHOP/HoC	HoC-SW	HOG
Misclassifications only	0.61	**0.96**	0.93	0.15	0.84	0.96	**0.98**
Misclassifications and phantom detections	0.28	**0.91**	0.83	0.07	0.83	0.92	**0.98**

Table 2: Results on the combined dataset of *Weizmann horses* [4] and *Leeds cows [34]* with reported separate results for category misclassifications only in the top row and with included phantom detections in the bottom row.

to 0.85. Note that, the results contain not only improvements due to lower phantom detections but also improvements in category misclassification. However, we can also observe a difference compared to the results in previous experiment where phantom detections were excluded and we can now deduce the improvements in the background misclassification. The results of both experiments are also summarized in Table 2. In both, the baseline method and the HoC classifier, the performance of detectors with background objects is reduced, but with the HoC classifier the reduction is only from a mAP of 0.96 and 0.84 to a mAP of 0.91 and 0.83 for the horse and the cow detector, respectively. While for the baseline method the reduction is from mAP of 0.61 and 0.15 to only 0.28 and 0.07 for the horse and the cow detector, respectively. Results indicate that HoC classifier is able to maintain much lower misclassification rate compared to the baseline LHOP. The HoC classifier was able to maintain a higher degree of discrimination despite of the added background objects while in the baseline method the performance dropped by more than half, when background detections were added.

Comparing to the HOG descriptor, we also observe similar pattern as in previous experiment, where HoC performed better in horse detection, while HOG performed better in cow detection due to the lower number of overall detected objects by the LHOP method. As in previous experiment replacing the LHOP hypothesis generation with sliding windows improves detection of the missing samples.

ETHZ Shapes Classes dataset

A further evaluation of category classification was performed on the *ETHZ Shape Classes* dataset [14] following the standard protocol for this dataset: half of the images that contained the selected category where randomly chosen for the training, while the other half were used for the testing only. All other images that did not contain category objects where used only for testing. Selected category training images were used to train LHOP library $\mathcal{L}$ for up to 7 layers. The experiment was repeated 10-times and performed independently for each category.

Based on the results reported in Table 3, a considerable improvements across all five categories can be noticed. HoC classifier outperformed the baseline method for all categories on average by around 20 percentage points, producing detection rate of 89% versus 67% for LHOP at 0.4 false positives per image (FPPI). The least discriminative improvement was achieved for apple logo and bottle categories where baseline already achieved detection rate of 90.4% and 86.7%, respectively. High results for this two categories are achieved due to a relatively small geometrical variations which enable

	Apple logo	Bottle	Giraffe	Mug	Swan	*Average*
Baseline LHOP	90.4 (5.7)	86.7 (8.7)	44.8 (12.4)	66.2 (7.6)	45.1 (14.2)	*66.7*
LHOP + HoC verification (our)	91.4 (4.6)	**96.6 (2.9)**	**85.7 (5.8)**	**92.3 (4.4)**	**79.3 (9.0)**	***89.0***
HOG [7] with linear SVM	**92.7 (6.7)**	90.7 (7.5)	71.3 (6.0)	88.2 (6.4)	68.9 (10.7)	82.4

Table 3: Evaluation result on *ETHZ Shape Classes [14]* with reported detection-rate (%) at 0.4 FPPI averaged over ten iterations (standard deviation values are shown in parentheses). Note, in the reported results detections were classified as positive if intersection-over-union with the groundtruth was more than 0.5.

good discrimination between false positive detections. Nevertheless, improvement on this two categories are 1% and 10% points for apple logo and bottle, respectively. The most discriminative improvement is achieved for the categories of giraffe and swan, where the detection rate at 0.4 FPPI is almost doubled from 44.8% and 45.1% to 85.7% and 79.3% for giraffe and swan, respectively. Both of those categories have a high degree of inner-class geometrical variability that prevents good discrimination in the baseline method. The results of the HoC classifier indicate that with the inclusion of the lower-layer parts we are now able to more easily distinguish between false positives in such categories.

Additionally, comparing to the HOG descriptor, the HoC classifier performed better in all categories except in apple logo, where HOG outperformed HoC by 1 percentage point.

6.4. Comparison to deep approaches

In this section we compare the proposed pipeline of the LHOP detector with our HoC descriptor against the current state-of-the-art deep approaches. Specifically, we compare it against the variant of convolutional neural network (CNN), the R-CNN [21], which recently achieved state-of-the-art results on PASCAL VOC datasets [10]. This method implements full detection pipeline with both localization and classification. Localization is implemented with the region proposals generator using selective search [39] and classification is implemented with the state-of-the-art convolutional neural network, AlexNet [26]. In our experiments the convolutional network was pre-trained with the ILSVRC2012 [36] dataset using the network proposed by Krizhevsky et al. [26] and the final classification layer was trained with the *MPEG-7 Shapes* training set. The evaluation is performed on the *MPEG-7 Shapes* [27] dataset where shape is a prominent feature. This ensures a fair comparison with the LHOP model, which currently ignores any color or texture features. The evaluation is carried out with a two sets of experiments: (a) one with original images from *MPEG-7 Shapes* and (b) one with corrupted images to evaluate the robustness to occlusions. Both sets of experiments were conducted under the same conditions as the experiment on *MPEG-7 Shapes* from Section 6.2.

In the first experiment with the original images the R-CNN achieved mAP of 0.95 while the LHOP/HoC pipeline achieved mAP of 0.72. Higher score of the R-CNN can be contributed to its learning process which maximizes the discriminative objective thus achieving better discrimination.

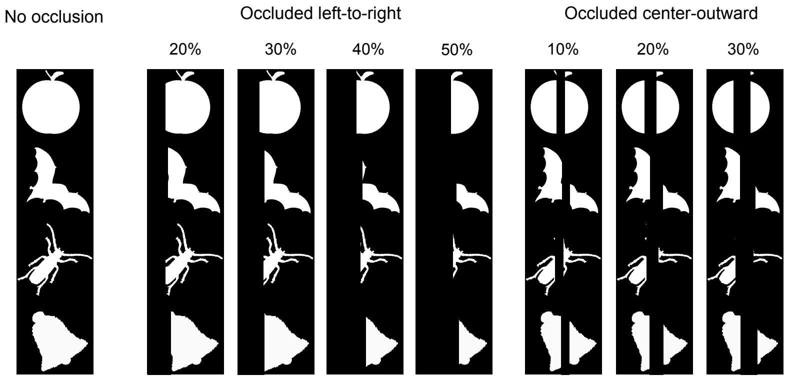


Figure 9: Examples of occlusions evaluated, with first column not being occluded, next four columns with left-to-right occlusions gradually increasing from 20% to 50%, and with the last three columns occlusions gradually increasing from center outwards from 10% to 30%.

mAP	Baseline/	Occluded left-to-right				Occluded center-outward		
	no occlusions	20%	30%	40%	50%	10%	20%	30%
R-CNN	**0.95**	**0.86**	**0.71**	**0.49**	0.003	0.09	0.04	0.02
LHOP + HoC verification (our)	0.72	0.62	0.47	0.31	**0.16**	**0.53**	**0.42**	**0.30**

Table 4: Results of evaluation on *MPEG-7 Shapes* [27] with additional occlusion either from left-to-right or from center-outwards. We report averaged mAP over all 70 classes for each experiment.

The second set of experiments evaluated robustness of the algorithm to various types and levels of occlusion. Two types of occlusions are evaluated: (i) one with object being occluded from left-to-right and (ii) one where objects are occluded from center outwards. Both types are varied with different levels of occlusion, ranging from 20% - 50% for left-to-right, and ranging from 10% - 30% for center-outwards occlusion (see, Figure 9). The occluded images are used only during testing while training was performed on original images. Results with both types and various levels of occlusion are shown in Table 4. The performance of both methods slowly decreases as the level of occlusion increases. The R-CNN becomes significantly prone to left-to-right occlusion when more than 50% of the object is not visible, achieving mAP of less than 0.01, while the LHOP/HoC pipeline was able to outperform it and achieve mAP of 0.15 at the same level of occlusion. The robustness of LHOP/HoC becomes even more apparent when object is occluded from center-outward, where LHOP/HoC consistently outperformed R-CNN at all levels of occlusion. The R-CNN was unable to achieve mAP higher than 0.10 even at the low level of occlusion, while LHOP/HoC achieved mAP of 0.53, 0.43 and 0.30 for 10%, 20% and 30% respectively for central-outwards occlusion.

The robustness of the LHOP/HoC to occlusion demonstrated with the second set of the experiments can be contributed to the generative learning of compositions used in the LHOP to clearly define the connections between a part, i.e. a feature, and its sub-compositions. This adds a straightforward way to perform a bottom-up as well as a top-down reasoning. While bottom-up reasoning is also used in CNN, the top-down is prohibited due to max-pooling operations. Some variants of neural networks, such as Convolutional Deep Belief Networks [30], have circumvented this by using probabilistic max-pooling, however, their top-down reasoning is not straightforward and they have to resort to Gibbs sampling to achieve it. On the other hand, the LHOP/HoC pipeline can take full advantage of the top-down reasoning to correctly predict the location of a partially occluded object by projecting the activated feature onto the first layer and correctly accounting for the missing features during location estimation. The HoC descriptor is then collected from the correct location which improves the detection performance of partially occluded objects. As opposed to LHOP/HoC pipeline, our results indicate, that R-CNN has difficulty at predicting the correct location of partially occluded object as it has to rely on a region proposals to generate the initial location and cannot implement the top-down reasoning in the neural network.

7. Discussion and conclusion

In this paper we have presented an improvement of discriminative power of hierarchical compositional model learnt-hierarchy-of-parts (LHOP) [18, 17]. We have provided an analysis behind the reasons for low discriminative power in LHOP model and have identified two principal types of problems: (a) similar category misclassifications and (b) phantom detections on background objects. To alleviate both of those problems we have proposed a novel descriptor called Histogram of Compositions (HoC), which is based on the same statistically relevant shapes as are used by the LHOP method. Lower-layer parts of the LHOP hierarchy were used for constructing HoC descriptor and an SVM classifier was further used to learn distinctive features important for discrimination. We have used HoC classifier to perform hypothesis verification of the LHOP detections and have avoided using computationally more intensive sliding windows.

An extensive evaluation was performed on five datasets where we have shown that proposed HoC descriptor performs significantly better than using similar descriptors such as HOG [7]. In the evaluation we have further provided evidence that applying HoC classifier to LHOP detections improves discriminative performance of LHOP method. We have achieved significant reduction of misclassifications between visually similar categories as well as a significant reduction of phantom detections around backgrounds.

Furthermore, we experimentally compared our approach to the state-of-the-art deep approach, the R-CNN [21]. While R-CNN performs better under clear conditions, we have shown that the LHOP/HoC pipeline is more robust to significant occlusions. We have identified top-down reasoning in LHOP steaming from its generative learning as an important factor in robustness to occlusion. Specifically, its ability to correctly predict location of an occluded object has proven important under such conditions. Another important advantage of LHOP compared to deep approaches is the explainability of the features. In the LHOP all features are modeled as sub-compositions with well understood local geometrical constraints. In convolutional neural networks features are just filters of previous layer features and their meaning is difficult to explain.

The clear meaning of features in the LHOP allows for easier integration with other modalities. For instance, 3D information around edges can be easily connected with any layer shape composition found at the same spatial position. At the same time one modality can add constraints to the learning process in other modalities and reduce the parameters search space, while in CNN modalities added as new first layer features significantly expend the search space. The clear meaning of features in LHOP also enables the learning process to be better adjusted for a given task. For instance, Aktas et al. [2] implemented the learning process with the graph based approach, while Ursic et al. [40] implemented learning of rotationally invariant features obtained from range scanner images of rooms. A clear separation and understanding of features allows to reason about individual modalities in separate channels, while at the same time enabling connections between them in the middle or higher layers.

Additionally, our proposed HoC descriptor is not restricted to the domain of shapes and can be applied to other modalities as well, such as 3D shape [24], motion, range images [40] or music [35]. HoC descriptor could also be applied to any other hierarchical approach that explicitly models compositions and, in particularly, any model that allows for top-down reasoning. This excludes pure discriminative deep networks such as CNN but it might be possible to apply it to generative Convolutional Deep Belief Networks [30], where certain degree of top-down reasoning is possible.

In our future work we would like to integrate the information provided by the HoC classifier into the LHOP method itself. Currently we rely on non-linear SVM to provide a good discrimination, but eliminating the need for the SVM would additionally speed-up the algorithm. We would also like to modify the response values to more accurately store a true discriminative information and use this information during the selection of compositions in the learning process to focus on compositions with a better discriminative properties. As the experiment with the R-CNN indicate that maximizing the discriminative objective during the learning can achieve excellent results on hierarchical approaches it would be beneficial to introduce similar discriminative learning of the LHOP vocabulary on top of the existing generative learning and combine the benefits of both generative and discriminative approaches.

References

[1] Agarwal, A., Triggs, B., 2006. Hyperfeatures - multilevel local coding for visual recognition. In: ECCV. Springer, pp. 30–43.

[2] Aktas, U., Ozay, M., Leonardis, A., Wyatt, J., 2014. A Graph Theoretic Approach for Object Shape Representation in Compositional Hierarchies Using a Hybrid Generative-Descriptive Model. European Conference on Computer Vision, 566–581.

[3] Amit, Y., Geman, D., 1999. A computational model for visual selection. Neural computation 11, 1691–1715.

[4] Borenstein, E., Ullman, S., 2002. Class-specific, top-down segmentation. European Conference on Computer Vision, 109–122.

[5] Bristow, H., Lucey, S., 2014. Why do linear SVMs trained on HOG features perform so well? arXiv:1406.2419.

[6] Chang, C. C., Lin, C. J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2 (3), 27:1—-27:27.

[7] Dalal, N., Triggs, B., 2005. Histograms of Oriented Gradients for Human Detection. In: CVPR. pp. 886–893.

[8] Dean, T., Ruzon, M., Segal, M., 2013. Fast, accurate detection of 100,000 object classes on a single machine. Computer Vision and Pattern Recognition.

[9] Enzweiler, M., Group, P. A., Sc, C., 2008. A Mixed Generative-Discriminative Framework for Pedestrian Classification, 1–8.

[10] Everingham, M., Van˜Gool, L., Williams, C. K. I., Winn, J., Zisserman, A., 2010. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision 88 (2), 303–338.

[11] Fawcett, T., Jun. 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861–874.

[12] Fei-Fei, L., Fergus, R., Perona, P., 2004. One-Shot learning of object categories. In: IEEE Transactions on Pattern Recognition and Machine Intelligence. IEEE Trans.

[13] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence 32, 1627–1645.

[14] Ferrari, V., Tuytelaars, T., Gool, L. V., 2006. Object Detection by Contour Segment Networks. In: Proceeding of the European Conference on Computer Vision. Vol. 3953 of LNCS. Elsevier, pp. 14–28.

[15] Fidler, S., Boben, M., Leonardis, A., 2009. Evaluating multi-class learning strategies in a generative hierarchical framework for object detection. In: Neural Information Processing Systems.

[16] Fidler, S., Boben, M., Leonardis, A., 2009. Learning Hierarchical Compositional Representations of Object Structure. Computer and Human Vision Perspectives, 1–18.

[17] Fidler, S., Boben, M., Leonardis, A., Aug. 2014. Learning a Hierarchical Compositional Shape Vocabulary for Multi-class Object Representation, 1–17.

[18] Fidler, S., Leonardis, A., 2007. Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts. In: CVPR. IEEE Computer Society.

[19] Fritz, M., Leibe, B., Caputo, B., Schiele, B., 2005. Integrating representative and discriminant models for object category detection. Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 1363–1370 Vol. 2.

[20] Fukushima, K., Miyake, S., Ito, T., 1983. Neocognitron: A neural network model for a mechanism of visual pattern recognition. Ieee Transactions On Systems Man And Cybernetics SMC-13, 826–834.

[21] Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Computer Vision and Pattern Recognition.

[22] Hinton, G., 2006. To Recognize Shapes , First Learn to Generate Images To Recognize Shapes , First Learn to Generate Images.

[23] Huang, F.-J., LeCun, Y., 2006. Large-Scale Learning with SVM and Convolutional Nets for Generic Object Categorization. In: CVPR. IEEE Press.

[24] Kramarev, V., Zurek, S., Wyatt, J. L., Leonardis, A., Aug. 2014. Object Categorization from Range Images Using a Hierarchical Compositional Representation. 22nd International Conference on Pattern Recognition, 586–591.

[25] Kristan, M., Boben, M., Tabernik, D., Leonardis, A., 2013. Adding discriminative power to hierarchical compositional models for object class detection. In: 18th Scandinavian Conference on Image Analysis. pp. 1–12.

[26] Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. Advances In Neural Information Processing Systems, 1–9.

[27] Latecki, L. J., Lak, R., Eckhardt, U., 2000. Shape Descriptors for Non-rigid Shapes with a Single Closed Contour. CVPR, 424–429.

[28] Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. Computer Vision and Pattern Recognition 2, 2169–2178.

[29] Lee, H., Grosse, R., Ranganath, R., Ng, A. Y., 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. Proceedings of the 26th Annual International Conference on Machine Learning ICML 09 2008, 1–8.

[30] Lee, H., Ng, A. Y., 2009. Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations.

[31] Leo, L., Yuanhao, Z., Alan, C., William, Y., 2010. Latent hierarchical structural learning for object detection. Computer Vision and Pattern Recognition.

[32] Leonardis, A., Fidler, S., 2011. Learning hierarchical representations of object categories for robot vision. Robotics Research.

[33] Lowe, D. G., 1999. Object Recognition from Local Scale-Invariant Features. In: Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2. ICCV '99. IEEE Computer Society, Washington, DC, USA, pp. 1150—-.

[34] Magee, D. R., Boyle, R. D., Jun. 2002. Detecting lameness using "Re-sampling Condensation" and "multi-stream cyclic hidden Markov models". Image and Vision Computing 20 (8), 581–594.

[35] Pesek, M., Marolt, M., 2013. Chord estimation using compositional hierarchical model. European Conference on Machine Learning: Workshop on Music abd Machine Learning.

[36] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei, L., Sep. 2014. ImageNet Large Scale Visual Recognition Challenge, 43.

[37] Si, Z., Zhu, S.-c., 2013. Learning AND-OR Templates for Object Recognition and Detection. IEEE Transactions on Pattern ananlysis and machine intelligence 35 (9), 2189–2205.

[38] Tabernik, D., Kristan, M., Boben, M., Leonardis, A., 2012. Learning statistically relevant edge structure improves low-level visual descriptors. In: International Conference on Pattern Recognition. pp. 1471 – 1474.

[39] Uijlings, J., van de Sande, K., 2013. Selective search for object recognition. International Journal of Computer Vision.

[40] Ursic, P., Kristan, M., Skočaj, M., Leonardis, A., 2012. Room Classification using a Hierarchical Representation of Space. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, October 7-12, 2012, Vilamoura, Algarve, Portugal.

[41] Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A., 2009. Multiple Kernels for Object Detection. In: Proceedings of the International Conference on Computer Vision.