Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **6**, November **2017**

Exploiting Semantic Context Relationships for Automatic Image Annotation

Dongping Tian

Institute of Computer Software
Baoji University of Arts and Sciences
No.1 Hi-Tech Avenue, Hi-Tech District, Baoji, Shaanxi 721013, P.R. China

Institute of Computational Information Science
Baoji University of Arts and Sciences
No.44 Baoguang Road, Weibin District, Baoji, Shaanxi 721007, P.R. China
tiandp@ics.ict.ac.cn, tdp211@163.com

Received December, 2016; revised May, 2017

Abstract. *Automatic image annotation has been an active topic of research in computer vision and multimedia areas for decades due to its potentials in semantic based image retrieval. In this paper, we present a semantic context model for automatic image annotation based on a novel conditional random field (CRF). To start with, a probabilistic latent semantic analysis (PLSA) with asymmetric modalities is built to predict a candidate set of annotations with confidence scores, through which the semantic contextual relations of image-to-image and image-to-word can be implicitly introduced into CRF. Followed by the contextual potential function is formulated based on the semantic context relationships between the annotation keywords by calculating their converted normalized Google distance, during which the semantic context of word-to-word relations can be explicitly brought in CRF. Finally, according to the criterion of the maximum a posteriori, the image annotation results can be obtained by inferring the indicator variables. The chief novelty of our work is to exploit PLSA to predict a candidate set of annotations as well as CRF to further explore the semantic context relationships for precise image annotation. Extensive experiments on the standard Corel5k dataset demonstrate that our model is much more effective than several state-of-the-art methods in dealing with the task of automatic image annotation and retrieval. In particular, this paper ends with a summary of some important conclusions and highlights the potential research directions of CRF in the field of semantic image analysis for the future.*
Keywords: Automatic image annotation, CRF, PLSA, Semantic gap, Image retrieval

1. **Introduction.** With the explosive growth of the world wide web and rapidly growing number of available digital color images, much research effort is devoted to the development of efficient semantic image analysis systems. It is witnessed that a considerable progress has been obtained in the past few years, yet, as a field, automatic image annotation (AIA) is still in its infancy and is not sophisticated enough to extract perfect semantic concepts according to image low-level features, facing many challenges and limitations. One of the main handicaps is the well-known semantic gap between low-level visual features and high-level semantic concepts. Fortunately, a huge number of advanced machine learning techniques have been proposed in the literature to narrow down it.

In general, most of the existing methods can be roughly classified into three categories, i.e., classification-based method, probabilistic modeling method and graph-based method.

To be specific, the classification-based method treats each semantic keyword or concept as an independent class and assigns each to one classifier. Work such as linguistic indexing of pictures [1], Bayes point machines [2] and supervised multiclass labeling [3] fall into this category. The probabilistic modeling method aims to learn a relevance model to represent the correlation between images and words. The early notable work includes the translation model [4] which treated AIA as a process of translation from a set of blob tokens to a set of keywords, cross-media relevance model (CMRM) [5] assumed the blobs and words were mutually independent given a specific image, continuous space relevance model (CRM) [6], multiple Bernoulli relevance model (MBRM) [7], dual cross-media relevance model [8], probabilistic latent semantic analysis related models [9-12], latent Dirichlet allocation (LDA) [13] and correlated topic model (CTM) [14], etc. In addition, the graph-based method has already achieved much success in the field of semantic image analysis. As the representative work, a graph model was developed to annotate images by exploring the pairwise connections in multiple full-length NSCs [15]. In [16], a generalized manifold ranking algorithm was put forward for image retrieval by representing images and their relationships as a graph and propagating labeled information among images through the graph structure. Subsequently the original manifold ranking algorithm was extended to a new framework for image retrieval from two aspects involving scalable graph construction and efficient computation [17], etc. Alternatively, it should be noted that the conditional random field that can be viewed as an undirected graphical model has been widely used in the area of computer vision. An early approach to image labeling was the multi-scale CRF by including contextual features [18]. Subsequent work [19] came up with the ensemble based on conditional random field (En-CRF) for multi-label image and video annotation. Especially in [20], a single-layered segment based CRF was proposed rather than multi-layered hierarchical CRF to integrate multi-scale features of pixels, segments and regions for image labeling. Besides, a tree-structured CRF was introduced for interactive image labeling [21] that explicitly took into account the dependencies among image labels. More reviews on CRF based semantic image analysis will be summarized in the next section.

The remainder of this paper is organized as follows. Section 2 succinctly reviews the CRF based semantic image analysis in the field of computer vision. Section 3 introduces the basic principles of CRF as well as its performance comparison with the hidden Markov models and maximum entropy Markov models respectively. In Section 4, the proposed semantic context model is elaborated in detail from three aspects of its potential function, parameter estimation and model inference. Section 5 reports and analyzes the experimental results on the general-purpose Corel5k dataset. At length, we end this paper with some concluding remarks and future work in Section 6.

2. **Related Work.**

In recent years, conditional random field has become an effective tool for a variety of different data segmentation and labeling tasks. In this section, CRF based semantic image analysis will be comprehensively reviewed from the aspects of image annotation, image classification, image segmentation, and other related applications.

2.1. **CRF based image annotation.**

In the context of AIA, the early notable work was the multi-scale CRF by including contextual features for image labeling [18]. Particularly, the interaction at pixel-level had the form of a restricted Boltzmann machine. However, it did not model the explicit notion of objects and its higher level nodes rather serve as switches to different context and object co-occurrences. In [22], CRF was employed for refining image annotation by incorporating the contextual relations between candidate set of annotations. To consider spatial adjacency information during the assignment of high-level objects to local image patches, a scene labeling model was developed in [23] by

incorporating both local features and features aggregated over the whole image or large sections for performing semantic region labeling. Besides, note that a series of CRF models were constructed for image labeling [24,25], the primary difference was the formulation of their potential functions by adopting Laplacian mixture and generalized Gaussian mixture potentials respectively. In the literature [26], maximal margin CRF was formulated to apply multiple visual features, but the feature weights were learned independently from the image labeling. To reduce the performance variance and to exploit the correlation between annotation words, a hierarchical two-stage CRF [27] was introduced to deal with the problem of labeling images of street scenes by several distinctive object classes. Subsequent work [28] integrated semantic context modeling and sparse multiple distance learning by using kernel logistic regression in CRF framework for AIA, etc. To sum up, all of the models mentioned above are able to obtain promising results. However, note that the potentials of CRF are not only selected empirically and hand-tuned to have better performance, but also heavily depend on the specific applications.

2.2. **CRF based image classification.** The aim of image classification, loosely speaking, is to decide whether an image belongs to a certain category or not. From the literature, it can be easily observed that CRF has been widely applied in the task of image classification. In [29], CRF was leveraged to address the spatial context from pairwise relations. In particular, a two-layer hierarchical formulation was presented to exploit different levels of spatial context in images for robust classification. Subsequently a hidden CRF was proposed for gesture recognition [30], which was for whole sequences classification by viewing the segment labels as hidden variables, but the hidden variable structure made the objective function non-convex and only local optimum could be achieved during the process of training. Based on this scenario, a hybrid approach of PLSA and a classification oriented CRF (COCRF) was constructed for natural scene categorization [31], in which a topic label was firstly assigned to each segment on the training data by PLSA and then a COCRF model was trained based on these topic labels. In the work of [32], a tree-structured conditional random field (TCRF) was formulated for image modeling, whose main advantage was its ability to incorporate contextual features at several levels of granularity that could be achieved by inducing a hierarchy of hidden variables over the given label field. It should be noted that the discriminative nature of CRF combined with the tree-like structure gave TCRF several advantages over other existing multi-scale models. Moreover, TCRF was able to yield a general framework that could be applied to a variety of image classification tasks. Beyond this, subsequent works [33,34] have also developed dynamic CRF models for object labeling, etc.

2.3. **CRF based image segmentation.** Image segmentation, more precisely, is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics. It has been shown that the addition of higher order potentials for CRF defined on cliques with more than two nodes can lead to better performance. As a result, many researchers have considered variants of the traditional CRF models to improve their performance [35-39]. In [35], a higher order CRF was formulated by defining potentials based on sets of pixels (image segments). Although this approach robustly penalized inconsistent labeling within a segment but it failed to model dependencies between segments. To solve this problem, an associative hierarchical CRF [36] was built to allow associations in both same layers and between layers. All of the models could successfully reason about pixels and/or segments, but they failed to incorporate the notion of object instances, their location and spatial extent, into the recognition framework. Furthermore, both of the two methods were highly dependent on the pixel-level annotations during the training stage. Alternatively, recent works [37-39] formulated the higher order potentials

of CRF for image segmentation in terms of the superpixel neighborhood. In particular, recent advances in deep learning reveal that it is becoming increasingly common to perform CRF inference within a deep neural network to facilitate joint learning of the CRF with a pixel-wise convolutional neural network (CNN) classifier [40-43]. Note that CRF inference is able to refine weak and coarse pixel-level label predictions to produce sharp boundaries and fine-grained segmentations. Based on this recognition, a recent work [40] exploited higher order potentials defined over superpixels but did not strictly assume that the segments shared boundaries with objects in an image. Afterwards CRF was applied to refine semantic segmentation obtained from a CNN classifier. In [41], Bell et al. focused on the material recognition and segmentation whereas in [42] Chen et al. combined the pixel-level CRF with deep CNN based unary terms for the task of image segmentation. Arguably, this did not fully harness the strength of CRF since it was not integrated with the deep network. As a result, Zheng et al.[43] combined the strengths of both CNN and CRF based graphical models in one unified framework for segmentation. In addition, CRF has also been extensively applied in many other tasks such as text recognition [44], language understanding [45], and 3D reconstruction [46], etc.

As previously reviewed, most of these models can achieve promising performance and motivate us to explore better image annotation methods with the help of their excellent experiences and knowledge. Thus in this paper, we present a novel semantic context model for automatic image annotation. The main contributions of this work can be summarized as follows. First, a PLSA model with asymmetric modalities is constructed to predict a candidate set of annotations to define the local evidence function for the conditional random field, through which the semantic contextual relations of image-to-image and image-to-word can be implicitly introduced into CRF. Second, a normalized Google distance based contextual potential function is formulated to explicitly bring in the semantic context relations of word-to-word in CRF. Extensive experiments on Corel5k validate that our model is superior or highly competitive to several state-of-the-art approaches.

3. **Conditional Random Field.** Conditional random field was first proposed by Lafferty [47] for labeling sequential data which was a linear-chain CRF. Subsequently it was extended to 2D for image labeling in [48], in which the image was first divided into regular grids and features of these grids were then extracted. The CRF model was built on these grid features with association and interaction potential functions. The association potentials denoted the likelihoods of the label given the observation of the grid while the interaction potentials were the likelihoods of the interaction between neighboring grid labels given the observation of neighboring grids. A CRF can be viewed as an undirected graphical model in which each vertex represents a random variable whose distribution is to be inferred, and each edge represents a dependency between two random variables. In CRF model, the distribution of each discrete random variable y_i in the graph is conditioned on an input sequence x. Mathematically, the conditional probability of $y = (y_1, y_2, \cdots, y_n)$ given x is formulated as below:

$$P(y|x) = \frac{e^{\phi(y,x;\Phi)}}{\sum_{y'} e^{\phi(y',x;\Phi)}} \tag{1}$$

where

$$\phi(y, x; \Phi) = \sum_i \sum_k \theta_{k1} f_{k1}(y_i, i, x) + \sum_{i,j} \sum_l \theta_{l2} f_{l2}(y_i, y_j, i, j, x) \tag{2}$$

Eq.(2) is the potential function. i,j are used to index the vertexes, $f_{k1}(y_i, i, x)$ and $f_{l2}(y_i, y_j, i, j, x)$ denote the node feature function and edge feature function respectively, $\Phi = (\theta_{k1}, \theta_{l2})$ indicates the model parameters to be learned.

Fig.1 depicts the graphical structures of CRF model. Especially for the 2D CRF model shown in Fig.1 (b), which is an undirected probabilistic graphical model that is an extension of the linear-chain CRF. The dash circles denote the observed features at the node, empty circles stand for the labels which are unknown for the test images, and the interactions between these random variables are displayed as edges in the figure.

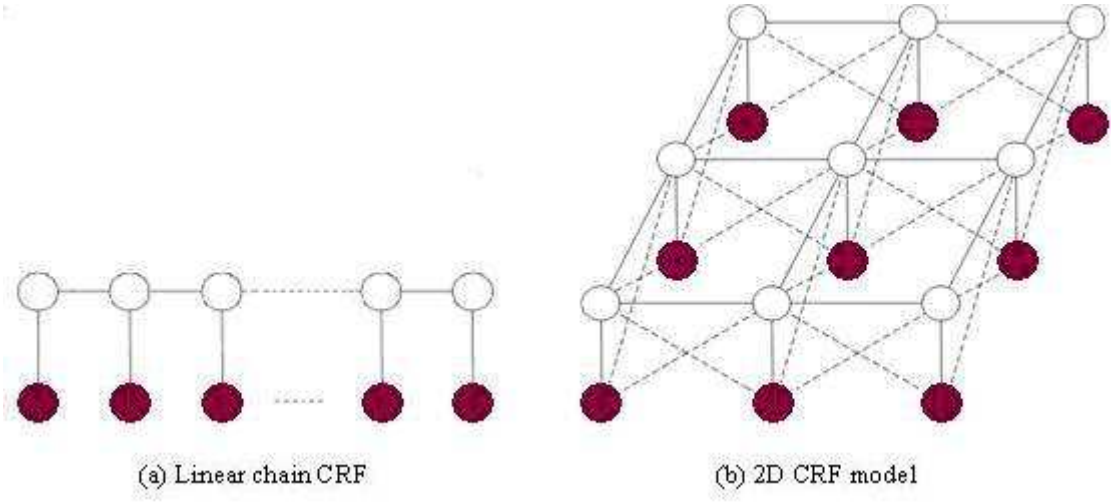

FIGURE 1. Graphical structures of the CRF model

In addition, it should be noted that the primary advantage of conditional random field over hidden Markov models (HMM) is its conditional nature, resulting in the relaxation of the independence assumptions required by HMM in order to ensure tractable inference. Besides, CRF can avoid the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMM) and other conditional Markov models based on directed graphical models. From the literatures, it can be clearly observed that CRF outperforms both MEMM and HMM on a number of real-world tasks in many fields, including pattern recognition, bioinformatics and computational linguistics.

4. **The Proposed Semantic Context Model.** In this section, we first provide a succinct introduction of PLSA model. Followed by the proposed semantic context model is elaborated from three aspects of its potential function, parameter estimation and model inference, respectively.

4.1. **PLSA model.** Probabilistic latent semantic analysis is a statistical latent class model that introduces a hidden variable (latent aspect) z_k in the generative process of each element x_j in a document d_i. Given this unobservable variable z_k, each occurrence x_j is independent of the document it belongs to, which corresponds to the following joint probability:

$$P(d_i, x_j) = P(d_i) \sum_{k=1}^{K} P(z_k|d_i) P(x_j|z_k) \tag{3}$$

The model parameters of PLSA are the two conditional distributions: $P(x_j|z_k)$ and $P(z_k|d_i)$. $P(x_j|z_k)$ characterizes each aspect and remains valid for documents out of the training set. On the other hand, $P(z_k|d_i)$ is only relative to the specific documents

and cannot carry any prior information to an unseen document. The expectation maximization (EM) algorithm is employed to estimate the parameters through maximizing the log-likelihood of the observed data. The steps of it can be succinctly described as follows.

E-step. The conditional distribution $P(z_k|d_i, x_j)$ is computed from the previous estimate of the parameters.

$$P(z_k|d_i, x_j) = \frac{P(z_k|d_i)P(x_j|z_k)}{\sum_{l=1}^{K} P(z_l|d_i)P(x_j|z_l)} \tag{4}$$

M-step. The parameters $P(x_j|z_k)$ and $P(z_k|d_i)$ are updated with the new expected values $P(z_k|d_i, x_j)$.

$$P(x_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, x_j)P(z_k|d_i, x_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, x_m)P(z_k|d_i, x_m)} \tag{5}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i, x_j)P(z_k|d_i, x_j)}{\sum_{j=1}^{M} n(d_i, x_j)} \tag{6}$$

Note that if one of the parameters ($P(x_j|z_k)$ or $P(z_k|d_i)$) is known, the other one can be inferred by using the folding-in method, which updates the unknown parameters with the known parameters kept fixed, so that it can maximize the likelihood with respect to the previously trained parameters. Given a new image visual features $v(d_{new})$, the conditional probability distribution $P(z_k|d_{new})$ can be inferred with the previously estimated model parameters $P(v|z_k)$, then the posterior probability of words can be computed by the following formula:

$$P(w|d_{new}) = \sum_{k=1}^{K} P(w|z_k)P(z_k|d_{new}) \tag{7}$$

From Eq.(7), a candidate set of annotations with confidence scores (i.e., the posterior probabilities of annotation keywords) can be easily obtained.

4.2. **Our proposed model.** It is known that the major advantage of CRF is that it can provide a coherent probabilistic fusion approach by simultaneously taking into account the individual probabilistic label assignment and the contextual relations between annotation keywords. Moreover, CRF has its own unique ability to express data long-range correlations and overlapping characteristics as well as to effectively overcome the label bias problem due to the use of global optimization techniques. Based on this recognition, we utilize CRF for automatic image annotation by incorporating the semantic context relationships between annotation keywords. Without loss of generality, the proposed CRF model can be described from three aspects of the potential function, parameter estimation and model inference, respectively.

- **Construction of the potential function**

Note that in CRF the selection of potential function is quite flexible, and different CRF models can adopt different forms of the potential function. Motivated by the work [12,22], the potential function is formulated as below:

$$\phi(y, x; \Phi) = \alpha_1 \times \sum_i \omega^1(y_i, i, x) + \alpha_2 \times \sum_{i,j} \omega^2(y_i, y_j, i, j) \tag{8}$$

where ω^1 indicates the local evidence of the state of y_i. It depends on the image observation x. ω^2 is a prior parameter that indicates the contextual potential between the states of

two variables y_i and y_j. Here, we take the local evidence as the logarithm of the confidence scores provided by the PLSA model.

$$\omega^1(y_i, i, x) = \begin{cases} \log P(c(y_i) = 1|x) \ , y_i = 1 \\ \log[1 - P(c(y_i) = 1|x)], y_i = 0 \end{cases} \tag{9}$$

where $c(y_i)$ is used to indicate the concept and P denotes the posterior probability of the annotation words according to Eq.(7). Besides, the normalized Google distance (NGD) [49] is a distance function between two words obtained by just typing them as the search terms in the Google search engine. It has a simple mathematical formulation but has a solid theory foundation. In actual fact, NGD is a measure of contextual relation of the words. Based on this recognition, the contextual potential is assumed to be independent of x and can be obtained as general knowledge by the converted NGD as below.

$$\omega^2(y_i, y_j, i, j) = \begin{cases} -\log NGD(y_i, y_j), y_i = y_j = 1 \\ 0 \ , otherwise \end{cases} \tag{10}$$

$$NGD(y_i, y_j) = \frac{\max(\log f(y_i), \log f(y_j)) - \log f(y_i, y_j)}{\log G - \min(\log f(y_i), \log f(y_j))} \tag{11}$$

where $f(y_i)$ and $f(y_j)$ denote the numbers of images containing labels y_i and y_j respectively, $f(y_i, y_j)$ is the number of images containing both y_i and y_j, G is the total number of images in the dataset.

In contrast with previous methods, the main advantages of this formulation based CRF lies in two aspects. On one side, the PLSA model based local evidence function is able to incorporate more rich context information, i.e., the semantic contextual relations of image-to-image and image-to-word, which can be implicitly introduced into CRF based on the bag-of-visual-words construction. On the other side, the NGD based contextual potential function is able to explicitly construct the semantic context relations of word-to-word. Subsequently this two kinds of contextual relationships are embedded into the basic CRF as its potential functions. It is worth noting that these two semantic relations are complementary to each other and the combination makes them also benefit from each other. Our primary goal is to solve the annotation problem by considering the trade off between the computational efficiency and the precision performance. Extensive experiments corroborate that the proposed model can markedly enhance the performance of image annotation and retrieval but at the same time retain its simplicity and elegance.

- **Parameter estimation of the model**

Note that in our CRF model (Eq.(8)), the weight parameters $\Phi = \{\alpha_1, \alpha_2\}$ to be learned are utilized to control the balance between local evidence and contextual potential.

$$L(\Phi) = \sum_k \log(y^k|x^k) - \frac{\alpha_1^2}{2\sigma^2} - \frac{\alpha_2^2}{2\sigma^2} \tag{12}$$

Since the task of CRF for image annotation is to infer the most probable labels given an input image and the model parameters which are learned from the training set. As can be seen from the above description, it is difficult to choose the weight parameters manually since the local evidence and contextual potential come from different sources. Hence the deepest gradient descent (DGD) algorithm is adopted to estimate the parameters Φ.

Algorithm 1 - Pseudocode of the parameter estimation

Input:

the training image set T, validation image set V.

Process:

Training PLSA model on T;

Selecting words with top confidence scores generated by the trained PLSA on V;
Constructing indicator vector y for all images in V;
Computing local evidence by Eq.(9) and the contextual potentials of CRF by Eq.(10);
Learning Φ *by maximizing the log posterior of Eq.(12) by the DGD algorithm;*

Output:

weight parameters $\Phi = \{\alpha_1, \alpha_2\}$.

- **Annotation inferring based on the model**

After learning the weight parameters of the proposed CRF model, the image annotation results can be obtained by inferring the most likely state of each variable as follows.

$$y_i^* = \arg\max_{y_i} P(y_i|x; \Phi^*), y_i \in \{0, 1\} \tag{13}$$

To be specific, its procedure can be described by the following Algorithm 2.

Algorithm 2 - Pseudocode of the CRF model inference

Input:

the learned weight parameter Φ*, test image I.*

Process:

Generating annotations of I by the trained PLSA;
Constructing the corresponding indicator vector;
Inferring the indicator variables by Eq.(13);

Output:

annotation results of I.

Note that the indicator vector in the pseudocode described above is constructed in such a way that variable y_i is true if the corresponding concept appears among the keywords with top-10 confidence scores and also in the ground truth labels, otherwise it is false.

5. **Experimental Results and Analysis.** To make a fair comparison, the general-purpose Corel5k dataset[1] is employed to validate the effectiveness of the proposed CRF model, which consists of 5,000 images from 50 Corel Stock Photo CD's provided by [4]. Each CD contains 100 images with a certain theme, of which 90 are designated to be in the training set and 10 in the test set, resulting in 4,500 training images and a balanced 500-image test collection. Besides, the dictionary contains 260 words that appear in both the training and testing set. It is worth noting that the normalized cuts algorithm (Ncuts) [50] rather than JSEG [51] is applied to segment images into a number of meaningful regions. The main reason is that JSEG only focuses on local features and their consistencies while Ncuts aims at extracting the global impression of an image data. So Ncuts, to some extent, can get a better segmentation result than that of JSEG (As can be seen from Fig. 2).

FIGURE 2. The segmentation results using Ncuts (mid) and JSEG (right)

[1]http://vision.sista.arizona.edu/kobus/research/data/eccv_2002/index.html

In addition, since the focus of this paper is not on image feature selection, for each image at most the 10 largest regions are selected and 809-dimensional visual features [2] (color, texture, shape and saliency) are extracted for each region, which include 81-dimensional grid color moment features, 59-dimensional local binary pattern (LBP) features, 120-dimensional Gabor wavelets texture features, 37-dimensional edge orientation histogram features and 512-dimensional GIST features respectively. Afterwards these features are employed to train PLSA based on the EM algorithm. Without loss of generality, the most commonly used metrics precision and recall of every word in the test set are calculated and the mean of these values is applied to summarize the model performance. Similar to [7], for a given semantic word, recall=B/C and precision=B/A, where A is the number of images automatically annotated with a given word in the top-5 returned word list, B is the number of images correctly annotated with that keyword in the top-5 returned word list, and C denotes the number of images having that word in the ground truth annotation. In addition, the mean average precision (mAP) is used to evaluate the retrieval performance of our model.

$$mAP = \frac{1}{N_w} \sum\nolimits_{w=1}^{N_w} AP(w) \tag{14}$$

with

$$AP(w) = \frac{\sum_{i \in relevant} precision(i)}{rel(w)} \tag{15}$$

Note that the AP of a query w is defined as the sum of the precisions of the correctly retrieved images at rank i divided by the total number of relevant images rel(w) for this query.

5.1. **Results of automatic image annotation.** To show the effectiveness of our model proposed in this paper, we compare it with several previous approaches [5,6,7,9,52]. The experimental results listed in Table 1 are based on two sets of words: the subset of 49 best words and the complete set of all 260 words that occur in the training set. From Table 1, it is clear to see that our model outperforms all the others, especially the first two approaches. Meanwhile, it is also superior to PLSA-WORDS, PLSA-FUSION, CRMR and MBRM by the gains of 18, 14, 7 and 4 words with non-zero recall, 18%, 18%, 13% and 4% mean per-word recall together with 79%, 32%, 14% and 9% mean per-word precision on the set of 260 words, respectively. Similarly, our model can also achieve consistent good performance on the set of 49 best words. Note that CRMR listed in Table 1 denotes CRM with rectangular regions as input. More details on it can be gleaned from reference [7].

Table 1. Performance comparison on Corel5k dataset

Models	CMRM	CRM	PLSA-WORDS	PLSA-FUSION	CRMR	MBRM	Ours
#words	66	107	108	112	119	122	126
Results on 49 best words							
Recall	0.48	0.70	0.76	0.76	0.75	0.75	0.78
Precision	0.40	0.59	0.58	0.65	0.72	0.73	0.75
Results on all 260 words							
Recall	0.09	0.19	0.22	0.22	0.23	0.25	0.26
Precision	0.10	0.16	0.14	0.19	0.22	0.23	0.25

In addition, Fig. 3 presents some examples of the annotations (only four cases are listed here due to the limited space) generated by RVM-CRF and our model, respectively. As

[2] http://appsrv.cse.cuhk.edu.hk/∼jkzhu/felib.html.

can be seen from Fig. 3, our model is able to generate more accurate annotation results compared with the original annotations as well as the ones provided in literature [22]. Note that the enriched and re-ranked annotations compared to those of the ground truth and RVM-CRF are underlined and italicized respectively. Taking the second image for example, there exist four tags in the original annotation list. However, after annotation by our model, its annotation is enriched by the other keyword "pergola", which is very appropriate and reasonable to describe the visual content of the image. Similarly, the enriched label "landscape" for the third image, and the enriched labels "plant & landscape" for the fourth image, which further demonstrates the effectiveness and efficiency of the proposed model in this paper.

Images				
Ground Truth Annotation	mountain, lake, water, grass, ocean	flower, plant, leaves, garden	building, city, sky, ocean	grass, tiger, cat, forest
RVM-CRF Annotation	mountain, lake, water, building, ocean	grass, plant, flower, garden, people	building, landscape, sky, city, ocean	grass, plant, tiger, people, landscape
Our Annotation	mountain, *water*, *lake*, grass, ocean	flower, leaves, garden, *plant*, pergola	building, *city*, *sky*, landscape, ocean	*tiger*, *grass*, plant, landscape, cat

FIGURE 3. Annotation comparison with RVM-CRF and Our model

To further illustrate the effect of the proposed model for automatic image annotation, Fig. 4 displays the average annotation precisions of the selected 10 words "house", "mountain", "snow", "tree", "building", "water", "beach", "sky", "cat" and "bear" based on RVM-CRF, PLSA-WORDS and our model, respectively. As shown in Fig. 4, the average precision of the proposed model is consistently higher than that of RVM-CRF and PLSA-WORDS.

Fig. 5 displays the precision-recall curves corresponding to the PLSA-WORDS and our model based on the Corel5k dataset, with the number of annotations from 2 to 10. As observed from Fig. 5, the performance of our model is apparently superior to that of the PLSA-WORDS, which further validates the effectiveness and efficiency of our model.

5.2. **Results of ranked image retrieval.** To further illustrate the effect of our model for image retrieval, mean average precision (mAP) is also applied as a metric to evaluate the retrieval performance. Here, we only compare our model with CMRM, CRM, MBRM, PLSA-WORDS and PLSA-RW due to the mAPs of other methods cannot be accessed directly from the literature. As shown in Table 2, our model is obviously superior to the other methods except for MBRM. Specifically, compared with MBRM, it can get 3% improvements on all the 260 words and words with positive recall, respectively.

From the perspective of probability theory, image retrieval can be seen as a procedure of ranking images in the database according to their posterior probabilities of being relevant to the query concept. To further stress the advantages of our model in image retrieval, Fig.

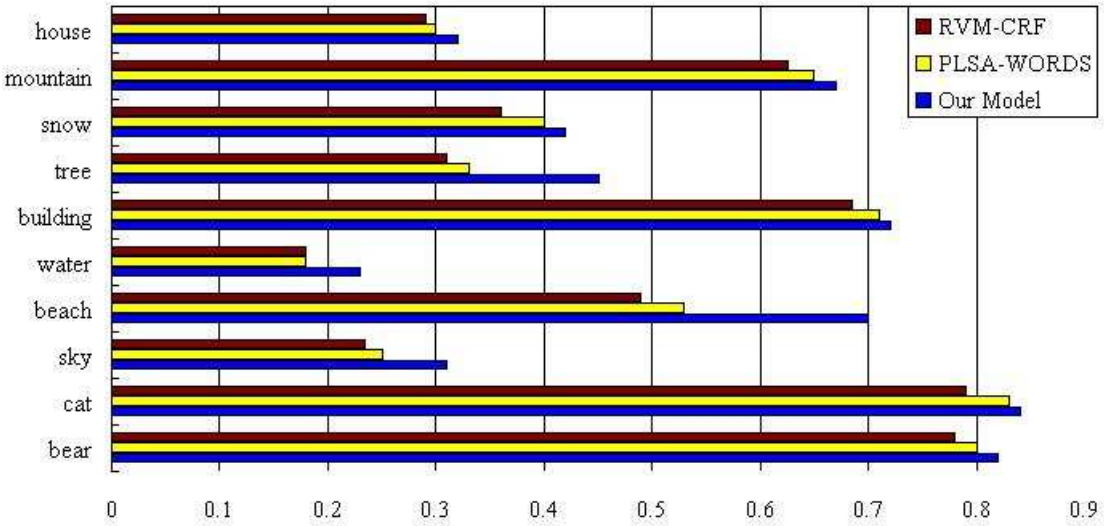

FIGURE 4. Average precision of RVM-CRF, PLSA-WORDS and Our model

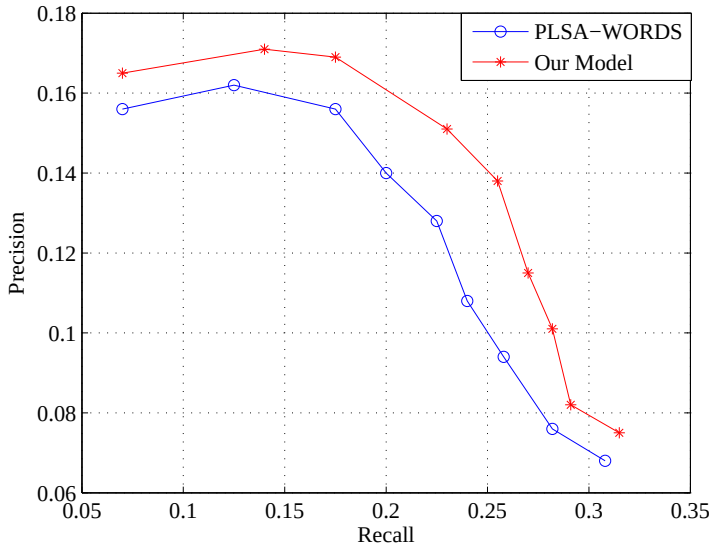

FIGURE 5. Precision-recall curves of PLSA-WORDS and Our model

TABLE 2. Comparison of ranked image retrieval on Corel5k

Models	All 260 words	Words with recall > 0
CMRM	0.17	0.20
CRM	0.24	0.27
MBRM	0.30	0.35
PLSA-WORDS	0.22	0.26
PLSA-RW	0.26	0.30
Our Model	0.31	0.36

6 presents the retrieval results obtained with single word queries on several challenging visual concepts being queries. Each row displays the top five matches to the semantic

query "flower", "coast", "tiger" and "mountain" from top to bottom, respectively. The diversity of visual appearance of the returned images further demonstrates that our model also has good generalization ability.

FIGURE 6. Semantic retrieval results on Corel5k

6. **Conclusions and Future Work.** In this paper, we have presented a semantic context model for the task of automatic image annotation based on a novel conditional random field. First of all, a PLSA model with asymmetric modalities is constructed to predict a candidate set of annotations with confidence scores. Afterwards the semantic relationships between these candidates are built by using the CRF, in which the unary and pairwise potential functions are formulated based on the posterior probabilities calculated by the PLSA model and the normalized Google distance, respectively. Subsequently the deepest gradient descent algorithm is adopted to estimate the parameters of the CRF. Finally, the annotation results are obtained by marginalized probability reasoning. Conducted experiments on Corel5k indicate that our model outperforms peer methods in the literature in terms of accuracy, efficiency and robustness.

As for future work, there are still several issues to be further explored. First, due to the advantage of CRF in general is that the conditional probability model can depend on arbitrary non-independent characteristics of the observation, unlike a generative image model which is forced to account for dependencies in the image, and therefore requires strict independence assumptions to make inference tractable. The down side of applying CRF is that inferring labels from the exact posterior distribution for complex graphs is intractable. So how to further relax the independence assumption of CRF is a worthy research direction. Second, the designs of association and interaction potentials are the main task of CRF research. But in general, the potentials are selected empirically and hand-tuned to have better performance. Moreover, this selection also heavily depends on the specific applications. So it is highly desirable to formulate a general and unified potential functions that can be applied to other multimedia related applications. Third, since the conventional unary and pairwise potentials are rudimentarily defined as summation of weighted feature functions. However, potentials of this form usually need an enormous number of features to render satisfactory results which makes their training and inference to be an exhaustively difficult task. In addition, conventional potentials are very sensitive to the parameter initialization and their training might get stuck in local

optima. So how to solve these issues is also well worth exploring. Fourth, as can be seen from the literature, most of the traditional CRF models are only up to second order and it is difficult to incorporate large-scale contextual information. Hence how to construct robust higher order potentials for CRF is also a very valuable research direction. Last but not the least, due to the complementary nature of integrating two or more machine learning techniques that can make them benefit from each other. Based on this recognition, how to efficiently hybridize CRF with other methods based on the tradeoff between computational complexity and model reconstruction error is a valuable research direction in the future.

Acknowledgment. The author would like to sincerely thank the anonymous reviewers for their valuable comments and insightful suggestions that have helped to improve the paper. Also, the author thanks Professor Zhongzhi Shi for stimulating discussions and helpful hints. This work is supported by the National Program on Key Basic Research Project (No.2013CB329502) and the Key Research Project of Baoji University of Arts and Sciences (No.ZK16047).

REFERENCES

[1] J. Li and J. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1075–1088, 2003.

[2] E. Chang, K. Goh, G. Sychay, et al., CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 1, pp. 26–38, 2003.

[3] G. Carneiro, A. Chan, P. Moreno, et al., Supervised learning of semantic classes for image annotation and retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

[4] P. Duygulu, K. Barnard, N. de Freitas, et al., Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, *Proc. of the European Conf. on Computer Vision (ECCV'02)*, pp. 97–112, 2002.

[5] L. Jeon, V. Lavrenko and R. Manmantha, Automatic image annotation and retrieval using cross-media relevance model, *Proc. of the 26th Int'l Conf. on Research and Development in Information Retrieval (SIGIR'03)*, pp. 119–126, 2003.

[6] V. Lavrenko, R. Manmatha and J. Jeon, A model for learning the semantics of pictures, *Advances in Neural Information Processing Systems 16 (NIPS'03)*, pp. 553–560, 2003.

[7] S. Feng, R. Manmatha and V. Lavrenko, Multiple Bernoulli relevance models for image and video annotation, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, pp. 1002–1009, 2004.

[8] J. Liu, B. Wang, M. Li, et al., Dual cross-media relevance model for image annotation, *Proc. of the 15th Int'l Conf. on Multimedia (MM'07)*, pp. 605–614, 2007.

[9] F. Monay and D. Gatica-Perez, Modeling semantic aspects for cross-media image indexing, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1802–1817, 2007.

[10] S. Nikolopoulos, S. Zafeiriou, I. Patras, et al., High order PLSA for indexing tagged images, *Signal Processing*, vol. 93, no. 8, pp. 2212–2228, 2013.

[11] D. Tian, W. Zhang, X. Zhao, et al., Employing PLSA model and max-bisection for refining image annotation, *Proc. of the 20th Int'l Conf. on Image Processing (ICIP'13)*, pp. 3996–4000, 2013.

[12] D. Tian, X. Zhao and Z. Shi, An efficient refining image annotation technique by combining probabilistic latent semantic analysis and random walk model, *Intelligent Automation & Soft Computing*, vol. 20, no. 3, pp. 335–345, 2014.

[13] D. Blei, A. Ng and M. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[14] D. Blei and J. Lafferty, Correlated topic models, *Annals of Applied Statistics*, vol. 1, no. 1, pp. 17–35, 2007.

[15] J. Liu, M. Li, W. Ma, et al., An adaptive graph model for automatic image annotation, *Proc. of the 8th Int'l Workshop on Multimedia Information Retrieval (MIR'06)*, pp. 61–70, 2006.

[16] J. He, M. Li, H. Zhang, et al., Generalized manifold-ranking-based image retrieval, *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 3170–3177, 2006.
[17] B. Xu, J. Bu, C. Chen, et al., Efficient manifold ranking for image retrieval, *Proc. of the 34th Int'l Conf. on Research and Development in Information Retrieval (SIGIR'11)*, pp. 525–534, 2011.
[18] X. He, R. Zemel and M. Carreira-Perpinan, Multiscale conditional random fields for image labeling, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, pp. 695–702, 2004.
[19] X. Xu, Y. Jiang, L. Peng, et al., Ensemble approach based on conditional random field for multi-label image and video annotation, *Proc. of the 19th Int'l Conf. on Multimedia (MM'11)*, pp. 1377–1380, 2011.
[20] L. Yu, J. Xie and S. Chen, Conditional random field based image labeling combining features of pixels, segments and regions, *IET computer vision*, vol. 6, no. 5, pp. 459–467, 2012.
[21] T. Mensink, J. Verbeek and G. Csurka, Tree-structured CRF models for interactive image labeling, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 2, pp. 476–489, 2013.
[22] Y. Wang and S. Gong, Refining image annotation using contextual relations between words, *Proc. of the 6th Int'l Conf. on Image and Video Retrieval (CIVR'07)*, pp. 425–432, 2007.
[23] J. Verbeek and B. Triggs, Scene segmentation with conditional random fields learned from partially labeled images, *Advances in Neural Information Processing Systems 21 (NIPS'08)*, pp. 1–8, 2008.
[24] X. Wang and X. Zhang, A new Laplacian mixture conditional random field model for image labeling, *Proc. of the 35th Int'l Conf. on Acoustics Speech and Signal Processing (ICASSP'10)*, pp. 2118–2121, 2010.
[25] M. Arani and X. Zhang, Generalized Gaussian mixture conditional random field model for image labeling, *Proc. of the IEEE Global Conf. on Signal and Information Processing (GlobalSIP'14)*, pp. 1068–1072, 2014.
[26] Y. Xiang, X. Zhou, Z. Liu, et al., Semantic context modeling with maximal margin conditional random fields for automatic image annotation, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'10)*, pp. 3368–3375, 2010.
[27] Q. Huang, M. Han, B. Wu, et al., A hierarchical conditional random field model for labeling and segmenting images of street scenes, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'11)*, pp. 1953–1960, 2011.
[28] C. Ji, X. Zhou, L. Lin, et al., Labeling images by integrating sparse multiple distance learning and semantic context modeling, *Proc. of the 12th European Conf. on Computer Vision (ECCV'12)*, pp. 688–701, 2012.
[29] S. Kumar and M. Hebert, A hierarchical field framework for unified context-based classification, *Proc. of the 10th Int'l Conf. on Computer Vision (ICCV'05)*, pp. 1284–1291, 2005.
[30] S. Wang, A. Quattoni, L. Morency, et al., Hidden conditional random fields for gesture recognition, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 1–7, 2006.
[31] Y. Wang and S. Gong, Conditional random field for natural scene categorization, *Proc. of the 18th British Machine Vision Conference (BMVC'07)*, pp. 1–10, 2007.
[32] P. Awasthi, A. Gagrani and B. Ravindran, Image modeling using tree structured conditional random fields, *Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence (IJCAI'07)*, pp. 2060–2065, 2007.
[33] C. Galleguillos, A. Rabinovich and S. Belongie, Object categorization using cooccurrence, location and appearance, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'08)*, pp. 1–8, 2008.
[34] C. Wojek and B. Schiele, A dynamic conditional random field model for joint labeling of object and scene classes, *Proc. of the 10th European Conf. on Computer Vision (ECCV'08)*, pp. 733–747, 2008.
[35] P. Kohli, L. Ladicky and P. Torr, Robust higher order potentials for enforcing label consistency, *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302–324, 2009.
[36] L. Ladicky, C. Russell, P. Kohli, et al., Associative hierarchical CRFs for object class image segmentation, *Proc. of the 14th Int'l Conf. on Computer Vision (ICCV'09)*, pp. 739–746, 2009.
[37] J. Shotton, J. Winn, C. Rother, et al., Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context, *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2–23, 2009.
[38] M. Ibrahim and M. El-Saban, Higher order potentials with superpixel neighbourhood for semantic image segmentation, *Proc. of the 18th Int'l Conf. on Image Processing (ICIP'11)*, pp. 2881–2884, 2011.
[39] P. Krahenbuhl and V. Koltun, Efficient inference in fully connected crfs with Gaussian edge potentials, *Advances in Neural Information Processing Systems 24 (NIPS'11)*, pp. 1–9, 2011.

[40] V. Vineet, J. Warrell and P. Torr, Filter-based mean-field inference for random fields with higher-order terms and product label-spaces, *International Journal of Computer Vision*, vol. 110, no. 3, pp. 290–307, 2014.

[41] S. Bell, P. Upchurch, N. Snavely, et al., Material recognition in the wild with the materials in context database, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'15)*, pp. 3479–3487, 2015.

[42] L. Chen, G. Papandreou, I. Kokkinos, et al., Semantic image segmentation with deep convolutional nets and fully connected crfs, *In: arXiv 1412.7062v3*, 2015.

[43] S. Zheng, S. Jayasumana, B. Romera-Paredes, et al., Conditional random fields as recurrent neural networks, *Proc. of the 20th Int'l Conf. on Computer Vision (ICCV'15)*, pp. 1–17, 2015.

[44] M. Jaderberg, K. Simonyan, A. Vedaldi, et al., Deep structured output learning for unconstrained text recognition, *In: arXiv:1412.5903v5*, 2015.

[45] K. Yao, B. Peng, G. Zweig, et al., Recurrent conditional random field for language understanding, *Proc. of the 39th Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP'14)*, pp. 4077–4081, 2014.

[46] B. Kim, P. Kohli and S. Savarese, 3D scene understanding by voxel-CRF, *Proc. of the 18th Int'l Conf. on Computer Vision (ICCV'13)*, pp. 1425–1432, 2013.

[47] J. Lafferty, A. McCallum and F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, *Proc. of the 18th Int'l Conf. on Machine Learning (ICML'01)*, pp. 282–289, 2001.

[48] S. Kumar and M. Hebert, Discriminative random fields, *International Journal of Computer Vision*, vol. 68, no. 2, pp. 179–201, 2006.

[49] R. Cilibrasi and M. Paul, The Google similarity distance, *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, 2007.

[50] J. Shi and J. Malik, Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[51] Y. Deng and B. Manjunath, Unsupervised segmentation of color-texture regions in images and video, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 800–810, 2001.

[52] Z. Li, Z. Shi, X. Liu, et al., Fusing semantic aspects for image annotation and retrieval, *Journal of Visual Communication and Image Representation*, vol. 21, no. 8, pp. 798–805, 2010.