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

Abstract

In this paper we introduce a new form of the Continuous Relevance Model (the BS-CRM) that captures the correlation between tags in a formal and consistent manner. We apply a beam search algorithm to find a near optimal set of mutually correlated tags for an image in a time that is linear in the depth of the search tree. We conduct an examination of the model performance under different kernels for the representation of the image feature distributions and suggest a method of adapting the kernel to the dataset. BS-CRM with a Minkowski kernel is found to significantly increase recall by 42% and precision by 38% over the original CRM model and outperforms more recent baselines on the standard Corel 5k dataset.

1 Introduction

Automatic image annotation is the process of associating relevant tags to images that define the high-level semantic content of the images. Generally speaking image annotation is a form of supervised classification of pictorial data. Each image class contains images, which are semantically similar and thus have at least one annotation in common. Furthermore an image usually can be provided with more than one annotation and hence most images can belong to multiple classes.

The typical approach taken to image tagging in the literature is to train a model on a small set of manually annotated images which can be used to assign an image to one or more classes. The training set provides a unique mapping between a textual annotation and the described semantic entities within the image. Given a novel image, the annotation model compares the visual words with an unknown image, annotating an image with a textual word in the case where the novel image contains the corresponding visual word. A commonality between most approaches to automatic image annotation is that they tend to treat each annotated word as an independent class therefore creating a different image classification model for every word in the keyword vocabulary. So, for example, as an annotation {*jungle, trees*} is just as plausible as {*jungle, snow*}.

In this paper we propose a formal and consistent method of adapting the CRM [8] to take account of the dependencies between annotation tags. This novel principled approach to finding optimal tag sets is able to increase the annotation accuracy in the event where the extracted image features are not of adequate quality to distinguish between annotation tags with sufficiently high probability. If we predict a set of tags together, rather than each tag independently there is the possibility that some tags in the set will boost the probability of correct, but otherwise low probability tags whilst suppressing the probability of irrelevant but higher probability tags. For example, consider the annotation tags "sky" and "ocean". As both refer to concepts that are some shade of the colour blue, it is difficult to differentiate between either based on extracted colour features. However, if we consider "airplane" and "bird" as part of the annotation set then we can differentiate more easily between these two concepts given that we expect "airplane" and "bird" to co-occur more frequently with respect to "sky" than to "ocean".

The key issue in predicting sets of tags in this manner is the exponential complexity that arises in finding the best (in terms of highest probability) set of tags for an image. For modest vocabulary sizes, a simple exhaustive search strategy over sets of tags is impossible. In this paper we take the novel approach of using a customized beam search algorithm in combination with the amended CRM model to efficiently search over sets of tags in a *quasi-greedy* fashion, only adding those tags that have the best chance of increasing the probability of the entire set of tags[1]. This amendment has the effect of reducing the exponential complexity to linear in the depth of the search tree, whilst finding a near-optimal set of tags. We refer to the novel beam search amalgamated algorithm as the *beam search CRM* or BS-CRM model.

2 Related Work

Despite still being in its relative infancy, the automatic image annotation field is extremely large and there exist many different techniques in the literature designed to tackle the problem.

The pioneering paper by Mori *et al*. described how candidate images were divided into a regular grid and a co-occurrence model applied to represent the co-occurrence of words with the image regions [11]. Duygulu *et al*. [5] utilize the statistical machine translation model of Brown *et al*. [1] and apply the EM algorithm to learn a maximum likelihood association of words to image regions using a bi-lingual corpus. More recently, Carneiro *et al*. [2] proposed a supervised multi-class labelling (SML) model which estimates the class density based on image-level and class-level Gaussian mixtures. Makadia et al. [10] introduced an approach consisting of colour and texture based features and a simple technique to combine distance computations on these features to create a nearest neighbour classifier for image annotation.

The most pertinent class of image tagging model for this paper is the relevance models [8][6] which were originally developed for information retrieval but have found great success in the image annotation field. A key idea behind this model is to find images that are most similar to the test image and then use their shared tags for annotation. The Continuous Relevance Model (CRM) [8] works with continuous image features directly using non-parametric kernel density estimators therefore avoiding the error prone k-means vector quantization step. This model was improved upon by the Multiple Bernoulli Relevance Model (MBRM) of Feng et al. [6] who demonstrated that a multiple Bernoulli distribution for the word-image probability distribution coupled with image features collected over a regular grid was able to provide a substantial increase in performance.

All of the aforementioned approaches predict image tags independently. Recently, researchers have turned to the question of how best to capture correlations between tags to

[1] Beam search has been applied with notable results to the decoding problem in the field of statistical machine translation [12].

enhance the performance of the annotation models. Given that we are essentially aiming to select the "best" set of tags that are most correlated with each other for a particular image, the question naturally arises on how one can find this best tag set out of the word vocabulary given the combinatorial explosion of possible sets of tags even for modest size vocabularies.

Liu *et al.* [9] modelled the relationship among the annotation words using word-based graph learning. Zhou *et al.* [14] overcome the combinatorial explosion by proposing a heuristic greedy iterative algorithm to estimate the keyword subset for a particular image which is found to significantly improve the performance of a state of the art image annotation algorithm. Nevertheless the objective function used by the authors has some notable drawbacks in the fact that it is both data-inconsistent[2] and relies on the use of a heuristic which is incompatible with the relevance model probabilistic framework.

Wang *et al.* [13] improve on this approach in their progressive image annotation model by applying a more powerful objective function in the form of the CRM to capture keyword correlation of tags[3]. The suggested greedy method involves adding successive tags to the set that have the largest joint probability with the tags already in the annotation, with the number of tags added to the set in this manner denoted as the progressive annotation length (hereby referred to as PAL). This method leads to the largest gain in performance for the first two tags and hurts performance for longer annotations.

3 Background

The Continuous Relevance Model CRM [8] is a statistical model for automatically assigning tags to unlabelled images. The CRM estimates the joint probability distribution of a set of tags $\mathbf{w} = \{w_1 \dots w_k\}$ together with an image $\mathbf{f}$ represented as a feature vectors $\mathbf{f} = \{\vec{f}_1 \dots \vec{f}_m\}$. The modelling of the joint distribution $P(\mathbf{w}, \mathbf{f})$ of tags and image regions in this manner is key to the model and gives it the ability to annotate images by searching for those tags $\mathbf{w}$ that maximize the conditional probability: $P(\mathbf{w}|\mathbf{f}) = P(\mathbf{w}, \mathbf{f})/P(\mathbf{f})$.

The probability $P(\mathbf{w}, \mathbf{f})$ is computed as joint expectation over the space of distributions $P(.|J)$ defined by annotated images J in the training set T:

$$P(\mathbf{w}, \mathbf{f}) = \sum_{J \in T} P(J) \prod_{i=1}^{k} P(w_i|J) \prod_{j=1}^{m} P(\vec{f}_j|J) \tag{1}$$

The annotation component $P(w_i|J)$ is modelled using a Dirichlet prior:

$$P(w_i|J) = \frac{\mu p_v + N_{v,J}}{\mu + \sum_{v'} N_{v',J}} \tag{2}$$

Here $N_{v,J}$ is the number of times the keyword v appears in the annotation of training image J and p_v is the relative frequency that the word v appears in the training set. μ is a smoothing parameter selected based on a held out validation set. In the original formulation

[2] As Zhou *et al.* assume that the word set is independent of the image and that the word set and image are conditionally independent given the current word, then it follows that either the current word is independent of the image or that the current word is independent of the word set. Both of these points contradict the premise of Zhou *et al.* [14]. See [3] for a complete description of data-inconsistency.

[3] In their paper the authors implement the CRM model with L_1 rather than L_2 distance.

of the CRM [8], the feature component $P(\vec{f}_j|J)$ uses a non-parametric density estimate based on Gaussian kernels:

$$P(\vec{f}_i|J) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\sqrt{2^k\pi^k\beta}}exp\left\{\frac{-||\vec{f}_i-\vec{f}_j||^2}{\beta}\right\} \tag{3}$$

Here the summation goes over the n regions in the training image J and $\vec{f}_j$ represents the feature vector of the j'th region. k denotes the dimensionality of the image feature vectors, and $||\vec{f}_i-\vec{f}_j||$ represents the Euclidean distance. β is the kernel bandwidth parameter of the model that is optimized on a held out validation set.

4 The BS-CRM Model

We introduce two improvements to the basic CRM [8]. 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.

4.1 Capturing Feature Covariance with Minkowski Kernels

We investigate replacing the Gaussian kernel in equation (3) with a generalised exponential kernel based on the Minkowski $\mathtt{p}$-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 modelling conjunctions of features than the standard Gaussian kernel. We define a *Minkowski kernel* based density estimate as follows:

$$P(\vec{f}_i|J) = \frac{1}{n}\sum_{j=1}^{n}c_{\mathtt{p}}exp\left\{\frac{-|\vec{f}_i-\vec{f}_j|^{\mathtt{p}}}{\beta}\right\} \tag{4}$$

Here $|\vec{f}_i-\vec{f}_j|^{\mathtt{p}} = \sum_{d=1}^{k}|f_{i,d}-f_{j,d}|^{\mathtt{p}}$ is a generalisation of the Euclidean norm, and the summation goes over the dimensions d of the feature vectors. $\mathtt{p}$ is a positive free parameter that is optimized on a held-out validation set. $c_{\mathtt{p}}$ is a constant that ensures that the kernel integrates to one, the exact value of which is unimportant given that conditional probabilities are computed as part of the BS-CRM model.

Figure 1 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 $\mathtt{p}<2$), allowing it to sense subtle differences in feature values in a way that mimics the operation of the human visual system [7]. Perhaps more importantly, the two kernel functions greatly differ in how they treat simultaneous deviation of multiple feature values from the mean. The right part of Figure 1 shows equidistant contours for the Gaussian kernel (dashed lines) and the Minkowski kernel (bold lines). The coordinates reflect variation in feature values 1 and 2 (e.g. colour and texture) between the training image $\mathtt{A}$ and three testing images $\mathtt{B}$, $\mathtt{C}$, $\mathtt{D}$. The Gaussian kernel has a spherical contour profile, so a large variation in the value of single feature 2 has a much greater effect

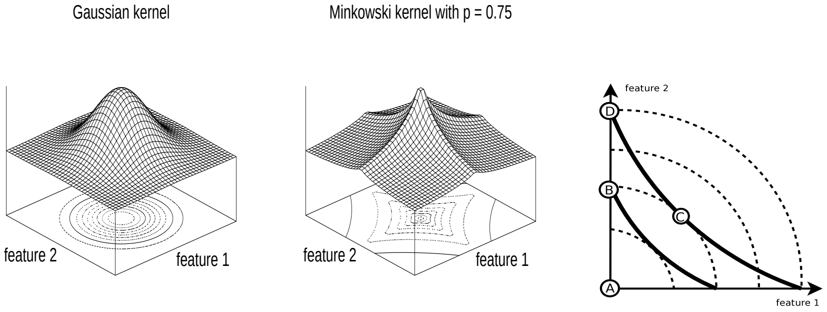


Figure 1: **Left:** density functions and equidistant contours for the Gaussian kernel. **Middle:** the proposed Minkowski kernel. **Right:** the Minkowski kernel is particularly sensitive when multiple feature values change at the same time (point `C`), whereas the Gaussian is more sensitive to large variations in any one feature (point `D`).

than simultaneous variation of feature 1 and feature 2. Under the Gaussian kernel, points `B` and `C` are equidistant from the mean `A`, whereas point `D` is much further. The Minkowski kernel (for `p`<1) behaves very differently: points `C` and `D` are equidistant and much further than `B`, so a simultaneous small change in several features is as important as large variations in a single feature. In other words, the Gaussian kernel can be thought of as mimicking a logical `OR` of variations in feature 1 and feature 2, whereas the Minkowski kernel is closer to a logical `AND`.

4.2 Capturing Keyword Correlation through Beam Search

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 [8] and many subsequent publications 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*}. In addition, selecting the most probable tags may lead to very frequent tags being included in the annotation.

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: $I(\mathbf{w}) = P(\mathbf{w}|\mathbf{f}) \cdot \log \frac{P(\mathbf{w}|\mathbf{f})}{P_0(\mathbf{w})}$

Here $P(\mathbf{w}|\mathbf{f})$ is a model of dependence between tags and image features defined by Equation 1 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 tag-set $\mathbf{w}$ 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.

Previous authors [13] [14] add tags to an existing set of tags using a formula that captures *pairwise* correlations between tags. They do so in a 'greedy' manner only adding the tag that leads to the maximum probability of having all of the tags together in the set. In this approach, we are not guaranteed that the next tag chosen, even if it does contribute the maximum gain to the selected subset, does not cause the probability mass of future tags to be skewed such that one or more relevant tags down the line are therefore missed. To overcome this issue, in this paper, we depart from previous approaches by integrating beam search into our model which allows us not only to consider the effect of adding the highest probability tag to the existing tag set but, crucially, the effect that is also brought about by the introduction of lower probability tags as well.

Specifically, in each layer of a breadth-first search graph, the BS-CRM expands only the B most promising nodes (tags), and discards the rest, where the integer B is called the *beam width*. The most promising nodes to branch are measured using $I(\mathbf{w})$. At any point during execution the algorithm will have B tag sets under consideration. This is an implementation of *beam search* [4] for image annotation. By bounding the width, the complexity of the search becomes linear in the depth of the search instead of exponential; the time and memory complexity of beam search is BD, where D is the depth of the search[4]. This flavour of beam search ensures that sub-optimal paths are quickly rejected during the search. At termination, the keyword set with the maximum probability is selected as the annotation of the image.

For the models of Wang *et al*. and Zhou *et al*., the width of the beam search is effectively one as they only keep one hypothesis (set of tags) at every step in the search tree. The BS-CRM model provides a principled generalisation utilizing beam search to maintain several hypotheses at each level in the search tree.

5 Experimental Results

5.1 Datasets

To provide a meaningful comparison with previously-reported results, we use, without any modification, the dataset provided by Duygulu *et al*.[5] [5]. This allows us to compare the performance of the model in a strictly controlled manner. The dataset consists of 5000 images. Each image contains an annotation of 1-5 tags. Overall there are 371 tags in the vocabulary.

In addition we test our model on the University of Washington dataset [6]. The dataset consists of 1109 images, with each image being annotation with 1-13 keywords. We manually removed function words and morphological variants (e.g. “runner” ⇒ “run”) to form a vocabulary of 320 words.

To demonstrate the scalability of our technique we also test the BS-CRM model on the IAPR TC-12 dataset. IAPR TC-12 is a collection of 19622 images of natural scenes, each of which is annotated with between 1-23 words. We use, without modification, the identical dictionary of 291 words and the set of 17662 training and 1960 testing images as used in [10].

For the UW and IAPR TC-12 datasets we extracted 41 dimensional features from each image across a regular grid consisting of normalized x,y grid centre coordinates, RGB, LAB

[4] Relating this to Image Annotation, in terms of the vocabulary size V and beam width B the complexity of the greedy beam search algorithm is DVB, whereas the non-greedy search is of complexity V^D. A substantial improvement.

[5] http://kobus.ca/research/data/eccv_2002/index.html

[6] http://www.cs.washington.edu/research/imagedatabase/

and HSV[7] average, standard deviation and skewness as well as the mean oriented energy in 30 degree increments.

5.2 Parameter Optimization

The Corel dataset is divided into 3 parts; namely 4000 training set images, 500 validation set images and 500 images in the test set. The validation set is used to find system parameters. After fixing the parameters, we merged the 4000 training set and 500 validation set images to make a new training set. This corresponds to the training set of 4500 images and the test set of 500 images used by Duygulu *et al.* [5].

We follow an identical procedure for UW and IAPR TC-12. UW was split into 609 training images, 300 validation images and 200 images in the test set. The IAPR TC-12 dataset was divided into 16000 training set images, 1662 validation images and 1,960 test set images. After parameter tuning, the validation set images are merged with the training set images giving 909 training images for UW and 17662 training images for IAPR TC-12.

To tune the parameters of the model we used the following optimization procedure: firstly, we perform a grid based search jointly over the β and μ parameters for the original CRM model. Holding β constant at the optimized value, we then optimize the BS-CRM model with respect to μ for varying widths. Holding β and μ constant we optimize the parameter p for the Minkowski kernel based density (Equation 4) on the held out validation set.

5.3 Results: Automatic Image Annotation

In this section we evaluate the performance of our model on the task of automatic image annotation on the Corel, UW and IAPR TC-12 testing datasets. We are given an un-annotated image I and are asked to automatically produce an annotation w_{auto}. The automatic annotation is then compared to the held-out human annotation w_I . We follow the experimental methodology used by [5]. Given a set of image regions r_I we use the BS-CRM algorithm to determine the set of words from that distribution and call them the automatic annotation of the image in question.

Then, following [5], we compute annotation recall and precision for every word in the testing set. Recall is the number of images correctly annotated with a given word, divided by the number of images that have that word in the human annotation. Precision is the number of correctly annotated images divided by the total number of images annotated with that particular word (correctly or not). Recall and precision values are averaged over the set of testing words.

For Corel we report the results on the complete set of all 260 words that occur in the testing set, for UW the set of 158 words and for IAPR TC-12 the set of 291 words. In addition we include the number of words with recall greater than zero: this metric seeks to measure the ability of the system to label images with rare keywords which are hard to annotate due to the small number of positive instances in the training set.

[7] HSV and LAB complement RGB by capturing different aspects of the image. HSV captures the amount of light illuminating a colour in the value channel and LAB captures human perception of brightness in the luminance channel.

5.3.1 Discussion of Results

The primary hypothesis of this paper is that using beam search to select the most informative set of tags will lead to more accurate image annotations. The secondary hypothesis is that determining an optimal kernel using the data itself, owing to its different geometry over the feature space, will outperform Gaussian kernels. In this section we discuss a set of experiments we carried out to test these hypotheses. In all experiments we use the models to annotate each image with 5 tags (for Corel and IAPR TC-12) and 7 tags (for UW), and then measure accuracy with respect to the human annotations.

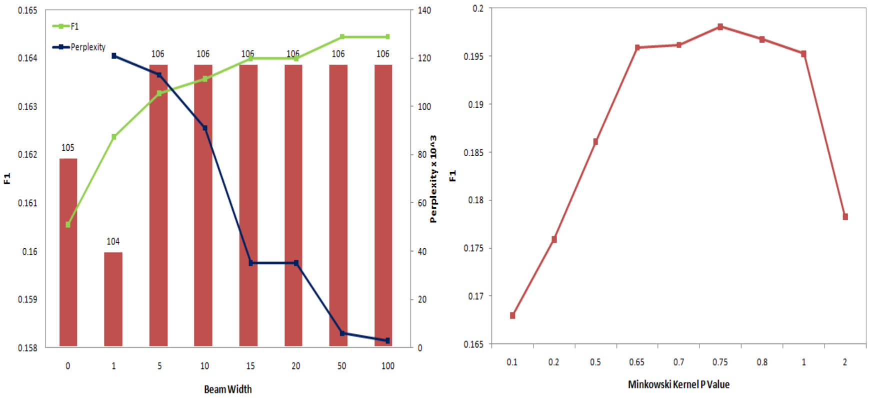


Figure 2: **Left:** The effect of beam search with the p=2 (Gaussian) kernel on the Corel *validation* set. F1 (green line) peaks at a beam width of 50 and remains constant suggesting no more benefit to wider beams. The perplexity (blue line) continues to decrease suggesting that the more probable sets are not necessarily the most accurate. **Right:** Finding the optimal value of the kernel parameter p on the Corel validation set. p peaks at a value of 0.75. For the IAPR TC-12 and UW datasets we, in both cases, find an optimal p of 0.70.

Beam search: In Figure 2 we compare the accuracy of selecting tags independently (CRM) against selecting them as a set (**Beam 1...100**) for the BS-CRM model with p=2 (Gaussian kernel). We observe that beam-based models outperform the models that assign the tags individually with respect to all evaluation measures. The improvements are observed for all beam widths, suggesting that even approximately-optimal tag sets are more accurate than independently selected tags. The greatest improvements were observed with a beam width of 50, with the F1 measure flattening out, suggesting that there is no additional gain to be realised over beam widths wider than 50.

The blue line in Figure 2 shows the *annotation perplexity*[8] of each model. We observe that wider beams allow the model to find more probable (less perplexing) sets of annotation keywords. However, the most probable set of tags is not always the most accurate one. The annotation accuracy of the beam-based model peaks around the beam-width of 50, whereas the perplexity continues to decrease with wider beams. A similar phenomenon observed by [13] and prompted the use of greedy term selection in their model.

[8] We define annotation perplexity as $\{\prod_I P(\mathbf{w_I}|\mathbf{f_I})\}^{-\frac{1}{n}}$. Intuitively, annotation perplexity of K means that for each testing image the model is as *perplexed* as if it were choosing between K equally attractive sets of keywords. Lower perplexity is better. N is the number of images in the test set.

Model	R	P	F1	N^+
COREL				
CRM	19	16	17	106
Zhou [14]	20	19	19	...
Lui [9]	24	19	21	125
CRM (p=0.75)	25	21	23	119
Wang [13]	23	23	23	123
BS-CRM (p=0.75)	27	22	24	130

Model	R	P	F1	N^+
UW				
CRM (p=0.70)	36	36	36	86
BS-CRM (p=0.70)	46	42	44	106
IAPR TC-12				
CRM (p=0.70)	15	23	19	202
BS-CRM (p=0.70)	22	24	23	250

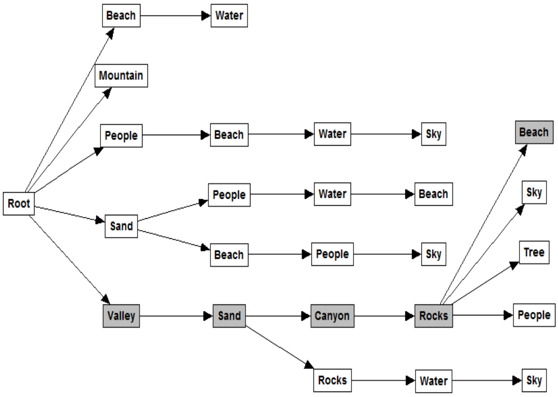
Table 1: Performance of the **BS-CRM** model on the Corel, IAPR TC-12 and UW datasets. R is % recall, P is % precision, N^+ is the number of words greater than zero. In all cases the BS-CRM model realises a significant increase in performance. Results for Wang [13] are for an identical setting of PAL=4. Model settings were: Corel dataset: BS-CRM: Beam width=5, PAL=4. UW Dataset: BS-CRM: Beam width=20, PAL=5. IAPR TC-12 Dataset: BS-CRM: Beam width=3, PAL=5.

Minkowski kernel based density: In Table 1 we further investigate the effect of replacing Gaussian kernels (Equation 3) with kernels based on the Minkowski kernel based density given by Equation 4 for the purpose of modelling image features. For the Corel dataset, p=0.75 leads to the highest performance on the validation set (see Figure 2). On the test set CRM (p=0.75) very substantially outperforms the original CRM formulation based on the Gaussian kernels. For CRM (p=0.75) this equates to a 32% increase in per-word recall, 31% increase in per-word precision and a 35% increase in F1. The improvement was statistically significant based on the paired t-test of per-word F1 (t-test: $p \leq 0.00004$).

Encouraged by these results, we test the effect of the p=0.75 kernel on the beam-based annotation models discussed above. BS-CRM (p=0.75) improvements over CRM (p=0.75) resulted in a confidence value of (t-test: $p \leq 0.08$) suggesting the result is statistically significant. For the IAPR TC-12 and UW datasets p=0.70 lead to the highest performance on the validation set. On the UW test set we realise a 22% increase in F1 (t-test: $p \leq 0.001$) with a 19% increase in F1 for IAPR TC-12 (t-test: $p \leq 2 \times 10^{-9}$). This suggests that, as for the case of Gaussian Kernels, the BS-CRM is also able to increase accuracy in the context of this particular kernel.

Comparison to literature: We note that BS-CRM model with p=0.75 on the Corel dataset fares well in comparison with results published by Zhou *et al.* [14], Liu *et al.* [9] and Wang *et al.* [13] showing improvements with respect to all accuracy measures. It is interesting to observe the interplay of the PAL technique of Wang et al. [13] and the BS-CRM algorithm. We find a benefit of using a PAL of 4 for BS-CRM (p=0.75) with no measurable benefit attained with the BS-CRM algorithm with lower settings for PAL. This suggests that the BS-CRM algorithm is able to effectively offset the rise in noisy tags being added for higher PAL settings as was experienced by Wang *et al.*

Minkowski vs. Gaussian kernels: In our experiments, the models based on the Minkowski kernel produced significantly more accurate annotations than the standard model based on the Gaussian kernel. On the Corel dataset the BS-CRM with a p=0.75 Minkowski kernel is found to be optimal gaining a 42% increase in recall, an 38% increase in precision and a 41% improvement in the F1 measure. The improvement is statistically significant with (t-test: $p \leq 0.00001$). This allows us to confidently conclude that the Minkowski kernel density is indeed superior for modelling image features in the context of the relevance-modelling framework of [8].



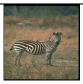
Model		
BS-CRM	Grass, Water, Bear Grizzly, Meadow	Grass, Tree, Water Plane, Zebra
CRM	Grass, Tree, Water Bear, Field	Grass, Tree, Cars Tracks, Prototype
Truth	Grass, Bear Meadow, Grizzly	Grass, Birds Plane, Zebra

Figure 3: **Left:** Example BS-CRM search tree. The first level corresponds to the annotation of the basic CRM. The BS-CRM refines this annotation by considering multiple hypotheses (defined by the beam width $B = 5$) for the most informative set of tags. 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 the grey path leads to the most informative hypothesis for the tag set. **Right:** The BS-CRM model is able to eliminate noisy tags produced by the CRM. For example, for the leftmost image the BS-CRM selects "grizzly and meadow" as more correlated to the existing labels of "bear, water, grass" than are "tree and field".

6 Conclusions and Future Work

In this paper we introduced the BS-CRM image annotation model which incorporates two novel contributions to the field. Firstly we investigated the effect of replacing the Gaussian kernel in the basic CRM with a generalised exponential kernel based on the Minkowski p-norm. Secondly we applied beam search to retrieve a near-optimal set of correlated tags. We showed that the BS-CRM model works significantly better than a number of other models for image annotation. In future it would be interesting to examine how the parameters of the model could be adapted dynamically to yield improved BS-CRM performance.

References

[1] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311, June 1993.

[2] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:394–410, March 2007.

[3] W. S. Cooper. Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *ACM Trans. Inf. Syst.*, 13(1):100–111, 1995.

[4] M. T. Dashti and A. J. Wijs. Pruning state spaces with extended beam search. In *Proceedings of the 5th international conference on Automated technology for verification and analysis*, ATVA'07, pages 543–552, 2007.

[5] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 97–112, London, UK, 2002. Springer-Verlag.

[6] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2:II–1002–II–1009 Vol.2, 2004.

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

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

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

[10] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pages 316–329, Berlin, Heidelberg, 2008. Springer-Verlag.

[11] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *MISRM'99: First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[12] C. Tillmann and H. Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Comput. Linguist.*, 29:97–133, March 2003.

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

[14] 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.