On a Quest for Image Descriptors Based on Unsupervised Segmentation Maps

Piotr Koniusz and Krystian Mikolajczyk
University of Surrey, Guildford, UK
p.koniusz, k.mikolajczyk @ surrey.ac.uk

Abstract

This paper investigates segmentation-based image descriptors for object category recognition. In contrast to commonly used interest points the proposed descriptors are extracted from pairs of adjacent regions given by a segmentation method. In this way we exploit semi-local structural information from the image. We propose to use the segments as spatial bins for descriptors of various image statistics based on gradient, colour and region shape. Proposed descriptors are validated on standard recognition benchmarks. Results show they outperform state-of-the-art reference descriptors with 5.6x less data and achieve comparable results to them with 8.6x less data. The proposed descriptors are complementary to SIFT and achieve state-of-the-art results when combined together within a kernel based classifier.

1. Introduction

Adequate image representations have been shown as crucial for the performance of image retrieval and recognition systems. State-of-the-art systems rely on interest point detectors such as MSER, Hessian or Harris [7] typically combined with descriptors derived from SIFT [5, **?**]. For category recognition, dense sampling has been advocated over key-point extraction [2]. Recent research [4] shows that unsupervised segmentation maps constitute a good alternative to both standard key-point detectors and dense sampling strategies. With less interest points derived from maps, they outperform the dense sampling approach which typically scores top in challenging classification problems [2]. This is due to the saliency of detected extremal curvature points along segment boundaries and full coverage of images with segments (in contrast to sparsely distributed interest points). This paper investigates direct application of segmentation maps in devising an image representation that covers all regions of processed images and makes use of semi-local structures formed by segments. In order to capture foreground/background boundaries of objects as well as the gradient within their areas, adjacent pairs of segments are processed. We argue they form good hypotheses for capturing an essential gradient-based object appearance. Further, multiple segmentation maps extracted with different parameters enrich such hypothesis space. To our best knowledge, there has been no attempt yet to use segmentations as spatial hypotheses for shape of the multiple descriptor cells.

Related work. Segmentation maps have been used widely as an auxiliary grouping cue in place of common bounding boxes [6]. It was also shown in [8] that enhancing foreground/background hypotheses improves classification results. Further, extremal curvatures of segments [4] were found to serve well as salient points outperforming dense sampling strategies. Optimal spatial arrangement of descriptor bins has also been explored recently in DAISY [10] which is designed deliberately for dense matching. It comprises several circular regions which are arranged in a polar manner resembling petals and a flower. Learning local image descriptors [13] can be structured into blocks concerning choice of gathered histogram evidence and spatial shape of bins. Blob representation proposed in [1] is somewhat similar in spirit to our work. Small quantity of segments corresponding to whole objects is described by colour and texture. Lastly, evaluation of colour descriptors can be found in [11].

2. Segmentation Based Image Descriptors

Segmentation maps act as spatial hypotheses highlighting distinct parts of objects as a whole. Multiple measurements can be taken from images within such defined areas. In nature, objects appear at different scales. Therefore, segmentation maps at different scales of observation were extracted and used to build more

Figure 1. Segmentations at few scales of observation (see text for details).

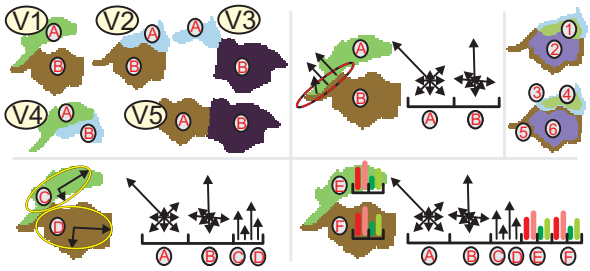

Figure 2. Architecture of the descriptors (see text for details).

accurate object representations. We used the implementation of Watershed segmentation reported in [4] as the top performer. Average numbers of segments per image were varied by factor of 1.6x between four consecutively coarser scales of observation $S_0, ..., S_3$ presented in figure 1 from top left towards bottom right.

Spatial arrangement. To establish a baseline system, we devised a basic descriptor such that each segment corresponded to one descriptor vector (single spatial bin). The statistics of orientations of image gradients were gathered within areas of segments including boundaries to form 12 dimensional vectors (we refer to it as V0). Next, in order to exploit semi-local image structures in the form of spatial arrangements of segments, all possible pairs (adjacent segments) were used to build descriptor comprising two spatial bins. Figure 2 (top left) illustrates how segments corresponding to the jockey's head displayed in figure 1 (bottom right) form pairwise combinations yielding vectors $\overrightarrow{V_1}, .., \overrightarrow{V_5}, ..., \overrightarrow{V_N}$. It is vital to ensure repeatability of such representation by preserving the order of spatial bins in descriptors. Therefore, segments are always grouped from top to bottom and from left to right. Figure 2 (top middle) depicts how pairwise statistics from regions A and B (note the order) are gathered to form a descriptor vector referred as V1. Numbers of orientation bins comprised on each spatial bin amounted to 8, 10, or 12 per experiment. Note, other combinations can be formed from pairs of segments. We investigated if including regions around boundaries of a segment pair independently from their interiors further improves representations. Therefore, vectors formed form segment pairs such as 3 and 5 in figure 2 (top right) were tested (V2). To measure levels of discriminative information within segment interiors only, another descriptor was designed by exploiting regions 4 and 6 as pairs of segments (V3). The statistics gathered only within small margins from boundaries of a joint segment $A \cup B$ capture primarily the edge between. Thus, influence of strong gradients along boundaries of $A \cup B$, except their common boundary, is decreased. Thus, we combined regions 1 and 2 only (V4). Lastly, we attempted to answer if boundaries and interiors of segments convey complementary information. Therefore, regions 3, 4, 5, and 6 were arranged into four spatial bins forming descriptor (V5).

Capturing shape of segments. Shape of segments is captured by the orientations of image gradients in particular from segment boundaries. Though, dominant shapes of segments and their relation can be also a valuable cue. We used eigenvectors of segments to represent them. Figure 2 (bottom left) shows ellipses fitted into adjacent segments to capture their dominant axes. Extracted eigenvectors and eigenvalues provided auxiliary bins to feature vectors. Threefold scenarios were investigated. 4, 6, or 8 orientation bins were yielded by angles of eigenvectors $\phi_k = \phi(\overrightarrow{E}_k)$ and incremented by corresponding eigenvalues $E_k = \| \overrightarrow{E}_k \|$. Additional two histograms are: 4 bins conveying phase $\phi_1, ..., \phi_4$ and 4 bins consisting of eigenvalues $E_1, ..., E_4$ only.

Colour statistics. Colour cues are complementary with orientation-based features as explained in [11]. To capture a gist of a semi-local colour profile, low dimensional colour histograms were collected within regions delineated by segments. We experimented with Opponent, YUV (also luminance-normalised) colour spaces. To decide how to quantise histogram bins, we estimated their distributions on Pascal 2008 training dataset and concluded marginal distributions of chromaticity components (C_1, C_2) were Laplacian shaped. This suggested 2 dimensional distributions seemed partially independent (both colour spaces). Thus, twofold ideas were examined: 5x5 bins joint statistics of C_1 and C_2 per segment, and separate 5 bins C_1 and 5 bins C_2 statistics per segment. Note, both colour spaces are light intensity and shift invariant [11] and resemble each other.

Data assignment and normalisation. The soft assignment amongst spatial bins based on bilinear approximation was examined. In case of segmentation-based descriptors, linear weighting depends upon the distance from the boundary between segments A and B. Also, soft assignment of orientation bins was based

on the same principle. Several different measurements are taken within each spatial bin, to wit: the orientations of image gradients, eigenvalues of segments, histograms of colours. Therefore, we experimented with normalising each measurement separately within each spatial bin as well as all bins together. The best results were achieved for each type of information normalised to unit vectors separately for each bin except histogram of eigenvalues which was normalised as a whole.

3. Evaluation Results

Initial tests were performed on Pascal 2008 data (2111 training and 2221 validation images for testing) with Pyramid Match Kernel (PMK) and SVM classifier from [3] before final tests on Pascal 2007 [2] (2501 training, 2510 validation, and 4952 testing images) with χ^2 kernel and KDA classifier [9]. Both systems were trained for the same 20 object classes. For PMK, the same environment as in [4] was created (4 pyramid levels with branch factor 20). For χ^2, hierarchical k-means clustering with 10x400 clusters and soft assignment [12] were applied. Further, dense feature sampling on a regular grid with the intervals of 8, 14, 20, and 26 pixels was applied to generate reference SIFT [5] descriptors with patch radii of 16, 24, 32, and 40 pixels.

Average numbers of features per image. The performed experiments aimed at using low numbers of features. Scales $S_0, ..., S_3$ yielded 596, 590, 353, and 199 feature vectors per image (on average). Combined scales S_{123}, S_{0123}, and S_{01234} produced 1148, 1738, and 2202 vectors which favorably compares to 3690 densely sampled SIFT features.

Experiments on Pascal 2008 data. Initial experiments were carried out on Pascal 2008 as it consisted of a smaller dataset. Our goal was to compare different approaches for fusing segment statistics. Table 1 (top) summarises them in terms of the mean average precision (MAP) for all 20 categories. The experiments were performed for scale S_1 until stated otherwise. Dimensionality of the proposed descriptor variants is indicated in brackets. V0 is a single spatial bin descriptor (12D) as elaborated earlier. $V1_S^{Do}$ stands for a segment pair descriptor with 2x12 bins (both normalised separately, soft assigned) with built-in orientation invariance based on dominant orientation mechanism. Although it is known such invariance decreases performance of PMK, it is useful to know the trade-offs for the applications that require it. $V1_S^{o8}$ to $V1_S^{o12}$ are variants of V1 with 2x8, 2x10, and 2x12 orientation bins. According to results, an increase in the number of orientation bins leads to a slight increase of scores, but saturates quickly.

$V1_H$ is a 2x12 bins hard assigned variant of V1.

variant MAP %	V0 23.88	$V1_S^{Do}$ 22.6	$V1_S^{o8}$ 24.91	$V1_S^{o10}$ 26.6
$V1_S^{o12}$ 27.43	$V1_H$ 27.78	$V1_{HSb}$ 28.12	$V1_{HSbt}$ 28.45	V2 27.05
V3 17.26	V5 28.09	$V1_{HSbt}^{Eg}$ 28.65	$V1_{Op}^{Eg}$ 30.62	$V1_{\widehat{Op}}^{Eg}$ 29.61
$V1_{UV}^{Eg}$ 30.67	$V1_{S_{03}}^{Eg}$ **32.32**	$V1_{OpS_{03}}^{Eg}$ **34.00**	DSIFT **33.77**	
$V1_{HSbt}$ 39.14	$V1^{Eg}$ 39.73	$V1_{Op}^{Eg}$ 43.39	$V1_{OpF}^{Eg}$ 43.00	$V1_{S_{03}}^{Eg}$ 43.44
$V1_{OpS_{13}}^{Eg}$ **45.26**	$V1_{OpS_{03}}^{Eg}$ **46.02**	$V1_{OpS_{04}}^{Eg}$ **47.54**	DSIFT **44.81**	OSIFT **46.56**
OS+V1 ***53.81***	OS+V1* ***57.8***	BK ***61.82***	BK+V1 ***63.34***	

Table 1. MAP for (top) Pascal 2008 and (bottom) 2007 benchmarks.

$V1_{HSb}$ and $V1_{HSbt}$ are descriptors comprising histograms using hard assignment and gradients obtained with Sobel operator. As to the latter variant, gradient magnitudes below an arbitrarily low threshold were not included into orientation bins. Interestingly, in contrast to the observations in [12], hard assignment outperformed soft assignment. We attribute this to the boundaries between pairs of segments being already good hypotheses distributing gradients proportionally amongst bins. As a number of spatial bins is low, smoothing should be avoided to keep them distinct.

Further, alternative spatial arrangements of pairs of segments were investigated. V2 to V4 comprise two and V5 four spatial bins (cf. section 2). They all use hard assignment, Sobel-based gradient and noise thresholding as in $V1_{HSbt}$. Removing gradients from segment interiors did not bring any benefit (case of V2). Retaining interiors and removing boundaries of segments forming pairs resulted in some information being conveyed as the results of V3 show. This is due to smooth edge transitions along boundaries of some objects and their texture. Variant V4 focusing on the boundary between segments A and B was a poorer performer than V2. Variant V5 did not deem descriptors any more descriptive than ordinary $V1_{HSbt}$ in spite of 48 dimensions.

The remaining experiments in this section were concerned with exploitation of segment shapes, their arrangements and colour information. We selected the most successful variant $V1_{HSbt}$ and combined it with 3 other variants of eigenvector based representations. Descriptors using orientations of eigenvectors decreased the results whilst histograms of eigenvalues (4D) added to $V1_{HSbt}$ brought a new information in $V1_{HSbt}^{Eg}$ (28D). For clarity, let us drop the subscript and call the most successful variant as $V1^{Eg}$. Its extensions with 2x2x5 bins of Opponent colour statistics are denoted as $V1_{Op}^{Eg}$ (48D) and $V1_{\widehat{Op}}^{Eg}$ (luminance-normalised), and for YUV

as $V1^{Eg}_{UV}$ (48D). The best results were delivered by $V1^{Eg}_{UV}$ and $V1^{Eg}_{Op}$. Finally, to benefit from multiple segmentations, feature vectors were appended across scales $S_0, ..., S_3$ to form $V1^{Eg}_{S_{03}}$ (28D) and $V1^{Eg}_{OpS_{03}}$ (48D). With 5.6x less data, the latter variant outperformed dense SIFT.

Experiments on Pascal 2007 data. Having identified the best configurations, further tests were performed on a larger dataset of Pascal 2007 (testing on the testing set). Apart from robust classification, this section is also concerned with gauging complementarity of the designed representations. If complementary, they can be fused together to score higher. For this purpose, χ^2 kernels built from the most promising descriptor variant and the state-of-the-art kernels from [9] were combined together by adding them. Table 1 (bottom) presents MAP classification results for both separate kernels and the most interesting fusions. As previously, $V1^{Eg}$ seemed to score a bit higher than $V1_{HSbt}$. This confirms that 4D histogram of eigenvalues conveys some information. Since Opponent and YUV spaces are similar, we report only results for $V1^{Eg}_{Op}$ (48D) and $V1^{Eg}_{OpF}$ which extends $V1^{Eg}$ with 2x5x5 colour bins (78D). In spite of higher dimensionality of joint distributions, no additional information was captured as they resembled the product of the marginal colour distributions.

To achieve better scale invariance, a collection of descriptor vectors at multiple scales was used. $V1^{Eg}_{OpS_{13}}$ is a collection of $V1^{Eg}_{Op}$ across scales $S_1, ..., S_3$. It performed on a par with dense SIFT with 8.6x less data. $V1^{Eg}_{OpS_{03}}$ turned out again the winning descriptor variant. $V1^{Eg}_{OpS_{04}}$ outperformed dense Opponent SIFT [11] with 13.4x less data. OS+V1 denote a kernel fusion of Opponent SIFT and our best descriptor. In spite of both using colours, their combination resulted in significant gain in performance. OS+V1* are the results of $V1$ merged with spatial version of kernel OS [11], which improves the results by 4%. Lastly, BK is a range of kernels built from multiple state-of-the-art descriptors [9]. BK+V1 is their fusion with our kernel based on $V1^{Eg}_{OpS_{04}}$, with further 5.5% improvements.

4. Conclusions

The experiments proved segmentation-based image descriptors proposed in this paper as highly informative, competitive and complementary to SIFT features. A clear benefit of such representation was noticeable also during k-means clustering. Reduction of dimensionality and feature numbers resulted in improved efficiency of the clustering approach. Unsupervised segmentations turned out to deliver good spatial hypotheses breaking objects down onto descriptive parts at multiple scales of observation. Further, such representation provided full coverage of images as opposed to sparse sampling. With 63.34%, the final kernel BK+V1 outperformed state-of-the-art systems scoring 62.2% in [14]. Such performance gain is significant compared to scores obtained by other systems for Pascal 2007 dataset.

Acknowledgements. This research was sponsored by the BBC Future Media and Technology and EPSRC EP/ F003420/1 research grants.

References

[1] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1026–1038, 1999.

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://pascallin.ecs.soton.ac.uk/challenges/VOC, 2007.

[3] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. *ICCV*, 2:1458–1465, 2005.

[4] P. Koniusz and K. Mikolajczyk. Segmentation based interest points and evaluation of unsupervised image segmentation methods. *BMVC*, 2009.

[5] D. G. Lowe. Object recognition from local scale-invariant features. *CVPR*, 2:1150–1157, 1999.

[6] T. Malisiewicz and A. Efros. Improving spatial support for objects vis multiple segmentations. *BMVC*, 2007.

[7] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. van Goll. A comparison of affine region detectors. *IJCV*, (65):43–72, 2005.

[8] P. Ott and M. Everingham. Implicit color segmentation features for pedestrian detection. *ICCV*.

[9] M. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, K. van de Sande, and T. Gevers. Visual category recognition using spectral regression and kernel discriminant analysis. *Subspace Workshop at ICCV*, 2009.

[10] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. *CVPR*, pages 1–8, 2008.

[11] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. A comparison of color features for visual concept classification. In *CIVR*, pages 141–149, July 2008.

[12] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. *PAMI*, 99(1), 5555.

[13] S. A. J. Winder and M. Brown. Learning local image descriptors. *CVPR*, pages 1–8, 2007.

[14] J. Yang, Y. Li, Y. Tian, L.-Y. Duan, and W. Gao. Group-sensitive multiple kernel learning for object categorization. *ICCV*, 2009.