Image Retrieval using Markov Random Fields and Global Image Features

Ainhoa Llorente
Knowledge Media Institute
The Open University
Milton Keynes, MK7 6AA
United Kingdom
a.llorente@open.ac.uk

R. Manmatha
Department of Computer Science
University of Massachusetts
Amherst, MA, 01003
manmatha@cs.umass.edu

Stefan Rüger
Knowledge Media Institute
The Open University
Milton Keynes, MK7 6AA
United Kingdom
s.rueger@open.ac.uk

ABSTRACT

In this paper, we propose a direct image retrieval framework based on Markov Random Fields (MRFs) that exploits the semantic context dependencies of the image. The novelty of our approach lies in the use of different kernels in our non-parametric density estimation together with the utilisation of configurations that explore semantic relationships among concepts at the same time as low-level features, instead of just focusing on correlation between image features like in previous formulations. Hence, we introduce several configurations and study which one achieve the best performance. Results are presented for two datasets, the usual benchmark Corel 5k and the collection proposed by the 2009 edition of the ImageCLEF campaign. We observe that, using MRFs, performance increases significantly depending on the kernel used in the density estimation for the two datasets. With respect to the the language model, best results are obtained for the configuration that exploits dependencies between words together with dependencies between words and visual features. For the Corel 5k dataset, our best result corresponds to a mean average precision of 0.32, which compares favourably with the highest value ever obtained, 0.35, achieved by Makadia et al. [22] albeit with different features. For the ImageCLEF09 collection, we obtained 0.32, as mean average precision.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object Recognition*

General Terms

Algorithms, Design, Experimentation

Keywords

Markov processes, Nonparametric statistics

CIVR '10, July 5-7, Xi'an China

1. INTRODUCTION

Automated image annotation refers to the process of learning statistical models from a training set of pre-annotated images in order to generate annotations for unseen images using visual feature extracting technology. This can be formulated in two ways, as direct image retrieval or as image annotation itself. Despite the fact that the work presented in this paper refers to direct retrieval, we discuss previous work done under both formulations.

The problem of modelling annotated images has been addressed from several directions in the literature. Initially, a set of generic algorithms were developed with the aim of exploiting the dependencies between image features and implicitly between words. However, many algorithms do not explicitly exploit the correlation between words. With respect to the deployed machine learning method, we can consider: co-occurrence models of low-level image features and words [26]; machine translation methods that translate image regions into words in the same way as words from French might be translated into English [6]; relevance models CRM [16], CMRM [12], and MBRM [10]; inference networks that connect image segments with words [24]; non-parametric density estimation [34]; supervised learning models [3, 4]; information-theoretic semantic indexing [21]; and [22] show that a proper selection of features could lead to very good results for a k-nearest neighbours algorithm.

The human understanding of a scene was a topic confronted by many researchers in the past. Authors like Biederman [1], and then, Torralba and Oliva [32] supported the hypothesis that objects and their containing scenes were not independent. For example, the prediction of the concept "beach" is usually followed by the presence of "water" and "sand". On the other hand, a "polar bear" should never appear in a "desert" scenario, no matter how high the probability of the prediction. As a result of this, a new collection of algorithms, devoted to exploring word-to-word correlations, shortly emerged. Thus, these methods relied on either filtering the results obtained by a previous baseline annotation method or on creating adequate language models as a way to boost the efficiency of previous approaches. In particular, some of them use co-occurrence information [19, 35]; others apply semantic measures to WordNet like [6, 17] or combine co-occurrence information with WordNet [18] and others build a concept hierarchy [29, 31] or use WordNet to induce hierarchies on annotation words [13]. However, the improvement in the performance obtained by word-to-word correlation methods built on top of individual concept detec-

tors might be hindered by error propagation of the baseline classifiers and by the lack of sufficient data, which can lead to over-fitting.

Nevertheless, Markov Random Fields (MRFs) provide a convenient way of modelling context-dependent entities like image content. This is achieved through characterizing mutual influences among such entities using conditional MRF distributions. The main benefit of using a MRF comes from the fact that we can model correlations between words explicitly. In this paper, we present a direct image retrieval framework that makes use of different configurations to model the image content. Besides that, the application of MRF theory allows us to easily formulate the joint distribution of the graph. The novelty of our approach lies in the use of different kernels, in our non-parametric density estimation, together with the utilisation of configurations that explore semantic relationships among concepts and low-level features instead of just focusing on correlation between image features like in previous formulations. Our focus is on the model and on obtaining a better kernel estimation. As Makadia et al. [22] show, a good choice of features can give very good results. Here our focus is not on the features. We use simple global features.

Tables 1 and 2 show comparative results for some of the previous approaches mentioned in Section 1 and in Section 2. However, we have only included those results obtained using the Corel 5k dataset (Section 4.1) and the evaluation measures considered in this paper (Section 4.3). Depending on the strategy adopted, direct retrieval or image annotation, the evaluation measures used are mean average precision (MAP) or the number of words with non-zero recall (NZR), precision (P), and recall (R). Section 2 discusses state-of-the-art automated image annotation algorithms. Section 3 introduces our Markov Random Field model. Section 4 explains the experiments undertaken, while Section 5 analyses our results. Finally, Section 6 explains the conclusions and plans for future work.

Table 1: Best performing automated image annotation algorithms expressed in terms of number of words with non-zero recall (NZR), recall (R), and precision (P) for the Corel 5k dataset. The first block represents classic probabilistic models, the second is devoted to the new generation of algorithms that incorporate language models, and the third depicts models based on MRF. Algorithms in each block are ordered according to the increasing value of F1. The evaluation is done using 260 words that annotate the test data. Algorithms marked with an asterisk use additional training data from an external corpus. (-) means numbers not available

Model	Author	NZR	R	P
CRM	Lavrenko et al. [16]	107	19	16
Npde	Yavlinsky et al. [34]	114	21	18
InfNet	Metzler&Manmatha [24]	112	24	17
CRM-Rect	Feng et al. [10]	119	23	22
MBRM	Feng et al. [10]	122	25	24
SML	Carneiro et al. [3]	137	29	23
Manifold	Loeff et al. [20]	-	40	21
JEC	Makadia et al. [22]	113	40	32
Anno-Iter	Zhou et al. [35]	-	18	21
CLM	Jin et al. [13]	-	21	18
ONT-500*	Srikanth et al. [31]	163	25	15
BHMMM*	Shi et al. [30]	153	34	16
DCMRM*	Liu et al. [19]	124	25	22
TMHD*	Jin et al. [14]	-	21	30
MRFA-region	Xiang et al. [33]	124	23	27
MRFA-grid	Xiang et al. [33]	172	36	31

2. RELATED WORK

Markov Random Fields have been widely used in computer vision applications to model spatial relationships between pixels. Feng and Manmatha [9] were the first to do direct retrieval (without an intermediate annotation step) using a MRF model. By ranking while maximising average precision the model is simplified due to the fact that the normaliser does not need to be calculated. They used discrete image features and obtained comparable results to the state-of-the arts algorithms. Later on, Feng presented a similar model [8] but applied it to the case of continuous image features. He achieved better performance with the continuous model than with the discrete model although the latter was more efficient in terms of speed. Both models were based on the Markov Random Field framework developed by Metzler and Croft [23], who modelled term dependencies in text retrieval. The novelty of their approach lies in training the model that maximises directly the mean average precision instead of maximising the likelihood of the training data.

Escalante et al. [7] proposed a MRF model as part of their image annotation framework, which additionally uses word-to-word correlation. Hernandez-Gracidas and Sucar [11] carried out another variation of the previous approach placing emphasis on the spatial information relation among objects. Both works are based on the MRF model proposed by Carbonetto et al. in [2], whose approach is considered to be out of the scope of this work as it is more aligned with the approaches usually adopted in the field of computer vision.

More recently, Xiang et al. [33] presented a new approach able to perform directly automated image annotation. They adopt a MRF to model the context relationships among semantic concepts with keyword subgraphs generated from training sample for each keyword. Thus, they defined two potential functions in cliques up to order two: the site potential and the edge potential. The former models the joint probability of an image feature and a word and was modelled using the Multiple Bernoulli Relevance Model (MBRM) [10]. The edge potential approximates the joint probability of an image feature and a correlated word. The parameter estimation is done adopting a pseudo-likelihood scheme in order to avoid the evaluation of the partition function. Finally, they showed significant improvement over six previous approaches for the Corel 5k dataset.

In [28], Qi et al. follow a similar approach applying a MRF to video annotation. Their method, the Correlative Multi-Label (CML) framework, simultaneously classifies concepts while modelling the correlations between them in a single step. They conduct their experiments on TRECVID 2005 dataset outperforming several algorithms.

3. MARKOV RANDOM FIELDS

For the basic Markov Random Field model we followed the approach and the notation used by Feng [8]. Let G

be an undirected graph whose nodes are called I and Q. A Markov Random Field (MRF) is an undirected graph G which allows the joint distribution between its two nodes to be modelled in terms of:

$$P_\Lambda(I,Q) = \frac{1}{Z_\Lambda} \prod_{c \in C(G)} \psi(c;\Lambda), \quad (1)$$

where $C(G)$ is the set of cliques defined in the graph G, $\psi(c;\Lambda)$ is a non-negative potential function over clique configurations parametrized by Λ, and Z_Λ is the value that normalised the distribution. When applied to the image retrieval case, the nodes of the graph, I and Q, represent respectively a image of the test set and a query. The image is represented by a set of feature vectors r and the query by a set of words w. Following the same reasoning as Feng in his continuous model developed in [8], we approximate the joint distribution using the following exponential form:

$$\psi(c;\Lambda) = e^{\lambda_c f(c)}. \quad (2)$$

Therefore, we arrive at the following model where images are ranked according to their posterior probability:

$$P_\Lambda(I|Q) \stackrel{rank}{=} \sum_{c \in C(G)} \lambda_c f(c), \quad (3)$$

where $f(c)$ is a real-valued feature function defined over the clique c weighed by λ_c.

Figure 1 shows a graph representing the dependencies explored in our model. The left side of the image illustrates the clique configurations considered in this research which contemplates cliques of up to third order. A 2-clique (r-w) consisting of a query node w and a feature vector r, followed by a 2-clique (w-w') representing the dependencies between words w and w', and, finally a 3-clique (r-w-w') capturing the relation between a feature vector r and two word nodes w and w'.

According to the graph, the posterior probability is expressed as:

$$P_\Lambda(I|Q) \stackrel{rank}{=} \sum_{c \in T} \lambda_T f_T(c) + \sum_{c \in U} \lambda_U f_U(c) + \sum_{c \in V} \lambda_V f_V(c), \quad (4)$$

where T is the set of 2-cliques containing a feature vector r and a query term w, U is the set of 2-clique (w-w') representing the dependencies between two words w and w' and V is the set of 3-cliques (r-w-w') capturing the relation between a feature vector r and two word nodes w and w'. Finally, and for simplicity, we make the assumption that all image features are independent of each other given some query Q.

The differences between this work and [8, 9] reside mainly in the divergent associations defined in our respective graphs. Both approaches investigate the dependencies between image regions and words (configuration r-w). However, their focus is on exploring the dependencies between various image regions while ours relies on the relationships between words. Thus, the rest of the configurations presented in this paper are new. Another differing point is that we work with feature vectors extracted from the entire image instead of with image regions. Additionally, both works differ on their selection of visual features. Finally, [8, 9] employ a Gaussian kernel in their density estimation while our strongest point is exploring additional kernels such as the "square-root" or the Laplacian kernel.

In what follows, we explain in detail the different configurations followed in this research.

3.1 Image-to-Word Dependencies

This configuration is formed by the set of 2-cliques r-w and it corresponds to the *Full Independence Model* developed by Feng in [8]. The potential function associated to this clique expresses the probability of generating the word w, for a given image feature, scaled by the prominence of the feature vector r in the test set image I, as shown in:

$$f_T(c) = P(w|r)P(r|I), \quad (5)$$

where $P(r|I)$ is set to be the inverse of number of features vector per image, as we make the assumption that the distribution is uniform. $P(w|r)$ is estimated applying Bayes' rule:

$$P(w|r) = \frac{P(w,r)}{\sum_w P(w|r)}, \quad (6)$$

where $P(w,r)$ is computed in a similar way to the continuous relevance model (CRM) developed by Lavrenko et al. [15]:

$$P(w,r) = \sum_{J \in \tau} P(J)P(w|J)P(r|J), \quad (7)$$

where τ represents the training set and J, a training image, and $P(J) \approx \frac{1}{|J|}$.

The function $P(r|J)$ is estimated using a non-parametric density estimation approach as represented in:

$$P(r|J) = \frac{1}{m} \sum_{t=1}^{m} k\left(\frac{|r - r_t|}{h}\right), \quad (8)$$

where r is a real-valued image feature vector of dimension d, m is the number of feature vectors representing the image J, t is an index over the set of biagrams in J. We propose as kernel function a Generalized Gaussian Distribution [5] whose probability density function (pdf) is defined as:

$$\text{pdf}(x;\mu,\sigma,p) = \frac{1}{2\Gamma(1+1/p)A(p,\sigma)} e^{-\frac{|x-\mu|^p}{A(p,\sigma)}}, \quad (9)$$

where $x, \mu \in \mathbb{R}$, p, $\sigma > 0$ and $A(p,\sigma) = \left[\frac{\sigma^2\Gamma(1/p)}{\Gamma(3/p)}\right]^{\frac{1}{2}}$. The parameter μ is the mean, the function $A(p,\sigma)$ is a scaling factor that allows the variance of x to take the value of σ^2, and p is the shape parameter that we will call norm. When $p = 1$, the pdf corresponds to a Laplacian or double exponential function and to a Gaussian when $p = 2$. Note that p can take any real value in $(0, \infty)$.

However, in this work, we will experiment with three types of kernels: a d-dimensional Laplacian kernel, which after simplification of Equation 9 yields

$$k_L(t;h) = \prod_{l=1}^{d} \frac{1}{2h_l} e^{-\left|\frac{t_l}{h_l}\right|}, \quad (10)$$

a Gaussian kernel, expressed as

$$k_G(t;h) = \prod_{l=1}^{d} \frac{1}{\sqrt{2\pi}h_l} e^{-\frac{1}{2}(\frac{t_l}{h_l})^2}, \quad (11)$$

and the "square-root" kernel (p=0.5)

$$k_{SQ}(t;h) = \prod_{l=1}^{d} \frac{1}{2h_l} e^{-\left|\frac{2t_l}{h_l}\right|^{\frac{1}{2}}}, \quad (12)$$

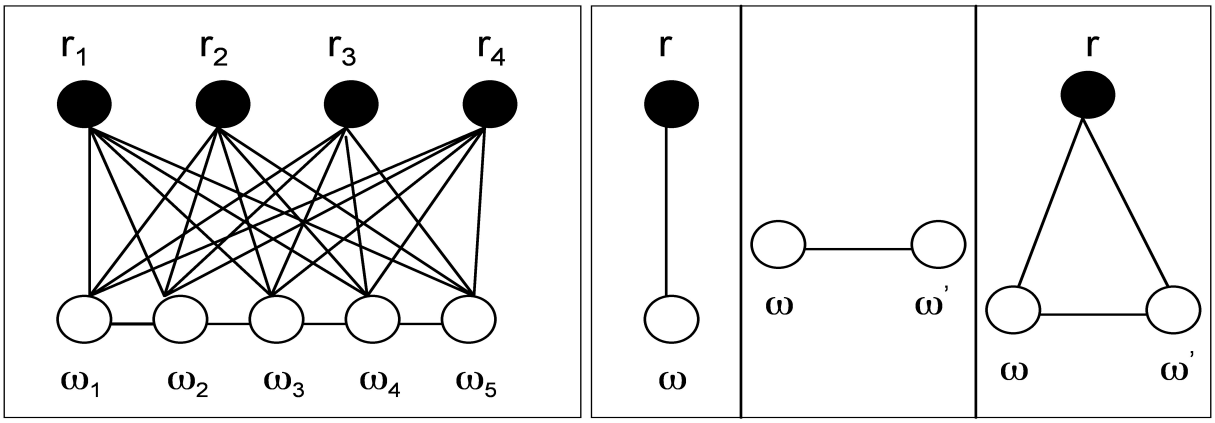

Figure 1: Markov Random Fields graph model. On the right-hand side, we illustrate the configurations explored in this paper: one representing the dependencies between image features and words (r-w), another between two words (w-w'), and the final one shows dependencies among image features and two words (r-w-w').

where $t = r - r_t$, and h_l is the bandwidth of the kernel, which is set by scaling the sample standard deviation of feature component l by the same constant *scale* (sc).

Finally, $P(w|J)$ is modelled using the same multinomial distribution as [15]:

$$P(w|J) = \lambda_1 \frac{N_{w,J}}{N_J} + (1 - \lambda_1)\frac{N_w}{N}. \quad (13)$$

$N_{w,J}$ represents the number of times w appears in the annotation of J, N_J is the length of the annotation, N_w is the number of times w occurs in the training set and N is the aggregate length of all training annotations. λ_1 is the smoothing parameter and together with the coefficient that scales the kernel bandwidth represents the two parameters that are estimated empirically using a held-out portion of the training set.

3.2 Word-to-Word Dependencies

The 2-clique w-w' models word-to-word correlation and is approximated by the following potential function:

$$f_U(c) = \gamma f(w, w') = \gamma \sum_{w'} P(w|w'), \quad (14)$$

$$P(w|w') = \frac{P(w, w')}{P(w')} = \frac{\#(w, w')}{\sum_w \#(w, w')}, \quad (15)$$

where $\#(w, w')$ denotes the number of times the word w co-occurs together with the word w' annotating an image of the training set. To avoid the problem of the sparseness of the data, we follow a smoothing approach:

$$P(w|w') = \beta \frac{\#(w, w')}{\sum_w \#(w, w')} + (1-\beta)\frac{\sum_w \#(w, w')}{\sum_J \sum_w \#(w, w')}, \quad (16)$$

where β is the smoothing parameter.

3.3 Word-to-Word-to-Image Dependencies

The model consists of 3-cliques formed by the words, w and w' and the feature vector r, and captures the dependencies among them. The underlying idea behind this model is that a feature vector representing two visual concepts should imply a degree of compatibility between the visual information and the concepts, and between the concepts themselves. This compatibility is measured by the potential function. For instance, assume that we have a marine scene representing a portion of the sea and a boat, the visual features should reflect the visual properties of the boat and the sea regarding colour and texture and, at the same time, the concepts "sea" and "boat" should pose a degree of semantic relatedness as both represent objects that share the same image context. Thus, the potential function over the 3-clique r-w-w' can be expressed as:

$$\lambda_V f_V(c) = \delta f((w, w'), r), \quad (17)$$

where δ is the weight of the potential function. This can be formulated as the possibility of predicting the pair of words (w, w') given the feature vector r, weighted by the importance of the vector in the image I:

$$f((w, w'), r) = P((w, w')|r)P(r|I), \quad (18)$$

where $P(r|I) \approx \frac{1}{|I|}$, and $|I|$ is set to the number of feature vectors that represent a test image. By applying Bayes formula and the continuous relevance model (CRM) developed by Lavrenko et al. [15] but adapted to $P((w, w'), r)$, we have the following:

$$P((w, w')|r) = \frac{\sum_{J\in\tau} P(J)P((w, w')|J)P(r|J)}{\sum_{(w,w')} P((w, w'), r)}, \quad (19)$$

where J refers to a training image, and τ to the training set. The rest of the terms are computed as follows. $P(J)$ is aproximated by $\frac{1}{|J|}$. $P((w, w')|J)$ is estimated following a generalisation of a multinomial distribution [15] as seen in Section 3.3.1. Finally, $P(r|J)$ is calculated following a Generalized Gaussian kernel estimation as in Equation 10, 11, and 12. In this model, we have three parameters: the smoothing parameter λ_2 of the multinomial distribution and two additional ones derived from the kernel estimation (scale *sc*, and γ) that are estimated during the training phase.

3.3.1 Multinomial Distribution of Pairs of Words

The multinomial distribution of pairs of words is modelled using the formula:

$$P((w, w')|J) = \sum_{w'} \lambda_2 \frac{N_{(w,w'),J}}{N_J} + (1 - \lambda_2)\frac{N_{(w,w')}}{N}. \quad (20)$$

The distribution measures the probability of generating the pair w and w', as annotation words, for the image J based on their relative frequency in the training set. Therefore, the first term reflects the preponderance of the pair of words (w, w') in the image J whereas the second is added as smoothing factor and registers the behaviour of the pair in the whole training set. Thus, $N_{(w,w'),J}$ represents the number of times, zero or one, (w, w') appears in the image J, N_J is the number of pairs that could be formed in the image J, $N_{(w,w')}$ is the number of times (w, w') occurs in the whole training set, $N = \sum_J N_J$, and λ_2 is the smoothing parameter.

For instance, when estimating the distribution of pairs of words formed by the term "tree" in an image annotated with the words "palm", "sky", "sun", "tree" and, "water", we should consider the weight of all pairs appearing in the image as well as in the rest of the training set:

$$\begin{aligned} P((\text{“}tree\text{”}, w')|J) &= \left[\lambda_2 \frac{1}{10} + (1-\lambda_2)\frac{221}{20,972}\right]_{tree-sky} + \\ &+ \left[\lambda_2 \frac{1}{10} + (1-\lambda_2)\frac{23}{20,972}\right]_{tree-sun} + \\ &+ \left[\lambda_2 \frac{1}{10} + (1-\lambda_2)\frac{143}{20,972}\right]_{tree-water} + \\ &+ \left[\lambda_2 \frac{1}{10} + (1-\lambda_2)\frac{22}{20,972}\right]_{tree-palm} + \\ &+ \sum_{w'} \left((1-\lambda_2)\frac{N_{(tree,w')}}{20,972}\right)_{tree-w'}, \end{aligned}$$

where w' represents the rest of vocabulary words that co-occur with "tree" in the rest of the training set, but not in J. Additionally, m is an integer value that represents the number of words annotating an image J, N_J is equal to $\binom{m}{2}$, and N is a constant for a given collection, and it is set to $20,972$ for the Corel 5k dataset. Even if there are images annotated by one single word or without annotations, $P((w, w')|J)$ might be different from zero due to the contribution of the second factor in Equation 20.

4. EXPERIMENTAL WORK

For our experiments, we have adopted a standard annotation database, the Corel 5k dataset, which is a considered benchmark in the field. Additionally, we use the collection provided by the Photo Annotation Task of the 2009 edition of ImageCLEF campaign. The ImageCLEF concepts are very different from the Corel annotations and, hence, the features used should be different for the two datasets. We, however, use the same features for both datasets as the focus of this paper is not on the features. We expect our results to, therefore, not be as good for the ImageCLEF dataset but we would like to show the power of the model and kernel estimates. Note that, although we did not participate in that edition, we compare our results with the other participants in order to provide an estimation of our performance.

4.1 Dataset

The Corel 5k dataset was first used by Duygulu et al. [6] and is a collection of 5,000 images coming from 50 Corel Stock Photo CDs that comprises a training set of 4,500 images and a test set of 500 images. Images of the training set were annotated by human experts using a set of keywords ranging from three to five from a vocabulary of 374 terms.

ImageCLEF is an image retrieval track part of the Cross Language Evaluation Forum (CLEF) campaign. In particular, the Photo Annotation Task [27] is a subtask inside ImageCLEF. The collection provided is a subset of the MIR Flickr 25k dataset. It is made up of a training set of 5,000 images manually annotated with words coming from a vocabulary of 53 visual concepts, and a test set of 3,000 images. It is worth noting that, while most of the vocabulary concepts corresponds to visual concepts, there exist others that not only are not visual but also are highly subjective as "Aesthetic_Impression", "Overall_Quality", and "Fancy".

4.2 Visual Features

The global visual features employed in this research correspond to a combination of 3x3 tiled marginal histogram of global CIELAB colour space computed across 2+2+2 bins, with a 3x3 tiled marginal histogram of Tamura texture across 2+2+2 bins with coherence of 6 and coarseness of 3, with a 3x3 tiled marginal histogram of global HSV colour space computed across 2+2+2 bins, and with a Gabor texture feature using six scales and four orientations.

CIE L*a*b* (CIELAB) is the most perceptually accurate colour space specified by the International Commission on Illumination (CIE). Its three coordinates represent the lightness of the colour (L*), its position between red/magenta and green (a*) and its position between yellow and blue (b*). The histogram was calculated over two bins for each coordinate. HSV is a cylindrical colour space with H (hue) being the angular, S (saturation) the radial and V (brightness) the height component. The H, S and V axes are subdivided linearly (rather than by geometric volume) into two bins each.

The Tamura texture feature is computed using three main texture features called "contrast", "coarseness", and "directionality". Contrast aims to capture the dynamic range of grey levels in an image. Coarseness has a direct relationship to scale and repetition rates, and finally, directionality is a global property over a region. The histogram was calculated over two bins for each feature. The final feature extracted is a texture descriptor produced by applying a Gabor filter to enable filtering in the frequency and spatial domain. We applied to each image a bank of four orientation and six scale sensitive filters that map each image point to a point in the frequency domain.

The image is tiled in order to capture a better description of the distribution of features across the image. Afterwards, the features are combined to maintain the difference between images with, for instance, similar colour palettes but different spatial distribution across the image.

4.3 Evaluation Measures

In this research, we present our results under the rank retrieval metric which consists in ranking the images according to the posterior probability value $P_\Lambda(I|Q)$ as estimated in Equation 3. Then, retrieval performance is evaluated with the Mean Average Precision (MAP), which is the average precision, over all queries, at the ranks where recall changes where relevant items occur. For a given query, an image is considered relevant if its ground-truth annotation contains the query. For simplicity, we employ as queries single words. For the Corel5k dataset we use 260 single word queries and 53 for the ImageCLEF09; in both cases we use all the words

Table 2: State-of-the-art of algorithms in direct image retrieval expressed in terms of mean average precision (MAP) for the Corel 5k dataset. Results with an asterisk show that the number of words used for the evaluation are 179, instead of the usual 260. The first block corresponds to the classic probabilistic models, the second illustrates models based on Markov Random Fields, and the last shows our best performing results

Model	Author	MAP
CMRM	Jeon et al. [12]	0.17*
CRM	Lavrenko et al. [16]	0.24*
CRM-Rect	Feng et al. [10]	0.26
LogRegL2	Magalhaes&Rüger [21]	0.28*
Npde	Yavlinsky et al. [34]	0.29*
MBRM	Feng et al. [10]	0.30
SML	Carneiro et al. [3]	0.31
JEC	Makadia et al. [22]	0.35
Discrete MRF	Feng&Manmatha [9]	0.28
MRF-F1	Feng [8]	0.30
MRF-NRD-Exp1	Feng [8]	0.31
MRF-NRD-Exp2	Feng [8]	0.34
MRF-Lplcn-rw	sc=7.4, λ_1=0.3	0.26
MRF-Lplcn-rw-ww'	sc=7.1,λ_1=0.9,γ=0.1,β=0.1	0.27
MRF-Lplcn-rww'	sc=7.1, λ_2=0.7	0.27
MRF-SqRt-rw-ww'	sc=9.6,λ_1=0.8,γ=0.1,β=0.9	0.29
MRF-SqRt-rw	sc=2, λ_1=0.3	0.32
MRF-SqRt-rw-ww'	sc=2.0,λ_1=0.3,γ=0.1,β=0.1	0.32
MRF-SqRt-rww'	sc=1.8, λ_2=0.3	0.32

that appear in the test set.

4.4 Model Training

The training was done by dividing the training set into two parts: the training set and the validation or held-out set. The validation test is used to find the parameters of the model. After that, the training and validation set were merged to form a new training set that helps us to predict the annotations in the test set. For the Corel 5k dataset, we partitioned the training set into 4,000 and 500 images. The ImageCLEF09 was divided into 4,000 as training set, and 1,000 as held-out data.

Metzler and Croft [23] argued that, for text retrieval, maximising average precision rather than likelihood was more appropriate. Feng and Manmatha [9] showed that this approach worked for image retrieval and we also maximised average precision. We followed a hill-climbing mean average precision optimisation as explained in [25].

5. RESULTS AND DISCUSSION

We analyse the behaviour of three models obtained by combining the clique configurations shown in Figure 1 for the two datasets. In particular, we join the image-to-word with the word-to-word model and investigate whether its performance is higher than the image-to-word and the word-to-word-to-image separately. We also explore the effect of using different kernels in the non-parametric density estimation. Finally, we study which combination of parameters achieves the best performance. The parameters under consideration depend on the selected language model.

Table 3: Top 20 best performing words in Corel 5k dataset ordered according to the columns

Word	Word
land	runway
flight	tails
crafts	festival
sails	relief
albatross	lizard
white-tailed	mule
mosque	sphinx
whales	man
outside	formula
calf	oahu

The name assigned to each of our models is made up of three parts. The first refers to the fact that it is a MRF model. The second applies to the kind of kernel considered: "Lplcn" for Laplacian, "Gssn" for Gaussian, and "SqRt" for the "square-root" kernel. The third part corresponds to the language model used: [-rw] refers to the image-to-word model, [-ww'] to the word-to-word model, and [-rww'] to the word-to-word-to-image model.

Our top results are represented in Table 2 for the Corel 5k dataset, and in Table 5 for the ImageCLEF09 collection. In Table 2, we have included other state-of-the-art algorithms for comparison purposes.

The "square root" kernel provides the best results in any configuration modelled for the two datasets. These results are followed by the Laplacian kernel whereas the Gaussian produces the lowest performance.

For the Corel 5k dataset, the best result corresponds to the word-to-word-to-image configuration, with a MAP of 0.32, closely followed by the image-to-word model, and by the combined image-to-word and word-to-word configuration. The kernel used in the three cases corresponds to the "square root". This result outperforms previous probabilistic methods, with the exception of the continuous MRF-NRD-Exp2 model of Feng [8], and the JEC system proposed by Makadia et al [22]. It is worth mentioning that the good results obtained by Makadia et al. are due to their careful use of visual features. Note that, the top 20 best performing words, which are represented in Table 3, have an average precision value of one. This means that the system is able to annotate these words perfectly.

For the ImageCLEF09 collection, the best performance is achieved by the image-to-word configuration. We consider that this behaviour is very revealing as the correlation between concepts is very rare in the collection, because of the nature of its vocabulary. Thus, as the correlation between words does not provide any added value to the model, the best performing is the image-to-word model, which detects concepts only based on low-level features. The corresponding MAP is of 0.32, which translated into the evaluation measures followed by ImageCLEF competition yields EER of 0.31 and AUC of 0.74. After comparing our results with the rest of the algorithms submitted to the competition, we are located in the position 21 (out of 74 algorithms). As mentioned before our choice of features for ImageCLEF is not optimal and with a better choice of features our model is expected to do a lot better. Again, best results were obtained using a "square root" kernel. Finally, we represent in

Table 4: Average Precision per Word for the top ten best performing words in ImageCLEF09

Word	Avg. Precision
Neutral_Illumination	0.97
No_Visual_Season	0.94
No_Blur	0.86
No_Persons	0.81
Sky	0.77
Outdoor	0.76
Day	0.74
No_Visual_Time	0.74
Clouds	0.61
Landscape_Nature	0.60

Table 4, the top ten best performing words for the image-to-word model. Not surprisingly, the best performing words correspond to visual concepts, while the worst performing correspond to the most subjective concepts.

6. CONCLUSIONS AND FUTURE WORK

We have demonstrated that Markov Random Fields provide a convenient framework for exploiting the semantic context dependencies of an image. In particular, we have formulated the problem of modelling image annotation as that of direct image retrieval. The novelty of our approach lies in the use of different kernels in our non-parametric density estimation together with the utilisation of configurations that explore semantic relationships among concepts at the same time as low-level features, instead of just focusing on correlation between image features like in previous formulations.

Experiments have been conducted on two datasets, the usual benchmark Corel 5k and the collection proposed by the 2009 edition of the ImageCLEF campaign. Our performance is comparable to previous state-of-the-art algorithms for both datasets. We observed that the kernel estimation has a significant influence on the performance of our model. In particular, the "square root" kernel provides the best performance for both collections. With respect to the language model, the best result corresponds to the configuration that exploits dependencies between words at the same time as dependencies between words and visual features. This makes sense as it is the configuration that makes use of the maximum amount of information from the image. However, the ImageCLEF achieves the best performance with the word-to-image configuration although closely followed by word-to-word-to-image model. We consider that this behaviour is very revealing as the correlation between concepts is very rare in the collection, as a result of the nature of its vocabulary. Thus, as the correlation between words does not provide any added value to the model, the best performing is the image-to-word model, which detects concepts only based on low-level features.

As for future work, we intend to consider other kernels to see whether we can improve our results even more. Additionally, we will study whether a better choice of features as in Makadia et al. [22] might improve our performance.

7. ACKNOWLEDGMENTS

This work was partially funded by EU-Pharos project (IST-FP6-45035) and by Santander Universities.

Table 5: Top performing results for the ImageCLEF09 dataset expressed in terms of Mean Average Precision using 53 words as queries. Note that parameters are not optimised completely due to time constraints. We believe that full optimisation cannot make the results worse and potentially can make them better

Model	MAP
MRF-Gssn-rw	0.30
MRF-Gssn-rw-ww'	0.30
MRF-Lplcn-rw-ww'	0.31
MRF-Lplcn-rw	0.31
MRF-SqRt-rw-ww'	0.32
MRF-SqRt-rww'	0.32
MRF-SqRt-rw	0.32

8. REFERENCES

[1] I. Biederman. On the semantics of a glance at a scene. In *Perceptual organization*. Erlbaum, 1981.

[2] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proceedings of the 8th European Conference on Computer Vision*, volume 1, pages 350–362, 2004.

[3] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised Learning of Semantic Classes for Image Annotation and Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[4] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 163–168, 2005.

[5] J. A. Domínguez-Molina, G. González-Farías, and R. M. Rodríguez-Dagnino. A practical procedure to estimate the shape parameter in the generalized Gaussian distribution. Technical report, Universidad de Guanajuato, 2003.

[6] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision*, pages 97–112, 2002.

[7] H. J. Escalante, M. Montes, and L. E. Sucar. Word Co-occurrence and Markov Random Fields for Improving Automatic Image Annotation. In *Proceedings of the 18th British Machine Vision Conference*, 2007.

[8] S. Feng. *Statistical models for text query-based image retrieval*. PhD thesis, University of Massachusetts Amherst, January 2008.

[9] S. Feng and R. Manmatha. A Discrete Direct Retrieval Model for Image and Video Retrieval. In *Proceedings of the International Conference on Content-based Image and Video Retrieval*, pages 427–436, 2008.

[10] S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1002–1009,

2004.

[11] C. Hernández-Gracidas and L. E. Sucar. Markov Random Fields and Spatial Information to Improve Automatic Image Annotation. In *IEEE Pacific-Rim Symposium on Image & Video Technology*, volume 4872, pages 879–892, 2007.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, pages 119–126, 2003.

[13] R. Jin, J. Y. Chai, and L. Si. Effective automatic image annotation via a coherent language model and active learning. In *Proceedings of the 12th International ACM Conferencia on Multimedia*, pages 892–899, 2004.

[14] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotations by combining multiple evidence & WordNet. In *Proceedings of the 13th International ACM Conference on Multimedia*, pages 706–715, 2005.

[15] V. Lavrenko, S. Feng, and R. Manmatha. Statistical Models For Automatic Video Annotation And Retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 17–21, 2004.

[16] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems*, 2003.

[17] W. Li and M. Sun. Automatic Image Annotation Based on WordNet and Hierarchical Ensembles. In *Proceedings of 7th International Conference on Intelligent Text Processing and Computational Linguistics*, volume 3878, pages 417–428, 2006.

[18] J. Liu, M. Li, W.-Y. Ma, Q. Liu, and H. Lu. An adaptive graph model for automatic image annotation. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 61–70, 2006.

[19] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. In *Proceedings of the 15th international conference on Multimedia*, pages 605–614, 2007.

[20] N. Loeff, A. Farhadi, I. Endres, and D. A. Forsyth. Unlabeled data improves word prediction. In *Proceedings of the 12th IEEE International Conference on Computer Vision*, 2009.

[21] J. Magalhães and S. Rüger. Information-theoretic semantic multimedia indexing. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, pages 619–626, 2007.

[22] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proceedings of the 10th European Conference on Computer Vision*, pages 316–329, 2008.

[23] D. Metzler and B. W. Croft. A Markov Random Field model for Term Dependencies. In *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 472–479, 2005.

[24] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, pages 42–50, 2004.

[25] W. Morgan, W. Greiff, and J. Henderson. Direct maximization of average precision by hill-climbing, with a comparison to a maximum entropy approach. In *Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 93–96, 2004.

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

[27] S. Nowak and P. Dunker. Overview of the CLEF 2009 Large Scale -Visual Concept Detection and Annotation Task. In *Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum*, 2009.

[28] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *Proceedings of the 15th international conference on Multimedia*, pages 17–26, 2007.

[29] R. Shi, T.-S. Chua, C.-H. Lee, and S. Gao. Bayesian learning of hierarchical multinomial mixture models of concepts for automatic image annotation. In *Proceedings of the Conference on Image and Video Retrieval*, pages 102–112, 2006.

[30] R. Shi, C.-H. Lee, and T.-S. Chua. Enhancing image annotation by integrating concept ontology and text-based Bayesian learning model. In *Proceedings of the 15th international conference on Multimedia*, pages 341–344, 2007.

[31] M. Srikanth, J. Varner, M. Bowden, and D. Moldovan. Exploiting ontologies for automatic image annotation. In *Proceedings of the 28th International ACM Conference on Research and Development in Information Retrieval*, pages 552–558, 2005.

[32] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, 2003.

[33] Y. Xiang, X. Zhou, T.-S. Chua, and C.-W. Ngo. A Revisit of Generative Model for Automatic Image Annotation using Markov Random Fields. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, pages 1153–1160, 2009.

[34] A. Yavlinsky, E. Schofield, and S. Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, pages 507–517, 2005.

[35] X. Zhou, M. Wang, Q. Zhang, J. Zhang, and B. Shi. Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, pages 25–32, 2007.