THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Sparse Kernel Learning for Image Annotation

Citation for published version:
Moran, S & Lavrenko, V 2014, Sparse Kernel Learning for Image Annotation. in *Proceedings of International Conference on Multimedia Retrieval.*, 113, ACM, New York, NY, USA. https://doi.org/10.1145/2578726.2578734

Digital Object Identifier (DOI):
10.1145/2578726.2578734

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of International Conference on Multimedia Retrieval

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 24. Sep. 2021

Sparse Kernel Learning for Image Annotation

Sean Moran
School of Informatics
The University of Edinburgh
EH8 9AB, Edinburgh, UK
sean.moran@ed.ac.uk

Victor Lavrenko
School of Informatics
The University of Edinburgh
EH8 9AB, Edinburgh, UK
vlavrenk@inf.ed.ac.uk

ABSTRACT

In this paper we introduce a sparse kernel learning framework for the Continuous Relevance Model (CRM). State-of-the-art image annotation models linearly combine evidence from several different feature types to improve image annotation accuracy. While previous authors have focused on learning the linear combination weights for these features, there has been no work examining the optimal combination of *kernels*. We address this gap by formulating a sparse kernel learning framework for the CRM, dubbed the SKL-CRM, that greedily selects an optimal combination of kernels. Our kernel learning framework rapidly converges to an annotation accuracy that substantially outperforms a host of state-of-the-art annotation models. We make two surprising conclusions: firstly, if the kernels are chosen correctly, only a very small number of features are required so to achieve superior performance over models that utilise a full suite of feature types; and secondly, the standard default selection of kernels commonly used in the literature is sub-optimal, and it is much better to adapt the kernel choice based on the feature type and image dataset.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Image Annotation, Visual Features, Statistical Models

1. INTRODUCTION

Over the past decade the number of images being captured and shared has grown enormously. There are several factors behind this remarkable trend. It is now commonplace for private individuals to own at least one digital camera, either attached to a mobile phone, or as a separate device in its own right. Digital cameras allow people to capture, edit, store and share high quality images with great ease. This factor, coupled with the low cost of memory and hard disk drives, has undoubtedly been a key driver behind the growth of personal image archives. Furthermore, the popularity of social networking websites such as Facebook, alongside image sharing websites such as Flickr, have given users an extra incentive to capture images to share and distribute amongst friends all over the world.

Substantial still image archives are also being amassed in the commercial domain. Markkula and Sormunen [14] studied the image archive of a Finnish Newspaper and described how archivists annotated pictures with keywords, with journalists searching the image collection based on those keywords. These companies typically employ teams of people to manually annotate the images. Correct annotation of images is crucial so as to maximise the efficiency in satisfying the needs of consumers; an incorrectly or insufficiently labelled image will be difficult to find in the archive. Manual image annotation, however, is infeasible for all but the smallest of image collections.

Recently there has been great interest amongst the computer vision and information retrieval community in the development of robust and efficient automatic image annotation systems. Automatic image annotation is the process of associating relevant labels to images that describe the semantic content of the images. Image annotation can be thought of as an instance of supervised classification for pictorial data [9]. The image annotation model learns an association between the visual content of images and their corresponding labels based upon a training dataset of manually annotated images. At test time, given a novel image, the annotation model tags the image with those labels most correlated with its visual appearance. The main purpose of annotating images in this manner is to allow for the retrieval of images based on natural language keywords.

Despite the popularity of automatic image annotation as a research topic the field is still very much an open research problem, mainly due to the fact that the analysis and understanding of images in unrestricted domains is an extremely challenging task. A balance has to be maintained by any algorithm between two conflicting goals: firstly the image representation chosen has to be very specific so as to be able to correctly differentiate between objects that may be easily confounded, such as sky and sea. On the other hand, any representation must be invariant to various confounding factors present in images such as occlusions, deformation,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ICMR '14, April 01 - 04 2014, Glasgow, United Kingdom
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2782-4/14/04 ...$15.00.
http://dx.doi.org/10.1145/2578726.2578734.

scale, background clutter, illumination and view point variations. These latter factors can make the same object look very different between images.

In this paper we introduce the *Sparse Kernel Learning Continuous Relevance Model* or SKL-CRM. We take the CRM image annotation model [11] as the basis of our proposed model and extend it in three ways: firstly, we introduce a framework for learning a sparse combination of base kernels in a data-driven manner. Secondly, we propose a simple technique for boosting the probability of rare words that are prevalent in benchmark image datasets. Lastly, we advocate the *Multinomial kernel*, a data-adaptive kernel that is capable of modelling discrete non-negative features in a probabilistically sound manner. We test the SKL-CRM on three standard datasets and show that it substantially outperforms almost every previously published image annotation model. We believe that this points to the importance of kernel selection, a problem that has been overlooked by previous researchers. We also draw two surprising conclusions: firstly, only a very small number of features are required so to achieve superior performance over models that utilise a full suite of feature types; and secondly, the standard default selection of kernels commonly used in the literature is sub-optimal, and it is much better to adapt the kernel choice based on the feature type and image dataset.

2. RELATED RESEARCH

Despite the difficulties inherent in the understanding of image content, substantial progress has still been realised in the area of automated image annotation over the past few years. Most if not all of the techniques suggested in the research literature approach the task by computing low-level image feature distributions for each concept of interest. This essentially reduces to the derivation of a probability model which links annotation keywords to image features. This probability model of associations between features and words can then be used to retrieve high probability keywords for a new feature set derived from a novel image. Recent results have shown that this approach is viable in improving retrieval results for a number of real-world image retrieval systems [10]. Early work on image annotation can usefully split according to the feature representation chosen: *Global feature-based* (also known as the scene-based approach) and *Block/region-based* [24].

Global feature based image annotation exploits the properties of global image features such as global colour and texture distributions. For example, Yavlinksy et al. [24] prepare a vector of real valued image features and a signature of image features to represent each image. A non-parametric kernel density estimator differentiates between the keywords by exploiting the irregularity in the distributions of image features. More recently [13] create a feature-set consisting of global colour and texture descriptors and apply a heuristic technique to combine distance computations on these features to create a nearest neighbour classifier for image annotation. The authors demonstrated a remarkable increase in image annotation accuracy as compared to previous work.

Block based image annotation applies an automatic segmentation step before the actual learning stage to identify salient objects within images. The general assumption is that feature computation based on a potentially strong segmentation better describes the visual objects, depicted in the image, than global features. This methodology depends highly on the performance of the selected segmentation algorithm to extract a good selection of coherent objects. Annotation quality is very sensitive to segmentation errors. Some authors have suggested that bypassing the segmentation step completely and computing features over a simple regular grid can in fact yield superior performance [5].

Research within the block-based image annotation branch can be broadly categorised by the type of model used: *generative* models, *discriminative* models and *nearest neighbour* based models. Generative modelling based approaches consist of *mixture models* and *topic models*. Mixture models formulate the image annotation task as the estimation of a joint likelihood over visual features and words. To annotate an unseen test image the model computes the conditional probability of each word in the vocabulary given the visual features of the image. A fixed number of the highest probability keywords are used as the annotation. Examples of such models include the Continuous Relevance Model (CRM) [11] and Multiple Bernoulli Relevance Model (MBRM) [5]. These models place a Gaussian non-parametric kernel density estimator over every training image to model the distribution of visual features, while words in the vocabulary are modelled using multinomial or Bernoulli distributions.

Examples of topic models adapted for image annotation include latent Dirichlet allocation [1], hierarchical Dirichlet processes [23], and machine translation methods [4]. The authors of [4] use a statistical machine translation model and apply EM to learn a maximum likelihood association of words to image regions using a bi-lingual corpus. In contrast, the Correspondence Latent Dirichlet Allocation (CorrLDA) [1] model uncovers this text-image link indirectly by appealing to a latent topic space in a generatively learned model. CorrLDA assumes a mixture of latent factors (topics) are responsible for generating words and image regions.

Rather than opting for a generative approach, other authors have proposed discriminative models for image annotation. In this scenario a separate classifier is built for each word in the vocabulary. Given a test image each word-specific classifier makes a judgement as to whether that image belongs to the class of images in our training dataset annotated with a particular word. Research in this area has investigated the application of support vector machines (SVMs), Bayes point machines and multiple instance learning-based models [7] [19]. For example [19] recently proposed a (one-vs-rest) SVM approach that employs a modified hinge loss to gain tolerance against confusing labels.

Nearest-neighbour (or local-learning) models predict keywords by taking a weighted combination of the keyword absence and presence among neighbouring images. Notable work in this area includes, Tagprop [8], short for tag propagation. In this model the weights of neighbouring images are based upon a set of distances computed using different similarity metrics across several feature types. The optimal weighted combination of these base distances is computed by maximising the log-likelihood of the word predictions on the training dataset. The direct integration of metric learning within the model was shown to substantially improve annotation performance over the state-of-the-art. Rather than solely build a model off either global or local image descriptors, the authors of [8] introduced the now *defacto* standard *multiple-feature* image annotation dataset. This dataset consists of 15 visual features ranging from local shape descriptors to global colour histograms.

The current state-of-the-art model for image annotation is the k-nearest neighbour model of [20]. There are two key ingredients to the success of this model: dealing with the severe class imbalance and exploiting the visual modality by learning an optimal weighted combination of base distances. To achieve keyword balance a unique and more balanced training dataset, referred to as a semantic neighbourhood, is crafted per test image based upon the visual similarity of a test image to the training dataset images. An optimal weighted combination of base distances and features is derived through a multi-label extension to the large-margin nearest-neighbour (LMNN) framework of [21].

3. BACKGROUND

The Continuous Relevance Model CRM [11] is a statistical model for automatically assigning words to unlabelled images using a set of N_J training images. The CRM estimates the joint probability distribution of a set of words $\mathbf{w} = \{w_1 \dots w_K\}$ from a vocabulary of size V together with an image $\mathbf{f}$ represented as a set of 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 (Equation 1).

$$P(\mathbf{w}|\mathbf{f}) = \frac{P(\mathbf{w}, \mathbf{f})}{P(\mathbf{f})} \tag{1}$$

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:

$$P(\mathbf{w}, \mathbf{f}) = \sum_{J \in T} P(J) \prod_{i=1}^{K} P(w_i|J) \prod_{i=1}^{M} P(\vec{f}_i|J) \tag{2}$$

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

$$P(w_i|J) = \frac{\mu p_v + N_{v,J}}{\mu + \sum_{v'} N_{v',J}} \tag{3}$$

Here $N_{v,J}$ is the number of times the keyword v appears in the annotation of training image J, p_v is the relative frequency that the word v appears in the training set and μ is a smoothing parameter selected based on a held out validation set. The CRM feature component $P(\vec{f}_j|J)$ is modelled with a kernel-based density estimator:

$$P(\vec{f}_i|J) = \frac{1}{R} \sum_{j=1}^{R} P(\vec{f}_i|\vec{f}_j) \tag{4}$$

Each region $j = 1 \dots R$ of the training image J instantiates a Gaussian kernel which has bandwidth β and is centered at the feature vector $\vec{f}_j$ of that region:

$$P(\vec{f}_i|\vec{f}_j) = \frac{1}{\sqrt{2^d \pi^d \beta}} exp\left\{ \frac{-||\vec{f}_i - \vec{f}_j||^2}{\beta} \right\} \tag{5}$$

Here d denotes the dimensionality of the image feature vectors and $||\vec{f}_i - \vec{f}_j||$ represents the Euclidean distance. The bandwidth parameter β is optimized on a held out portion of the training set.

4. THE SKL-CRM MODEL

4.1 Promoting the probability of rare words

The SKL-CRM models the distribution of image tags differently to the CRM. Due to the high imbalance between annotation keywords in many image datasets, recent annotation models [8] [20] attempt to boost the probability for rare words and decrease it for very frequent tags. To achieve a similar effect in the SKL-CRM we regularise the output annotation probability $P(\mathbf{w}|\mathbf{f})$ (Equation 1) using *max-min normalisation* (Equation 6).

$$\hat{P}(\mathbf{w}|\mathbf{f}) = \frac{P(\mathbf{w}|\mathbf{f}) - min_{\mathbf{f}'} P(\mathbf{w}|\mathbf{f}')}{max_{\mathbf{f}'} P(\mathbf{w}|\mathbf{f}') - min_{\mathbf{f}'} P(\mathbf{w}|\mathbf{f}')} \tag{6}$$

4.2 Kernel-Feature Alignment Algorithm

4.2.1 Problem Overview

Recent image annotation models employ the feature set introduced by [8], which consists of a mixture of local (SIFT, robust hue) and global (Gist, colour histograms) image features. Previous work use a Gaussian kernel for Gist features, a Laplacian kernel for the global colour histograms and a χ^2 kernel for the local SIFT based features [8]. To the best of our knowledge there has been no systematic study as to whether or not this assignment of kernels to feature types is in fact optimal across different image datasets. As different kernels correspond to different notions of similarity we hypothesise that assigning a specific kernel function to a feature type has an important impact on the quality of the resulting annotations. We argue that this commonly accepted setting of kernels to feature types is sub-optimal and it is better to *learn* the optimal kernel for each feature type.

To test our hypothesis we propose a kernel learning framework for the CRM [11] model, dubbed the Sparse Kernel Learning (SKL) CRM. We frame the learning problem as that of finding an optimal *alignment* between a given feature type (for example, an RGB colour histogram) and a particular kernel (for example, a Laplacian kernel). In principle the set of kernels could contain any valid kernel function. In this paper, we consider the χ^2 kernel (Section 4.3.3), Hellinger kernel (Section 4.3.3) and also two *data-adaptive* kernels: the Generalised Gaussian (Section 4.3.1) and our proposed *Multinomial kernel* for count-based image features (Section 4.3.2). Given a set of image features of size A and a set of kernels of size B, we wish to find a matrix $\mathbf{\Psi} \in \mathbf{\Pi}$ that specifies an optimal alignment between the two sets (Equation 7).

$$\mathbf{\Pi} := \left\{ \mathbf{\Psi} \in \{0,1\}^{A \times B} \text{ and } \forall i \sum_{j} \mathbf{\Psi}_{ij} = 1 \right\} \tag{7}$$

The alignment matrix $\mathbf{\Psi}$ specifies a mapping between elements of our feature set and kernel set. We find the best alignment $\mathbf{\Psi}^*$ by directly optimising the quality of the image annotations it yields (Section 4.2.2).

4.2.2 Optimising annotation F_1 score

Rather than optimise a convenient objective such as the log-likelihood [8], we directly optimise annotation accuracy as measured by the mean per word F_1 score computed on a held-out validation dataset. This F_1 score is computed as

follows: firstly, we use the SKL-CRM to annotate the validation dataset images. The predicted tags are determined by selecting 5 keywords per validation image that have the highest $\hat{P}(\mathbf{w}|\mathbf{f})$ (Equation 6) with the visual feature probability $P(I|J)$ given as in Equation 8.

$$P(I|J) = \prod_{i=1}^{M} \sum_{j=1}^{R} exp\left\{ -\frac{1}{\beta} \sum_{u,v} \mathbf{\Psi}_{u,v} k^v(\vec{f}_i^u, \vec{f}_j^u) \right\} \quad (8)$$

Here $k^v(\vec{f}_i^u, \vec{f}_j^u)$ denotes the v-th kernel function operating on the u-th feature type. Equation 8 is a principled generalisation of the CRM visual feature probability (Equation 4) to handle a bag of distinct feature types. The predicted annotations can be compared to the ground-truth annotations to compute the F_1 score: if a word w_i is present in the ground-truth of n_{i1} images, and it is predicted for n_{i2} images out of which n_{i3} of the predictions are correct - precision is therefore n_{i3}/n_{i2} and recall is n_{i3}/n_{i1}. The F_1 score over the entire vocabulary is subsequently given as in Equation 9.

$$F_1 = \frac{2}{V} \times \frac{\sum_{i=1}^{V}(n_{i3}/n_{i2}) \times \sum_{i=1}^{V}(n_{i3}/n_{i1})}{\left\{\sum_{i=1}^{V}(n_{i3}/n_{i2}) + \sum_{i=1}^{V}(n_{i3}/n_{i1})\right\}} \quad (9)$$

We optimise the objective function $F_1(\mathbf{G}, \hat{\mathbf{P}}_{\mathbf{\Psi}})$ which takes as input a ground-truth matrix $\mathbf{G} \in \Re^{N_I \times V}$ and a label prediction matrix $\hat{\mathbf{P}}_{\mathbf{\Psi}} \in \Re^{N_I \times V}$ where each element is $\hat{P}(\mathbf{w}|\mathbf{f})$ (Equation 6), N_I is the number of testing images, and returns the corresponding F_1 score. The kernel-feature alignment $\mathbf{\Psi}$ is now represented implicitly by the annotations $\hat{\mathbf{P}}_{\mathbf{\Psi}}$ resulting from that alignment. The ground-truth matrix specifies the true labels for each test image while the prediction matrix $\hat{\mathbf{P}}_{\mathbf{\Psi}}$ gives the SKL-CRM predicted labels for a specific kernel-feature alignment $\mathbf{\Psi}$. Our optimisation objective can be compactly stated as in Equation 10.

$$\begin{aligned} \underset{\mathbf{\Psi}}{\text{maximize}} \quad & F_1(\mathbf{G}, \hat{\mathbf{P}}_{\mathbf{\Psi}}) \\ \text{where} \quad & \hat{\mathbf{P}}_{\mathbf{\Psi}} = \texttt{promote}(\mathbf{P}_{\mathbf{\Psi}}) \\ \text{and} \quad & \mathbf{P}_{\mathbf{\Psi}} = \hat{\mathbf{S}}\mathbf{W} \end{aligned} \quad (10)$$

The function promote(.) applies Equation 6 to each element of the label prediction matrix $\mathbf{P}_{\mathbf{\Psi}}$, $\mathbf{W} \in \Re^{N_J \times V}$ holds the image-word probabilities $P(w|J)$ and $\hat{\mathbf{S}} \in \Re^{N_I \times N_J}$ is the matrix of Bayesian posterior probabilities $P(J|I)$.

4.2.3 Greedy Set-Based Alignment Algorithm

The consequence of directly optimising the annotation F_1 score is that the objective $F_1(\mathbf{G}, \hat{\mathbf{P}}_{\pi})$ is both non-smooth and non-convex making it difficult to maximise via gradient ascent. To circumvent this issue, we introduce a deterministic greedy approach to aligning each feature type with a kernel that leads to maximisation of the F_1 score. Our proposed optimisation strategy is presented in Algorithm 1. Starting with an empty set, this algorithm, at each iteration, *greedily* adds the feature-kernel combination that maximises the F_1 annotation score with respect to the features and kernels already present in the set. The parameters β (Equation 8) and μ (Equation 3) are optimised individually as each new feature-kernel combination is considered for addition to the set. We observe rapid convergence to a local optima typically only after five feature-kernel combinations have been added to the set (Section 5.3).

Algorithm 1 Greedy kernel-feature alignment algorithm

Input: Ground-truth label matrix $\mathbf{G}$.
Output: Optimal kernel-feature alignment matrix $\mathbf{\Psi}^*$
$\mathbf{\Psi}^* = 0$
while $\mathbf{\Psi}^*$ changes **do**
  $\mathbf{\Psi} = \mathbf{\Psi}^*$
  //Find the best kernel-feature to add to the set//
  for each a s.t. $\forall i$ $\mathbf{\Psi}(a,i) = 0$ **do**
    for for each b,μ,β **do**
      $\mathbf{\Psi}(a,b) = 1$
      if $F_1(G, \hat{\mathbf{P}}_{\mathbf{\Psi}}) > F_1(G, \hat{\mathbf{P}}_{\mathbf{\Psi}^*})$ **then**
        $\mathbf{\Psi}^* = \mathbf{\Psi}$
      end if
      $\mathbf{\Psi}(a,b) = 0$
    end for
  end for
  //Optimise selected kernel-features in the set//
  for each a s.t. $\exists i$ $\mathbf{\Psi}^*(a,i) = 1$ **do**
    $\mathbf{\Psi} = \mathbf{\Psi}^*$
    $\mathbf{\Psi}(a,i) = 0$
    for for each b,μ,β **do**
      $\mathbf{\Psi}(a,b) = 1$
      if $F_1(G, \hat{\mathbf{P}}_{\mathbf{\Psi}}) > F_1(G, \hat{\mathbf{P}}_{\mathbf{\Psi}^*})$ **then**
        $\mathbf{\Psi}^* = \mathbf{\Psi}$
      end if
      $\mathbf{\Psi}(a,b) = 0$
    end for
  end for
end while

4.3 Discrete and Real-Valued Kernels

In this section we describe the set of kernels we use in our SKL-CRM model. The kernels under consideration can be categorised into two groups: those specialised for real-valued features (Section 4.3.1) and kernels better able to model discrete count-based features (Section 4.3.2, Section 4.3.3).

4.3.1 Generalised Gaussian Kernel

This kernel is similar to the Minkowski kernel of [16] and, as such, 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. The generalised Gaussian kernel parametrised by a shape factor $\mathtt{p}$ is defined as follows:

$$P(\vec{f}_i|\vec{f}_j) = \frac{\mathtt{p}^{1-1/\mathtt{p}}}{2\beta\Gamma(1/\mathtt{p})} \exp\left[-\frac{1}{\mathtt{p}} \frac{|\vec{f}_i - \vec{f}_j|^{\mathtt{p}}}{\beta^{\mathtt{p}}} \right], \quad (11)$$

Here $\Gamma(\cdot)$ denotes the gamma function and $|\vec{f}_i - \vec{f}_j|^{\mathtt{p}} = \sum_{d=1}^{D} |f_{i,d} - f_{j,d}|^{\mathtt{p}}$ is a generalisation of the Euclidean norm. The summation goes over the dimensions d of the feature vectors. $\mathtt{p}$ and β are positive free parameters set on a held-out validation set. By varying the value of $\mathtt{p}$ we can obtain a broad range of different kernel functions: if $\mathtt{p} \to 0$ a Dirac delta function appears, if $\mathtt{p} = 1$ we obtain the Laplacian, if $\mathtt{p} = 2$ a Gaussian is the result and if $\mathtt{p} \to \infty$ a uniform

kernel is revealed. For fractional values ($0 < \mathtt{p} < 1$) we have the Minkowski family of kernels.

4.3.2 Multinomial Kernel

In this paper we advocate a *Multinomial* kernel for image annotation that is specifically optimised for *count-based* descriptors, and defined as follows:

$$P(\vec{f}_i|\vec{f}_j) = \frac{(\sum_d f_{i,d})!}{\prod_d (f_{i,d}!)} \prod_d (p_{j,d})^{f_{i,d}} \tag{12}$$

Here the products go over the bins d in the histograms, $f_{i,d}$ represents the count for bin d in the unlabelled image i, and $f_{j,d}$ is the corresponding count for the training image j. The Multinomial coefficient in front of the product is independent of the training image j, and cancels out when we compute the conditional probability $P(\mathbf{w}|\mathbf{f})$. We use Jelinek-Mercer smoothing for estimating the parameters $p_{j,d}$ of the Multinomial kernel:

$$p_{j,d} = \lambda \frac{f_{j,d}}{\sum_d f_{j,d}} + (1-\lambda) \frac{\sum_j f_{j,d}}{\sum_{j,d} f_{j,d}} \tag{13}$$

The smoothing parameter λ is optimized on a held-out portion of the training set. We believe that Multinomial kernels offer a probabilistically sound way of modelling histogram-based feature vectors, because they are specifically designed for discrete observations (counts), do not suffer from model deficiency [3] and properly estimate the likelihood of low and zero counts.

4.3.3 Additive Homogeneous Kernels

In addition to the Generalised Gaussian and Multinomial Kernels, we also study two additive homogeneous kernels. Specifically, we consider the Hellinger kernel (Equation 14).

$$k(\vec{f}_i, \vec{f}_j) = \sum_d \sqrt{f_{i,d} f_{j,d}} \tag{14}$$

for two L_1 normalised feature vectors $\vec{f}_i$ and $\vec{f}_j$ (i.e. $\sum_d f_{i,d} = 1$ and $f_{i,d} \geq 0$). In addition, we also consider the χ^2 kernel (Equation 15).

$$k(\vec{f}_i, \vec{f}_j) = \sum_d \frac{2 f_{i,d} f_{j,d}}{f_{i,d} + f_{j,d}} \tag{15}$$

Both kernels are commonly used for computing histogram distance due to their higher sensitivity to smaller bin values as compared to the Gaussian kernel.

5. EXPERIMENTS

5.1 Datasets

We evaluate on three standard image annotation datasets. The datasets cover a diverse range of different image topics from natural scenes to personal photos, logos and drawings thereby providing a challenging test suite for evaluation. All datasets are identical to those used in most recent image annotation publications [20] [8], thereby permitting direct comparison.

Corel 5K: has for a long time been a standard benchmark for image annotation. The dataset consists of 5,000 images from 50 Corel Stock Photo CDs. Each CD includes 100 images on the same topic. Each image contains an annotation of 1-5 keywords. Overall there are 371 tags of which 260 occur in the test set. In our evaluation a fixed set of 499 images are used for testing with the remainder used for training. This split corresponds to previous related work [8].

IAPR TC12: is a collection of 19,627 images of natural scenes that include different sports and actions, photographs of people, animals, cities and landscapes. The vocabulary consists of 291 tags, with an average of 4.7 keywords per image. There are 17,665 training images with the remaining 1,962 used for testing.

ESP Game: consists of 20,768 images collected in the ESP collaborative image labelling task. In ESP game two players assign labels to the same image without communicating. Only common labels are accepted thereby enticing players to provide accurate tags to the images. We use the identical image subset as [8]. There are 18,687 images in the training dataset and 2,081 in the test dataset.

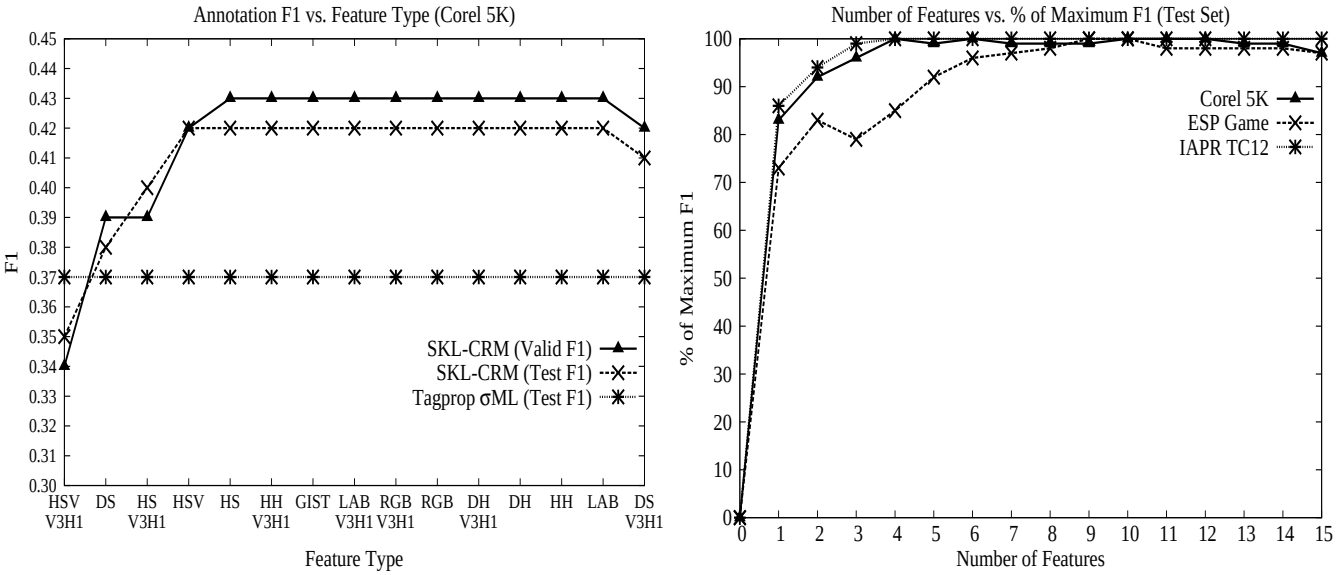
5.2 Experimental Methodology

Features: To fairly compare our model performance to recent work we use, without modification, the feature set introduced by [8] in the context of their Tagprop model for image annotation. The feature set consists of a mixture of 15 distinct local and global descriptors. The local descriptors include SIFT and local hue histograms both of which are extracted densely on a multiscale grid or for Harris-Laplacian interest points. The local descriptors are quantized using k-means with each image being represented as a bag-of-visual-words histogram. Global features consist of Gist features which encode the layout of the image and colour histograms with 16 bins in each colour channel for the RGB, LAB, HSV colour spaces. All descriptors except for Gist are L_1-normalised. Furthermore all features (except for Gist) are computed in a spatial arrangement[1]. In all, there is one Gist descriptor, six colour histograms and eight bag-of-features.

Parameter Optimization: The parameter optimisation strategy is identical for each dataset. We divide each dataset into three parts: a training dataset, a validation dataset and a testing dataset. The validation dataset is used to find learn the optimal kernel-feature combination (Section 4.2) and is a randomly selected subset of the training dataset. After fixing the parameters, we merge the training and validation datasets to make a new training dataset. The training and testing dataset splits for all three datasets are identical to previously published work [8] [20]. Our final reported test F_1 score is determined as follows: we take the parameter configuration at the point where *validation* dataset F_1 score is maximised and then run that instance of the SKL-CRM model on the test dataset, reporting the resulting F_1 score.

Evaluation Procedure: We are given an un-annotated image I and are asked to automatically produce an annotation $\mathbf{w}_{auto}$. The automatic annotation is then compared to the held-out human annotation $\mathbf{w}_I$. We follow the experimental methodology used by [4]. Given a test image we use the SKL-CRM algorithm to determine the 5 words with the highest conditional probability (Equation 6) and call them the automatic annotation of the image in question. Then, following [4], we compute annotation recall and precision for

[1]Features computed in a spatial arrangement are denoted with a *V3H1* suffix in this paper.



(a) Corel 5K F_1 optimisation

(b) All datasets F_1 optimisation

Figure 1: (a) Corel 5K annotation F_1 score versus the contents of the feature set. (b) The SKL-CRM is able to reach the maximum annotation F_1 score across all three datasets with only a very small subset of the available image features.

	Dataset		
Model	Corel 5K	IAPR TC12	ESP Game
CRM†	17	–	–
CRM‡	29	26	19
SKL-CRM	42	38	32

Table 1: Annotation F_1 scores for various incarnations of the CRM model. CRM† is the original CRM as reported in [11] using the feature set of [4]. CRM‡ is the CRM model using all 15 tagprop-based features [8] and default kernel selection. SKL-CRM is our proposed model with adaptive kernel allocation.

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. In addition we include the number of words with recall greater than zero (denoted as $N+$): this metric seeks to measure the ability of the system to label images with rare keywords.

5.3 Results

In this section we evaluate the performance of our model on the task of automatic image annotation. We examine one primary hypothesis, namely learning an optimal combination of kernels using the data itself, owing to its different geometry over the feature space, will outperform the standard (default) assignment of kernels to feature types often found in the literature [8] [20]. In this section we discuss a set of experiments we carried out to test this hypothesis.

Standard vs. Data-Driven Kernel Assignment: Table 1 shows how our proposed model compares against the original CRM model (CRM†) and against the CRM model using the full 15 Tagprop based features and default kernel assignments (CRM‡). It is clear from this table that the SKL-CRM model substantially outperforms both models, across all three datasets. For example, on the Corel 5K dataset the SKL-CRM attains a 147% increase over CRM†. Against CRM‡ the SKL-CRM realised a 45% increase in F_1 measure. There are two interesting conclusions from this experiment: firstly, the Tagprop based features are clearly a more powerful set of features than those of Duygulu et al. [4] - simply using the CRM with these features, we obtain a substantial increase in performance over CRM†. Secondly, it is more effective to adapt the kernels based on the data itself, rather than opt for the default selection of kernels suggested in the literature. This is vividly demonstrated by the large increase in performance of the SKL-CRM versus CRM‡.

For the Corel 5K dataset we find a Multinomial Kernel (MK) ($\lambda = 0.99$) optimal for the HSV feature type, a Generalised Gaussian (GG) kernel ($\mathtt{p} = 0.9$) for HSV_V3H1, a GG ($\mathtt{p} = 0.1$) for Harris Hue (HH_V3H1), a Gaussian for Harris SIFT (HS), a GG ($\mathtt{p} = 0.7$) for HS_V3H1, and a Laplacian for Dense SIFT (DS). This result provides two interesting conclusions: firstly, notice the prevalence (4 out of 6) of data-driven kernels amongst the alignments, including our proposed Multinomial kernel - data-adaptive kernels are clearly more effective than standard kernels. Secondly, we observe that no kernel-feature assignment agrees with the standard assignment recommended in the literature. This observation demonstrates that is it difficult to predict, a-priori, which kernel is best for a given feature, justifying the need for our greedy kernel-feature alignment algorithm. We make the same general observations for the larger IAPR TC12 and ESP Game datasets.

Greedy Optimisation Algorithm: In Figure 1(a), for Corel 5K, we show the progress of our greedy optimisation algorithm as each new feature-kernel alignment is added

	COREL 5K					IAPR TC12					ESP Game			
Model	R	P	F_1	N^+		R	P	F_1	N^+		R	P	F_1	N^+
CRM [11]	19	16	17	107		-	-	-	-		-	-	-	-
MBRM [5]	25	24	25	122		23	24	23	223		19	18	18	209
InfNet [15]	24	17	20	112		-	-	-	-		-	-	-	-
NPDE [24]	21	18	19	114		-	-	-	-		-	-	-	-
SML [2]	29	23	26	137		-	-	-	-		-	-	-	-
TGLM [12]	29	25	27	131		-	-	-	-		-	-	-	-
JEC [13]	32	27	29	139		29	28	28	250		25	22	23	224
Tagprop SD [8]	33	30	31	136		20	50	29	215		19	48	27	212
MRFA [22]	36	31	33	172		-	-	-	-		-	-	-	-
GS [25]	33	30	31	146		29	32	30	252		-	-	-	-
RF-opt [6]	40	29	34	157		31	44	36	253		26	41	32	235
CCD (SVRMKL+KPCA) [17]	41	36	38	159		29	44	35	251		24	36	29	232
KSVM-VT [19]	42	32	36	179		29	47	36	268		32	33	33	259
Tagprop ML [8]	37	31	34	146		25	48	33	227		20	49	29	213
Tagprop σML [8]	42	33	37	160		35	46	40	266		27	39	32	239
SKL-CRM (this work)	**46**	**39**	**42**	**184**		32	47	38	**274**		26	41	32	248

Table 2: Performance of the SKL-CRM model against a wide range of recent annotation models on three benchmark image annotation datasets (Corel, IAPR TC12 and ESP game).

to the set. Remarkably the SKL-CRM model attains the maximum annotation performance of 0.434 F_1 on the validation set (0.420 F_1 on the test set) after only *six* feature types (HSV and HSV_V3H1, Dense SIFT (DS), Harris SIFT (HS and HS_V3H1) and Harris Hue (HH_V3H1)) have been added to the set. Furthermore, with just *two* features the SKL-CRM reaches 90% performance, surpassing Tagprop σ-ML. These results demonstrate that further features are detrimental and our greedy optimisation algorithm is able to effectively identify a small subset of features that jointly maximise annotation performance. This trend is repeated on the IAPR TC12 and ESP Game datasets where we also find sparse optimal solutions (Figure 1(b)): for IAPR TC12, only 3 features are required to reach the maximum annotation F_1, whereas 9 features are required for ESP Game.

Interestingly, we found no additional benefit in using a weighted combination of the optimal kernel-feature alignments. We hypothesise that the kernels themselves are acting as "weights" either up-weighting the effect of a feature type that is added in the initial stages of the optimisation procedure, while down-weighting the contribution of those features added towards the end. In our experimental results we noticed that a Generalised Gaussian with $\mathtt{p} = 0.1$, effectively a Dirac spike, was frequently aligned to those features added in the latter stages. In contrast, Generalised Gaussian kernels with a higher value of $\mathtt{p}$ (or a Multinomial Kernel with a high setting of λ) were assigned to features in the early part of the optimisation procedure. As the initial features added to the set are responsible for the vast majority of the annotation performance we believe that the higher p-norm Generalised Gaussian kernels (or the Multinomial kernel) are up-weighting those features, whereas the low p-norm kernels are suppressing the influence of those latter, and less effective, features.

SKL-CRM Performance vs. the Literature: Table 2 presents the results of the SKL-CRM model against a broad selection of image annotation models recently proposed in the literature. Encouragingly, on the Corel 5K dataset we find a substantial increase in annotation F_1 measure with respect to nearly all recently proposed image annotation models. For example, with improve annotation F_1 by 14% with respect to the strong baseline of Tagprop σ-ML - a local learning model that employs metric learning to find an optimal combination of base kernels and word-specific logistic sigmoids to boost the probability of rare words [8]. Our superior performance to Tagprop σ-ML on this dataset demonstrates that learning an optimal combination of kernels can be more effective than learning an optimal combination of weights for the default base kernels. Table 2 also presents results IAPR TC12 and ESP Game datasets. We find that the SKL-CRM is also competitive to recently proposed models on these two larger datasets.

Finally, Table 3 demonstrates how our proposed model compares against the current best performing image annotation model, the 2PKNN model of [20]. We believe our proposed algorithm, while not reaching the annotation quality attained by 2PKNN, offers significant computational advantages. Firstly, feature sparsity increases model robustness and hinders over-fitting while also substantially reducing the computational complexity of the model [25]. We sacrifice no annotation performance through this high level of sparsity, and we are further encouraged by the fact that our model substantially outperforms the group sparsity image annotation model of [25]. In addition to sparsity, our greedy optimisation algorithm does not require the computation of a gradient. Our optimisation technique, specifically lines 7-15 in Algorithm 1, is therefore amenable to massive parallelisation for big data image annotation [18].

6. CONCLUSIONS

In this paper we introduced a sparse kernel learning (SKL) framework for the Continuous Relevance Model (CRM). The SKL-CRM model incorporates a greedy kernel-feature alignment algorithm which, at each iteration, determines the best kernel for a given image feature type. The alignment is chosen based on how well, in terms of annotation F_1 measure, that feature-kernel alignment performs in combination with a set of previously aligned features. In our experimental validation we observed that this greedy alignment algorithm is

	COREL 5K					IAPR TC12					ESP Game			
Model	R	P	F_1	N^+		R	P	F_1	N^+		R	P	F_1	N^+
2PKNN [20]	40	39	40	177		32	49	39	274		23	51	32	245
2PKNN-ML [20]	46	44	45	191		37	54	44	278		27	53	36	252
SKL-CRM (this work)	46	39	42	184		32	47	38	274		26	41	32	248

Table 3: Comparison of the SKL-CRM model against the current state-of-the-art model (2PKNN).

able to reach an impressive level of annotation performance by using only a sparse subset of the available features. This sparse feature representation provides storage and processing advantages over comparable models at test time, while in many cases surpassing recent image annotation models.

Experimental validation of the SKL-CRM brought three further interesting findings: firstly, data-adaptive kernels, such as the Generalised Gaussian and our proposed Multinomial kernel are more effective for image annotation than standard kernels such as the Gaussian or χ^2 kernels. Secondly, it is impossible to predict a-priori which particular kernel is appropriate for a given feature type. In most previous work it is assumed, for example, that the Gaussian kernel is most appropriate for the Gist feature, while colour histogram features can be best exploited with the Laplacian kernel. In this paper we demonstrated that this assumption is flawed, and in fact it is much better to learn the appropriate kernel for a given feature based on the image data itself. Lastly, we found no additional benefit in learning a weighted combination of the optimal kernel-feature alignments.

The SKL-CRM aligns a feature to a single kernel. In the future we will investigate a *continuous relaxation* of this discrete alignment constraint. A continuous relaxation would allow a feature type to be represented as a weighted superposition of kernels, which may lead to enhanced accuracy.

7. ACKNOWLEDGEMENTS

We thank Yashaswi Verma for helpful discussions.

8. REFERENCES

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *In JMLR'03.*

[2] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *In PAMI'07.*

[3] W. S. Cooper. Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *In TOIS'95.*

[4] 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.*

[5] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. *In CVPR'04.*

[6] H. Fu, Q. Zhang, and G. Qiu. Random forest for image annotation. In *ECCV'12.*

[7] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *In PAMI'08.*

[8] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV'09.*

[9] C. Hentschel, S. Stober, A. Nürnberger, and M. Detyniecki. Automatic image annotation using a visual dictionary based on reliable image segmentation. In *AMR'08.*

[10] V. Lavrenko, S. Feng, and R. Manmatha. Statistical models for automatic video annotation and retrieval. *In ICASSP'04.*

[11] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS'03.*

[12] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *In JPR'09.*

[13] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV '08.*

[14] M. Markkula and E. Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *In IR'00.*

[15] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *CIVR'04.*

[16] S. Moran and V. Lavrenko. Optimal tag sets for automatic image annotation. In *BMVC'11.*

[17] H. Nakayama. *Linear distance metric Learning for large-scale generic image recognition.* PhD thesis, The University of Tokyo, Japan, 2011.

[18] P. Richtárik and M. Takác. Distributed coordinate descent method for learning with big data. *In CoRR'13.*

[19] Y. Verma and C. V. Jawahar. Exploring svm for image annotation in presence of confusing labels. In *BMVC'13.*

[20] Y. Verma and C. V. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *ECCV'12.*

[21] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *In JMLR'09.*

[22] Y. Xiang, X. Zhou, T.-S. Chua, and C.-W. Ngo. A revisit of generative model for automatic image annotation using markov random fields. In *CVPR'09.*

[23] O. Yakhnenko and V. Honavar. Annotating images and image objects using a hierarchical dirichlet process model. In *MDM '08.*

[24] A. Yavlinsky, E. Schofield, and S. Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *CIVR'05.*

[25] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas. Automatic image annotation using group sparsity. In *CVPR'10.*