Unsupervised Learning of Categorical Segments in Image Collections

Marco Andreetto* Lihi Zelnik-Manor† Pietro Perona*

*Dept. of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, United States

†Dept. of Electrical Engineering
Technion - Israel Institute of Technology
Haifa, 32000, Israel

Abstract

Which one comes first: segmentation or recognition? We propose a probabilistic framework for carrying out the two simultaneously. The framework combines an LDA 'bag of visual words' model for recognition, and a hybrid parametric-nonparametric model for segmentation. If applied to a collection of images, our framework can simultaneously discover the segments of each image, and the correspondence between such segments. Such segments may be thought of as the 'parts' of corresponding objects that appear in the image collection. Thus, the model may be used for learning new categories, detecting/classifying objects, and segmenting images.

1 Introduction

Grouping (or image segmentation) and recognition have long been associated in the vision literature. Three views have been entertained on their relationship: (a) grouping is a useful preprocessing step for recognition: first you divide up the image into homogeneous regions, then recognition proceeds by classifying and combining these regions [1, 2, 3, 4], (b) segmentation is a by-product of recognition: once we know that there is an object in a given position, we may posit the components of the object and this may help segmentation [5, 6], (c) segmentation and recognition can be performed independently; in particular, recognition does not require segmentation nor grouping [7, 8, 9, 10, 11, 12]. It is important to note that these views are all true: segmentation and recognition are not necessary for each other, but both benefit from each other. It is therefore intuitive that recognition and grouping/segmentation might have to be carried out simultaneously, rather than in sequence, in order to obtain the best results. Furthermore: it would be advantageous to combine both model learning, as well as recognition, with segmentation. We explore here the idea of jointly carrying out recognition and segmentation – we propose and study what is perhaps the simplest such probabilistic formulation.

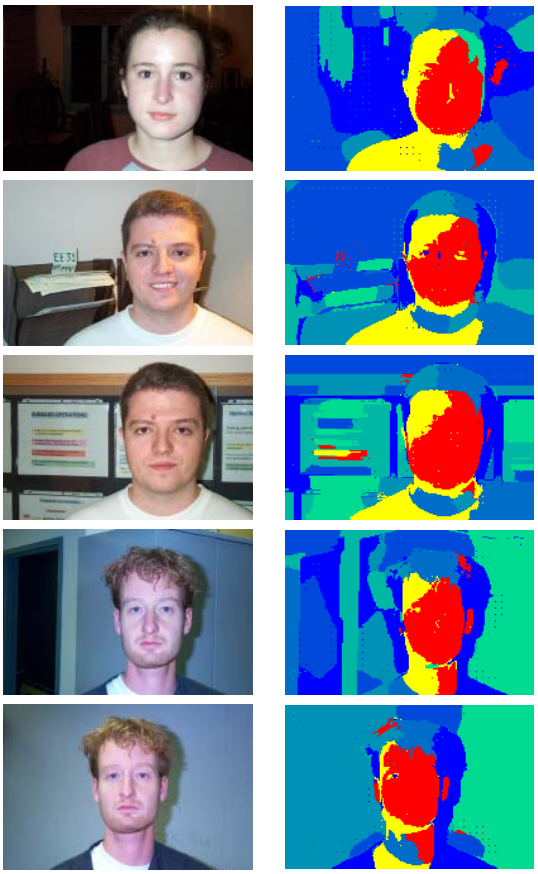

Figure 1: Categorical segments/parts are learned across multiple images in a collection. As can be seen segments on the background are colored blue while facial topics are marked in yellow and red. The background topics are spread over the entire image space and thus their shape densities are not too informative. The face topics, however, appear with consistently the same shape. The red topic captures the left side of the face while the yellow one captures the top right and neck.

Our work builds upon recent work on recognition and segmentation. First, we choose to categorize images us-

ing simple statistics of 'visual words' as features. Using 'bags of visual words' to characterize the appearance of an image, or an image patch, combines an idea coming from the literature on texture, where Leung and Malik [13] proposed vector-quantizing image patches to produce a small dictionary of 'textons', and an idea from the literature on document retrieval, where statistics of words are used to classify documents [14]. Early visual recognition papers using 'bags of visual words' considered the image as a single bag [15, 16, 12], while recently we have seen efforts either to classify independently multiple regions per image, after image segmentation [3, 4] or to force nearby visual words to have the same statistcs [17]. Second, recent literature on image segmentation successfully combines the notion that images are 'piecewise smooth' with the notion that segments shapes are more often than not 'simple'. These insights have been pursued both with parametric probabilistic models [18, 19], with non-parametric deterministic models [20] and with probabilistic hybrid parametric-nonparametric models [21]. The latter is a very simple probabilistic formulation which, as we shall see, combines gracefully with the popular LDA model for visual recognition.

Our work most closely builds upon two papers. Russell et al. [3] first proposed to apply the 'bags of visual words' point of view to image segments, rather than to the entire image, in the hope of discovering multiple objects in each image. Our work combines segmentation and category model learning in one step, rather than first carrying out segmentation and then categorizing the segments. While Russell et al.'s segmentation is independent for each image, in our work segmentation is carried out simultaneously and each segment's definition benefits from related segments being simultaneously discovered in other images. Conversely, Andreetto et al. [21] segment an entire collection of images simultaneously, while discovering the correspondence between homologous segments. However, the characteristics that pair segments are restricted to size, shape and average color. Associating bags of visual words to each segment allows us to discover more interesting visual connections between corresponding segments, and thus discover visual categories.

Objects are often of complex appearance, therefore, one cannot hope to always be able to segment images into complete objects. Instead, we follow the approach of Todorovic and Ahuja [22] that proposed to first segment images into object parts with consistent appearance (e.g., eye, eyebrow, nose, etc.) and then build object models by learning the geometric arrangement of parts. Since in this work the segmentation of each image was performed independently, the detected parts of the same object in two different images could lack consistency. For example, in one image one could get the nose as a segment while in another image the nose could be joined into a single segment together with the cheeks. Our goal is to simultaneously segment and learn categorical object parts, therefore, obtaining consistent image segmentations of entire collections (see Fig. 1).

In Section 2 we introduce our probabilistic model. In Section 3 we describe how to carry out inference. Section 4 presents a number of experiments. Our concluding remarks are in Section 5.

2 Learning categorical segments

Our goal is to learn categorical segments in a collection of images. We achieve that by simultaneously segmenting images and discovering correspondences between segments appearing in different images. To this end, we propose a model for image formation (see Fig. 2), which can be viewed as a probabilistic formulation of the model of Russell et al. [3], but here segmentation and recognition happen simultaneously. It can also be seen as an extension of the image segmentation model proposed by Andreetto et al. [21] where image segments are represented by a distribution of visual words, on top of local appearance.

The model represents a collection of M images I_m. An image is represented by N regularly spaced sample pixels (e.g. one sample per pixel). At each sample pixel n we measure a feature vector x_n, for example the pixel's position and RGB values, i.e., $x_n = [row, column, R, G, B]$. We further extract a fixed size 'visual word' w_n centered at pixel n. In our implementation, visual words are represented as vector-quantized filter-responces as in [23], but other discrete representations can be used as well.

Each image is formed of K regions (segments) whose statistics are shared across images. Each segment k in image m has a probability distribution $f_{k,m}$ of feature vector values x_n, and a probability distribution ϕ_k of the visual words w_n. Note, that the distributions $f_{k,m}$ of feature vectors are independent for each image while the distributions of visual words ϕ_k are shared across images. This is because we assume that the appearance of an object part, which is captured by the visual words distribution, is similar in all images, while the position and colors of the part in a particular image are independent of the other images. For example, a car can be of many different colors and appear in various image locations, however, its overall appearance, as described by the visual words, is the same in all images. We model the RGB distributions $f_{k,m}$ with the hybrid non-parametric model proposed by Andreetto et al. [21], while for the ϕ_k we use an LDA model, as proposed by Fei-Fei et al. [12] and Sivic et al. [24]. In other words, if we switch off the x_n component, the model reverts to an LDA one, while, if we switch off the w_n component, the model reverts to model of Andreetto at al. [21]

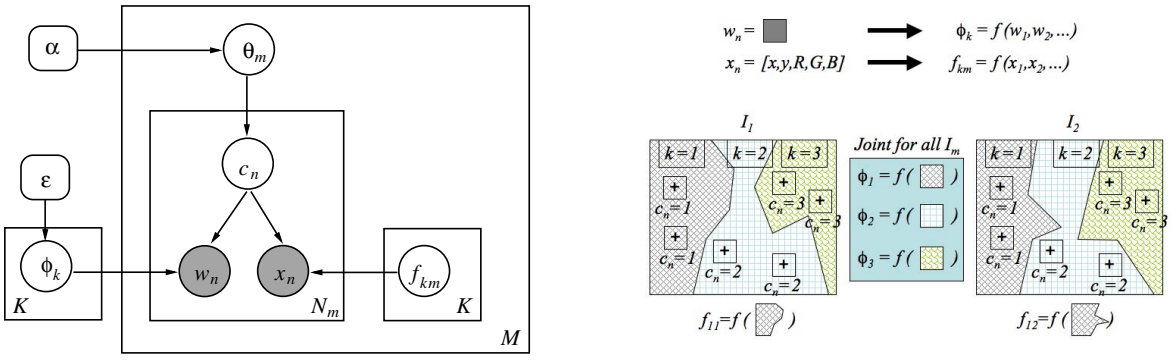

Figure 2: *Left*: Basic generative model for cluster data. The two gray nodes x_n and w_n represent the only two observed quantities in the model: the feature vectors (position and color) and the visual words associated with each pixel. The nodes c_n, $f_{k,m}$, ϕ_k, and θ_m are hidden quantities that represent the segment assignment for x_n and w_n, the probability density of the feature vectors in segment k of image I_m, the visual word distribution for segment k, and the size of the segments in image I_m, respectively. Finally the two squares with rounded corners α and ϵ represent the hyperparameters for the Dirichlet distributions over θ_m and ϕ_k respectively. *Right*: Visualization of image representation. An image is represented by a collection of N patches with their corresponding positions and colors x_n and visual words w_n. The densities f_{11} and f_{12} represent the shape and appearence of two corresponding segment in image 1 and 2. The distribution ϕ_1 represents the visual words statistics for topic 1 for all the images.

In this model, visual words are grouped by segments. This enables learning topics that are related to object parts rather than to whole scenes, as is the case in the standard "bag of words" representation over the entire image. A key aspect of the proposed model is that the densities $f_{k,m}$ allow grouping into a single image segment all the visual words generated from the corresponding topic distribution ϕ_k. Moreover, depending on the characteristics of the densities $f_{k,m}$ it is possible to enforce different constraints. For example, assuming a Gaussian distribution over the pixel positions in the image, as in Sudderth et al.[25], results in a spatial grouping of visual word generated from the topic ϕ_k.

Our model assumes that the feature vectors x_n and the visual words w_n associated with pixel are independent given the hidden variables c_n, the topic assignment for the pixel. This assumption is not necessarily correct since in practice both may depend on the local pattern of intensities.

2.1 Modeling segment shape and appearance

The densities $f_{k,m}$ are the glue that binds together the elements within the segments in the image collection. The quality of these models will help group visual words belonging to the same topic. However, the segments in an image can have complex shapes and are not easily modeled by a Gaussian or similar density.

A probabilistic model for representing image segments that does not constrain the segment to have a particular shape or appearance is the one proposed by Andreetto et al. [21]. In this model the densities are estimated in a non parametric way [26]. Given a kernel function $\mathcal{K}(u, w)$ and a set of N_k feature vectors $x_1, x_2, ..., x_{N_k}$ drawn from the density $f_{k,m}$, it is possible to obtain an estimate of the true density as:

$$\hat{f}_k(x) = \frac{1}{N_k} \sum_{j=1}^{N_k} \mathcal{K}(x, x_j) \tag{1}$$

A typical choice for the kernel function between two points is the Gaussian kernel:

$$\mathcal{K}_{\sigma_j}(x, x_j) = \frac{1}{2\pi\sigma_j^{2\,D/2}} \exp\left(-\frac{||x - x_j||^2}{2\sigma_j^2}\right)$$

where the scale parameter σ_j may be set according to local analysis as suggested in [27]. Given two feature vectors x_i and x_j, the kernel function $\mathcal{K}(x_i, x_j)$ measures the "affinity" between two points, i.e. how likely it is that the two vectors to be in the same segment. These affinities are the same as those used in other segmentation methods like spectral clustering [20] and mean-shift [28].

2.2 Modeling topics of visual words

Following the LDA model [14], each visual word w_n in a given segment k is assumed to be sampled from a topic/segment multinomial distribution with parameters ϕ_k. All the ϕ_k multinomial coefficients are sampled from the same prior distribution: a symmetric Dirichlet distribution

with scalar parameter ϵ:

$$\phi_k \sim \text{Dir}(\epsilon) \qquad (2)$$
$$w_n|\phi_k \sim \text{Multinomial}(\phi_k)$$

The K topic/segment distributions are not image specific like the densities $f_{k,m}$, but are rather shared within the entire collection. This allows coupling segment statistics across different images based on the distribution of visual words that they contain.

2.3 Sharing information on shape and appearance

The model we presented so far assumes the densities $f_{k,m}$ are independent for each image. Sometimes, however, the shape and appearance of segments are also consistent across images. For example, faces have similar shapes in all images. In these cases, it can be useful to share some information between the densities $f_{k,m}$. To share information on shape and colors we adopt the hybrid semi-parametric density representation proposed in [21]. This models densities as a weighted sum of a Gaussian term and a non-parametric term. We keep the center of the Gaussian and the non-parametric term independent for each image. Only the covariance of the shape densities is then shared across images. This is because the actual position of the segment is independent in each image, while the shape (captured by the covariance) is shared. We can further decide that a certain number of segments are shared while the rest are independent. This corresponds to data where only part of the content is shared across images. For example, in a collection of face images the background of each image can be different and independent from all other images.

3 Inference

We denote by boldface letter vectors of all values, e.g., $\mathbf{c} = [c_1, \ldots, c_N]$. To estimate the posterior distribution $p(\mathbf{c}|\mathbf{x}, \mathbf{w})$ we use a Gibbs sampling inference algorithm. Let $p(c_i|\mathbf{c}_{-i}, \mathbf{x}, \mathbf{w})$ be the posterior distribution of the hidden class label c_i of the i'th pixel given the class labels $\mathbf{c}_{-i}$ of all pixels but i, all feature vectors $\mathbf{x}$ and all visual words $\mathbf{w}$. This yields:

$$p(c_i = k|\mathbf{c}_{-i}, \mathbf{x}, \mathbf{w}) \propto$$
$$p(x_i, w_i|c_i = k, \mathbf{x}_{-i}, \mathbf{w}_{-i}, \mathbf{c}_{-i}) p(c_i|\mathbf{c}_{-i}). \qquad (3)$$

The feature vectors x_i and visual words w_i are assumed to be independent given c_i. We can, therefore, decompose the likelihood term as the product:

$$p(x_i, w_i|c_i = k, \mathbf{x}_{-i}, \mathbf{w}_{-i}, \mathbf{c}_i) =$$
$$p(x_i|c_i = k, \mathbf{x}_{-i}, \mathbf{c}_{-i}) p(w_i|c_i = k, \mathbf{w}_{-i}, \mathbf{c}_{-i}). \quad (4)$$

The first term of of Eq. 4 is the likelihood of the feature vector x_i to be in the k-th segment. Using the non-parametric approximation of Eq. 1 this term can be approximated as:

$$p(x_i|c_i = k, \mathbf{x}_{-i}, \mathbf{c}_{-i}) = \frac{1}{N_k} \sum_{j \in S_k} \mathcal{K}(x_i, x_j) \qquad (5)$$

where the kernel values $\mathcal{K}(x_i, x_j) = A_{ij}$ represent the affinity between x_i, and x_j[1], S_k is the set of feature vectors in segment k, excluding the vector i, and N_k the cardinality of segment S_k.

The second term of Eq. 4 is the likelihood of the visual word w_i to belong to the topic distribution ϕ_k. Given the conjugate prior over ϕ_k (see Eq. 3) we obtain:

$$p(w_i, |c_i = k, \mathbf{w}_{-i}, \mathbf{c}_{-i}) = \frac{N_{w_i,k} + \epsilon}{N_k + \epsilon V}, \qquad (6)$$

where $N_{w_i,k}$ is the number of pixels with visual word w_i assigned to segment k in all the images of the collection, and ϵ is the hyperparameter of the Dirichlet prior over the topic distributions ϕ_k's.

Similarly, the prior term of Eq. 3 can be written as:

$$p(c_i = k|\mathbf{c}_{-i}) = \frac{N_k + \alpha_k}{(N_m - 1) + \sum_k \alpha_k}, \qquad (7)$$

where N_k if the cardinality of segment S_k in image m, N_m is the number of pixels in image m and α_k are the hyperparameters of the Dirichlet prior over θ_m.

Combining Eq.5, Eq. 6, and Eq. 7 we obtain the following expression for the conditional probabilities used by the Gibbs Sampling algorithm:

$$p(c_i = k|\mathbf{x}, \mathbf{w}, \mathbf{c}_{-i}) \propto \qquad (8)$$
$$\frac{1}{N_k} \sum_{j \in S_k} K(x_i, x_j) \frac{N_{w_i,k} + \epsilon}{N_k + \epsilon V} \frac{N_k + \alpha_k}{(N_m - 1) + \sum_k \alpha_k}.$$

All the quantities in Eq. 8 can either be precomputed, like the affinities $K(x_i, x_j) = A_{ij}$, or updated very efficiently.

Given the samples from $p(\mathbf{c}|\mathbf{x}, \mathbf{w})$ by Gibbs sampling, it is possible to assign each pixel to a segment using the MAP estimator. The segment distributions $f_{k,m}$ and the topic distributions ϕ_k can be estimated given the assignment, (see Sections 2.1,2.2).

4 Empirical evaluation

In all our experiments image dimensions ranged between 100×100 to 240×320 depending on the database. Following the approach of Fei-Fei et al. [12] we extract the

[1] The A_{ij} are the entries of the affinity matrix used by the Normalized Cut segmentation algorithm. They can be precomputed before the inference step.

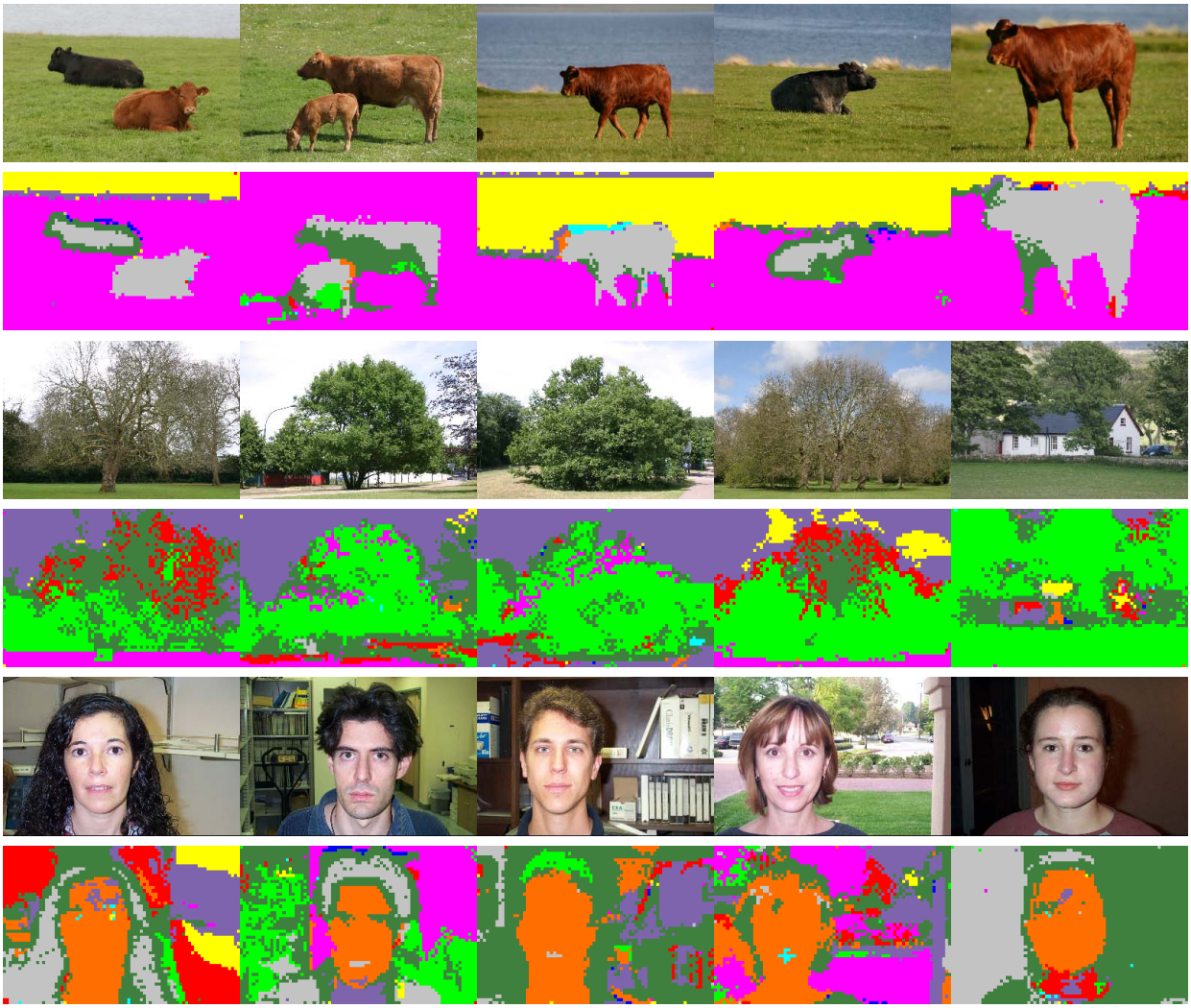

Figure 3: Segmentation-recognition results on the MSRC data-set of [29]. Each topic is marked by a different color. Our system automatically segmented image of the same category consistently. Grass is marked by magenta, sky with yellow and purple, trees with green, etc.

visual words sampling with a dense grid of 5 pixels step. At each location of the grid a patch descriptor is computed by considering the response of a filter bank [23]. This descriptor has dimension 17. A subset of the extracted descriptors is used to construct a visual dictionary using K-means (see Sivic et al. [24]). In all our experiments we use a dictionary of 256 words.

In all our experiments the shape and appearance densities $f_{k,m}$ are independent for each image while topics of visual words are shared. We use the intervening countours method [31] to compute the affinities between pixels.

We tested our system on three public databases: MSRC [29] database, the LabelMe [32] subset used by [3] and the scene database of [30]. Note, that we do not use any supervision. Figures 3,4 and 5 present results of learning categorical segments with $K = 10$. Our system discovers shared segments such as grass, faces, and sky. Note, that methods such as [3] that first segment images and then classify segments are error prone due to inconsistent segmentation. For example, in [3] cars are often merged with the road and trees are merged with skies. Our method learns shared categorical segments and therefore provides consistent segmentation of entire collections.

When using 20 topics, the running time of our system on the MSRC dataset (240 images) is about 4350 seconds on a Pentium 4 CPU (18 second per image).

For quantitative evaluation we computed the success rate of *fully unsupervised* classification on the scene and MSRC

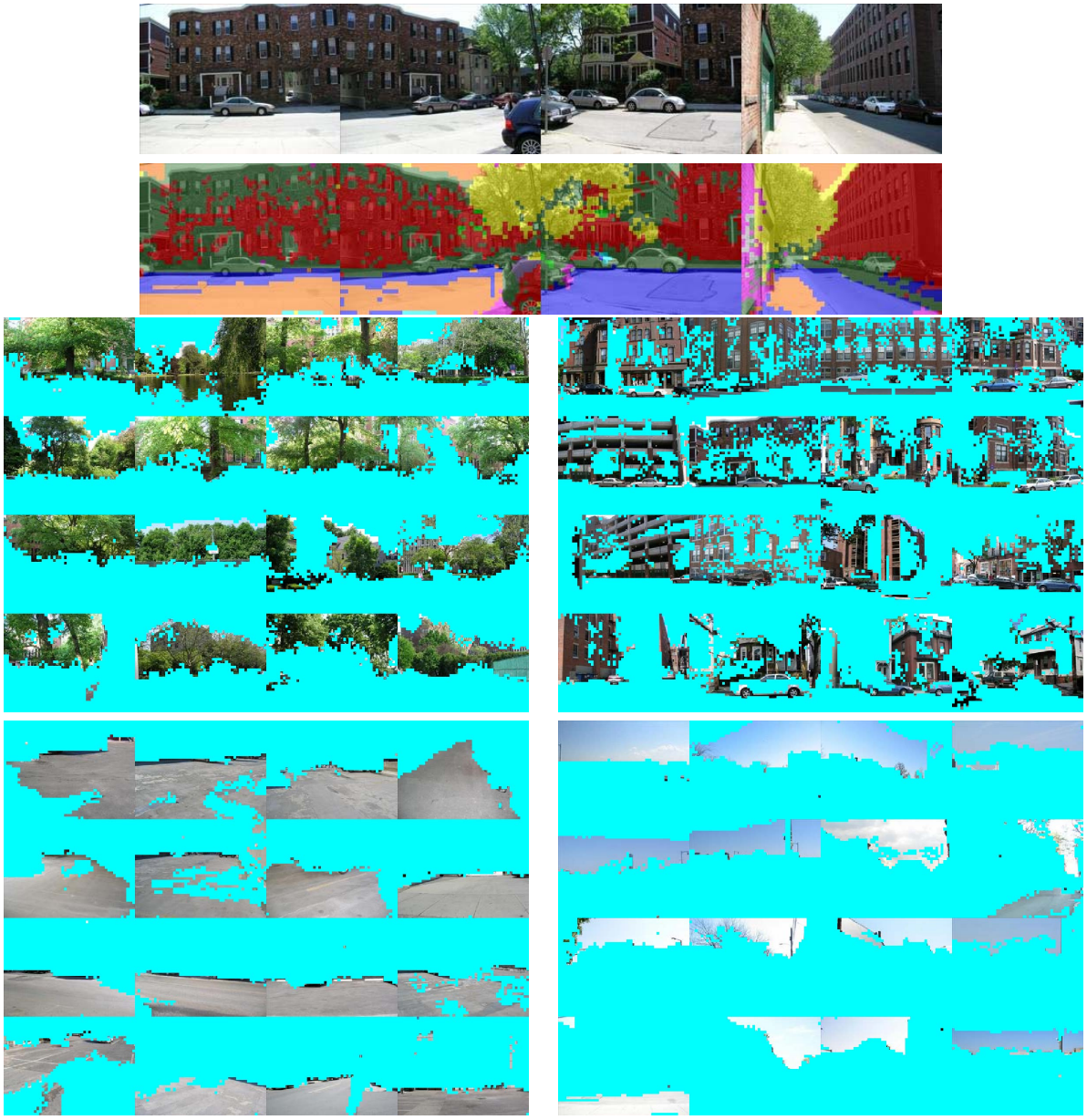

Figure 4: Top two rows: sample images and corresponding segmentation. Bottom: Four sample topics/segments learned from the LabelMe database. Each images contains 16 segments from a specific topic. The four topics represent four different elements of a possible street scene: "tree/foliage", "buildings", "street pavement", and "sky", The segmented images, as well as the topics show the consistency we obtain across images in the collection.

databases. Each image was represented by its distribution over topics. We then applied the approach of [27] to cluster the images and adopted the best bipartite graph match to find the optimal correspondence between the obtained clusters and the ground truth ones. The resulting overall classification precision was $\sim 38\%$ for the scene database and $\sim 48\%$ for the MSRC database. While lower than the reported values in [4, 12], note, that there 100 images were used for training, while our approach is *fully unsupervised*. Furthermore, it was shown in [12] that reducing the number of training images to 5 dropped the precision to below 30%. Our approach significantly outperforms this result.

Figure 1 shows what can be obtained by our system when both topics and shape are shared across images. The collection included 30 images of faces from the Caltech-4 database. The images were taken with varying backgrounds

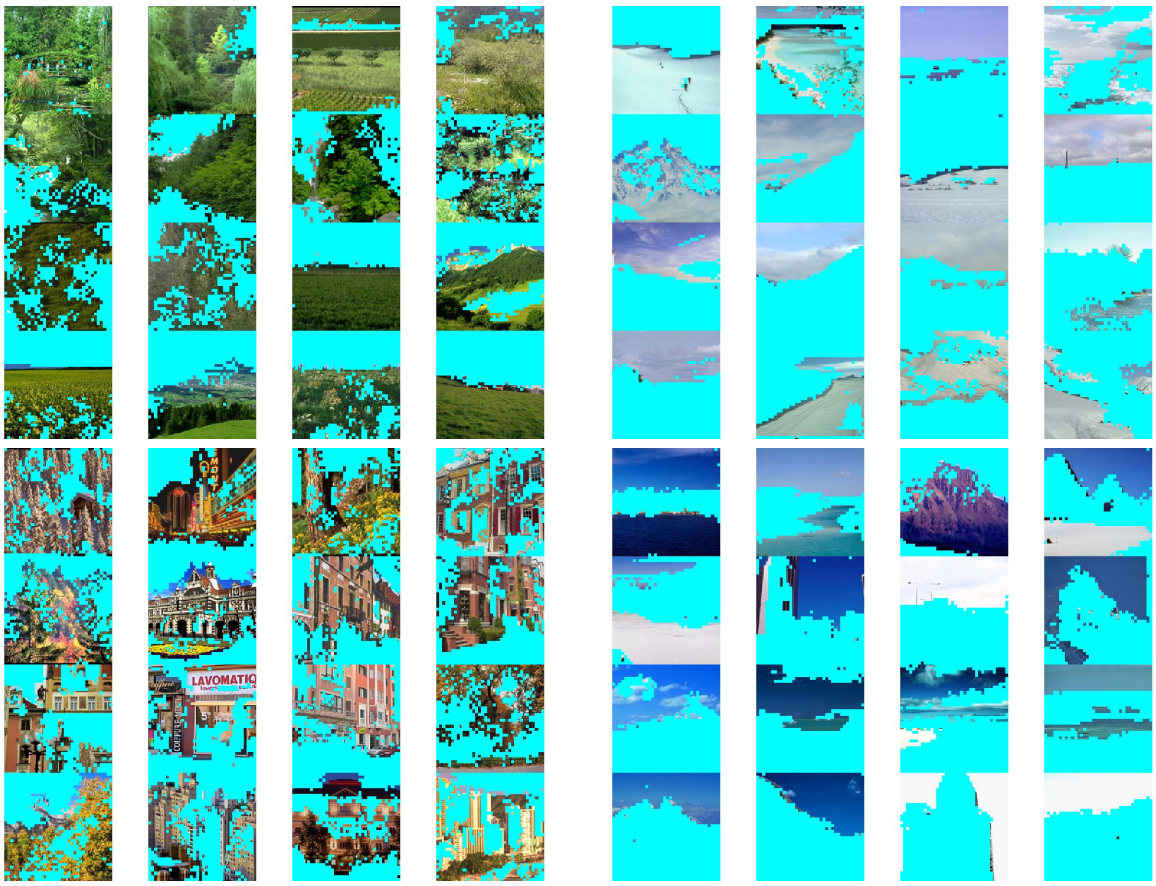

Figure 5: Four sample topics/segments learned from the Scene database [30]. The variance within each category here is high therefore detection of categorical segments is challenging. Our visual words representation incorporates color information, therefore skies were assigned to two topics, light blue and dark blue.

thus the topics corresponding to the background are varied in shape and appearance, while the faces are consistent across images. We show segmentation results where topics corresponding to the background are colored in various shades of blue, while topics corresponding to the face are colored in red and yellow. Since both topics of visual words and shape and appearance are shared across images in this case, the "red" and "yellow" topics have consistently similar shapes in all images. See Fig. 1.

5 Conclusions

We proposed a simple probabilisitic approach to segment and recognize simultaneously consistent object parts. Our experiments are the first to obtain simultaneously segmentation and categorization without supervision in a consistent one-step process.

Our system differs from previous work, which either cascaded or interleaved segmentation and recognition, while we integrate them into a single process.

Our results should be seen as a proof of principle. More informative features, such as texture, stereoscopic disparity and motion flow could be added. The consistent probabilistic model lends itself easily to semi-supervised and supervised learning as well.

Acknowledgments

Funding for this research was provided by ONR-MURI Grant N00014-06-1-0734. Lihi Zelnik-Manor is supported by FP7-IRG grant 2009783.

References

[1] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Henry Holt and Co., Inc., New York, NY, USA, 1982.

[2] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 7–27, 2001.

[3] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[4] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent object segmentation and classification," in *International Conference on Computer Vision (ICCV)*, 2007.

[5] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *ECCV'04 Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic, May 2004, pp. 17–32.

[6] E. Borenstein and S. Ullman, "Class-specific, top-down segmentation," in *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part II*, London, UK, 2002, pp. 109–124.

[7] M. C. Burl and P. Perona, "Recognition of planar object classes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '96)*, Washington, DC, USA, 1996, p. 223.

[8] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part I*, London, UK, 2000, pp. 18–32.

[9] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," *Proc of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[12] Fei-Fei Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR '05: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, Washington, DC, USA, 2005, pp. 524–531.

[13] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vision*, vol. 43, no. 1, pp. 29–44, 2001.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[15] M. Vidal-Naquet and S. Ullman, "Object recognition with informative features and linear classification," in *ICCV*, 2003, pp. 281–288.

[16] Gy. Dorkó and C. Schmid, "Selection of scale-invariant parts for object class recognition," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2003, p. 634.

[17] X. Wang and E. Grimson, "Spatial latent dirichlet allocation," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2007.

[18] Z. Tu and S.-C. Zhu, "Image segmentation by data-driven markov chain monte carlo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 657–673, 2002.

[19] P. Orbanz and J. M. Buhmann, "Nonparametric bayesian image segmentation," *International Journal of Computer Vision*, 2007.

[20] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[21] M. Andreetto, L. Zelnik-Manor, and P. Perona, "Non-parametric probabilistic image segmentation," in *International Conference on Computer Vision (ICCV)*, 2007.

[22] S. Todorovic and N. Ahuja, "Extracting subimages of an unknown category from a set of images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06) - Volume 1*, New York, NY, USA, 2006, pp. 927–934.

[23] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, Beijing, China, 2005.

[24] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *International Conference on Computer Vision (ICCV)*, 2005.

[25] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Learning hierarchical models of scenes, objects, and parts," in *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, Beijing, China, 2005, pp. 1331–1338.

[26] L. Wasserman, *All of Nonparametric Statistics*, Springer, 2006.

[27] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems (NIPS)*, 2005, pp. 1601–1608.

[28] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[29] A. Criminisi, "Microsoft research cambridge object recognition image database, version 1.0," 2004.

[30] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, , no. 42, 2001.

[31] Timothee Cour, Florence Benezit, and Jianbo Shi, "Spectral segmentation with multiscale graph decomposition," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, Washington, DC, USA, 2005, pp. 1124–1131, IEEE Computer Society.

[32] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," 2005.