A Multi-stage Segmentation based on Inner-class Relation with Discriminative Learning

Haoqi Fan[1], Yuanshi Zhang[2] and Guoyu Zuo[3]
[1]Department of Computer Science, Beijing University of Technology, Beijing, China
[2]Department of Statistics, Columbia University, New York, U.S.A.
[3]School of Electronics Information and Control Engineering, Beijing University of Technology, Beijing, China

Keywords: Segmentation, Discriminative Model.

Abstract: In this paper, we proposed a segmentation approach that not only segment an interest object but also label different semantic parts of the object, where a discriminative model is presented to describe an object in real world images as multiply, disparate and correlative parts. We propose a multi-stage segmentation approach to make inference on the segments of an object. Then we train it under the latent structural SVM learning framework. Then, we showed that our method boost an average increase of about 5% on ETHZ Shape Classes Dataset and 4% on INRIA horses dataset. Finally, extensive experiments of intricate occlusion on INRIA horses dataset show that the approach have a state of the art performance in the condition of occlusion and deformation.

1 INTRODUCTION

Image segmentation is a fundamental and long-standing problem in computer vision, which aims to cluster pixels in an image into distinct, semantically coherent and salient regions. Solutions to image segmentation serve as the basis for a broad range of applications, which include content-based image retrieval, object detection, video surveillance and object tracking. Unfortunately, the work to segment an image is found difficult and challenging for two main reasons (Maji et al., 2009). One is the fundamental complexity of modeling a vast amount of visual pattern, and the other is the intrinsic ambiguities in image perception. In recent researches, there are mainly three ways to solve the problem pertaining to segmentation and recognition:

First method is commonly known as the Top-down segmentation, which is a method using prior knowledge of an object, such as its characteristic to propose plausible pixels that may compose a certain object. The principal difficulty in top-down segmentation stems from the large variability in the appearances and shapes of objects within a given class. Unfortunately, recent extensive experiments in (Li et al., 2012) demonstrated that a single region generated by image segmentation can rarely be equated with a physical object. Since there is a variety of object shapes, only an approximate edge alignment between the training masks and new object instances can be predicted. The approach adopted in this paper is more general and does not rely on its coherent appearance of the entire object but rather on the semantic parts. It can also comprehend the novel inner-class relationship between parts of the object, which makes the representation of a movable joints object in relative positions more naturally and powerful.

The second method, Bottom-up approach, is to over-segment the image into regions or pixels and then identify them as corresponding labeled object. This approach mainly relies on continuity principles and joint local features such as SIFT, color and texture. The Bag of Regions model (Hu et al., 2011), representing object shape at multiple scales and encoding shapes even in the presence of adjacent clusters, has recently delivered impressive performances. In the subsequent work, a semantic context was joined in recent works (Chen et al., 2011) and has a good performance. In particular, the bottom-up approach is incapable of capturing the segmentation of an object from its background, thus an object might be segmented into multiple regions. The approach in this paper differs from this line of work, because the model not only takes advantage of local features, but also avails itself of the shape and

Fan H., Zhang Y. and Zuo G..
A Multi-stage Segmentation based on Inner-class Relation with Discriminative Learning .
DOI: 10.5220/0004717404860493
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 486-493
ISBN: 978-989-758-004-8

clique features.

In order to address these shortcomings appeared in top-down and bottom-up approach, recent works have started to study the integration of higher-order potentials into random field models. For reasons of computational tractability, successful approaches rely on relatively small clique sizes (Sun et al., 2011), despite their strongly increasing expressive power. These models still mainly focus on encode local image properties. Unfortunately, because of the approach's lack of consideration for the global features such as shape of the integral object, overcoming these limitations to build effective structural object descriptions has proven to be quite challenging. Our approach is more along random field model as we focus on both the representation and classification of individual regions and modeling the relations between regions in the intact object. Our discriminative model can be interpreted as powerful pluralistic potentials with graph-based models. It uses features and relationship between parts to represent objects in images. In particular, a novel inner-class relationship is proposed to describe the relationship between different parts in an object.

We describe an object in real world images as an assembly of flexible skeleton and parts based on the skeleton, i.e., an object that is composed of multiply, disparate and correlative parts. A discriminative model is proposed and inferenced by a multi-stage segmentation algorithm. It is performed as a multi-stage segmentation via the maximization of a discriminative function, which is similar to the two-stage-segmentation showed in (Gould et al., 2008). The function formulates local features, including the local shape and visual appearance, and sets the possible skeletal shape of the object, which is performed by the inner-class pairwise features. Our method can not only segment an interest object but also label different semantic parts of the object. We validate our approach by conducting extensive experiments on ETHZ Shape Classes and INRIA horses dataset, and we also test our method in intricate occlusion on INRIA horses dataset and in real world. The results show that our approach has a satisfying performance.

Our main contributions in the paper are as follows:

1. We propose a discriminative model that can not only segment an interest object but also label different semantic parts of the object, especially the various parts of articulated body animals' bodies. As we know, this is a frontier in recognition and has been left largely unexplored as (Arbelaez et al., 2012) touch upon it.
2. We propose inner-class and clique features to describe the relationship between different body parts and a discriminative model to semantically label different parts of an object. Specifically, we solve the problem of object proposal, a NP-hard problem, via our multi-stage segmentation algorithm, and we train the model via the latent structural SVM learning framework.
3. We validate our approach by conducting extensive experiments on two acknowledged datasets and showed that our method boost an average increase of about 5% on ETHZ Shape Classes Dataset and 4% on INRIA horses dataset. Above all, we conduct extensive experiments of intricate occlusion on INRIA horses dataset shows that the approach have a state of the art performance in the condition of occlusion and deformation.

The rest of this paper is organized as follows; our proposed discriminative model is described in Section 2. We discuss the features we used and how to parameterize them in Section 3. We detailed the training and inference process in Section 4 and Section 5, respectively. In Section 6, we show complex and comprehensive experimental results on real world and two Datasets. Then we conclude the result in Section 7.

2 APPROACH

We design a discriminative model that can not only segment an interest object but also label different semantic parts of the object. The model takes visual appearance, inner-class relationship and the shape of the object into account, which make our model ability to interprete as powerful pluralistic potentials. Furthermore, we take latent variables into account for different inner-class parts, which describe an object by a set of parts of the object (e.g. a giraffe is represents by the set {head, neck, body, forelegs, hindlegs}). We use a multi-stage segmentation algorithm to inference labels in our approach, and it is fast and effective because it avoids combinatorial computation in optimization. In addition, the model is trained within the latent structural SVM learning framework.

Given an image, our approach starts with an initial over-segmentation of the image by partitioning it into multiple homogeneous regions. To make certain that pixels in a region belong to the same label and abstain from obtaining regions which are larger than the object we intended, we over-segment the image using NCuts (Cour et al., 2005).

Fig.1. illustrates the graphic model in our approach; a graph model is used to describe an image, where each over-segmented region corresponds to a node. Let the set of nodes that we want to label be denoted by x. Then, we associate a hidden part label h_i with each node x_i, and represent these hidden part labels as a connected graphical model. Every h_i represents one part of the object. Each hidden node can take a label from a label set L. For example, nodes in Fig.1. have their class labels {Giraffe} and their hidden label must be $L^{\text{Giraffe}} = \{\text{head, neck, body, forelegs, hind legs}\}$.

Figure 1: Graphical representation of the graphic model based on the discriminative model. Spots denote the regions we want to segment, the hidden nodes which speculate the regions' part category, denoted by h_i (Here the redundant undirected edges are not show but all the nodes h_i adjacent to each other with the same y are connected). And each hidden node has its own object label y_i. The most left part showed the object proposal of the neck part of the giraffe.

In particular, a node x_i which belongs to class y_i has its own hidden part label h_i, and h_i is an element of L^{y_i}. For simplicity, we fix the relationship between the object labels and hidden part labels as our priority. In particularly, we make a clear demand in the number of parts for each compound class, and do not share parts between classes. In other words, we restrict a regions' part category h to a class label y only from a subset of values so that h uniquely determines y. For example, a part of {Giraffe} must choose a hidden label from $\{\text{head, neck, body, forelegs, hind legs}\}$, and it is impossible that a region with a hidden label of mug-body belongs to {Giraffe}. We denote this deterministic mapping from parts to objects by $y(h_i)$. τ denotes the object proposal, which means it defines merger of the over-segmented regions in an image, because the images merge sets of regions which belong to the same semantic hidden part together. For example, the object proposal τ in the neck part is $\{x1, x2, x3\}$, which means our approach would merge $\{x1, x2, x3\}$ together to be a proposed region r, and the region r is labeled with $\{h: \text{neck}; y: \text{Giraffe}\}$.

Our goal is to predict both the semantic labels of objects and the labels of each parts (hidden labels). We define a discriminant function $f_w(x, y, h, \tau)$ which is parameterized by a set of weight w. The optimal object proposal and corresponding label and hidden label are given by

$$(y^*, h^*, \tau^*) = \text{argmax}_{y,h,\tau} f_w(x, y, h, \tau)$$

Where $f_w(x, y, h, \tau)$ includes terms for unary and pairwise potentials. More specifically, $f_w(x, y, h, \tau)$ is defined as

$$f_w(x, y, h, \tau) = \sum_{i \in \tau} w_\alpha^{y,h^T} \alpha(r_i) + \sum_{i,j \in \tau} w_\beta^T \beta(h_i, h_j, y_i, y_j, r_i, r_j) + \sum_{i,j,\ldots,l \in \tau} w_\rho^T \rho(h_i, h_j, \ldots, h_l, y_i, y_j, \ldots, y_l) \quad (1)$$

Where $\alpha(x_i)$ represents unary features of x_i and $\beta(h_i, h_j, y_i, y_j, r_i, r_j)$ represents pairwise features of inner-class between two regions x_i, x_j with their labels y_i, y_j and their hidden labels h_i, h_j. $\rho(h_i, h_j, \ldots, h_l, y_i, y_j, \ldots, y_l)$ is the clique feature to ensure that there is no two regions correspond to the same hidden label, which we will explain at the end of this section. The vectors $w = (w_\alpha, w_\beta, w_\rho)$ are the corresponding weight. Also, r_i represent the object regions given by the object proposal τ. $\alpha(r_i)$ is a unary observation feature function and $\beta(h_i, h_j, y_i, y_j, r_i, r_j)$ is a pairwise feature function. In particular, $\alpha(r_i)$ is used to describe the features that belong to single object, including color, texture, shape, size, etc. And $\beta(h_i, h_j, y_i, y_j, r_i, r_j)$ is used to depict the pairwise spatial relationship and the visual appearance differences.

The weights w are the parameters to be learned in our model. The unary weights $w_\alpha^{y,h}$ consist of a few components and each component corresponds to a specific subclass of a certain object category. Determining the number of subclasses for each object is described in section 4.

We will explain about the clique feature function $\rho(h_i, h_j, \ldots, h_l, y_i, y_j, \ldots, y_l)$ in the following way: In

the model, different hidden nodes of pixels or super-pixels can belong to the same label. However, there is much more restriction in our model. As a cup cannot have more than one cup handles and a human being just have one head, one certain class of object in the picture has no reason to contain two regions with the same hidden part label. Even though there are some regions with the same hidden part label which are adjacent to each other, they must be merged together to be an integral region, and that is what τ is defined (To make sure that any label corresponds to only one region instead of multiple regions, and regions with same hidden label should be merged together). Because compared to the separate regions, an integral region has a very different shape features. The shape features play a very important role in our model. We will discuss it in Section 3.

To calculate $(y^*, h^*, \tau^*) = argmax_{y,h,\tau} f_w(x, y, h, \tau)$ is a NP-hard problem with three variables τ^*, h_i^* and y_i^*, so we calculate the (y_i^*, h_i^*) and τ^* individually and repetitively, and we will explain how to do it in Section 4.

3 FEATURES

3.1 Unary Features

Interactions between the image content and the variables of interest are described by the unary observation factors performed by g_i. For each over-segmented region, we extract multiple types of region-level features representing appearance statistics based on shape, color and texture.

The shape features include two parts. One is the size of the region, and the other is the shape feature based on tensor scale which is a morphometric histogram presented by (Andalóet et al., 2010). The shape feature unifies the representation of local structure thickness, orientation, and anisotropy. It is archived by using Image Foresting Transform (IFT), a salience detector and a shape descriptor, both based on tensor scale. The color features include the mean HSV value, its standard deviation, and a color histogram. The texture features are average responses of filter banks in each region. To do this, we utilize textons (Liu et al., 2010) which have been proven effective in categorizing materials. A dictionary of textons is learned by convolving a 17-dimensionalflter bank (Winn et al., 2005) with all the training images and running K-means clustering (using Mahalanobis distance) on the filter responses.

3.2 Pairwise Features

In the past, a significant amount of work (Gould et al., 2008) has been done on proving that the semantic context is very useful for image categorization. In our work, we use two parts of pairwise features.

One part is the similarity of two regions, i.e. the visual appearance differences between two parts. We measure the color and texture difference using χ^2 distance of color and texture histogram.

The other part is to measure the context between two regions, and we proposed an inner-class pairwise feature that models the interaction of regions' hidden labels (h_i, h_j) that belong to the same class. We define the pairwise features f_{ij} between two regions r_i and r_j ,and their corresponding y_i and y_j values in Eq. (2) as equal.

$$Distance = \begin{cases} 0, h_i = h_j,\ r_i\ is\ adjacent\ to\ r_j \\ f_{i,j}^1(d), h_i \neq h_j \end{cases} \quad (2)$$

$$Angle = \begin{cases} 0, h_i = h_j,\ r_i\ is\ adjacent\ to\ r_j \\ f_{i,j}^2(\theta), h_i \neq h_j \end{cases}$$

d is the normalized distance between two regions' centroid and θ ($\theta \in [0,\pi)$) is the angle between them. The normalized distance d is calculated by the formula $\frac{d_{original}}{\sqrt{area\ of\ regions\ i+j}}$. $f_{i,j}^1$ and $f_{i,j}^2$ are the probability function which obtains by look-up relative distance and angle histograms. Indirect adjacency means r_i and r_j are in same clique that all regions in the clique are interconnected and have the same label y. We construct relative distance and angle probability histograms that encode offset preference between regions with different hidden labels. We first over-segment images and merged segments manually on the training dataset (only merge segments for training processes, not for the inference step) to make sure the merged fractions belong to the same part semantically and combine them into an integrated region. Then we use the EM algorithm to label different regions' hidden parts h based on unary features $\alpha(r_i)$, such as shapes, colors, without the pairwise and clique features. The next step involves calculating the distance and angle relationship between regions with different hidden parts label h.

The process of constructing the relative distance and angle probability histograms is simple. We calculate all pairs of regions' centers' probability of

distances and angles in all images, which belong to different labels h, and draw them as histograms. The relative location and angle probability histograms are blurred by a Gaussian filter with variance equal to 10% of the histograms, which reduces the bias of center location shift in the training data.

At testing time, we first predict the class y_i and the h_i for each region independently using the unary features based on the simplified discriminative function (only with unary features) in Eq. (3), and then use the multi-stage segmentation algorithm to calculate the approximate solution of $\text{argmax}_{y_i,\tau} P(y_i, \tau | x, \theta)$ and $\text{argmax}_{h_{i,\tau}} P(h_i, \tau | x, \theta)$. Figure 2 is an example of probability of distances and angles.

$$y, h = \text{argmax}_{y,h} w_\alpha^{y,h^T} \alpha(r_i) \quad (3)$$

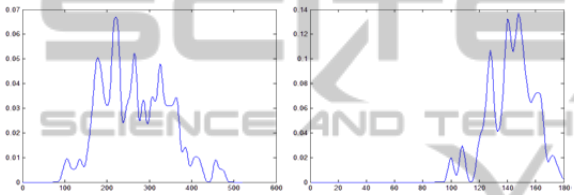

Figure 2: The relative distance (left) and angle (right) probability histograms $f_{i,j}^1$ and $f_{i,j}^2$, i is giraffe's head and j is the giraffe's neck.

3.3 Clique Features

As we mentioned above, there is more restriction in our model. On one hand, because the shape features play a very important role in our model, and compared to the separate regions, an integral region has a very different shape features. So we wish all adjacent regions with the same hidden part label merging together to be an integral one. On the other hand, since the regions with same hidden label are merged, one certain class of object in the picture has no reason to contain two regions with the same hidden part label (e.g. a giraffe never has two heads). If there are two regions that share the same hidden part label in one certain object, the segmentation and labeling must be wrong from semantic point of view. In order to avoid such situations, we set the clique features as a penalty function. The features' function are $\rho(h_i, h_j, \dots, h_l, y_i, y_j, \dots, y_l) = w_\rho^1 \rho^1 + w_\rho^2 \rho^2$, and $w_\rho = [w_\rho^1, w_\rho^2]$, and the value ρ^1 and ρ^2 are shown in Eq. (4).

$$\rho^1 = \begin{cases} -\text{Max}, h_i = h_j, \ r_i \text{ is adjacent to } r_j \\ 0, others \end{cases}$$
$$\rho^2 \quad (4)$$
$$= \begin{cases} -\text{Max}, h_i = h_j, r_i \text{ is adjacent to } r_j, r_i \ ar \\ 0, others \end{cases}$$

As one certain class of object in the picture has no reason to contain two regions with the same hidden part label. If the adjacent regions share the same label, it would be merged together by the multi-stage segmentation algorithm. Therefore, nonadjacent regions belong to the same hidden part label will not be tolerated. In order to prevent this situation, we set them to –Max.

4 TRAINING

We want the area of the hidden nodes representing an intuitive sense, which means that each hidden layer represents an integrated semantic part of an object. For example, a mug is composed of one mug body part and at most one mug handle part, rather than a large number of insignificant fractions. Thus, we over-segment images and merge segments on the training dataset manually, which merged fractions semantically belonging to the same part together and combined into an integrated region. In other words, we define the τ^* manually in the training part. Since we don't have the hidden labels h, we can't calculate the pairwise and clique features, thus we should calculate the h first. To do that, we extract unary features $\alpha(r_i)$ from the integrated regions, then we use the EM algorithm to get the different regions' hidden parts h based on unary features $\alpha(r_i)$, such as shapes, colors, without the pairwise and clique features. When we get the regions' hidden part labels h, we extract not only the unary features, but also the pairwise and clique features. Finally we train the feature vectors with both their labels y and hidden labels h.

We now present the process to learn the weights in the model with the N training images $\{(y_1, x_1, \tau_1), \dots, (y_N, x_N, \tau_N)\}$. We merge regions manually and get the hidden labels of each regions following the above processes. In order to explain the notation in a simple way, we concentrate the features in Eq. (1) to $\Phi(x, y) = (\alpha, \beta, \rho)$, and denote $f_w(x, y, h)$ as $= w^T \Phi(x, y)$. Then via the structural SVM learning framework (Yu et al., 2009), we train the weights. The formulate the large margin could be trained as the problem following,

$$\min_{w,\xi_n \geq 0} \frac{1}{2} w^T w + C \sum_{n=1}^{N} \xi_n$$

$$\text{s.t.} \forall n, \quad y \neq y_n, \max_h w^{y_n,h^T} \phi(x_n, y_n) - \max_h w^{y_n,h^T} \phi(x_n, y) \geq \Delta(y, y_n) - \xi_n \tag{5}$$

Note there is a constraints in Eq. (5), it requires that the any wrong labeling by at least a loss $\Delta(y, y_n) = \lambda \sum_i I(y^i, y_n^i)$ is lesser than discrimination score of ground truth labeling. Also I is the indicator function (means $I(a, a) = 0$ and $I(a, b) = 1$, where $a \neq b$) and y^i is label of region i. How much these constraints are violated is measured by the slack variables $\{\xi_n\}$. The numbers of constraints in Eq. (5) can be solved efficiently via the cutting-plane algorithm (Joachims et al., 2009). It works by finding the most violated constraints, then using them as active ones. the most violated constraints for the nth image amounts to computing could be found by Eq.(6).

$$(y_n^*, h_n^*) = \text{argmax}_{y,h} w^{y,h^T} \phi(x_n, y) + \Delta(y, y_n) \tag{6}$$

Once the most violated constraints are found, they become the only that remain active. Then the Eq. (6) could be rewrote as an unconstrained problem in Eq. (7).

$$\min_w \frac{1}{2} w^T w + C \cdot \sum_n \left(w^{y_n^*, h_n^{*T}} \phi(x_n, y_n^*) - \max_h w^{y_n,h^T}(x_n, y_n) \right) \tag{7}$$

Here the summation is over RGB images for which $w^{y_n^*,h_n^{*T}} \phi(x_n, y_n^*) - \max_h w^{y_n,h^T}(x_n, y_n) > \Delta(y_n^*, y_n)$, and the slack variables for other images could be directly set to zero. Eq. (7), a difference of two convex functions, can be solved via the Concave-Convex Procedure (CCCP).

5 INFERENCE

The problem of finding the most violated constraints for the image amounts to computing $(y^*, h^*, \tau^*) = \text{argmax}_{y,h,\tau} f_w(x, y, h, \tau)$, but this is a NP-hard problem obviously. The multi-stage segmentation algorithm is used to solve the problem approximately by inferring the y_i^* h_i^* and τ_i^* in different steps: Predict the y_i^* h_i^* in step 1, and then predict the τ^* in step 2, and then repeat this processes again until the algorithm converges.

In step 1, we are typically interested in predicting labels for new data x. We predict them by averaging out the hidden variables and all label variables but one, to calculate the maximum marginal

$$y_i^* = \text{argmax}_{y_i} f_w(x, y, h, \tau)$$

Alternatively, we can calculate the most likely joint configuration of labels by taking the maximal simultaneously over all y. Although such configurations are global consistent, the per fragment error tends to be slightly worse. To see what parts the algorithm has learned, we can look at the most likely parts:

$$h_i^* = \text{argmax}_{h_i} f_w(x, y_i^*, h, \tau)$$

In step 2, we calculate the approximate $\tau^* = \text{argmax}_\tau f_w(x, y_i^*, h_i^*, \tau)$, in order to make the shape and pairwise feature discrimination effectively. Finally each region that belongs to one certain semantic part should not be over-segmented. To ensure that there is not more than one region belongs to the same hidden label, all regions with the same hidden label cannot be simply merged. For example, two are considered as parts of the giraffe's neck merged together and come out to be a body part. We designed a multi-stage segmentation algorithm on the thinking of greedy algorithm, which searches and merges regions most likely belonging to the same part step by step. For the sake of – Max in pairwise features, we are not worry about the algorithm which will not converge.

The multi-stage segmentation algorithm.

```
Calculate all y,h = argmax_{y,h} w_α^{y,h^T} α(r_i)
with only unary features
Repeat
Step1.Calculate
(y*, h*) = argmax_{y,h} f_w(x, y, h, τ) with all
features
Step2.
For r_i in all regions
    For r_j border to r_i
        If h_i = h_j
Merge r_j and r_i together, Calculate
        (y*, h*) = argmax_{y,h} f_w(x, y, h, τ)
        Else
Set hidden label of r_j to h_i,
Calculate (y*, h*) = argmax_{y,h} f_w(x, y, h, τ)
        End
    End
 End
 i, j = arg max_{i,j} E_ij
 if h_i = h_j
    Merge r_j and r_i together
 Else
    Set hidden label of r_j to h_i
 End
Until labels are unchanged.
```

6 EXPERIMENTS

The ETHZ Shape Dataset (Ferrari et al., 2010) contains 255 images of 5 different object classes—Applelogos (40 images), Bottles (48 images), Mugs (48 images), Giraffes (87 images) and Swans (32 images). The dataset is designed in a way that the selected object classes do not have a distinctive appearance and the only representation, which can be used to detect object class instances, is their shape.

We over-segment the image using NCuts (Cour et al., 2005) with n = 50 segments, and calculate features vectors following Section 3 from each segments. Actually, we are not interested in what hidden label the background belongs to, thus we don't merge but simple recognize regions that are labeled {background} . It makes our approach running faster. In most cases, algorithm terminates before 30 iterations.

As a result, we report our result on the dataset for evaluation, and we follow the test settings of Ferrari et al. (Ferrari et al., 2010). In Table 1 we show our model's recall and precision of the segmented boundaries based on correct segmentation.

Table 1: our model's recall and precision of the segmented boundaries, and Compared our discriminative model with traditional HCRF model and work of Ferrari et al..

Precision (%)	our model learn by HCRF	Our approach	Ferrari et al.
Applelogos	90.2/94.2	93.1/93.5	91.6/93.9
Bottles	86.6/83.9	90.3/84.8	83.6/84.5
Giraffes	75.7/77.3	79.5/77.7	68.5/77.3
Mugs	78.9/77.3	83.6/77.2	84.4/77.6

We achieve higher recall at higher precision compared to previous work (Ferrari et al., 2010). We use the first half of images in each class for training, and test on the second half of this class as positive images plus all images in other classes as negative images. In our approach we only use the ground truth outlines of objects present in the first half of images for each class. These statistics show results in precise boundaries of segmentations. The improvement in giraffe demonstrates the efficiency of our discriminative model, especially for the movable joints object like giraffes and swans. The slightly lower percentages of Mugs are due to the mug handles which cannot be fully captured in the over-segmentation step.

To show that our discriminative model has a good performance in the experiences, we use the traditional HCRF model training on the dataset,

Figure 3: Segmentation on ETHZ Shape Dataset. For each example, we show on the left side of the input image and in the middle and the right side of the segmentation for the best matching model. In particular, we mask regions with different hidden part label of different colors. On the right of the selected object mask and the best matching model is displayed.

and compare with our model in Table 1. We can now safely draw the conclusion that our model which trained on the latent structural SVM framework has a better performance than traditional HCRF model on the ETHZ Shape Dateset. In particular, our approach improves the precision of recognition about 5 percent to Ferra et al. We also evaluate our approach on the second dataset, INRIA horses, which has 340 images and half of them contain horses. We archive a high detection rate of 89.7% at 1.0 fppi, which is higher than (Maji et al., 2009) about 4.3%.

In order to show the ability of the model to recognizing the occluded object, we add occlusion picture to the INRIA horses dataset, since the occlusion condition in the real word is complex, we use rainbow ribbon with different directions to block the part of the target object in the image. In the experiment process, we train the model with the normal dataset, and segment the blocked dataset. We could find that even a major part of the target object is blocked by completely unrelated object, our model could label almost all valid parts. The performance of our model in the occlusion condition is superior to other state of the art works (Nearly all of other works couldn't identify these horses). Examples are showed in Fig. 4.

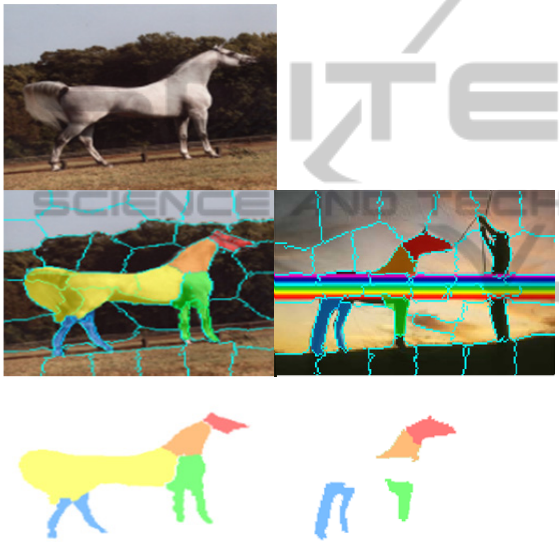

Figure 4: (left) Examples of detections of horses in different poses. (right) Segmentation on INRIA horses Dataset blocked with rainbow ribbon.

7 CONCLUSIONS

In this paper, we have presented a discriminative model, and have achieved more accurate segmentation results. Our method is able to comprehend the relationship between parts of an object, and make the representation of an articulated object in different poses more efficiently and naturally. However, our method computes features on regions instead of single pixels and thus it has become a weakness of our model. Therefore, we will focus on the edge and superpixel-level feature in the future.

REFERENCES

Maji, Subhransu, and Jitendra Malik, 2009. Object detection using a max-margin hough transform. In *Computer Vision and Pattern Recognition. CVPR 2009. IEEE Conference on, pp. 1038-1045.* IEEE.

Li, Zhenguo, Xiao-Ming Wu, and Shih-Fu Chang, 2012. Segmentation using superpixels: a bipartite graph partitioning approach. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 789-796.* IEEE.

Hu, Rui, Tinghuai Wang, and John Collomosse, 2011. A bag-of-regions approach to sketch-based image retrieval. In *Image Processing (ICIP), 2011 18th IEEE International Conference on.* IEEE.

Gould, Stephen, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller, 2008. Multi-class segmentation with relative location prior. In *International Journal of Computer Vision 80, no. 3: 300-316.* Springer.

Sun, Jian, and Marshall F. Tappen. Learning non-local range markov random field for image restoration, 2011. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conf. on, pp. 2745-2752.* IEEE.

Yu, Chun-Nam John, and Thorsten Joachims, 2009. Learning structural SVMs with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1169-1176.*

Joachims, Thorsten, Thomas Finley, and Chun-Nam John Yu, 2009. Cutting-plane training of structural SVMs. In *Machine Learning 77, no. 1 (2009): 27-59.*

Cour, Timothee, Florence Benezit, and Jianbo Shi, *2005.* Spectral segmentation with multiscale graph decomposition. In *Computer Vision and Pattern Recognition, vol. 2, pp. 1124-1131.* IEEE.

Andaló, F. A., P. A. V. Miranda, R. da S. Torres, and A. X. Falcão, 2010. Shape feature extraction and description based on tensor scale. In *Pattern Recognition 43, no. 1: 26-36.*

Liu, Guang-Hai, Lei Zhang, Ying-Kun Hou, Zuo-Yong Li, and Jing-Yu Yang, 2010. Image retrieval based on multi-texton histogram. In *Pattern Recognition 43, no. 7 (2010): 2380-2389.*

Ferrari, Vittorio, Frederic Jurie, and Cordelia Schmid, 2010. From images to shape models for object detection. In *International Journal of Computer Vision 87, no. 3: 284-303.*

Winn, John, Antonio Criminisi, and Thomas Minka, 2005. Object categorization by learned universal visual dictionary. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 2, pp. 1800-1807.* IEEE.

Arbeláez, Pablo, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir Bourdev, and Jitendra Malik, 2012. Semantic segmentation using regions and parts. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 3378-3385.* IEEE.

Chen, Xi, Arpit Jain, Abhinav Gupta, and Larry S. Davis, 2011. Piecing together the segmentation jigsaw using context. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.* IEEE.