6-2010

Semantic context modeling with maximal margin conditional random fields for automatic image annotation

Yu XIANG

Xiangdong ZHOU

Zuotao LIU

Tat-Seng CHUA

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Semantic Context Modeling with Maximal Margin Conditional Random Fields for Automatic Image Annotation

Yu Xiang Xiangdong Zhou Zuotao Liu
Fudan Unviersity
Shanghai, China
{072021109, xdzhou, 082024020}@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

Context modeling for Vision Recognition and Automatic Image Annotation (AIA) has attracted increasing attentions in recent years. For various contextual information and resources, semantic context has been exploited in AIA and brings promising results. However, previous works either casted the problem into structural classification or adopted multi-layer modeling, which suffer from the problems of scalability or model efficiency. In this paper, we propose a novel discriminative Conditional Random Field (CRF) model for semantic context modeling in AIA, which is built over semantic concepts and treats an image as a whole observation without segmentation. Our model captures the interactions between semantic concepts from both semantic level and visual level in an integrated manner. Specifically, we employ graph structure to model contextual relationships between semantic concepts. The potential functions are designed based on linear discriminative models, which enables us to propose a novel decoupled hinge loss function for maximal margin parameter estimation. We train the model by solving a set of independent quadratic programming problems with our derived contextual kernel. The experiments are conducted on commonly used benchmarks: Corel and TRECVID data sets for evaluation. The experimental results show that compared with the state-of-the-art methods, our method achieves significant improvement on annotation performance.

1. Introduction

Context modeling for Vision Recognition and Automatic Image Annotation (AIA) receives increasing attentions nowadays due to the progress in human vision understanding and the encouraging results of preliminary studies on context modeling [16]. Such approaches can be broadly classified into object based contextual models and holistic image based contextual models [20]. The former accounts for object co-occurrence [19] or spatial relationships between objects [23, 17] based on object segmentation, while the latter treats an image as a whole and utilizes the statistical summary of the scene [15, 13]. In the research of automatic image annotation, holistic image based contextual models attract more attentions due to the potential for large scale image and video search [20].

Figure 1. Illustration of semantic context by example images from Corel image data set and their human annotations.

Semantic concepts co-occur frequently in an image, such as "bird" and "tree", "car" and "track", and so on. Some illustrative images of Corel data set [4] are presented in Figure 1. Intuitively, knowing an image labeled with "bird" or "car" provides hint of labeling it with "tree" or "track" respectively. Similarly, if two semantic concepts never occur together, exploiting the contextual relationships between them helps reducing the number of "false positives". A few recent works have been proposed to utilize contextual relationships between semantic concepts in AIA. One paradigm performs annotation refinement by modeling semantic context using independent multi-layer model, e.g. semantic context layer and visual perception layer, such as Dirichlet mixture model [20] and Markov Random Field (MRF) [24], which are based on some previous AIA methods. The other paradigm casts AIA into structural classification problem, where an image is annotated with multiple concepts simultaneously and concept correlations are utilized in the annotation process [18]. The structural SVM [22] was adopted to solve the classification problem. However, training the

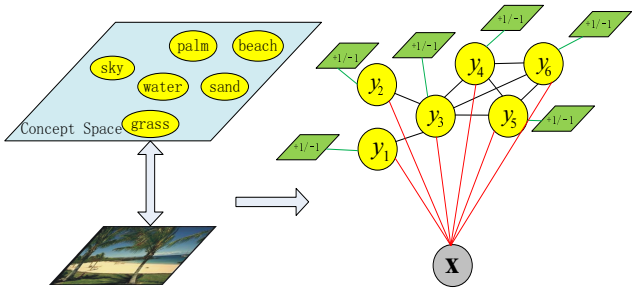


Figure 2. The framework of our model for semantic context modeling in AIA.

structural SVM results in a quadratic programming (QP) problem with the number of constraints exponential in the number of concepts, which is not scalable to the concept space (the set of concepts and their interactions).

In this paper, we propose a novel discriminative Conditional Random Field (CRF) [10] model for semantic context modeling in AIA. Different from the previous CRFs in object based contextual models [19, 17], our CRF model is built over semantic concepts to model the interactions between them. Compared with the previous semantic contextual models [18, 20, 24], our method captures the interactions between semantic concepts from both semantic level and visual level in an integrated manner, which leverages the ability of semantic context modeling. Specifically, the sites in our CRF model correspond to semantic concepts and the edges represent the correlations between the concepts. A binary label is associated to each site to indicate the presence/absence of the corresponding concept in an image. The potentials are designed based on linear discriminative models, where the edge potential is formulated as a visual dependent smoothing function to facilitate context modeling. We design a novel margin loss function for maximal margin parameter estimation, where the traditional hinge loss is decoupled into a set of sub-hinge losses. The parameter estimation is performed by solving a set of independent QP problems with our derived contextual kernel, which makes our model more scalable to the concept space. Figure 2 graphically illustrates the framework of our CRF model. We apply the proposed Maximal Margin Conditional Random Field (MMCRF) model to AIA and conduct experiments on Corel and TRECVID-2005 data sets. Compared with both contextual and non-contextual the state-of-the-art methods in AIA, our model achieves significant improvement on annotation performance.

The rest of the paper is organized as follow: Section 2 reviews some related work. Section 3 presents the model setting for our MMCRF model. Section 4 and 5 detail the maximal margin parameter estimation and model inference respectively. Section 6 presents the experiments, and Section 7 concludes the paper.

2. Related Work

A significant amount of work have been devoted to the problem of AIA. Generative models [7, 11, 5] focus on learning the correlations between images and semantic concepts, while discriminative models formulate AIA as a classification problem and apply classification techniques to AIA, such as Support Vector Machine (SVM) [3] and Gaussian mixture model [2]. Yang et al. [25] proposed an Asymmetrical Support Vector Machine-based Multiple-Instance Learning (ASVM-MIL) algorithm for AIA. Recently, some efforts have been devoted to semantic context modeling in AIA. Rasiwasia and Vasconcelos [20] used mixtures of Dirichlet distributions to model the correlations between semantic concepts. Xiang et al. [24] employed Markov Random Field (MRF) to boost the potential of traditional generative models. Qi et al. [18] proposed a Correlative Multi-Label (CML) annotation framework for video annotation. Guillaumin et al. [6] proposed the TagProp algorithm based on KNN method and achieved very competitive annotation performance on Corel data set.

Conditional Random Field (CRF) [10] is proposed for segmenting and labeling 1-D sequence data initially. Its 2-D version is called Discriminative Random Field (DRF) [9], which is used to model the spatial dependencies in images. Taskar et al. [21] proposed the Max-Margin Markov Network (M^3N) for multi-label classification by exploiting structure among class labels in the maximal margin framework. Our work differs from DRF and M^3N in that we introduce the decoupled hinge loss into the maximal margin learning of CRF and solve the problem by solving a set of independent QPs with the derived contextual kernel.

3. Conditional Random Fields

A set of random variables $\mathbf{y} = \{y_1, y_2, \cdots, y_m\}$ is said to be a conditional 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, given an observation $\mathbf{x} \in \mathcal{X}$ if and only if the following two conditions are satisfied:

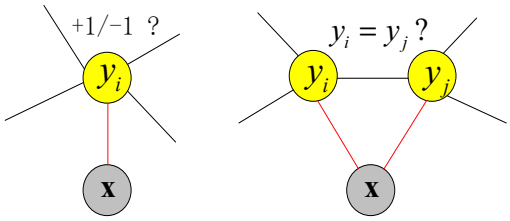
$$P(\mathbf{y}|\mathbf{x}) > 0, \forall \mathbf{y} \in \mathcal{Y}, \tag{1}$$

$$P(y_i|\mathbf{x}, y_{\mathcal{S}-\{i\}}) = P(y_i|\mathbf{x}, y_{\mathcal{N}_i}), \forall i \in \mathcal{S}, \tag{2}$$

where $y_{\mathcal{A}} = \{y_i | i \in \mathcal{A}\}$. Equation (2) indicates that a random variable only interacts with its neighboring random variables given an observation. The Hammersley-Clifford theorem [14] states that every CRF obeys the following distribution:

$$p(\mathbf{y}|\mathbf{x}) = Z^{-1} \times e^{-U(\mathbf{x},\mathbf{y})}, \tag{3}$$

where Z is a normalizing constant called the partition function and $U(\mathbf{x}, \mathbf{y})$ is the energy function, which is the sum of clique potentials $V_c(\mathbf{x}, \mathbf{y})$ over all possible cliques $\mathcal{C}$. In



(a) (b)

Figure 3. The potential design of our MMCRF model: (a) site potential, (b) edge potential.

this paper, we only consider cliques of order up to two, so the energy function can be reduced to

$$U(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{S}} V_1(\mathbf{x}, y_i) + \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{N}_i} V_2(\mathbf{x}, y_i, y_j). \quad (4)$$

3.1. Site Potential

In our CRF framework, $\mathbf{y}$ represents labels of semantic concepts and $\mathbf{x}$ represents image features. The site potential $V_1(\mathbf{x}, y_i)$ is modeled using a local discriminative model which decides the label of $y_i \in \{-1, +1\}$ based on the observation $\mathbf{x}$ ignoring its neighbors. Figure 3(a) illustrates the idea for site potential design. We employ linear models for their well studied theory and define the site potential as:

$$V_1(\mathbf{x}, y_i) = -y_i(\mathbf{w}_i^T \phi_i(\mathbf{x}) + b_i), \quad (5)$$

where $\mathbf{w}_i$ and b_i are the parameters associated with site i, and ϕ_i is a function that maps the observation $\mathbf{x}$ on a feature space related with class i. Note that $y_i(\mathbf{w}_i^T \phi_i(\mathbf{x}) + b_i)$ is the functional margin of the example $(\phi_i(\mathbf{x}), y_i)$ with respect to hyperplane $(\mathbf{w}_i, b_i)$, so increasing the margin lows down the potential.

3.2. Edge Potential

The edge potential $V_2(\mathbf{x}, y_i, y_j)$ is also modeled using a linear model. But different from site linear models, edge linear models work to decide whether the labels of a pair of sites should be the equal or not based on the observation $\mathbf{x}$. They can be considered as data (visual) dependent smoothing functions. Figure 3(b) shows the design of edge potential in our model. By fixing the bias parameter in edge linear model to zero, we have edge potential as:

$$V_2(\mathbf{x}, y_i, y_j) = -y_i y_j \mathbf{w}_{ij}^T \phi_j(\mathbf{x}), \quad (6)$$

where $\mathbf{w}_{ij}$ is the parameter associated with edge (i, j), and ϕ_j is a function that maps the observation $\mathbf{x}$ on a feature space related with class j. The edge parameter $\mathbf{w}_{ij}$ is not symmetric in our model. If we consider $y_i y_j$ as the label of the observation $\phi_j(\mathbf{x})$ in edge linear model, then large margin of the edge linear model corresponds to small potential too. By substituting Equations (5) and (6) into (4), we get the energy function:

$$U(\mathbf{x}, \mathbf{y}, \mathbf{w}, \mathbf{b}) = -\sum_{i \in \mathcal{S}} y_i(\mathbf{w}_i^T \phi_i(\mathbf{x}) + b_i) - \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{N}_i} y_i y_j \mathbf{w}_{ij}^T \phi_j(\mathbf{x}), \quad (7)$$

where $\mathbf{w}$ and $\mathbf{b}$ denote the parameters of the CRF model.

3.3. Concept Graph

In our MMCRF model, the construction of the graph structure is based on the the co-occurrences of concepts in a training data set $\mathcal{T} = \{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^T$, where T denotes the size of the training set. Two concepts co-occur if they are associated with the same observation in the training set. We then use concept co-occurrences to define a correlation measure between concepts:

$$P(y_j | y_i) = \frac{|y_i \bigcap y_j|}{|y_i|}, \quad (8)$$

which is the estimation of the prior conditional probability of observing y_j on condition of y_i. Based on the above measure, we define site j is a neighbor of site i, i.e. $j \in \mathcal{N}_i$, if and only if $P(y_j|y_i) \geq P_0$, $\forall i, j \in \mathcal{S}$, where P_0 is a predefined threshold constant. The constructed neighborhood system is not symmetric since the interaction between two concepts is not mutually equal.

4. Maximal Margin Parameter Estimation

In energy based learning [12], there is no requirement for proper normalization. Parameter estimation seeks for an energy function that ensures the labels corresponding to the minimum value of the energy function is the correct label configuration of a given observation. In order to evaluate the quality of a specific energy function, we define a loss function which incorporates the loss on a training data set and our prior knowledge about the task. Therefore, the aim of parameter estimation becomes to finding the parameters which produce the lowest value of the loss function.

4.1. Decoupled Hinge Loss

We utilize hinge loss to perform parameter estimation in the proposed CRF model. Hinge loss is known as a margin loss which creates an energy gap between the correct answer and the incorrect ones. It is used in support vector machines which are recognized as the state-of-the-art classifiers. The hinge loss of a training sample $(\mathbf{x}^t, \mathbf{y}^t)$ can be defined as:

$$L_{\text{hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) = \max\Big(0, m + U(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) - U(\mathbf{x}^t, \bar{\mathbf{y}}^t, \mathbf{w}, \mathbf{b})\Big), \quad (9)$$

where m is the positive margin and $\bar{\mathbf{y}}^t$ is the most offending incorrect answer:

$$\bar{\mathbf{y}}^t = \arg \min_{\mathbf{y}\in\mathcal{Y} \text{ and } \mathbf{y}\neq\mathbf{y}^t} U(\mathbf{x}^t, \mathbf{y}, \mathbf{w}, \mathbf{b}). \quad (10)$$

The hinge loss is linear with the difference between the energies of the correct answer and the most offending incorrect answer when it is larger than $-m$. By substituting Equations (7) into (9) and rearranging the sums, we can get:

$$\begin{aligned} &L_{\text{hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) \\ &= \max\Big(0, m + \sum_{i\in\mathcal{S}} (\bar{y}_i^t - y_i^t)(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i) \\ &+ \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{N}_i} (\bar{y}_i^t \bar{y}_j^t - y_i^t y_j^t) \mathbf{w}_{ij}^T \phi_j(\mathbf{x}^t)\Big). \end{aligned} \quad (11)$$

In the above loss function, the most offending incorrect answer is unknown. Thus, it needs to explore the label space to ensure the margin, which results in an optimization problem with the number of constraints exponential in the number of labels. To deal with the problem, structural SVM [22] maintains a working set of active constraints, while M^3N [21] utilizes the graph structure to reduce the constraints into polynomial size. However, both of them lead to complicated optimization problems. Here, based on the design of our potential functions, we propose a decoupled hinge loss where the most offending incorrect answers of the site linear models and the edge linear models are simply the negative of the corresponding true labels:

$$\begin{aligned} &L'_{\text{hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) \\ &= \sum_{i\in\mathcal{S}} \max\Big(0, m_i - 2y_i^t(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i)\Big) \\ &+ \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{N}_i} \max\Big(0, m_{ij} - 2y_i^t y_j^t \mathbf{w}_{ij}^T \phi_j(\mathbf{x}^t)\Big), \end{aligned} \quad (12)$$

where m_i, $i \in \mathcal{S}$ are the margins of the site linear models and m_{ij}, $i \in \mathcal{S}$, $j \in \mathcal{N}_i$ are the margins of the edge linear models. According to the following proposition, we can use the loss of Equation (12) instead of the loss of Equation (11) for parameter estimation.

Proposition 1. If $m \leq \sum_{i\in\mathcal{S}, \bar{y}_i^t = -y_i^t} m_i + \sum_{i\in\mathcal{S}, j\in\mathcal{N}_i, \bar{y}_i^t \bar{y}_j^t = -y_i^t y_j^t} m_{ij}$, then $L'_{\text{hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b})$ is an upper bound of $L_{\text{hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b})$.

Indicated by the proposition, if we set m smaller than some threshold, decreasing the decoupled loss L'_{hinge} (12) will also diminish the original loss L_{hinge} (11), since L'_{hinge} is an upper bound of L_{hinge}. Using L'_{hinge} enables us to find the parameters by searching for the maximal margin hyperplane of each linear model separately. However, the site and edge linear models interact with each other during inference. So the error propagation among the linear models can degrade the model performance. To alleviate the problem, we adopt a compromise between L_{hinge} and L'_{hinge}. That is we decouple L_{hinge} so that the parameters are estimated site by site, where a site linear model $(\mathbf{w}_i, b_i)$ and its neighboring edge linear models $\mathbf{w}_{ij}, j \in \mathcal{N}_i$ are learnt simultaneously. The corresponding decoupled hinge loss is

$$\begin{aligned} &L''_{\text{hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) \\ &= \sum_{i\in\mathcal{S}} \max\Big(0, m_i - 2y_i^t(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i) \\ &- \sum_{j\in\mathcal{N}_i} 2y_i^t y_j^t \mathbf{w}_{ij}^T \phi_j(\mathbf{x}^t)\Big). \end{aligned} \quad (13)$$

In L''_{hinge}, the margins of edge linear models are not specified. So the upper bound of m cannot be formulated as in Proposition 1. However, the existence of the upper bound of m can still ensure L''_{hinge} is an upper bound of L_{hinge}. So we can use it to perform parameter estimation.

4.2. Biased Regularization

To keep the model from overfitting, we add a regularization term in our loss function based on prior knowledge about AIA task. Since the interaction parameters on the edges tend to be overestimated [9], we need to penalize the edge linear models more in practice. Therefore, we introduce two different parameters in the regularization term corresponding to the two kinds of linear models. The biased regularization term is defined as follows:

$$R(\mathbf{w}) = \lambda_1 \sum_{i\in\mathcal{S}} \|\mathbf{w}_i\|^2 + \lambda_2 \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{N}_i} \|\mathbf{w}_{ij}\|^2, \quad (14)$$

where λ_1 and λ_2 are two constants controlling the penalty of the site linear models and edge linear models respectively.

4.3. Parameter Estimation Framework

Given a training data set $\mathcal{T} = \{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^T$, we combine the per-sample decoupled hinge loss (13) and the biased regularization (14) to obtain the loss function:

$$L(\mathcal{T}, \mathbf{w}, \mathbf{b}) = \frac{1}{T} \sum_{t=1}^{T} L''_{\text{hinge}}(\mathbf{x}^t, \mathbf{y}^t, \mathbf{w}, \mathbf{b}) + R(\mathbf{w}). \quad (15)$$

Following the methodology of support vector machines, we obtain the primal form of the optimization problem for parameter estimation:

$$\begin{aligned} &\min_{\mathbf{w},\mathbf{b},\boldsymbol{\xi}} \frac{1}{2}\Big(\sum_{i\in\mathcal{S}} \|\mathbf{w}_i\|^2 + \lambda \sum_{i\in\mathcal{S}} \sum_{j\in\mathcal{N}_i} \|\mathbf{w}_{ij}\|^2\Big) + C \sum_{i\in\mathcal{S}} \sum_{t=1}^{T} \xi_i^t \\ &\text{s.t. } y_i^t\big(\mathbf{w}_i^T \phi_i(\mathbf{x}^t) + b_i + \sum_{j\in\mathcal{N}_i} y_j^t \mathbf{w}_{ij}^T \phi_j(\mathbf{x}^t)\big) \geq 1 - \xi_i^t, \\ &\quad \xi_i^t \geq 0, \ \forall i \in \mathcal{S}, \forall t, \end{aligned} \quad (16)$$

where $C = \frac{1}{\lambda_1 T}$ and $\lambda = \frac{\lambda_2}{\lambda_1}$ are two constants and $\boldsymbol{\xi}$ denotes the introduced slack variables.

4.4. Algorithm for Solving the Optimization Problem with Contextual Kernel

Accordingly, the problem of our maximal margin parameter estimation becomes the optimization problem (16), which can be decoupled into $|\mathcal{S}|$ subproblems, one for each site. Therefore, we can perform parameter estimation site by site. The subproblem for site i can be formulated as:

$$\begin{aligned}
&\min_{\mathbf{w},\mathbf{b},\boldsymbol{\xi}} \frac{1}{2}\Big(\|\mathbf{w}_i\|^2 + \lambda \sum_{j\in\mathcal{N}_i} \|\mathbf{w}_{ij}\|^2 \Big) + C\sum_{t=1}^{T_i} \xi_i^t \\
&\text{s.t. } y_i^t\big(\mathbf{w}_i^T\phi_i(\mathbf{x}^t) + b_i + \sum_{j\in\mathcal{N}_i} y_j^t \mathbf{w}_{ij}^T \phi_j(\mathbf{x}^t)\big) \geq 1 - \xi_i^t, \forall t \\
&\quad \xi_i^t \geq 0, \forall t. \qquad (17)
\end{aligned}$$

A dedicated training set $\mathcal{T}_i = \{(\mathbf{x}^t, \mathbf{y}^t)\}_{t=1}^{T_i}$ is selected from the global training set $\mathcal{T}$ for site i, which enables us to use more balanced positive and negative samples in $\mathcal{T}_i$. By performing Lagrangian transformation, we obtain the dual form of (17):

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} -\frac{1}{2}\sum_{t=1}^{T_i}\sum_{t'=1}^{T_i} \alpha_i^t \alpha_i^{t'} y_i^t y_i^{t'} K_i(\mathbf{x}^t, \mathbf{x}^{t'}) + \sum_{t=1}^{T_i} \alpha_i^t \\
&- \frac{1}{2\lambda}\sum_{t=1}^{T_i}\sum_{t'=1}^{T_i}\sum_{j\in\mathcal{N}_i} \alpha_i^t \alpha_i^{t'} y_i^t y_j^t y_i^{t'} y_j^{t'} K_j(\mathbf{x}^t, \mathbf{x}^{t'}) \\
&\text{s.t.} \quad \sum_{t=1}^{Ti} \alpha_i^t y_i^t = 0,\ C \geq \alpha_i^t \geq 0,\ \forall t, \qquad (18)
\end{aligned}$$

where $\boldsymbol{\alpha}$ denotes the dual variables. As in the conventional SVM, we have replaced the inner products of two observations $\langle \phi_i(\mathbf{x}^t) \cdot \phi_i(\mathbf{x}^{t'})\rangle$ and $\langle \phi_j(\mathbf{x}^t) \cdot \phi_j(\mathbf{x}^{t'})\rangle$ with kernel functions $K_i(\mathbf{x}^t, \mathbf{x}^{t'})$ and $K_j(\mathbf{x}^t, \mathbf{x}^{t'})$ respectively. So we need not perform the mapping ϕ explicitly, but only design the kernels for different sites. A close examination of the dual form (18) reveals that it is the dual form of the maximal margin classifier of site i with the derived kernel function:

$$K(\mathbf{x}^t, \mathbf{x}^{t'}) = K_i(\mathbf{x}^t, \mathbf{x}^{t'}) + \frac{1}{\lambda}\sum_{j\in\mathcal{N}_i} y_j^t y_j^{t'} K_j(\mathbf{x}^t, \mathbf{x}^{t'}). \qquad (19)$$

Note that the above derived kernel function utilizes not only the observation features but also the labels in the site's neighborhood. Therefore, we refer it as *Contextual Kernel*. The pairwise terms in (19) function as "smooth" kernels which are controlled by a sufficiently large parameter λ to ensure the whole contextual kernel is Positive-Semi-Definite (PSD). The derived contextual kernel enables us to utilize the ordinary SVM algorithms to solve the dual problem (18) while exploiting the semantic context.

4.5. Discriminative Metric Learning for Kernel Construction

Different kinds of visual features, such as color histogram, texture, local appearance, etc, of the observation contribute differently to the semantics. Therefore, we construct semantic oriented kernels using discriminative metric learning technique to combine multiple visual features of the observations [6]. We calculate a base distance between two observations on each kind of features. Finally, the distance between two observations is defined as the weighted sum of the base distances:

$$d_{\mathbf{w}}(\mathbf{x}^t, \mathbf{x}^{t'}) = \mathbf{w}^T \mathbf{d}_{tt'}, \qquad (20)$$

where $\mathbf{d}_{tt'}$ is a vector of base distances between $\mathbf{x}^t$ and $\mathbf{x}^{t'}$, and $\mathbf{w}$ is the weight to be learnt from a training set. Generalized Gaussian kernel is used based on the distance:

$$K_{\mathbf{w}}(\mathbf{x}^t, \mathbf{x}^{t'}) = e^{-g d_{\mathbf{w}}(\mathbf{x}^t, \mathbf{x}^{t'})}, \qquad (21)$$

where g is the width of the Gaussian kernel. To learn the parameter $\mathbf{w}$ in the kernel, we utilize the metric learning technique in nearest neighbor models [6]. But the difference is that we learn kernel K_i using positive samples in each class i. So we obtain different semantic oriented kernels for all the sites.

5. Model Inference

The inference problem in CRFs is to find the most compatible configuration of the sites with a given observation. Specifically, inference produces the labels with the smallest energy value given an observation $\mathbf{x}$:

$$\mathbf{y}^* = \arg\min_{\mathbf{y}\in\mathcal{Y}} U(\mathbf{x}, \mathbf{y}), \qquad (22)$$

where $U(\mathbf{x}, \mathbf{y})$ is defined in (4). Exhaustive search is intractable in practice, since the number of configurations is exponential to the size of the sites. So we employ an algorithm called iterated conditional modes (ICM) [1] for inference. ICM updates the labels sequentially by maximizing local conditional probabilities, which is equivalent to minimizing the following local energy functions for each site:

$$\begin{aligned}
U_i(\mathbf{x}, y_i, y_{\mathcal{N}_i}) &= V_1(\mathbf{x}, y_i) + \sum_{j\in\mathcal{N}_i} V_2(\mathbf{x}, y_i, y_j) \\
&= -y_i(\mathbf{w}_i^T\phi_i(\mathbf{x}) + b_i) - \sum_{j\in\mathcal{N}_i} y_i y_j \mathbf{w}_{ij}^T \phi_j(\mathbf{x}). \qquad (23)
\end{aligned}$$

In the $(t+1)$th step, given the observation $\mathbf{x}$ and the neighboring labels $y_{\mathcal{N}_i}^{(t)}$, the algorithm sequentially updates each $y_i^{(t)}$ into $y_i^{(t+1)}$ using the following rule:

$$y_i^{(t+1)} = \arg\min_{y_i} U_i(\mathbf{x}, y_i, y_{\mathcal{N}_i}^{(t)}), \qquad (24)$$

which is equivalent to

$$y_i^{(t+1)} = \begin{cases} +1, \text{ if } \mathbf{w}_i^T \phi_i(\mathbf{x}) + b_i + \sum_{j \in \mathcal{N}_i} y_j^{(t)} \mathbf{w}_{ij}^T \phi_j(\mathbf{x}) \geq 0 \\ -1, \text{ otherwise.} \end{cases} \tag{25}$$

The rule using dual variables and kernels is

$$y_i^{(t+1)} = \begin{cases} +1, \text{ if } \sum_{t=1}^{T_i} \alpha_i^t y_i^t K_i(\mathbf{x}^t, \mathbf{x}) + b_i \\ + \frac{1}{\lambda} \sum_{t=1}^{T_i} \sum_{j \in \mathcal{N}_i} \alpha_i^t y_i^t y_j^t y_j^{(t)} K_j(\mathbf{x}^t, \mathbf{x}) \geq 0 \\ -1, \text{ otherwise,} \end{cases} \tag{26}$$

where $\boldsymbol{\alpha}$ and b_i are the estimated parameters. Starting from an initial configuration (all labels are set to -1), the iteration continues until convergence. Then we obtain the most compatible label configuration of the observation.

6. Experiments

6.1. Experimental Datasets

Corel Dataset: The Corel dataset [4] is widely used in AIA for performance comparison. It contains 5,000 images, where 4,500 images are used for training and the rest 500 images for testing. Each image is labeled with 1-5 keywords, and there are totally 374 keywords used in the dataset. But most of the keywords have few positive samples. For instance, only 70 of the 374 keywords have positive samples more than 60.

TRECVID-2005 Dataset: The TRECVID-2005 dataset contains about 108 hours of multi-lingual broadcast news, which is more diverse and represents the real world scenario. 61,901 keyframes are extracted from these videos and annotated by 39 concepts. Working with the whole dataset is time consuming. Therefore, we select training and testing data from 90 videos and the other 47 videos respectively. For each concept, we randomly select no more than 500 and 100 positive samples for training and testing respectively. As a result, we have 6,657 keyframes for training and 1,748 keyframes for testing.

6.2. Feature Extraction

We extract different kinds of features commonly used for image search and classification . We use two types of global features: Gist features [15] and color histograms. The color histograms are calculated with 8 bins in each color channel for RGB, LAB and HSV representations, which results in three 512-dimensional feature vectors for each image. For local feature, we use SIFT and adopt the soft-weighting scheme [8] for bag-of-features. All feature vectors but Gist are L1-normalized. To compute the base distances between different types of features, we use L2 as the metric for Gist, L1 for color histograms and χ^2 for SIFT [6].

Table 1. Performance comparison with SVM on Corel dataset. N+, Length, R and P denote the number of keywords with non-zero recall value, average annotation length, average recall and average precision respectively. 263 and 70 denote the 263 keywords appearing in the test set and the largest (most frequent) 70 keywords in the dataset respectively.

Models	SVM	$\lambda = 1$	$\lambda = 60$	$\lambda = 80$	$\lambda = 140$
N+ of 263	81	**146**	97	99	87
Length	4.33	33.85	5.15	4.97	4.90
R of 70	0.5447	0.4725	0.5226	**0.5554**	0.5393
P of 70	0.3983	0.1450	**0.4409**	0.4373	0.3982
N+ of 70	63	63	**68**	67	64

6.3. Evaluation Measures

We use recall, precision and F1 to measure the annotation performance as previous methods. 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, precision and F1 are defined as $recall = \frac{|W_C|}{|W_G|}$, $precision = \frac{|W_C|}{|W_M|}$ and $F1 = \frac{2 \times recall \times precision}{recall + precision}$. We compute recall and precision for each keyword and then average them to measure the annotation performance.

6.4. Experimental Results

6.4.1 Evaluation of Semantic Context Modeling

We choose SVM as the baseline method and compare MMCRF with it on the Corel dataset to evaluate the semantic context modeling. We trained a binary SVM for each keyword in the dataset. It is important to construct different training sets for discriminative models, since the data sets used are always imbalance. To capture the semantics of different keywords, we used all the positive samples for each keyword. Using balanced negative and positive samples for each keyword will predict lots of false positives for keywords with small number of positive samples. So we used more negative samples for these keywords. In order to demonstrate the effect of semantic context and achieve a fair comparison, we fixed the negative sample selection strategy for both models. We used at least 200 negative samples for all the keywords. We also used the same kernels learnt from multiple features by discriminative metric learning for both models. The experimental results are shown in Table 1. We evaluate our MMCRF model at different values of the parameter λ, which controls the interactions between semantic concepts. The larger the value of λ, the less influence of the semantic context. We only present some representative

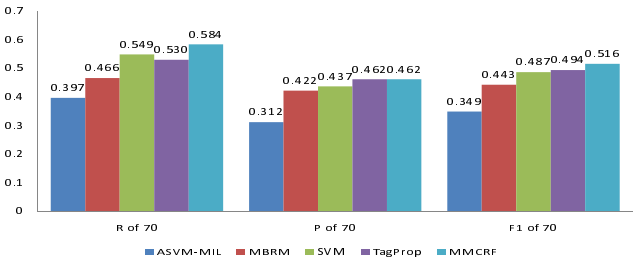


Figure 4. Performance comparison with non-contextual methods: ASVM-MIL, MBRM, SVM and TagProp on Corel dataset.

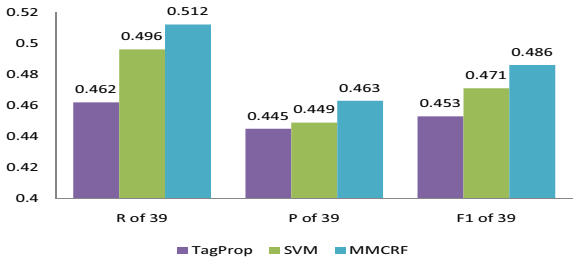


Figure 5. Performance comparison with TagProp and SVM on TRECVID-2005 dataset.

Table 2. Performance comparison with contextual methods: HCM and MRFA on Corel dataset.

Models	MRFA	MMCRF	HCM	MRFA	MMCRF
	Results on 70 keywords		Results on 104 keywords		
N+	69	**70**	87	**97**	88
R	0.518	**0.584**	0.433	0.427	**0.444**
P	0.448	**0.462**	0.359	**0.445**	0.402
F1	0.480	**0.516**	0.393	**0.437**	0.422

results in the table. When $\lambda = 1$, our model predicts 146 of the 263 keywords in the test set, much more than 81 of SVM. However, the average annotation length is more than 30, which means the correlations between keywords dominate and degrade the performance. When $\lambda = 140$, the annotation performance of our model is nearly the same as SVM, which indicates that the semantic context is hardly utilized. For λ values between the two extremes and the average annotation length near 5, our model can predict 99 of the 263 keywords and 68 of the largest 70 keywords, while the result of SVM is 81 and 63 respectively. The average recall and average precision on the largest 70 keywords are improved by 2% and 10% respectively when $\lambda = 80$. So by modeling the semantic context, our model has strong ability to improve annotation accuracy for large keywords, as well as the ability to label rare keywords.

6.4.2 Comparison of AIA Performance on Corel

To further evaluate the effectiveness of our model, we compare it with four non-contextual AIA methods: SVM (with metric learning for kernel generation), ASVM-MIL [25], MBRM [5] and TagProp [6], and two contextual methods: HCM [20] and MRFA [24]. The experimental results of non-contextual methods and ours are shown in Figure 4. We compare the annotation performance of the methods on the largest 70 keywords, since seldom used keywords cannot be effectively learnt by discriminative models due to insufficient positive training samples as noted in [25]. For SVM, TagProp and our MMCRF, we performed metric learning on the five kinds of visual features as mentioned in Section 6.2. For SVM and MMCRF, we determined the number of negative samples for each keyword through cross validation. From the figure we can see that our model achieves the best F1 score. Compared with the second best method TagPop, MMCRF improves the average recall by 10%, while achieving the same average precision as it.

The comparison with the two contextual models: HCM and MRFA is shown in Table 2. Our model outperforms MRFA by 13% and 3% on average recall and average precision of the largest 70 keywords respectively. For HCM, since only the results on the largest 104 keywords is provided in [20], we also evaluate MRFA and MMCRF on the largest 104 keywords. It shows that our MMCRF gains 3% on average recall and 12% on average precision respectively compared with HCM. MMCRF is more sensitive to the number of positive training images due to the adopted max-margin learning approach. Therefore MRFA achieves better F1 score on the largest 104 keywords compared with MMCRF. This is mainly because the 71th to 104th keywords have fewer positive training images, which results in the overall performance degradation on MMCRF. We list some annotation examples of MMCRF compared with the ground-truth annotations in Figure 6. The images are chosen to display different scenes. The annotations of our method are satisfactory in capturing the gist of the images.

6.4.3 Comparison of AIA Performance on TRECVID

We evaluate our model for video annotation on the TRECVID-2005 dataset. Since SVM is regarded as the state-of-the-art method for concept detection in videos and TagProp [6] achieves competitive performance on image annotation, we compare our method with them on the TRECVID-2005 dataset. The experimental results are shown in Figure 5. From the figure we can see that by utilizing the correlations between concepts, our MMCRF model outperforms SVM (with metric learning) by 3% on both average recall and average precision of the 39 concepts. The improvement is not as great as in Corel dataset because the smaller concept space limits the contributions of semantic context. Note that both MMCRF and SVM outperform TagProp on the TRECVID-2005 dataset. In particular, our MMCRF outperforms TagProp by 11% on average recall and 4% on average precision of the 39 concepts. With increase in content diversity, such as in TRECVID-

Figure 6. Comparison of MMCRF annotations with ground-truth annotations on Corel dataset and TRECVID-2005 dataset.

2005 [8], the performance of nearest neighbor based methods degrade. Whereas discriminative models, such as SVM and MMCRF, are more powerful to handle the diversity which often occurs in practical applications. Figure 6 presents some annotation examples of MMCRF compared with ground-truth annotations. For instance, we have perfect matching for the second and fifth keyframes. MMCRF even predicts the ignored concepts by human, such as "crowd" in the third keyframe. It shows that MMCRF has strong ability of handling different scene annotation.

7. Conclusion

In this paper, we proposed a novel discriminative Maximal Margin Conditional Random Field model for semantic context modeling in AIA. Our model inherits the merits of maximal margin learning methods and captures the correlations between semantic concepts during annotation. By designing a novel decoupled hinge loss, our model can be solved by a set of ordinary SVMs with the derived contextual kernel. Extensive experiments conducted on commonly used benchmarks for image and video keyframe annotation show that our model is more capable of utilizing semantic context and handling diverse data in AIA. For the future work, we plan to highlight the contextual kernel by kernel learning framework to further extend our work.

Acknowledgments

This work was partially supported by the NSFC under grant No.60773077.

References

[1] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 1986.

[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] C. Cusano, G. Ciocca, and R. Schettini. Image annotation using svm. *Proceedings of Internet Imaging IV*, 2004.

[4] 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.

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

[6] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. *ICCV*, 2009.

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

[8] Y. Jiang, C. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. *CIVR*, 2007.

[9] S. Kumar and M. Hebert. Discriminative fields for modeling spatial dependencies in natural images. *NIPS*, 2004.

[10] J. Lafferty, A. McCallum, and F. Pereira. Conditonal random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, 2001.

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

[12] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting Structured Data, MIT Press*, 2006.

[13] F.-F. Li and P. Perona. A bayesian hierarchical model for learning nature scene categories. *CVPR*, 2005.

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

[15] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[16] A. Oliva and A. Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, December 2007.

[17] J. Poyway, K. Wang, B. Yao, and S. Zhu. A hierarchical and contextual model for aerial image understanding. *CVPR*, 2008.

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

[19] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. *ICCV*, 2007.

[20] N. Rasiwasia and N. Vasconcelos. Holistic context modeling using semantic co-occurences. *CVPR*, 2009.

[21] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. *NIPS*, 2003.

[22] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. *ICML*, 2004.

[23] L. Wolf and S. Bileschi. A critical view of context. *IJCV*, pages 251–261, 2006.

[24] Y. Xiang, X. Zhou, T. Chua, and C. Ngo. A revisit of generative model for automatic image annotation using markov random fields. *CVPR*, 2009.

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