

Non-Parametric Probabilistic Image Segmentation

Marco Andreetto Lihi Zelnik-Manor Pietro Perona
Department of Electrical Engineering
California Institute of Technology
Pasadena, CA, 91125, USA
{marco, lihi, perona}@caltech.edu

Abstract

We propose a simple probabilistic generative model for image segmentation. Like other probabilistic algorithms (such as EM on a Mixture of Gaussians) the proposed model is principled, provides both hard and probabilistic cluster assignments, as well as the ability to naturally incorporate prior knowledge. While previous probabilistic approaches are restricted to parametric models of clusters (e.g., Gaussians) we eliminate this limitation. The suggested approach does not make heavy assumptions on the shape of the clusters and can thus handle complex structures. Our experiments show that the suggested approach outperforms previous work on a variety of image segmentation tasks.

1. Introduction

Image segmentation is a hard task and unsupervised segmentation often results in unsatisfactory results. Incorporating higher level information on the image content can significantly improve the segmentation quality. It has been shown (e.g., [10] [2]) that when one knows which objects are present in an image, significantly improved segmentation results can be obtained. These methods, however, are highly specific for segmenting familiar objects in images, and do not generalize to other types of knowledge. For example, when segmenting a video sequence, one knows that consecutive frames are highly correlated even when their content is unknown. Image collections often share repeated parts, e.g., beach pictures will all include water, sand, and sky, but the shared parts are not always known a-priori. We thus seek a general segmentation framework that allows natural incorporation of such information to boost segmentation quality.

To date, numerous clustering algorithms have been proposed. We categorize them into three broad classes. The first consists of non-probabilistic algorithms, such as k-means [1], mean-shift [3], and agglomerative methods [4] that are based on some useful heuristics. When the heuristics match the data they perform well. For example,

k-means provides good results when the data is blob-like and the agglomerative approach succeeds when clusters are dense and there is no noise. However, more complex data structures such as those usually observable in typical images generally result in the failure of these methods.

The second class consists of probabilistic algorithms, that estimate explicit parametric models of the data, such as expectation maximization for fitting Gaussian mixture models (GMM-EM) [13]. The probabilistic methods are principled, however, when the data is arranged in complex and unknown shapes, as is the case for images, they tend to fail.

Complex data are handled well by the third class of algorithms (which has made them the state-of-the-art in image segmentation) that consists of the many variants of spectral factorization [8, 14, 17, 15, 21, 12]. These algorithms do not make strong assumptions on the shape of clusters, and thus generally perform better on images. Unfortunately, spectral factorization lacks a full probabilistic interpretation, which makes incorporation of high-level knowledge quite difficult, if not impossible.

Here we propose a general probabilistic framework that allows a straightforward incorporation of various types of prior information, such as those described above. The proposed method can handle complex data structures and does not require restricting assumptions on cluster distributions. Furthermore, we show that our approach is a generalization of GMM-EM that eliminates the constraint of using purely parametric models, while maintaining all the advantages of a full probabilistic formulation.

We start by presenting a probabilistic model for clustering in Section 2. We then show, in Section 2.1, how parametric, non-parametric and semi-parametric models can be used to represent cluster distributions. In Section 2.3 we suggest a sampling-based technique to infer the underlying cluster assignment. Finally, empirical evaluation on image segmentation with varying levels of prior knowledge is provided in Section 3.

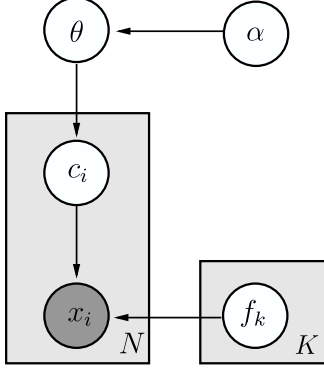


Figure 1. Basic generative model for cluster data.

2. A Probabilistic Model for Clustering

Let x_1, \dots, x_N be a set of observations in \mathbb{R}^D generated from K independent processes $\{C_1, \dots, C_K\}$. Each process C_k is described by a density function $f_k(x)$. An observation x_i is generated in the following way:

- Select one of the K processes C_k by sampling the hidden variable c_i according to some multinomial distribution $p(k)$ with parameters θ .
- Draw a point x_i according to the process-specific probability density function $f_k(x)$.

Each observation is sampled independently. Figure 1 presents the plate diagram of our model.

Let \mathbf{c}_{-i} and \mathbf{x}_{-i} be the set of labels and observations excluding point i . We assume that given cluster assignments \mathbf{c}_{-i} for all but the i th point, it is possible to obtain estimates $\hat{f}_k(x)$ for the cluster distributions $f_k(x)$, $k = 1, \dots, K$. We can then compute the conditional probability that a new point x originated from each of the K processes:

$$\begin{aligned}
 p(c_i = k \mid \mathbf{c}_{-i}, \mathbf{x}) & \\
 & \propto p(x_i | c_i = k, \mathbf{x}_{-i}, \mathbf{c}_{-i}) p(c_i = k | \mathbf{c}_{-i}, \mathbf{x}_{-i}) \\
 & = p(x_i | c_i = k, \mathbf{x}_{-i}, \mathbf{c}_{-i}) p(c_i = k | \mathbf{c}_{-i}) \\
 & \approx \hat{f}_k(x) p(c_i = k | \mathbf{c}_{-i}), \tag{1}
 \end{aligned}$$

The prior on the within-cluster distribution $p(c_i = k | \mathbf{c}_{-i}, \mathbf{x}_{-i})$ was reduced to $p(c_i = k | \mathbf{c}_{-i})$ since, according to the model, if x_i is unknown then the cluster assignment c_i depends only on the cluster assignment \mathbf{c}_{-i} of the rest of the data. So far we have not made any assumptions about the global structure of the clusters. Our only assumption is that the data each process $f_k(x)$ generates tends to cluster.

In the following sections we describe how the cluster distributions $f_k(x)$ and prior $p(c_i = k | \mathbf{c}_{-i})$ are modeled.

2.1. Modeling Cluster Distributions

2.1.1 A Non-Parametric Model

When no information is available on the shape of clusters we avoid parametric distributions (which would impose a possibly incorrect shape for the cluster) and adopt a non-parametric approach to estimate the densities $f_k(x)$. This is equivalent to placing a little probability ‘‘bump’’ around each data point and approximating the cluster distribution as the (normalized) ‘‘sum’’ of the ‘‘bumps’’.

Given a kernel function $\mathcal{K}(x_i, x_j)$ [19] which measures the affinity between points x_i and x_j (i.e., how much we believe the two points originated from the same process when all we know is their coordinates x_i and x_j) and a set of N_k points drawn from the unknown distribution $f_k(x)$, a non-parametric density estimator for $f_k(x)$ is defined as:

$$\hat{f}_k(x) = \frac{1}{N_k} \sum_{j=1}^{N_k} \mathcal{K}(x, x_j) \tag{2}$$

The kernel is chosen according to what is known on the data. A typical choice would be a Gaussian kernel

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

where σ_j may be set according to local analysis as suggested in [22]. Other kernels can be used as well. For example, in intensity based image segmentation we may wish to set to zero the connectivity between far away pixels to enforce a (strong) locality of the segmentation and to obtain a sparse problem. The kernel in this case will be a product of a Gaussian kernel and two ‘‘box kernels’’:

$$\mathcal{K}(x, x_j) = \mathcal{K}_L(r, r_j) \mathcal{K}_L(s, s_j) \mathcal{K}_{\sigma_j}(l, l_j) \tag{3}$$

where r_j, s_j are the image coordinates of the j ’th pixel and l_j is its intensity. The box kernel is defined as: $\mathcal{K}_L(r, r_j) = \frac{I((y-y_j)/2L)}{2L}$ and $I(a) = 1$ for $|a| \leq 1$ and 0 otherwise. L is the radius of the box kernel and \mathcal{K}_{σ_j} is as defined above.

2.1.2 Parametric Clusters

When it is known a-priori that part (or all) of the clusters are distributed according to some parametric form one should incorporate this information. This is easily done within the proposed framework by using parametric models for the cluster densities $f_k(x)$. For example, when it is believed the data generated by one cluster is ‘‘lumpy’’, it can be described by a Gaussian distribution: $\hat{f}_k(x) = \mathcal{N}(\mu, \Sigma)$. Random outlier points can be represented as a cluster with uniform distribution: $\hat{f}_k(x) = \frac{1}{V}$, where V is the volume of the data bounding box. Since generally the model parameters

are unknown, we will estimate them simultaneously with the assignment. Note, that in the suggested framework the densities of different clusters are independent, thus different models can be used for each (i.e., we can have a mixture of parametric and non-parametric clusters and a variety of parametric models).

2.1.3 Hybrid Clusters

The use of parametric models can lead to good segmentation results, which explains the popularity of methods like GMM-EM; however, in many cases such assumptions are too strong. For example, spectral clustering (which does not assume any structure) outperforms the parametric methods in most image segmentation tasks. In many cases assuming a specific parametric model is too restrictive. For example, a clump can be well represented *globally* by a Gaussian density; yet, this description could be too crude and inaccurate *locally*, e.g., if it has a jagged boundary.

We thus suggest using a hybrid representation which combines a parametric and a non-parametric component. The parametric component captures the known underlying global structure, while the non-parametric component captures the local deviation from it. The simplest such representation is a convex combination:

$$\hat{f}_k(x) = (1 - \lambda) \frac{1}{N_k} \sum_{j=1}^{N_k} \mathcal{K}(x, x_j) + \lambda g_k(x) \quad (4)$$

where $g_k(x)$ is a parametric model, e.g., a Gaussian or a uniform, and $\lambda \in [0, 1]$ represents the relative influence between the two terms (recall that both terms are normalized and sum to 1). In all of our experiments we used $\lambda = 0.1$.

2.2. Cluster probabilities

We next suggest an approach to modeling the prior probabilities of the K processes through the distribution of the hidden variable θ . A common choice for this type of mixture model is to assume θ distributed as a Dirichlet random variable:

$$(\theta_1, \theta_2, \dots, \theta_k) \sim Dir(\alpha_1, \alpha_2, \dots, \alpha_K)$$

Under this assumptions the relative size of the Dirichlet hyper-parameters α_k 's are the suggested ratios between the cluster sizes, while their sum $\sum_k \alpha_k$ represents the level of confidence in this information. The larger $\sum_k \alpha_k$ the stronger is the belief in the prior about the size of the clusters. Setting all α_k to the same value suggests all clusters have equal probability. Prior knowledge on the distribution of clusters can easily be incorporated by setting α_k accordingly.

Following simple calculations it is possible to derive the expression for the conditional prior term in eq. (1) as:

$p(c_i = k | \mathbf{c}_{-i}) = \frac{N_k + \alpha_k}{N - 1 + \sum_k \alpha_k}$, where N_k is the size of cluster k without counting observation i , N is the total number of observations and the α_k 's are the hyperparameters of the Dirichlet distribution for θ .

Other choices for the distribution of the random variable θ are possible. Of particular interest are non-parametric priors such as the *Chinese Restaurant Process* [18], in which the number of clusters is automatically discovered.

2.3. Inference

To estimate the posterior distribution $p(c | \mathbf{x})$ we propose using a *Gibbs Sampler*. This leads to:

Algorithm

1. Initialize by assigning at random each pixel to a cluster $c_i = k$.
2. For each pixel i in the image:
 - a. Compute the unnormalized probabilities $\tilde{p}(c_i = k | \mathbf{c}_{-i}, \mathbf{x})$ for all k using Eq. (1).
 - b. Normalize: $p(c_i = k | \mathbf{c}_{-i}, \mathbf{x}) = \frac{\tilde{p}(c_i = k | \mathbf{c}_{-i}, \mathbf{x})}{\sum_k \tilde{p}(c_i = k | \mathbf{c}_{-i}, \mathbf{x})}$
 - c. Sample from the multinomial distribution $p(c_i = k | \mathbf{c}_{-i}, \mathbf{x})$ to obtain new hard cluster assignment for the label variable c_i of point x_i .
 - d. Update model parameters of f_k for parametric and hybrid cluster densities.
 - e. Move to pixel $i + 1$.
3. Repeat Step 2 until some stopping criterion (i.e., convergence) has been reached. In our experiments we stop after 1000 iterations.
4. Repeat Step 2 100 times and collect the hard cluster assignments as samples for each image pixels.
5. For each data-point its final cluster assignment is taken as the maximum a-posteriori assignment (i.e., the most frequent one) out of all the collected samples¹. Assignment probabilities are set according to label frequencies in the collected samples.

Replacing step 2 with a Maximization of the likelihood yields a ML estimate of the clustering. If one further constraints the clusters to be parametric Gaussians the algorithm is equivalent to EM for Gaussian mixture model. We have experimented with both options (i.e., sampling and maximization) and found the sampling based approach to be preferable. While likelihood maximization works well in cases where there is one global maxima, it can easily get

¹If the distributions that describe the probabilistic model are symmetric with respect to a permutation of the labels, we sometimes observed the so called *label switching problem* [5]. In this case we use k-means instead of the MAP to detect common change patterns in the label assignments.

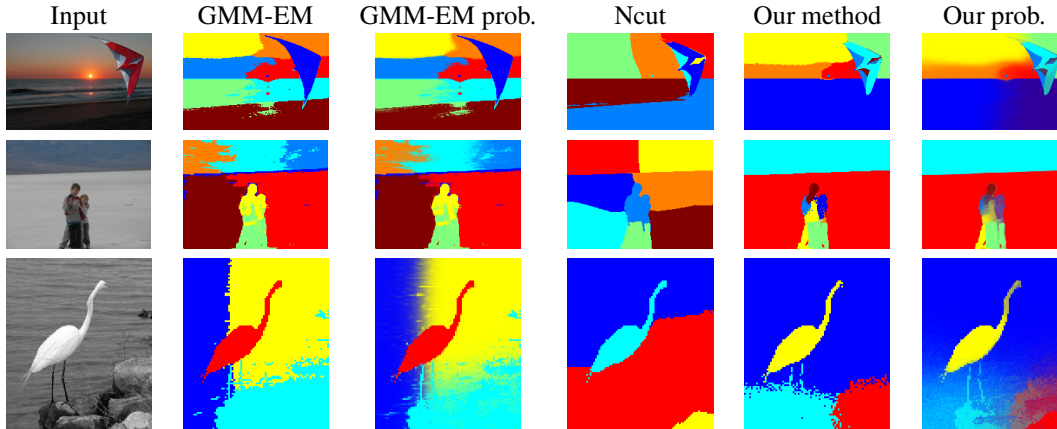


Figure 2. **Unsupervised image segmentation.** Example results from the two data sets we experimented on. Top three are part of the 16 general pictures and the bottom one is part of the 100 Egret images. Number of clusters was set to 8 general images, and 4 for the Egrets (all other results are provided in the supplemental material). The manual ratings obtained for the complete data-sets are presented in Fig. 8. The third and last columns show assignment probabilities, where the color of pixel is a linear combination of the segment markers according to cluster assignment probabilities. The probabilities obtained using non-parametric representation show smooth gradual transitions between clusters corresponding to the same image region, e.g., the sky in the top two images.

trapped in local maximum. We have found such a behavior to be highly recurrent in cases where cluster shapes are complex and their models require a non-parametric component.

3. Empirical Evaluation with Varying Levels of Prior Information

In the previous sections we proposed an unsupervised approach to clustering. We next wish to evaluate the performance of this approach, the ease at which prior knowledge can be incorporated and its benefits. This is done via a series of experiments, each utilizing different type of prior information. We start with fully unsupervised image segmentation. We then show how small amount of partial labeling can be easily incorporated and how the quality of results is boosted. We conclude with the segmentation of video frames and image collections, where information is shared across consecutive frames or similar images.

We compare our results with those of EM² on a mixture of Gaussians [1] and normalized-cuts [22] since they provide a probabilistic clustering framework and a state-of-the-art image segmentation approach, respectively. In all of the experiments affinities were computed using local scaling [22] for all affinity based methods. To evaluate segmentation quality and to compare between methods we designed a tool for manually rating results. The tool shows a human rater the original image, as well as all corresponding seg-

mentation results. The order in which the results are sorted is random for each image thus the rater cannot correlate the displayed results with a certain algorithm. Each result needs to be rated as “good”, “o.k.” or “bad”. All of the presented experiments were evaluated by 6 unbiased raters. Summary of rating statistics is provided in Fig. 8.

3.1. Unsupervised image segmentation

Experiments were performed on two image data sets: The first is a set of 100 images of Egrets [11] where only image intensities were used to compute affinities. The second is a set of 16 general color images (mostly of swimmers, since one of the authors loves swimming). Figure 2 shows a few representative image segmentation results (The results for all other images are provided in the supplemental material).

Figure 8 shows the rating statistics obtained for both data-sets. The performance of EM fitting of a mixture of Gaussians (in both coordinate space and color space) are of the lowest quality, because assuming image segments can be approximated by Gaussians is inaccurate. The results by normalized cut and our suggested approach are comparable with slight preference to our method. Our approach, as well as EM, naturally provides soft assignment of pixels to segments (see Figure 2). Such soft assignments often make more sense, e.g., in ambiguous cases where the transition between segments is gradual. Furthermore, they provide more information than hard decisions do. An attempt at obtaining soft assignments from normalized cuts was proposed in [6]. Their approach lacks, however, a complete probabilistic interpretation.

²Our experiments show that inference with EM and with Gibbs sampling on a Gaussian mixture model provide similar segmentation results. This suggests that the performance of GMM are intrinsic of the probabilistic model and not a consequence of the particular inference method used.

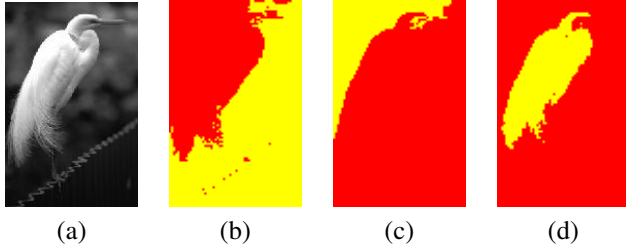


Figure 3. **Partial Labeling.** A typical result of intensity based image segmentation into 2 clusters (out of a 100 images in the Egret set of [11]). (a) Original image (b) GMM-EM clustering (c) Normalized cuts (d) Our result with partial labeling. Boundary pixels were constrained to the background cluster.



Figure 4. **Partial Labeling, comparison with GrabCut.** Left: input image. Right: Our segmentation result, obtained by manually labeling part of the image as background (similar to [16]).

3.2. Partial Labeling

While our general framework is unsupervised, some partial information on the assignment of points to clusters is often available. Such information can be provided in one of three forms: partial labeling, “must-link” constraints and “cannot-link” constraints. We next explore all three.

Partial assignment of points to clusters is equivalent to having observed the labels of some of the (usually hidden) random variables c_i of the model. Such type of constraints are thus incorporated by fixing the corresponding observed labels c_i during the inference process on the model (described in Section 2.3). This leads to a more stable solution and faster convergence. Figure 3 shows how minimal partial labeling can significantly improve image segmentation results. The segmentations obtained by our method are of higher quality than those of GMM-EM (using the same constraints). Comparison to spectral factorization is impossible since labels cannot be fixed. We thus compare our results to those of graph-cuts methods. Graph-cuts [16] are somewhat similar in spirit to spectral factorization but require significant user interaction and are thus generally of less interest to us. Figure 4 shows our approach provides comparable results to those of Rother et. al [16] when the same amount of user intervention is utilized.

Constraints which force points to reside in the same cluster (“must-link”) can be incorporated by estimating the labels of those points jointly. This corresponds to a modifi-

cation of the model of Fig. 1 where an edge (conditional dependence) is added between the constrained points. The “cannot-link” constraints can (in theory) be incorporated, in a similar manner, by estimating the labels for these points jointly while enforcing exclusion. While in our inference method this is easily achievable if the “cannot-link” constraints involve only pairs of separated points, it is difficult to consider exclusion dependencies over a larger number of points, since the number of possible assignments would grow exponentially.

Incorporating labeling constraints (of any type) is not trivial in non-probabilistic methods such as spectral clustering. Yu and Shi [20] showed how “must-link” constraints on pairs of points can be incorporated, albeit with some additional computational cost. It has not been shown how to incorporate “cannot-link” constraints or partial labeling in spectral clustering.

3.3. Temporal coherence in videos

In the previous sections we evaluated performance in the unsupervised and partially supervised cases. But other types of prior information are often available. In this section we examine segmentation of video frames. Adjacent video frames are known to be highly correlated regardless of their content. In this section we show how this can be incorporated into our segmentation framework and improve segmentation quality. A related idea was proposed by Jojic and Frey [7] who separated video frames into layered sprites. Their underlying assumption was that all layers are shared among the video frames and each layer can undergo only limited transformations such as translation and occlusion. This does not apply to general videos where the camera moves significantly, resulting in large changes in background, as well as complex motion of articulated objects, such as human bodies, which imply large changes in appearance and shape across video frames. We thus propose an approach that assumes coherence only across consecutive frames and not throughout the sequence.

Pixel-level segmentation of video sequences is a high dimensional problem, since the data-set size equals the overall number of pixels. Therefore, one has to resort to segmenting separately small portions of the video. We will assume here the video portions are individual frames. This can result in a set of independent segmentations even for consecutive frames which are highly correlated. To obtain a globally consistent segmentation one needs to enforce spatiotemporal coherence across frames. This can be done by first segmenting each frame independently and afterwards matching segments across frames. Alternatively, coherence could be enforced directly during the segmentation task. The latter is impossible for methods like spectral clustering, which do not allow incorporating prior information.

On the contrary, our framework is particularly suitable

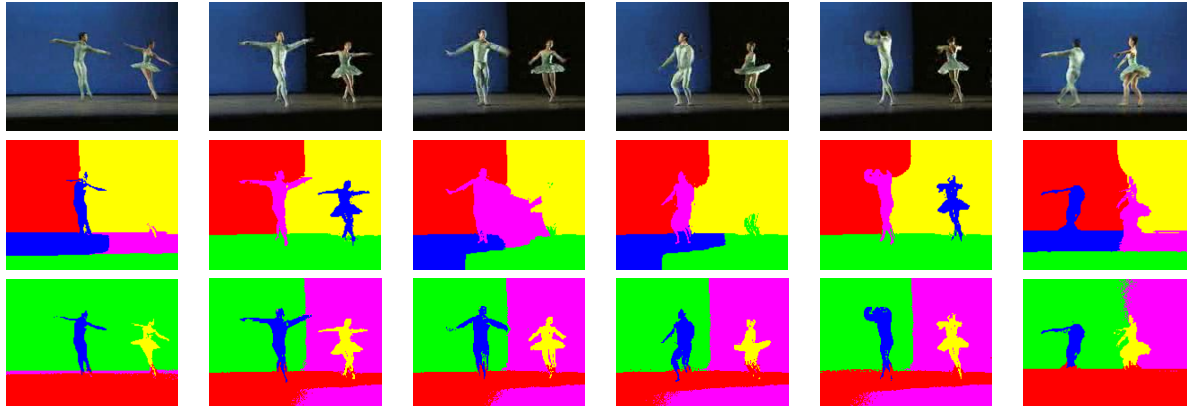


Figure 5. **Video sequence segmentation.** *Top:* Frames 218, 280, 282, 284, 286, and 329 out of a 343 frame long video. *Middle:* Ncut segmentation results. *Bottom:* Our result while enforcing spatiotemporal coherence across frames is significantly better, see also Fig. 8 for statistics of manual rating. The complete video as well as results on a different video are provided in the supplemental material.

for this purpose. We segment videos frame-by-frame while propagating information from one frame to the next. We initialize the segmentation of each frame with the segmentation result of the previous frame. Since consecutive frames are highly correlated this on its own speeds-up the computation (by reducing the number of iterations of the sampler) and promotes more consistent results. Furthermore, since our clustering provides cluster assignment probabilities for each pixel, we detect high confidence pixels and fix their labels for some iterations. This constrains the segmentation of each frame to be highly similar to that of its predecessor. We then release the labels of all pixels and collect samples. This procedure localizes slowly changing parts of the video, such as the background, and reduces the computational cost by speeding up convergence. Following is a short summary of the proposed video segmentation approach:

1. Segment the first frame of the sequence and obtain cluster assignment probabilities for each pixel.
2. For all the remaining frames $f = 2, \dots, F$:
 - a. For frame $f - 1$, compute the confidence R_i of segment assignment of the i 'th pixel as: $R_i = (p(c_i = k_v | \mathbf{x}) - p(c_i = k_w | \mathbf{x})) / p(c_i = k_v | \mathbf{x})$ where $p(c_i = k_v | \mathbf{x})$ and $p(c_i = k_w | \mathbf{x})$ are the highest and the second highest cluster assignment probabilities for pixel i .
 - b. Initialize the sampler for frame f with cluster assignment and confidence weights of frame $f - 1$.
 - c. Run the sampler for N_1 iterations while fixing the labels of the high confidence pixels, $R_i > 0.9$.
 - d. Run the sampler for further N_1 iterations with all labels free to change, and collect samples.
 - e. Set cluster assignment of frame f as MAP estimator and keep cluster assignment probabilities.

Even though this is a very simple way to impose temporal coherence algorithm it still show that higher-level information can significantly improve the quality of segmentation. Using more complex (possibly probabilistic) models for the motion of the object in the video is likely to further improve the segmentation results. Figure 5 compares the results of the proposed approach with those of normalized cuts with post-segmentation segment matching. The segmentation obtained by normalized cuts is inconsistent across frames. Our method significantly outperforms both normalized cuts and GMM-EM³ and returns video segmentations that are both of high quality and consistent across frames (i.e. the same object is consistently assigned to the same cluster, denoted by same color, throughout the whole video sequences). For sanity check, we also compared segmentation results of our method with and without temporal coherence. Using temporal coherence significantly improved the segmentation quality. Results are omitted due to lack of space. Please refer to supplemental material for the complete video sequence as well as other videos.

3.4. Part sharing in image collections

In videos, consecutive frames are known to be highly correlated. A less constrained scenario is that of segmenting images of a collection where images are known to share parts. For example, a collection can be a set of face images with varying background, see Figure 7. Although the pictures are generally different, all of them show a person's face in the foreground. We next show how utilizing this can improve segmentation quality as well as provide coherent

³The GMM-EM model we use for comparison in our experiment is closely related to the model of Khan and Shah [9], with the main difference that no information of local velocity is used in the clustering. The segmentation obtained by GMM-EM in our comparisons is consistent across frames but is of poor quality due to complex shapes of the segments.

results where the distinction between foreground and background is consistent.

One can address this problem in a number of ways. We chose a simple solution, as our goal is to show the versatility of our approach and the ease at which it can be tuned for specific tasks.

Given a collection of F images, where all images show a face (of different subjects) with varying background, our goal is to obtain foreground/background segmentation of each, as well as to identify the shared foreground (i.e., the face). The naive solution would be to segment each image independently into two parts and then match segments to identify the foreground. Instead, we simultaneously segment all the images in the collection using the hybrid-model described Sec. 2.1.3 to represent each segment. We then enforce one cluster to share information from all the images. That is, the global part of its distribution f_k is estimated by gathering information from all the pixels assigned to the cluster in all the images. For all the other segments the global term is computed using only the information of the corresponding image. The local term of each segment (shared or not) is independent for each image. Note, that we do not tell the system which part is to be shared. Instead, this emerges automatically as the most consistent part, in appearance and average position, across images. In the image collection we experimented with images share only the face, while the background varies significantly, thus the face emerges as the shared part.

Figure 7 compares results of foreground/background segmentation with and without part sharing. As expected, enforcing coherence across images results both in consistently assigning the person to the foreground as well as improving the quality of segmentation.

3.5. Noisy data

As was discussed in Section 2.1, a benefit of the proposed framework is that different models can be used for cluster densities. Figure 6 presents an example where this becomes useful. The data contains three spiral clusters and random outlier points. Clearly, fitting a mixture of Gaussians will not work on such data. Using normalized cuts to segment into three clusters discovers the dense spiral clusters but the outliers are arbitrarily assigned to the closest spiral. Switching to 4 clusters splits one of the spirals. Applying the suggested probabilistic approach with three non-parametric clusters and one parametric with a uniform distribution results in discovering the three spirals and collecting all the outliers into the uniform distribution cluster.

4. Discussion and Conclusions

We have presented a principled probabilistic approach to clustering that can handle complex cluster shapes by using a

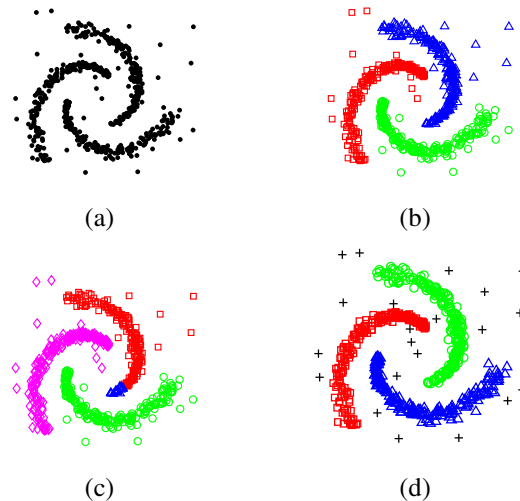


Figure 6. Noisy data. (a) Input. (b) Segmentation by normalized cuts using 3 clusters: the outliers are arbitrarily assigned to the 3 clusters. (c) Segmentation by normalized cuts into 4 clusters: even with an additional cluster the outliers are assigned to the main clusters and one of the three clusters is randomly split. (d) Our segmentation: using a parametric (uniform) distribution for one cluster results in correctly identifying the three clusters and the outliers (crosses).

non-parametric representation as well as hybrid parametric and non-parametric models. The proposed method can incorporate prior knowledge on cluster structure and assignment constraints in a principled way and provides probabilistic assignment into clusters. When no prior information is available our results are comparable to those of spectral clustering (which is considered state-of-the-art in image segmentation). However, in cases where some knowledge is available our method significantly out-performs others.

Acknowledgments

Funding for this research was provided by ONR-MURI Grant N00014-06-1-0734.

References

- [1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, England: Oxford, 1995.
- [2] E. Borenstein and S. Ullman. Learning to Segment. In *Proc. 8th European Conference on Computer Vision (ECCV)*, May 2004.
- [3] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.
- [4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.
- [5] A. Jasra, C. C. Holmes, and D. A. Stephens. MCMC Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 20(1), 2005.
- [6] R. Jin, C. Ding, and F. Kang. A probabilistic approach for optimizing spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.



Figure 7. **Image collection.** Six examples out of a collection of 30 images of 5 different people. Top: Original images. Second row: Ncut segmentation. Third row: Our approach segmenting each image independently. Bottom: Our result with part sharing provides higher quality segmentation. See also Fig. 8

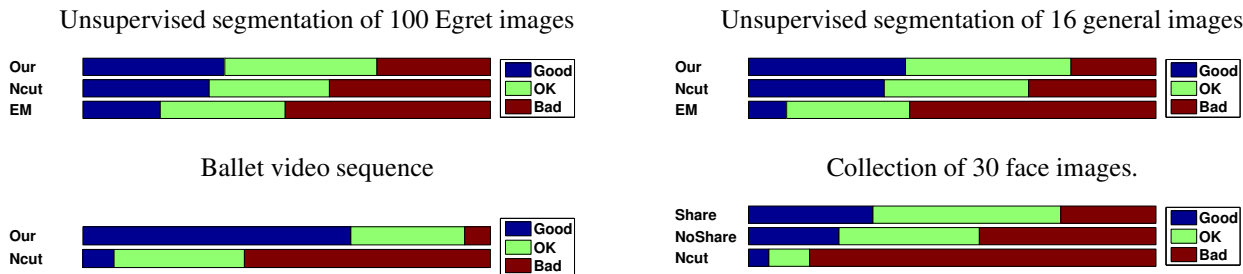


Figure 8. **Rating.** Six people rated the segmentation results of all the images in our data sets. The plots show the rating statistics of each experiment and each method. Each bar is split into three parts whose size correspond to the fraction of images assigned the corresponding label. Good results correspond to little of the red and a lot of blue. Our method outperforms other methods in all of the experiments (on the ballet video sequence we don't have results of GMM-EM since it failed to converge.)

[7] N. Jovic and B. Frey. Learning Flexible Sprites in Video Layers. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[8] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.

[9] S. Khan and M. Shah. Object Based Segmentation of Video Using Color, Motion and Spatial Information. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[10] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Obj cut. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 18–25, Washington, DC, USA, 2005. IEEE Computer Society.

[11] S. Lazebnik, C. Schmid, and J. Ponce. A Maximum Entropy Framework for Part-Based Texture and Object Recognition. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 832–838, Beijing, China, October 2005.

[12] M. Meila and J. Shi. Learning Segmentation by Random Walks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 873–879, 2000.

[13] T. K. Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, November 1996.

[14] A. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, 2001.

[15] P. Perona and W. T. Freeman. A Factorization Approach to Grouping. In *Proc. 5th European Conference on Computer Vision (ECCV)*, volume 1, pages 655–670, 1998.

[16] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[17] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[18] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes, 2003.

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

[20] S. X. Yu and J. Shi. Segmentation Given Partial Grouping Constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):173–183, 2004.

[21] R. Zass and A. Shashua. A Unifying Approach to Hard and Probabilistic Clustering. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 294–301, October 2005.

[22] L. Zelnik-Manor and P. Perona. Self-Tuning Spectral Clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1601–1608, 2005.