Optimal Tag Sets for Automatic Image Annotation

Sean Moran
http://homepages.inf.ed.ac.uk/s0894398/
Victor Lavrenko
http://homepages.inf.ed.ac.uk/vlavrenk/

School of Informatics
The University of Edinburgh
10 Crichton St
Edinburgh EH8 9AB

In this paper we introduce the Beam Search CRM (BS-CRM) model. This model implements two novel improvements to the basic CRM [2]. First, we argue that using a *Minkowski kernel* allows us to capture the covariance of visual features more effectively than the standard Gaussian kernel. Second, we advocate a procedure that selects the *most informative* subset of tags as the image annotation. Our procedure captures the mutual dependence within a set of tags, and naturally prevents noisy tags from being assigned during the search procedure.

In automatic image annotation the basic objective is to find the set of tags $\mathbf{w} = \{w_1 \dots w_k\}$ that serves as the best annotation for the test image represented with a set of feature vectors $\mathbf{f} = \{\vec{f}_1 \dots \vec{f}_m\}$. The traditional approach used by [2] and many subsequent publications [3] [5] [4] involves estimating the marginal probability distribution over individual tags $P(w|\mathbf{f})$ and annotating the image with top-ranked tags from that distribution. This approach however does not take into consideration any correlation between the tags: the top-ranked tags could be incohesive and contradictory, e.g. {*tropical, blizzard, supernova*}.

Beam Search: To address both of the above issues, we propose to annotate images with the *most informative* subset of tags. We define the amount of information $I(\mathbf{w})$ present in a set of tags $\mathbf{w}$ as the expected excess number of bits required to encode this set with the background model: $I(\mathbf{w}) = P(\mathbf{w}|\mathbf{f}) \cdot \log \frac{P(\mathbf{w}|\mathbf{f})}{P_0(\mathbf{w})}$.

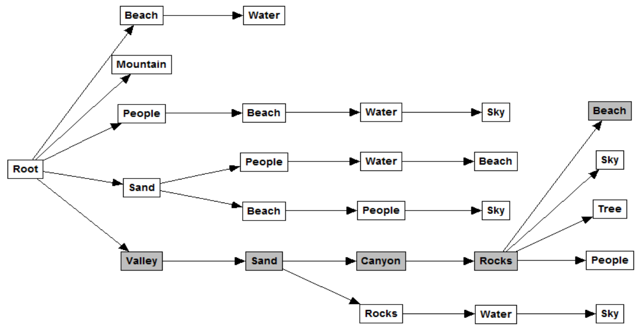

Figure 1: Example search tree for the BS-CRM algorithm on the Corel dataset. The first level of the tree corresponds to the annotation of the basic CRM. The BS-CRM refines this set of annotation keywords by considering multiple hypotheses for the most informative set of tags at each level of the tree. Only the most informative tags are added to the set of B hypotheses at each iteration. Less promising nodes are pruned, thereby constraining the search space.

Here $P(\mathbf{w}|\mathbf{f})$ is a model of dependence between tags and image features and $P_0(\mathbf{w})$ is a background model that treats every tag as an isolated event, independent of all other tags and image features: $P_0(\mathbf{w}) = p_{w_1} \times p_{w_2} \times \dots \times p_{w_k}$. $I(\mathbf{w})$ can be interpreted as the contribution of tagset $\mathbf{w}$ to the Kullback-Leibler divergence between the relevance model P and the background model P_0. We propose to annotate the image $\mathbf{f}$ with a set of tags $\mathbf{w}$ that has the largest information content $I(\mathbf{w})$. Since this procedure involves optimisation over the universe of all possible tag-sets, we resort to an efficient approximation procedure based on the *beam search* algorithm as illustrated in Figure 1.

Minkowski Kernel: In addition to this we investigate replacing the Gaussian kernel with a generalised exponential kernel based on the Minkowski p-norm. We will argue that the proposed kernel is more sensitive to subtle changes in the visual appearance of an image region and better capable of modeling conjunctions of features than the standard Gaussian kernel. We define a *Minkowski kernel* based density estimate as follows:

$$P(\vec{f}_i|J) = \frac{1}{n}\sum_{j=1}^{n} c_{\mathrm{p}} exp\left\{\frac{-|\vec{f}_i - \vec{f}_j|^{\mathrm{p}}}{\beta}\right\} \tag{1}$$

Here $|\vec{f}_i - \vec{f}_j|^p = \sum_{d=1}^{k} |f_{i,d} - f_{j,d}|^p$ is a generalisation of the Euclidean norm, and the summation goes over the dimensions d of the feature vectors. p is a positive free parameter that is optimized on a held-out validation set. c_{p} is a constant that ensures that the kernel integrates to one.

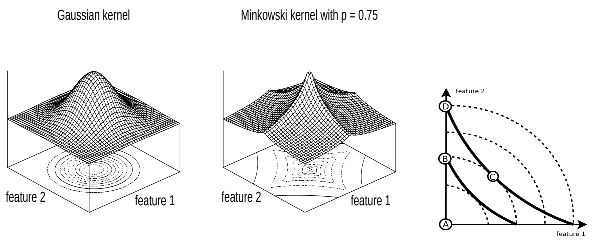

Figure 2: **Left:** density functions and equidistant contours for the Gaussian kernel. **Middle:** the proposed Minkowski kernel. **Right:** the Minkowski kernel is particularly sensitive when multiple feature values change at the same time (point C), whereas the Gaussian is more sensitive to large variations in any one feature (point D).

Figure 2 highlights the difference between the Gaussian kernel and the proposed Minkowski kernel. The Gaussian density on the left is convex around the mean, which makes it insensitive to small differences between the training and testing feature regions. The Minkowski kernel in the middle is concave (for p<2), allowing it to sense subtle differences in feature values in a way that mimics the operation of the human visual system [1]. Perhaps more importantly, the two kernel functions greatly differ in how they treat simultaneous deviation of multiple feature values from the mean: the Gaussian kernel has a spherical contour profile, so a large variation in the value of single feature 2 has a much greater effect than simultaneous variation of feature 1 and feature 2. The Minkowski kernel (for p<1) behaves very differently: a simultaneous small change in several features is as important as large variations in a single feature.

Experiments: We present a comprehensive evaluation of the proposed model in relation to the basic Gaussian kernel based CRM model [2] and models recently proposed in the literature that specifically attempt to capture keyword correlation. The experiments show that the beam-based CRM model with a Minkowski kernel density significantly outperforms the same model based on Gaussian kernels, producing a 42% increase in recall, a 38% increase in precision on the standard Corel dataset. We note that BS-CRM model on the Corel dataset fares well in comparison with results published by Zhou *et al.* [5], Liu *et al.* [3] and Wang *et al.* [4] showing improvements with respect to all accuracy measures. This allows us to confidently conclude that the BS-CRM model exhibits superior performance in the context of the relevance-modelling framework of [2].

[1] P. Howarth and S. M. Rüger. Fractional distance measures for content-based image retrieval. In *ECIR*, pages 447–456, 2005.

[2] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*. MIT Press, 2003.

[3] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recogn.*, 42:218–228, February 2009.

[4] B. Wang, Z. W. Li, N. Yu, and M. Li. Image annotation in a progressive way. In *proc. ICME*, pages 811–814, 2007.

[5] X. Zhou, M. Wang, Q. Zhang, J. Zhang, and B. Shi. Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 25–32, 2007.