Object Matching Using Boundary Descriptors

Ognjen Arandjelović Swansea University, UK
ognjen.arandjelovic@gmail.com

Abstract

The problem of object recognition is of immense practical importance and potential, and the last decade has witnessed a number of breakthroughs in the state of the art. Most of the past object recognition work focuses on textured objects and local appearance descriptors extracted around salient points in an image. These methods fail in the matching of smooth, untextured objects for which salient point detection does not produce robust results. The recently proposed bag of boundaries (BoB) method is the first to directly address this problem. Since the texture of smooth objects is largely uninformative, BoB focuses on describing and matching objects based on their post-segmentation boundaries. Herein we address three major weaknesses of this work. The first of these is the uniform treatment of all boundary segments. Instead, we describe a method for detecting the locations and scales of salient boundary segments. Secondly, while the BoB method uses an image based elementary descriptor (HoGs + occupancy matrix), we propose a more compact descriptor based on the local profile of boundary normals' directions. Lastly, we conduct a far more systematic evaluation, both of the bag of boundaries method and the method proposed here. Using a large public database, we demonstrate that our method exhibits greater robustness while at the same time achieving a major computational saving – object representation is extracted from an image in only 6% of the time needed to extract a bag of boundaries, and the storage requirement is similarly reduced to less than 8%.

1 Introduction

The problem of recognizing 3D objects from images has been one of the most active areas of computer vision research in the last decade. This is a consequence not only of the high practical potential of automatic object recognition systems but also significant breakthroughs which have facilitated the development of fast and reliable solutions [6, 10, 11]. These mainly centre around the detection of robust and salient image loci (keypoints) or regions [6, 7] and the characterization of their appearance (local descriptors) [6, 8]. While highly successful in the recognition of textured objects even in the presence of significant viewpoint and scale changes, these methods fail when applied on texturally smooth (i.e. nearly textureless) objects [4]. Unlike textured objects, smooth objects inherently do not exhibit appearance from which well localized keypoints and thus discriminative local descriptors can be extracted. The failure of keypoint based methods in adequately describing the appearance of smooth objects has recently been demonstrated by Arandjelović and Zisserman [4] using images of sculptures [1].

http://dx.doi.org/10.5244/C.26.85

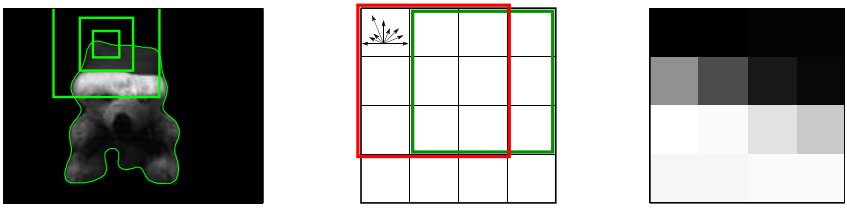
Smooth objects. Since their texture is not informative, characteristic discriminative information of smooth objects must be extracted from shape instead. Considering that it is not possible to formulate a meaningful prior which would allow for the reconstruction of an accurate depth map for the general class of smooth 3D objects, the problem becomes that of matching apparent shape as observed in images. This is a most challenging task because apparent shape is greatly affected by out of plane rotation of the object as shown in Figure 1. What is more, the extracted shape is likely to contain errors when the object is automatically segmented out from realistic, cluttered images, which is also illustrated in Figure 1. The bag of boundaries (BoB) method of Arandjelović and Zisserman was the first to address this problem explicitly; their approach is described in detail in Section 2.

Figure 1: As seen on the example in this figure (the second object from the Amsterdam Library of Object Images [5]), the apparent shape of 3D objects changes dramatically with viewpoint. Matching is made even more difficult by errors introduced during automatic segmentation. The leftmost image in the figure also shows automatically delineated object boundaries – one external boundary is shown in red and one internal boundary in green.

2 Bag of boundaries

The bag of boundaries method of Arandjelović and Zisserman [2] describes the apparent shape of an object using its boundaries, both external and internal as shown in Figure 1. The boundaries are traced and elementary descriptors extracted at equidistant points along the boundary. At each point at which descriptors are extracted, three descriptors of the same type are extracted at different scales, computed relative to the foreground object area, as shown in Figure 2(a).

Baseline descriptor. The semi-local elementary descriptor is computed from the image patch centred at a boundary point and it consists of two parts. The first of these is similar to the HoG representation of appearance [3]. Arandjelović and Zisserman [2] compute a weighted histogram of gradient orientations for each 8×8 pixel cell of the image patch which is resized to the uniform scale of 32×32 pixels, and concatenate these for each 3×3 cell region as illustrated in Figure 2(b). These region descriptors are then concatenated themselves, resulting in a vector of dimension 324 (there are 4 regions, each with 9 cells and each cell is represented using a 9 direction histogram) which is L_2 normalized. The second part of the descriptor is what Arandjelović and Zisserman term the occupancy matrix. The value of each element of this 4×4 matrix is the proportion of foreground (object) pixels in the corresponding region of the local patch extracted around the boundary, as shown in Figure 2(c). This matrix is rasterized, L_2 normalized and concatenated with the corresponding HoG vector to produce the final 340 dimension descriptor.



(a) Multiscale (b) HoG (c) Occupancy

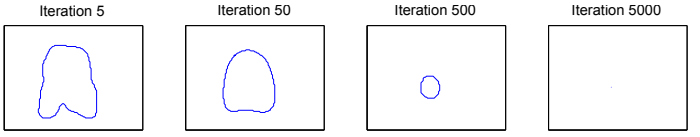
Figure 2: The bag of boundaries method of Arandjelović and Zisserman (a) extracts boundary descriptors at three scales (fixed relative to total object/foreground area), each descriptor consisting of (b) a HoG-like representation of the corresponding image patch and (c) the associated occupancy matrix.

Matching. Arandjelović and Zisserman apply their descriptor in the standard framework used for large scale retrieval. First, the descriptor space is discretized by clustering the descriptors extracted from the entire data set. The original work used 10,000 clusters. Each object is then described by the histogram of the corresponding descriptor cluster memberships. This histogram is what the authors call a bag of boundaries. Note that the geometric relationship between different boundary descriptors is not encoded and that the descriptors extracted at different scales at the same boundary locus are bagged independently. Finally, retrieval ordering is determined by matching object histograms using the Euclidean distance following the usual *tf-idf* weighting [13].

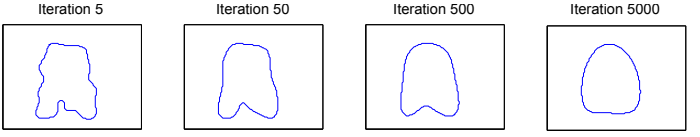
Limitations. The first major difference between the BoB method and that proposed in the present work lies in the manner in which boundary loci are selected. Arandjelović and Zisserman treat all segments of the boundary with equal emphasis, extracting descriptors at equidistant points. However, not all parts of the boundary are equally informative. In addition, a dense representation of this type is inherently sensitive to segmentation errors even when they are in the non-discriminative regions of the boundary. Thus we propose the use of a sparse representation which seeks to describe the shape of the boundary in the proximity of salient boundary loci only. We show how these loci can be detected automatically. The second major difference between the BoB method and ours, is to be found in the form that the local boundary descriptor takes. The descriptor of Arandjelović and Zisserman is image based. The consideration of a wide image region, as illustrated in Figure 2(a), when it is only a characterization of the local boundary that is needed is not only inefficient, but as an explicit description also likely not the most robust or discriminative representation. In contrast, our boundary descriptor is explicitly based on local shape.

3 Boundary keypoint detection

The problem of detecting characteristic image loci is well researched and a number of effective methods have been described in the literature; examples include approaches based on the difference of Gaussians [6] and wavelet trees [4]. When dealing with keypoints in images, the meaning of saliency naturally emerges as a property of appearance (pixel intensity) which is directly measured. This is not the case when dealing with curves for which saliency has to be defined by means of higher order variability which is computed rather than



(a) Simple smoothing



(b) Proposed method

Figure 3: (a) Boundary curve smoothing as Gaussian-weighted averaging produces the shrinking artefact, which eventually collapses the contour into its centre of mass. (b) In contrast, the proposed method preserves the circumference of the curve, only smoothing its curvature. In both cases shown is the effect of repeated smoothing with a Gaussian kernel with the standard deviation of 0.8% of the initial boundary circumference.

directly measured. In this paper we detect characteristic boundary loci as points of local curvature maxima, computed at different scales. Starting from the finest scale after localizing the corresponding keypoints, Gaussian smoothing is applied to the boundary which is then downsampled for the processing at a coarser scale. Having experimented with a range of factors for scale-space steps, we found that little benefit was gained by decreasing the step size from 2 (i.e. by downsampling finer than one octave at a time).

We estimate the curvature at the i-th vertex by the curvature of the circular arc fitted to three consecutive boundary vertices: $i-1$, i and $i+1$. The method used to perform Gaussian smoothing of the boundary is explained next.

Boundary curve smoothing. The most straightforward approach to smoothing a curve such as the object boundary is to replace each of its vertices c_i (a 2D vector) by a Gaussian-weighted sum of vectors corresponding to its neighbours:

$$c'_i = \sum_{j=-w}^{w} G_j \times c_{i+j} \tag{1}$$

where G_j is the j-th element of a Gaussian kernel with the width $2w+1$. However, this method introduces an undesirable artefact which is demonstrated as a gradual shrinkage of the boundary. In the limit, repeated smoothing results in the collapse to a point – the centre of gravity of the initial curve. This is illustrated in Figure 3(a). We solve this problem using an approach inspired by Taubin's work [12]. The key idea is that two smoothing operations are applied with the second update to the boundary vertices being applied in the "negative" direction. The second smoothing is applied on the results of the first smoothing:

$$c''_i = \sum_{j=-w}^{w} G_j \times c'_{i+j}, \tag{2}$$

resulting in the vertex differential:

$$\Delta c''_i = c''_i - c'_i. \tag{3}$$

The final smoothing result $\tilde{c}_i$ is computed by subtracting this differential from the result of the first smoothing, weighted by a positive constant K:

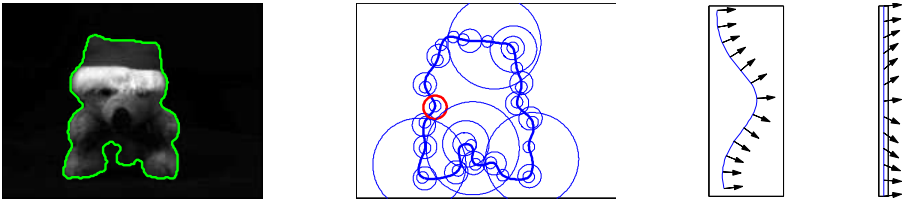
$$\tilde{c}_i = c'_i - K \times \Delta c''_i. \tag{4}$$

We determine the constant K by requiring that in the limit, repeated smoothing does not change the circumference of the boundary. In other words, repeated smoothing should cause the boundary to converge towards a circle of the radius $l_c/(2\pi)$ where l_c is the circumference of the initial boundary. For this to be the case, smoothing should leave the aforesaid circle unaffected. It can be shown that this is satisfied iff:

$$K = \frac{1}{\sum_{j=-w}^{w} G_j \times \cos(j\phi)} \tag{5}$$

where $\phi = 2\pi/n_v$ and n_v is the number of boundary vertices. The effects of smoothing a boundary using this method are illustrated on an example in Figure 3(b).

An example of a boundary contour and the corresponding interest point loci are shown respectively in Figures 4(a) and 4(b).



(a) Contour (b) Keypoints (c) Descriptor

Figure 4: (a) Original image of an object overlaid with the object boundary (green line), (b) the corresponding boundary keypoints detected using the method proposed in Section 3 and (c) an illustration of a local boundary descriptor based on the profile of boundary normals' directions (the corresponding interest point is shown in red in (b)).

4 Local boundary descriptor

Following the detection of boundary keypoints, our goal is to describe the local shape of the boundary. After experimenting with a variety of descriptors based on local curvatures, angles and normals, using histogram and order preserving representations, we found that the best results are achieved using a local profile of boundary normals' directions.

To extract a descriptor, we sample the boundary around a keypoint's neighbourhood (at the characteristic scale of the keypoint) at n_s equidistant points and estimate the boundary normals' directions at the sampling loci. This is illustrated in Figure 4(c). Boundary normals are estimated in a similar manner as curvature in Section 3. For each sampling point, a circular arc is fitted to the closest boundary vertex and its two neighbours, after which the desired normal is approximated by the corresponding normal of the arc, computed analytically. The normals are scaled to unit length and concatenated into the final descriptor with $2n_s$ dimensions. After experimenting with different numbers of samples, from as few as 4 up to 36, we

found that our method exhibited little sensitivity to the exact value of this parameter. For the experiments in this paper we use a conservative value from this range of $n_s = 13$.

We apply this descriptor in the same way as Arandjelović and Zisserman did their in the BoB method [2], or indeed a number of authors before them using local texture descriptors [11]. The set of training descriptors is first clustered, the centre of each cluster defining the corresponding descriptor word. An object is then represented by a histogram of its descriptor words. Since we too do not encode any explicit geometric information between individual descriptors we refer to our representation as a bag of normals (BoN).

5 Evaluation

In the original publication in which the bag of boundaries representation was introduced, evaluation was performed on a data set of sculptures automatically retrieved from Flickr [1, 2]. These experiments were successful in demonstrating the inadequacy of image keypoint based methods for the handling of smooth objects and the superiority of the BoB approach proposed by the authors. However, we identify several limitations of the original evaluation. Firstly, the results of Arandjelović and Zisserman offer limited insight into the behaviour of the representation with viewpoint changes. This is a consequence of the nature of their data set which was automatically harvested from Flickr and which contains uncontrolled viewpoint, of variable extent for different objects. In contrast, in this paper we perform evaluation using a data set which contains controlled variation, allowing us to systematically investigate the robustness of different representations to this particular nuisance variable. In addition, while the sculptures data set is indeed large, the number of objects which were actually used as a retrieval query was only 50. This reduces the statistical significance of the results. Herein we use a database of 1000 objects and query the system using each of them.

Data set. As the evaluation data set, we used the publicly available *Amsterdam Library of Object Images* (ALOI) [5]. This data set comprises images of 1000 objects, each imaged from 72 different viewpoints, at successive 5° rotations about the vertical axis (i.e. yaw changes). We used a subset of this variation, constrained to viewpoint directions of 0–85°. The objects in the database were imaged in front of a black background, allowing a foreground/background mask to be extracted automatically using simple thresholding, as illustrated using the first 10 objects in the database in Figure 5. This segmentation was performed by the authors of the database, rather than the authors of this paper. It should be emphasized that the result of aforesaid automatic segmentation is not perfect. Errors were mainly caused by the dark appearance of parts of some objects, as well as shadows. This is readily noticed in Figure 5 and in some cases, the deviation from the perfect segmentation result is much greater than that shown and, importantly, of variable extent across different viewpoints.

It is important to emphasize that the ALOI data set contains a variety of object types, some smooth and others which are not. This means that better matching results on this data set could be obtained by not ignoring textural appearance. Thus it should be understood that the results reported herein should not be compared to non-boundary based methods. Rather, the purpose of our evaluation should be seen specifically in the context of approaches based on apparent shape only.

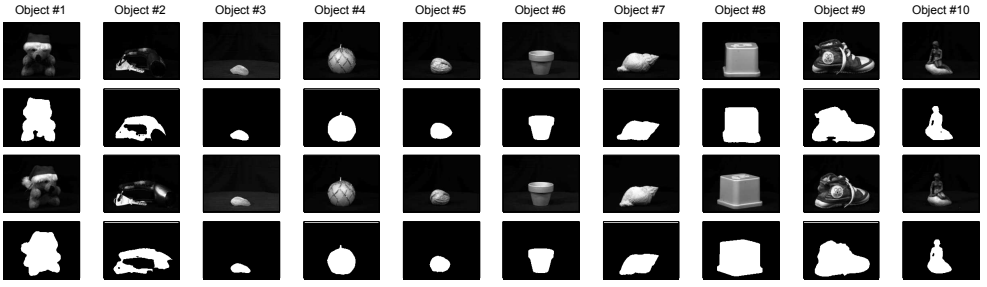


Figure 5: The first 10 objects in the *Amsterdam Library of Object Images* (ALOI) [5] seen from two views 30° apart (first and third row) and the corresponding foreground/background masks, extracted automatically using pixel intensity thresholding. Notice the presence of segmentation errors when a part of the object has dark texture, or when it is in the shadow.

Methodology. For both the BoB and BoN methods, we learn the vocabularies of the corresponding descriptor words using the 1000 images of all objects from the 0° viewpoint. We used a 5000 word vocabulary for the former method. Because the descriptor proposed herein is contour based, it inherently captures a smaller range of variability (not necessarily variability of interest) than the image based descriptor of Arandjelović and Zisserman, and is of a lower dimension, a smaller vocabulary of 3000 words was used for the BoN based method.

We perform three experiments:

- In the first experiment we compare the BoB and BoN representations in terms of their robustness to viewpoint change. The representations of all 1000 objects learnt from a single view are matched against the representations extracted from viewpoints at 5–85° yaw difference. Each object image is used as a query in turn.

- In the second experiment we compare the BoB and BoN representations in terms of their robustness to segmentation errors. The representations of all 1000 objects learnt from a single view are matched against the representations extracted from the same view but using distorted segmentation masks. In this experiment we distort the segmentation mask by morphological erosion using a 3×3 'matrix of ones' structuring element, as shown in Figure 6. Results are reported for 1–4 iterations of erosion and, as before, each object image is used as a query in turn.

- In the third experiment we also compare the BoB and BoN representations in terms of their robustness to segmentation errors. This time we distort the segmentation mask by morphological dilation using a 3×3 'matrix of ones' structuring element, as shown in Figure 6. As before, results are reported for 1–4 iterations of erosion and each object image is used as a query in turn.

Results. The key results of our first experiment are summarized in Figures 7(a) and 7(d). These plots show the variation in the average rank-N recognition rate for $N \in \{1, 5, 10, 20\}$ across viewpoint variations between training and probe images of 5–85°. Overall, the performance of the BoB and BoN representations was found to be quite similar. Some advantage of the proposed BoN representation was observed in rank-1 matching accuracy. For example,

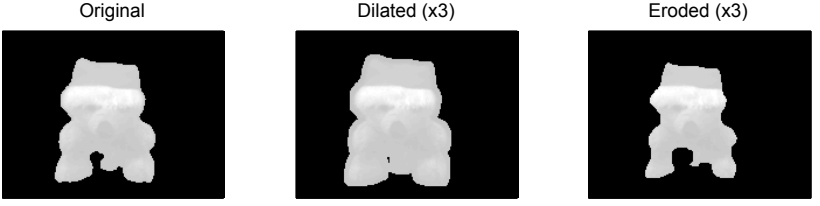
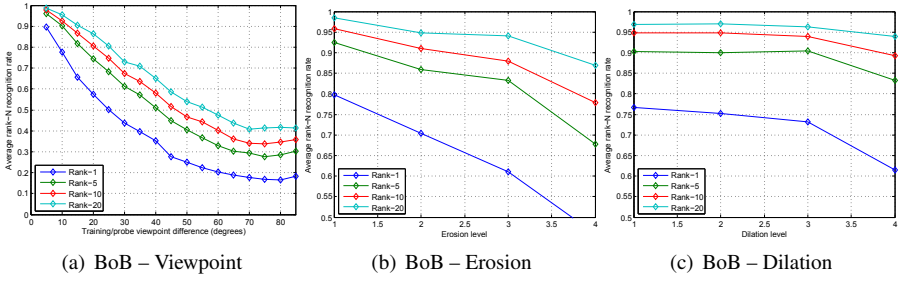
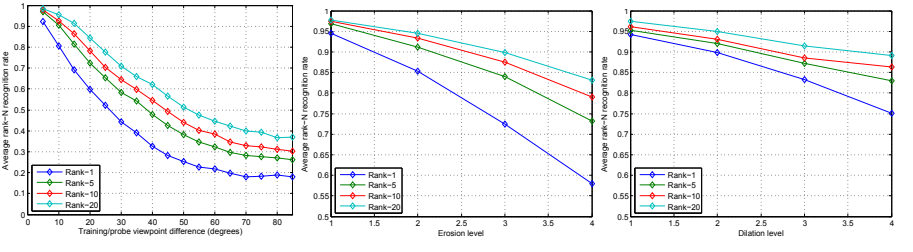


Figure 6: The robustness of boundary based representations of the apparent shape of objects to segmentation error is evaluated by matching objects using the initial, automatically extracted segmentation masks, against the same set of objects seen from the same view, but with a distorted mask. We examined segmentation mask distortions using 1–4 times repeated erosion or dilation, using a 'matrix of ones'p structuring element.

at 5° viewpoint difference between training and probe images, the average rank-1 matching rate of the BoB is 89.7% and that of BoN 91.5%. At 10° difference between training and probe, the average rank-1 matching rate of the BoB is 77.6% and that of BoN 80.3%. Using the results of matching at 5–30° viewpoint difference, by applying the least squares estimator on the logarithm transformed recognition rates, each 5° change in yaw can be estimated to decrease the BoB performance by approximately 12% and the BoN performance by approximately 10%.



(a) BoB – Viewpoint (b) BoB – Erosion (c) BoB – Dilation



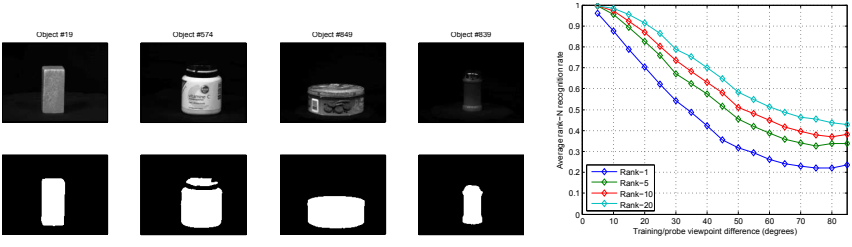
(d) BoN – Viewpoint (e) BoN – Erosion (f) BoN – Dilation

Figure 7: Summary of the results of the three experiments.

The results of the second and third experiments are summarized in Figures 7(b) and 7(e), and Figures 7(c) and 7(f) respectively. In these experiments, the superiority of the proposed BoN representation is more significant. For example, the distortion of the segmentation mask by two erosions reduces the rank-1 matching rate of the BoB by 30% and that of the BoN by half that i.e. 15%. The negative effects of dilation of the mask were less significant for both representations but qualitatively similar: repeated twice, dilation reduces the rank-1

matching rate of the BoB by 25% and that of the BoN by only 10%.

The objects which were most often difficult to match correctly and the corresponding objects that they were confused with are shown in Figure 8(a) for the BoB and in Figure 8(b) for the BoN. The pairs of mistaken objects can be seen to have similar local boundary shapes, but rather different global shapes. This suggests that one of the weaknesses of both the BoB and BoN representations is their lack of explicit encoding of the geometric relationship between different descriptor words. Similar findings have been reported in the context of local descriptor based representations of textured objects [9].



(a) Confusion – BoB (b) Confusion – BoN (c) BoB + BoN – Viewpoint

Figure 8: The two most confused objects for the (a) BoB and (b) BoN representations (shown are the corresponding raw images in the top row and their segmentations masks in the bottom row). (c) Viewpoint invariance is improved by a simple additive decision level combination of BoB and BoN representations.

We also investigated the possibility of a simple decision level combination of the two representations. Since in both BoB and BoN histograms are L_2 normalized, so are their Euclidean distances and we combine the corresponding BoB and BoN matching scores by simple summation. Viewpoint invariance, evaluated using the protocol of the first experiment described previously, is shown in Figure 8(c). From this plot the improvement is readily apparent. The average drop in rank-1 matching rate for each 5° change in yaw between the training and probe image set is reduced from 12% and 10% for BoB and BoN representations respectively, to 7%.

Lastly, we analyzed the computational demands of the two representations. The proposed BoN is superior to BoB in every stage of the algorithm. Firstly, the time needed to extract the proposed descriptors from a boundary is dramatically lower than those of Arandjelović and Zisserman – approximately 16 times in our implementation[1]. The total memory needed to store the extracted descriptors per object is also reduced, to approximately 8%. Unlike the descriptor of Arandjelović and Zisserman which is affected by the confounding image information surrounding the boundary, the proposed descriptor describes local boundary shape direction. Thus, the size of the vocabulary of boundary features need not be as large. This means that the total storage needed for the representations of all objects in a database is smaller, and their matching faster (the corresponding histograms are shorter).

[1] Matlab, running on an AMD Phenom II X4 965 processor and 8GB RAM.

6 Conclusions

In this paper we described a novel method for matching objects using their apparent shape i.e. the shape of the corresponding boundaries between the segmented foreground and background image regions. The proposed method is sparse because each object is represented by a collection of local boundary descriptors extracted at salient loci only. We proposed a method for detecting salient boundary loci based on local curvature maxima at different scales and circumference preserving smoothing, and a novel descriptor which comprises a profile of sampled boundary normals' directions. Evaluated on a large data set, the proposed method was shown to be superior to the state of the art both in terms of its robustness to viewpoint and segmentation mask distortions, as well as its computational requirements (time and space). Our results suggest that future work should concentrate on the representation of the global geometric relationship between local descriptors.

References

[1] *http://www.robots.ox.ac.uk/~vgg/research/sculptures/*, Last accessed May 2012.

[2] R. Arandjelović and A. Zisserman. Smooth object retrieval using a bag of boundaries. *In Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 375–382, November 2011.

[3] N. Dalai and B. Triggs. Histograms of oriented gradients for human detection. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1:886–893, 2005.

[4] J. Fauqueur, N. Kingsbury, and R. Anderson. Multiscale keypoint detection using the dual-tree complex wavelet transform. *In Proc. IEEE International Conference on Image Processing (ICIP)*, pages 1625–1628, 2006.

[5] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The Amsterdam library of object images. *International Journal of Computer Vision (IJCV)*, 61(1):103–112, 2005.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2003.

[7] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(10):1615–1630, 2004.

[8] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision (IJCV)*, 65(1/2):43–72, 2005.

[9] D. Parikh. Recognizing jumbled images: The role of local and global information in image classification. *In Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 519–526, 2011.

[10] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. *In Proc. IEEE International Conference on Computer Vision (ICCV)*, 2: 1470–1477, 2003.

[11] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. *In Proc. IEEE International Conference on Computer Vision (ICCV)*, 2005.

[12] G. Taubin. Curve and surface smoothing without shrinkage. *In Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 852–857, 1995.

[13] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems*, 26(3):1–37, 2008.