Novel Image Feature Alphabets for Object Recognition

Martin Lillholm and Lewis Griffin
Dept. of Computer Science, University College London
m.lillholm@cs.ucl.ac.uk

Abstract

Most successful object recognition systems are based on a visual alphabet of quantised gradient orientations. Here, we introduce two richer image feature alphabets for use in object recognition. The two alphabets are evaluated using the PASCAL VOC challenge 2007 dataset. The results show that both alphabets perform as well as or better than the 'standard' gradient orientation based one.

1. Introduction

Object recognition has progressed rapidly during the last ten years. Performance scores are good [8, 11, 9] and several challenging benchmark databases have been established [2, 4]. Much of the success can be attributed to the use of of bag-of-words like approaches, expressive encodings of local image structure such as the SIFT descriptor [7], and modern machine learning tools. A 'standard' approach typically comprises interest point detection, SIFT-based patch description, a visual vocabulary derived from a quantisation of SIFT-space, a histogram-of-visual-words based image description, and finally a classification scheme. In this pipeline, the SIFT-descriptor uses a visual alphabet of quantised gradient orientation to encode local image structure and the subsequent quantisation provides a set of visual words — typically in the order of hundreds or thousands.

In this paper, we investigate to what extent more complex visual alphabets can benefit and simplify object recognition. We introduce two novel feature alphabets and evaluate them using the dataset provided by the PASCAL VOC Challenge 2007 [2].

The rest of this paper is organised as follows. In section 2, we briefly discuss visual alphabets and introduce Basic Images Features and oriented Basic Image Features. Section 3 focuses on visual words and present three simple visual word encoding schemes. Section 4 reviews the PASCAL VOC Challenge and presents our classification framework. Sections 5 and 6 contain results and discussion respectively.

2. Visual Alphabets

Image intensities are the simplest visual alphabet. Although raw pixel intensities have been found useful for e.g., texture classification [10], they inherently suffer from lack of invariance and robustness to simple image perturbations and transformations. Several alternatives have been suggested for object recognition such as quantised gradient orientation [7]. The bulk of related research contributions [11], however, focus on the subsequent encoding of simple alphabets into region descriptions or visual words.

We suggest using a richer and more invariant feature alphabet even at this early stage of the processing pipeline.

2.1. Basic Image Features

Rich labelling of image positions according to its type of geometrical structure is well established in the computer vision literature. Typical schemes detect and *sparsely* label feature types as e.g., edges and corners [6].

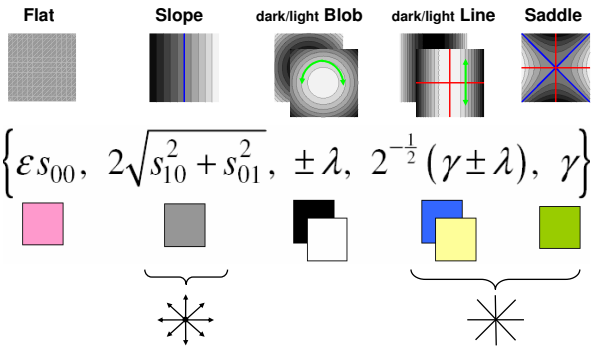

Figure 1. Basic Image Features.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

In [5, 1], the authors presented a *dense* labelling scheme that labels all image positions as one of seven possible types of local image symmetry. This alphabet of Basic Image Features (BIFs) is based on a group theoretical derivation of local image symmetry sensitivities of a 2^{nd} order filter bank of Gaussian derivatives [5]. The seven BIFs are listed and illustrated in the top two rows of Figure 1. An image location is classified as the symmetry type with the largest of the seven expressions in Figure 1 row 3. Here, $\lambda = \sigma^2(s_{20} + s_{02})$, $\gamma = \sigma^2((s_{20} - s_{02})^2 + 4s_{11}^2)^{\frac{1}{2}}$, and $s_{mn} = \left\langle \frac{\partial^{m+n}}{\partial x^m \partial y^m} G_\sigma^{(m,n)} \mid I \right\rangle$ is the inner product between the image I and a two-dimensional Gaussian Derivative. For simplicity, it is assumed that the probed image position is at the origin. The parameter, ϵ, controls the fractional contrast below which the image is considered flat. The fourth row in Figure 1 is the colour scheme used for the seven possible types of symmetry — an example of an image and its BIF map can be seen in Figure 2. The parameter σ controls the scale of the Gaussian kernel and can be used to generate a 'scale space' of increasingly blurred BIF maps. The BIF labelling is invariant with respect to image translations/rotations and intensity scalings.

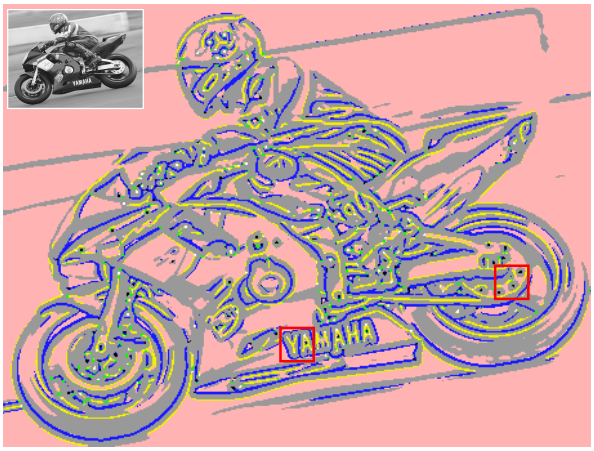

Figure 2. BIF example.

BIFs have successfully been used as the basis for a state-of-the-art multi-scale texture classification algorithm [1]. In the following, we assess this alphabet's utility for object recognition.

2.2. Oriented Basic Image Features

Given the success [8, 9] of gradient orientation based visual alphabets it is natural to consider local orientation in the context of BIFs. We initially note that a pure gradient based orientation scheme only makes sense for non-singular points and is likewise dubious for e.g., near-singular but noisy locations. Both caveats can be addressed satisfactory if local orientation is introduced as an augmentation to a BIF based categorisation: oriented Basic Image Features (oBIFs).

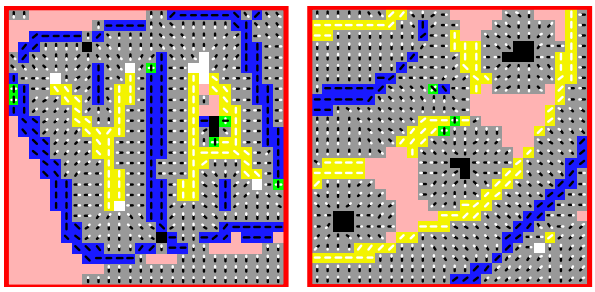

Figure 3. Oriented Basic Image Features.

Flat or near-flat locations (pink BIFs) should not be assigned an orientation at all. Slope-like points (grey BIFs) are dominated by 1^{st} order structure and can be assigned a meaningful gradient orientation — we use eight quantised orientation as suggested in [7]. The rotationally symmetric blob-like points (white or black BIFs) should, like flat point, not be assigned an orientation. Finally, line and saddle-like points (blue, yellow, and green BIFs) are dominated by their 2^{nd} order structure and the natural choice of orientation is the direction perpendicular to the largest eigenvalue of the Hessian (4 unoriented directions). This gives a total of 23 oBIFs as illustrated in the bottom two rows of Figure 1. Figure 3 shows a zoomed version of the two red squares marked in Figure 2 using oBIFs. If we only base oBIFs on 1^{st} order structure i.e., only use oriented pink and grey BIFs, we effectively get the visual alphabet used for SIFT features and as such this can be viewed as a 1^{st} order oBIFs.

3. Visual Words

The next step in a standard object recognition pipeline is to encode visual words using the visual alphabet of choice. Several encodings schemes or descriptors have been suggested in the literature such as SIFT, RIFT, and SPIN [11]. Slightly simplified, most descriptors lie somewhere in the spectrum between a template and a histogram with e.g., templates of histograms or histograms of templates as examples of in-between descriptors as illustrated in Figure 4 a). As an example, the popular SIFT descriptor is, roughly, a 4×4 template of histograms (of gradient orientation) each with a 4×4 pixel support. Although several details are omitted for clarity, we use this as the basis for

the evaluation of the two suggested BIF based visual alphabets and present results for the two extremes of the spectrum: templates and histograms.

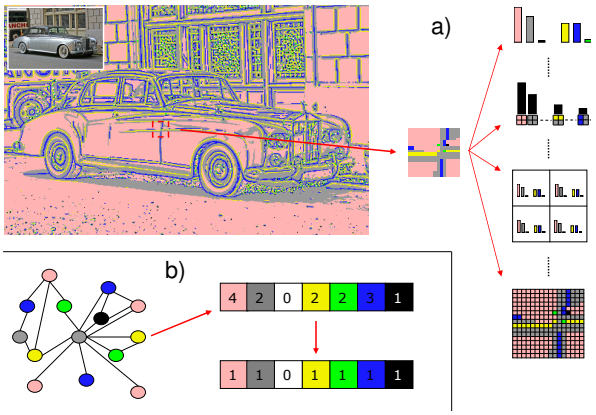

Figure 4. Encoding of Visual Words.

Figure 4 b) illustrates another potential descriptor: a region based graph which contrary to histogram or template based descriptors encodes the full topology within the region of support. Although a desirable property, this is also a computationally very demanding descriptor and we suggest using a simple approximation. The first numbered row next to the graph gives the corresponding region count for each type of BIF and finally the second row contains binary indications of the presence of a type of region. We will also evaluate this region presence descriptor to get a first indication of the potential of a graph based descriptor. In the following, we will use Basic Image Patterns (BIPs) for visual words based on BIF or oBIFs.

4. Evaluation Framework

The PASCAL VOC Challenge 2007 [2] dataset contains 9963 images and has been manually annotated for 20 object classes. The dataset is split into roughly 25% training images, 25% validation images, and 50% test images. An evaluation framework has been trained and tuned on the training and validation sets respectively and results are presented for the test set. Results are averaged over all 20 classes and presented as the area under the ROC curve (AUC-ROC).

For each image, we calculate a scale space of BIFs (or oBIFs) and extract BIPs *densely* for all combinations of position and scale. We use $\epsilon = 0.05$ throughout. The union of BIPs from all images form the list of uniquely occurring BIPs. This list is pruned by removing all BIPs that occur in only one image scale space. For each of the 20 object classes, the 1000 most informative BIPs are selected as features. Information gain is measured using mutual information for the initial selection and additional mutual information once one or more features are selected. This iterative feature selection scheme is adopted from Fleuret [3]. Subsequently a Näive Bayes Classifier is trained for each class and the initial list of 1000 features selected is reduced to the most informative prefix using the validation set.

Note that all calculations are based on exact feature matches. There is no data-driven quantisation step and mutual information is thus based solely on the absence and presence counts of the features relative to the class in question.

We evaluate the BIF based alphabets using each of the suggested BIP-schemes: templates, histograms, and presences. We also include a pure gradient orientation based (SIFT-like) alphabet for comparison. Each alphabet/BIP combination is evaluated using regions of support from 2×2 up to 16×16. For this initial study, all results are obtained using down-sampled images. The average image size used is approximately 40×30 whereas the original mean image size is approximately 460×380. We cannot expect to obtain competitive [8, 9] scores using images that small; but the comparative nature of this study is unaffected.

5. Results

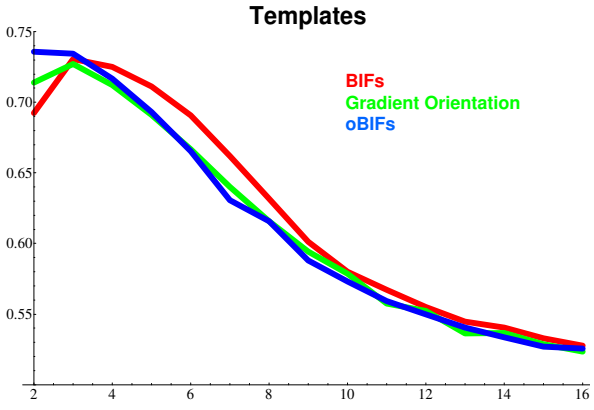

Figure 5. Template results.

The results for templates, histograms, and presences are shown in Figures 5, 6, and 7 respectively. Each plot gives the AUC-ROC as a function of the size of the region of support for the BIP encoding and contains a graph for each of the three evaluated alphabets: BIFs, oBIFs, and gradient orientation.

The template results show that the three alphabets have similar peak performances when encoded with no spatial slack. BIFs slightly outperform gradient orien-

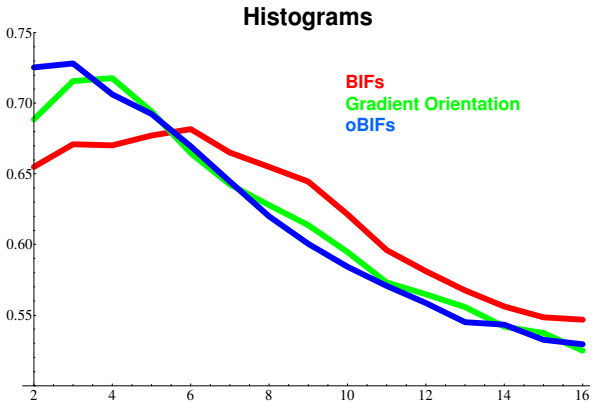

Figure 6. Histogram results.

tation and oBIFs outperform both. oBIF, BIFs, and gradient orientation obtain optimal scores for supports of 2×2, 3×3, and 3×3 respectively.

For histograms, the most notable results are that BIFs are outperformed by both gradient orientation and oBIFs and that oBIFs outperform gradient orientation. With no spatial constraints in the encoding, the importance of local orientation becomes clear. oBIF, BIFs, and gradient orientation obtain optimal scores for supports of 3×3, 4×4, and 6×6 respectively.

The presence results further emphasise oBIFs as the better of the three alphabets but perhaps more interestingly the best oBIF presence score is also the best score across all BIP and BIF types.

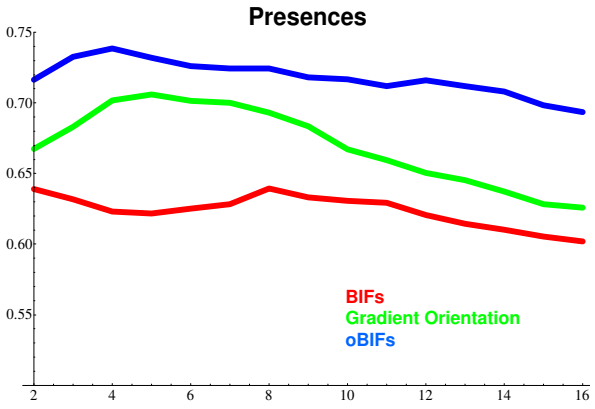

Figure 7. Presence results.

Initial experiments suggest that the described framework would yield results at the median level of the PASCAL 2008 entrants if applied to full-resolution images, but a more complete study is needed to confirm this.

6. Discussion

We have presented two novel feature alphabets in the context of object recognition and compared their performance to the 'standard' gradient orientation alphabet. The first, BIFs, achieve overall performance as good as or better than gradient orientations; specifically for templates. BIFs do, however, not perform as well as the other two alphabets using local descriptors with less spatial grip. The second alphabet, oBIFs, can be seen as a natural 2^{nd} order generalisation of gradient orientation. In terms of performance, oBIFs deliver the best overall score which is achieved for the simplest of the three tested descriptors. We conclude that larger feature alphabets can lead to both better performance and simpler encodings of visual words. Furthermore, the simple graph-approximating presence BIPs perform best and shows promise for future research.

References

[1] M. Crosier and L. Griffin. Texture classification with a dictionary of basic image features. In *CVPR*, 2008.

[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://www.pascal-network.org/.

[3] F. Fleuret. Fast Binary Feature Selection with Conditional Mutual Information. *The Journal of Machine Learning Research*, 5:1531–1555, 2004.

[4] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[5] L. Griffin and M. Lillholm. Symmetry-sensitivities of derivative-of-gaussian filters. Submitted IEEE PAMI.

[6] T. Lindeberg. Scale-space: A framework for handling image structures at multiple scales. *In Proc. CERN School of Computing*, 1996.

[7] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[8] M. Marszalek, C. Schmid, H. Harzallah, J. van de Weijer, and M. Vision. Learning Object Representations for Visual Object Class Recognition. http://lear.inrialpes.fr/pubs/2007/MSHV07/Marszalek Schmid-VOC07-LearningRepresentations-slides.pdf.

[9] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. *European Conference on Computer Vision*, 4:464–475.

[10] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *CVPR*, volume 2, pages 691–698, June 2003.

[11] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *International Journal of Computer Vision*, 73(2):213–238, 2007.