Image Parsing: Segmentation, Detection, and Recognition.

Zhuowen Tu
Department of Statistics
University of California at Los Angeles
Los Angeles, CA 90095

Alex Chen
Department of Statistics
University of California at Los Angeles
Los Angeles, CA 90095

Alan Yuille
Department of Statistics
University of California at Los Angeles
Los Angeles, CA 90095
yuille@stat.ucla.edu

Song-Chun Zhu
Department of Statistics
University of California at Los Angeles
Los Angeles, CA 90095

In International Conference of Computer Vision. Nice, France. pp 18-25. October. 2003.

Image Parsing: Unifying Segmentation, Detection, and Recognition

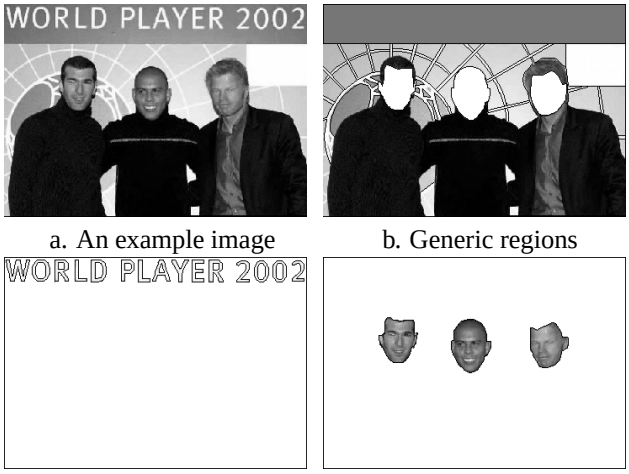
Zhuowen Tu, Xiangrong Chen, Alan L. Yuille, Song-Chun Zhu
University of California, Los Angeles
Los Angeles, CA, 90095
{ztu,xrchen,yuille,sczhu}@stat.ucla.edu

Abstract

We propose a general framework for parsing images into regions and objects. In this framework, the detection and recognition of objects proceed simultaneously with image segmentation in a competitive and cooperative manner. We illustrate our approach on natural images of complex city scenes where the objects of primary interest are faces and text. This method makes use of bottom-up proposals combined with top-down generative models using the Data Driven Markov Chain Monte Carlo (DDMCMC) algorithm which is guaranteed to converge to the optimal estimate asymptotically. More precisely, we define generative models for faces, text, and generic regions– e.g. shading, texture, and clutter. These models are activated by bottom-up proposals. The proposals for faces and text are learnt using a probabilistic version of AdaBoost. The DDMCMC combines reversible jump and diffusion dynamics to enable the generative models to explain the input images in a competitive and cooperative manner. Our experiments illustrate the advantages and importance of combining bottom-up and top-down models and of performing segmentation and object detection/recognition simultaneously.

1. Introduction

This paper presents an framework for parsing images into regions and objects. We demonstrate a specific application on outdoor/indoor scenes where image segmentation, the detection of faces, and the detection and reading of text are combined in an integrated framework. Fig. 1 shows an example in which a natural image is decomposed into generic regions (e.g. texture or shading), text, and faces. The tasks of obtaining these three constituents have traditionally been studied separately sometimes with detection and recognition being performed after segmentation [10], and sometimes with detection being a separate process, see for example [20]. But there is no commonly accepted method of combining segmentation with recognition. In this paper we show that our image parsing approach gives a principled way for addressing all three tasks simultaneously in a common framework which enables them to be solved in a cooperative and competitive manner. There are clear advantages to solving these tasks at the same time. For example, examination of the Berkeley dataset [11] suggests that human observers sometimes use object specific knowledge to perform segmentation but this knowledge is not used by current computer vision segmentation algorithms [9, 18]. In addition, as we will show, segmentation algorithms can help object detection by "explaining away" shadows and occluders. The application in this paper is motivated by the goal of designing a computer vision system for the blind that can segment images and detect and recognize important objects such as faces and text.



a. An example image b. Generic regions

c. Text d. Faces

Figure 1: Illustration of parsing an image into generic regions (e.g. texture and shading) and objects. An example image (a) is decomposed into two layers: (b). the region layer and the object layer which is further divided into text (c) and faces (d).

We formulate the problem as Bayesian inference. Top-down generative models are used to describe how objects and generic region models (e.g. texture and shading) generate the image intensities. The goal of image parsing is to invert this process and represent an input image by the parameters of the generative models that best describe it together with the boundaries of the regions and objects. It is crucial that all the generative models generate raw image intensi-

ties. This enables us to directly compare different models (e.g. by model selection) and thereby treat segmentation, detection and recognition in an integrated framework. For example, this requirement prevents us from using Hinton *et al*'s generative models for text [14] because these models generate image features and not raw intensities.

In order to estimate these parameters we use bottom-up proposals, based on low-level cues, to guide the search through the parameter space. More specifically, we combine bottom-up and top-down cues using the Data Driven Markov Chain Monte Carlo (DDMCMC) algorithm [18, 19] which is, in theory, guaranteed to converge to the MAP estimate asymptotically.

The bottom-up proposals for faces and text are learnt from training data by using a variant of the AdaBoost algorithm that outputs conditional probabilities [5] instead of classifications [20]. The use of conditional probabilities means that we do not have to make a firm decision based on AdaBoost and can instead use evidence from the generative models to resolve difficult cases. This improves performance particularly in the presence of occluders and shadows (which can be explained away by the other region models). The top-down generative models for faces and text are based on models with parameters estimated from training data. The bottom-up proposals and top-down generative models for generic regions are those used in previous work [18, 19] where they were tested on several hundred images.

The structure of this paper is as follow. Section (2) briefly reviews previous work on segmentation, face detection, and text detection and reading. In section (3), we describe the representation and the DDMCMC algorithm. Section (4) describes the generative models for faces and text. In section (5), we describe the use of AdaBoost algorithm to learn conditional probabilities distributions. DDMCMC jump and diffusion dynamics design is briefly discussed in section (6). Section (7) shows the results of using AdaBoost by itself and then the results obtained by our image parsing approach.

2. Related Work on Segmentation, Detection and Recognition

No existing work, to the best of our knowledge, combines segmentation, detection, and recognition in an integrated framework. These tasks have often been treated independently and/or sequentially. For example, Marr ([10]) proposed performing high-level tasks, such as object recognition, on intermediate representations obtained by segmentation and grouping.

Current segmentation algorithms [9, 18] perform well on large datasets although they do not yet achieve the ground truth results obtained by human subjects [11]. From one perspective, the work in this paper extends the DDMCMC segmentation algorithm ([18]) by introducing object specific models.

There has also been impressive work using image features for face detection [3, 15, 17, 21, 22, 20] and for text detection and recognition [8, 16, 1]. These approaches can all be used to specify bottom-up proposals for object detection in DDMCMC. It is most convenient for us to use the AdaBoost approach ([20]) because of it effectivness and its probabilistic interpretation, see section (5).

The generative models we use are based on generic region models (e.g. texture and shade) [18] and deformable templates [6, 7]. Similar models were proposed for text ([14]) but cannot be used here because they generate image features and not intensities.

3. Bayesian Formulation

We formulate image parsing as Bayesian inference. A scene interpretation includes a number of generic regions, letters and digits, and faces denoted by W^r, W^t, and W^f respectively. The region representation includes the number of regions K^r, and each region R_i has a label $\ell_i \in \{1, 2\}$ and parameter θ_i for its intensity model

$$W^r = (K^r, \{\mathbf{R}_i : i = 1, 2, ..., K^r\}),$$

where $\mathbf{R}_i = (R_i, \theta_i, \ell_i)$. Similarly, we have

$$W^t = (K^t, \{T_i : i = 1, 2, ..., K^t\}), and$$

$$W^f = (K^f, \{F_i : i = 1, 2, ..., K^f\}),$$

where $T_i = (L_i, \vartheta_i)$ and $F_i = (R_i, \varrho_i)$.

Thus, the solution vector is of the form

$$W = (W^r, W^t, W^f).$$

The goal is to estimate the most probable interpretation of an input image $\mathbf{I}$. This requires computing the W^* that maximizes *a posteriori* probability over, Ω, the solution space of W,

$$W^* = \arg\max_{W \in \Omega} p(W|\mathbf{I}) = \arg\max_{W \in \Omega} p(\mathbf{I}|W)p(W). \quad (1)$$

The likelihood $p(\mathbf{I}|W)$ specifies the image generating processes from W to $\mathbf{I}$ and the prior probability $p(W)$ represents our prior knowledge of the world. By assuming the mutual independence between W^r, W^t, W^f we have the prior model

$$p(W) = \left(p(K^r)\prod_i^{K^r} p(\mathbf{R}_i)\right)\left(p(K^t)\prod_i^{K^t} p(T_i)\right)\left(p(K^f)\prod_i^{K^f} p(F_i)\right).$$

To make generic regions, text, and faces directly comparable, we define

$$p(R_i) \propto exp\{-\gamma|R_i|^\alpha - \lambda|\partial R_i|\}. \quad (2)$$

Details about the definition of region model can be found in [18]. We define $p(\mathbf{R}_i) = p(R_i)$, $p(F_i) = p(R_i)$, and $p(T_i) = p(L_i)$.

The likelihood function can be written as

$$p(\mathbf{I}|W) = (\prod_{i}^{K^r} p(\mathbf{I}_{R_i}; \theta_i, \ell_i) \prod_{i}^{K^t} p(\mathbf{I}_{L_i}; \vartheta_i) \prod_{i}^{K^f} p(\mathbf{I}_{R_i}; \varrho_i).$$

We use the DDMCMC algorithm for estimating W^*. DDMCMC [18] is a version of the Metroplis-Hastings algorithm and hence is guaranteed to converge to samples from the posterior. It employs data-driven bottom-up proposals $q(W \mapsto W'|\mathbf{I})$ to drive the convergence of top-down generative models. Moves are selected by sampling from $q(W \mapsto W'|\mathbf{I})$ and they are accepted with probability $\alpha(W \mapsto W')$:

$$\alpha(W \to W') = \min(1, \frac{p(W'|\mathbf{I})}{p(W|\mathbf{I})} \cdot \frac{q(W' \to W|\mathbf{I})}{q(W \to W'|\mathbf{I})}).$$

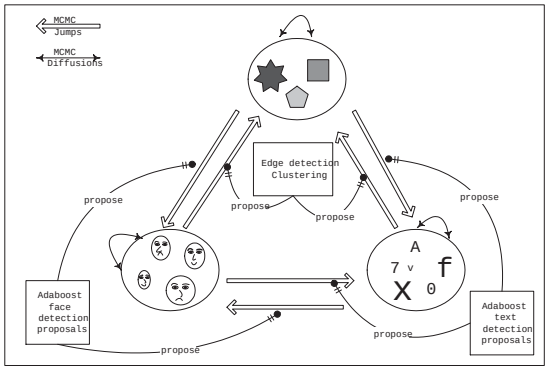


Figure 2: Illustration of the DDMCMC approach for segmentation, detection, and recognition.

These moves can be subdivided into two basic types, jumps which realize moves between different dimensions and diffusion which realizes moves within fixed dimension. Firstly, *jump moves* which are discrete and correspond to the birth/death of region hypotheses, splitting and merging of regions, and switching the model for a region (e.g. changing from a texture model to a spline model), changing a generic region into a face, creating a letter, etc. Secondly, *diffusion processes* which correspond to continuous changes such as altering the boundary shape of a region, text or a face and changing the parameters of a model used to describe a region. Fig. 2 gives a schematic illustration of how the jump and diffusion dynamics proceed driven by bottom-up proposals.

The bottom up proposals for faces and text are learnt using a probabilistic version of AdaBoost, see section (5). The bottom up proposals for generic regions (e.g. shading and texture) were described in [18].

In summary, bottom-up proposals drive top-down generative models which compete with each other to explain the image.

4. Generative Models

This section describes our generative models. We will concentrate on our text model for space. The models will be used for text detection and reading.

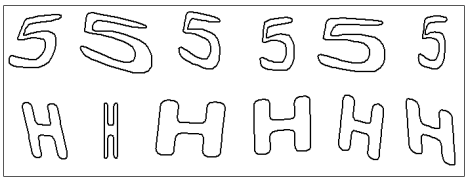


Figure 3: Random samples drawn from the generative models for letters and digits.

In natural scenes, text such as street signs and store names are usually painted in regular fonts, which can be modeled by deformable templates. We define a set of templates, $TP = \{TP(i), i = 1..62\}$, corresponding to ten digits and twenty six letters in upper case and lower case. Each template $TP(i)$ is represented by an outer boundary and 0 or up to 2 inner boundaries, each of which is modeled by twenty five control points. Given an input image, we need to inference how many text symbols there are, which type they are and what deformations they have. From the standard shape of each text, we denote its shape by

$$L_i = (c_i, S_i, M_i),$$

where $c_i \in \{1..62\}$ is the index of template $T(c_i)$, S_i includes positions of control points, and M_i denotes the affine transformation of S_i. Thus, the prior distribution on L_i can be specified as

$$p(L_i) = p(c_i)p(S_i|c_i)p(M_i).$$

Here $p(c_i)$ is a uniform distribution on all the digits and letters. $p(S_i|c_i)$ is the probability of perturbation of control points S_i w.r.t. the template $TP(c_i)$ and it is computed by the distance between contour points of S_i and the template $TP(c_i)$. Using quadratic B-Splines, the contour points can be computed as $G_{TP(c_i)} = U \times M_s \times T(c_i)$ and $G_{S_i} = U \times M_s \times S_i$. Thus the distribution are expressed as

$$p(S_i|c_i) \propto exp\{-\gamma Area(G_{L_i})^{\alpha} - \int_s \frac{D(G_{S_i}(s), G_{T(c_i)}(s))^2}{2\sigma^2} ds\},$$

where $D(G_{S_i}(s), G_{TP(c_i)}(s))$ is the distance between contour point $G_{S_i}(s)$ and $G_{T(c_i)}(s)$. The prior on affine transformation M_i is defined such that severe rotation and distortion are penalized. Figure (3) shows some samples drawn

from the above model. The intensities of the text exhibit smooth shading pattern and we use a quadratic form

$$J(x, y; \vartheta) = ax^2 + bxy + cy^2 + dx + ey + f,$$

with parameters $\vartheta = (a, b, c, d, e, f, \sigma)$. Therefore, the generative model for pixel (x, y) on the text is

$$\mathbf{I}(x, y) = \mathbf{J}(x, y) + N(0, \sigma^2).$$

Figure 4: Samples drawn from the PCA face model.

The generative model for faces is simpler and uses techniques like Principal Component Analysis (PCA) to obtain representations of the faces. Lower level features, also modeled by PCA, can be added [12]. Fig. 4 shows some faces sampled from the PCA model. We also add other features such as occlusion process, as described in Hallinan et al [7].

5. AdaBoost and Conditional Probabilities

The standard AdaBoost algorithm, see for example [20], produces a binary decision – e.g. face or non-face. Here we follow Friedman *et al* [5] and allow AdaBoost to estimate the conditional probabilities instead.

Standard AdaBoost learns a "strong classifier" $H_{\text{Ada}}(\mathbf{I})$ by combining a set of T "weak classifiers" $\{h_t(\mathbf{I})\}$ using a set of weights $\{\alpha_t\}$:

$$H_{\text{Ada}}(\mathbf{I}) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{I})).$$

where the selection of features and weights are learned through supervised training off-line [4].

Our variant of AdaBoost outputs conditional probabilities and is based on the following theorem [5].

Theorem. *The AdaBoost algorithm trained on data from two classes A, B converges, in probability, to estimates of the conditional distributions $q(A|\mathbf{I}), q(B|\mathbf{I})$ of the data $\mathbf{I}$:*

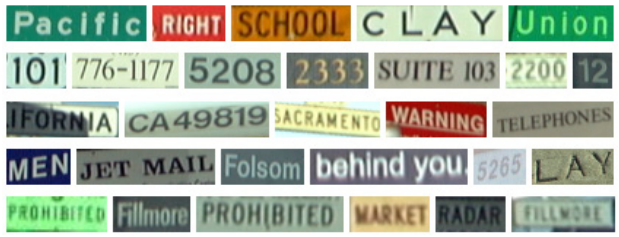
$$\frac{e^{\sum_{t=1}^{T} \alpha_t h_t(\mathbf{I})}}{\exp^{\sum_{t=1}^{T} \alpha_t h_t(\mathbf{I})} + \exp^{-\sum_{t=1}^{T} \alpha_t h_t(\mathbf{I})}} \mapsto q(A|\mathbf{I}) \qquad (3)$$

$$\frac{\exp^{-\sum_{t=1}^{T} \alpha_t h_t(\mathbf{I})}}{\exp^{\sum_{t=1}^{T} \alpha_t h_t(\mathbf{I})} + \exp^{-\sum_{t=1}^{T} \alpha_t h_t(\mathbf{I})}} \mapsto q(B|\mathbf{I}), \qquad (4)$$

We use AdaBoost to learn these conditional probability distributions so that they can activate our generative models (in practice, the conditional probabilities are extremely small for almost all parts of an image). This allows us to avoid premature decisions about the presence or absence of a face. By contrast, standard AdaBoost can be thought of as using these conditional distributions for classification by the log-likelihood ratio test.

5.1. AdaBoost Training

We used standard AdaBoost training methods [4, 5] combined with Viola and Jones' cascade approach using asymmetric weighting [20]. The cascade enables the algorithm to rule out most of the image as face, or text, locations with a few tests and allows computational resources to be concentrated on the more challenging parts of the images (i.e. in our terminology, regions where the conditional probabilities are non-negligible).



a. Text (From these, we extracted text segments.)

b. Faces

Figure 5: Positive training examples for AdaBoost.

Our text database contains 561 text images, some of which can be seen in Fig. 5. They are extracted by hand from 162 static images of San Francisco street sceens. More than half of the images were taken by blind volunteers (so as to simulate the conditions under which our system will eventually be used). We divided each text image into several overlapping text segments with fixed width-to-height ration 2:1. There are in total 7,000 text segments in the positive training set. The negative examples were obtained by a bootstrap process similar to Drucker et al [2]. First we selected negative examples by randomly sampling from windows in the image dataset. After training with these samples, we applied the AdaBoost algorithm to classify all windows in the training images (at a range of sizes). Those misclassified as text were then used as negative examples

for learning conditional distributions. The image regions most easily confused with text were vegetation, repetitive structures such as railings or building facades, and some chance patterns. The features used for AdaBoost were image tests corresponding to the statistics of elementary filters – see technical report for more details.

The AdaBoost for faces was trained in a similar way. This time we used Haar basis vectors [20] as elementary features. We used the FERET [13] database for our positive examples, see Fig. 5, and by allowing small rotation and translation transformation we had 5,000 positive examples. We used the same strategy as described above (for text) to obtain negative examples.

In both cases, we tested AdaBoost for detection (i.e. for classification) using a number of different thresholds. In agreement with previous work on faces [20], AdaBoost gave very high performance with low false positives and false negatives, see table (1). But the low error rates are slightly misleading because of the enormous number of windows in each image, see table (1). This means that by varying the threshold, we can either eliminate the false positives or the false negatives but not both at the same time. We illustrate this by showing the face regions and text regions proposed by AdaBoost in figure (6). If we attempt classification by putting a threshold then we can only correctly detect all the faces at the expense of false positives.

Object	False Positive	False Negative	Images	Subwindows
Face	65	26	162	355,960,040
Face	918	14	162	355,960,040
Face	7542	1	162	355,960,040
Text	118	27	35	20,183,316
Text	1879	5	35	20,183,316

Table 1: Performance of AdaBoost at different thresholds.

Instead, we prefer to use AdaBoost as proposals to generative models. Also, generic region proposals can find text that AdaBoost misses, for example, the '9' in the bottom panel of figure (6) will fail to be detected by AdaBoost for text, but will be detected as a generic "shading region" and later recognized as a '9'.

6. Computation and algorithm

Given the mixture models in the formulation and our interest in obtaining nearly globally optimal solutions, we design Markov chains to simulate walks in the solution space Ω.

6.1. Diffusion equations

Given W with fixed number of generic regions, text, and faces, and their model parameters, the interactions between these elements are governed by PDEs for the boundary and template deformation. Fig.7 illustrates the motion. The

Figure 6: The boxes show faces and text as detected by AdaBoost. Observe the false positives due to vegetation, tree structure, and random image patterns. It is impossible to select a threshold which has no false positives and false negatives for this image. Instead we use AdaBoost to output conditional probabilities, which will take their biggest values in the boxes, which are used in the DDMCMC algorithm.

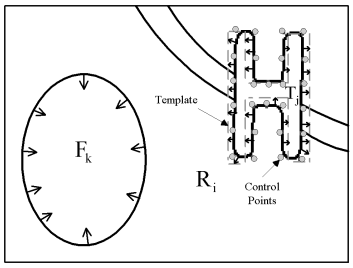


Figure 7: The diffusion and evolution of the boundaries is driven by the competition PDEs between regions.

PDEs are derived as greedy steps for minimizing the energy functions (or minus log-posterior probability) through variational calculus, especially the Green's theory. For a boundary whose left and right components are regions or faces, its motion equation is similar as the one in the region competition algorithm [23]. There are three energy terms for region R_i: one for the likelihood, and two for the prior on area $|R_i|$ and perimeter $|\partial R_i|$ defined in eqns.(2).

$$E(\mathbf{R}_i) = \int\int_{R_i} -\log p(\mathbf{I}(x,y)|\theta_{\ell_i})dxdy + \gamma|R_i|^{\alpha} + \lambda|\partial R_i|.$$

Likewise, for a letter T_j

$$E(T_j) = \int\int_{L_j} \log p(\mathbf{I}(x,y)|\theta_{\ell_j})dxdy + \gamma|L_j|^{\alpha} - logp(L_j).$$

Let v be a point on the boundary of R_i and T_j, i.e.

$$v(s) = (x(s), y(s)) \text{ on } \Gamma(s) = \partial R_i \cap \partial T_j.$$

The motion equation for control points can be obtained as

$$\begin{aligned}
\frac{dS_m}{dt} &= -\frac{\delta E(\mathbf{R}_i)}{\delta S_m} - \frac{\delta E(T_j)}{\delta S_m} \\
&= \int [-\frac{\delta E(\mathbf{R}_i)}{\delta v} - \frac{\delta E(T_j)}{\delta v}] \frac{1}{|\mathbf{J}(s)|} ds \\
&= \int \mathbf{n}(v)[\log \frac{p(\mathbf{I}(v);\theta_{\ell_i})}{p(\mathbf{I}(v);\theta_{\ell_j})} + 0.9\gamma(\frac{1}{|D_j|} - \frac{1}{|D_i|}) \\
&+ \frac{D(G_{S_j}(v) - G_T(v))^2}{2\sigma^2}] \frac{1}{|\mathbf{J}(s)|} ds,
\end{aligned}$$

where $\mathbf{J}(s)$ is the Jacobian matrix for the spline function. Thus, control points are moved by the forces transferred from boundary points through this motion equation.

6.2. Jump dynamics

Structural changes in the solution W are realized by Markov chain jumps (see [18]). We design the following reversible jumps between:

(i) two regions – model switching: $\theta_1 \leftrightarrow \theta_2$
(ii) a region R and a text T: $\mathbf{R} \leftrightarrow T$
(iii) a region R and a face F: $\mathbf{R} \leftrightarrow F$
(iv) split or merge a region: $(\mathbf{R}_k) \leftrightarrow (\mathbf{R}_i, \mathbf{R}_j)$
(v) birth or death of a text: $T\ \{,\} \leftrightarrow \{,T\}$.

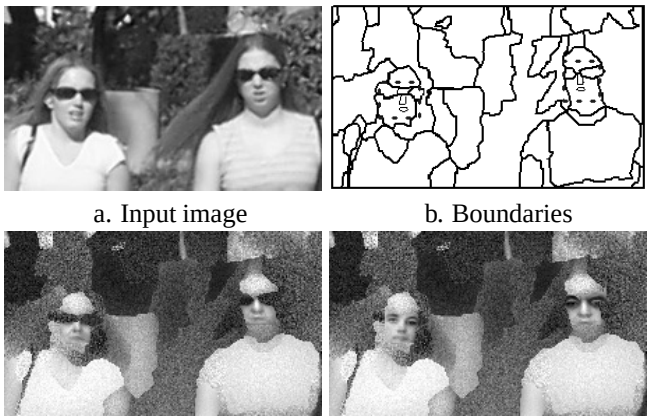
The Markov chain selects one of the above moves at each time, triggered by bottom-up compatibility conditions.

7. Experiments

We test the proposed image parsing algorithm on a number of outdoor/indoor images. The speed is comparable to segmentation methods such as normalized cuts [9]. A detailed description and demonstrations of convergence of the basic DDMCMC paradigm can be seen in [18].

The results of our experiments are shown in three ways: (i) synthesized images sampled from $P(\mathbf{I}|W^*)$ using the parameters and boundaries W^* estimated by the DDMCMC algorithm, (ii) the segmentation boundaries of the image, and (iii) the text and faces extracted from the image, with text symbols indicating the text that has been correctly read by the algorithm. Fig. 9 shows that we can obtain segmentation, face detection (at a range of scales), and text detection and correct text reading. Moreover, the synthesized images are fairly realistic.

High-level knowledge helps segmentation to overcome problem of oversegmentation and provides better synthesis in comparison to [18]. Segmentation supports the recognition of objects. Intuitively, the generative models for faces, text, texture, and shading compete to explain the image data. But this competition also enables cooperation. For example, the dark glasses on the two women in Fig. 8.a are detected as generic "shading regions" and *not* as part of the faces. They are then treated as "outlier" data which the face model does not need to explain and hence increases the robustness of the face detection. In Fig. 8.d, we show the synthesised faces by removing the sun-glasses. The Parking image in the third row of Fig. 9 also illustrates another example of cooperativity. For this image, where the bottom-up text AdaBoost model failed to propose the digit "9" as a text region, see Fig. 9. However, the generic region processes detected it as a homogeneous image region and then proposed it as a letter "9" which was confirmed by the generative model.



a. Input image b. Boundaries

c. Synthesis 1 d. Synthesis 2

Figure 8: Parsing a close-up of the Parking Image. Generic "shading region" processes detect the dark glasses and so the face model does not have to explain this part of the data. Otherwise the face model would have difficulty because it would try to fit the glasses to eyes. Standard AdaBoost would only correctly classify these faces at the expense of false positives, see Fig. 6.

The Street Image, see the forth row of Fig. 9, shows an example where the generative models for faces were required to reject face regions wrongly proposed by AdaBoost, see Fig. 6. Moreover, this example shows cooperatively because the shaded regional models were used to "explain away" shadows that otherwise would have disrupted the detection and reading of the text (observe the heavy shading patterns on the text "Heights Optical").

The ability to synthesize the image after estimating the parameters W^* is an advantage of our Bayesian approach, see [18]. The synthesis helps illustrate the successes, and

sometime the weaknesses, of our generative models. Moreover, the synthesized images show how much information about the image has been captured by our models. In table (2), we show the number of bytes used in our representation W^* and compare them to the jpeg compression for the equivalent images. Image encoding is not the goal of our current work, however, and more sophisticated generative models would be needed to synthesize very realistic images. Nevertheless, our synthesized images are fair approximations and we could reduce the coding of W^* substantially by encoding the boundaries more efficiently (at present, we code boundary pixels independently).

Image	Stop	Soccer	Parking	Street	Westwood
jpg bytes	23,998	19,563	23,311	26,170	27,790
W^*	4,886	3,971	5,013	6,346	9,687

Table 2: Comparison of bytes required by jpg and W^* for each image.

8. Summary and Conclusions

This paper has introduced a framework for image parsing by defining generative models for the processes that create images including specific objects and generic regions such as shading and texture. Bottom-up proposals are learnt by the AdaBoost algorithm which provides conditional probabilities for the presence of objects in the image. These conditional probabilities enable inference by rapid search through the parameters of the generative models, and the segmentation boundaries, using the DDMCMC algorithm.

We implement our system using generative models for text and faces combined with generic models for shaded and textured regions. Our approach enables these different models to compete and cooperate to describe the input images. We were able to segment the images, detect faces, and detect and read text in city scenes. Our experiments showed several cases where the shaded models helped face and text detection by explaining away shadows and occluders (sun-glasses). In turn, the text and face models improved the quality of the segmentations.

The current limitations of our approach lie in the limited class of objects we currently model. This limitation was motivated by our application goal of detecting text and faces for the visually disabled. But, in principle, our approach can include broad types of objects.

Acknowledgments

This work is supported by the National Institute of Health (NEI) RO1-EY 012691-04 and an NSF grant 0240148. The authors thank the Smith-Kettlewell research institute for providing us with text training images.

References

[1] S. Belongie, J. Malik, and J. Puzicha, "Matching shapes", *Proc. of ICCV*, 2001.

[2] H. Drucker, R. Schapire, and P. Simard, "Boosting performance in neural networks," *Intl J. Pattern Rec. and Artificial Intelligence*, vol. 7, no. 4, 1993.

[3] F. Fleuret, and D. Geman, "Coarse-to-Fine face detection", *IJCV*, June, 2000.

[4] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm", *Proc. of ICML*, 1996.

[5] J. Friedman, T. Hastie and R. Tibshirani. "Additive logistic regression: a statistical view of boosting", Dept. of Statistics, Stanford Univ. Technical Report. 1998.

[6] U. Grenander, Y. Chow, and D.Keenan. *HANDS: A Pattern Theoretic Study of Biological Shapes*. Springer-Verlag, 1990.

[7] P.Hallinan, G.Gordon, A.Yuille, P.Giblin, and D. Mumford, "Two and Three Dimensional Patterns of the Face", AKPeters, 1999.

[8] A. K. Jain and B. Yu, "Automatic text localication in images and video frames", *Pattern Recognition*, 31(12), 1998.

[9] J. Malik, S. Belongie, T. Leung and J. Shi, "Contour and texture analysis for image segmentation", *IJCV*, vol.43, no.1, 2001.

[10] D. Marr. *Vision*. W.H. Freeman and Co. San Francisco, 1982.

[11] D. Martin, C. Fowlkes, D. Tal and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics", *Proc. of ICCV*, 2001.

[12] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation", *IEEE Trans. PAMI*, vol.19, no.7, 1997.

[13] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms", *Image and Vision Computing J*, vol. 16, no. 5, 1998.

[14] M. Revow, G.K.I. Williamst and G.E. Hinton, "Using generative models for handwritten digit recognition", *IEEE Trans. PAMI*, vol.18, 1996.

[15] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection", In *IEEE Trans. PAMI*, vol. 20, 1998.

[16] T. Sato, T. Kanade, E. Hughes, and M. Smith, "Video OCR for Digital News Archives," *IEEE Intl. Workshop on Content-Based Access of Image and Video Databases*, Jan., 1998.

[17] H. Schniederman and T. Kanade, "A Statistical method for 3D object detection applied to faces and cars", *Proc. of Computer Vision and Pattern Recognition*, 2000.

[18] Z. Tu and S.C. Zhu, "Image segmentation by Data Driven Markov chain Monte Carlo", *IEEE Trans. PAMI*, vol. 24, no. 5, 2002.

[19] Z. Tu and S.C. Zhu, "Parsing images into regions and curve processes", *Proc. of ECCV*, June, 2002.

[20] P. Viola and M. Jones, "Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade", In *Proc. of NIPS01*, 2001.

[21] M. Weber, W. Einhuser, M. Welling, P. Perona, "Viewpoint-invariant learning and detection of human heads", *Proc. of Int. Conf. Automatic Face and Gesture Recognition*, 2000.

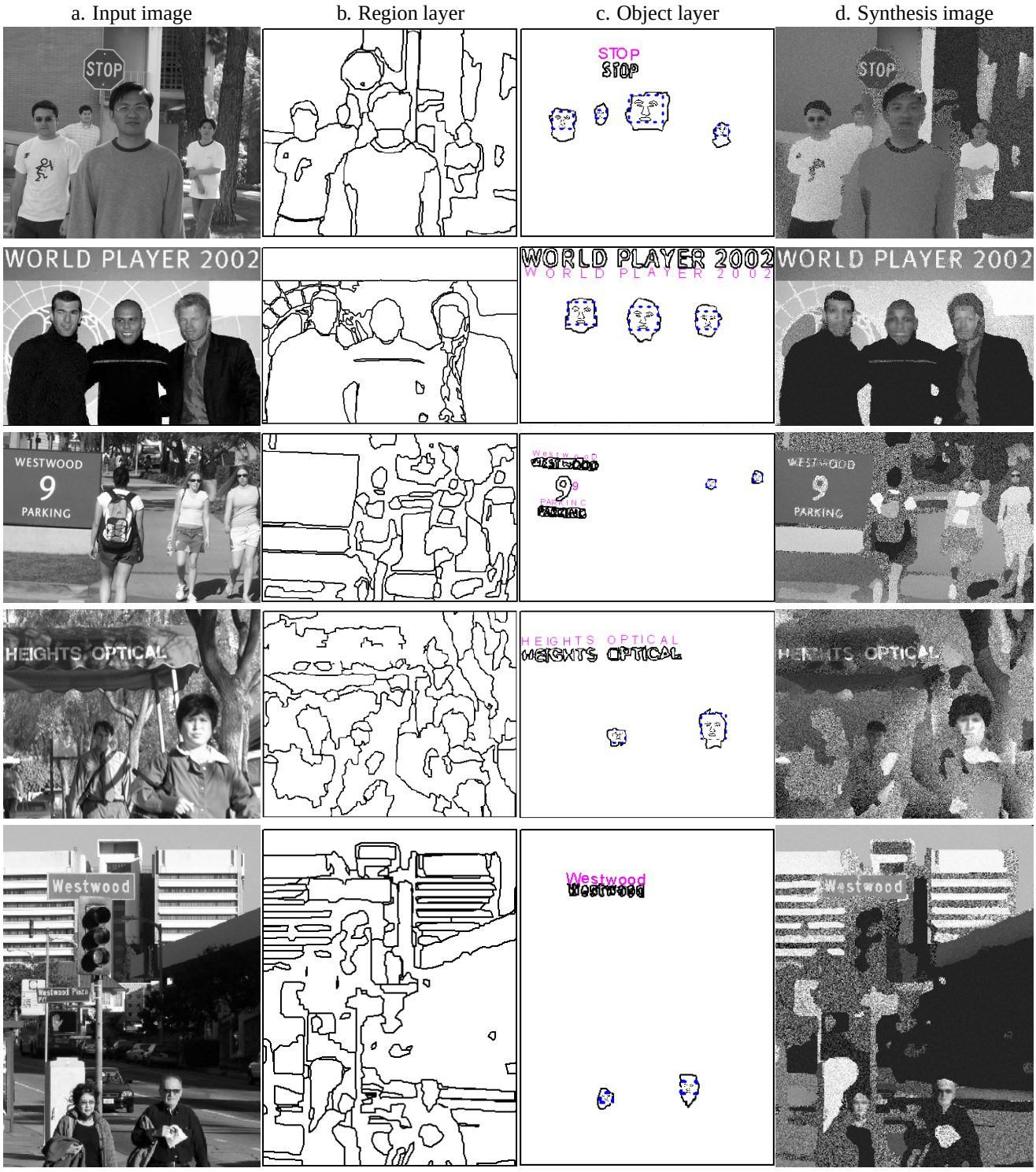


Figure 9: Results of segmentation and recognition on several outdoor/indoor images: Stop sign (row 1), Soccer (row 2), Parking (row 3), Street (row 4), and Westwood (row 5).

[22] Ming-Hsuan Yang, N. Ahuja, D. Kriegman, "Face detection using mixtures of linear subspaces", In *Proc. of Int. Conf. Automatic Face and Gesture Recognition*, 2000.

[23] S. C. Zhu and A. L. Yuille, "Region competition," *IEEE Trans. PAMI*, vol. 18, no. 9, 1996.