A Generative-Discriminative Hybrid Method for Multi-View Object Detection

Dong-Qing Zhang
Department of Electrical Engineering
Columbia University, New York, NY 10027
dqzhang@ee.columbia.edu

Shih-Fu Chang
Department of Electrical Engineering
Columbia University, New York, NY 10027
sfchang@ee.columbia.edu

Abstract

We present a novel discriminative-generative hybrid approach in this paper, with emphasis on application in multi-view object detection. Our method includes a novel generative model called Random Attributed Relational Graph (RARG) which is able to capture the structural and appearance characteristics of parts extracted from objects. We develop new variational learning methods to compute the approximation of the detection likelihood ratio function. The variaitonal likelihood ratio function can be shown to be a linear combination of the individual generative classifiers defined at nodes and edges of the RARG. Such insight inspires us to replace the generative classifiers at nodes and edges with discriminative classifiers, such as support vector machines, to further improve the detection performance. Our experiments have shown the robustness of the hybrid approach – the combined detection method incorporating the SVM-based discriminative classifiers yields superior detection performances compared to prior works in multi-view object detection.

1. Introduction

Part-based object detection by learning from example images have been a popular computer vision research topic in recent years. Much of the previous research has been focusing on single view object detection. In contrast, multi-view object recognition aims at recognizing objects and learning object models in images under different views. Multi-view object detection is more challenging as the view point change could result in the spatial constellation change of parts, as well as significant changes in part appearances.

This material is based upon work funded in whole by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

Previous approaches for part-based object detection lie in three categories: generative methods, discriminative methods and hybrid methods that combine generative and discriminative approaches.

Generative method assume an object instance is generated from a statistical model that captures the variations of object structure and appearance. Object detection is realized by calculating the likelihood ratio under positive and negative hypotheses based on the statistical model. One of the well-known generative models is the constellation model [4][12], which captures the spatial relationship among the parts by a joint Gaussian density function, and part appearance by another set of Gaussian density functions. Part correspondence between the object and model is established before detection or learning using a state-space search algorithm, known as A-star algorithm. Learning is realized by a E-M like algorithm by iteratively finding the correspondence and re-estimating the parameters of Gaussian functions. Another type of generative model is pictorial structure , which models the relations among parts locally as pairwise relations. Originally, the pictorial structural model was used to locate the objects and their parts, but more recently [2], they have been extended to generic object detection.

Both the constellation model and the pictorial structure model were only applied to single view object detection. For multi-view object detection, an effective object model should be able to model the occlusion of parts as well as partial relationship among parts resulting from occlusion. Although the constellation model can handle part occlusion, it models the part relations as a global constellation. Whereas, pictorial structure has the capability of modeling partial relationship, yet it does not provide the explicit model for learning the occlusion statistics of individual parts.

While in generative models, learning is conducted in positive examples and negative examples separately to maximize the likelihood, discriminative models directly minimizes the relaxed error function, which makes discriminative models often more accurate. The boosting-based

method is an instance of the discriminative models. Viola and Jones first applied boosting to real-time face detection [10]. Opelt *et al.*[1] then extended boosting to generic mutli-view object detection. In contrast to the generative models, boosting based methods cannot model the relationship among parts. Despite this limitation, the boosting-based methods have performed very well in multi-view object recognition.

It is not difficult to imagine that combining the generative and discriminative approaches could complement two methods. Recently, there have been several attempts to combine the generative and discriminative approaches. For instance, Holub and Perona [7] has developed Fisher kernels based on the constellation model. For every input, Fisher kernel method calculates the Fisher score of the input, and Support vector Machine (SVM) is applied to classification in the Fisher score space. Fisher kernel method is a convenient way to construct a valid kernel for SVM. Yet, it remains unclear whether fisher kernel is an optimal way for SVM based classification, as the optimal decision boundary in the Fisher score space is not necessarily optimal for the original object instances. Another path towards generative and discriminative classification is through boosting, Bar-Hillel *et al.* [6] has presented a boosting-based method based on their own generative model, which, similar to the constellation model, models part relations as a global distribution function.

In this paper, we first propose a new method for generative part-based object modeling, called Random Attributed Relational Graph (RARG or Random ARG). The model is similar to the pictorial structure model, but with extended capabilities in modeling graph topological change, part occlusion and unsupervised learning. The modeling of the topological variation and the node attribute variation make the model an extension of the traditional random graph [3], in which only variation of graph topology is modeled. This is where the term RARG comes from. We will show later the RARG model captures the advantages of both the constellation model and the pictorial structure model. It accommodates part occlusion and models the partial relationship among object parts. Such unique strength makes it a strong candidate for multi-view object detection.

Under the RARG framework, object instances and images are modeled as Attributed Relational Graphs(ARGs). An image ARG consists of an object ARG generated from RARG and additional background parts. Object detection is realized by computing the likelihood ratio of positive and negative hypotheses under this generative model. We realize likelihood computation by constructing a Markov Random Field based on the model RARG and the image ARG. We then show an important relation between the likelihood ratio and the partition functions of the Markov Random Fields. By exploiting the log convexity of the partition functions of MRFs, the logarithm of a partition function can be approximated using variational methods, avoiding exponential complexity of exhaustive search. The approximated partition functions can then be used to compute the likelihood ratio needed for object detection.

The variational approximation/representaiton of the likelihood function is a linear combination of the likelihood ratio functions defined at the individual vertices and edges of the RARG. Our key insight is that each individual likelihood ratio function can be thought of as an classifier. Note for generative models, these likelihood functions are specified by the part appearance distribution functions, which are often Guassian density functions. For multi-view object recognition, the Guassian density or similar single-modal distributions are often inaccurate. However, we can replace the individual likelihood ratio functions with more powerful discriminative classifiers, for instance Support Vector Machines, to obtain better performance. In the learning process, such replacement results in an iterative procedure of estimating part correspondences and training the discriminative classifiers, similar to the original Expectation-Maximization process used in generative learning.

We conduct experiments using the Graz data set [1] to evaluate the performance of the proposed technique and competing solutions for multi-view object detection. The performance of our proposed method is notably better than that of the pure generative approach. It is also better than the performance achieved by the previous boosting-based method. These results confirm the capability of our new generative model of learning informative object model from training data with unlabeled parts, and the superior performance of the incorporated discriminative learning methods.

The paper is organized as follow. In Sec. 2, we introduce the new generative model based on RARG with details of the learning and detection procedures. In Sec. 3, we describe the discriminative extension of the proposed generative model to make it a hybrid approach. In Sec. 4, we present the experimental results and comparisons with previous work.

2. Random ARG for Object Recognition

In our system, an object instance or image is represented as an Attributed Relational Graph [5], formally defined as

Definition 1 *An Attributed Relational Graph(ARG) is defined as a triplet $O = (V, E, Y)$, where V is the vertex set, E is the edge set, and Y is the attribute set that contains attribute y_u attached to each node $n_u \in V$, and attribute y_{uv} attached to each edge $e_w = (n_u, n_v) \in E$.*

For an object instance, a node in the ARG corresponds to one part in the object. The node attribute y_u is a feature vector consisting of appearance features such as moments,

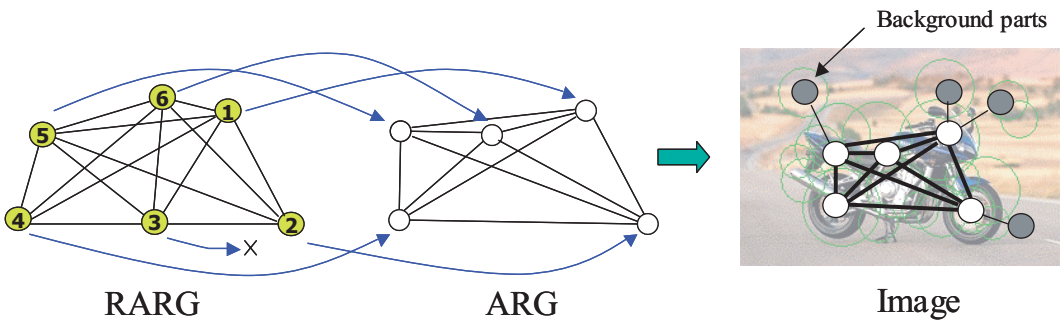

Figure 1. A two-step generative model of objects. Nodes and attributes are first sampled from RARG to form an ARG instance. Then background parts are added to form an image.

colors and spatial features such as spatial coordinates. The edge attribute y_{uv} may include relational features such as coordinate difference or neighborhood adjacency.

For an object model, we use a graph-based representation similar to the ARG but attach random variables to the nodes and edges of the graph. We call such model as Random Attributed Relational Graph

Definition 2 *A Random Attributed Relational Graph (RARG) is defined as a quadruple $R = (V, E, A, T)$, where V is the vertex set, E is the edge set, A is a set of random variables consisting of A_i attached to the node $n_i \in V$ with pdf $f_i(.)$, and A_{ij} attached to the edge $e_k = (n_i, n_j) \in E$ with pdf $f_{ij}(.)$. T is a set of binary random variables, with T_i attached to each node (modeling the presense/absence of nodes).*

$f_i(.)$ is used to capture the statistics of part appearances. $f_{ij}(.)$ is used to capture the statistics of part relational features. T_i is used to handle part occlusion. $r_i = p(T_i = 1)$ is referred to as the *occurrence probability* of part i in the object model.

An ARG hence can be considered as an instance generated from RARG by two steps: first draw samples from $\{T_i\}$ to determine the topology of the ARG, then draw samples from A_i and A_{ij} to obtain the attributes of the ARG and thus the appearance of the object instance.

2.1 Bayesian Classification under Random ARG Framework

We follow the previous work to formulate the object detection problem as a binary classification problem with two hypotheses: $H = 1$ indicates that the image contains the target object (e.g. bike), $H = 0$ otherwise. Let O denote the ARG representation of the input image. Object detection problem therefore is reduced to the likelihood ratio test. An instance of object is said to be detected if

$$\frac{p(O|H=1)}{p(O|H=0)} > \frac{p(H=0)}{p(H=1)} = \lambda \qquad (1)$$

Where λ can be empirically set to adjust the precision and recall performance. The main problem now is thus to compute the positive likelihood $p(O|H = 1)$ and the negative likelihood $p(O|H = 0)$. $p(O|H = 0)$ is the likelihood assuming the image is a *background* image without the target object. Due to the diversity of the *background* images, we adopt a simple decomposable *i.i.d.* model for the background parts. We factorize the negative likelihood as

$$\begin{aligned} p(O|H=0) &= \prod_u p(y_u|H=0) \prod_{uv} p(y_{uv}|H=0) \\ &= \prod_u f^-_{B_1}(y_u) \prod_{uv} f^-_{B_2}(y_{uv}) \end{aligned} \qquad (2)$$

where $f^-_{B_1}(\cdot)$ and $f^-_{B_2}(\cdot)$ are *pdf*s modeling the statistics of the appearance and relations of the parts in the *background* images, referred to as background *pdf*s. The minus superscript indicates that the parameters of the *pdf*s are learned from the negative data set. To compute the positive likelihood $p(O|H = 1)$, we assume that an image is generated by the following generative process (Figure 1): an ARG is first generated from the RARG, additional patches, whose attributes are sampled from the background *pdf*s, are independently added to form the final part-based representation O of the image. In order to compute the positive likelihood, we further introduce a variable X to denote the correspondences between parts in the ARG O and parts in the RARG R. Treating the correspondence X as a hidden variable, we have

$$p(O|H=1) = \sum_X p(O|X, H=1)p(X|H=1) \qquad (3)$$

Where the correspondence X can be represented in different manner. In the previous papers [4][7], X is represented as a integer-valued vector, in which the value of component is the index of the object part. However, X can be also represented as a binary matrix, with $x_{iu} = 1$ if the part i in the object model corresponds to the part u in the image, $x_{iu} = 0$ otherwise. If we assign each x_{iu} a node,

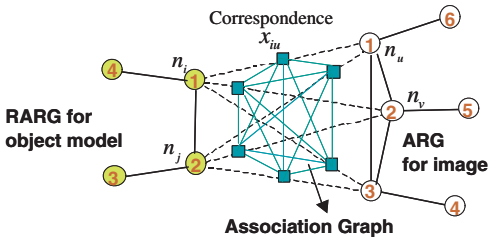

Figure 2. ARG, RARG and the Association Graph.

then these nodes form an Association Graph as shown in Figure 2. The Association Graph can be used to define an undirected graphical model (Markov Random Field) for computing the positive likelihood in Eq. (3). In the rest of the paper, iu therefore is used to denote the index of the nodes in the Association Graph. Compared with the integer-valued representation, binary representation allows more choices of inference algorithms, for instance Belief Optimization algorithms[11][13]. Furthermore, if the size of ARG is large, binary representation allows us to easily prune the MRF by discarding vertices in the association graph that correspond to pairs of dissimilar parts so that the inferencing process can be made faster.

2.2 Design MRF by RARG parameters

The factorization in Eq. (3) requires computing two components $p(X|H=1)$ and $p(O|X,H=1)$. These two terms shall be designed according to the parameters of the RARG.

First, $p(X|H=1)$, the prior probability of the correspondence, is designed in a way that a part in the object can only match at most one part in the model, vice versa. Furthermore, $p(X|H=1)$ is also used to encode the *occurrence probability* r_i. To achieve these, $p(X|H=1)$ is designed as a binary pairwise MRF with the following Gibbs distribution

$$p(X|H=1) = \frac{1}{Z}\prod_{iu,jv}\psi_{iu,jv}(x_{iu},x_{jv})\prod_{iu}\phi_{iu}(x_{iu}) \quad (4)$$

Where Z is the normalization constant, a.k.a the partition function. $\psi_{iu,jv}(x_{iu},x_{jv})$ is the two-node potential function defined as

$$\begin{aligned}\psi_{iu,jv}(1,1) &= \varepsilon, \quad for \quad i=j \;\; or \;\; u=v;\\ \psi_{iu,jv}(x_{iu},x_{jv}) &= 1, \quad otherwise\end{aligned} \quad (5)$$

where ε is set to 0 or a small positive number. Therefore, if the part correspondences violate the one-to-one constraint, the prior probability would be zero (or near zero). $\phi_{iu}(x_{iu})$ is the one-node potential function. Adjusting $\phi_{iu}(x_{iu})$ affects the distribution $p(X|H=1)$, therefore it is related to the *occurrence probability* r_i. Each potential $\phi_{iu}(.)$ defined at vertex iu in the association graph has two parameters $\phi_{iu}(1)$ and $\phi_{iu}(0)$. Yet, it is not difficult to show that different $\phi_{iu}(1)$ and $\phi_{iu}(0)$ with the same ratio $\phi_{iu}(1)/\phi_{iu}(0)$ would result in the same distribution $p(X|H=1)$ (but different partition function Z). Therefore, we can let $\phi_{iu}(0)=1$ and $\phi_{iu}(0)=z_i$. z_i is independent of u, as we treat every image node that possibly matches the model part i equally.

Second, we need to derive the conditional density $p(O|X,H=1)$. Assuming that y_u and y_{uv} are independent given the correspondence, we have

$$p(O|X,H=1) = \prod_{uv}p(y_{uv}|X,H=1)\prod_{u}p(y_u|X,H=1)$$

Furthermore, y_u and y_{uv} should only depends on the RARG nodes that are matched to image parts u and v. Thus

$$\begin{aligned}p(y_u|x_{11}=0,...,x_{iu}=1,...,H=1) &= f_i(y_u)\\ p(y_{uv}|x_{11}=0,...,x_{iu}=1,x_{jv}=1,...,H=1) &= f_{ij}(y_{uv})\end{aligned}$$

Also, if there is no node in the RARG matched to u and v, this means image parts u and v are extra parts coming from the background. Then y_u,y_{uv} should be sampled from the background *pdf*s, i.e.

$$\begin{aligned}p(y_u|x_{11}=0,x_{iu}=0,...,x_{NM}=0,H=1) &= f^+_{B_1}(y_u)\\ p(y_{uv}|x_{11}=0,x_{iu}=0,...,x_{NM}=0,H=1) &= f^+_{B_2}(y_{uv})\end{aligned}$$

where $f^+_{B_1}(\cdot)$ and $f^+_{B_2}(\cdot)$ is the background *pdf* trained from the positive data set. Note that here we use two sets of background *pdf*s to capture the difference of the background statistics in the positive data set and that in the negative data set. Such distinction is reasonable as a specific class of objects often occur more frequently in certain types of backgrounds, rather than completely random backgrounds. These two distributions can be set to be equal, if we assume the background statistics of the positive and negative sets are the same.

Combining all these elements together, we can prove (proof in [15]) the following relationship between the detection likelihood ratio and the partition function.

Theorem 1 *The likelihood ratio is related to the partition functions of MRFs as the following*

$$\frac{p(O|H=1)}{p(O|H=0)} = \sigma\frac{Z'}{Z} \quad (6)$$

where Z is the partition function of the Gibbs distribution $p(X|H=1)$. Z' is the partition function of the Gibbs distribution of a new MRF, which happens to be the posterior probability of correspondence $p(X|O,H=1)$, with the following form

$$p(X|O,H=1) = \frac{1}{Z'}\prod_{iu,jv}\varsigma_{iu,jv}(x_{iu},x_{jv})\prod_{iu}\eta_{iu}(x_{iu}) \quad (7)$$

where the one-node and two-node potential functions have the following forms

$$\eta_{iu}(1) = z_i \frac{f_i(y_u)}{f^+_{B_1}(y_u)}; \varsigma_{iu,jv}(1,1) = \psi_{iu,jv}(1,1) \frac{f_{ij}(y_{uv})}{f^+_{B_2}(y_{uv})} \quad (8)$$

As shown above, the potential functions depend on the appearance features of the nodes and relational features of the edges. All other values of the potential functions are set to 1 (e.g. $\eta_{iu}(x_{iu} = 0) = 1$). σ is a correction term

$$\sigma = \prod_u f^+_{B_1}(y_u)/f^-_{B_1}(y_u) \prod_{uv} f^+_{B_2}(y_{uv})/f^-_{B_2}(y_{uv})$$

2.3 Computing the Partition Functions

Theorem 1 reduces the likelihood ratio calculation to the computation of the partition functions.

The partition function of the prior correspondence distribution Z can be computed by closed form in polynomial time or using the following upper bound approximation (see proof in [15])

Lemma 1 *The log partition function satisfies the following inequality*

$$\ln Z \leq \sum_{i=1}^{N} \ln(1 + M z_i)$$

and the equality holds when N/M tends to zero (N and M are the numbers of parts in the object model and image respectively). Is the MRF is pruned, the upper bound is changed to

$$\ln Z \leq \sum_{i=1}^{N} \ln(1 + d_i z_i)$$

where d_i is the number of the nodes in the ARG that are allowed to match to the node i in the RARG after pruning the Association Graph.

The challenge is to compute the partition function Z', which is a summation over all correspondences, whose number is exponential in MN.

There are several ways to approximate Z'. The simplest method is using Viterbi approximation, which first finds the most likely correspondence, then calculating the partition function by using the obtained maximum likelihood correspondence and discarding all other correspondence. This approximation may not be accurate as it does not consider alternative part correspondences. A more accurate approach is using variational approximation. Because the log partition function of a MRF is convex, the convex duality property [11] can be used to represent the log partition function as a variational form. Equivalently, Jensen's inequality can be applied to find the lower bound of the partition function for approximation as the following

$$\ln Z' \geq \sum_{(iu,jv)} \hat{q}(x_{iu}, x_{jv}) \ln \varsigma_{iu,jv}(x_{iu}, x_{jv}) + \sum_{(iu)} \hat{q}(x_{iu}) \ln \eta_{iu}(x_{iu}) + \mathcal{H}(\hat{q}(X)) \quad (9)$$

Where $\hat{q}(x_{iu})$ and $\hat{q}(x_{iu}, x_{jv})$ are known as one-node and two-node beliefs, which are the approximated marginal of the Gibbs distribution $p(X|O, H = 1)$. $\mathcal{H}(\hat{q}(X))$ is the approximated entropy, which can be approximated by Bethe approximation[14]. Calculating beliefs, known as probabilistic inference, can be realized by different methods. We have used two methods to perform inference : Gibbs sampling scheme and Loopy Belief Propagation (LBP). It turns out that the LBP algorithm only works when the part relations are not taken into account. And LBP usually does not converge. To solve this problem, we have developed a special LBP convergence scheme. Consequently, LBP based method yields quite good performance, as reported in our previous technical report. Gibbs sampling scheme can be also used in the viterbi search. It turns out that a very small number of samples already yields very good performance. More accurate methods, such as semideiminite relaxation, can also be applied, but their computational cost is too high in our application.

2.4 Learning Random Attributed Relational Graph

In the generative model, Gaussian density functions are used for all random numbers defined at the RARG and the background model. Therefore, the Gaussian parameters for the vertex i and edge ij in the RARG and the background need to be learned, including μ_i,Σ_i,μ_{ij},Σ_{ij};$\mu^+_{B_1}$,$\Sigma^+_{B_1}$,$\mu^+_{B_2}$,$\Sigma^+_{B_2}$;$\mu^-_{B_1}$,$\Sigma^-_{B_1}$,$\mu^-_{B_2}$,$\Sigma^-_{B_2}$. and z_i.

Learning the RARG can be realized by Maximum Likelihood Estimation (MLE). Directly maximizing the positive likelihood with respect to the parameters is intractable, instead we can maximize the lower bound of the positive likelihood through Jensen's inequality lower bound, resulting in a scheme known as variational Expectation-Maximization (Variational E-M).

Variational E-Step: Perform inference or Gibbs sampling search to obtain the one-node and two-node beliefs.

M-Step: Maximize the overall log-likelihood with respect to the parameters, which results in a set of parameter update equations (see [15] for detailed equations).

For the background Gaussian function parameters $\mu^-_{B_1}$,$\Sigma^-_{B_1}$,$\mu^-_{B_2}$,$\Sigma^-_{B_2}$, the maximum likelihood estimation re-

sults in the sample mean and covariance matrix of the part attributes and relations of the images in the negative data set.

Besides the Gaussian function parameters, we also need to learn the parameter z_i, which is related to the *occurrence probability* r_i and required by the computation of Eq.(8). However, it turns out that directly learning z_i is difficult. Instead, we can learn r_i first and then convert r_i back to z_i. Learning r_i requires the following lemma (proof in [15])

Lemma 2 *r_i and z_i is related by the following equation:*

$$r_i = z_i \frac{\partial \ln Z}{\partial z_i}$$

where Z is the partition function defined in Eq. (4).

The *occurrence probability* then can be learned using the following simple equation

$$r_i = \frac{1}{K} \sum_k \sum_u \hat{q}(x_{iu}^k = 1) \tag{10}$$

Where K is the number of the positive training data. Combining lemma 1 and lemma 2, we can convert r_i back to z_i using the following equation $z_i = r_i/((1 - r_i)M)$ (for the complete MRF, likewise for the pruned MRF).

2.5 Spanning Tree Approximation for Spatial Relational Features

Our approaches described so far assume the RARG is fully-connected and the parts and edges are independent. However, this assumption may not be valid in some cases. For example, if the relational feature y_{ij} represents coordinate difference. the relational features among three nodes (y_{12}, y_{23}, and y_{31}) are mutually dependent. For graphs with dependent features, the factorization process mentioned above for deriving the likelihood function needs to be modified. For this, we adopt a pruned tree representation in the E step of the inferencing process, while still keeping the full graph in the M step. We prune the fully-connected RARG to a tree by a spanning tree approximation algorithm, which discards the edges whose variations of the coordinate differences are high. This results in a modified variational E-M scheme, which ensures the independence assumption is correct. Note the above method is not equivalent to a tree-based model, since a fully-connected graph is still used in the M step and the discarded edges are dynamically selected in each iteration.

3 Discriminative Learning and Classification

Under the generative learning frame work, the positive and negative hypotheses are learned separately by maximum likelihood estimation. If the probabilistic density functions of the positive and negative hypotheses are precise, then the classification based on likelihood ratio is optimum according to the Bayesian classification theory. However, for multi-view object detection, modeling the distributions of part appearances as Gaussian functions are inaccurate due to large variations. Therefore, the performance of the generative classification often degrades if data distribution is complex. In contrast, discriminative methods directly maximize the classification errors, thus often yielding superior performance against generative methods.

Motivated by the above, we further explore a key insight about the inferencing processing of RARG in order to enhance its robustness over multi-view objects. The insight reveals that the variational approximation (in Eq.(9) or the Viterbi approximation of the detection likelihood ratio can actually be considered as a linear aggregation of the individual classifiers defined on nodes and edges in the RARG. The classifier at node n_i in the generative model is the log likelihood ratio function $\ln \eta_{iu}(x_{iu})$ and the classifier at edge e_{ij} is the log likelihood ratio function $\ln \varsigma_{iu,jv}$. In the generative classifiers, these likelihood ratio functions are specified by the Gaussian distributions. In order to increase the discriminative power, we can replace the log likelihood ratio functions with discriminative classifiers, such as SVMs, resulting in a generative and discriminative hybrid method. The new classifier can be written as

$$C(O) = \sum_{(iu,jv)} \hat{q}(x_{iu}, x_{jv}) C_{ij}(y_{uv}) + \sum_{(iu)} \hat{q}(x_{iu}) C_i(y_u)$$

where C_i is the discriminative classifier at node n_i, and C_{ij} is the discriminative classifier at edge e_{ij}. $\hat{q}(x_{iu}$ and $\hat{q}(x_{iu})$ are beliefs computed from the MRF. And C_i and C_{ij} are related to the earlier potential functions by $\eta_{iu}(1) = \exp(C_i(y_u))$ and $\varsigma_{iu,jv}(1,1) = \exp(C_{ij}(y_{uv}))$.

The learning process now consists of two passes: generative learning initialization and discriminative learning. Generative initialization, similar to the learning process described in Sec 2.4, is intended to roughly discover the structure of the object model, and learn an approximate part appearance and relation distribution. The learned generative models are used to find the probabilities of the initial part correspondences, with which the discriminative learning step is conducted by using a new E-M procedure:

Variational E-Step: Perform inference or Gibbs sampling search (Viterbi approximation) to obtain $\hat{q}(x_{iu})$ and $\hat{q}(x_{iu})$.

M-Step: Perform discriminative learning of each individual classifiers to minimize the classification error.

In the M-step, for vertex i of the RARG model, the parts in the positive images that are matched to i (in the case of viterbi approximation) and all parts from the negative

Figure 3. Examples of object images in Graz02 data set

Figure 4. Part detection examples using salient region detector

images are used for training. If variational approximation is used, then positive parts are sampled according to the matching probabilities computed from inference. Sometimes, the negative data set may be large, resulting in high computational cost. In that case, a sampling process (currently uniform) is used to obtain a smaller subset of the negative samples.

4 Experiments

In the experiments, we compare the performance of our method with the generative methods and previous methods for multi-view object detection. We use the Graz data Set [1] which contains images of objects under different views with large variations in scale, orientation, color and shape. Graz data set consists of two data sets: Graz01 and Graz02. Graz01 data set contains two object classes, "bike" and "person", and one background class. Graz02 data set contains three object classes, "car", "bike" and "person", and one background class. Objects in the Graz02 data set have more variations in scale and spatial position. Graz data sets have been used for evaluating object detection in [1] and [7]. Figure 3 shows some sample images from the Graz02 data set.

We use Kadir's Salient Region Detector [9] with the same parameters across all classes to extract object parts in the images . Such detector has been successfully used in other object detection work, such as [4]. The maximum

Grz01	Gen	Hyb	std
Car	76.5	77.3	1.4
Person	68.7	71.8	0.8

Grz02	Gen	Hyb	std
Car	71.0	72.8	1.1
Bike	74.7	76.7	1.3
Person	74.8	79.8	1.4

Table 1. Performance comparison between the generative and hybrid approaches (std: standard deviation of the performance of the hybrid approach)

number of parts in each image is restricted to 100 in order to control the computational cost. Each data set for a class is randomly partitioned with equal size into training and testing sets. Each test set is further randomly partitioned into two equal subsets for conducting two-fold cross validation. Namely, one sub set is used for tuning parameters, including the size of the RARG and the SVM RBF kernel width, while the other is used for the final testing. The averaged Equal Error Rate (EER) [4][1] then is computed. To test the sensitivity of the performance to the training data selection, we repeat the above process four times, and compute the mean value and standard deviation of EER. For SVM, we use an implementation known as OSU-SVM [8], by which we need to tune the parameter λ , which is related to the kernel width of the RBF kernel. The generative learning process normally converges in 15 to 20 iterations, which is significantly smaller than that of the method used in [4]. The E-M iterations in the discriminative learning normally converges in 8 to 12 steps.

Features extracted from image parts include regular color moments (ten components), size (output from region detector), and spatial coordinates. Relational features include spatial coordinate differences. In our experiments, we have found that the color moments slightly outperform PCA coefficients features used in [4].

Choosing the size of the RARG (number of nodes) remains a challenging problem. Currently, we use an exhaustive search method to find the optimal choice among some discrete choices (5, 10, 15) by comparing the performance over the validation data set. It turns out that in most cases, RARGs with 15 nodes yield the best results, while a size larger than 15 nodes actually does not yield better performance.

We compare the performance difference between the hybrid method and the pure generative method (Table 1). The standard deviation of the performance is also shown to assess the significance of the performance gain. In the case of the generative method, the likelihood ratio in Eq.(1) is used for classification. In some separate experiments reported in [15], we have found our RARG-based generative model achieves about the same accuracy as that using the well-known constellation model. As shown in Table 1, the performance of the new discriminative-generative hybrid ap-

Graz01	boosting	ours	Graz02	boosting	ours
Bike	76.5	77.3	Car	70.2	72.8
Person	68.7	71.8	Bike	76.5	76.7
			Person	77.2	79.8

Table 2. Performance comparison of boosting[1] and our method.

proach is consistently better than the generative method. The improvement is particularly significant for the "person" class.

Finally, we compare the performance of our hybrid method with that by the boosting-based method reported in [1]. In [1], the experiments are conducted under different settings using different features. We compare our method with one of their settings with the most similar features. For Graz01, we compare with their system using moment invariant features (Basic moments were not used in their experiments. And for Graz02, we compare with their system using basic moments. Other more sophisticated features such as SIFT features are also used in [1]. However, their advantages in detection performance are not consistent across different data sets.

The above table shows that our approach outperforms the prior work consistently across all classes. The improvement is significant for classes like "car", but minor for classes like "bike". This may be because it is difficult to capture the structure of the "bike" object, as many bike images in the training set include just bike wheels instead of full bike.

Previous work using discriminative and generative learning with Fisher kernel [7] has reported their results in Graz01 data set. Their performance in Graz02, which is more difficult, is nevertheless unavailable. Directly comparing our experiments with [7] may be incomplete. Here, we list the performance of our system and theirs (single model with 3-part) in Graz01 as a reference

Graz01	boosting [1]	Fisher [7]	ours
Bike	76.5	76.4	77.3
Person	68.7	74.9	71.8

Table 3. Performance comparison using Graz01 data set

5 Conclusion

We have presented a new generative-discriminative model for multi-view object detection. We develop a rigorous generative component based on Random Attributed Relational Graph and derive the fundamental relations between MRF partition functions and likelihood ratio functions for learning and detection. The model is effective and intuitive - it automatically learns the structure and appearance variations of an object class in an unsupervised manner. We then incorporate a discriminative learning scheme into the generative framework. Our experiments have confirmed the power of the generative model and the robustness of the hybrid approach in object detection, outperforming the pure generative approaches as well as previous work.

References

[1] A. P. A. Opelt, M. Fussenegger and P. Auer. Generic object recognition with boosting. In *Technical Report TR-EMT-2004-01, EMT, TU Graz, Austria,*, 2004.

[2] D. Crandall and P. Felzenszwalb. Spatial priors for part-based recognition using statistical models. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, 2005.

[3] P. Erdös and J. Spencer. *Probabilistic Methods in Combinatorics*. New York: Academic Press, 1974.

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, pages 66–73. IEEE, 2003.

[5] H.G.Barrow and R.J.Popplestone. Relational descriptions in picture processing. In *Machine Intelligence*, pages 6:377–396, 1971.

[6] A. B. Hillel, T. Hertz, and D. Weinshall. Efficient learning of relational object class models. In *IEEE International Conference on Computer Vision (ICCV)*, 2005.

[7] A. Holub and P. Perona. A discriminative framework for modeling object class. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, 2005.

[8] http://svm.sourceforge.net/docs/3.00/api/.

[9] T. Kadir and M. Brady. Scale, saliency and image description. *In International Journal of Computer Vision*, pages 45(2):83–105, 2001.

[10] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision - to appear*, 2002.

[11] M. J. Wainwright and M. I. Jordan. Semidefinite methods for approximate inference on graphs with cycles. In *UC Berkeley CS Division technical report UCB/CSD-03-1226*, 2003.

[12] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, pages 101–109. IEEE, 2000.

[13] M. Welling and Y. W. Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Uncertainty of Artificial Intelligence*, 2001, Seattle, Washington.

[14] J. S. Yedidia and W. T. Freeman. Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium*, pages pp. 239–236,Chap. 8,, Jan. 2003.

[15] D.-Q. Zhang. *Statistical Part-Based Models: Theory and Applications in Image Similarity, Object Detection and Region Labeling*. PhD Thesis,Graduate School of Arts and Sciences, Columbia University, 2005.