Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation

Alexei Yavlinsky[1], Edward Schofield[1,2] and Stefan Rüger[1]

[1]Department of Computing, South Kensington Campus
Imperial College London, London SW7 2AZ, UK
[2]Telecommunications Research Center Vienna
{alexei.yavlinsky, s.rueger}@imperial.ac.uk, schofield@ftw.at

Abstract. This paper describes a simple framework for automatically annotating images using non-parametric models of distributions of image features. We show that under this framework quite simple image properties such as global colour and texture distributions provide a strong basis for reliably annotating images. We report results on subsets of two photographic libraries, the Corel Photo Archive and the Getty Image Archive. We also show how the popular Earth Mover's Distance measure can be effectively incorporated within this framework.

1 Introduction

Automated image annotation has arisen as a recent alternative to querying databases of natural images directly by image content, with the benefit that the content of a desired image can often be specified most conveniently with keywords or natural language. Such a facility can be helpful for users wishing to search increasingly large collections of unlabelled images available on the web and elsewhere.

One of the first attempts at image annotation was reported by Mori *et al.* [1], who tiled images into grids of rectangular regions and applied a co-occurence model to words and low-level features of such tiled image regions. Since then researchers have looked at the problem in two different ways. The first way has been to use an image segmentation algorithm to divide images into a number of irregularly shaped 'blob' regions and to operate on these blobs. This has been pursued by several researchers recently. Duygulu *et al.* [2] created a discrete 'vocabulary' of clusters of such blobs across an image collection and applied a model, inspired by machine translation, to translate between the set of blobs comprising an image and annotation keywords. Jeon *et al.* [3] recast image annotation into a problem in cross-lingual information retrieval, applying a cross-media relevance model to perform image annotation and ranked retrieval, obtaining better retrieval performance than in the translation model of [2]. Lavrenko *et al.* [4] adapted the model of [3] to use continous probability density functions to describe the process of generating blob features, hoping to avoid the loss of

information related to quantization; they achieve substantially better retrieval performance on the same dataset. Metzler and Manmatha [5] likewise segmented training images, connecting them and their annotations in an inference network, whereby an unseen image is annotated by instantiating the network with its regions and propagating belief through the network to nodes representing the words. Feng *et al.* [6] replace blobs with rectangular blocks and model image keywords using a multiple Bernoulli distribution thus achieving better results than in [4] and [5]. Other relevant research is that of Blei and Jordan [7], proposing an extension of the Latent Dirichlet Allocation (LDA) model [8], which assumes that a mixture of latent factors are used to generate words and blob features; the authors then show how the model can be used to assign words to individual blobs.

A second way is a simpler scene-oriented approach. This was explored by Oliva and Torralba, who showed that images can be described with basic scene labels such as 'street', 'buildings' or 'highways', using a selection of relevant low-level global filters [9, 10]. They further showed how simple image statistics can be used to infer the presence and absence of objects in the scene [11].

This paper follows the second approach and explores the possibility of using 'global' features for automated image annotation, which are simpler still than those used in [9–11]. Our modelling framework is based on nonparametric density estimation, using the technique of 'kernel smoothing'. We investigate how well such an approach works with various global image features and show how the popular Earth Mover's Distance metric can be effectively incorporated within this framework. We evaluate our aproach on two image collections: the 5,000-image subset of the Corel Image Archive originally used by Duygulu *et al.* in [2], which makes our results comparable to several recent works on the subject [2–6], and our own set of about 7,500 images from the Getty Image Archive.

2 A simple framework for image annotation

Suppose a human annotator is prompted for a single annotation word for the image x, and that he chooses word w with probability $p(w|x)$. We wish to model this process. We use Bayes' Theorem to invert the conditional dependence as:

$$p(w|x) = \frac{f(x|w)p(w)}{f(x)}, \tag{1}$$

where we interpret $f(x)$ as the probability density of image x and $f(x|w)$ as the density of x conditional upon the assignment of annotation w.

We now wish to model $f(x|w)$ for each possible annotation word w by collecting a sample T_w of images with each label w as a training set. A critical factor in modelling the densities $f(x|w)$ will be choosing a representation x for the images. This paper considers two different representations: as a vector of real-valued image features $x = (x_1, \ldots, x_d)$, $x_i \in \mathbb{R}$; and as a 'signature' of image features, defined later in this section. In general we want a representation for which the densities are as separable as possible for different annotation classes

w, yet are dense enough for reliable inference from a small sample of images for each class.

One method of inference is to specify a parametric form *a priori* for the true distributions of image features for the annotation class w and then estimate the parameters using the methods of classical statistics. Another method is to encode all our knowledge about the true distribution as constraints on the model and choose the model subject to these constraints with maximum entropy (the 'flattest') or minimum relative entropy to some prior density. A third method is to adopt a nonparametric estimator of the true density that makes no prior assumptions about the true density.

The first method is less appropriate within this framework than the second two. In general, the distributions of image features will have shapes that are irregular, not resembling any simple parametric form. Instead we hope this irregularity will be helpful in characterizing and distinguishing the distributions under different word classes. This paper considers the third method, nonparametric estimation.

2.1 Nonparametric Density Estimation

The simplest nonparametric estimator of a distribution function is the empirical distribution function, but it is known that smoothing can improve efficiency for finite samples [12]. 'Kernel smoothing', first used by Parzen in [13], is a general formulation of this. Where x is a vector $(x_1, \ldots, x_d)$ of real-valued image features, we define the kernel estimate of $f_w(x) = f(x|w)$ as

$$\hat{f_w}(x) = \frac{1}{nC} \sum_{i=1}^{n} k\left(\frac{x - {x_w}^{(i)}}{h}\right), \tag{2}$$

where $x_w^{(1)}, \ldots, x_w^{(n)}$ is the sample of images with label w in the training set T_w, where k is a kernel function that we place over each point $x^{(i)}$, and where $C = \int k(t)dt$ so that $\hat{f}(x)$ integrates to 1 and is itself a probability density. We omit the subscripts w for the rest of this section to simplify the notation. Here the positive scalar h, called the bandwidth, reflects how wide a kernel is placed over each data point. Under some mild conditions [14], $\hat{f}$ converges to f in probability as $n \to \infty$.

We experiment with two types of kernels. The first is a d-dimensional Gaussian kernel

$$k_G(t; h) = \prod_{l=1}^{d} \frac{1}{\sqrt{2\pi}h_l} e^{-\frac{1}{2}\left(\frac{t_l}{h_l}\right)^2}, \tag{3}$$

where $t = x - x^{(i)}$, and where we set each bandwidth parameter h_l by scaling the sample standard deviation of feature l by the same constant λ.

Friedman *et al.* [15] point out that kernel smoothing may become less effective in high-dimensional spaces due to the problem known as the *curse of dimensionality*. They examine a projection pursuit method for reducing the effective dimensionality of a space by projecting it onto a single dimension in a way

that preserves its most salient characteristics. This is one way of sidestepping the problem, but this paper considers another way based on comparing image *signatures* under the Earth Mover's Distance (EMD) measure [16], which has found several applications in image retrieval [17].

A signature is a representation of clustered data defined as $s = \{(c_1, m_1), \ldots, (c_d, m_d)\}$, where, for a cluster i, c_i is the cluster's centroid and m_i is the number of points belonging to that cluster or its mass. Given two such signatures, EMD is defined as the minimum amount of work necessary to transform one signature into the other (see [16, 18] for details). One can create a signature for an image by grouping its colours into k clusters. Rubner *et al.* [16] report that using EMD on images represented with as few as 8 clusters of CIE*Lab* colour outperforms the traditional distance measures applied to high-dimensional colour features.

We use this advantageous property of EMD for density estimation by defining our second kernel as

$$k_E(s, s^{(i)}; h) = \frac{1}{h} e^{-\frac{d(s, s^{(i)})}{h}}, \tag{4}$$

where $d(s, s^{(i)})$ is the EMD between signatures s and $s^{(i)}$, and where h is the kernel bandwidth. The above kernel function exploits the fact that EMD is a true metric [16, 18] to yield a density centered on each signature $s^{(i)}$ in the signature space; this allows us to estimate probability density functions of image signatures for a particular word class. We shall refer to k_E as the *EMD kernel* throughout the rest of this paper.

Several methods have been studied for choosing the optimal bandwidth h for a given kernel and density estimation task. [19] and [20] give a good overview. For this paper we use the simple method of cross-validation, choosing the bandwidth that maximizes performance on a withheld data set. The precise performance measures are described in Section 4.

2.2 Bayesian Image Annotation

We now define the terms of the Bayesian model in Equation (1) for assigning the probabilities of a word w to an unseen image x. In the case where x is a d-dimensional feature vector, we model the probability density function $f(x|w)$ as

$$\hat{f}(x|w) = \frac{1}{|T_w|} \sum_{x^{(i)} \in T_w} k_G(x - x^{(i)}; h). \tag{5}$$

Similarly, for the signature case, we model f as

$$\hat{f}(s|w) = \frac{1}{|T_w|} \sum_{s^{(i)} \in T_w} k_E(s, s^{(i)}; h). \tag{6}$$

We then model the prior probability $p(w)$ of the word w as

$$\hat{p}(w) = \frac{|T_w|}{\sum_w |T_w|}, \tag{7}$$

where $|T_w|$ is the size of the training sample for the word w. Finally, we make the approximation $f(x) \approx \sum_w f(x|w)p(w)$ for simplicity.

Computational complexity. Using this model requires $O\left(\sum_w |T_w|\right)$ time to annotate a new image x. This is suitable for annotating images offline.

Relationship to other models. We make a note that our framework is different to the Continous Relevance Model (CRM) by Larvernko *et al.* [4], which also uses kernel smoothing for image features. CRM uses kernel density estimation to define a generative model for observing a set of blobs in a training image, which is then used as part of that image's relevance model. In our approach kernels are simply used for estimating densities of features conditional on each keyword.

3 Image Features

Global Features. We attempt to model image densities using two simple classes of global image features: the distribution of pixel colour in CIE space, and a subset of perceptual texture features proposed by Tamura [21] and adapted for image retrieval by Howarth and Rüger [22]. For each pixel in the image, we compute CIE*Lab* colour values and the coarseness, contrast and directionality texture properties obtained using a sliding window. This results in a 6-channel image representation. For each channel, the mean, second, third and fourth central moments are computed resulting in a 24-dimensional feature vector combining colour and texture. Additionally, this feature is split into two separate 12-dimensional colour and texture features, which are then evaluated independently.

Locally Sensitive Features. We designed a tiled image feature to investigate whether performance can be gained by looking at spatial configuration of colour and texture properties. Each image is split into $3 \times 3 = 9$ equal rectangular tiles; within each tile the mean and the second moment are computed for each of the above 6 channels. This results in a 108-dimensional feature vector. Note that this image segmentation is not context driven, i.e., we are not trying to detect the presence of any object boundaries, so one can still argue that this is a global feature.

Image Signatures. We used colour-only signatures for EMD computations, which were extracted for each image by applying simple k-means clustering to pixels in CIE*Lab* space and setting k to 16.

4 Performance Evaluation

4.1 Image and Caption Data

The Corel Dataset. One of the datasets we use is the one by Duygulu *et al.* [2]. The dataset consists of 5,000 images from Corel Stock Photo library. Each image was also assigned 1–5 keywords from a vocabulary of 371 words. To make our results comparable to those recently published in [2–5] we use the same training and test dataset partition as in [2], where there are 4,500 training images and 500 test images. To optimise the kernel bandwidth parameters for different features we randomly divide the training set into 3,800 training images and 700 images on which different bandwidth settings are evaluated.

The Getty Dataset. In the past the Corel photo collection has been criticized that for being an easy collection from an image retrieval point of view. For instance, Müller *et al.* observed that image retrieval performance can be substantially improved if the right image subset is selected for evaluation [23]. We attempted to build a more realistic dataset for our experiments by downloading 7,560 medium-resolution thumbnails of photographs from the Getty Image Archive website[1], together with the annotations assigned by the Getty staff to catalogue those pictures. The selection of photographs was obtained by submitting the following query to the Getty website: "`photography, image, not composite, not enhancement, not 'studio setting', not people`", with the additional search option to exclude illustrations. With this query we sought to obtain a random selection of photos, which excludes any non-photographic content, any digitally composed or enhanced photos and any photos taken in unrealistic studio settings. The constraint to exclude people is imposed to reduce the semantic ambiguity of annotations. The resulting dataset contains pictures from a number of different photo vendors, which – we hope – reduces the chance of unrealistic correlations between keywords and image contents.

Keywords for Getty images come in three different flavours: subjects (e.g. '`tiger`'), concepts (e.g. '`emptyness`') and styles (e.g. '`panoramic photograph`'). We created our vocabulary using subject keywords only, of which there were over 6,000. We restricted the range of keywords to those, which occur in fewer than 10% of the images and those, which occur more than 50 times. We then pruned references to specific locations (e.g. '`europe`', '`japan`'), descriptions of dominant image colour, verbs and abstract nouns (e.g. '`flying`', '`close-up`'). This resulted in a final list of 184 words ranging from specific objects (e.g. '`insect`', '`church`') to more general object categories (e.g. '`building structure`') and scene properties (e.g. '`urban scene`', '`autumn`', '`illuminated`').

We randomly split the dataset into 5,000 training and 2,560 test images. The list of Getty image IDs used to make up the dataset, the vocabulary and the annotations can be downloaded[2].

[1] http://creative.gettyimages.com

[2] http://mmir.doc.ic.ac.uk/www-pub/civr2005

4.2 Image Annotation

The first task we evaluate is automated image annotation. Our approach is the same as in [3–5], where top 5 most probable words are assinged to each unseen test image after which mean word precision and recall are found. For each feature we found the kernel scaling factor λ (and the bandwidth h for the EMD kernel) that maximized precision and recall figures on the withheld evaluation set. We compare our results on the Corel dataset with the Continuous Relevance Model (CRM) [4], the Inference Network Model (InfNet) [5] and the Multiple Bernoulli Relevance Model (MBRM) [6]. Note that in this and the following sections we do not set out to establish the relative merits of these models as compared to ours. Rather, we use the published results to investigate whether comparable performance can be achieved in principle using our approach.

	# words w/ recall > 0	Precision	Recall
Random	15	0.01	0.02
Tamura	50	0.04	0.05
CIE	96	0.13	0.16
TamuraCIE	105	0.15	0.18
EMD	104	0.16	0.19
CRM	107	0.16	0.19
TamuraCIE-3×3	114	0.18	0.21
InfNet	112	0.17	0.24
MBRM	122	0.24	0.25

Table 1. Precision and recall results on the Corel dataset

As the table shows, the combined colour/texture feature (TamuraCIE) performs comparably to CRM and the tiled colour/texture feature (TamuraCIE-3×3) does somewhat better and gets close to the Inference Network performance. This shows that retaining some structural information about the scene is helpful and that kernel smoothing works well for this feature despite its high dimensionality. The EMD kernel does as well as CRM, which is particularly encouraging as it only uses global colour information; this confirms our initial hypothesis which led to the design of this kernel. All reported figures are significantly better than what would be obtained if the top 5 captions were assigned by chance.

4.3 Ranked Retrieval

We use the same experimental setup as in [3] to evaluate ranked retrieval performance. For the Corel dataset all 1– 2– and 3-word queries were generated that would yield at least 2 relevant images in the test set. For the Getty dataset we required at least 6 relevant images for any given query (to cut down the greater number of queries due to the larger size of the test set), and generated all possible 1–4 word queries under this constraint. Given an m-word query $Q = \{q_1, q_2, \ldots, q_m\}$ the retrieval score for an image x is defined as:

$$p(q_1, q_2, \ldots, q_m|x) = \prod_{i=1}^{m} p(q_i|x) \tag{8}$$

Query results are then evaluated using the standard average precision metric. As before, we optimised the kernel bandwidths for this task on the withheld set. Results on the Corel dataset, presented in Table 2, show that TamuraCIE has a reasonable performance compared to CRM and that TamuraCIE-3×3 outperforms both CRM and the Inference Network. The colour-only EMD kernel performs slightly better than CRM and rivals the performance of the Inference Network. All reported figures are significantly above random chance. The features have a slightly different behaviour on the Getty dataset (Table 3), where the EMD kernel comes top for queries longer than 1 word. The results show that – despite Getty being an undoubtedly harder dataset – good retrieval performance can be achieved using our framework in tandem with the simple features we have chosen; they also highlight the robust performance of the EMD kernel.

Query Length	1 word	2 words	3 words
Number of Queries	179	386	178
Relevant Images	1675	1647	542
Random	0.0293	0.0198	0.0228
Tamura	0.0969	0.0871	0.1013
CIE	0.1963	0.1979	0.2325
TamuraCIE	0.2450	0.2450	0.2761
CRM	0.2353	0.2534	0.3152
EMD	0.2683	0.2734	0.3250
InfNet	0.2633	0.2649	0.3288
TamuraCIE-3×3	0.2861	0.2922	0.3301
MBRM	0.3000	—	—

Table 2. Mean average precision for ranked retrieval on the Corel dataset

Query Length	1 word	2 words	3 words	4 words
Number of Queries	184	967	655	297
Relevant Images	9255	10722	4970	1950
Random	0.0233	0.0070	0.0063	0.0070
Tamura	0.0473	0.0225	0.0257	0.0276
CIE	0.0624	0.0411	0.0496	0.0520
TamuraCIE	0.0788	0.0613	0.0891	0.1109
TamuraCIE-3×3	0.0921	0.0907	0.1670	0.2412
EMD	0.0827	0.0917	0.1803	0.2759

Table 3. Mean average precision for ranked retrieval on the Getty dataset

4.4 Kernel bandwidth optimisation

At this point it is worth mentioning the motivation behind using two different bandwidth setings for the ranked retrieval and image annotation tasks. Figure 1 shows how performance is affected by the choice of the kernel scaling factor for the TamuraCIE-3×3 feature on the withheld set. One can see that wider kernels seem to be more suitable for ranked retrieval, whereas narrower kernels appear to be more favourable for automated annotation. This can be explained by the different nature of the two tasks. In the first task we are interested in ranking images as accurately as possible given a particular keyword and therefore require individual keyword densities to be robust to noise in the high-dimensional feature space. Increasing the kernel bandwidth achieves this goal by making the estimated keyword densities smoother. However, it also has the effect of making

them less separable. This is detrimental for the second task, in which we are interested in obtaining the most accurate ranking of keywords given an image. This necessitates the use of different bandwidth values for the two tasks to achieve optimal performance in both.

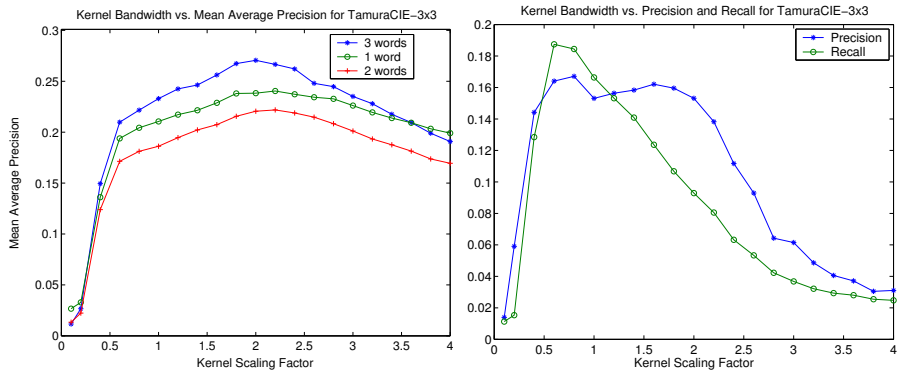

Fig. 1. Kernel bandwidth effects on the withheld set from Corel

5 Conclusions and Future Work

We have presented a simple framework for automated image annotation based on nonparametric density estimation. We have shown that under this framework very simple global image properties can yield reasonable annotation accuracies. A surprising finding is that using merely colour information can achieve 'state of the art' performance for the Corel dataset and good performance for the more difficult Getty collection. We attribute this result to the robustness of the EMD kernel and note that this kernel may be useful when one intends to use other sparse image features within this framework. Our experiments have shown that global colour is a strong basis for modelling keyword densities. This may be due to the general homogeneity of photographic collections. We look forward on this basis to exploring image features outside the colour domain.

Acknowledgements. We would like to thank R Manmatha and David Forsyth for helpful comments and discussions of the subject. The first author is partially funded by the Overseas Research Scholarship award.

References

1. Y Mori, H Takahashi, and R Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of the International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.
2. P Duygulu, K Barnard, N de Fretias, and D Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision*, pages 97–112, 2002.

3. J Jeon, V Lavrenko, and R Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the ACM SIGIR Conference on Research and Development in Infrmation Retrieval*, pages 119–126, 2003.
4. V Lavrenko, R Manmatha, and J Jeon. A model for learning the semantics of pictures. In *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems NIPS*, 2003.
5. D Metzler and R Manmatha. An inference network approach to image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 42–50, 2004.
6. S Feng, R Manmatha, and V Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1002–1009, 2004.
7. D Blei and M Jordan. Modeling annotated data. In *Proceedings of the ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134, 2003.
8. D Blei, A Ng, and M Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
9. A Oliva and A Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.
10. A Oliva and A Torralba. Scene-centered representation from spatial envelope descriptors. In *Proceedings of Biologically Motivated Computer Vision*, 2002.
11. A Torralba and A Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14:391–412, 2003.
12. R Reiss. Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, 8:116–119, 1981.
13. E Parzen. On estimation of a probability density and mode. *Annals of Mathematical Statistics*, 35:1065–1076, 1962.
14. W Härdle. *Applied Nonparametric Regression*. Cambridge University Press, 1992.
15. J Friedman, W Stuetzle, and A Schroeder. Projection pursuit density estimation. *Journal of the American Statistical Association*, 79:599–608, 1984.
16. Y Rubner. The earth-mover's distance as a metric for image retrieval. Technical Report STAN-CS-TN-98-86, Stanford University, 1998.
17. Y Rubner, J Puzicha, C Tomasi, and J Buhmann. Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84:25–43, 2001.
18. E Levina and P Bickel. The earth mover's distance is the Mallows distance: Some insights from statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 251–256, 2001.
19. M Jones, J Marron, and S Sheather. A brief survey of bandwidth selection for density estimation. *Journal of American Statistics Association*, 91:401–407, 1996.
20. R Loader. Bandwidth selection: classical or plug-in? *The Annals of Statistics*, 27(2):415–438, 1999.
21. H Tamura. Texture features corresponding to visual perception. *IEEE Transactions. Systems, Man and Cybernetics*, 8(6):460–473, 1978.
22. P Howarth and S Rüger. Evaluation of texture features for content-based image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 326–334, 2004.
23. H Müller, S Marchand-Maillet, and T Pun. The truth about Corel - evaluation in image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, pages 38–49, 2002.