Int J Multimed Info Retr
DOI 10.1007/s13735-014-0063-y

REGULAR PAPER

A sparse kernel relevance model for automatic image annotation

Sean Moran · Victor Lavrenko

Received: 28 June 2014 / Accepted: 17 August 2014

Abstract In this paper, we introduce a new form of the continuous relevance model (CRM), dubbed the SKL-CRM, that adaptively selects the best performing kernel per feature type for automatic image annotation. Previous image annotation models apply a standard selection of kernels to model the distribution of image features. Popular examples include a Gaussian kernel for modelling GIST features or a Laplacian kernel for global colour histograms. In this work, we demonstrate that this standard assignment of kernels to feature types is sub-optimal and a substantially higher image annotation accuracy can be attained by adapting the kernel-feature assignment. We formulate an efficient greedy algorithm to find the best kernel-feature alignment and show that it is able to rapidly find a *sparse subset* of features that maximises annotation F_1 score. In a second contribution, we introduce two data-adaptive kernels for image annotation—the *generalised Gaussian* and *multinomial kernels*—which we demonstrate can better model the distribution of image features as compared to standard kernels. Evaluation is conducted on three standard image datasets across a selection of different feature representations. The proposed SKL-CRM model is found to attain performance that is competitive to a suite of state-of-the-art image annotation models.

Keywords Image annotation · Object recognition · Kernel density estimation

S. Moran (✉) · V. Lavrenko
University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, Scotland, UK
e-mail: sean.moran@ed.ac.uk

V. Lavrenko
e-mail: vlavrenk@inf.ed.ac.uk

1 Introduction

Automatic image annotation describes the algorithmic process of assigning one or more descriptive keywords to a digital image. For example, if we have an image of a tiger in a jungle setting, we might expect the algorithm to assign the keywords "Tiger", "Grass", "Tree" to the image. This research field has become increasingly popular over the years due in part to the massive growth in the availability of large still image archives, both within industry and in private personal collections. To facilitate search and exploration of digital still image archives, each image requires a set of labels that describe the high-level semantic content of the image. Unfortunately, manual labelling suffers from the disadvantages of not only being slow, expensive and highly subjective, but also impossible to scale to modern multi-million image libraries.

The social study conducted by Ames and Naaman [2] provides an insight into the motivations that drive private individuals to annotate their images. This study revealed a changing opinion with regard to the usefulness of manual image annotation, from it being nearly completely avoided for personal offline collections through to it being warmly embraced for online collections such as those on Flickr. The authors revealed a taxonomy of reasons behind this increase in motivation, with one of the most interesting being the social incentives brought about by online libraries, where, for example, a photographer may obtain the "satisfaction" of having made available a highly popular (or most viewed) photograph on the website. Despite this, many images posted online lack high-quality descriptive labels forcing the major search engines to rely on analysing the text in the associated webpage to discover the semantic content of an image.

Large organisations such as newspapers and broadcasters also maintain substantial still image archives. Markkula

Published online: 19 September 2014

and Sormunen [31] 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. The alternative search techniques of query by sketch and query by example have been cited as less flexible and user friendly[1] means of querying image libraries than the familiar query by text already employed to search document collections.

Over the past decade, the computer vision community has attempted to address the problem of automatic annotation of images. The depth and breadth of the research have been astounding, with suggested algorithms cutting right across the entire taxonomy of machine learning models. We conduct a thorough review of the research field in Sect. 2. In general terms, however, image annotation can be thought of as a form of multi-label classification of image-based data [18]. We are given an unannotated image which is typically parsed into a set of discriminative image features. The features from our novel image are compared to the features of a large training set of manually labelled images. Images in our training set that have highly similar features, as defined by some notion of similarity (e.g. a kernel function), have a greater chance of propagating their associated keywords to our novel test image.

Automatic image annotation 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. Any algorithm has to surmount the so-called *semantic gap* between low-level image features (GIST, global colour histograms) and high-level concepts (Tiger, Grass) in the images. To have any hope of doing so, an annotation model must maintain a fine balance between two conflicting goals: firstly, the image representation 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, scale, background clutter, illumination and view point variations. These latter factors can make the same object look very different between images.

Even if we can attain a reliable mapping between low-level image features and real-world concepts, authors such as Enser [12] highlight fundamental limitations of any means of automatic image annotation. Firstly, the vocabulary keywords relate to entities that are visible within the image, whereas real users tend to submit search queries related to more abstract scenarios that involve the depicted objects. Enser illustrates this point with the query "the first public engagement of Prince Charles" which would be difficult to identify from content extractable by automated algorithms. Furthermore, the generic object limitation questions the usefulness of generic labels for images such as "sun", "grass" and "tiger": "...they have the common property of visual stimuli which require a minimally-interpretive response from the viewer". That is to say, users typically submit queries referring to objects by proper name which typically have limited associated visual features in images. As a consequence Enser argues that any defining textual annotations may always necessarily have to be manually assigned to images.

Despite the computational and philosophical difficulties in automating the process of image annotation, it is widely regarded that semantic indexing of images using the current breed of annotation algorithms, while not entirely capturing the full conceptual properties, is still a considerable advance over relying solely on manual annotation or having no associated labels at all. In this paper, we focus on advancing the state-of-the-art in automatic image annotation by proposing a set of novel contributions that increase the labelling accuracy of the continuous relevance model (CRM) of [26]. Our new probabilistic model of image annotation is dubbed the Sparse Kernel Learning Continuous Relevance Model or SKL-CRM and makes two novel contributions to the field.

Our first contribution is a greedy algorithm for automatically discovering the best kernel (e.g. Laplacian, Gaussian) to model the similarity between a given set of image features (e.g. GIST, global colour histograms). Our experimental results show that this greedy algorithm is effective in finding a *sparse subset* of features that maximise the image annotation performance while at the same time not requiring the computation of the derivative of the objective function. Evaluation of this algorithm leads to two surprising conclusions: firstly, with just a small subset of the available features, our model can equal and in some cases outperform competing image annotation models in the literature; and secondly, the standard default selection of kernels for each feature type as commonly used in the literature is sub-optimal, and it is much better to adapt the kernels to a given feature type for a particular image dataset.

Our second contribution is the introduction of two data-adaptive visual kernels for automatic image annotation: the *generalised Gaussian* and *multinomial kernels*. The kernels are parametrised by a shape factor that permits both to be adapted to a specific feature set. In our experimental evaluation, we show that the data-adaptive capability permits both kernels to be substantially more effective at modelling the distribution of image features as compared to standard kernels such as the Gaussian or χ^2 kernel. We envisage these data-adaptive kernels being valuable not only for image anno-

[1] Users are known to find it particularly difficult to represent their image needs via abstract image features [23].

tation but in a plethora of other domains that employ kernels for modelling the similarities between vectorial data [44].

The SKL-CRM combines these novel contributions in a single, probabilistically sound model of image annotation. We thoroughly evaluate the performance of our model against competing models in the literature and find that it outperforms a wide selection, while reaching a competitive level of performance with the state-of-the-art.

2 Related research

Automatic image annotation has been the subject of an intense level of research over the past decade and a diverse set of models have been suggested to tackle the problem. The field can be usefully divided into the type of feature representation used (global feature-based or block-based), followed by the type of statistical model (local learning, generative, discriminative etc). All models are united in their attempt to formulate a mapping between the low-level image features and annotation keywords.

The *global feature*-based branch, also known as the scene-based approach, exploits the properties of global image colour and texture distributions to differentiate between annotation keywords. These methods sidestep the commonly error-prone segmentation of an image into distinct sub-regions and effectively exploit earlier pyschovisual studies that found evidence of a link between the scene category of an image and the associated colour distribution [39]. The global feature representation is reported to perform well in classification tasks when the distinctive visual properties are distributed equally throughout the image, for example, in a city scene where strong vertical and horizontal edges are dominant.

Prominent research within the global feature branch include [7,20,30] and [53]. Early work, such as that of [7], explored the applicability of a support vector machine (SVM) classifier to predict the keyword class of images based on extracted HSV colour histograms. In contrast, Huang et al. [20] employed a classification tree to model the spatial correlation of colours in images. More recently, Yavlinksy et al. [53] construct a simple non-parametric kernel density model (NPDE) based on CIELab colour features and Tamura and Log-Gabor texture filters and demonstrate that these basic global image properties can attain reasonable levels of annotation accuracy. In a similar vein, Makadia et al. [30] introduced a different feature set also consisting of colour (RGB, HSV, LAB) and texture-based descriptors (Haar and Gabor filters) and described a heuristic nearest neighbour-based model for feature fusion and label prediction. The authors established a new baseline for image annotation that exhibited significantly improved performance over the then state-of-the-art in the field.

In contrast, *block/region-based* image annotation introduces an automatic segmentation step, such as normalised cuts [42], before the learning stage so as to isolate real-world objects within the images. The hope is that a good segmentation can better resolve the presence of visual objects within the image versus the sole use of global features. Unfortunately, segmentation in unconstrained images is a challenging problem and the process may sometimes fail to extract a set of coherent objects. As a second step, the features resulting from the segmented regions are subsequently clustered into a discrete visual vocabulary. As indicated by some authors [26], annotation quality is sensitive to clustering errors and depends on being able to a priori select the right cluster granularity: selecting too many clusters results in extreme sparseness of the space, while too few will lead us to confuse different objects in the images. These issues with automatic segmentation have led some authors to bypass this step entirely and compute features over a simple regular grid, which can in fact yield superior performance [13].

There are many examples in the literature of models that rely on a segmentation step prior to learning. It is useful to split this branch by the type of model used: *generative* models, *discriminative* models, *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. Influential models in this category include the Co-occurrence model [37], the cross media relevance model (CMRM) [22], CRM [26] and multiple Bernoulli relevance model (MBRM) [13].

One of the earliest examples of a probabilistic approach to image annotation is the co-occurrence model of Mori et al. [37]. The authors segment images using a simple regular grid and a probabilistic generative model is learnt based on the co-occurrence statistics of vocabulary keywords and the clusters derived from segmented image regions. In other early work, the authors of [11] 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. A notable feature of this approach is the association of words to image regions, in comparison to most other models which do not specify which image sub-structure gave rise to which word. The original pre-processed Corel 5K dataset first made available by [11] has become a widely used and popular benchmark of annotation systems in the literature. We denote the features arising from this dataset as the *Duygulu features* in this work.

This early work was subsequently advanced by the CMRM, a modification of relevance-based language mod-

els within information retrieval (IR) to the task of image annotation. The CMRM is a *discrete* analogue of the later CRM and MBRM models and relies on a vector quantisation step to cluster the set of image features to form a visual codebook. The probabilities of drawing a word and a "blob" (clustered image features) are computed using smoothed maximum likelihood estimates in this model. The CMRM demonstrated impressive gains in annotation accuracy versus the co-occurrence model.

The CRM [26] replaced the CMRM vector quantisation step with direct modelling of the continuous image features using non-parametric kernel density estimation with the Gaussian kernel. The probability of drawing a vocabulary keyword in this model was computed using a multinomial distribution. The avoidance of an error prone intermediate clustering step permitted the CRM to substantially outperform the CMRM on the benchmark Corel dataset, both for image annotation and retrieval. However, the use of non-parametric kernel density estimators placed over every training image requires large kernel distance matrices to be manipulated and stored, leading to a substantial increase in the computational load required for this model.

Feng et al. [13] later argued that the multinomial distribution was inappropriate for modelling the distribution of words for image annotation. Specifically, the multinomial distribution focuses on the *prominence* rather than the *presence* of words in the annotation. In benchmark datasets, a word only occurs at most once in the annotation of an image; therefore, modelling the frequency (prominence) of words is unnecessary. MBRM [13] tackled this issue by replacing the multinomial distribution with multiple Bernoulli models that naturally incorporate multi-keyword annotations. In addition, the authors partition each image into a regular grid and compute continuous image features over these regions. This latter technique avoids the computational expense of a dedicated image segmentation algorithm and provides the model with a larger set of image regions for learning the association between regions and words. The authors report an advancement in performance over the CRM. More specifically, a large boost in performance is realised by the image grid, with a more modest gain through the Bernoulli distribution. In a subsequent paper, [25] capture the benefits of the MBRM in the CRM by padding the annotations of each image to a fixed length in the so-called *normalised* CRM model.

A well-studied avenue of research has explored effective techniques for refining the labels produced by relevance modelling-based approaches to image annotation by capturing keyword correlation. For example, taking account of keyword correlation should make {*jungle, trees*} more plausible than {*jungle, snow*}. Wang et al. [46] use the CRM to capture keyword correlation of tags. The suggested greedy method involves adding successive tags to the set that have the largest joint probability with the tags already in the annotation. The authors of [55] build a graph with nodes representing the candidate annotations output by the CRM with weights reflecting the similarity between the nodes of the graph. The random walk with restarts algorithm is then applied to this fully connected graph to re-rank the candidate annotations. In later work, [50] incorporate the MBRM in a Markov random field framework (MRFA model) so as to capture the relationship between keywords. More recently, [34] extended the model of [46] in their BS-CRM model, demonstrating how the right combination of visual kernel and keyword correlation measure could yield improved annotation accuracy.

In other related work, Carneiro et al. [6] proposed the supervised multiclass labelling (SML) technique which aims to learn class conditional densities from the training data. The authors frame image annotation from a multiple instance learning perspective—in this paradigm, a model is learnt from positive and negative bags of examples, where a bag is a collection of localised image features. A bag is deemed a positive exemplar if only one of the examples in the bag is positive, and negative otherwise. The principle behind the annotation model is that, for a given image, the positive (keyword related) features follow a specific distribution while the negative features (pertaining to unrelated keywords) are uniformly distributed. By averaging the density estimates for a specific class, the keyword-related densities are reinforced while the non-keyword-based densities are curtailed. Images are modelled using a Gaussian mixture with model averaging performed using mixture hierarchies to yield the required class-conditional densities.

Topic model inspired image annotation models have been another popular area of research within the probabilistic modelling branch of the field. Annotated images are modelled as samples from a mixture of topics, with each topic being a distribution over visual features and words. In these models, an indirect association between blobs and words is derived via a latent topic space. Typically, a multinomial distribution is used for modelling the distribution of words and a Gaussian for visual features. Exact inference in these models is usually intractable with most approaches appealing to inference and learning approximations to find the topic mixture per image and the data distribution of the given topics.

The seminal work in this area can be traced back to the Correspondence Latent Dirichlet Allocation (CorrLDA) [5], a model that finds conditional relationships between latent variable representations of image regions and words. The authors demonstrate the flexibility of CorrLDA by evaluating against the tasks of automatic image annotation, automatic region annotation, and text-based image retrieval. This work was further described in [4] amongst several other models for annotated data. Unfortunately, sensitivity to model initialisation and the simplifying assumptions needed to make inference tractable, such as assuming that the generation

of an image region given a topic is Gaussian, have caused these models to somewhat lag behind the competition. In subsequent work, [51] introduced the MoM-HDP, a non-parametric generalization of the LDA model of [4,5] using a hierarchical Dirichlet process (HDP) that automatically adapted the number of clusters based on the training data.

In image annotation, the overall end goal is to find the conditional probability of a word given an image. This suggests that direct modelling of the decision boundary using discriminative approaches might be particularly useful for this task. In this scenario, the task of image annotation is framed as a classification problem where, for each word in the vocabulary, a binary decision is made as to whether it should appear as the annotation of a given image. Many types of discriminative models have been used for image annotation including support vector machines (SVMs) [52], random forest classifiers [14] and passive aggressive classifiers [15].

Grangier and Bengio [15] introduced a passive aggressive model for image retrieval (PAMIR) that directly optimises an image ranking loss inspired by the ranking SVM [24]. Fu et al. [14] explored the applicability of a random forest-based framework for image annotation. This model sorts the training dataset images into the leaf nodes of multiple random trees based on the associated visual features and keywords. For a given test image, training images that fall into the same leaf nodes across multiple trees form the so-called "semantic nearest neighbours", the top-k of which are used for labelling. SVMs have also been popular [10,52]. For example, in [52], a (one-vs-rest) SVM is adapted with a novel hinge loss so as to gain specific tolerance to what the authors dub "confusing labels", i.e., labels with similar semantics, for example {*flower*, *plant*} that should be treated as positive exemplars in learning the decision boundary. This so-called KSVM-VT model demonstrates competitive results to the current suite of state-of-the-art image annotation models.

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 [17], short for tag propagation. In this model, the weights of neighbouring images are based on 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. Rather than solely build a model off either global or local image descriptors, the authors of [17] introduced the now *de-facto* standard multiple-feature image annotation dataset. This dataset consists of 15 visual features ranging from local shape descriptors to global colour histograms providing a powerful standard feature set for annotation. We refer to this feature set as the *Tagprop features* in this work.

In subsequent related work, Zhang et al. [54] applied group sparse coding to the feature set of [17] in their group sparse (GS) image annotation model. The authors demonstrated that by carefully pruning non-informative or redundant features, image annotation accuracy can be further increased over and above the performance of [17].

The current state-of-the-art model for image annotation is the two-pass k-nearest neighbour (2PKNN) model of [45]. Two key ingredients contribute to the success of this model: dealing with the severe keyword frequency imbalance inherent in the benchmark annotation datasets and maximally leveraging the visual modality by learning a weighted combination of base distances *and* features. To achieve keyword balance, a unique and more balanced training dataset, referred to as a semantic neighbourhood, is crafted per test image based on the visual similarity of a test image to the training dataset images. The optimal weighted combination of feature distances and individual feature dimensions is derived through a multi-label extension to the large-margin nearest neighbour (LMNN) framework of [48]. The 2PKNN model demonstrates impressive levels of annotation accuracy through its ability to effectively exploit the visual and textual modalities.

3 Background

The CRM [26] 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 $\mathbf{w} = \{w_1 \ldots w_K\}$ from a vocabulary of size V together with an image $\mathbf{f}$ represented as a set of feature vectors $\mathbf{f} = \{\mathbf{f}_1 \ldots \mathbf{f}_M\}$. 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 (Eq. 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:

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

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 $P(\mathbf{f}_j|J)$ is modelled with a kernel-based density estimator:

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

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

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

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

4 The SKL-CRM model

In this section, we present the proposed sparse kernel learning (SKL) framework for the CRM, dubbed the SKL-CRM. Our method consists of *three* parts: investigation into a method for promoting the probability of rare tags in the context of relevance modelling (Sect. 4.1), a greedy optimisation algorithm for finding a pairing of features to kernels (Sect. 4.2) and the application of the generalised Gaussian (Sect. 4.3.1) and multinomial kernel (Sect. 4.3.2) to model the distribution of image features.

4.1 Promoting the probability of rare words

Many benchmark image datasets have a substantial imbalance between the frequent and rare keywords in the vocabulary (Fig. 1). This imbalance causes many image annotation models to bias their prediction to the more dominant keywords. In the case of the CRM, this means that, when predicting the keywords for a novel test image, the rare keywords do not appear within the nearest neighbours of the test image with sufficient weight and hence have a very low likelihood of being assigned to the image. All state-of-the-art image annotation models have a mechanism for promoting the probability of rare words. For example, [17] train word-specific logistic sigmoid models, one per word while [45] craft a more balanced semantic neighbourhood for each test image.

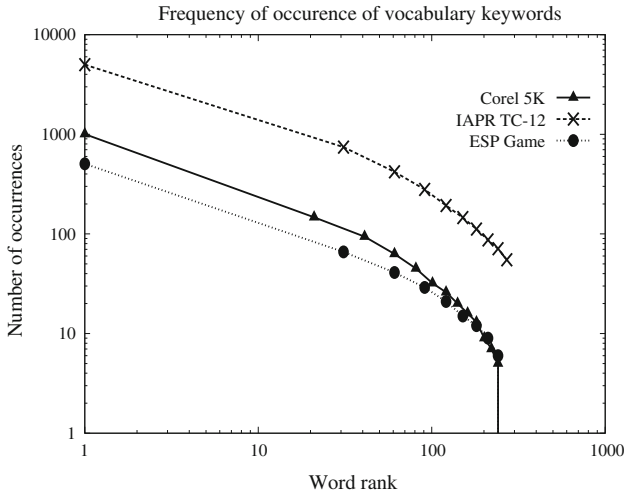


Fig. 1 Zipfian distribution of vocabularly keywords in the benchmark image datasets (log-log scale). The relative differences between the IAPR TC-12 versus ESP-Game and Corel 5K become important in Sect. 5.3.1

In this work, we explore a simple technique, *max–min normalisation*, for promoting the probability of rare words in the context of relevance modelling (Eq. 6).

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

Equation 6 regularises the SKL-CRM conditional probability $P(\mathbf{w}|\mathbf{f})$ (Eq. 1) and is an adaptation of a similar technique first suggested by [45] for application to local-learning-based models. The lower the occurrence of a word in the training dataset, the greater is the amplification given by Eq. 6 to the probability of that word's presence in each testing image. This technique tempers the over-confident prediction, ensuring that rare words have a greater chance of appearing as the annotation of an image.[2] While conceptually simple, our experiments show that this method of promoting the probability of rare words is just as effective as more computationally expensive techniques such as learning a logistic sigmoid model per word class as in [17].

4.2 Kernel-feature alignment algorithm

4.2.1 Problem overview

Recent image annotation models employ the feature set introduced by [17], which consists of a mixture of local (SIFT, robust hue) and global (GIST, colour histograms) image fea-

[2] In preliminary experiments, we also found that *z-score* normalisation has a similar effect, but for simplicity we report the max–min normalisation results in this paper.

tures. Previous work uses 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 [17]. 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 [26] model, dubbed the 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 (Sect. 4.3.3), Hellinger kernel (Sect. 4.3.3) and also two *data-adaptive* kernels: the generalised Gaussian (Sect. 4.3.1) and our proposed *multinomial kernel* for count-based image features (Sect. 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 $\Psi \in \Pi$ that specifies an optimal alignment between the two sets (Eq. 7).

$$\Pi := \left\{ \Psi \in \{0,1\}^{A \times B} \quad \text{and} \quad \forall i \sum_j \Psi_{ij} = 1 \right\} \tag{7}$$

The alignment matrix Ψ specifies a mapping between elements of our feature set and kernel set. We find the best alignment Ψ^* by directly optimising the quality of the image annotations it yields (Sect. 4.2.2).

An intuitive overview of our proposed algorithm is depicted in Fig. 2. In this toy example, we wish to compute the similarity between two images based on four different feature types (GIST, SIFT, LAB, HSV). The contents of the optimal alignment matrix Ψ are also shown for clarity. To compute the similarity between the image of the tiger and the image of the city scene, we need to compute a kernel function on the corresponding image features. Our greedy algorithm finds a kernel per feature type that leads to a local maximum in the image annotation quality. For example, in this setup a Laplacian kernel is aligned with the LAB colour feature which is indicated by a 1 in the appropriate cell of

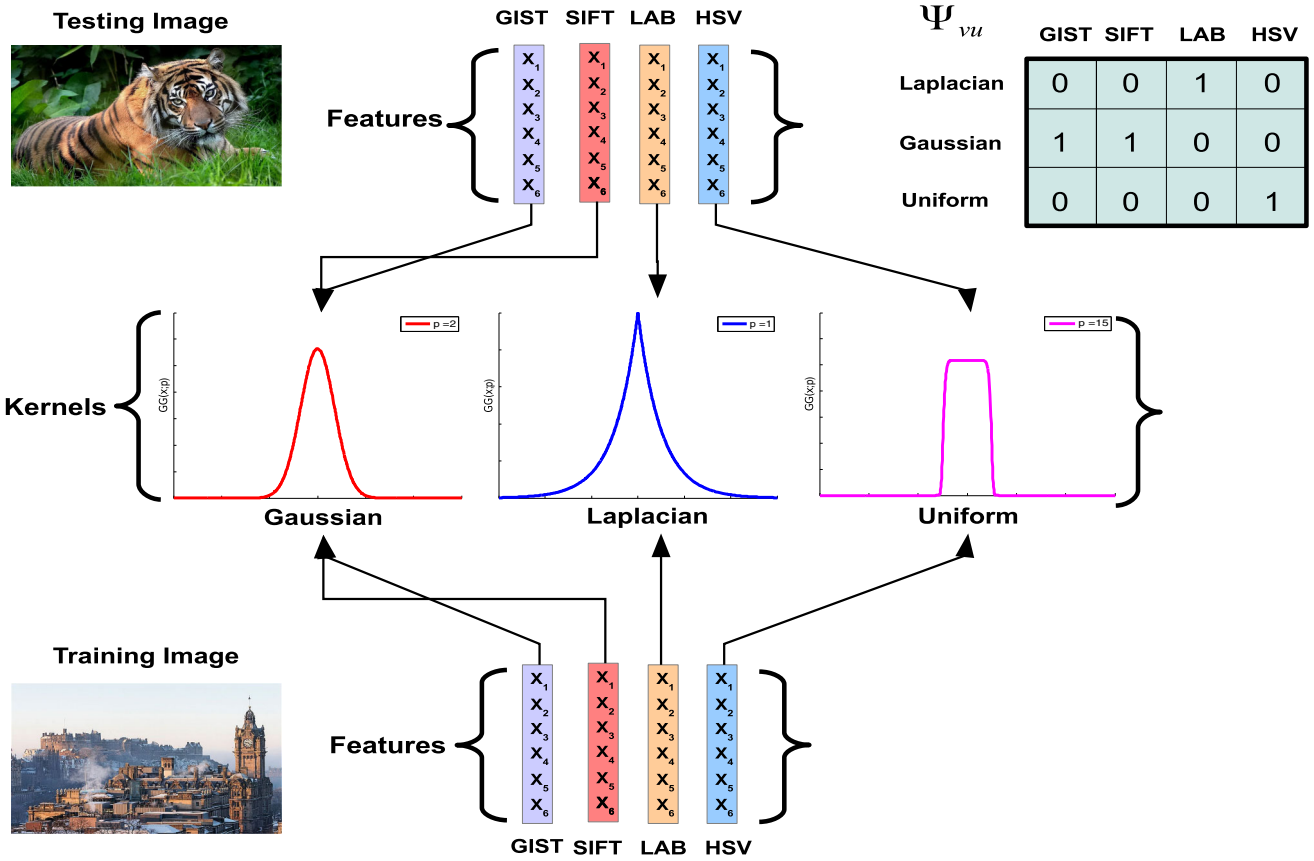


Fig. 2 Illustration of the greedy kernel-feature alignment algorithm on a toy example. See text for a description

matrix Ψ. The algorithm terminates when all available feature types have been aligned to a kernel.

4.2.2 Optimising annotation F_1 score

Rather than optimise a convenient objective such as the log-likelihood [17], 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 five keywords per validation image that have the highest $\hat{P}(\mathbf{w}|\mathbf{f})$ (Eq. 6) with the visual feature probability $P(I|J)$ given as in Eq. 8.

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

Here, $k^v(\mathbf{f}_i^u, \mathbf{f}_j^u)$ denotes the vth kernel function operating on the uth feature type. Equation 8 is a principled generalisation of the CRM visual feature probability (Eq. 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 Eq. 9.

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

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

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

The function `promote(.)` applies Eq. 6 to each element of the label prediction matrix $\mathbf{P}_\Psi$, $\mathbf{W} \in \Re^{N_J \times V}$ 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 $P(J|I)$ given by Eq. 11.

$$\hat{\mathbf{S}} = \exp\left\{\mathbf{S} - \left\{\mathbf{Z} \times \mathbf{1}_{1\times N_J}\right\}\right\} \tag{11}$$

Here, $\mathbf{S} \in \Re^{N_I \times N_J}$ is the matrix of image-image likelihoods $log\{P(I|J)\}$ and $\mathbf{Z} \in \Re^{N_I \times 1}$ represents a normalization vector which is also computed in log-space (Eq. 12).

$$\mathbf{Z}_{ij} = \log\left\{\sum_{J\in T}\exp\{\log\{P(I|J)\}\}\right\} \tag{12}$$

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 (Eq. 8) and μ (Eq. 3) are optimised individually as each new feature-kernel combination is considered for addition to the set. We observe rapid convergence to a local optimum typically only after five feature-kernel combinations have been added to the set (Sect. 5.3). To lighten the computational load at runtime, we pre-compute the set of $A \times B$ kernel matrices so that they may be simply looked up during the optimisation procedure.

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 (Sect. 4.3.1) and kernels better able to model discrete count-based features (Sects. 4.3.2, 4.3.3).

4.3.1 Generalised Gaussian kernel

We investigate replacing the Gaussian kernel in Eq. (5) with a generalised exponential kernel based on the Minkowski p-norm. This kernel is similar to the Minkowski kernel of [34] and is also known as the generalised Gaussian in the statistics community.[3] We will argue that the proposed kernel is more sensitive to subtle changes in the visual appearance

[3] We use Minkowski kernel and generalised Gaussian interchangeably to refer to the same kernel in this work.

Algorithm 1 Greedy kernel-feature alignment algorithm

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

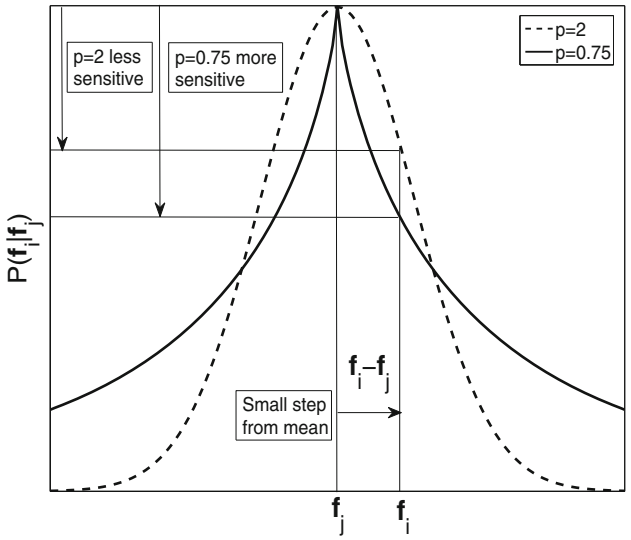


Fig. 3 The generalised Gaussian ($p = 0.75$) kernel is more sensitive to small changes in one feature compared to the Gaussian ($\mathtt{p} = 2$) kernel. If we take a small step from the mean, the $p = 0.75$ kernel will undergo a larger change in output as compared to the $p = 2$ kernel

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:

$$P(\mathbf{f}_i|\mathbf{f}_j) = \frac{\mathtt{p}^{1-1/\mathtt{p}}}{2\beta\Gamma(1/\mathtt{p})} \exp\left[-\frac{1}{\mathtt{p}}\frac{|\mathbf{f}_i - \mathbf{f}_j|^{\mathtt{p}}}{\beta^{\mathtt{p}}}\right], \tag{13}$$

Here, $\Gamma(\cdot)$ denotes the gamma function and $|\mathbf{f}_i - \mathbf{f}_j|^{\mathtt{p}} = \sum_{d=1}^{D} |f_{i,d} - f_{j,d}|^{\mathtt{p}}$ is a generalisation of the Euclidean norm. The summation goes over the dimensions d of the feature vectors while $\mathtt{p}$ 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. The normalising constant ensures that the kernel integrates to one but is not required when implementing this kernel in the CRM given that it drops out of the equation during the conditional probability computation (Eq. 1).

Figures 3 and 4 highlight the difference between the familar Gaussian kernel and the generalised Gaussian kernel. The generalised Gaussian kernel for fractional values of $\mathtt{p}$ is much more sensitive to the change in a feature value compared to the Gaussian kernel. Figure 3 illustrates this most clearly: if we take the same step away from the mean ($\mathbf{f}_i - \mathbf{f}_j$), the output of the generalised Gaussian kernel (in this case with $\mathtt{p} = 0.75$) will be correspondingly larger compared to the output of the Gaussian kernel. This is further illustrated in Fig. 4. Here, the Gaussian density on the left is concave around the mean, which makes it insensitive to small differences between the training and testing feature regions. The generalised Gaussian (Minkowski) kernel in the middle is convex (for $\mathtt{p} < 1$), allowing it to sense subtle differences in feature values in a way that mimics the operation of the human visual system [19].

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 Fig. 4 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 A and three testing images B, C, 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 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 generalised Gaussian kernel (for $\mathtt{p} < 1$) 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 generalised Gaussian kernel is closer to a logical AND.

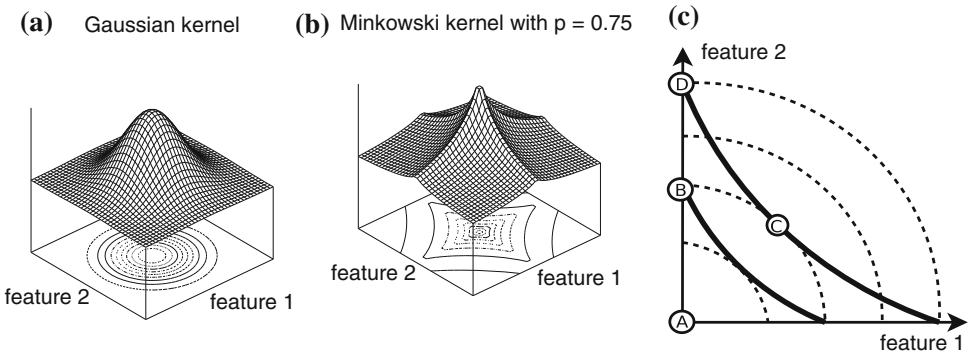


Fig. 4 **a** Density functions and equidistant contours for the Gaussian kernel. **b** The generalised Gaussian (Minkowski) kernel for $p = 0.75$. **c** The Minkowski kernel is particularly sensitive when multiple feature values change at the same time (point C), whereas the Gaussian is more sensitive to large variations in any one feature (point D)

4.3.2 Multinomial kernel

The generalised Gaussian kernel described in the previous section forms a flexible and powerful family of distributions for modelling the real-valued image features, such as the Corel features of Duygulu et al. [11]. However, as we argue in this section, this kernel is not appropriate for modelling *count-based* features, which are becoming more prevalent for describing the visual appearance of images. A particularly relevant example of *count-based* descriptors are the Tagprop features [17], which can be seen as a concatenation of 15 different types of histograms.

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

$$P(\mathbf{f}_i|\mathbf{f}_j) = \frac{\Gamma(\sum_d f_{i,d} + 1)}{\prod_d (\Gamma(f_{i,d} + 1))} \prod_d (p_{j,d})^{f_{i,d}} \tag{14}$$

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:

$$p_{j,d} = \lambda \frac{f_{j,d}}{\sum_d f_{j,d}} + (1 - \lambda) \frac{\sum_j f_{j,d}}{\sum_{j,d} f_{j,d}} \tag{15}$$

The smoothing parameter λ is optimized on a held-out portion of the training set.

We believe that there are two reasons why the generalised Gaussian kernel is not appropriate for modelling count data: (1) it is probabilistically deficient [9] and (2) it tends to underestimate low and zero counts. We discuss both of these reasons below, and refer to Fig. 5 for a visual illustration.

(i) *Model deficiency* The generalised Gaussian kernel (of which the popular Gaussian kernel is a special case) allocates the probability mass over both positive and negative numbers. However, negative numbers cannot possibly result from count-based observation. When the kernel bandwidth β is large in relation to the mean $\mathbf{f}_j$, a significant proportion of the probability density will be wasted on events that can never be observed in our data (*negative counts*). As we increase the number of features (dimensions) in our vectors $\mathbf{f}$, the probability mass allocated to non-negative count vectors becomes vanishingly small, so the model is increasingly deficient. On the other hand, the multinomial kernel assigns all the probability mass to observable events (non-negative counts).

(ii) *Low counts* Count-based observations often follow power–law behaviours and are typically distributed with a positive skew. The generalised Gaussian kernel is naturally symmetric (zero-skew), so it will always underestimate the likelihood of low-positive counts. The multino-

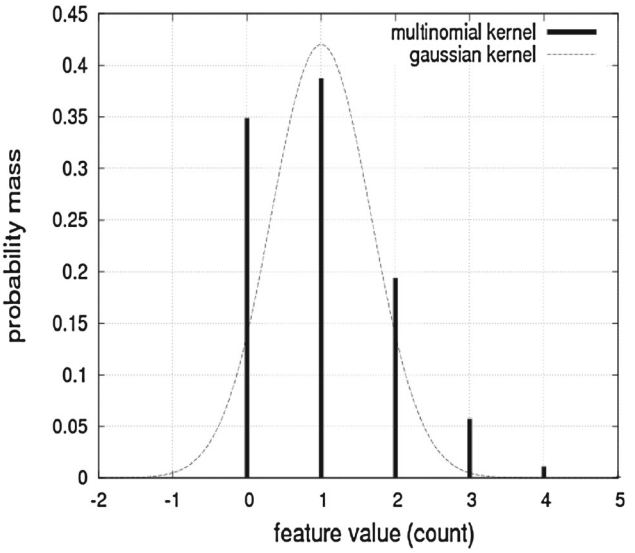


Fig. 5 A comparison of the multinomial kernel against a Gaussian (p-norm) kernel with the same mean and variance. The Gaussian kernel underestimates the likelihood of low counts (e.g. zero count), and devotes a significant amount of its mass to impossible observations (negative counts), forming a probabilistically deficient model of the data

mial kernel has a positive skew, and will assign a higher probability to low and zero counts (see Fig. 5 for a count of zero).

In summary, we believe that the multinomial kernel offers a superior way of modelling histogram-based feature vectors, because it is specifically designed for discrete observations (counts), does not suffer from model deficiency and properly estimates 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 (Eq. 16).

$$k(\mathbf{f}_i, \mathbf{f}_j) = \sum_d \sqrt{f_{i,d} f_{j,d}} \tag{16}$$

for two L_1 normalised feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ (i.e. $\sum_d f_{i,d} = 1$ and $f_{i,d} \geq 0$). In addition, we also consider the χ^2 kernel (Eq. 17).

$$k(\mathbf{f}_i, \mathbf{f}_j) = \sum_d \frac{2 f_{i,d} f_{j,d}}{f_{i,d} + f_{j,d}} \tag{17}$$

Both kernels are commonly used for computing histogram distance due to their higher sensitivity to smaller bin values as compared to the Gaussian kernel [3].

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 [17,45], thereby permitting direct comparison. The statistics of the datasets are summarised in Table 2.

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 with an average of 3.4 keywords per image. In our evaluation, a fixed set of 500 images are used for testing with the remaining 4,500 images being used for training. This split corresponds to previous related work [17].

ESP Game was originally built by Von Ahn and Dabbish [1] from images on the Internet. The images are diverse and cover topics such as logos, drawings, personal photos and web page decorations (Table 1). The quality of the images is also highly variable making this a particularly challenging dataset for automatic annoation. There are 20,770 images in the dataset the annotations of which were 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 subset of images as [45]. There are 18,689 images in the training dataset and 2,081 in the test dataset, with an average of 4.7 keywords per image.

IAPR TC-12 is a collection of 19,627 images of natural scenes that include different sports and actions, photographs of people, animals, cities and landscapes from South America (Table 1). This dataset does not exhibit variety to the same extent as the ESP Game dataset. IAPR TC-12 was initially published for cross-lingual retrieval [16] with each image being accompanied by descriptions in several languages. The raw dataset was originally processed by [30] for use in image annotation evaluation by extracting common nouns in English. The resulting vocabulary consists of 291 keywords, with an average of 5.7 keywords per image. There are 17,665 training images with the remaining 1,962 being used for testing.

5.2 Experimental methodology

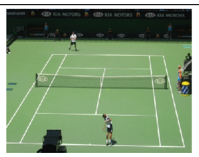
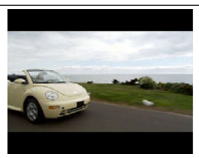
To fairly compare our model performance to previously published figures we use the identical feature set, parameter optimisation strategy and evaluation procedure of previous relevant work [17,26,45].

5.2.1 Features

We use, without modification, the feature set introduced by [17] 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 [29] and local hue histograms [47] 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 [40] 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[4] [27]. In all, there is one GIST descriptor, six colour histograms and eight bag-of-features.

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

Table 1 A selection of example test images from the IAPR TC-12 and ESP Game datasets

IAPR TC-12					
					
Racing (1.00) Spectator (0.55) **Car** (0.45) **Tree** (0.44) Man (0.39)	Tree (0.31) **House** (0.20) Pond (0.18) Palm (0.17) People (0.12)	**Tree** (0.29) **Palm** (0.17) Sky (0.14) House (0.13) **People** (0.13)	**Camera** (1.00) **Tennis** (0.73) **Court** (0.73) **Player** (0.73) **Chair** (0.61)	**Grave** (0.39) **Mummy** (0.37) Pot (0.36) **Brick** (0.32) **Bone** (0.18)	Cup (0.37) Lookout (0.30) Group (0.26) **Front** (0.26) Tourist (0.23)
ESP Game					
					
Sky (0.27) **Grass** (0.22) **Mountain** (0.18) **Green** (0.12) Hill (0.09)	Person (0.51) **Hat** (0.26) Boy (0.26) Sky (0.22) Nose (0.19)	Shop (0.28) Anime (0.18) **Man** (0.18) **Woman** (0.12) **People** (0.11)	**Pole** (1.00) **Grass** (0.75) **Dog** (0.50) White (0.50) Man (0.00)	**Road** (0.59) **Sky** (0.36) **Car** (0.30) **Grass** (0.29) **Yellow** (0.28)	**Blue** (0.25) Logo (0.13) Circle (0.11) White (0.11) Letter (0.11)

Note the wide diversity of images between the two datasets. Below each image, we show the annotations produced our proposed SKL-CRM model. Words in bold case appear in the ground truth annotations and the regularised scores assigned by our model are shown in brackets. Notice that in many cases, our model predicts relevant words for an image but those words do not exist in the ground truth. This is the well-known problem of weak-labelling that plagues the benchmark image datasets

Table 2 Statistics of the Tagprop datasets used in our experimental evaluation

Dataset	# Images	# Labels	# Training images	# Testing images	Labels per image	Images per label
Corel 5K	5,000	260	4,500	500	3.4, 4, 5	58.6, 22, 1004
ESP Game	20,770	268	18,689	2,081	4.7, 5, 15	326.7, 172, 4553
IAPR TC-12	19,627	291	17,665	1,962	5.7, 5, 23	347.7, 153, 4999

In columns 6 and 7, the entries are in the format: *mean, median, maximum*. Adapted from a similar table in [45]

5.2.2 Parameter optimization

The parameter optimisation strategy is identical for each dataset. We are given a training dataset and a test dataset as is used in previous work. To make clear our parameter optimisation strategy, we denote the training dataset provided by [17] as the *full training dataset*. To create a validation dataset for parameter tuning, we split the full training dataset into two portions: a *reduced training dataset* and a validation dataset both of which are randomly sub-sampled from the full training dataset. The reduced training dataset and validation dataset are used to find the optimal kernel-feature combination (Sect. 4.2).

For Corel 5K, we use the full training dataset (4,500 images) to derive the reduced training dataset (4,000 images) and validation dataset (500 images). To ensure that the cross-validation is computationally tractable, on the larger ESP Game and IAPR TC-12 datasets we do not use the full training datasets (18,869 and 17,665 images, respectively) to derive the reduced training dataset and validation dataset. Rather, for finding the optimal parameters for our model variants on ESP Game and IAPR TC-12, we take a random sample of 5,000 images from the full training dataset in both cases and use 4,500 of these images as the reduced training dataset the remaining 500 as the validation dataset.

After fixing the parameters, we again use the full training dataset to compute the annotations on the test images. The training and testing dataset splits for all three datasets are identical to previously published work [17,45]. The final reported F_1 score is computed by taking the parameter configuration at the point where *validation* dataset F_1 score is maximised and then running that instance of the SKL-CRM model on the test dataset, reporting the resulting F_1 score.

5.2.3 Evaluation procedure

We follow the standard recall and precision-based evaluation paradigm extensively used in the literature [11]. In this scenario, we are given an unseen test 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 $\mathbf{w}_I$. Given a test image, we use the SKL-CRM algorithm to determine the five words with the highest conditional probability (Eq. 6) and call them the automatic annotation of the image in question. Then, following [11], 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. 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 Experimental results

In this section, we evaluate the performance of our model on the task of automatic image annotation. We examine *three* main hypotheses:

1. *HYP-1* Regularising the conditional probability using max–min normalisation is effective at improving the recall of rare words in the vocabulary.
2. *HYP-2* 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 [17,45].
3. *HYP-3* Greedy kernel-feature alignment is more effective than learnt distance weights for combining different disparate feature types for the purposes of image annotation.

In this section, we discuss a set of experiments that we carried out to test the hypotheses. For each experiment, we compute a measure of statistical significance—a *paired t test*—based on mean per word F_1 [43]. In other words, when comparing System A to System B, the harmonic mean of precision and recall is computed per word and the F_1 score of identical words from System A and B form the pairs for the t test. The t test permits us to eliminate random chance as being responsible for any observed increase in performance. We measure statistical significance at the 1 % level.

5.3.1 Effect of conditional probability regularisation

In Table 3, we demonstrate the effect of regularising the CRM conditional probabilities with max–min normalisation (Eq. 6). CRM (T) is the standard CRM model (standard kernel alignment, no conditional probability regularisation) learnt on the Tagprop feature set (denoted by T). CRM ($T + P$) denotes the same model but with conditional probability regularisation applied (denoted by P). We notice that across all three datasets the mean per word recall, precision and number of words with recall greater than zero all substantially increase after applying this regularisation. Importantly, the precision does not suffer at the expense of boosting the recall of the rarer words.

More specifically on the Corel 5K dataset, we find a 20 % relative increase in recall, a 14 % increase in precision and a 19 % increase in the number of words with recall greater than zero for CRM ($T + P$) versus CRM (T). To test the significance of this result, we compute a paired t test based on per word F_1 yielding a p value of $p \leq 0.00003$. CRM ($T+$

Table 3 Annotation performance scores for various incarnations of the CRM model

	COREL 5K				IAPR TC-12				ESP Game			
Model	R	P	F_1	N^+	R	P	F_1	N^+	R	P	F_1	N^+
CRM	19	16	17	107	–	–	–	–	–	–	–	–
CRM (T)	30	28	29	135	21	32	26	229	12	42	19	202
CRM ($T + P$)	36	32	$34^{\dagger}$	161	25	50	$34^{\dagger}$	266	16	42	$23^{\dagger}$	234
SKL-CRM	46	39	$42^{\ddagger}$	184	32	51	$39^{\ddagger}$	274	26	41	$32^{\ddagger}$	248

CRM is the original model as reported in [26] using the feature set of [11]. CRM (T) is the CRM model using all 15 tagprop-based features [17] (denoted with T) and default kernel selection. CRM ($T + P$) is the CRM model with Tagprop features (T) default kernel allocation and max–min conditional probability regularisation (P). SKL-CRM is our proposed model with max–min regularisation and the adaptive kernel allocation mechanism. There are two points to note regarding these results versus those in [35]. Firstly, through a more extensive parameter sweep, we find a slightly better local maxima for IAPR TC-12 leading to an F_1 of 39. Secondly for ESP Game, the F_1 of 25 for CRM (T) on ESP Game should in fact be an F_1 of 19. This erratum is corrected in the above table. Finally, † indicates that the result is statistically significant based on a paired t test versus CRM (T), while ‡ indicates statistical significant versus CRM ($T + P$)

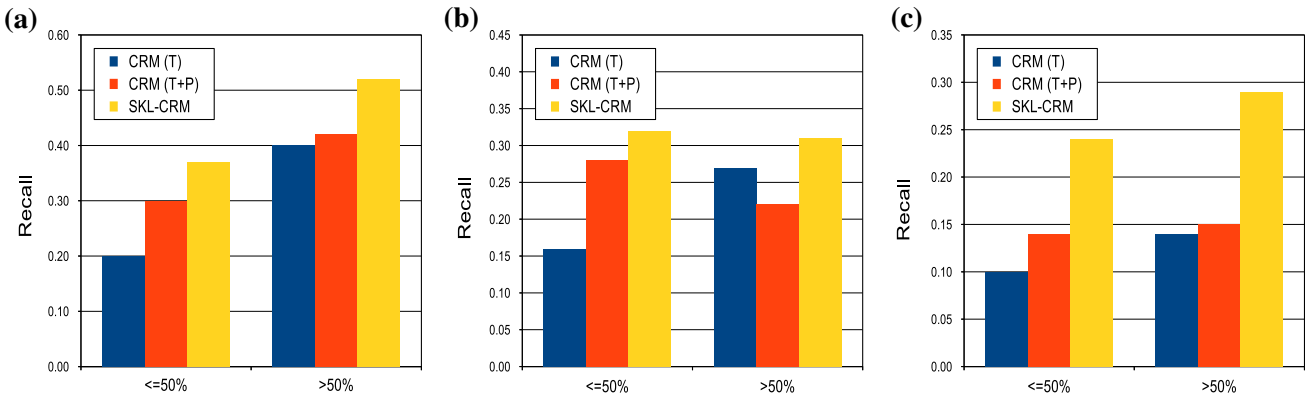


Fig. 6 Annotation quality measured in terms of mean recall (*y*-axis) for various incarnations of the CRM model. The *labels* are grouped based on their frequency in the dataset (*x*-axis). The first bin corresponds to the 50 % least frequent labels and the second bin corresponds to the 50 % most frequent labels

P), therefore, gives a statistically significant improvement in annotation quality versus CRM (T) on the Corel 5K dataset. For the IAPR TC-12 dataset, we find a 19 % relative increase in recall, a 56 % increase in precision and a 16 % increase in the number of words with recall greater than zero for CRM ($T + P$) versus CRM (T). The increase in mean per word F_1 is statistically significant based on a paired t test p value: $p \leq 0.0001$. Lastly for the ESP Game dataset, we also observe a statistically significant (p value: $p \leq 1.0 \times 10^{-7}$) increase in per word F_1 with increases of 33 % for recall and 16 % for the number of words with recall greater than zero.

This experiment and the statistical significance of the gains in F_1 score suggest that it can be beneficial to perform max–min normalisation on the conditional probabilities arising from a relevance model learnt on keyword imbalanced image datasets. The impressive annotation quality gains resulting from this regularisation are particularly appealing given the simplicity of the method. For example, we avoid the computational expense of having to learn one logistic regressor per keyword as in [17] but nevertheless attain similar annotation quality.

To gain further insight into the effect of the regularisation mechanism, we follow [45] and compute the mean recall of the 50 % most and least frequent words in the vocabulary both before and after applying min-max regularisation. The results of this experiment are shown in Fig. 6. Partitioning the vocabulary in this manner demonstrates that for Corel 5K (Fig. 6a) and ESP Game (Fig. 6c) the regularisation [CRM ($T + P$)] is both benefitting the recall of the rarer words and, to a lesser extent, the more frequent words in the vocabulary.

In the case of IAPR TC-12 (Fig. 6b), regularisation considerably improves the recall of the rarer words (by 75 %) but at the detriment of the recall on the more frequent words (which falls by 19 %). The effect of regularisation appears worse for the more frequent words on IAPR TC-12. A possible explanation arises from Fig. 1—from this chart it is clear that words considered frequent for Corel 5K and ESP Game (e.g. frequency 50 …200) are in fact infrequent for IAPR TC-12. Therefore, we expect that the fall in mean recall for more frequent words on IAPR TC-12 is due to where (50 % most frequent and least frequent words) we decided to split the vocabularly in Fig. 6.

Across all three datasets, we can see, however, that the SKL-CRM model, combining both the regularisation for rare words and the adaptive kernel allocation, obtains the highest recall across both the rarer words and the more frequent words. We examine the annotation performance of the SKL-CRM model in more detail in Sect. 5.3.2. Given these results we can confirm our first hypothesis (HYP-1) that max–min normalisation is an effective technique for regularising the over-confident conditional probabilities so as to lend more weight to the rarer words in the vocabularly.

5.3.2 Standard versus data-driven kernel assignment

We now turn to our second hypothesis (HYP-2) that the standard allocation of kernels to feature types is sub-optimal. As a reminder, it is accepted practice in the literature to compute similarities for SIFT and hue histograms using a χ^2 kernel while a Gaussian kernel is used for GIST and a Laplacian kernel for global colour histogram-based features. We refer to this alignment as the *standard alignment* and challenge the notion that it is optimal for all datasets in this particular experiment.

Our results are presented in Table 3 where we show the proposed model, the SKL-CRM, measured against three baselines: the original CRM model with Duygulu features [26], the CRM model using the full 15 Tagprop-based features and standard kernel assignments [CRM (T)] and the CRM model with max–min conditional probability regularisation and standard kernel assignment [CRM ($T + P$)]. It should be noted that our proposed model, the SKL-CRM, combines both greedy kernel-feature selection and max–min conditional probability regularisation. From this table,

we firstly observe, across all three benchmark datasets, that the SKL-CRM outperforms the three CRM baselines across all three evaluation metrics. For example, on the Corel 5K dataset, the SKL-CRM attains a substantial 147 % increase in annotation F_1 over the original CRM model. Against the CRM model on the same set of features, CRM (T), the SKL-CRM realises a 45 % increase in annotation F_1 measure.

There are three factors that contribute to the increased performance of the SKL-CRM: (1) the Tagprop feature set, (2) the max–min conditional probability mechanism, and (3) the adaptive kernel allocation algorithm. It is insightful to ascertain the proportion of this performance gain that arises from each of these factors. Comparing CRM to CRM (T) we can clearly see the effect of the different feature sets on annotation performance: it is obvious that the Tagprop-based features are a more powerful set of features than those of Duygulu et al. [11]—simply using the CRM with these features, we obtain a substantial increase (71 %) in annotation F_1 over the CRM with Duygulu features on Corel 5K. This observation further lends weight to the notion that to fairly ascertain the performance of a new model it must be learnt on the same feature set as previously published models so as to zero out the effect of a more powerful set of features.

We will examine the Corel 5K dataset first: if we compare CRM (T) to the CRM ($T + P$) model on this dataset we can isolate the effect of the conditional probability regularisation technique (Eq. 6). Here, we find an increase in annotation performance over CRM (T) across all three evaluation metrics as has been previously noted in some detail in Sect. 5.3.1. In a similar manner, if we compare CRM ($T + P$) to the SKL-CRM, we can isolate the effect of the greedy kernel-feature alignment algorithm. For the Corel 5K dataset, the SKL-CRM obtains a 24 % increase in F_1 measure versus CRM ($T + P$) which is statistically significant p value: $p \leq 3.0 \times 10^{-7}$. This result demonstrates the effectiveness of our proposed greedy algorithm. In the case of the Corel 5K dataset, it is interesting to note that 38 % of the total increase in F_1 arises from the max–min regularisation while the remaining 62 % is due to picking the best kernel per feature type. Using both techniques (max–min normalisation, adaptive kernel allocation) together in the SKL-CRM yields the best overall annotation quality.

We now provide a breakdown of the performance on the larger IAPR TC-12 and ESP Game datasets: as can be observed in Table 3, the max–min regularisation yields a statistically significant increase in annotation F_1 of 31 % for IAPR TC-12 and 21 % for ESP Game. Comparing CRM ($T + P$) to the SKL-CRM in Table 3, we observe that the adaptive kernel allocation again yields a further increase in annotation quality: recall for IAPR TC-12 increases by 28 % (63 % for ESP Game), precision by 2 % and the number of words with recall greater than zero by 3 % (6 % for ESP Game) over CRM ($T + P$). For both datasets the increase in per word F_1 is statistically significant: IAPR-TC 12 p value: $p \leq 8.0 \times 10^{-18}$ and ESP Game p value: $p \leq 2.0 \times 10^{-29}$.

For IAPR TC-12, we can attribute 62 % of the gain in F_1 measure of the SKL-CRM versus the CRM (T) to max–min normalisation while 38 % of the gain arrives from the adaptive kernel-feature alignment. In this case, properly accounting for the rare vocabulary keywords is marginally more important than exploiting the visual modality. In contrast, for ESP Game 31 % of the increase in F_1 for the SKL-CRM arises from the regularisation while 69 % comes from the adaptive kernel allocation mechanism. As for the Corel 5K dataset, the highest annotation quality for IAPR TC-12 and ESP Game is attained using both max–min normalisation and adaptive kernel allocation together in the form of the SKL-CRM model. These results suggest that both the visual modalities need to be maximally exploited by an adaptive kernel allocation mechanism in tandem with a method which ensures that the rarer keywords are not suppressed by the more frequent keywords in the vocabulary. It is clear that using either method alone is *not* sufficient for best performance as has been previously noted in related work [17,45]. This fact is most vividly demonstrated by the statistically significant increase in annotation quality of the SKL-CRM versus CRM ($T + P$) and CRM (T) across all three benchmark image datasets.

We have so far determined that adaptive kernel allocation can lead to gains in image annotation quality. To fully confirm our second hypothesis, we must ascertain the identity of the kernel-feature alignments that give rise to the SKL-CRM performance and compare these alignments to the standard alignment advocated in the literature. Table 4 lists the optimal feature-kernel alignments found by our greedy algorithm for all three benchmark image datasets. The observed kernel-feature alignments provide two interesting conclusions: firstly, we note the prevalence (4 out of 6 on Corel 5K, 2 out of 4 on IAPR TC-12 and 5 out of 9 for ESP Game) of our proposed data-driven kernels amongst the alignments, including our proposed multinomial kernel. This observation indicates that data-adaptive kernels are much more effective than standard kernels for computing the similarities between the visual features. While clearly useful for image annotation, given their inherent generality, we envisage that such kernels will also find much wider application to other areas of Computer Vision (and beyond) where feature similarity needs to be computed.

Secondly and perhaps more importantly, 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, therefore, confirm our second hypothesis (HYP-2) that the standard feature-kernel assignment advocated in the literature is sub-

Table 4 Optimal kernel-feature alignments for the three benchmark image datasets

Feature	Dataset		
	Corel 5K	IAPR TC12	ESP Game
RGB	–	GG (0.70)	–
RGB_V3H1	–	–	–
LAB	–	–	MK (0.99)
LAB_V3H1	–	–	–
HSV	MK (0.99)	–	LP
HSV_V3H1	GG (0.90)	–	LP
HH		–	–
HH_V3H1	GG (0.10)	–	GG (0.10)
HS	GA	GG (0.70)	LP
HS_V3H1	GG (0.70)	–	–
DH	–	–	–
DH_V3H1	–	–	GG (0.10)
DS	LP	χ^2	GG (0.70)
DS_V3H1	–	–	GA
GIST	–	LP	GG (0.70)

MK multinomial kernel, *GG* generalised Gaussian kernel, *GA* Gaussian kernel, *LP* Laplacian kernel. The parameter settings for the given data-adaptive kernel are given in brackets

optimal and better performance can be realised by an adaptive allocation.

5.3.3 Greedy optimisation algorithm performance

In this section, we test our third and final hypothesis (HYP-3) namely that adaptively assigning kernels to features is more effective than taking a weighted sum of standard kernel distances. The latter technique is employed in all previous related works to combine the distances resulting from multiple different feature types [17,45] and it is, therefore, instructive to examine how our proposed adaptive kernel-feature alignment strategy fairs in comparison. To examine this hypothesis, we firstly investigate the optimisation profile of our greedy kernel-feature alignment algorithm as proposed in Sect. 4.2.3.

In Fig. 7a, for Corel 5K, we show the progress of our greedy optimisation algorithm as each new feature-kernel alignment is added to the set. Each point on the *x*-axis proceeding from left to right indicates the feature that has been selected at that particular iteration of the algorithm: so, for example, the colour histogram HSV_V3H1 is an important feature for Corel 5K given that it has been selected first. Each subsequent feature can be considered the next most important feature, and so forth. In this way, our algorithm can essentially be interpreted as an instance of a greedy feature selection mechanism.

We observe from Fig. 7a that 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, and quite remarkably, with just *two* features the SKL-CRM reaches 90 % performance, surpassing Tagprop σ-ML. This trend is repeated on the IAPR TC-12 and ESP Game datasets where we also find sparse optimal solutions: for IAPR TC-12, only 4 features are required to reach the maximum annotation F_1, whereas 9 features are required for ESP Game. We also show in Fig. 7b the value

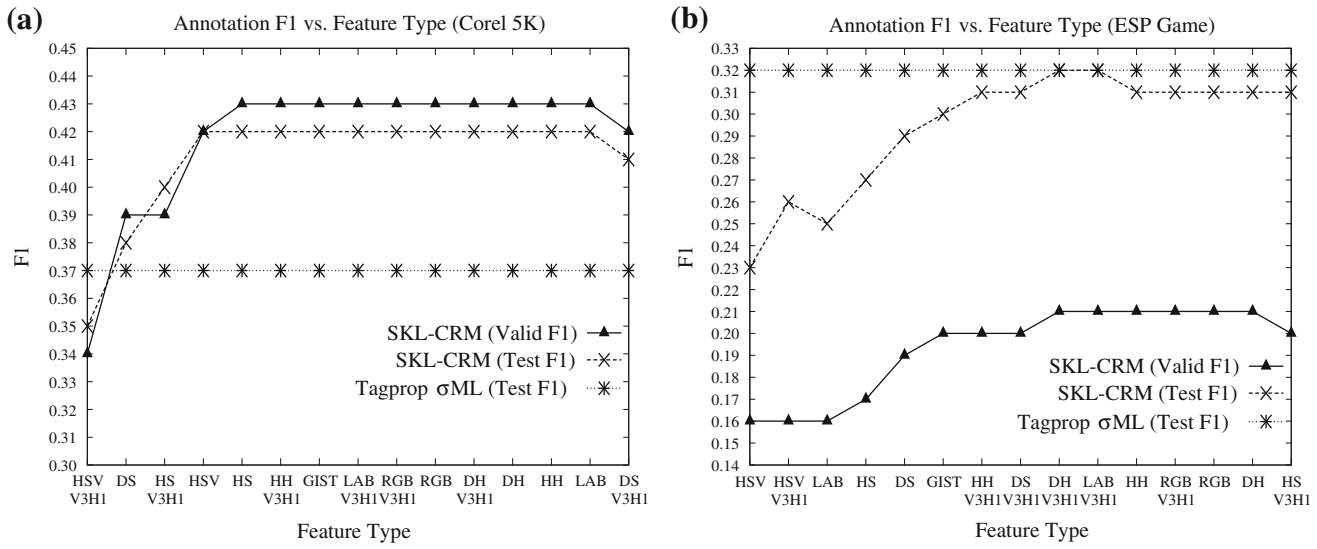


Fig. 7 **a** Corel 5K annotation F_1 score versus the contents of the feature set. **b** ESP Game annotation F_1 score versus the contents of the feature set. Note that in **b** the validation dataset F_1 score is substantially lower than the test dataset F_1 due to the sub-sampling—see Sect. 5.2.2

Table 5 Annotation performance resulting from adaptive weight allocation ($T + P + W$) versus adaptive kernel allocation (SKL-CRM)

Model	COREL 5K				IAPR TC-12				ESP Game			
	R	P	F_1	N^+	R	P	F_1	N^+	R	P	F_1	N^+
CRM ($T + P + W$)	39	37	38	166	28	52	37	271	23	41	30	248
SKL-CRM	46	39	42†	184	32	51	39†	274	26	41	32†	248

The symbol † indicates that the result is statistically significant based on a paired t test versus CRM ($T + P + W$)

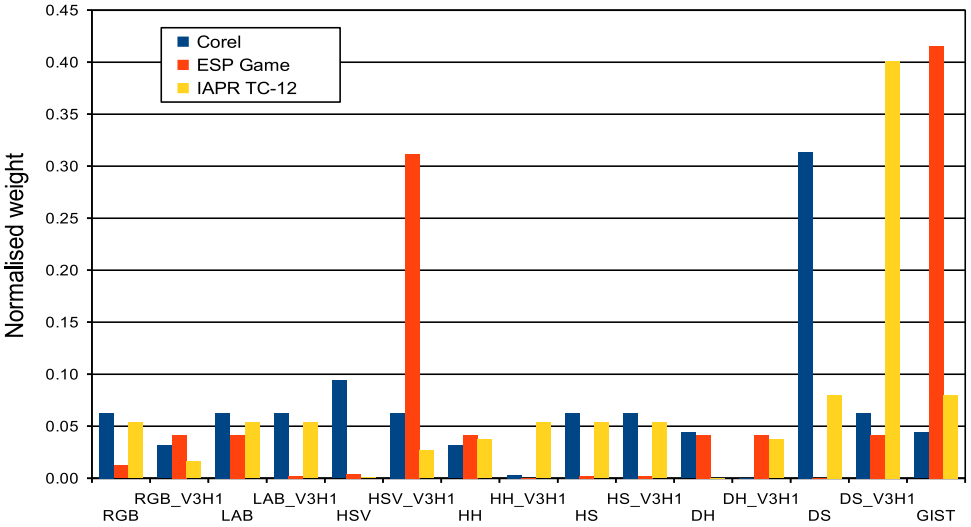


Fig. 8 The normalised values of weights found by coordinate descent for each of the 15 Tagprop features. These weights are used to combine the distances arising from the standard kernels in a weighted summation in a similar manner to the Tagprop model [17]

of F_1 measure versus the set of selected features on the ESP Game dataset while Table 4 lists the optimal alignments for all three datasets. These results demonstrate that further features are detrimental and our greedy optimisation algorithm is able to effectively identify a *sparse subset* of features that jointly maximise annotation performance.

We now investigate how taking a weighted sum of the distances arising from the *standard* feature-kernel assignment performs with respect to our proposed data-adaptive kernel allocation. Table 5 presents the results of this experiment. CRM ($T + P + W$) denotes the CRM model with conditional probability regularisation and an aggregate kernel distance derived as a weighted summation of the distances arising from the standard assignment of kernels. The weights for each kernel were learnt by coordinate descent optimisation based on maximisation of the annotation F_1 measure on a held-out validation dataset. To ensure a fair comparison, the validation datasets were identical to those used for the adaptive-kernel feature alignment algorithm. To mitigate the effect of local minima, we used five randomly selected coordinate sweep patterns and selected the run that led to the maximum annotation F_1 on the validation dataset.

The results of this experiment are shown in Table 5. We observe that the SKL-CRM outperforms the adaptive weighting scheme across all three benchmark datasets in terms of annotation F_1 measure, with a 11 % relative increase on Corel 5K and a more modest 5 % increase of IAPR TC-12 and 7 % increase on ESP Game. We test the statistical significance of these results based on a paired t test of per-word F_1. For the SKL-CRM versus adaptive weighting [CRM ($T + P + W$)] on Corel 5K, we find a p value: $p \leq 0.0003$, for IAPR TC-12 a p value: $p \leq 5 \times 10^{-9}$ and for ESP Game a p value: $p \leq 0.005$. We confirm our third hypothesis (HYP-3) that the SKL-CRM adaptive kernel assignment mechanism is a more effective means of exploiting the visual modality as compared to a weighted summation of the distances arising from a set of standard kernels.

In addition, we show in Fig. 8 the value of the *normalised weights* per feature type across the three datasets. From this chart, we can see the features that are most important for each dataset: for example, for Corel 5K the colour histograms (particularly HSV) and Dense SIFT are given a high weighting. Interestingly, the greedy kernel-feature alignment algorithm (Fig. 7a) similarly selects both Dense SIFT and the spatial variant of HSV, HSV_V3H1, as the first two features, and hence two most important features, for this dataset. For IAPR TC-12, the Dense SIFT and GIST appear to have a high weighting suggesting they are particularly important for performance on this dataset. Our greedy algorithm also finds both of these features to be within the set of four most important features. Finally, for ESP Game, both GIST and the HSV_V3H1 colour histogram are assigned the highest

weight and we can see in Fig. 7b that our greedy algorithm also considers these features key to performance on ESP Game.

The natural question arises as to whether or not taking a weighted summation of the distances arising from the adaptive kernels can yield increased performance over and above the SKL-CRM model. In other words, would a combination of adaptive kernels and adaptive weights reap higher annotation quality? Unfortunately, we found that the annotation F_1 score from both techniques combined was indistinguishable from simply using adaptive kernel allocation alone. Given the near equivalence of the features chosen by the greedy algorithm and those features assigned a high weight in Fig. 8, we believe that the kernels themselves are acting in some sense as weights on the features, either up-weighting the effect of a feature type that is added in the initial early stages of the optimisation procedure, while down-weighting the contribution of those features added towards the end of the optimisation.

More specifically, 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.

5.3.4 SKL-CRM performance versus the literature

In Table 6, we present the annotation performance of the SKL-CRM against a broad selection of image annotation models recently proposed in the literature. The models we compare to span the full range of different model types covered in Sect. 2—from generative to discriminative to local-learning-based models. Encouragingly, across all three standard image annotation benchmark datasets, we find that the SKL-CRM either decisively outperforms or is competitive to a wide range of existing models. This demonstrates that our model not only outperforms previous incarnations of the CRM but also models employing many other techniques from the field of machine learning.

For example, on the Corel 5K dataset, we improve recall by 9.5 %, precision by 18 % and the number of words with recall greater than zero by 15 % with respect to the strong baseline of Tagprop σ-ML. Tagprop σ-ML is 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 [17]. Our superior perfor-

Table 6 Performance of the SKL-CRM model against a wide range of recent annotation models on three benchmark image annotation datasets (Corel, IAPR TC-12 and ESP game)

Model	COREL 5K				IAPR TC-12				ESP Game			
	R	P	F_1	N^+	R	P	F_1	N^+	R	P	F_1	N^+
CRM [26]	19	16	17	107	–	–	–	–	–	–	–	–
MBRM [13]	25	24	25	122	23	24	23	223	19	18	18	209
InfNet [32]	24	17	20	112	–	–	–	–	–	–	–	–
NPDE [53]	21	18	19	114	–	–	–	–	–	–	–	–
BS-CRM [34]	27	22	24	130	22	24	23	250				
SML [6]	29	23	26	137	–	–	–	–	–	–	–	–
TGLM [28]	29	25	27	131	–	–	–	–	–	–	–	–
JEC [30]	32	27	29	139	29	28	28	250	25	22	23	224
Tagprop SD [17]	33	30	31	136	20	50	29	215	19	48	27	212
MRFA [50]	36	31	33	172	–	–	–	–	–	–	–	–
GS [54]	33	30	31	146	29	32	30	252	–	–	–	–
RF-opt [14]	40	29	34	157	31	44	36	253	26	41	32	235
CCD (SVRMKL+KPCA) [38]	41	36	38	159	29	44	35	251	24	36	29	232
KSVM-VT [52]	42	32	36	179	29	47	36	268	32	33	33	259
FastTag [8]	43	32	37	166	26	47	34	280	22	46	30	247
Tagprop ML [17]	37	31	34	146	25	48	33	227	20	49	29	213
Tagprop σML [17]	42	33	37	160	35	46	40	266	27	39	32	239
SKL-CRM (this work)	46	39	42	184	32	51	39	274	26	41	32	248

Table 7 Comparison of the SKL-CRM model against the current state-of-the-art model (2PKNN)

Model	COREL5K				IAPRTC-12				ESPGame			
	R	P	F_1	N^+	R	P	F_1	N^+	R	P	F_1	N^+
2PKNN [45]	40	39	40	177	32	49	39	274	23	51	32	245
2PKNN-ML [45]	46	44	45	191	37	54	44	278	27	53	36	252
SKL-CRM (this work)	46	39	42	184	32	51	39	274	26	41	32	248

mance to Tagprop σ-ML on this dataset further 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 6 also presents results on the IAPR TC-12 and ESP Game datasets. We find that the SKL-CRM is also competitive to recently proposed models on these two much larger datasets. For example, on IAPR TC-12 we obtain very similar performance to Tagprop σ-ML, while decisively outperforming the strong benchmark random forest model (RF-opt) [14] by 8 % F_1, the FastTag model [8] by 15 % F_1 and the more recently proposed SVM-based annotation model (KSVM-VT) [52] by 8 % F_1. On ESP Game results are competitive to Tagprop σ-ML, RF-opt and KSVM-VT while outperforming FastTag by 7 % F_1.

The state-of-the-art image annotation model is currently the two-pass KNN (2PKNN) model of [45]. As touched upon in Sect. 2, this model employs a large-margin nearest neighbour metric learning algorithm to learn weights on both the features and distances while balancing the distribution of words by inducing a unique, balanced training set, per test image. While 2PKNN achieves impressive levels of annotation quality (Table 7), the methods employed for exploiting the visual and textual modalities exhibit some disadvantages. For the textual modality, a test image-specific subset of the training dataset has to be constructed for every test image leading to an increase in the computational demands when annotating novel images. The SKL-CRM uses a single training dataset for all test images eliminating this issue. In addition, for the visual modality, the 2PKNN metric learning algorithm requires multiple random initialisations (five are used in the original work) making it computationally expensive to obtain the best quality annotations—in contrast our greedy algorithm is deterministic. Furthermore, it is not clear how sparse the weights are in 2PKNN and whether a significant proportion of the features can be discounted. Our greedy algorithm explicitly targets sparsity by greedily finding the best performing subset of features. This sparsity substantially reduces the complexity of our model at test time while also outperforming other sparsity-based annotation models such as the model of [54]. Finally, as our greedy algorithm is coordinate descent based it is, therefore, straightforward to parallelise (namely lines 7–15 in Algorithm 1) for application to larger image datasets [41].

6 Conclusions

In this paper, we introduced a Sparse Kernel Learning (SKL) framework for the 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 able to reach an impressive level of annotation performance 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 four further several 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 the 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. Thirdly, in the context of relevance model-based image annotation, we found that a data-adaptive kernel-feature alignment was more effective than taking a weighted sum of distances arising from the standard set of kernels. Lastly, we found no additional benefit in learning a weighted combination of the optimal kernel-feature alignments.

There are a number of fruitful avenues for future research in this area. Firstly, the SKL-CRM is currently limited to a single kernel per feature type. We would like to investigate a continuous relaxation of this discrete alignment constraint that permits more than one kernel to be assigned to a given feature. In addition, the SKL-CRM alignment is currently identical for every keyword in the vocabulary. Having a spe-

cific alignment of kernels to feature types for each word in the vocabulary would be an interesting route for future research.

Secondly, the SKL-CRM is limited to fairly small image datasets of the order of 20k images. This limitation arises from the need to store and manipulate large kernel distance matrices arising from placing a non-parametric kernel density estimator over every training image. At the same time, multi-million image datasets such as ImageNet are becoming popular in the literature [49]. To annotate at such scale approximate nearest neighbour (ANN), search techniques could be employed to more efficiently find the k-nearest neighbours of each test image [21,36]. It would be interesting to determine the accuracy scalability trade-off resulting from a hashing-based ANN search method.

Lastly, it is becoming more apparent that the Tagprop feature set of [17], while being powerful in its own right, has been exploited to its near-fullest by the current breed of annotation models. A new, more powerful, set of image features may be needed now so as to push the mean per word recall and precision on the three standard benchmark datasets to higher levels. For example, it would be interesting to explore mid-level or high-level structured image features (in terms of parts for example) that capture semantic concepts beyond the basic visual cues offered by low-level features such as SIFT [33].

Acknowledgments We thank the anonymous reviewer for their helpful comments.

References

1. von Ahn L, Dabbish L (2005) Esp: labeling images with a computer game. In: AAAI spring symposium: knowledge cfrom volunteer contributors, pp 91–98
2. Ames M, Naaman M (2007) Why we tag: motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '07ACM, New York, NY, USA, pp 971–980
3. Arandjelovic R, Zisserman A (2012) Three things everyone should know to improve object retrieval. In: CVPR. IEEE, New York, pp 2911–2918
4. Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM, Jordan MI (2003) Matching words and pictures. J Mach Learn Res 3:1107–1135
5. Blei DM, Jordan MI (2003) Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval, SIGIR '03ACM, New York, NY, USA, pp 127–134
6. Carneiro G, Chan AB, Moreno PJ, Vasconcelos N (2007) Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans Pattern Anal Mach Intell 29(3):394–410
7. Chapelle O, Haffner P, Vapnik VN (1999) Support vector machines for histogram-based image classification. Trans Neural Netw 10(5):1055–1064
8. Chen M, Zheng A, Weinberger KQ (2013) Fast image tagging. In: Dasgupta S, Mcallester D (eds) Proceedings of the 30th international conference on machine learning (ICML-13), vol 28, pp 1274–1282. JMLR workshop and conference proceedings
9. Cooper WS (1995) Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. ACM Trans Inf Syst 13(1):100–111
10. Cusano C, Ciocca G, Schettini R (2003) Image annotation using SVM. In: Santini S, Schettini R (eds) Internet imaging V, society of photo-optical instrumentation engineers (SPIE) conference Series, vol 5304, pp 330–338
11. Duygulu P, Barnard K, de Freitas JFG, Forsyth DA (2002) Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Proceedings of the 7th European conference on computer vision-part IV, ECCV '02. Springer, London, pp 97–112
12. Enser P, Sandom C, Lewis P (2005) Automatic annotation of images from the practitioner perspective. In: Image and video retrieval, pp 497–506
13. Feng SL, Manmatha R, Lavrenko V (2004) Multiple bernoulli relevance models for image and video annotation. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, CVPR'04. IEEE Computer Society, Washington, DC, pp 1002–1009
14. Fu H, Zhang Q, Qiu G (2012) Random forest for image annotation. In: Proceedings of the 12th European conference on computer vision, , ECCV'12, vol Part VI. Springer, Berlin, pp 86–99
15. Grangier D, Bengio S (2008) A discriminative kernel-based approach to rank images from text queries. IEEE Trans Pattern Anal Mach Intell 30(8):1371–1384. doi:10.1109/TPAMI.2007.70791
16. Grubinger M (2007) Analysis and evaluation of visual information systems performance. PhD thesis, School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University, Melbourne, Australia
17. Guillaumin M, Mensink T, Verbeek J, Schmid C (2009) Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation. In: International conference on computer vision, pp 309–316
18. Hentschel C, Stober S, Nrnberger A, Detyniecki M (2007) Automatic image annotation using a visual dictionary based on reliable image segmentation. In: Adaptive multimedia retrieval. Lecture Notes in Computer Science, vol 4918. Springer, Berlin, pp 45–56
19. Howarth P, Rüger S (2005) Fractional distance measures for content-based image retrieval. In: Proceedings of the 27th European conference on advances in information retrieval research, ECIR'05. Springer, Berlin, pp 447–456
20. Huang J, Kumar SR, Zabih R (1998) An automatic hierarchical image classification scheme. In: Proceedings of the Sixth ACM international conference on multimedia, MULTIMEDIA '98. ACM, New York, pp 219–228
21. Indyk P, Motwani R (1998) Approximate nearest neighbors: Towards removing the curse of dimensionality. In: Proceedings of the thirtieth annual ACM symposium on theory of computing, STOC '98. ACM, New York, pp 604–613
22. Jeon J, Lavrenko V, Manmatha R (2003) Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in Information retrieval, SIGIR '03. ACM, New York, pp 119–126
23. Jeon J, Manmatha R (2004) Using maximum entropy for automatic image annotation. In: CIVR. Lecture Notes in Computer Science, vol 3115. Springer, Berlin, pp. 24–32
24. Joachims T (2002) Optimizing search engines using clickthrough data. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '02. ACM, New York, pp 133–142
25. Lavrenko V, Feng S, Manmatha R (2004) Statistical models for automatic video annotation and retrieval. ICASSP 3:1044–1047

26. Lavrenko V, Manmatha R, Jeon J (2003) A model for learning the semantics of pictures. NIPS
27. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE computer society conference on computer vision and pattern recognition, CVPR '06, vol 2. IEEE Computer Society, Washington, DC, pp 2169–2178
28. Liu J, Li M, Liu Q, Lu H, Ma S (2009) Image annotation via graph learning. Pattern Recognit 42(2):218–228
29. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
30. Makadia A, Pavlovic V, Kumar S (2008) A new baseline for image annotation. In: Proceedings of the 10th European conference on computer vision: part III, ECCV '08. Springer, Berlin, pp 316–329
31. Markkula M, Sormunen E (2000) End-user searching challenges indexing practices in the digital newspaper photo archive. Inf Retr 1(4):259–285
32. Metzler D, Manmatha R (2004) An inference network approach to image retrieval. In: Proceedings of the international conference on image and video retrieval. Springer, Berlin, pp 42–50.
33. Mittelman R, Lee H, Kuipers B, Savarese S (2013) Weakly supervised learning of mid-level features with beta-bernoulli process restricted boltzmann machines. In: Proceedings of the 2013 IEEE conference on computer vision and pattern recognition, CVPR '13. IEEE Computer Society, Washington, DC, pp 476–483
34. Moran S, Lavrenko V (2011) Optimal tag sets for automatic image annotation. In: Proceedings of the British machine vision conference. BMVA Press, London, pp 1.1–1.11
35. Moran S, Lavrenko V (2014) Sparse kernel learning for image annotation. In: Proceedings of international conference on multimedia retrieval, ICMR '14. ACM, New York, pp 113:113–113:120
36. Moran S, Lavrenko V, Osborne M (2013) Variable bit quantisation for lsh. In: Proceedings of the 51st annual meeting of the association for computational linguistics (vol 2: short papers). Association for Computational Linguistics, Sofia, pp. 753–758
37. Mori Y, Takahashi H, Oka R (1999) 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
38. Nakayama H (2011) Linear distance metric learning for large-scale generic image recognition. PhD thesis, The University of Tokyo, Japan
39. Oliva A, Schyns P (2000) Diagnostic colors mediate scene recognition. Cogn Psychol 41(2):176–210
40. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
41. Richtárik P, Takác M (2013) Distributed coordinate descent method for learning with big data. In: CoRR'13
42. Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905
43. Smucker MD, Allan J, Carterette B (2007) A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the sixteenth ACM conference on information and knowledge management, CIKM '07. ACM, New York, pp 623–632
44. Ulz MH, Moran SJ (2013) Optimal kernel shape and bandwidth for atomistic support of continuum stress. Model Simul Mater Sci Eng 21(8):085, 017
45. Verma Y, Jawahar CV (2012) Image annotation using metric learning in semantic neighbourhoods. In: Proceedings of the 12th European conference on computer vision, ECCV'12, vol Part III. Springer, Berlin, pp 836–849
46. Wang B, Li ZW, Yu N, Li M (2007) Image annotation in a progressive way. In: Proceedings of ICME, pp 811–814
47. van de Weijer J, Schmid C (2006) Coloring local feature extraction. In: Proceedings of the 9th European conference on computer vision, ECCV'06, vol Part II. Springer, Berlin, pp 334–348
48. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10:207–244
49. Weston J, Bengio S, Usunier N (2010) Large scale image annotation: learning to rank with joint word-image embeddings. Mach Learn 81(1):21–35
50. Xiang Y, Zhou X, Unviersity F, seng Chua T, wah Ngo C (2009) A revisit of generative model for automatic image annotation using markov random fields. In: Proceedings of IEEE computer vision and pattern recognition, pp 1153–1160
51. Yakhnenko O, Honavar V (2008) Annotating images and image objects using a hierarchical dirichlet process model. In: Proceedings of the 9th international workshop on multimedia data mining: held in conjunction with the ACM SIGKDD 2008, MDM '08. ACM, New York, pp 1–7
52. Yashaswi Verma CJ (2013)Exploring svm for image annotation in presence of confusing labels. In: Proceedings of the British machine vision conference. BMVA Press, London
53. Yavlinsky A, Schofield E, Rüger S (2005) Automated image annotation using global features and robust nonparametric density estimation. In: Proceedings of the 4th international conference on image and video retrieval, CIVR'05. Springer, Berlin, pp 507–517
54. Zhang S, Huang J, Huang Y, Yu Y, Li H, Metaxas DN (2010) Automatic image annotation using group sparsity. In: CVPR. IEEE, New York, pp 3312–3319
55. Zhu S, Liu Y (2008) Image annotation refinement using semantic similarity correlation. In: ICPR'08