Increased complexity of low-level structures improves histograms of compositions

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

While low-level visual features, such as histogram of oriented gradients (HOG), have been successfully used for object detection and categorization, we have been able to improve upon their performance by introducing histogram of compositions (HoC) in our previous work. In this paper we propose an extended version of HoC descriptor that uses additional layers from hierarchical model. We experimentally show that extended HoC surpasses the performance of the original descriptor by approximately 5% *as additional layer provides higher complexity of compositions. Furthermore, with additional layer we show to produce competitive results to original HoC descriptor combined with HOG and can even further increase performance by adding HOG on top of HoC with additional layer.*

1 Introduction

Common to many of the approaches for solving the problem of object detection and categorization is the use of low-level features such as [10, 3]. They enable to effectively describe object's visual properties using simple features and produce state-of-the-art results. Particularly, Histogram of Oriented Gradients [3] has been extensively used by many different researchers to produce excellent results. Its popularity can be attributed to simple design where objects are represented by small local gradient structures with histogram of orientations, normalized to their local surroundings. Small histograms are able to represent local shapes through local gradient distributions that are invariant to small local geometric transformations. As [3] have shown, these sets of features have proven to be highly effective for representation of people in upright position.

While HOG was originally proposed as descriptor for people detection, it was quickly applied to other object categories as well. It was used in a well-known Elastic Bunch Graph Matching (EBGM) [1] to produce better results then other approaches, while in [13], it was successfully applied to leaf detection. In [9], authors used it in combination with a bag-of-features, such as SIFT, and additional context information and achieved state-of-the-art results on more then half of categories in PASCAL VOC 2007 and 2008 dataset [4]. Additionally, many participants of the PASCAL challenge [5] incorporated HOG descriptor and produced top results. In [14], a boosted HOG-LBP was used with additional multi-context approach and their method ranked first and second in many classes. [7] incorporated HOG as a low-level feature with a weak geometrical model in a deformable parts model to produce state-of-the-art results.

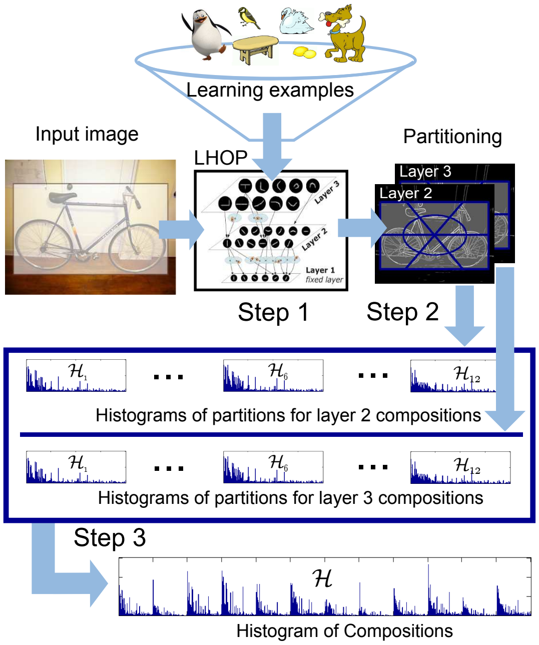


Figure 1: Extraction of multiple layer HoC. Significant edges are first detected as compositions for each layer from learnt-hierarchy-of-parts, i.e. LHOP (step 1). The object region is then divided into several partitions, and next, histogram over compositions is extracted for each partition and for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

Even though HOG descriptor is capable of achieving state-of-the-art results we have shown in [11] that there are still ways of improving it. Improvements can be made due

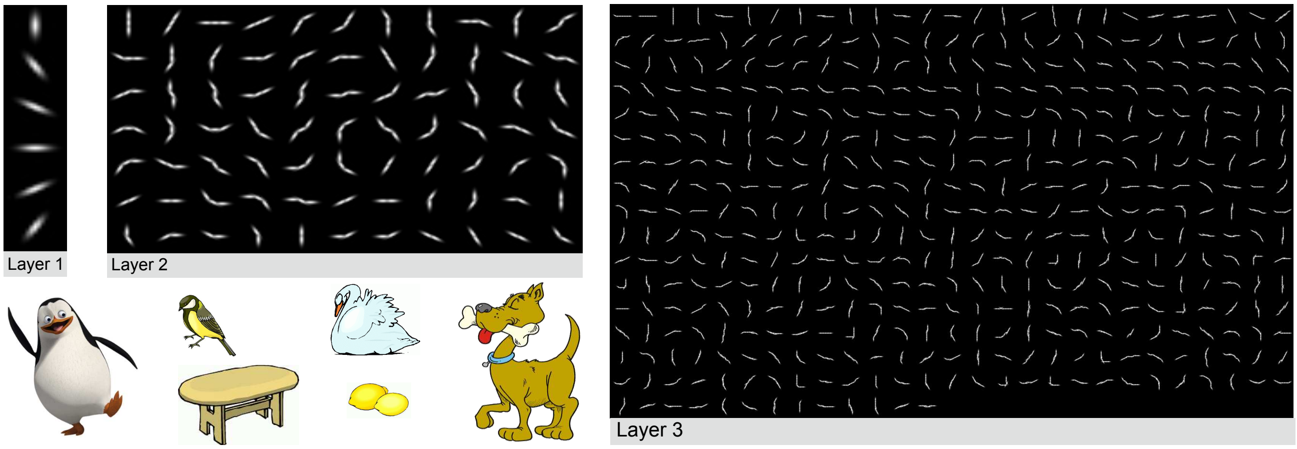


Figure 2: The library of parts per each layer with some samples of training examples

to certain shortcoming of HOG descriptor such as the use of a densely sampled local structures which can capture unnecessary shapes and thus lead to noisier representation of object's appearance. In [11], we have introduced a HOG-like edge based descriptor called histogram of compositions (HoC). Instead of densely sampling all local shapes and structures, as in the HOG, this descriptor relies upon sparser set of hierarchical compositions produced by Learned Hierarchy of Parts (LHOP) [8]. The LHOP method enables learning of small local shapes through statistics of general set of images and thus producing a vocabulary of compositions that are statistically relevant for description of various object categories. By removing statistically irrelevant shapes and structures we have achieved better performance on Caltech-101 dataset [6] compared to HOG descriptor while at the same time reducing dimensionality of the descriptor.

We have shown in [11] that both descriptors can be complementary to each other as performance increased considerably when descriptors were combined together. This indicates that both descriptors, to certain degree, encode different set of information. HoC descriptor used in [11] relies only on second layer of hierarchical model, which allows the use of compact set of compositions with smaller vocabulary size thus producing descriptor with lower dimensionality but still superior performance. But using only one hierarchical layer allows for only limited complexity of compositions. In this paper we explore the effect of using additional higher layer compositions in HoC descriptor. With higher layer compositions we can model parts with higher complexity and can therefore capture more complex shapes and structures. We show that by using additional third-layer compositions we can achieve better performance than we would with only second-layer compositions. We also show that adding third-layer compositions to only HoC descriptor can achieve similar performance as adding only HOG to original HoC descriptor. This allows us to use a single framework (HoC) and can reduce certain disadvantages brought by HOG. For instance, such as the problem of dimensionality, which in HOG descriptor is determined by preselected window size. Additionally, if we are willing to tolerate problems of HOG, we show that adding HOG descriptor produces even higher performance as information contributed by third-layer compositions is still more shape-based while HOG captures more texture-like features.

The remainder of the paper is structured as follows. In Section 2, we briefly describe the application of the histogram of compositions to multiple layers. In Section 3, we conduct the experiments and discuss the results of comparing the HoC computed from different set of layers to HOG and their combinations. We draw conclusions in Section 4 and provide several venues for further research.

2 Multiple layer Histogram of Compositions

We use existing HoC descriptor introduced in [11] and extend it to additional layers with ability to use arbitrary combination of layers (see, Figure 1). We form multiple layer HoC descriptor $\mathcal{H}$ from specific image region by starting with a simple descriptor $\mathcal{H}^{(\mathcal{L})}$ for each layer $\mathcal{L}$ obtained through processing of image with an LHOP method and extracting HoC descriptor according to procedure described in [11]. We use the same weighting function and same region split so we can define $\mathcal{H}^{(\mathcal{L})}$ as

$$\mathcal{H}^{(\mathcal{L})} = \alpha^{(\mathcal{L})}[\mathcal{H}_1^{(\mathcal{L})}, \dots, \mathcal{H}_M^{(\mathcal{L})}], \qquad (1)$$

where $\mathcal{H}_m^{(\mathcal{L})}$ is histogram of compositions within m-th region on $\mathcal{L}$-th layer and $\alpha^{(\mathcal{L})}$ is a normalization factor such that the histogram of cells at $\mathcal{L}$-th layer sum to one. We form final descriptor $\mathcal{H}$ by simple concatenation of histograms from desired layers :

$$\mathcal{H} = [\mathcal{H}^{(\mathcal{L}_1)}, \dots, \mathcal{H}^{(\mathcal{L}_N)}], \qquad (2)$$

where $[\mathcal{L}_1, \mathcal{L}_2, ..., \mathcal{L}_N]$ is a set of layer numbers used for final descriptor. In general we can combine arbitrary set of layers but we focused only on combining second and third layer parts as compositions in higher layers represent parts tuned to more category specific shapes. Such specific parts are not desired for general based vocabulary.

3 Experiments and results

We evaluated the use of higher layer compositions by conducting three sets of experiments. In first set we used original HoC with only second-layer compositions. This experiments formed a baseline comparison for our proposed extended descriptor. In second set of experiments we evaluated HoC with only third-layer compositions, while in third set we evaluated the use of HoC descriptor with combined second-layer and third-layer compositions. All experiments were conducted on the Caltech-101 [6] dataset by extracting descriptor on whole image and following the methodology of [12]: we trained using different number of examples, tested on randomly selected 15 examples from the rest of the set and finally repeated this process 5-times.

To learn libraries of compositions with second and third layer, we trained LHOP method using reference implementation from [8] on approximately 250 *general images* thus producing library with 77 compositions on the second layer and 445 compositions on the third layer. All learnt compositions per layer are shown in Figure 2 together with some samples of training examples.

We used the binary from [3] for HOG descriptor with the same parameters as in [11], i.e. 8×8 pixels wide cells and 16×16 pixels wide blocks, where all images were resized to 64×64 pixels. For classification we used one-versus-rests LIBSVM [2] with an RBF kernel using chi-squared distance function (RBF-$\mathcal{X}^2$).

3.1 Results

Looking at the results of descriptor with second-layer parts, shown in Figure 3, we can notice performance that is consistent with the results in [11]. HoC descriptor outperforms HOG by approximately 4% in all cases, while combination of HoC with HOG boosts performance by additional 7%.

Focusing next on HoC descriptor with third-layer parts (see, Figure 4) we notice improvement of approximately 2% across different number of training examples when compared with second-layer HoC, while concatenation of second and third-layer adds additional 3% increase of performance, therefore achieving overall 5% increase compared to second-layer compositions. Improved performance can be attributed to more richer vocabulary at third layer where there are many more parts with more complex compositions. But with increased complexity we cannot capture some simpler and smaller shapes therefore by adding second-layer parts we achieve even greater performance.

Note that HoC with the second and third-layer compositions achieves nearly equivalent performance compared to combined second-layer HoC and HOG. Although latter has by around 1% to 2% better performance, we are still able to achieve comparable performance without relaying on different type of descriptor.

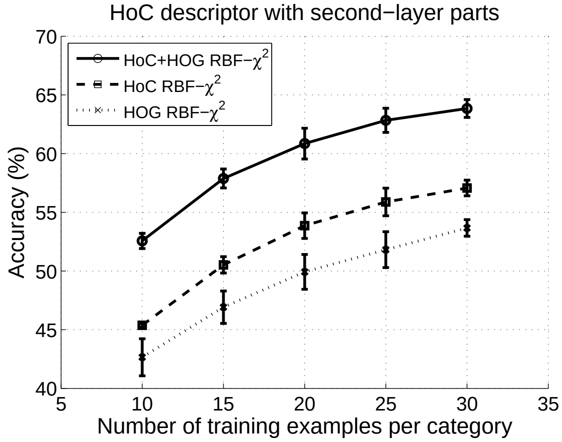
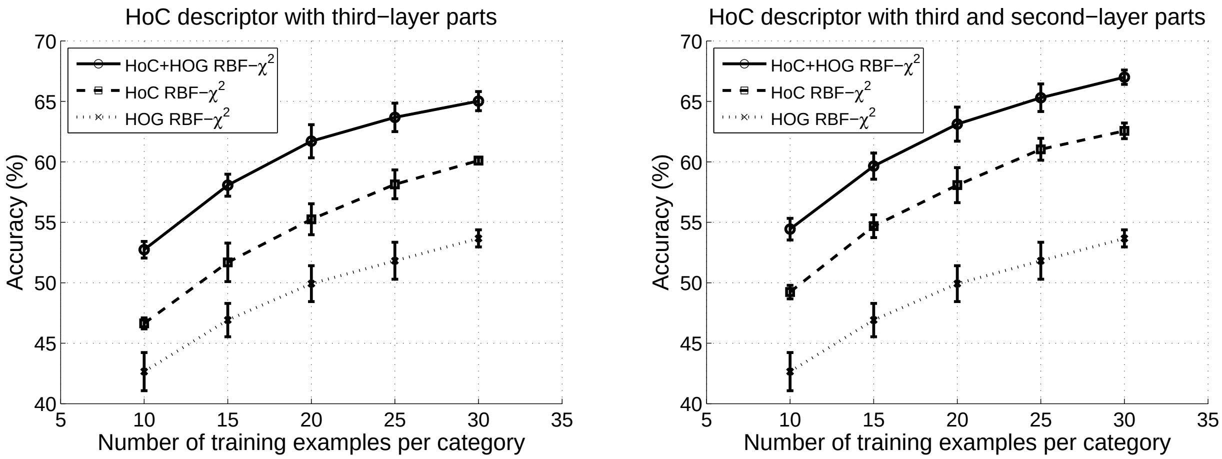


Figure 3: Repeated baseline results on Caltech-101 database using only second-layer HoC descriptor

By adding HOG descriptor to HoC with compositions from different layers we see consistently higher performance as additional layers are used. With HoC constructed from second and third-layer parts and additionally combined with HOG producing best results that have accuracy at around 67% at 30 training examples per category. While achieving best performance using this combination of descriptors, we notice from Figure 4b, that this value is only by 5% better relative to HoC without HOG, but when using only second-layer parts (see, Figure 3) we notice that increase of performance by adding HOG is around 7%. This difference in increased performance is a indication that by adding third-layer composition we were now able to capture certain local structures that were originally captured by HOG.

4 Conclusion

In this paper we presented an extended version of histogram of compositions, originally introduced by [11]. Our extended version includes additional layers and have shown to produce better performance by adding third-layer compositions. Additional compositions provide a richer vocabulary and are capable of capturing more complex shapes and structures. We have evaluated two types of HoC descriptors on Caltech-101 dataset; one with only third-layer parts and one with second-layer and third-layer parts combined. Both of them have exceeded the performance of original descriptor. Additionally, we have shown that HoC descriptor with second and third-layer compositions can produce similar results to combined second-layer HoC descriptor with HOG, thus reducing dependency on using different framework. But, by adding HOG descriptor to HoC with second and third layer we've shown to achieve even higher results. In our future work we will investigate the impact of combining more texture-based features with HoC descriptor and evaluate the use of HoC for detection and localization of



(a)

(b)

Figure 4: Results on Caltech-101 database with (a) third-layer and (b) combination of third and second-layer

objects using a sliding-window technique.

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] Alberto Albiol, David Monzo, Antoine Martin, Jorge Sastre, and Antonio Albiol. Face recognition using hog-ebgm. *Pattern Recogn. Lett.*, 29:1537–1543, July 2008.

[2] Chih-Chung Chang and Chih-Jen 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] Navneet Dalal and Bill 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 Challenge Results. http://www.pascal-network.org/challenges/VOC.

[5] 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.

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

[9] Hedi Harzallah, Frédéric Jurie, and Cordelia Schmid. Combining efficient object localization and image classification. In *International Conference on Computer Vision*, sep 2009.

[10] David 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] D. Tabernik, M. Kristan, M. Boben, and A. Leonardis. Learning statistically relevant edge structure improves low-level visual descriptors. In *International Conference on Pattern Recognition*, 2012.

[12] 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.

[13] Xue-Yang Xiao, Rongxiang Hu, Shan-Wen Zhang, and Xiao-Feng Wang. Hog-based approach for leaf classification. In *ICIC*, pages 149–155, Berlin, Heidelberg, 2010. Springer-Verlag.

[14] Yinan Yu, Junge Zhang, Yongzhen Huang, Shuai Zheng, Weiqiang Ren, and Chong Wang. Object detection by context and boosted hog-lbp. In *Visual Recognition Challange workshop, European conf. Comp. Vision*, 2010.