Learning statistically relevant edge structure improves low-level visual descriptors

Domen Tabernik[1], Matej Kristan[1], Marko Boben[1], Aleš Leonardis[1,2]
[1]*Faculty of Computer and Information Science, University of Ljubljana*
[2]*CN-CR Centre, School of Computer Science, University of Birmingham*
{domen.tabernik},{matej.kristan},{marko.boben},{ales.leonardis}@fri.uni-lj.si

Abstract

Over the recent years, low-level visual descriptors, among which the most popular is the histogram of oriented gradients (HOG), have shown excellent performance in object detection and categorization. We form a hypothesis that the low-level image descriptors can be improved by learning the statistically relevant edge structures from natural images. We validate this hypothesis by introducing a new descriptor called the histogram of compositions (HoC). HoC exploits a learnt vocabulary of parts from a state-of-the-art hierarchical compositional model. Furthermore, we show that HoC is a complementary HoC descriptor to HOG. We experimentally compare our descriptor to the popular HOG descriptor on the task of object categorization. We have observed approximately 4% *improved categorization performance of HoC over HOG at lower dimensionality of the descriptor. Furthermore, in comparison to HOG, we show a categorization improvement of approximately* 10% *when combining HOG with the proposed HoC.*

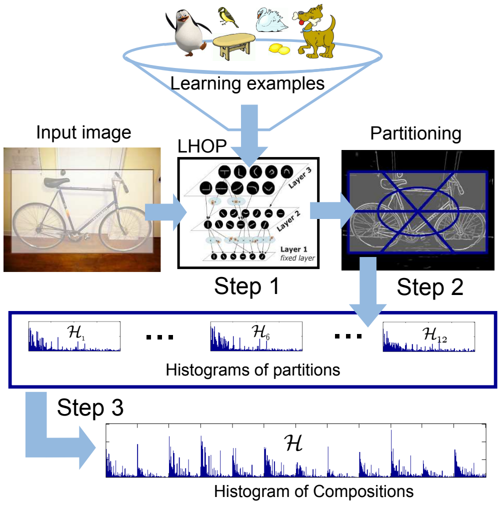

Figure 1. Extraction of HoC. Significant edges are first detected as compositions from learnt-hierarchy-of-parts, i.e. LHOP (step 1). The object region is then divided into several partitions, and next, histogram over compositions is extracted for each partition (step 2). Our descriptor $\mathcal{H}$ is formed by concatenating all histograms into a single final histogram (step 3).

1 Introduction

Low-level features have lately been successfully used for object detection and categorization. Features such as [10, 3] have been used to produce state-of-the-art results on many different problem domains. Particularly, Histogram of Oriented Gradients [3] has been able to achieve excellent results for detection and categorization of different objects categories. As originally proposed by its authors, it has first been used for pedestrian detection but has then been quickly applied to detection of many different object categories. Its relatively simple design and good performance have contributed to high popularity of this method and many researchers incorporated it into their own designs.

In [12], the HOG was used for leaf detection, while in [1], the HOG was applied to a well-known Elastic Bunch Graph Matching (EBGM). Both produced better results than other approaches. Lately many participants of the PASCAL challenge [4] have demonstrated state-of-the-art results using HOG descriptor. In [13], a boosted HOG-LBP was used with additional multi-context approach and their method ranked first in 6 classes and second in 5 classes. The deformable parts

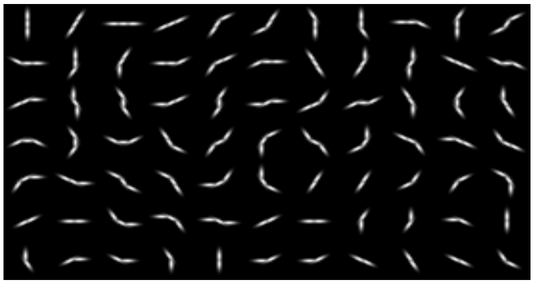

Figure 2. The library of parts from layer 2.

models were introduced in [7], which use the HOG as a low-level feature in conjunction with a weak geometrical model. Together with a latent SVM for learning deformation models this method, along with its recent derivations, also shows the best results in detection of many object classes.

Despite its high popularity the HOG descriptor still suffers from several disadvantages. Firstly, dimensionality of HOG descriptor depends on the image size due to the fixed-size blocks. Another disadvantage of HOG are orientations that are densely sampled and are therefore likely to capture additional structures, which are irrelevant for object representation, thus leading to a poorer object detection. Additionally, to improve performance various researchers have been using HOG in combination with other features that specifically encode texture or color [14, 5, 13].

We propose a new descriptor that is based on learning of local structures which are statistically relevant for object description. For learning of these structures we utilize a state-of-the-art learnt-hierarchy-of-parts (LHOP) [9] algorithm, that uses a layered hierarchical representations of edge structures. Our proposed descriptor is related to [8], but rather than having the algorithm greedily search for the terminal nodes from the entire hierarchy (shapinals), we use only a single layer (second) in the hierarchy as a learned shape structure vocabulary. In addition, we learn shapes on independent set of general images, making the descriptor category and database independent and apply weighting functions on part locations. The resulting descriptor, which is our first contribution, is called *histogram of compositions* (HoC). A pipeline for extraction of HoC is shown in Figure 1. As our descriptor uses only relevant shapes and structures we achieve lower dimensionality and eliminate the problem of irrelevant shapes found in HOG.

As our second contribution we show that HoC descriptor is complementary to HOG, and a combination of both descriptors considerably boosts detection performance. Their combination captures more shape specific structures represented by the LHOP compositions in HoC descriptor on one hand, and on the other hand we capture more texture-like structures found by densely sampled orientations in HOG descriptor that HoC does not capture.

The remainder of the paper is structured as follows. In Section 2, we briefly describe the LHOP model and introduce the histogram of compositions in Section 3. In Section 4, we experimentally compare the HoC to HOG, as well as their combination, and discuss the results. We draw conclusions in Section 5 and provide several venues for further research.

2 Learned Hierarchy of Parts

In this section we briefly describe the learnt-hierarchy-of-parts (LHOP) [9] that forms a basis for our descriptor and refer the reader to [9] for more details.

The LHOP is a recursive hierarchical compositional vocabulary of shape parts. Each layer in the hierarchy contains a library of parts and each part is a composition of parts from a lower layer. The lowest layer contains Gabor filters oriented in six directions, and by virtue of composition, the complexity of the parts in each higher layer gradually increases, ending at the top layer with a library of categorical shapes. During detection, only the parts of the lowest layer are compared to the image intensity, and the higher-layer parts are detected layer-by-layer through efficient indexing scheme [9]. The library in the first three layers is learnt jointly for all categories and the remaining layers are learnt sequentially one category at a time. This results in a significant sharing of parts across various categories. Since the parts at the first layer are fixed, the parts at the second layer are learnt by recording frequent activations (compositions) of layer-one parts within a receptive field from a large number of natural images. Once the parts at the second layer are learnt, this procedure is repeated layer by layer until all libraries of all layers are learnt. Since the LHOP constructs the parts by focusing on frequent edge structures in natural images, we can view it as a machine learning algorithm for learning edge structures of increasing complexity, tuned to natural images.

We define the trained LHOP library as a set of L compositions $\mathcal{L} = \{P_l\}_{l=1:L}$, where P_l is an identifier of l-th composition in the library. Applying the library $\mathcal{L}$ on an image I, the LHOP algorithm infers a set of K detected parts, $\mathcal{C}(I, \mathcal{L})$,

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

where the k-th detected part $\pi_k = [P_k, \mathbf{c}_{\pi_\mathrm{k}}, \lambda_k]$ is defined by its library identifier P_k, its location $\mathbf{c}_{\pi_k}$ in the image and its detection score λ_k. In general, the library $\mathcal{L}$ refers to the parts from all layers, but we consider only the learnt parts from the second layer and in the remainder of the paper, $\mathcal{L}$ will refer only to the library from the second layer. For better intuition, we show the learnt second-layer library in Figure 2.

3 Histogram of Compositions

We define the HoC descriptor $\mathcal{H}$ within image region Ω by calculating a histogram of detected compositions over the library of compositions $\mathcal{L}$. For better encoding of spatial layouts we split region into M-partitions as shown in Figure 1. From each partition m, a histogram $\mathcal{H}_m$ over the entire library of compositions is extracted. The value of the histogram bin corresponding to part identifier P_l is defined as

$$\mathcal{H}_m(P_l) = \sum_{\pi_k} \lambda_k \frac{\Phi(w^{(m)}_{\pi_k};\varphi)}{\sum\limits_n \Phi(w^{(n)}_{\pi_k};\varphi)} \delta_{P_l}(P_k), \tag{2}$$

where λ_k is a composition's response, $\delta_{P_l}(P_k)$ is the Kronecker delta centered at P_l, $\Phi(x;\varphi)$ is a gate function that returns zero when $x < \varphi$, and $w^{(m)}_{\pi_k}$ is a weighting function that assigns a weight by which a detected composition contributes to the histogram bin P_l. The gate function is required for a stricter localization of a part's response. In our experiments we have used $\varphi = 0.1$. The weighting function is defined as

$$w^{(m)}_{\pi_k} = \frac{\mathcal{N}(\mathbf{c}_{\pi_k};\mathbf{c}_m,\Sigma_m)}{\sum\limits_n \mathcal{N}(\mathbf{c}_{\pi_k};\mathbf{c}_n,\Sigma_n)}, \tag{3}$$

where $\mathcal{N}(\cdot;\mathbf{c},\Sigma)$ is a Gaussian with mean $\mathbf{c}$ and covariance Σ. Σ_m is covariance of m-th partition in the partitioning. We have found that robustness is increased if the covariances are slightly reduced. In our experiments we reduced them by a factor of 0.25. The final descriptor of a region Ω is then defined as a concatenation of all histograms from the partitions, i.e., $\mathcal{H} = \alpha[\mathcal{H}_1,\ldots,\mathcal{H}_M]$, thus forming a $M \cdot L$-dimensional feature vector, where α is a normalization factor such that the histogram cells sum to one.

4 Experiments and results

We have compared HoC to the closely-related HOG descriptor on its own and as a complementary descriptor. For implementation of HoC descriptor we obtained a reference implementation of LHOP from the authors of [9] and implemented the HoC descriptor in C++[1]. To learn libraries of compositions we trained LHOP method on approximately 250 *general images* thus producing library with 77 compositions on the second layer which gives us a 924-dimensional HoC descriptor.

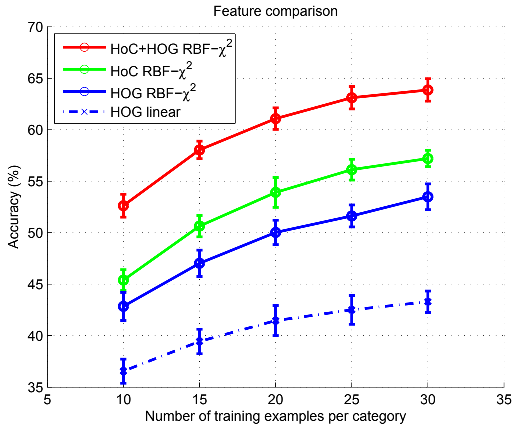

Figure 3. Results on Caltech-101 database

For HOG descriptor we used the binary from [3] with 8×8 pixels wide cells and 16×16 pixels wide blocks. For HOG, all images were resized to 64×64 pixels as it proved to perform best in our experiments. This produces a 1764-dimensional HOG descriptor.

For classification we used the one-versus-rests LIBSVM [2] with two different kernels: (i) a linear kernel and (ii) an RBF kernel with chi-squared distance function (RBF-$\mathcal{X}^2$).

All experiments were conducted on the Caltech-101 [6], following the methodology from [11]: we randomly selected specific number of examples per category and tested on randomly selected 15 examples from the rest of the set. The experiments were repeated 10-times and we reported the mean averaged classification rate over all 102 object categories.

4.1 Results

Looking at the results on Caltech-101 in Figure 3 we notice that out of linear and non-linear RBF-$\mathcal{X}^2$ kernels, RBF-$\mathcal{X}^2$ produces classification accuracy which is around 10 percent better in almost all cases. Focusing on RBF-$\mathcal{X}^2$ kernel, we see that HoC descriptor produces results which are consistently better then results of HOG descriptors. At 30 training examples per category HOG achieves average classification accuracy of

[1] We plan to make the binaries for construction of the HoC descriptor available to facilitate other researchers in applying our descriptor to their own research.

53.4 percent while HoC produces by a 4 percent better result at accuracy of 57.2. Note that all improvements are statistically significant at 0.05 level. Higher performance of HoC can be attributed to LHOP compositions that are learnt from general images and therefore allow HoC to encode only shapes and structures relevant for object description. Additionally, the HoC descriptor achieved this better performance at lower dimensional descriptor than HOG.

From the results we can also see that a combination of both descriptors increases performance by 6 to 8 percent compared to HoC and by almost 10 to 12 percent compared to HOG. At 30 training examples HoC+HOG descriptor with RBF-$\mathcal{X}^2$ kernel produced classification accuracy of 63.9 percent which is by 7 percent better then using only HoC and by 10 percent better then only HOG descriptor. The increase of performance in combined features indicates that both descriptors capture a complementary shape and texture information. While HoC is sparser descriptor that can better represent more complex shapes, HOG on the other hand is a dense descriptor which captures many different local structure including more texture-like structures that are missing in HoC. By combining both descriptors, we can therefore better capture complex shapes and texture-like structures jointly and thus achieve performance that is considerably better from the individual descriptors.

Note that we achieved significant improvement with the library that was learnt from a set of natural images that were not a part of Caltech-101 dataset, which indicates the generality of our descriptor.

5 Conclusion

In this paper we presented a new edge based descriptor termed *histogram of compositions* (HoC) which, in contrast to histogram of oriented gradients (HOG), uses only shapes and local structures relevant for object representation. By using the learnt-hierarchy-of-parts (LHOP) [9] as the basis for learning those relevant shapes, our method produces a descriptor of lower dimensionality and at the same time delivers improved performance over HOG descriptor. More importantly, we have also evaluated combined HoC and HOG descriptors, which we proved can achieve even grater performance then each descriptor achieves individually. This indicates that descriptors encode complementary information. In our future work we will explore combining HoC with other low-level descriptors, that encode complementary features, in particular, texture-based descriptors such as local-binary-pattern (LBP). Additionally we would also like to apply the descriptor within more advanced object detection schemes that currently utilize the HOG descriptor (e.g., [7] and its derivations).

Acknowledgments. This work was supported in part by ARRS research program P2-0214 and ARRS research projects J2-4284, J2-3607 and J2-2221.

References

[1] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol. Face recognition using hog-ebgm. *Pattern Recogn. Lett.*, 29:1537–1543, July 2008.

[2] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[5] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR*, 2009.

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

[7] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4. http://www.cs.brown.edu/ pff/latent-release4/.

[8] S. Fidler, M. Boben, and A. Leonardis. Similarity-based cross-layered hierarchical representation for object categorization. In *CVPR*. IEEE Computer Society, 2008.

[9] S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR*. IEEE Computer Society, 2007.

[10] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[11] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.

[12] X.-Y. Xiao, R. Hu, S.-W. Zhang, and X.-F. Wang. Hog-based approach for leaf classification. In *ICIC*, pages 149–155, Berlin, Heidelberg, 2010. Springer-Verlag.

[13] Y. Yu, J. Zhang, Y. Huang, S. Zheng, W. Ren, and C. Wang. Object detection by context and boosted hog-lbp. In *Visual Recognition Challange workshop, European conf. Comp. Vision*, 2010.

[14] C. Zeng and H. Ma. Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. In *ICPR*, pages 2069–2072. IEEE, 2010.