Unsupervised Feature Selection and Learning for Image segmentation

Mohand Saïd Allili
Université du Québec en Outaouais
Department of Computer Science and Engineering
J8X 3X7, Gatineau, Québec, Canada.

Djemel Ziou
Université de Sherbrooke
Department of Computer Science
J1K 2R1, Sherbrooke, Québec, Canada.

Nizar Bouguila
Institute for Information Systems Engineering
H3G 2W1, Concordia University, Montreal, Canada.

Sabri Boutemedjet
Université de Sherbrooke
Department of Computer Science
J1K 2R1, Sherbrooke, Québec, Canada.

Abstract

In this paper we investigate the integration of feature selection in segmentation through an unsupervised learning approach. We propose a clustering algorithm that efficiently mitigates image under/over-segmentation, by combining generalized Gaussian mixture modeling and feature selection. The algorithm is based on generalized Gaussian mixture modeling which is less prone to region number over-estimation in case of noisy and heavy-tailed image distributions. On the other hand, our feature selection mechanism allows to automatically discard uninformative features, which leads to better discrimination and localization of regions in high-dimensional spaces. Experimental results on a large database of real-world images showed us the effectiveness of the proposed approach.

Keywords: *Segmentation, feature selection, mixture of generalized Gaussian distributions.*

1. Introduction

Segmentation is an important topic in computer vision and image/video processing. In recent years, supervised and unsupervised feature-space clustering has been extensively investigated in image (video) segmentation. In this approach, image features (color, texture, motion, etc.) are assumed to have similar values within each region and different values between different regions, where segmentation amounts to dividing the feature space into different compact clusters. The clusters are then supposed to correspond to the regions in the (spatial) image (video) space. In most clustering-based segmentation algorithms, the number of regions is assumed to be known or estimated from the data. In the latter case, no universal method exists for efficient estimation of the number of regions. Rather, criteria mainly based on information theory, such as minimum description length (MDL), Akaike (AIC) or Bayesian information criterion (BIC), are used [1, 2, 13]. In most of these methods, the presence of noisy features (i.e., not discriminating enough the salient regions) or outlying data, generally, influence badly this estimation.

In what follows, we give a brief review on some recent clustering-based image segmentation methods. They are divided into two main groups: parametric and non-parametric methods. In the parametric group, we find mainly approaches based on mixture models. For instance, the authors in [4, 14] used the maximum likelihood estimation to fit a Gaussian mixture model to the image histogram, where the number of regions is estimated using the MDL principle. This approach, however, treats color and texture images in the same way by fusing several features from the image, and therefore tends to over-segment images. In the same vein, [1, 14] use active contours to capture simultaneously all the regions of an image. In the non-parametric group, we find mainly approaches based on the K-means algorithm and its variants [7], and kernel-based approaches [8]. In both of the above groups, two main assumptions are made about the data. First, most often the region data distribution are supposed to be Gaussian. Although this assumption is valid for most of images, it loses its validity if a region is not perfectly homogeneous (e.g., region not uniformly illuminated, or contaminated by noise/outliers). Recently, we proposed in [2] a new mixture model based on the generalized Gaussian distribution (GGD) (called MoGG), which has the good property of mitigating region over-fitting of images induced by features with non-Gaussian distribution. We recall that non-Gaussian data arise generally in heavy-tailed image histograms, with sharped or flat-shaped modes.

978-0-7695-4040-5/10 $26.00 © 2010 IEEE
DOI 10.1109/CRV.2010.44

Since the number of mixture components is determined automatically, such histograms can be easily over-fitted using the Gaussian mixture model (i.e., over-estimation of the number of regions), which causes over-segmentation [2].

The second assumption of these methods is about grouping features. Most of the methods combine several (color, texture) features in multi-dimensional vectors, where it is supposed that the more features we have, the better the segmentation is expected to be. It has been shown in the past that the performance of mixture modeling may be substantially deteriorated in the presence of many non-informative features [11]. In segmentation, for instance, texture features on one particular orientation may discriminate a texture pattern in an image while features on other orientations are uniform (i.e. irrelevant) and useless for this segmentation. Therefore, introducing a *feature selection* (FS) mechanism, to remove irrelevant features, is of prominent importance to estimate the real number of regions and achieve better segmentation. To our knowledge, this issue has not been considered for segmentation in the past.

In this work, we propose an efficient segmentation framework which incorporates unsupervised FS in the MoGG model. This combination aims at enhancing the representation of non-Gaussian data and eliminating irrelevant features, leading to significant reduction of over-segmentation in images. Our model performs automatically the selection of the optimal number of regions with their parameters, and the subset of relevant features, by optimizing a single objective based on the minimum message length (MML) criterion [13]. This unified objective allows for accurate discrimination and identification of the real image regions inside high-dimensional feature vectors, while penalizing mixture over-fitting.

This paper is organized as follows: In Section 2, an outline of the proposed model for segmentation is presented. In Section 3, we present the details of our learning algorithm. Section 4 shows some experiments image segmentation. We end the paper with a conclusion and some future work perspectives.

2. The proposed model

We proposed in [2] a multi-dimensional version of the GGD, where each feature distribution is allowed to have its own shape parameter. Given a D-dimensional feature space, the GGD distribution of a feature vector $\vec{x} = (x_1, ..., x_D) \in \mathbb{R}^D$ is defined as follows:

$$p(\vec{x}|\vec{\mu}, \vec{\sigma}, \vec{\lambda}) = \prod_{l=1}^{D} \frac{\lambda_l K(\lambda_l)}{2\sigma_l} \exp\left\{ -A(\lambda_l) \left| \frac{x_l - \mu_l}{\sigma_l} \right|^{\lambda_l} \right\}, \quad (1)$$

where $\vec{\mu} = (\mu_1, ..., \mu_D)$, $\vec{\sigma} = (\sigma_1, ..., \sigma_D)$ and $\vec{\lambda} = (\lambda_1, ..., \lambda_D)$ are vectors of location, scale and shape parameters. Also, we have $K(\lambda) = \frac{\left[\frac{\Gamma(3/\lambda)}{\Gamma(1/\lambda)}\right]^{\frac{1}{2}}}{\Gamma(1/\lambda)}$ and $A(\lambda) = \left[\frac{\Gamma(3/\lambda)}{\Gamma(1/\lambda)}\right]^{\frac{\lambda}{2}}$, with $\Gamma(\cdot)$ denoting the gamma function. Each parameter $\lambda_l \geq 0$, $l = 1, \ldots, D$, controls the shape of the GGD and determines whether it is peaked or flat in the lth dimension of the image feature vector [2].

Suppose that the image contains M regions. The goal of segmentation is to assign the ith pixel described by a visual feature vector $\vec{x}_i$ to one of the M regions. This assignment is defined by a label $\vec{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iM})$ such that $z_{ij} \in \{0, 1\}$ and $\sum_{j=1}^{M} z_{ij} = 1$, where $z_{ij} = 1$ if $\vec{x}_i$ belongs to region j, and 0, otherwise. Since each region is defined by a single component in the mixture model, the conditional distribution of $\vec{x}_i$ given the region label $\vec{z}_i$ is given by a MoGG, as follows [2]:

$$p(\vec{x}_i|\vec{z}_i, \vec{\Theta}) = \prod_{j=1}^{M} \left(\prod_{l=1}^{D} p(x_{il}|\vec{\theta}_{jl}) \right)^{z_{ij}}, \quad (2)$$

where $\vec{\theta}_{jl} = (\lambda_{jl}, \mu_{jl}, \sigma_{jl})$, and $p(x_{il}|\theta_{jl})$ denotes a univariate GGD component. Since the z_{ij}'s constitute the missing information, they can be identified for a given $\vec{\Theta}$ (the set of MoGG parameters) using the Bayes rule as $p(z_{ij}|\vec{x}_i, \vec{\Theta}) = p(\vec{x}_i, z_{ij}|\vec{\Theta})/p(\vec{x}_i|\vec{\Theta}) \propto p_j p(\vec{x}_i|z_{ij}, \vec{\Theta})$ with $p_j = p(z_{ij})$. Therefore, for a given number of components M, the core step of the segmentation process is to estimate $\vec{\Theta}^* = (p_j^*, \theta_{jl}^*), (j = 1, \ldots, M \text{ and } = 1, \ldots, D)$, as the optimum of a certain objective such as the likelihood of all image data [4].

Note that the MoGG defined by Eq. (2) assumes that all the features have equal importance. However, with high-dimensional descriptors such as color, texture, etc., features do not contribute equally in discriminating among the exiting regions. Indeed, some features may be uniform or unimodal and their distributions independent of the region labels [12]. These "noisy" features may, on the one hand, confuse the inference by increasing the complexity of the model (i.e., overfitting the model) [9], and, on the other hand, compromise the distinction between the real regions. Therefore, we extend the MoGG model by modeling the discrimination power of each feature separately. Let ϕ_l be a binary variable, set to 0 when the lth feature is irrelevant (i.e., uniform) and to 1, otherwise. Then, the distribution of each x_{il}, given the component label z_{ij}, can be approximated as follows [3, 9]:

$$p(x_{il}|\theta_{jl}^*, \vec{\varphi}_l^*, \phi_l) \simeq \left(p(x_{il}|\theta_{jl}) \right)^{\phi_l} \left(p(x_{il}|\vec{\varphi}_l) \right)^{1-\phi_l}, \quad (3)$$

where now $\vec{\Theta}^* = (p_j^*, \theta_{jl}^*, \vec{\varphi}*_l)$. The star superscript denotes the unknown true distribution of the lth feature of the region, and both $p(.|\theta_{jl})$ and $p(x_{il}|\varphi_l)$ are univariate GGDs. In Eq. (3), ϕ_l is a hidden variable set to 1 from

the data in every case where the lth feature is multimodal. However, this model leads to false positives (i.e., uninformative features that are identified as relevant) when a feature is defined by only overlapped components that cannot distinguish among the real image regions [3]. Therefore, we generalize the definition of feature's relevance by considering the irrelevant component $p(.|\vec{\varphi}_l)$ as a common mixture of GGDs independent of the region labels $\vec{z}_i$. This choice is also motivated by the ability of the mixture to approximate almost any arbitrary distribution of the irrelevant features. We consider K as the number of components in this common mixture, with the parameters $\varphi_{1l}, \ldots, \varphi_{Kl}$ for each feature, respectively. By supposing each ϕ_l as a Bernoulli variable with parameters $p(\phi_l = 1) = \rho_{l1}$ and $p(\phi_l = 0) = \rho_{l2}$, such that $\rho_{l1} + \rho_{l2} = 1$, it is straightforward to show that the distribution $p(\vec{x}_i|\vec{\Theta})$ is given by:

$$p(\vec{x}_i|\mathbf{\Theta}) = \sum_{j=1}^{M} p_j \prod_{l=1}^{D} \left(\rho_{l1} p(x_{il}|\theta_{jl}) + \rho_{l2} \sum_{k=1}^{K} \pi_{kl} p(x_{il}|\varphi_{kl}) \right) \tag{4}$$

where π_{kl} (with $\sum_{k=1}^{K} \pi_{kl} = 1$) denotes the prior probability that x_{il} is generated by the kth component of the common mixture, given that the lth feature is irrelevant (i.e., ϕ_l=0).

In what follows, we consider the notations $\mathbf{p} = (p_1, ..., p_M)$, $\vec{\rho}_l = (\rho_{l1}, \rho_{l2})$ and $\vec{\pi}_l = (\pi_{l1}, ..., \pi_{lK})$. Image segmentation with FS can now be achieved by optimizing a certain objective with respect to the set of all model's parameters $\vec{\Theta} = (\mathbf{p}, \theta_{jl}, \varphi_{kl}, \vec{\pi}_l)$, where $j = 1, ..., M$ and $l = 1, ..., D$. Note that both M and K are unknown parameters and need to be identified from the data.

3. Model learning

Maximum likelihood method (ML) is the most commonly used approach for the estimation of $\vec{\Theta}$ [2, 4]. The natural choice for finding ML estimates for models with missing information ($\vec{z}$,and $\vec{\Phi}$ in our case) is the expectation-maximization (EM) algorithm [5]. The log-likelihood of N independent and identically distributed feature vectors $\mathcal{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$ in the image is given by:

$$\log p(\mathcal{X}|\vec{\Theta}) = \sum_{i=1}^{N} \log \left(p(\vec{x}_i|\mathbf{\Theta}) \right) \tag{5}$$

To achieve a fully unsupervised segmentation and penalize very complex models, we follow a similar approach to [3, 9] by optimizing the message length of the data [13]. Therefore, the optimal number of components $\hat{M}$ and $\hat{K}$ and model's parameters $\hat{\vec{\Theta}}$ correspond to the MML. Furthermore, we initialize the EM algorithm with a fixed number of potential components (M and K). During EM iterations, the weights p_j, ρ_{l1} and ϕ_{kl} corresponding to unwanted components are forced to zero. By this way, our approach will be less sensitive to initialization and will penalize overfitting during estimation. The message length of the image data is given by:

$$MML \simeq -\log p(\vec{\Theta}) + \frac{1}{2}\log|I(\vec{\Theta})| + \frac{c}{2}(1 + \log\frac{1}{12}) - \log p(\mathcal{X}|\vec{\Theta}) \tag{6}$$

where $p(\vec{\Theta})$, $|I(\vec{\Theta})|$ and $p(\mathcal{X}|\vec{\Theta})$ denote the prior distribution, the Fisher information and the likelihood, respectively. The constant $c = M + 3D + KD + 3DM + 3DK$ in Eq. (6) is the total number of parameters. To calculate the MML, we assume the independence of the different groups of parameters, which factorizes both $|I(\vec{\Theta})|$ and $p(\vec{\Theta})$ over the Fisher and prior distribution of these groups. Given that the parameters $\mathbf{p}$, $\vec{\rho}_l$ and $\vec{\pi}_l$ are defined on the simplexes $\{(p_1, \ldots, p_M) : \sum_{j=1}^{M-1} p_j < 1\}$, $\{(\rho_{l1}, \rho_{l2}) : \rho_{l1} < 1\}$, and $\{(\pi_{l1}, \ldots, \pi_{lK}) : \sum_{k=1}^{K-1} \pi_{lk} < 1\}$, respectively, a natural choice as a conjugate prior for these vectors is the Dirichlet distribution with hyper-parameters set to 0.5 (i.e., uninformative Jeffrey's prior [5]). These distributions are given by:

$$p(\mathbf{p}) \propto \frac{1}{\prod_{j=1}^{M} p_j^{1/2}}, \quad p(\vec{\rho}_l) \propto \frac{1}{\rho_{l1}^{1/2}\rho_{l2}^{1/2}}, \quad p(\vec{\pi}_l) \propto \frac{1}{\prod_{k=1}^{K} \pi_{kl}^{1/2}} \tag{7}$$

We summarize the algorithm for the minimization of the MML objective with the constraints $0 < p_j \leq 1$, $0 < \rho_{l1} \leq 1$, $0 < \pi_{kl} \leq 1$ and $\sum_{j=1}^{M} p_j = 1$, $\sum_{k=1}^{K} \pi_{kl} = 1$. This algorithm takes as input an image feature set $\mathcal{X}$ and the numbers M_{min} (resp. K_{min}) and M_{max} (resp. K_{max}), which denote the lower and upper bounds of M (resp. K). We optimize the objective MML using the EM algorithm. Note that the convergence rate of the EM algorithm in our case is quadratic. Finally, to initialize the EM we use the Fuzzy C-means algorithm. The following script summarizes the main steps of model selection for segmentation.

4. Experimental results

To show the performance of our approach, we conducted experiments on several examples of image segmentation. Real-world images contain in general regions with different content of color and texture. Therefore, to have a good representation of regions, it is preferable to use a feature vector $\vec{x}(u, v)$ for each pixel (u, v), which combines color and texture information of the image [1]. For color, we use the RGB color space. For texture, we use 24 features calculated from the color correlogram (CC) of the pixel neighborhood [1]. We recall that an element $C^{d,\omega}(c_i; c_j)$ of the CC matrix gives the probability that two pixels (u_1, v_1) and

-**foreach** M *and* K *(*$M_{min} \leq M \leq M_{max}$, $K_{min} \leq K \leq K_{max}$*)* **do**

1. Initialize the parameters using the Fuzzy C-means.
2. **while** *not converged* **do**
   a. Compute $p(j|\vec{x}_i)$ using Eq. (9).
   b. Calculate feature relevance using Eq. (12).
   c. Update the weights using Eqs. (10) and (13).
   d. Update θ_{jl} using Eqs. (14) and (15).
   e. Update the parameters λ_{jl} using Eqs. (16).

   end
3. Calculate the MML criterion using Eq. (6)

end

-Select the optimal model M^* and K^*, such that $(M^*, K^*) = \text{argmin}_{M,K}(MML)$.

-Build the segmentation map by labeling each pixel in the image according to the optimal model.

Algorithm 1: The segmentation algorithm using the proposed model.

(u_2, v_2), where (u_2, v_2) is at distance d and orientation ω from (u_1, v_1), are of color c_i and c_j, respectively. We calculate the CC for 2 distances, $d \in \{1, 3\}$, and 4 orientations, $\omega \in \left\{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\right\}$. For each orientation, we took the average of the two distances. We derive from each CC the following texture features: *Inverse-Difference-Moment (IDM)*, *Energy (EN)*, *Contrast (CT)*, *Variance (VR)*, *Homogeneity (HG)* and *Correlation (CR)*.

To test the performance of the proposed method, we compared the segmentation accuracy of all the tested models against the ground truth (i.e., segmentation performed manually). For this purpose, we use the Berkeley Benchmark color image dataset [10], where the manual segmentations are available. We enlarged the dataset with images downloaded from the Corel and Freefoto datasets, making a database of 2000 images (Fig. (3) gives a sample of these images). To quantitatively measure how the segmentation model output is meaningful and close to the human segmentation, we use the following objective criteria:

- *The boundary localization error* (ϵ_1): This criterion measures the misalignment of the region contours between each tested segmentation and the ground truth. Let us define by $\mathbf{S} = \{S_1, ..., S_L\}$ and $\mathbf{S'} = \{S'_1, ..., S'_{L'}\}$ the sets of segments[1] composing the segmentation of a tested method and the ground truth, respectively. The region boundary localization error is defined by [10] as:

$$\epsilon_1 = \frac{1}{N} \sum_{(u,v)} \min\left\{\epsilon_{(u,v)}(TS, GT), \epsilon_{(u,v)}(GT, TS)\right\} \tag{8}$$

where $S_i(u, v)$ and $S'_j(u, v)$ are the segments that contain the pixel (u, v) in a tested segmentation model (TS) and the ground truth (GT), respectively. The errors $\epsilon_{(u,v)}(TS, GT)$ and $\epsilon_{(u,v)}(GT, TS)$ are given by: $\epsilon_{(u,v)}(TS, GT) = \frac{|S_i - S'_j|}{|S_i|}$, $\epsilon_{(u,v)}(GT, TS) = \frac{|S'_j - S_i|}{|S_j|}$, where " $-$ " is the set difference operator, and N is the number of pixels in the image.

- *The amount of over/under-segmentation* (ϵ_2): This criterion measures the amount of region over/under-segmentation produced by each tested segmentation, by comparison to the ground truth. According to previous definition of segments, a set of segments $S_{m1}, ..., S_{m\ell}$ $(2 \leq \ell \leq L)$ in the TS over-segment a segment S'_m in the GT *iff* [6]: $\forall i \in \{1, ..., \ell\}$: $|S_{mi} \cap S'_m| \geq k|S_{mi}|$ and $\sum_{i=1}^{\ell} |S_{mi} \cap S'_m| \geq k|S'_m|$, where k is a threshold that we set here to 0.75 as suggested in [6]. We define the error ϵ_2 as the sum of the number of segments in the GT that are over-segmented in the TS, and the number of segments in the TS that are over-segmented in the GT (i.e., instances of *under-segmentation*).

In the segmentation examples presented below, for each image we run the tested models 10 times and we report the average ($\bar{\epsilon}_{i=1,2}$) of the obtained errors values. To illustrate the capacity of our approach to accurately identify the number of regions, Fig. 1 shows two examples of typical MML calculation and image segmentation from our dataset. The real number of regions is 5 for both images. The final optimal number of components minimizing the MML for our model is $\hat{M} = 5$ and $\hat{K} = 2$ for both images; whereas when fixing $K = 1$, we obtained $\hat{M} = 6$ and $\hat{M} = 7$ for the first and second images, respectively. As expected, the errors ϵ_1 and ϵ_2 confirm the amount of over/under-segmentation avoided by our model.

We performed the segmentation on the 2000 images using the five aforementioned models. Figs. 4 shows the result of segmentation for a sample of images chosen randomly from the dataset. Each pixel in a given segmentation map takes the color mean value in the mixture component (region) to which the pixel is assigned. Tab. 1 gives the values of $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ for the shown examples, and the average of these errors in the whole dataset. Two conclusions can be drawn from the table. First, the MoGG (resp. MoGG+FS) generally outperforms the MoG (resp. MoG+FS) in both performance criteria. We noticed in some images that over-segmentation can be caused by specularities, self-shadowing or/and non-uniform illumination. In those images, homogeneity of color inside a region can be altered due to these effects. The flexibility of the MoGG allows to include those areas inside the main regions, while the MoG has a tendency to create small regions to fit them.

[1] A segment S, where $|S| > \tau$, is a connected set of equally labeled pixels. The symbol "$|\cdot|$" denotes the cardinality of a set. τ is a threshold set to 05.

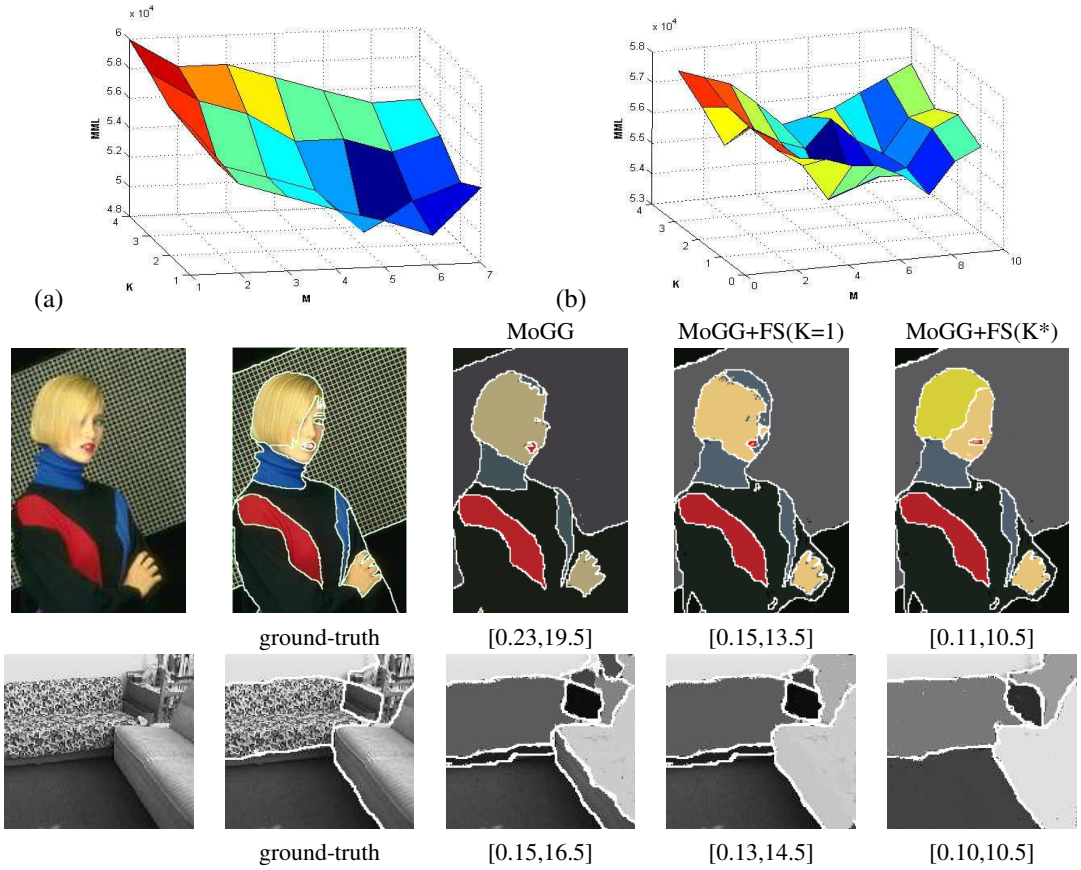

Figure 1. A typical example of image segmentation comparing the MoGG and MoGG+FS models against the ground truth: (a) and (b) represent the MML calculated for the MoGG+FS for the example 1 and 2 respectively; (c) to (g) represent, respectively, the original image, the ground truth, segmentation using MoGG, using MoGG+FS(K=1) and MoGG+FS(K=*). Below each segmentation, we show the errors [$\bar{\epsilon}_1$,$\bar{\epsilon}_2$]

Second, using FS in both MoGG and MoG models yields better performance than without using FS (see the values of the errors $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$). Finally, using a mixture of arbitrary number of components for irrelevant features enhances the segmentation accuracy compared to using only a fixed number of components (i.e., $K = 1$).

The last row of Tab. 1 shows the values of $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ obtained by each model for the whole dataset. We can see that the added performance using the proposed model, with respect to ϵ_1, is approximately 51% against MoG, 34% against MoG+FS, 48% against MoGG, and 24% against MoGG+FS with fixed $K = 1$. With respect to ϵ_2, the added performance is 52% against MoG, 34% against MoG+FS, 45% against MoGG, and 20% against MoGG+FS with fixed $K = 1$. These results clearly demonstrate the advantage of combining FS and GGD mixture modeling for reducing over/under-segmentation.

5. Conclusion

We proposed a new model which combines the GGD formulation and FS in robust mixture modeling for segmentation. We gave a principled framework for the estimation of the model parameters in an unsupervised fashion. The shown results demonstrate the usefulness and the effectiveness of the proposed model in reducing over/under-segmentation, and yielding accurate image segmentation, when using heavy-tailed and high-dimensional data. Future work will investigate other applications involving heavy-tailed and multiple-features data modeling (e.g., collection of images, filter-banks, wavelets, etc.). Also, adding spatial information is an important issue to handle toward yielding more semantically meaningful segmentation. Finally, the output of segmentation can be enriched by including multi-resolution analysis and hierarchical segmentation schemes,

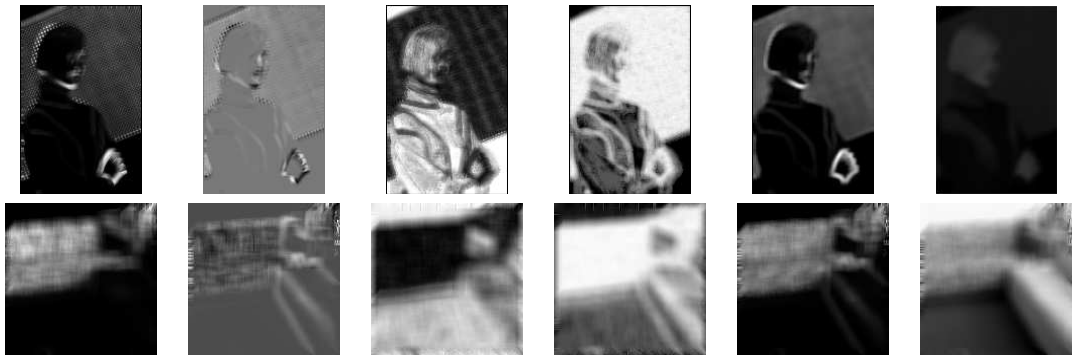

Figure 2. Texture features calculated for the images in Fig. 1.

Image	$[\bar{\epsilon}_1,\bar{\epsilon}_2]$				
	MoG	MoG+FS	MoGG	MoGG+FS(K=1)	MoGG+FS(K=*)
1 to 10	[0.21,22.2]	[0.12,15.5]	[0.18,19.4]	[0.12,14.9]	[0.10,10.5]
All the 2000 images	$[\mathbf{0.21}, \mathbf{23.4}]$	$[\mathbf{0.15}, \mathbf{16.8}]$	$[\mathbf{0.19}, \mathbf{20.7}]$	$[\mathbf{0.13}, \mathbf{13.9}]$	$[\mathbf{0.10}, \mathbf{11.2}]$

Table 1. Values of the errors ϵ_1 and ϵ_2 for the compared segmentation models.

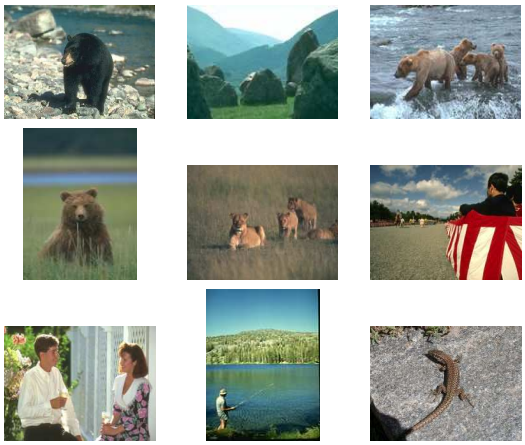

Figure 3. A sample of images in our test dataset.

where multiple outputs of segmentation can be used, for example, in object recognition.

6 Appendix

We minimize iteratively the objective function Eq. (6) using EM algorithm given by:

E-step:

$$p(j|\vec{x}_i) = \frac{p_j \prod_{l=1}^{D} \left[\beta_j(x_{il})\right]}{\sum_{j=1}^{M} p_j \prod_{l=1}^{D} \left[\beta_j(x_{il})\right]}, \tag{9}$$

where $\beta_j(x_{il}) = \rho_{l1} p(x_{il}|\theta_{jl}) + \rho_{l2} p(x_{il}|\vec{\varphi}_l)$, and $p(x_{il}|\vec{\varphi}_l) = \sum_{k=1}^{K} \pi_{kl} p(x_{il}|\varphi_{kl})$.

We estimate the parameters Θ in the M-step as

M-step:

$$\hat{p}_j = \frac{\max\left(\sum_{i=1}^{N} p(j|\vec{x}_i) - \frac{3D}{2}, 0\right)}{\sum_{j=1}^{M} \max\left(\sum_{i=1}^{N} \hat{z}_{ij} - \frac{3D}{2}, 0\right)} \tag{10}$$

$$\frac{1}{\hat{\rho}_{l1}} = 1 + \tag{11}$$

$$\frac{\max\left(\sum_{i=1}^{N}\sum_{j=1}^{M} p(j|\vec{x}_i) \frac{\rho_{l2} p(x_{il}|\vec{\varphi}_l)}{\beta_j(x_{il})} - \frac{3K}{2}, 0\right)}{\max\left(\sum_{i=1}^{N}\sum_{j=1}^{M} p(j|\vec{x}_i) \frac{\rho_{l1} p(x_{il}|\theta_{jl})}{\beta_j(x_{il})} - \frac{3M}{2}, 0\right)} \tag{12}$$

$$\hat{\pi}_{kl} = \frac{\max\left(\sum_{i=1}^{N}\sum_{j=1}^{M} p(j|\vec{x}_i) \frac{\rho_{l_2} \pi_{kl} p(x_{il}|\varphi_{kl})}{\beta_j(x_{il})} - \frac{3}{2}, 0\right)}{\sum_{k=1}^{K} \max\left(\sum_{i=1}^{N}\sum_{j=1}^{M} p(j|\vec{x}_i) \frac{\rho_{l2} \pi_{kl} p(x_{il}|\varphi_{kl})}{\beta_j(x_{il})} - \frac{3}{2}, 0\right)} \tag{13}$$

$$\hat{\mu}_{jl}^{\theta} = \frac{\sum_{i=1}^{N} p(j|\vec{x}_i) \frac{\rho_{l1} p(x_{il}|\theta_{jl}) |x_{il} - \mu_{jl}^{\theta}|^{\lambda_{jl}^{\theta}-2}}{\beta_j(x_{il})} x_{il}}{\sum_{i=1}^{N} p(j|\vec{x}_i) \frac{\rho_{l1} p(x_{il}|\theta_{jl}) |x_{il} - \mu_{jl}^{\theta}|^{\lambda_{jl}^{\theta}-2}}{\beta_j(x_{il})}} \quad (14)$$

$$\hat{\sigma}_{jl}^{\theta} = \sqrt[\lambda_{jl}^{\theta}]{\frac{\sum_{i=1}^{N} \frac{p(j|\vec{x}_i) \rho_{l1} p(x_{il}|\theta_{jl}) \lambda_{jl}^{\theta} A(\lambda_{jl}^{\theta}) |x_{il} - \mu_{jl}^{\theta}|^{\lambda_{jl}^{\theta}}}{\beta_j(x_{il})}}{\sum_{i=1}^{N} \frac{p(j|\vec{x}_i) \rho_{l1} p(x_{il}|\theta_{jl})}{\beta_j(x_{il})}}} \quad (15)$$

Finally, we estimate the parameters $\vec{\lambda}_j^{\theta}$ and $\vec{\lambda}_k^{\varphi}$, with $j = 1, ..., M$ and $k = 1, ..., K$ using the Newton-Raphson method:

$$\hat{\lambda}_{\circ l}^{\star} \simeq \hat{\lambda}_{\circ l}^{\star} - \left[\frac{\partial^2 MML(M,K)}{\partial {\lambda_{\circ l}^{\star}}^2} \right]^{-1} \left[\frac{\partial MML(M,K)}{\partial \lambda_{\circ l}^{\star}} \right] \quad (16)$$

where the symbol $\circ$ refers to either the index j or k, and the symbol $\star$ for either θ or φ.

References

[1] M.S. Allili and D. Ziou. Globally Adaptive Region Information for Color-Texture Image Segmentation. *Pattern Recognition Letters*, 28(15):1946-1956, November 2007.

[2] M.S. Allili, N. Bouguila and D. Ziou. Finite General Gaussian Mixture Modelling and Application to Image and Video Foreground Segmentation. *Journal of Electronic Imaging*, 17:013005.1-013005.13, January-March 2008.

[3] S. Boutemedjet, N. Bouguila, and D. Ziou. Feature Selection for Non-Gaussian Mixture Models. *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, 69-74, August 27-29, 2007.

[4] C. Carson, S. Belongie, H. Greenspan and J. Malik. Blobworld: Image Segmentation Using Expectation-Maximization and its Application to Image Querying. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(8):1026-1038, August 2002.

[5] T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning. New York: Springer. pp. 236243. 2001.

[6] A. Hoover et al. An Experimental Comparison of Range Image Segmentation Algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(7):673-689, July 1996.

[7] H.D. Cheng, X.H. Jiang and Y. Sun. Color image segmentation: advances and prospects. *Pattern Recongition*, 34(12):2259-2281, 2001.

[8] J. Kim et al. A Nonparametric Statistical Method for Image Segmentation Using Information Theory and Curve Evolution. *IEEE Trans. on Image Processing*, 14(10): 1486-1502, 2005.

[9] M.H.C. Law, M.A.T. Figueiredo and A.K. Jain. Simultanious Feature Selection and Clustering Using Mixture Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(9):1154-1166, September 2004.

[10] 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. *IEEE Int'l Conf. on Computer Vision (ICCV)*, 416-423, July 9-12, 2001.

[11] A.Y. Ng. On Feature Selection: Learning with Exponentially many Irrelevant Features as Training Examples. *15'th International Conf. on Machine Learning (ICML)*, 404-412, July 24-27, 1998.

[12] J. Novovicová et al. Divergence Based Feature Selection for Multimodal Class Densities. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(2):218-223, February 1996.

[13] C. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Information Science and Statistics, Springer, 2005.

[14] S. Zhu and A. Yuille. Region Competition: Unifying Snakes, Region Growing and Bayes/MDL for Multiband Image Segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(9):884-900, September 1996.

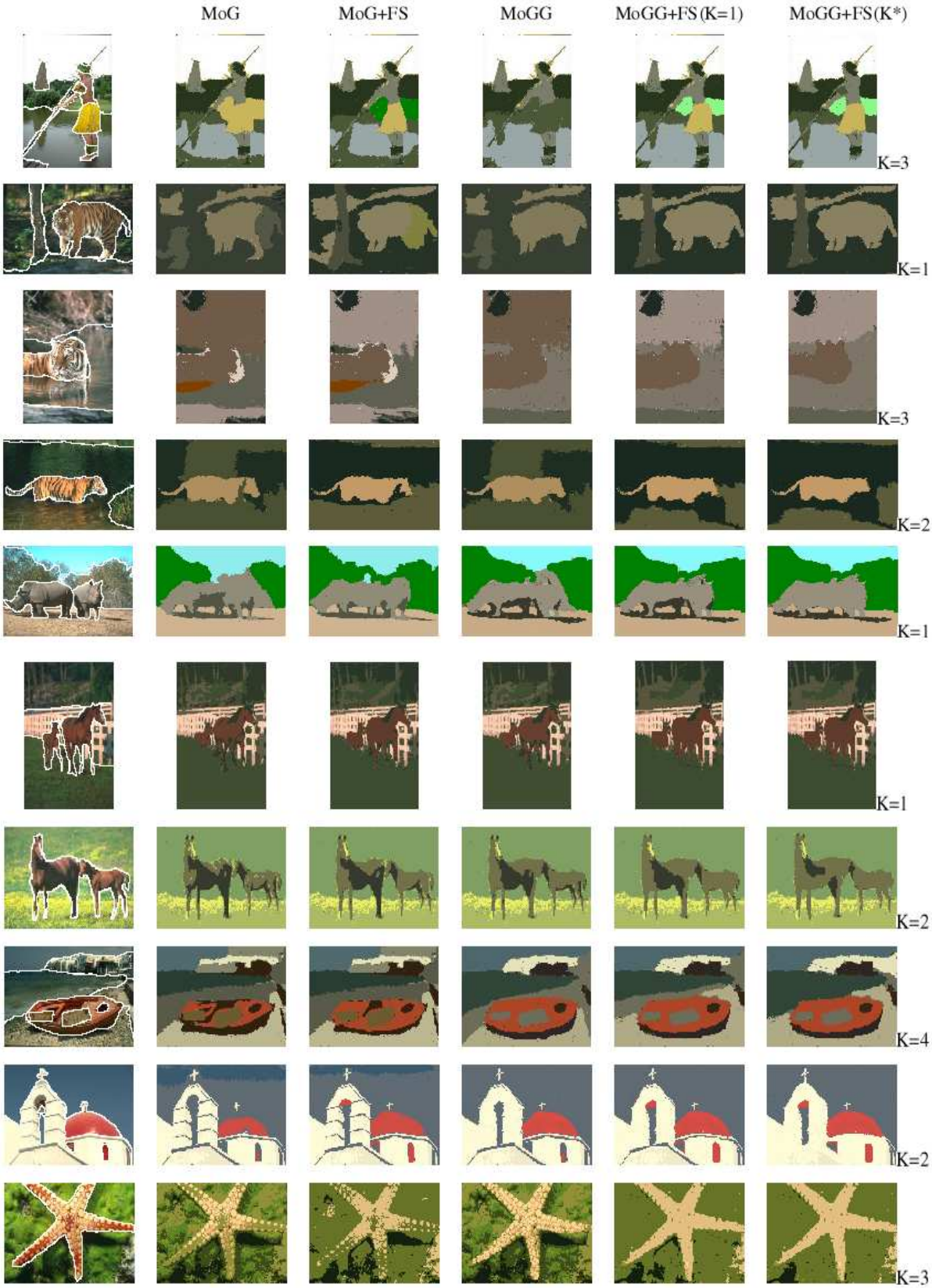

Figure 4. Example 1 to 10 of real-world image segmentation. In each row, the first column represents the ground truth. We show the optimal value of K found for each example.