A Revisit of Generative Model for Automatic Image Annotation using Markov Random Fields

Yu Xiang Xiangdong Zhou
Fudan Unviersity
Shanghai, China
{072021109, xdzhou}@fudan.edu.cn

Tat-Seng Chua
National University
Singapore
chuats@comp.nus.edu.sg

Chong-Wah Ngo
City University
HongKong, China
cwngo@cs.cityu.edu.hk

Abstract

Much research effort on Automatic Image Annotation (AIA) has been focused on Generative Model, due to its well formed theory and competitive performance as compared with many well designed and sophisticated methods. However, when considering semantic context for annotation, the model suffers from the weak learning ability. This is mainly due to the lack of parameter setting and appropriate learning strategy for characterizing the semantic context in the traditional generative model. In this paper, we present a new approach based on Multiple Markov Random Fields (MRF) for semantic context modeling and learning. Differing from previous MRF related AIA approach, we explore the optimal parameter estimation and model inference systematically to leverage the learning power of traditional generative model. Specifically, we propose new potential function for site modeling based on generative model and build local graphs for each annotation keyword. The parameter estimation and model inference is performed in local optimal sense. We conduct experiments on commonly used benchmarks. On Corel 5000 images [3], we achieved 0.36 and 0.31 in recall and precision respectively on 263 keywords. This is a very significant improvement over the best reported result of the current state-of-the-art approaches.

1. Introduction

Automatic Image Annotation (AIA) becomes increasingly important due to its potential in many interesting applications, such as keyword based image and video retrieval and browsing. However, a major bottleneck of AIA is the so-called semantic gap problem due to the mismatch between visual perception and high-level semantics. To deal with this challenge, various AIA models, mostly based on the discriminative models and the generative probabilistic models, have been proposed in the current literature. Discriminative model treats AIA as a classification problem, by treating each semantic concept or keyword as a class. Earlier studies were devoted to develop binary classifiers, while most recent works viewed the problem as a multi-class classification. Generative model, on the other hand, focuses on learning the correlations between visual features and semantic concepts. An influential work is the Cross-Media Relevance Model (CMRM) [5], which estimates the joint probability of visual-based keywords and text-based semantic keywords from training samples. CMRM was subsequently improved by Continuous Relevance Model (CRM) [8] and Multiple Bernoulli Relevance Model (MBRM) [4], which are recognized as the state-of-the-art approaches in AIA.

In addition to learning from visual features, the context relationship among semantic concepts is another vivid clue which could be employed for inferring the semantics of images. For instance, “bird” and “tree” are co-occurred frequently as the semantic labels of images. Intuitively speaking, this hints higher confidence of labeling a new image as “bird”, if knowing that there is also a high probability for “tree” presents in the image. Such context relationship has indeed been exploited in both discriminative and generative models. The former extends AIA as a multi-label classification problem [13], while the later exploits the correlations between keywords [11][16].

While generative model such as CRM and MBRM have shown very competitive performance, the learning ability, specifically when context relationship being considered, remains limited. The weak learning ability is mainly due to the lack of proper parameter setting for modeling semantic context. On one hand, most approaches emphasize model simplicity by using fewer parameters [8][4], resulting in over abbreviation of the model for context estimation. On the other hand, it becomes natural to expect that parameter optimization can pose serious computational problem if more parameters are included. While there is a trade-off between model simplicity and annotation effectiveness, existing approaches, such as CLM [6] and DCMRM [11] developed based upon CRM for modeling semantic context, adopt simple parametric model and offer only limited performance improvement as compared to CRM and MBRM.

Different from previous studies [11][6][13][14], we revisit the generative model by addressing the learning of semantic context when more parameters mandatory for modeling the relationship are considered. We adopt Multiple Markov Random Field (MRF) to boost the potential of the traditional generative model for AIA problem. Specifically, we model the context relationship among semantic concepts with keyword subgraphs generated from training samples for each keyword. We present new site potential function based on generative model for adaptively label prediction. The model parameters are learnt by maximum pseudo-likelihood with Gaussian prior for regularization. In addition, our model determines the number of semantic labels of an image automatically and is robust to the inherent data imbalance problem – a challenge often comes alongside with most training sets with semantic labels.

Differing from previous MRF related AIA, such as CML [13] which focuses on global keyword graph building and ignores the parameter estimation of MRF, our main contribution is that *we fully explore the learning ability of Multiple MRFs to realize the full potentials of the widely studied traditional generative models for AIA*. Our approach provides a better mean of modeling when more parameters are indeed mandatory for characterizing the underlying semantic context. Therefore, we achieved very significant improvement on annotation performance. In our experiment on Corel dataset [3] we achieved 0.36 and 0.31 respectively in recall and precision, which is a significant improvement over the best reported results. We also reported very encouraging results on TRECVID dataset.

The rest of the paper is organized as follows: Section 2 reviews the related work. Section 3 presents the model setting for MRF, while sections 4 and 5 outline our approaches for parameter estimation and model inference respectively. Section 6 details the AIA procedure using MRF. Section 7 presents the experimental results, and Section 8 concludes this paper.

2. Related Work

A significant amount of research efforts have been devoted to the problem of AIA. Generative model based methods attempt to estimate the joint probability of image and keywords. Duygulu et al. [3] used a machine translation model to link keywords and image regions. Jeon et al. [5] proposed cross-media relevance model (CMRM) to estimate the joint probability of keywords and image using discrete blobs to represent regions. It was subsequently improved by continuous relevance model (CRM) [8] and multiple Bernoulli relevance model (MBRM) [4]. Liu et al. [11] proposed a dual cross-media relevance model (DCMRM), which integrates keyword relationship, image retrieval, and web search techniques together to infer the semantics of image. Wang et al. [14] proposed a Markov model-based image annotation (MBIA) method, in which keywords are treated as the states of a Markov chain. Discriminative model based methods apply classification techniques to train classifiers for image labeling. Yang et al. [15] proposed an asymmetrical support vector machine for region-based image annotation. Carneiro et al. [2] proposed a supervised multi-class labeling (SML) approach, which estimates the class density based on image-level and class-level Gaussian mixtures. To utilize keyword correlation in the annotation process, multi-label classification techniques receive more attentions nowadays. Kang et al. [7] extended the standard label propagation algorithms to propagate multiple labels.

Markov random fields are widely used in many computer vision problems, such as image segmentation [12], object detection [10], etc. In these applications, MRFs are used for modeling the spatial relationships between pixels. Recently, Cao et al. [1] applied conditional random fields (CRF) based on event and scene model for photo annotation. Qi et al. [13] proposed a correlative multi-label (CML) annotation framework which simultaneously classifies concepts and models their correlations for video annotation. It is related to MRF, but is limited to global keyword graph building while lacking focus on MRF model estimation.

3. Multiple Markov Random Fields Based Automatic Image Annotation

In this section, we first give a brief introduction to MRF theory, and then detail the construction of our MRFs for image annotation.

3.1. Markov Random Field

A set of random variables $\mathcal{F} = \{f_1, f_2, \cdots, f_m\}$ is said to be a Markov random field on sites $\mathcal{S} = \{1, 2, \cdots, m\}$ with respect to a neighborhood system $\mathcal{N} = \{\mathcal{N}_i | i \in \mathcal{S}\}$, where $\mathcal{N}_i$ is the set of sites neighboring i, if and only if the two following conditions are satisfied:

$$P(\mathbf{f}) > 0, \forall \mathbf{f} \in \mathbb{F}, \tag{1}$$

$$P(f_i | f_{\mathcal{S}-\{i\}}) = P(f_i | f_{\mathcal{N}_i}), \forall i \in \mathcal{S}, \tag{2}$$

where $\mathbf{f} = (f_1, f_2, \cdots, f_m)^T$ is a random variable vector and $f_{\mathcal{A}} = \{f_i | f_i \in \mathcal{F} \text{ and } i \in \mathcal{A}\}$. Equ. 2 indicates that a random variable only interacts with its neighboring variables. The Hammersley-Clifford theorem states that every MRF obeys the following distribution:

$$P(\mathbf{f}) = Z^{-1} \times e^{-U(\mathbf{f})}, \tag{3}$$

where

$$Z = \sum_{\mathbf{f}} e^{-U(\mathbf{f})} \tag{4}$$

is a normalizing constant called partition function, and $U(\mathbf{f})$ is the energy function. It is the sum of clique potentials $V_c(\mathbf{f})$ over all possible cliques $\mathcal{C}$. In this paper, we only consider cliques of order up to two. So the energy function can be reduced to

$$U(\mathbf{f}) = \sum_{i \in \mathcal{S}} V_1(f_i) + \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}). \qquad (5)$$

Detailed introduction about MRFs and their applications in computer vision can be found in [9].

3.2. Keyword Graph

In our framework, the construction of the graph structure of MRF is based on the keyword correlations extracted from training set $\mathcal{T} = \left\{(\mathbf{d}^k, \mathbf{f}^k)\right\}_{k=1}^{K}$, where $\mathbf{d}^k$ is the feature vector of the kth image, $\mathbf{f}^k$ is the corresponding label vector, and K is the size of the training set. $\mathbf{f}^k = (f_1^k, f_2^k, \cdots, f_{|\mathcal{V}|}^k)^T$, where $f_i^k \in \{-1, +1\}$ indicates the absence or presence of keyword w_i in a pre-defined vocabulary set $\mathcal{V}$. In the training set, each image is associated with a set of keywords, which is similar to the so called “bag-of-words” text representation model in text retrieval. We consider each training image as a document, and the associated keywords as the words in the document. Thus the training set can be viewed as a corpus. We then use keyword co-occurrence in the corpus to define the correlations between keywords. Specifically, if two keywords co-occur in the corpus, we consider them to be correlated. Based on the so-defined correlations between keywords, we build a keyword graph as follows. Let the keyword set be $\mathcal{S} = \{1, 2, \cdots, m\}$, where $i \in \mathcal{S}$ corresponds to keyword w_i in vocabulary $\mathcal{V}$. We construct a graph $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ on keyword set $\mathcal{S}$, where $(i, i') \in \mathcal{E}$ if and only if i and i' are correlated.

3.3. Generative Model based Potential Function

Instead of building a single MRF on the keyword graph $\mathcal{G}$ as in [13], we construct MRFs one for each keyword in the vocabulary $\mathcal{V}$ to capture different semantics among keywords. In order to define the sites and neighborhood system of the MRF for keyword w_i, we extract a subgraph $\mathcal{G}_i = (\mathcal{S}_i, \mathcal{E}_i)$ from $\mathcal{G}$, where $\mathcal{S}_i = \{i\} \cup \mathcal{N}_i$, and $\mathcal{E}_i = \{(i,j) | i,j \in \mathcal{S}_i \text{ and } (i,j) \in \mathcal{E}\}$. We treat the keywords in $\mathcal{S}_i$ as the sites, and two sites are neighbors to each other if there is an edge between them. Thus the MRF takes all the keywords correlated with w_i into consideration. In the rest of this section, we discuss the MRF for a single keyword w_i. We still use $\mathcal{S}$ to denote the sites of the single keyword MRF for clarity.

For image annotation task, we employ random variable f_i which takes value from $\{-1, +1\}$ to indicate the absence or presence of keyword w_i for an image, $\forall i \in \mathcal{S}$. The value of f_i is said to be the label of site i. We define the **site potential** as:

$$V_1(f_i) = f_i(\lambda_i + \alpha_i P(\mathbf{d}, w_i)), \qquad (6)$$

where $P(\mathbf{d}, w_i)$ is the joint probability of image feature $\mathbf{d}$ and keyword w_i, which can be obtained from a generative model based image annotation method. And λ_i, α_i are the parameters to be estimated. The motivation of Equ. 6 is, if $\alpha_i < 0$, the more probable label for high $P(\mathbf{d}, w_i)$ is $+1$, which corresponds to lower site potential. We define the **edge potential** as:

$$V_2(f_i, f_{i'}) = \beta_{ii'} f_i f_{i'} P(\mathbf{d}, w_{i'}), \qquad (7)$$

where $\beta_{ii'}$ is the parameter to be estimated. The edge potential incorporates the joint probability of image feature $\mathbf{d}$ and correlated keyword $w_{i'}$. By substituting Equ. 6 and Equ. 7 into Equ. 5, we get the energy function:

$$\begin{aligned} U(\mathbf{f}|\theta) \quad = \quad & \sum_{i \in \mathcal{S}} f_i \left(\lambda_i + \alpha_i P(\mathbf{d}, w_i)\right) + \\ & \sum_{i \in \mathcal{S}} \sum_{i' \in \mathcal{N}_i} \beta_{ii'} f_i f_{i'} P(\mathbf{d}, w_{i'}), \end{aligned} \qquad (8)$$

where θ denotes the parameters of the MRF. Noting that in Equ. 8, we assume image feature $\mathbf{d}$ has been observed.

Most existing approaches based on generative model can be directly incorporated into the proposed MRF framework. In our case, we employ MBRM [4] to estimate $P(\mathbf{d}, w)$, which is the expectation computed over the images in the training set. Since each keyword appears in an image only once, it is more appropriate to describe annotation keywords with Bernoulli distribution. Meanwhile, a beta prior (conjugate to a Bernoulli) is applied for smoothing. For details please refer to [4].

Up to now, we have outlined the construction of MRF for depicting the semantic context of keyword. We will further present the estimation of parameters for the energy function in next section.

4. Parameter Estimation

4.1. Pseudo-likelihood

The widely used technique for parameter estimation in MRFs is maximum likelihood, which chooses the parameters that maximize the joint probability (Equ. 3) of labels (likelihood of parameters). However, evaluating the partition function (Equ. 4) is intractable in practice, because the number of configurations is exponential to the size of the sites. So we adopt an approximation scheme called pseudo-likelihood to avoid the evaluation of the partition function [9]. The pseudo-likelihood is defined as

$$PL(\mathbf{f}) = \prod_{i \in \mathcal{S}} P(f_i | f_{\mathcal{N}_i}) = \prod_{i \in \mathcal{S}} \frac{e^{-U_i(f_i, f_{\mathcal{N}_i})}}{\sum_{f_i} e^{-U_i(f_i, f_{\mathcal{N}_i})}}, \qquad (9)$$

where

$$U_i(f_i, f_{\mathcal{N}_i}) = V_1(f_i) + \sum_{i' \in \mathcal{N}_i} V_2(f_i, f_{i'}), \qquad (10)$$

is the energy introduced by site i. Because f_i and $f_{\mathcal{N}_i}$ are not independent, the pseudo-likelihood is not the true likelihood. Substituting Equ. 6 and Equ. 7 into Equ. 10, we can get:

$$\begin{aligned} U_i(f_i, f_{\mathcal{N}_i}) &= f_i\left(\lambda_i + \alpha_i P(\mathbf{d}, w_i)\right) + \\ & \sum_{i' \in \mathcal{N}_i} \beta_{ii'} f_i f_{i'} P(\mathbf{d}, w_{i'}). \qquad (11) \end{aligned}$$

Let

$$\theta_i = (\lambda_i, \alpha_i, \beta_{ii' \forall i' \in \mathcal{N}_i})^T, \qquad (12)$$

$$\mathbf{x}_i = (1, P(\mathbf{d}, w_i), f_{i'} P(\mathbf{d}, w_{i'})_{\forall i' \in \mathcal{N}_i})^T, \qquad (13)$$

then we can rewrite Equ. 11 to

$$U_i(f_i, f_{\mathcal{N}_i}) = f_i \theta_i^T \mathbf{x}_i, \qquad (14)$$

where θ_i is the parameter associated with site i, and $\mathbf{x}_i$ is the training data constructed for site i. Substituting Equ. 14 into Equ. 9, the pseudo-likelihood is given by

$$PL(\mathbf{f}) = \prod_{i \in \mathcal{S}} \frac{e^{-f_i \theta_i^T \mathbf{x}_i}}{e^{-\theta_i^T \mathbf{x}_i} + e^{\theta_i^T \mathbf{x}_i}}. \qquad (15)$$

The parameters $\theta = (\theta_1^T, \theta_2^T, \cdots, \theta_{|\mathcal{S}|}^T)^T$ are estimated by maximizing the pseudo-likelihood with regularization on the training images.

4.2. Maximum Pseudo-likelihood with Regularization

Suppose we have constructed a training data set $\mathcal{T} = \{(\mathbf{x}^k, \mathbf{f}^k)\}_{k=1}^K$ for the working MRF, where $\mathbf{x}^k = \{\mathbf{x}_1^k, \mathbf{x}_2^k, \cdots, \mathbf{x}_{|\mathcal{S}|}^k\}$, $\mathbf{x}_i^k$ is defined as in Equ. 13 for the kth image, and $\mathbf{f}^k = (f_1^k, f_2^k, \cdots, f_{|\mathcal{S}|}^k)^T$, f_i^k is the label of site i for the kth image. Then the pseudo-likelihood on the training set $\mathcal{T}$ is

$$\begin{aligned} \prod_{k=1}^K PL(\mathbf{f}^k) &= \prod_{k=1}^K \prod_{i \in \mathcal{S}} P(f_i^k | f_{\mathcal{N}_i}^k) \\ &= \prod_{i \in \mathcal{S}} \prod_{k=1}^K P(f_i^k | f_{\mathcal{N}_i}^k) = \prod_{i \in S} PL_i, \quad (16) \end{aligned}$$

where

$$PL_i = \prod_{k=1}^K P(f_i^k | f_{\mathcal{N}_i}^k) \qquad (17)$$

is the pseudo-likelihood on site i. Because there is no shared parameter between any PL_i, the maximum pseudo-likelihood estimation $\theta = (\theta_1^T, \theta_2^T, \cdots, \theta_{|\mathcal{S}|}^T)^T$ of Equ. 16 can be obtained by maximize PL_i to get the parameters θ_i (Equ. 12) for each sites. Note, this property not only speeds up the parameter estimation process significantly, but also enables us to estimate the parameters on different sites with their own training data sets. With the specific training set for each site of the MRF, the problem of data imbalance can be mitigated in some extent. Now we concentrate on maximizing PL_i to get the pseudo-likelihood estimation of θ_i.

Suppose we have constructed a training set $\mathcal{T}_i = \{(\mathbf{x}_i^k, f_i^k)\}_{k=1}^{K_i}$ for site i, then the log pseudo-likelihood on site i is

$$\begin{aligned} \ln PL_i &= \sum_{k=1}^{K_i} \ln P(f_i^k | f_{\mathcal{N}_i}^k) \\ &= \sum_{k=1}^{K_i} \left\{ (1 - f_i^k) \theta_i^T \mathbf{x}_i^k - \ln(1 + e^{2\theta_i^T \mathbf{x}_i^k}) \right\}. \quad (18) \end{aligned}$$

The excessive number of parameters can cause over-fitting problem when there is insufficient training examples available. To deal with this problem, we penalize the log pseudo-likelihood Equ. 18 with a spherical Gaussian weight prior:

$$L_i(\theta_i) = \sum_{k=1}^{K_i} \left\{ (1 - f_i^k) \theta_i^T \mathbf{x}_i^k - \ln(1 + e^{2\theta_i^T \mathbf{x}_i^k}) \right\} - \frac{\|\theta_i\|^2}{2\sigma^2}, \qquad (19)$$

where the value of σ is chosen empirically and constrained to be the same for all sites. To maximize Equ. 19, we set its derivatives to zero. These score equations are

$$\frac{\partial L_i(\theta_i)}{\partial \theta_i} = \sum_{k=1}^{K_i} \left\{ \mathbf{x}_i^k \left(1 - f_i^k - 2P(\mathbf{x}_i^k; \theta_i)\right) \right\} - \frac{\theta_i}{\sigma^2}, \quad (20)$$

where

$$P(\mathbf{x}_i^k; \theta_i) = \frac{e^{2\theta_i^T \mathbf{x}_i^k}}{1 + e^{2\theta_i^T \mathbf{x}_i^k}}. \qquad (21)$$

To solve the score equations Equ. 20, we employ the Newton-Raphson algorithm, which requires the Hessian matrix

$$\frac{\partial^2 L_i(\theta_i)}{\partial \theta_i \partial \theta_i^T} = -4 \sum_{k=1}^{K_i} \left\{ \mathbf{x}_i^k \mathbf{x}_i^{k^T} P(\mathbf{x}_i^k; \theta_i) \left(1 - P(\mathbf{x}_i^k; \theta_i)\right) \right\} - \frac{\mathbf{I}}{\sigma^2}, \qquad (22)$$

where $\mathbf{I}$ is the identity matrix. Starting with θ_i^{old}, a single Newton-Raphson update is

$$\theta_i^{\text{new}} = \theta_i^{\text{old}} - \left(\frac{\partial^2 L_i}{\partial \theta_i \partial \theta_i^T} \right)^{-1} \frac{\partial L_i}{\partial \theta_i}, \qquad (23)$$

where the derivatives are evaluated at θ_i^{old}. The Newton-Raphson algorithm will converge, because the penalized log pseudo-likelihood Equ. 19 is concave.

5. Model Inference

The inference problem in MRFs is to find the most probable configuration of the sites:

$$\mathbf{f}^* \leftarrow \arg\max_{\mathbf{f}} P(\mathbf{f}), \tag{24}$$

where $P(\mathbf{f})$ is given by Equ. 3. We employ an algorithm called iterative conditional modes (ICM) for inference, which maximizes local conditional probabilities sequentially. In the $(k+1)$th iteration step, given the image feature $\mathbf{d}$ and the other labels $f_{\mathcal{S}-\{i\}}^{(k)}$, the algorithm sequentially updates each $f_i^{(k)}$ into $f_i^{(k+1)}$ by maximizing the conditional probability $P(f_i|\mathbf{d}, f_{\mathcal{S}-\{i\}}^{(k)})$. Because in a MRF, f_i only depends on the labels in its neighborhood, we can equivalently maximize

$$P(f_i|\mathbf{d}, f_{\mathcal{N}_i}^{(k)}). \tag{25}$$

Maximizing Equ. 25 is equivalent to minimizing the corresponding potential using the following rule

$$f_i^{(k+1)} \leftarrow \arg\min_{f_i} U_i(f_i, f_{\mathcal{N}_i}), \tag{26}$$

which is equivalent to

$$f_i^{(k+1)} = \begin{cases} 1, & \text{if} \quad \theta_i^T \mathbf{x}_i \leq 0 \\ -1, & \text{if} \quad \theta_i^T \mathbf{x}_i > 0 \end{cases}, \tag{27}$$

where θ_i is the estimated parameter of site i, and $\mathbf{x}_i$ is the training data constructed for site i based on the image feature. Starting from an initial configuration, the iteration continues until convergence, and then we can get the most probable labels of the sites.

6. Image Annotation

In this section, we outline the algorithms for MRF learning and image annotation.

6.1. Training Set Construction

In order to perform parameter estimation, we construct training data for each site of the MRF from training data set $\mathcal{T}$. Suppose we want to build a training set $\mathcal{T}_i$ for site i, which is corresponding to keyword w_i. We first sample the training set $\mathcal{T}$ to get a new set $\mathcal{T}_i^{'}$ of size K_i with a more balanced positive and negative samples for keyword w_i, where the positive samples are images labeled with keyword w_i. Sampling is helpful to deal with the data imbalance problem in the training set, because in practical systems, there are far more negative samples than the positive ones. We utilize all the positive samples of a keyword and randomly select a subset of negative samples whose size is larger than the positive sample set by a small factor δ, where $\delta = 1$ in our experiment. The reason is that if we have sufficient positive samples, the additional negative sample would have little effect on the built model. On the other hand if the semantic is hard to capture because of the lack of enough positive samples, then the extra negative sample can prevent the model from generating excessive false positives. Second, for each image $\mathbf{d}^k$ in the training set $\mathcal{T}_i^{'}$, we extract the labels corresponding to site i and all its neighboring sites $i' \in \mathcal{N}_i$, and calculate the joint probabilities $P(\mathbf{d}^k, w_i)$ and $P(\mathbf{d}^k, w_{i'})$ on these sites. Finally, we combine the labels and the joint probabilities to form a training set $\mathcal{T}_i = \{(\mathbf{x}_i^k, f_i^k)\}_{k=1}^{K_i}$, where $\mathbf{x}_i^k$ is defined as in Equ. 13 for the kth image, and f_i^k is the label of site i for the kth image. Algorithm 1 is the procedure for training set construction.

Algorithm 1 Training Set Construction

```
Input: global training set $\mathcal{T}$, working MRF $MRF$
Output: training set $\mathcal{T}^{''}$ for $MRF$
for each site $i$ of $MRF$ do
    Sample $\mathcal{T}$ to get a much balanced data set $\mathcal{T}_i^{'}$
    for each $\mathbf{d}^k \in \mathcal{T}_i^{'}$ do
        Extract labels $f_i^k$ and $f_{i'}^k, \forall i' \in \mathcal{N}_i$
        Calculate $P(\mathbf{d}^k, w_i)$ and $P(\mathbf{d}^k, w_{i'}), \forall i' \in \mathcal{N}_i$
        Calculate $\mathbf{x}_i^k = (1, P(\mathbf{d}^k, w_i), f_{i'}^k P(\mathbf{d}^k, w_{i'})_{\forall i' \in \mathcal{N}_i})^T$
    end for
    $\mathcal{T}_i = \{(\mathbf{x}_i^k, f_i^k)\}_{k=1}^{K_i}$
end for
$\mathcal{T}^{''} = \bigcup_{i=1}^{|\mathcal{S}|} \mathcal{T}_i$
```

6.2. Annotation Algorithm

After parameter estimation on the constructed training set, the annotation process is straightforward. Note, for an input image I, each MRF will output a label vector, but only the corresponding label, say the w_i for the ith MRF, will be considered as the most confidential one and treated as the label for I. After performing inference on all the MRFs, we obtain the annotation of the image. Our Markov Random Fields based Image Annotation method- MRFA is summarized in Algorithm 2. Note if we annotate a collection of

Algorithm 2 MRFA: Markov Random Field Image Annotation Process

```
Input: an unlabeled image $I$, keyword vocabulary $\mathcal{V}$,
    training set $\mathcal{T}$, constructed keyword graph $\mathcal{G}$
Output: labels of image $I$
for each $w \in \mathcal{V}$ do
    Extract a subgraph $\mathcal{G}_w$ from $\mathcal{G}$ for $MRF_w$
    Construct training set $\mathcal{T}_w^{''}$ for $MRF_w$ by Alg. 1
    Estimate the parameters of $MRF_w$ based on $\mathcal{T}_w^{''}$
    Perform inference of $I$ on $MRF_w$ to get the label
end for
```

images, the keyword subgraphs, training sets construction and parameter estimation for each MRF only need perform once.

7. Experiment

7.1. Experimental Dataset and Evaluation

Corel Dataset: We use Corel image dataset [3] for experiments. The dataset is widely used in AIA for performance comparison. It consists of 5000 images , where 4500 images are for training and the rest for testing. Each image is labeled with 1-5 keywords, and a total of 374 different keywords are in the dataset. Each image is segmented into 1-10 regions. For each region, a 36-dimensional feature vector is extracted [3]. In addition to region-based features, grid-based features are also used by CRM and MBRM. Here we also introduce a new grid feature. We partitioned each image into 26 rectangular grids (5×5 plus one extra center grid), and extracted 528 dimensions feature vector for each grid, namely 448 color features (including local and global color histogram) and 80 edge features extracted according to MPGE7. In the experiment, we perform testing using both region-based and grid-based features. We append the name of an approach with '-grid', if our grid-based feature is used. For example, MBRM-grid means MBRM using our grid-based features.

TRECVID Dataset: To evaluate our approach for video annotation, we also conduct experiments on the benchmark TRECVID 2005 dataset, which contains about 170 hours of multi-lingual broadcast news. These videos are automatically segmented into 61,901 shots. Each shot is further segmented into 5 grid, and a 45-dimensional visual feature vector is extracted for each grid. Thus each shot has a 225-dimensional feature vector. There are 39 different keywords in the dataset, and each shot is associated with 0-11 keywords. We construct the training set with 9,000 randomly sampled shots and the test set with another 1,000 randomly sampled shots. Every sampled shot is labeled with at least one keyword.

Evaluation Measures: Similar to previous work for image annotation, we use recall and precision to measure the annotation performance. Given a query word w, let $|W_G|$ be the number of human annotated images with label w in the test set, $|W_M|$ be the number of annotated images with the same label of the annotation algorithm, and $|W_C|$ be the number of correct annotations of our algorithm, then recall and precision are defined as $recall = \frac{|W_C|}{|W_G|}$ and $precision = \frac{|W_C|}{|W_M|}$.

Table 1. Performance comparison with MBRM on Corel dataset using region-based features

Models	MBRM	MRFA
#words with recall > 0	109	124
Average #words/image	5	4.3
Results on all 263 words		
Mean Per-word Recall	0.20	0.23
Mean Per-word Precision	0.19	0.27
Results on 49 best words		
Mean Per-word Recall	0.68	0.67
Mean Per-word Precision	0.64	0.76

Table 2. Performance comparison with MBRM and single MRF on Corel dataset using grid-based features

Models	MBRM	MRFA	MRF-s
#words with recall > 0	123	172	136
Average #words/image	5	5.2	9.6
Results on all 263 words			
Mean Per-word Recall	0.25	0.36	0.28
Mean Per-word Precision	0.23	0.31	0.20
Results on 49 best words			
Mean Per-word Recall	0.75	0.79	0.69
Mean Per-word Precision	0.73	0.80	0.63

7.2. Experiments Results

7.2.1 Comparison on Corel Dataset

Since MBRM is the representative generative model based AIA approach with very competitive performance, we first compare our annotation framework with MBRM on the Corel dataset using region-based features [3]. Because most previous work cannot automatically determine the optimal annotation length, for MBRM, we fix the size of each image annotations to 5 as in [4], it shows best performance in experiment. While our approach can automatically decide the size of the annotation. The results are shown in Table 1. From the table, we can see that as compared with MBRM, our proposed MRFA method improves the annotation performance significantly. For all 263 words appearing in the test set, it gains 15% on average recall and 42% on average precision respectively. For the best 49 keywords with largest F1 scores, it gains 19% on average precision while the average recall is nearly the same. Overall, our method labels 4.3 keywords for each image on average, which is less than MBRM of 5. Also, our method has 124 keywords with recall > 0 as compared with 109 of MBRM, which means that our method has better performance on labeling rare keywords which are hard to annotate due to the small number of positive instances in the training set.

By using the grid-based visual features, both the performance of MBRM and our MRFA improved significantly as compared to using region features. The results are shown

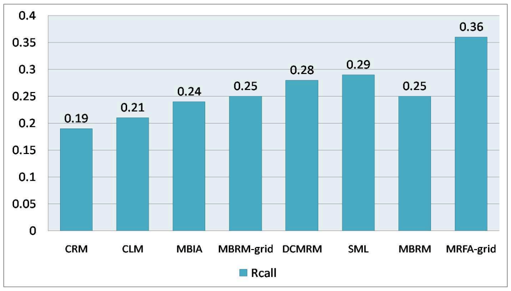

Figure 1. The annotation performance compared with other methods by Recall

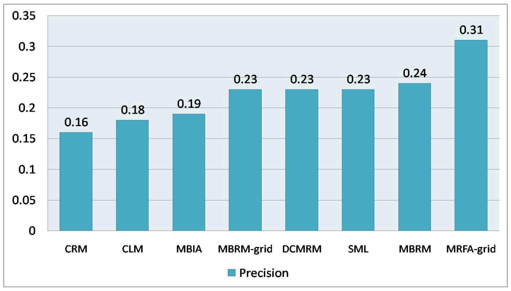

Figure 2. The annotation performance compared with other methods by Precision

in Table 2. For all 263 keywords, our method has 172 keywords with recall > 0, which is a significant 40% improvement over MBRM. The average recall and average precision of MRFA is 0.36 and 0.31 which again indicates significant improvement of 44% and 35% respectively over MBRM. For the best F1 49 words, our model also has significant improvement on average recall and average precision. Overall, the experimental results demonstrate our approach has strong ability to improve annotation accuracy and label rare keywords. Our analysis shows that the performance improvement of our method is mainly contributed by our proposed new MRF model instead of our grid-based visual features. To compare our multiple MRF with the method of using global graph MRF, we also show the annotation performance of using a single MRF (denoted by MRF-s in the experiment) for all the 374 keywords in Table 2, which indicates that by training multiple MRFs, MRFA avoids a global optimal parameter setting which is hard to estimate, so achieves better annotation performance.

Besides of MBRM [4], we also compare our approach to five other different state-of-the-art AIA methods, including generative model: CRM [8], CLM [6], DCMRM [11], and discriminative model: MBIA [14], and SML [2]. Figure 1 and 2 show the comparative performance in terms of recall and precision between our MRFA method and the state-of-the-art approaches. Our method achieves the best precision and recall, and the improvement is more than 24% as compared with the second best performing system.

Table 3. Performance comparison with MBRM on TRECVID dataset

Models	MBRM	MRFA
#words with recall > 0	32	39
Average #words/image	5	3.62
Results on all 39 words		
Mean Per-word Recall	0.39	0.47
Mean Per-word Precision	0.32	0.45

Figure 3 gives some examples of annotation results of our method and MBRM on Corel dataset. It shows that our method not only covers the correct annotation keywords labeled by MBRM, but also labels more true keywords and avoids some false alarms. For example, the annotation result of MRFA for the first image and the last image are the same as the ground-truth, while MBRM has false alarms. For the third image, our MRFA even labeled a keyword “caribou ”, which should be the true keyword for this image, but was ignored by the human annotators.

7.2.2 Comparison on TRECVID Dataset

For video data, we compare our method with MBRM on TRECVID 2005 dataset. We fix the number of annotation keywords per video shot for MBRM to be 5, which achieves the best performance in our experiments. The experimental results are given in Table 3. From the Table, we can see that as compared to MBRM, our method can predict all the 39 words in the annotation vocabulary, and it achieves improvement of 21% and 41% respectively on average recall and average precision, while labeling each shot with fewer keywords. Figure 4 gives details of annotation performance of each keyword as compared to MBRM. It shows that for most keywords our method has significant improvement on precision as compared with MBRM. For recall, we have 14 keywords better than MBRM, 17 keywords equal to MBRM. MRFA performs satisfactorily for rare keywords such as ”Mountain”, ”Prisoner” and ”Truck” that cannot be predicted by MBRM.

8. Conclusion

We have presented the formulation of Markov Random Fields to empower the learning ability of generative model for AIA problem. Such formulation is demonstrated to be appropriate for learning the context relationship of semantic concepts. The newly proposed potential function for optimal parameter estimation and model inference, in particular, shows significant impact on the learning ability. Our approach also offers great ability in labeling rare keywords and adaptive determination of the number of keywords for image annotation. We verified the performance of our approach through extensive experiments on commonly used benchmarks. Particularly, we reported the state-of-the-art

	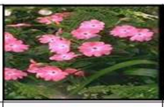	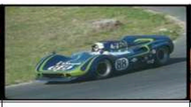	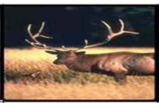	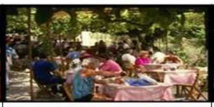	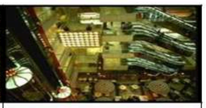
MRFA-grid	**leaf, flowers, petals, stems**	**grass, cars, tracks, prototype**	**grass, cow, bulls, antlers, elk, caribou**	**people, flowers, restaurant, shops, street, festival**	**light, shops**
MBRM-grid	**sky, water, flowers, bush, petals**	**water, grass, cars, tracks, prototype**	**sky, water, grass, antlers, elk**	**water, flowers, display, shops, street**	**sky, water, tree, light, shops**
Ground-truth	**leaf, flowers, petals, stems**	**cars, tracks, turn, prototype**	**cow, bulls, antlers, elk**	**tree, people, restaurant, tables**	**light, shops**

Figure 3. Some annotation examples on Corel dataset

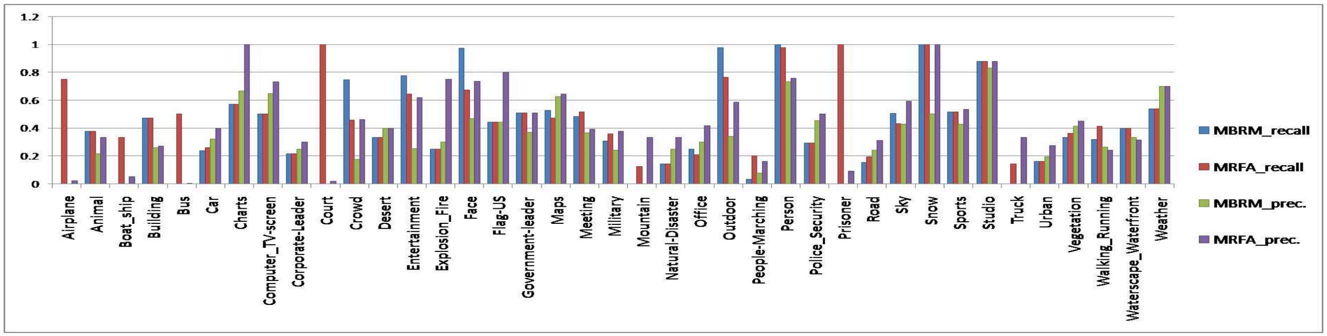

Figure 4. Comparison of MRFA and MBRM on TRECVID dataset for 39 keywords. Please see color version for more clarity

performance on Corel dataset, showing significant improvement over six other existing approaches based on generative and discriminative models.

For future work, we will focus on two directions. One direction investigates the scalability issue when there are thousands keywords to be annotated. One possibility is to explore the use of one keyword subgraph for a class of keywords rather than one graph per keyword as it is currently done with great effectiveness. Another direction is to improve annotation performance by leveraging on WordNet or Web resource in building keyword graph.

9. Acknowledgment

This work was partially supported by the NSFC under Grant No.60403018 and No.60773077.

References

[1] L. Cao, J. Luo, H. Kautz, and T. Huang. Annotating collections of photos using hierarchical event and scene models. *CVPR*, 2008.

[2] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE PAMI*, 29, 2007.

[3] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. *ECCV*, 2002.

[4] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. *CVPR*, 2004.

[5] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. *SIGIR*, 2003.

[6] R. Jin, Y. Chai, and L. Si. Effective automatic image annotation via a coherent language model and active learning. *ACM Multimedia*, 2004.

[7] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. *ACM Multimedia*, 2006.

[8] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. *NIPS*, 2004.

[9] S. Z. Li. Markov random field modeling in computer vision. *Springer-Verlag Press*, 1995.

[10] Y. Li, Y. Tsin, Y. Genc, and T. Kanade. Object detection using 2d spatial ordering contraints. *CVPR*, 2005.

[11] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. *ACM SIGMM*, 2007.

[12] B. Micusik and T. Pajdla. Multi-label image segmentation via max-sum solver. *CVPR*, 2007.

[13] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, and H. Zhang. Correlative multi-label video annotation. *ACM SIGMM*, 2007.

[14] C. Wang, L. Zhang, and H. Zhang. Scalable markov model-based image annotation. *CIVR*, 2008.

[15] C. Yang and M. Dong. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. *CVPR*, 2006.

[16] X. Zhou, M. Wang, J. Zhang, Q. Zhang, and B. Shi. Automatic image annotation by an iterative approach: Incorporating keyword correlations and region matching. *CIVR*, 2007.