Single-Histogram Class Models for Image Segmentation

F. Schroff[1], A. Criminisi[2], and A. Zisserman[1]

[1] Dept. of Engineering Science, University of Oxford,
Parks Road, Oxford, OX1 3PJ, UK
{schroff, az}@robots.ox.ac.uk
[2] Microsoft Research Ltd, Cambridge, CB3 0FB, UK
antcrim@microsoft.com

Abstract. Histograms of visual words (or textons) have proved effective in tasks such as image classification and object class recognition. A common approach is to represent an object class by a set of histograms, each one corresponding to a training exemplar. Classification is then achieved by k-nearest neighbour search over the exemplars.

In this paper we introduce two novelties on this approach: (i) we show that new compact *single* histogram models estimated optimally from the entire training set achieve an equal or superior classification accuracy. The benefit of the single histograms is that they are much more efficient both in terms of memory and computational resources; and (ii) we show that bag of visual words histograms can provide an accurate pixel-wise segmentation of an image into object class regions. In this manner the compact models of visual object classes give simultaneous segmentation and recognition of image regions.

The approach is evaluated on the MSRC database [5] and it is shown that performance equals or is superior to previous publications on this database.

1 Introduction

Segmenting natural images automatically in a bottom up fashion has a long history but has not been that successful – see [16] for a recent example and earlier references. Two more recent and fruitful trends are *class driven segmentation*, where object class models propose object localisations that can then refine a more local (bottom up) image segmentation [1, 2, 9, 11, 12, 17], and *interactive segmentation* in which a human supplies approximate segmentations and then refines and groups automatically generated image based segmentations [4, 15]. For example, consider a colour based segmentation of a patchy cow – a purely bottom up segmentation will tend to separate the image into many different regions rather than recognising the cow as a single, coherent object – there is a clear need for segmentation and recognition to work together.

Many class driven recognition and segmentation algorithms represent the object class or texture using multiple exemplars [1, 2, 9, 11, 12, 20]. One contribution of this paper is to show that equal or superior recognition results can be

P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 82–93, 2006.

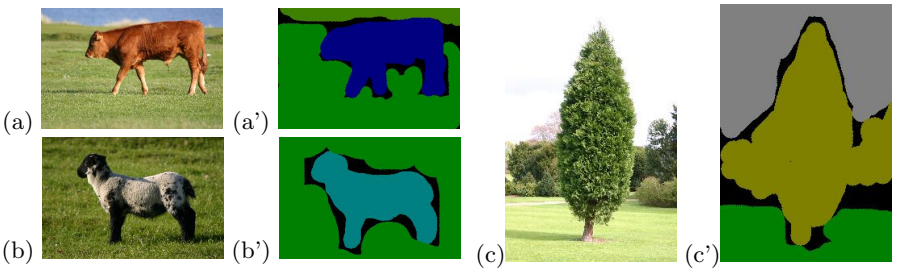


Fig. 1. The image database. (a–c) Example images from the MSRC database and (a'–c') their ground-truth class segmentation maps. The colour indicates the object class, with different kinds of grass, for example, associated with a common "grass" label. Note the presence of unlabelled (black) pixels.

obtained by a single class model if an appropriate distance measure is used, and also to explain why this result comes about. A second contribution of this paper is to show that pixel-wise *segmentations* can be obtained from sliding windows using class models.

In more detail we represent an object category by a single histogram of dense visual words, and investigate the effectiveness of this representation for segmentation. The advantage of a single class histogram is a very compact, and consequently computationally efficient, representation. Histograms of visual words have been used previously for region or image level classification [6, 8, 14, 18, 23], though for the most part based on sparse descriptors. Others that have used dense descriptors [3, 22, 8] have only considered soft segmentations based on the support of the visual words, rather than explicit pixel-wise classification.

Previous authors [22] have also investigated representing each class in a compact way using a single Gaussian model of each category. The class models explored here are even simpler and more efficient since they consist of simple histograms without a covariance. We compare the performance of our class models with those of [22] using the same data sets.

For the experiments in this paper we use the MSRC image database [5] (see figure 1). The database contains many classes including grass, trees, sheep, buildings, bicycles and others, seen from different viewpoints and under general illumination conditions. A coarse region level ground truth labelling is available, and this is used to learn the class histograms, and also to assess the pixel level and region level classifications during testing. For example, in the 120 training images there are 64 labelled grass regions and 22 labelled cow regions.

2 Background: Features, Visual Words and Histograms

This section illustrates the basic algorithms for estimating the object class models and the intermediate data representation necessary for classification. The training and testing steps have much in common and are briefly described next.

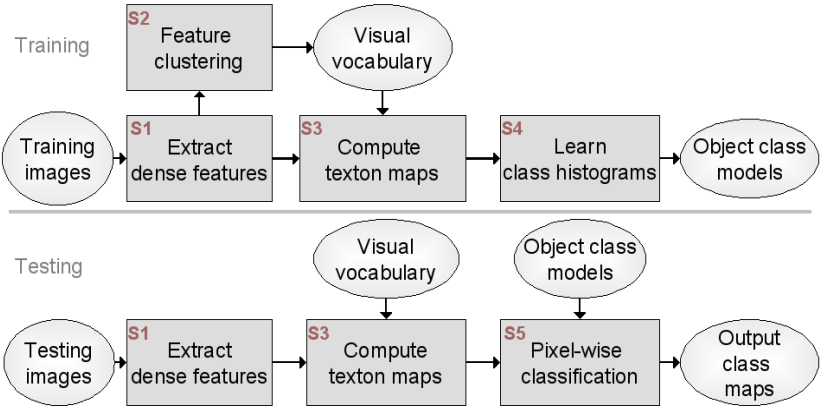


Fig. 2. Flow diagram for the training and testing algorithms

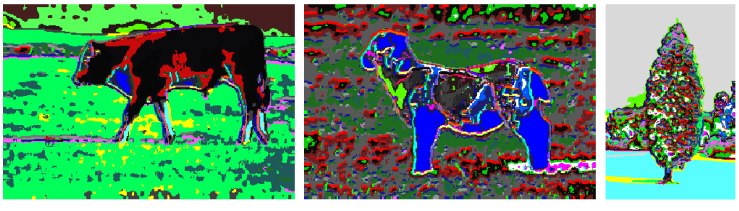


Fig. 3. Texton maps for the images of figure 1. Different colours uniquely identify different textons/visual words. A small visual vocabulary with only 50 words has been used here for illustration.

In this paper, feature vectors are estimated densely, i.e. at each pixel location. The actual feature vectors (step `S1` in figure 2) are raw 3×3 or 5×5 colour patches [3, 20] in the *CIE-LAB* colour space. Thus, their dimensionality is 27 or 75, respectively.

During training a vocabulary of V visual words (also called textons, [13, 20, 21, 22]) is built by clustering the feature vectors extracted from many training images (step `S2`). Feature clustering is performed by K-means on a randomly sampled 25% subset of feature vectors using equal numbers from each training image. Note, a suitable degree of invariance (to lighting, rotations, scale etc) is learnt implicitly from the training images (since these provide examples of lighting changes etc), and no additional invariances are built in.

Given the set of cluster centres (these are the visual words or textons), it is now possible to associate each pixel in the training images with the closest visual word in the vocabulary (step `S3`). The result may be visualised by generating colour-coded word maps such as in figure 3.

Finally, we compute histograms of visual words for each of the training regions (step `S4`). Those histograms can then be (i) stored separately as training exemplars, or (ii) combined together to produce compact and yet discriminative

models of object categories. Here we estimate such *single-histogram* class models and demonstrate classification accuracy comparable to standard k-nearest-neighbour classification (k-NN) on the exemplars.

During testing (step `S5`), an input image is converted into its corresponding texton map. Then, pixel-wise classification is obtained by means of a sliding window technique. A window of dimension $(2w + 1) \times (2w + 1)$ is slid across the image to generate a histogram of visual words for each position. The centre pixel is then classified according to the closest class histogram. In this manner an image can be segmented into the various classes it contains, for example into pixels arising from grass, trees or sheep.

3 Single-Histogram Models for Efficient Classification

This section describes details of our class model estimation algorithm (step `S4` in figure 2). During training the histograms corresponding to different training regions (exemplars) belonging to the same class are combined together into a single, optimally estimated class histogram. During testing for pixel-wise classification, a histogram is computed for *each* pixel of the test image using a sliding window, and this histogram is then compared to each of the C (the number of classes) class histograms (as opposed to each of the (possibly many) training regions/exemplars in the case of k-NN classification). The use of single class histograms clearly reduces the classification cost. The class models are used both for the aforementioned pixel-wise classification, via a sliding window, and for region level classification, explained later on.

The key question then is how to compute such single-histogram models. Let $\mathbf{p}$ be one of the exemplar histograms and $\mathbf{q}$ the single histogram model that we seek. Histograms are represented as V-vectors, with V the vocabulary size. For a given class c, the "optimal" class histogram $\mathbf{q}$ is the one which minimises the overall distance to all the N_c exemplar histograms $\mathbf{p}^j$, as this minimises intra-class variability. Ideally, for best discrimination, one would also like to maximise the inter-class variability, and we return to this point later. The optimal solution $\hat{\mathbf{q}}$ depends on the histogram distance function used during classification. In this paper we analyse and compare the two most common alternatives: (i) a Kullback-Leibler divergence (D_{KL}), and (ii) a Euclidean distance (D_{L2}). The same framework may also be applied to other distance measures, such as histogram intersection, χ^2, Bhattacharyya or Alpha-Divergence.

Kullback-Leibler divergence: The KL divergence between the two normalised histograms $\mathbf{a}$ and $\mathbf{b}$ is defined as:

$$D_{KL}(\mathbf{a} \parallel \mathbf{b}) = \sum_i a_i \log \frac{a_i}{b_i} \quad .$$

The subscript i labels the bins (a_i or b_i), with $i = 1 \ldots V$.

Given a class c we seek the model histogram $\hat{\mathbf{q}}$ which minimises the following cost:

$$E_{KL} := \sum_{j=1}^{N_c} n^j D_{KL}(\mathbf{p}^j \parallel \mathbf{q}) \quad \text{subject to } \parallel \mathbf{q} \parallel_1 = 1,\ q_i \geq 0\ \forall i \quad , \tag{1}$$

where n^j denotes the number of pixels in the j^{th} exemplar region, and is used as a weight to each exemplar histogram. N_c is the number of exemplar regions for the object category c. The normalised histogram for the j^{th} exemplar image region in class c is denoted $\mathbf{p}^j$. Note that the weighting factors n^j could be set to one, thus treating all training exemplars equally. Both versions were explored and gave comparable results.

Standard manipulation yields the global minimum of (1) as:

$$\hat{\mathbf{q}} := \frac{\sum_j n^j \mathbf{p}^j}{\sum_j n^j} \quad . \tag{2}$$

It can be shown [7] that $\hat{\mathbf{q}}$, with n^j as defined, corresponds to the maximum likelihood estimate of the visual word distribution for class c given its N_c training region visual words. In other words, $\hat{\mathbf{q}}$ describes the overall visual word distribution in all training regions.

During classification, given a query image sliding window, or region and its corresponding histogram $\mathbf{p}$, the closest class model $\tilde{\mathbf{q}} = \arg\min_{\mathbf{q}} D_{KL}(\mathbf{p} \parallel \mathbf{q})$ is chosen, i.e. $\tilde{\mathbf{q}}$ is the model that best explains $\mathbf{p}$ and the corresponding class the most likely one.

Euclidean Distance: The Euclidean distance between the two histograms $\mathbf{a}$ and $\mathbf{b}$ is defined as:

$$D_{L2}(\mathbf{a}, \mathbf{b}) = \sum_{i=1} (a_i - b_i)^2 \quad .$$

Once again, given the class c and its exemplar histograms $\mathbf{p}^j$ we seek the histogram $\hat{\mathbf{q}}$ which minimises the following cost:

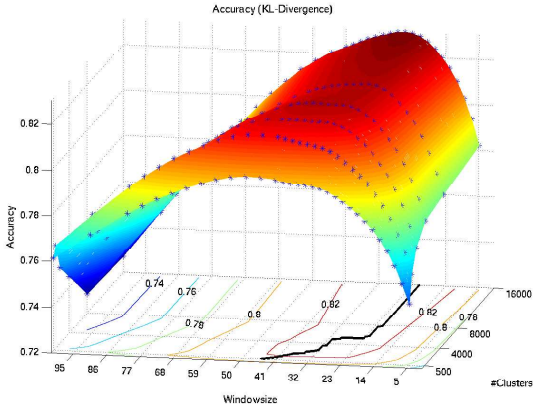
$$E_{L2} := \sum_{j=1}^{N_c} n^j D_{L2}(\mathbf{p}^j, \mathbf{q}) \quad \text{subject to} \quad \parallel \mathbf{q} \parallel_1 = 1,\ q_i \geq 0\, \forall i \quad . \tag{3}$$

Standard manipulation leads to the same $\hat{\mathbf{q}}$ as obtained by minimising (1), i.e. as given in (2).

Next we assess the discrimination power of the learnt class models by measuring pixel-wise classification performance.

4 Results and Comparative Evaluation

In this section we assess the validity of our models by measuring accuracy of segmentation/recognition against two subsets of the MSRC database [5]: a six class subset, *6-class* = { *cow, sheep, dog, cat, bird, grass* }; and a nine class subset, *9-class* = { *building, grass, tree, cow, sky, aeroplane, face, car, bicycle* } [22].



V ($w = 11$)	Acc. (%)	w ($V = 8000$)	Acc. (%)
500	79.1	5	80.3
1000	80.7	11	**82.4**
2000	81.7	15	**82.4**
4000	82.3	20	82.1
8000	82.4	26	81.1
16000	**83.0**	30	80

Fig. 4. Accuracy analysis on the 6-class set. Pixel-wise classification performance as a function of the size w of the sliding window and the size V of the visual vocabulary. The features are 27-dimensional 3×3 CIE-LAB patches. The vocabulary is learnt by K-means clustering run for 500 iterations. KL divergence is used for histogram comparisons.

The databases are split into 125 training and 50 test images for the *6-class* set, and 120 training and 120 test images for the *9-class* set. The visual vocabulary and class models are learnt from the training data only. As mentioned before, during testing a window of dimension $(2w+1) \times (2w+1)$ is slid across the image to generate a histogram of visual words for each pixel, and thereby classifying the centre pixel.

Accuracy of segmentation/recognition is measured by the proportion of test pixels correctly classified according to ground truth. Only the pixels belonging to one of the aforementioned classes are taken into consideration. In the remainder we refer to this accuracy as pixel-wise classification performance, as opposed to region-wise classification performance which is introduced later. In the following we first evaluate performance using the *6-class* set together with single class histograms over the system parameters: features (3×3, 5×5); number of iterations in K-means; vocabulary size V; and window size w. We then compare the performance of the single class histogram to that of using k-NN over all the exemplars.

The effect of the window and vocabulary sizes: The first set of experiments are designed to evaluate optimal values for the size of the sliding window w, the vocabulary size V, and the best feature clustering technique.

Figure 4 plots the pixel-wise classification accuracy as a function of both the window size w and the vocabulary size V. Two cross-sections of the accuracy function through the maximum are shown in the table. The maximum performance is reached for $w = 11 - 15$ and $V = 16,000$. Accuracy does not vary much over the range $V = 8,000$ – $128,000$, so from here on a vocabulary of size $V = 8,000$ is used to reduce computational cost. The optimal value $w = 12$

Table 1. Variations of K-means clustering. The mean (± one standard deviation) pixel-wise classification accuracy computed over multiple runs of K-means; with the number of runs used in each case shown in brackets. Different numbers of iterations of K-means for constructing the visual vocabulary on the *6-class* and *9-class* sets are compared. KL divergence together with single class histograms on 5×5 patches was used.

	KM, 0 iters	KM, 1 iter	KM, 10 iters	KM, 500 iters
6-class	$81.96 \pm 0.20\%$ (50)	$82.24 \pm 0.20\%$ (10)	$82.54 \pm 0.15\%$ (5)	$\mathbf{82.56} \pm 0.13\%$ (5)
9-class	$74.72 \pm 0.22\%$ (10)	$74.92 \pm 0.17\%$ (10)	$\mathbf{75.07} \pm 0.15\%$ (10)	–

Table 2. k-NN vs. single-histograms. Comparing the *pixel*-wise classification performance obtained by our single-histogram class models with that obtained from conventional nearest neighbour. In this case we used $V = 8000$ and 5×5 patches as features. K-means with 10 iterations was used to construct the visual vocabulary. For the 6-class set the best performing k out of $k = 1 \ldots 100$ and for the 9-class the performance for k-NN with $k = 1$ is reported. Using single-histogram class models in conjunction with KL divergence produces the best results.

	D_{KL}(6-class)	D_{L2}(6-class)	$D_{\chi 2}$(6-class)	D_{KL}(9-class)	D_{L2}(9-class)	$D_{\chi 2}$(9-class)
k-NN	82.1%	76.6%	78.7%	71.6%	65.1%	72.0%
single hist.	**82.4%**	77.0%	–	**75.2%**	58.7%	–

is also used. The performance is found not to depend much on the size of the feature (i.e. size of colour patch), 5×5 colour patches are used from here on.

The effects of different clustering techniques: In table 1 we compare the influence of different numbers of iterations in K-means clustering for the construction of the visual vocabulary. Zero iterations denote randomly sampled cluster centres from the feature space, which is how K-means is initialised in all cases. Interestingly the performance is only slightly affected by the number of iterations. In particular there is only a small gain in increasing from 10 to 500 iterations. From here on we use 10 iterations as a trade off between performance and computational time for the experiments.

Keeping all exemplar histograms vs. single-histogram class models: Next we compare the performance of single-histogram models with respect to conventional k-NN classification, and provide evidence for the main claim of the paper.

Table 2 summarises the results of applying a k-NN approach, i.e. maintaining all the exemplar histograms of each class separately, and our single-histogram class models. Classification performance is measured for both KL and L2 distance. In all cases the accuracy obtained by the proposed class models is comparable (if not superior) to that obtained by k-NN. Experiments were carried out for the 6-class and 9-class datasets, as shown (the optimal k in the k-NN was $k = 1$ for KL divergence, and $k = 3, 4$ for L2; for the 9-class set only $k = 1$ was used). Substituting L2 distance for KL divergence reduces the performance by nearly 6%. This confirms the better suitability of the KL divergence *for* single class histograms (see following discussion).

Table 3. Confusion matrices for the single class histogram method (see table 2). (a) for the 6-class set; achieving an overall pixel-wise classification accuracy of **82.4%**. (b) for the 9-class set; achieving a pixel-wise classification accuracy of **75.2%**. KL divergence is used in both cases.

GT\Cl	grass	cow	sheep	bird	cat	dog
grass	95.61	2.0	1.2	1.2		0.1
cow	3.8	71.9	6.4	1.0	5.4	11.5
sheep	3.2	12.0	62.7	4.3	4.9	13.0
bird	5.5	27.1	24.0	27.7	10.4	5.4
cat		5.5	12.4	6.9	69.8	5.5
dog	1.1	24.7	2.3	6.5	18.2	47.2

(a) conf. mat. for 6-class set

GT\Cl	build.	grass	tree	cow	sky	plane	face	car	bike
build.	56.7	0.0	4.8	3.0	2.2	12.8	1.4	11.6	7.5
grass	0.5	84.8	9.7	3.9		1.2			
tree	6.4	5.6	76.4	1.2	0.3	1.3		2.4	6.5
cow	1.9	2.4	2.7	83.8		0.2	4.5	3.7	0.8
sky	6.5		2.1		81.1	6.4		3.9	
plane	16.8	0.8	5.0	3.4	0.1	53.8		16.6	3.5
face	4.6	0.0	0.4	19.1		0.6	68.5	3.6	3.2
car	7.4		1.1	3.4	0.7	2.6	2.0	71.4	11.6
bike	9.9	0.1	4.8	2.9		1.5	0.1	8.8	72.0

(b) conf. mat. for 9-class set

Table 4. Region-wise classification. Comparing the *region*-wise classification performance obtained by our single-histogram class models with that obtained from conventional nearest neighbour. Shown are the *best* results if V is varied (V is shown in brackets). Results are comparable to previous published performances for this dataset [22].

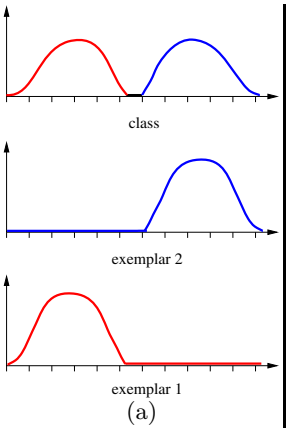
	1-NN (χ^2)	cl-Hist (KL)	1-NN ([22])	1-NN T ([22])
9-class	92.34 (for V = 4000 and V =32000)	93.43 (V=64000)	93.4	92.7

Table 3 shows the confusion matrices for selected experiments of table 2. The matrices are row normalised (so that the percentages in each row sum to 100%). Only pixels belonging to one of the classes are considered. For the 6-class set, the grass class is recognised most reliably, followed by cows, cats and sheep. This provides us with an idea of the relative difficulty of modelling each class. At this point one may think that our models work well only with texture-defined objects (grass, woolly sheep...). However, we also include classification of man made (less texture-like) objects such as cars and bicycles in the 9-class database (as also used in [22]). Table 3b presents the confusion matrix. The performance is still well above 70%, thus confirming the modelling power of the proposed class histograms (see following discussion).

4.1 Region Level Classification

Next we compare the accuracy of discrimination of our models with that achieved by the Gaussian models proposed in Winn et al. [22]. Following their evaluation methodology, we classify each input test *region*[1] as belonging to one of the classes in the database and measure the error with respect to ground truth. Table 4 shows that the proposed, simpler class models perform comparably. For this comparison the exact training/test splits were provided by the authors of [22].

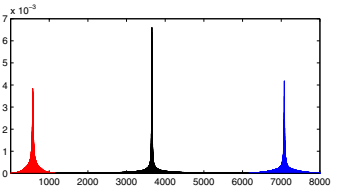
[1] The area of the region and its ground truth label is known.



(a)

(b)

	top	bottom
KL	2.29 (**cow**)	1.65 (**cow**)
	6.96 (dog)	3.75 (sheep)
	11.66 (bird)	4.83 (dog)
L2	0.019 (**cow**)	0.0028 (**dog**)
	0.202 (grass)	0.0032 (cat)
	0.202 (sheep)	0.0034 (cow)



(c)

Fig. 5. Advantages of KL. Different instances of cows induce different proportions of visual words. A unified "cow" model histogram (c) will contain different "modes" for the different visual aspects and species of the instances. (a) provides a schematic visualisation. In (c) the mode corresponding to the top cow in (b) is shown in red (left), and for the bottom cow in blue (right). The remaining visual words of the *cow*-model are shown in black in the middle. Note that a simple sorting of the visual words has been employed to bring out the different modes. The table shows the distances of the cow exemplars in (b) to the class models (showing the nearest class in bold). KL divergence ignores zero bins in the query histograms and is thus better suited for this scenario (note the wrong classification with L2 for the bottom cow).

Each of the methods (k-NN using χ^2 on exemplars, and KL for single-class histograms) are optimised separately over the size of the vocabulary V, and the best result is reported. χ^2 is reported for k-NN as this gives superior results to L2 and it is the standard distance measure for region classification on exemplars [20]. In both cases the features are 5×5 patches and the visual vocabulary was constructed with K-means (10 iterations). In addition to the results given in the table we experimented on the 6-class database using $V = 8000$. The result is similar in that the 1-NN χ^2 performance was 79.5% and the single-class histogram reached 85.5%.

4.2 Discussion

As the experiments demonstrate, KL divergence is superior to both $L2$ and χ^2 distance when the single-histogram models are used ([19] uses KL for similar reasons). This observation can be explained by the fact that the KL divergence does not penalise zero bins in the query histogram (which are non-zero in the model) as much as the other two distances. As a result of the way our class models are learnt, they are likely to have many non-zero bins due to the contribution of all training images' visual words to the model histogram. Query histograms that stem from a very specific object instance are very likely to have many zero bins. Consider the three schematic histograms shown in figure 5a. If $L2$ (or χ^2)

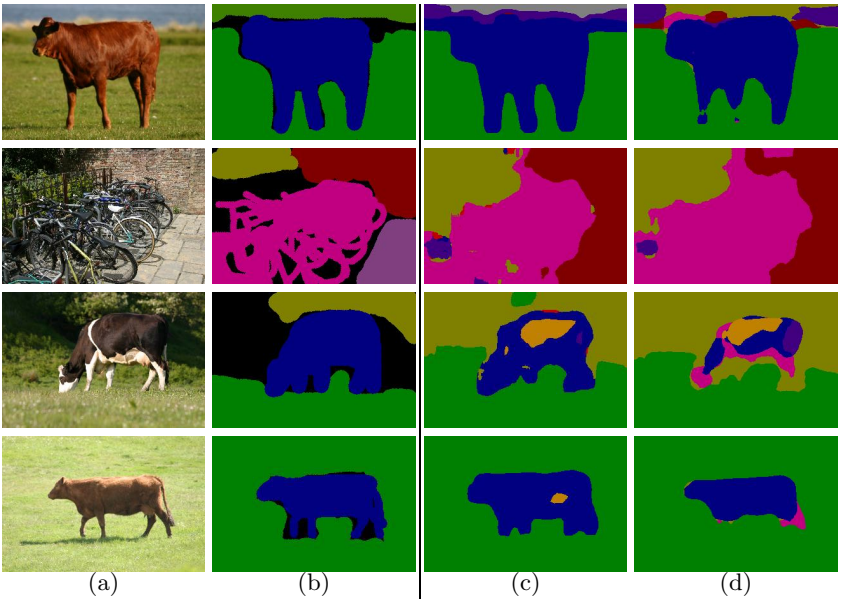


Fig. 6. Class segmentation results. (a) Original photographs. (b) Ground-truth class labels. (c) Output class maps obtained with KL divergence. (d) Output class maps obtained with L2 distance. In most cases L2 gives less accurate segmentation. In all cases our single-histogram class models were used, together with 5×5 patch features, $V = 8000$ and K-means clustering.

distances are used then each exemplar histogram will have a large distance from the class histogram (due to bins q_i of the mode in the class histogram which are not present in each of the exemplars). However, the KL divergence ignores all the null bins of the exemplar histograms (as these are zero p_i values in $p_i \log \frac{p_i}{q_i}$), thus making it a better suited distance. Figure 5c provides an example of such a multi-modal class histogram (here the cow model), and two exemplar regions inducing modes in the class model. The table shows the actual distances of the two cow regions to the three closest class models. In this case the bottom cow would be classified incorrectly as dog if L2 was used.

The optimal estimation of a class histogram is related to the topic vectors of Probabilistic Latent Semantic Analysis (pLSA) used in statistical text analysis [10]. Using the common terminology, each exemplar region represents a document by its word frequencies, visual words in our case. In the pLSA "learning" stage each exemplar is modelled by a topic distribution and each topic by a visual word distribution. In our case we use the additional information provided by the training data and hence define the topics to correspond to the object categories. Furthermore, each exemplar is constrained to be modelled by one topic only – the class assigned to it by the training annotation. Consequently, our method directly corresponds to pLSA in that it also minimises the KL divergence of the modelled data to the given data. The model is just more constrained in our case.

As mentioned earlier it would be desirable to maximise the inter-class distance when building the single-histograms. Maximising the inter-class distance or generally merging the class histograms in a discriminative way is left for future research. See [7] for related approaches.

Finally, figure 6 shows results of class segmentations of images. Note that the (visual) accuracy of the L2 classification results is inferior to that obtained with KL divergence.

5 Conclusion

This paper has introduced a new technique for the estimation of compact and efficient, generative single-histogram models of object classes. The models are applied to simultaneously segment and recognise images.

Despite their simplicity, our single-histogram class models have proved as discriminative as keeping around *all* exemplar histograms (and classifying via nearest neighbour approaches). The main advantage being their storage economy, computational efficiency and scalability. Note, the computational efficiency is a significant advantage since methods for speeding up nearest neighbour search, such as k-D trees, do not perform well in high dimensions. Here the number of dimensions equals the number of histogram bins and is of the order of thousands. Thus, finding the closest exemplar (in k-NN classification) reduces to a linear search through all the exemplars, whilst for single class histograms the search is only linear in the number of classes.

Different histogram similarity functions have been compared. In the case of single-histogram class models, the KL divergence has been demonstrated to achieve higher accuracy than widely used alternatives such as L2 and χ^2 distances.

The pixel labelling results demonstrate that our class histograms can also be used to segment out objects. A natural next step is to combine such labellings with a contrast dependent prior MRF in the manner of [4] in order to obtain crisp segmentation boundaries. Alternatively the resulting pixelmaps can be used to initialise graph-cuts methods automatically rather than manually as in [15].

In future work we will compare performance of the single class histograms against other standard discriminative classifiers trained on the exemplars. For instance, an SVM could be trained on sliding-window histograms for pixel-wise classification or, as in the work of [17], weak classifiers can be built from histograms of visual words within sliding rectangular regions, and then combined into a discriminative classifier using boosting.

Bibliography

[1] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *Proc. ECCV*, pages 109–124, 2002. 82

[2] E. Borenstein and S. Ullman. Learning to segment. *ECCV*, 2004. 82

[3] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *Proc. ECCV*, 2006. 83, 84
[4] Y. Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *Proc. ICCV*, 2001. 82, 92
[5] A. Criminisi. Microsoft research cambridge object recognition image database. version 1.0, 2004. `http://research.microsoft.com/vision/cambridge/recognition/`. 82, 83, 86
[6] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Stat. Learning in CV, ECCV*, pages 1–22, 2004. 83
[7] I. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. J. Machine Learning Research, 2003. 86, 92
[8] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, Jun 2005. 83
[9] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. *Proc. CVPR*, 2004. 82
[10] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 43:177–196, 2001. 91
[11] M. P. Kumar, P. H. S. Torr, and A. Zisserman. OBJ CUT. In *Proc. CVPR*, 2005. 82
[12] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, volume 2, pages 264–271, 2003. 82
[13] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, Jun 2001. 84
[14] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *Proc. ICCV*, pages 883–890, 2005. 83
[15] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *Proc. ACM SIGGRAPH*, 23(3):309–314, 2004. 82, 92
[16] E. Sharon, A. Brandt, and R. Basri. Segmentation and boundary detection using multiscale intensity measurements. In *Proc. CVPR*, volume 1, pages 469–476. IEEE Computer Society, 2001. 82
[17] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint Appearance, Shape and context Modeling for Multi-Class Object Recognition and Segmentation. *Proc. ECCV*, 2006. 82, 92
[18] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proc. ICCV*, 2005. 83
[19] E. Spellman, B. C. Vemuri, and M. Rao. Using the KL-center for Efficient and Accurate Retrieval of Distributions Arising from Texture Images. In *Proc. CVPR*, 2005. 90
[20] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *Proc. CVPR*, volume 2, pages 691–698, Jun 2003. 82, 84, 90
[21] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *IJCV*, 62(1–2):61–81, Apr 2005. 84
[22] J. Winn, Criminisi, A., and T. Minka. Object Categorization by Learned Universal Visual Dictionary. *Proc. ICCV*, 2005. 83, 84, 86, 89
[23] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classifcation of texture and object categories: An in-depth study. Technical Report RR-5737, INRIA Rhône-Alpes, Nov 2005. 83