Learning to Detect Scene Text Using a Higher-order MRF with Belief Propagation

Dong-Qing Zhang and Shih-Fu Chang
Department of Electrical Engineering, Columbia University,
New York, NY 10027, USA.
{dqzhang, sfchang}@ee.columbia.edu

Abstract

Detecting text in natural 3D scenes is a challenging problem due to background clutter and photometric/gemetric variations of scene text. Most prior systems adopt approaches based on deterministic rules, lacking a systematic and scalable framework. In this paper, we present a parts-based approach for 3D scene text detection using a higher-order MRF model. The higher-order structure is used to capture the spatial-feature relations among multiple parts in scene text. The use of higher-order structure and the feature-dependent potential function represents significant departure from the conventional pairwise MRF, which has been successfully applied in several low-level applications. We further develop a variational approximation method, in the form of belief propagation, for inference in the higher-order model. Our experiments using the ICDAR'03 benchmark showed promising results in detecting scene text with significant geometric variations, background clutter on planar surfaces or non-planar surfaces with limited angles.

1. Introduction

Text detection in natural 3D scenes is an important but challenging problem. Scene text provides direct information about the scene and event. It therefore can be used as effective features for image recognition, search and retrieval. Figure 1 shows some examples of scene text, illustrating the variations of 3D shape, lighting, and background cluttered-ness.

There have been much prior work on text detection, but most of them use *ad hoc* rules, lacking a systematic framework. Such approaches are difficult to generalize and achieve robust performance. They can be classified as texture based [10][1], region based [9][12], or hybrid [7]. Spatial layout analysis is also used in some of the systems in a rule based setting.

Text lines or words can be modeled as multi-part objects, where characters are disconnected parts. There has been some prior work on parts-based object detection and motion analysis. For example, in [2][5], a part constellation model is proposed to detect multi-part object with supervised and unsupervised learning. Spatial relations of parts are modeled using covariance matrix. In [4], Objects are modeled as trees. Detecting objects is realized by matching model trees and input pictures. In [13], human motion detection is realized by a parts-based approach, where the parts modeling is limited to triangulated decomposable graphs. In [8], a parts-based approach is proposed to detect human body. Boosting is applied to combine weak classifiers corresponding to different body part assemblies. In [15], a graph partitioning approach is developed to group individual parts into objects. However, no probabilistic structure is presented to support systematic learning.

Figure 1: The examples of scene text in images

Markov Random Field (MRF) is an undirected graphical model, having widespread applications in computer vision. MRF with pairwise potential and belief propagation has been applied in many low-level vision applications [6]. However, in order to detect multi-part objects, pairwise potential is often inadequate since it only captures two-node constraints. For example, in the text detection task, the pairwise potential cannot capture the unique spatial relationship that every three characters should be aligned on a straight line or a smooth curve. Another limitation of the traditional pairwise MRF model is that the state potential function does not incorporate the observed features. This makes it difficult to model the parts relations for general applications. For example, if we need to enforce that the "land" region should locate below the "sky" region in a natural image, the coordinate difference of the two regions is necessary to be taken into account.

In this paper, we propose a parts-based object detection system via learning a high-order MRF model. The methodology is applied to detect scene text in images. The problem is formulated as calculating the beliefs (the marginalized probability) at nodes that correspond to automatically segmented regions. In order to realize efficient probabilistic inference, a variatioanl method similar to Bethe approximation [14] is developed, which is converted into higher-order belief propagation equations. Supervised learning of this high-order MRF model is realized by maximum likelihood estimation.

Compared with prior systems. The proposed generative statistical framework incorporates higher-order constraints and takes advantage of the efficient inference algorithms. The proposed higher-order MRF model is also unique in that it uses potential functions considering inter-part relational attribute.

The higher-order MRF model is evaluated against the pairwise MRF model using a set of public benchmark images. The experiments show a substantial performance improvement accredited to the adoption of the higher-order statistical model. Moreover, the results also show that the presented method is extraordinarily robust even for text in severely cluttered background or with significant geometric variations. These evidences confirm the advantage of the higher-order MRF model for parts-based detection of scene text and probably broader categories of objects.

The paper is organized as follows: Section 2 describes the formation of the region adjacency graph. Section 3 formulates the text detection problem using MRF model. Section 4 presents the approach for designing potential functions, which is followed by the learning approaches described in section 5. Section 6 discusses the problem of multiple text lines with its solution. Experimental setting and results are described in section 7. Finally, section 8 summarizes the contribution and future work.

2. Region adjacency graph formation

Region adjacency graph (RAG) is used to model the properties of parts and parts relations. In this model, each node represents a segmented region, and each edge represents the likely relations between two regions. Region detection is realized by a mean-shift segmentation algorithm [3].

The edges between nodes are established according to the spatial positions of the regions. An edge is established only if the minimum distance between two regions is less than a predetermined threshold. The value of the minimum distance threshold (MDT) should allow three consecutive characters form a three-clique (i.e. triangle). Larger MDT would yield denser graph and more cliques, resulting in more computation cost. The optimal selection of MDT remains an unsolved issue for future exploitation. A straightforward method is to use a multi-pass detection procedure, in which a small MDT is started and subsequently increased until text is detected.

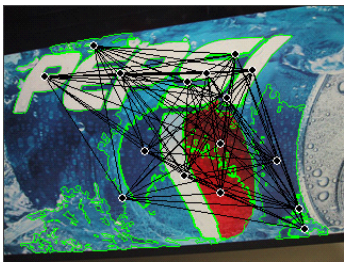

Figure 2: Region segmentation and adjacency graph. Segmented regions are indicated with green borders.

Nested regions, such as a bounding box and its encompassed characters, would not be connected by edges, in order to prevent unnecessary computation. Moreover, the regions that touch image boundaries are assumed to be background. They are therefore eliminated to save computation resources. One example of RAG is shown in the Figure 2.

3. Formulating text detection using MRF

Based on a RAG, the corresponding Markov Random Field (MRF) is constructed by attaching each node i a state random variable X_i taking value from a label set. In text detection, the label set consists of two labels: "text" ($X_i = 1$) or "non-text" ($X_i = 0$). The observed features include one-node features y_i extracted from each region i, and three-node features y_{ijk} extracted from every three connected regions (or a three-clique in RAG). Text detection therefore can be modeled as the probabilistic inference problem given all observation features. The overall relations can be modeled as a joint probability $p(x, y)$, with $x = \{x_i | 1 \leq i \leq N\}$ and $y = \{y_i, y_{ijk} | 1 \leq i, j, k \leq N\}$ where N is the region number. Text detection is therefore the problem of computing the marginal (or belief)

$$p(x_i|y) = \sum_{x \setminus x_i} p(x, y)/p(y) \tag{1}$$

Labeling a region as text or non-text is realized by likelihood ratio rest of the two opposite hypotheses ($x_i = 1$,text;$x_i = 0$,non-text):

$$\frac{p(x_i = 1|y)}{p(x_i = 0|y)} = \frac{p(x_i = 1, y)}{p(x_i = 0, y)} \geq \lambda \tag{2}$$

where λ is a threshold, which can be adjusted to vary the precision and recall rate.

3.1 Pairwise MRF

Pairwise MRF has been applied in a variety of low-level vision applications. The joint probability of a pairwise MRF can be written as

$$p(x,y) = \frac{1}{Z}\prod_{ij}\psi_{ij}(x_i,y_j)\prod_{i}\phi_i(x_i,y_i) \qquad (3)$$

where Z is the normalization constant, $\psi_{ij}(x_i,y_j)$ is the state comparability function, $\phi_i(x_i,y_i)$ captures the compatibility between the state and observation. The marginal probability of MRF can be calculated by Belief Propagation[14].

For multi-part object detection in cluttered background, one needs to identify the parts and group them into assemblies by accommodating the relations of the parts. This requires identifying structures in the adjacency graph, not only verifying the compatibility between two nodes. For example, in text detection, we need to verify if three regions are aligned on a straight line approximately. These constraints cannot be addressed by pairwise potentials and require functions involving more than two nodes.

3.2 Higher-Order MRF with belief propagation

To overcome the limitation of the pairwise MRF, we attempt to utilize MRF model with higher-order potentials while keeping computational efficiency of the belief propagation.

We adopt a unique generative model accommodating higher-order constraints, as visualized in Figure 3(Left), in which the observation features are not only defined at node but also three-cliques. Here we omit two-node potentials in order to simplify the computation and due to the fact that two-node constraints can be also incorporated in the three-node potentials if the graph is dense. It is not difficult to show that this model can be factorized as following:

$$p(x,y) = \frac{1}{Z}\prod_{ijk}\psi_{ijk}(x_i,x_j,x_k)p(y_{ijk}|x_i,x_j,x_k)\prod_{i}p(y_i|x_i) \qquad (4)$$

Where y_i is the observation feature vector at node n_i. $y_{ijk}$is the clique-level relational feature, which is extracted from the entire set of nodes in the clique and is used to characterize the attribute relations of the three nodes in the same clique. Examples of clique features may include the relations of locations, shapes, and symmetry among the nodes. The higher-order potentials and clique features allow this model perform local pattern matching and evolve towards higher-scale hidden structures. The potential function containing the clique features is crucial for multi-part relationship modeling. $\psi_{ijk}(x_i,x_j,x_k)$ is the potential imposing prior constraint,and $p(y_{ijk}|x_i,x_j,x_k)$,$p(y_i|x_i)$ is the probability density functions at three-cliques and nodes respectively. Here we implicitly assume that the observation features y_{ijk},y_i are independent.

By combining the prior constraints and emission probabilities, this model is equivalent to the following MRF with inhomogeneous potentials:

$$p(x,y) = \frac{1}{Z}\prod_{ijk}\psi'_{ijk}(x_i,x_j,x_k,y_{ijk})\phi'_i(x_i,y_i) \qquad (5)$$

where $\psi'_{ijk}(x_i,x_j,x_k,y_{ijk})$ and $\phi'_i(x_i,y_i)$ are the inhomogeneous potential functions.

In the rest of the paper, we use shorthand $\psi_{ijk}(x_i,x_j,x_k)$ and $\phi_i(x_i)$ for $\psi'_{ijk}(x_i,x_j,x_k,y_{ijk})$ and $\phi'_i(x_i,y_i)$ to simplify notations.

It has been shown that the belief propagation (BP) in pairwise MRF is equivalent to the Bethe approximation [14], a type of variational approximation. For higher-order MRF, we can use a similar variational approximation to obtain a higher-order version of the belief propagation. The detailed derivation is described in the Appendix.

The message passing rule for higher-order BP is as following (also illustrated in Figure 3(Right))

$$m_{jki}(x_i) \longleftarrow \lambda\sum_{x_j}\sum_{x_k}\phi_j(x_j)\phi_k(x_k)\psi_{ijk}(x_i,x_j,x_k) \prod_{(l,n)\in N_p(k)\setminus(i,j)} m_{lnk}(x_k) \prod_{(l,n)\in N_p(j)\setminus(i,k)} m_{lnj}(x_j) \qquad (6)$$

where λ is a normalization factor so that the message computation will not cause arithmetic overflow or underflow. $N_p(i)$ is the node pair set in which each node pair forms a three-clique with the node i. Once the messages converge, the beliefs are computed using

$$b_i(x_i) = k\phi_i(x_i)\prod_{(j,k)\in N_p(i)} m_{jki}(x_i) \qquad (7)$$

Where k is a normalization factor. Messages are uniformly initialized as a constant, typically 1.

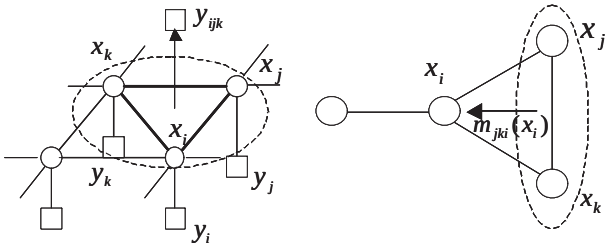

Figure 3: (Left) MRF with higher-order potential, node features, and clique-level relational features (Right) The message passing of the high-order belief propagation

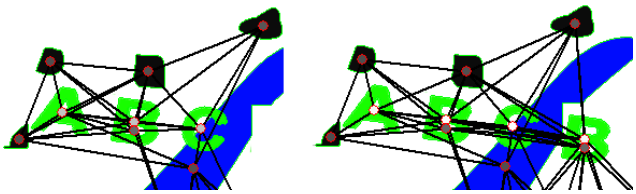

Figure 4: The reinforcement of the beliefs as the number of characters increases

Besides using the proposed higher-order BP, an alternative approach is to reduce the higher-order MRF to a pairwise MRF by clustering nodes and inserting additional nodes [16]. This process needs careful redesign of the potential functions and has to introduce extra delta-function-like potential functions, which may cause unstable message updates. It is therefore more straightforward to use the above higher-order version of belief propagation to perform inference.

Intuitively, the higher-order BP rules perform local pattern matching (by three-node potential with clique-level relational features) and pass around the evidences to the neighboring nodes to enhance or diminish the beliefs. To show this, Figure 4 shows the inference results from inputs with different numbers of characters. The brightness of each node (corresponding to a character) shown in the figure represents the belief of being "text" object. We note that more characters result in higher beliefs of the individual characters due to the interactions of the nodes.

Because the region adjacency graph is automatically generated, the topology of the graph is often loopy. Thus, in theory, the convergence of the BP cannot be guaranteed. However, our experiments on actual images so far have not observed significant divergences of message updates. This is probably due to the appropriate designs of the potential functions, or because the magnitudes of oscillations are too small to be observed.

4. Design of the Potential Functions

In order to effectively detect text, we need to carefully design the potential functions and emission probabilities in Eq. (4). The prior potentials are discrete probability mass functions. For emission probabilities, we have to adopt parametric probability density functions so that the model can be properly learned. In our system, we assume that the $p(y_{ijk}|x_i, x_j, x_k)$,$p(y_i|x_i)$ both have the form of Gaussian function or mixture of Gaussians.

In the following, we describe a few features for one-node and 3-node potential functions. Note the functions are general and other features can be added when useful, not only limited to the set we currently include in the implementation.

4.1 The one-node potential

In our current implementation, only aspect ratio is used as the feature for one-node potential. The distribution of the aspect ratio is modelled as Gaussian functions. There are two Gaussian *pdf*s: one for state 0 and another one for state 1, denoted as $G_0(y_i) = \mathcal{N}(\mu_0, \Sigma_0)$ and $G_1(y_i) = \mathcal{N}(\mu_1, \Sigma_1)$ respectively.

This model is accurate in the absence of segmentation errors. However, in many cases, multiple character regions may be merged due to poor region segmentation. To accommodate the mixed types of regions (single character regions and merged regions), we can use mixture of Gaussians to model the distribution.

4.2 The three-node potential

Three-node potential functions are used to enforce the spatial and visual relationship constraints on the cliques. The clique feature vector is extracted from every three-clique, the component of this vector is described as follows.

a) Minimum Angle

The feature is defined as the sinusoid of the minimum angle of the three-clique, i.e.:

$$y_{ijk}(1) = \sin(min_m \theta_m), m = 1, 2, 3.$$

where θ_m is one of the angles of the three-clique. For a text line, the minimum angle should be close to 0. For text on a non-planar surface, the angle is assumed to be small (e.g., text on a cylindrical surface). Note that the statistical modelling approach allows for soft deviation from a fixed value, and thus non-planar text with small angles can also be detected.

b) Consistency of the region inter-distance

For most scene text in an image, the difference of the character inter-distance is approximately the same. The feature is defined as ,

$$y_{ijk}(2) = \|\mathbf{v}_1\| - \|\mathbf{v}_2\|$$

where $\mathbf{v}_1$,$\mathbf{v}_2$ are the two laterals with the maximum angle in the triangle.

c) Maximum color distance

The feature is defined as the maximum pairwise color distance of the three regions. The use of this feature is based on the fact that the text regions in a text line have near uniform color distribution. The color distance is defined in the HSV space. For greyscale images, we can replace the color distance with the intensity difference although it may not be as robust as using color.

d) Height consistency of the character

The constraint enforces that the heights of the three character regions are approximately the same. The height divergence ratio is defined as

$$y_{ijk}(4) = (h_{max} - h_{min})/h_{min}$$

where h_{min} and h_{max} are the minimum and maximum height of the three regions. English characters usually are written with fixed discrete levels of height. Thus a mixture of Gaussian s model would be adequate.

5 Learning the Higher-Order MRF

Learning the Higher-Order MRF is realized by the maximum likelihood estimation. Suppose M images are used in training. We want to estimate the optimal parameter set $\hat{\theta}$ to maximize the likelihood of the whole set of images.

$$\hat{\theta} = argmax_\theta \sum_{m=1}^{M} \ln p(x^m, y^m|\theta) \tag{8}$$

x^m,y^m is the state vector and observation feature vector in the mth image, where x^m is labelled by annotators. According to Eq.(4), the joint log likelihood of x,y in one image can be factorized as

$$\ln p(x,y) = \sum_{ijk} \ln \psi(x_i, x_j, x_k|\theta_x) + \tag{9}$$

$$\sum_{ijk} \ln p(y_{ijk}|x_i, x_j, x_k, \theta_{y3}) + \sum_i \ln p(y_i|x_i, \theta_{y1}) - \ln Z$$

Where θ_x is the parameter for the state prior probability mass function. θ_{y3} is the parameter of the probability density function for the three-clique relational feature. θ_{y1} is for the one-node observation density. Since these three functions have independent parameters, the learning process can be carried out separately. The maximum likelihood estimates of θ_{y3},$\theta_{y1}$are obtained by simply calculating the mean and variance (or covariance matrix) of the Gaussian functions using the labeled data. θ_x is the prior distribution parameter, which can be calculated by counting the number of the state configurations in the training data.

The features presented in Section 3 require the potential functions of each clique invariant to permutation of label assignments of the states in the same clique. For a three-clique, there are 8 different configurations, but due to the permutation invariance, there will be only 4 different configurations $(x_i, x_j, x_k) = (111)$,$(x_i, x_j, x_k) = (100)$,$(x_i, x_j, x_k) = (100)$, $(x_i, x_j, x_k) = (000)$. As an example, $(x_i, x_j, x_k) = (111)$ means all three nodes in the clique are text regions. Correspondingly, we have Gaussian *pdf*s:

$$G_{111}(y_{ijk}) = p(y_{ijk}|x_i = 1, x_j = 1, x_k = 1) = \mathcal{N}(\mu_{111}, \Sigma_{111})$$

Figure 5: (Left) The miss in detecting multiple text lines due to cross-text-line (CTL) cliques. (Right) the results after potential function modification.

$$G_{110}(y_{ijk}) = p(y_{ijk}|x_i = 1, x_j = 1, x_k = 0) = \mathcal{N}(\mu_{110}, \Sigma_{110})$$
$$G_{100}(y_{ijk}) = p(y_{ijk}|x_i = 1, x_j = 0, x_k = 0) = \mathcal{N}(\mu_{100}, \Sigma_{100})$$
$$G_{000}(y_{ijk}) = p(y_{ijk}|x_i = 0, x_j = 0, x_k = 0) = \mathcal{N}(\mu_{000}, \Sigma_{000})$$

6 Modification of the potential functions for multiple text lines

The above detection algorithm works well when the image only contains single text line or the text lines are apart far away. However, if two or more text lines are close to one another, the algorithm will miss one or more text lines, as shown in the Figure 5. Such miss of detection is due to the negative constraint produced by the cross-text-line cliques (marked as dashed red lines in the Figure 5(Left)). In this case, the value of $G_{110}(y_{ijk})$,$G_{100}(y_{ijk})$,$G_{000}(y_{ijk})$ may be much larger than $G_{111}(y_{ijk})$ for a cross-text-line clique. The one-dimensional illustration of this situation is shown in the Figure 6, where the red (blue) curve indicates the potential function trained from "text"-"text"-"non-text" (text-text-text) cliques. Consequently, assigning the "non-text" label to one of the nodes in the cross-text-line three-clique will yield higher overall likelihood (as shown in the dashed line). One way to fix this problem is to modify the $G_{111}(y_{ijk})$ potential function such that it only has positive constraint effect within the desired feature range by the fol-

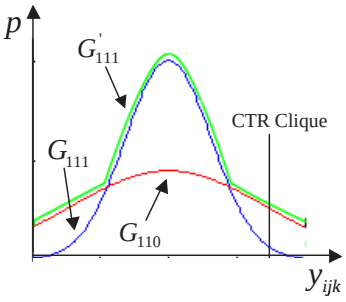

Figure 6: The potential functions and the modified version of $G_{111}(y_{ijk})$

lowing operator

$$G'_{111}(y_{ijk}) = \sup\{G_{110}(y_{ijk}), G_{100}(y_{ijk}), G_{000}(y_{ijk})\}$$

The resulting function is shown in Figure 6. Therefore if the three-node feature is far from the mean of the Gaussian, it no longer gives higher value for $G_{110}(y_{ijk})$,$G_{100}(y_{ijk})$,$G_{000}(y_{ijk})$ compared with $G_{111}(y_{ijk})$. This modification shows very significant improvement in the experiments while it does not significantly impact the non-text regions. Figure 5(Right) shows the results by using the modified potentials. One potential drawback of the above modification is that it may raise the belief of the non-text region and thus increase false alarms. However, if the text line has enough characters, the likelihood ratio test with higher threshold will correctly reject those non-text regions. Another problem is that some singleton regions disconnected with any other region may exist in image. No three-node potential constraint is imposed on these nodes. Consequently, the beliefs are totally determined by the one-node potential function, which is often inaccurate. To handle this problem, we can let the one-node potential only give negative constraint to non-text region if the features are quite different from the learned Gaussian mean. Thus, the one-node potential is modified using:

$$G'_1(y_i) = \sup\{G_1(y_i), G_0(y_i))\}$$

7 Experiments and results

To evaluate the proposed approach, we evaluate the performance using a public dataset used in the scene text detection competition in ICDAR 2003 [11]. The dataset contains 20 images with different natural conditions, for example, outdoor/indoor, background clutter, geometric variations, lighting variation, etc. All are colored images in the RGB format.. The resolution of these images is very high with a typical size 1280x960. To reduce the computation cost, we resize these images to about 640x480. This test set is limited since only images containing text are included. In order to evaluate the capability of the system in rejecting false regions in the cluttered images, another ten images with cluttered background but without text are added to the data set.

A cross-validation procedure is used to test the algorithm: the data is divided into two subsets, each of which alternates as training and testing set in a two fold cross-validation process. In the learning stage, each image first segmented by the mean-shift algorithm, and the segmented regions are manually labeled as text or non-text. Cross-text-line cliques are excluded from training to avoid confusion. We measure the precision and recall of the text region detection. Recall is the percentage of the ground truth text

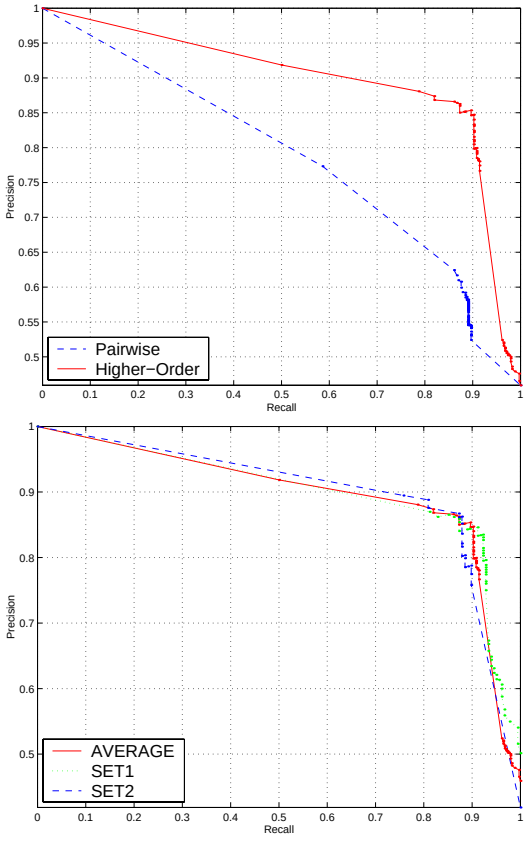

Figure 7: Precision recall curve, (Top) The comparison of ROC curve by using conventional pairwise MRF (dashed blue) and proposed method (red). (Bottom) The ROC curve of detection in set 1(green) set 2(blue) and average(red).

regions that are detected, while precision is the percentage of the correct text regions in the detected regions. The accuracy is measured at the character level.

We use the MRF model with pairwise potential as the baseline for comparison. The relational features are added into the pairwise model. It uses two features in the two-node potential - the color difference and height consistency. The one-node potential is the same as that used in the proposed higher-order MRF. The potentials are learned from labeled data. Inference is realized by standard belief propagation. A precision-recall curve (ROC curve) is generated by varying the threshold of the likelihood ratio, as shown in Eq.(2).

The performance comparison is shown in Figure 7(Top), which indicates that the higher-order MRF model significantly outperforms MRF with pairwise potential. Note interestingly there seems to be a turning point at 0.85/0.85 as precision/recall. The performance variance when using the cross-validation process is shown in Figure 7(Bottom), showing that the proposed method is stable over different training/testing partitions. Unfortunately, to the best

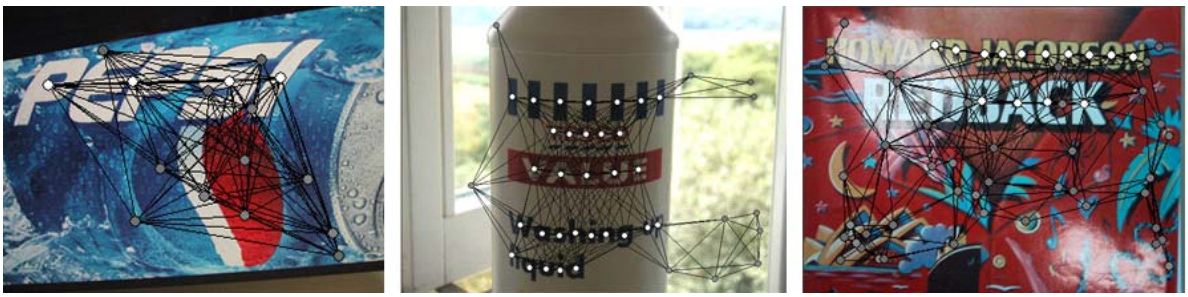

Figure 8: Example results from the higher-order MRF model (Brightness of nodes represents the probability as "text")

of our knowledge, there is no public-domain performance data over the same benchmark set that we can compare.

Note that these results have not included the text regions missed in the automatic segmentation process. The miss rate of region detection is about 0.33. This makes the optimal recall (including segmentation and detection) about 0.67. The region detection miss is mainly due to the small text size. The inference computation speed excluding segmentation varies from 0.5 second to 30 second per image on a Pentium III 800MHz PC depending on the number of the cliques. The average inference speed is 2.77 second per image. The speed of segmentation and region formation is about 0.2 second to 5 second per image, depending on the image size and the content complexity of the image. The speed is improvable, since no code optimization and look-up-table is used currently.

Figure 8 shows some detection results by the proposed higher-order MRF model. The results show that the method is very robust to background clutteredness and geometric variations, and is able to detect text on curved as well as planar surfaces. Detecting text on curved surfaces is hard to achieve by conventional systems using fixed rules, where hard constraints are usually used. Our system achieves improved performance in this aspect by using soft constraints captured by the statistical method. Furthermore, the issue of character merging is successfully handled if the merged regions remain on the same planar or curve surfaces. To compare with MRF with pairwise potential, Figure 9 shows its output, which illustrates that without using the higher-order constraints, the pairwise MRF is very vulnerable to the clutter.

Figure 9: Output from the pairwise MRF model (brightness of nodes represents the probability as "text")

8 Conclusion

We have presented a statistical method to detect text on planar or non-planar with limited angles in natural 3D scenes. We propose a MRF model with higher-order potential and incorporate intra-part relational features at the clique level. The proposed method is systematic, learnable, and robust to the background clutter and geometric variations. The system can be readily modified for the general multi-part object detection, for instance human body detection. We also plan to add more features and constraints into the system to further boost the detection performance.

Acknowledgments

We thank Shahram Ebadollahi, Lexing Xie, Winston Hsu, Yong Wang, Tian-Tsong Ng for valuable comments and discussions.

References

[1] L. Agnihotri and N. Dimitrova. Text detection for video analysis. In *Workshop on Content Based Image and Video Libraries*, pages 109–113, January Colorado, 1999.

[2] M. C. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry, vol 2. ECCV 1998.

[3] D. Comaniciu and P. Meer. Robust analysis of feature spaces: Color image segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 750–755, San Juan, Puerto Rico, June 1997.

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *Proceedings of the IEEE Com-*

puter Vision and Pattern Recognition Conference, pages 66–73, 2000.

[5] 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*, pages 66–73, 2003.

[6] W. Freeman, E.C.Pasztor, and O.T.Carmichael. Learning low-level vision. In *International Journal of Computer Vision,Vol 40, Issue 1*, pages 24–57, October 2000.

[7] J. Gao and J. Yang. An adaptive algorithm for text detection from natural scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Hawaii,2001.

[8] S. Ioffe and D. Forsyth. Probabilistic methods for finding people. In *International Journal of Computer Vision,Volume 43, Issue 1*, pages 45–68, June 2001.

[9] A. Jain and B.Yu. automatic text location in images and video frames. In *Pattern Recognition,vol.31,no.12*, pages 2055–2076, 1998.

[10] H. Li, D. Doermann, and O. Kia. Automatic text detection and tracking in digital video. In *IEEE Trans. on Image processing,Vol 9, No. 1*, January 2000.

[11] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. Icdar 2003 robust reading competitions. In *International Conference on Document Analysis and Recognition*, pages 682 – 687, 2003.

[12] J. Shim, C. Dorai, and R. Bolle. Automatic text extraction from video for content-based annotation and retrieval. In *Proc. 14th International Conference on Pattern Recognition,vol. 1*, pages 618–620, Brisbane, Australia,August 1998.

[13] Y. Song, X. Feng, and P. Perona. Towards detection of human motion. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition,Vol I*, pages 810–817, Hilton Head Island, South Carolina, June 2000.

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

[15] S. X. Yu and J. Shi. Object-specific figure-ground segregation. In *Computer Vision and Pattern Recognition (CVPR)*, Madison, Wisconsin, June 16-22 2003.

[16] Y.Weiss and W.T.Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. In *Neural Computation, Vol 13*, pages 2173–2200, 2001.

9 Appendix

Let $b_i(x_i)$ denotes the one-node belief and $b_{ijk}(x_i, x_j, x_k)$ denotes three-node belief. Let $N_p(i)$ be the node pair set in which each node pair forms a three-clique with the node i.

The energies associated with nodes and cliques can be define as

$$\begin{aligned} E_i(x_i) &= -\ln\phi_i(x_i) \\ E_{ijk}(x_i, x_j, x_k) &= -\ln\psi_{ijk}(x_i, x_j, x_k) - \ln\phi_i(x_i) \\ &\quad -\ln\phi_j(x_j) - \ln\phi_j(x_j). \end{aligned}$$

Then the Gibbs free energy [14] is

$$G = \sum_{ijk}\sum_{x_i x_j x_k} b_{ijk}(x_i, x_j, x_k)\big(E_{ijk}(x_i, x_j, x_k) + \ln b_{ijk}(x_i, x_j, x_k)\big) - \sum_i (q_i - 1)\sum_{x_i} b_i(x_i)\big(E_i(x_i) + \ln b_i(x_i)\big)$$

Where q_i is the degree of the node i. Therefore the Lagrangian multipliers and their corresponding constraints are

$$r_{ijk}: \sum_{x_i, x_j, x_k} b_{ijk}(x_i, x_j, x_k) - 1 = 0, \quad r_i: \sum_{x_i} b_i(x_i) - 1 = 0$$

$$\lambda_{jki}(x_i): \quad b_i(x_i) - \sum_{x_j}\sum_{x_k} b_{ijk}(x_i, x_j, x_k) = 0$$

The Lagrangian L is the summation of the G and the multiplier terms. To maximize L, we have

$$\begin{aligned} &\frac{\partial L}{\partial b_{ijk}(x_i, x_j, x_k)} = 0 \quad \Rightarrow \\ &\qquad \ln b_{ijk}(x_i, x_j, x_k) = E_{ijk}(x_i, x_j, x_k) + 1 + \lambda_{jki}(x_i) \\ &\qquad\qquad +\lambda_{kij}(x_j) + \lambda_{ijk}(x_k) + r_{ijk} \\ &\frac{\partial L}{\partial b_i(x_i)} = 0 \quad \Rightarrow \\ &\qquad \ln b_i(x_i) = -E_i(x_i) + \frac{1}{q_i - 1}\sum_{(j,k)\in N_p(i)} \lambda_{jki}(x_i) + r_i' \end{aligned}$$

where r_i' is the rearranged constant.

By using change of variable or defining message as:

$$\lambda_{jki}(x_i) = \ln \prod_{(l,n)\in N_p(i)\setminus(j,k)} m_{lni}(x_i)$$

We obtain the following equations

$$\begin{aligned} &b_i(x_i) = k\phi_i(x_i) \prod_{(j,k)\in N_p(i)} m_{jki}(x_i), \\ &b_{ijk}(x_i, x_j, x_k) = k\psi_{ijk}(x_i, x_j, x_k)\phi_i(x_i)\phi_j(x_j)\phi_k(x_k) \\ &\prod_{l,n\in N_p(i)\setminus j,k} m_{lni}(x_i) \prod_{l,n\in N_p(j)\setminus i,k} m_{lnj}(x_j) \prod_{l,n\in N_p(k)\setminus i,j} m_{lnk}(x_k) \end{aligned}$$

Apply the constraint $b_i(x_i) = \sum_{x_j}\sum_{x_k} b_{ijk}(x_i, x_j, x_k)$, we obtain

$$\begin{aligned} m_{jki}(x_i) \longleftarrow\!\lambda &\sum_{x_j}\sum_{x_k} \phi_j(x_j)\phi_k(x_k)\psi_{ijk}(x_i, x_j, x_k) \\ &\prod_{(l,n)\in N_p(k)\setminus(i,j)} m_{lnk}(x_k) \prod_{(l,n)\in N_p(j)\setminus(i,k)} m_{lnj}(x_j) \end{aligned}$$

Which is exactly the message passing rule in Eq. (6).