Bag of Textons for Image Segmentation via Soft Clustering and Convex Shift*

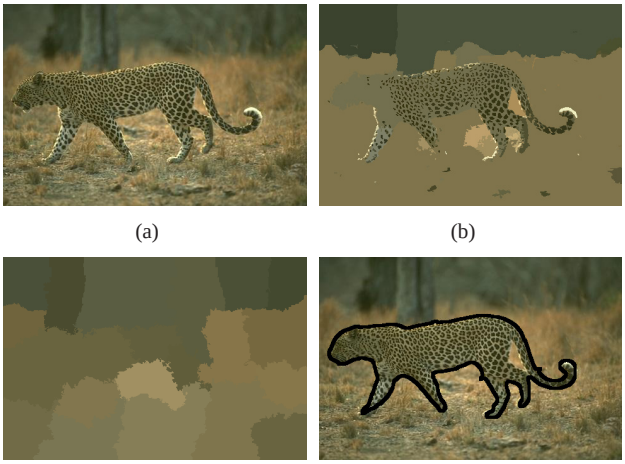
Zhiding Yu*† Ang Li† Oscar C. Au* Chunjing Xu‡#
*Dept. of Elec. & Comp. Eng., HKUST, HK †Robotics Inst., CMU, Pittsburgh, PA 15213
‡Multimedia Lab, SIAT, CAS, China 518055 #Dept. of Information Eng., CUHK, HK
{zdyu, eeau}@ust.hk, {yzhiding, angli123}@andrew.cmu.edu, cj.xu@siat.ac.cn

Abstract

We propose an unsupervised image segmentation method based on texton similarity and mode seeking. The input image is first convolved with a filter-bank, followed by soft clustering on its filter response to generate textons. The input image is then superpixelized where each belonging pixel is regarded as a voter and a soft voting histogram is constructed for each superpixel by averaging its voters' posterior texton probabilities. We further propose a modified mode seeking method - called convex shift - to group superpixels and generate segments. The distribution of superpixel histograms is modeled nonparametrically in the histogram space, using Kullback-Leibler divergence (K-L divergence) and kernel density estimation. We show that each kernel shift step can be formulated as a convex optimization problem with linear constraints. Experiment on image segmentation shows that convex shift performs mode seeking effectively on an enforced histogram structure, grouping visually similar superpixels. With the incorporation of texton and soft voting, our method generates reasonably good segmentation results on natural images with relatively complex contents, showing significant superiority over traditional mode seeking based segmentation methods, while outperforming or being comparable to state of the art methods.

1. Introduction

Before the introduction of "Bag of words model" (BoW) into computer vision, one could find the early applications of BoW in natural language processing (NLP) [9]. The BoW in NLP is a popular method that ignores the word orders for representing documents. The BoW model allows a dictionary-based modeling, and each document looks like a "bag" which contains some words from the dictionary. Computer vision researchers use a similar idea for image representation. To represent an image using BoW model, an image can be treated as a document. And features extracted from the image are considered as the "words". Extraction of words often includes following three steps: feature detection, feature description and codebook generation.[10] A definition of the BoW model can be the "histogram representation based on independent features" [11]. It is a widely used basic element for further processing in computer vision, especially in object categorization. Content based image indexing and retrieval (CBIR) is also an early adopter of this image representation technique [12].



(a) (b) (c) (d)

Figure 1. Segmentations of an image from the Berkeley Segmentation Dataset. (a) The original image. (b) Segmentation generated by mean shift. (c) Segmentation generated by quick shift. (d) Result obtained by the proposed algorithm, showing considerable improvement in terms of segmentation quality. Notice that although there is no human interaction, the produced foreground object segment highly overlaps the groundtruth.

Our method shares similar idea with BoW except that the "word" we extract is textual information. What we need is a compact representation for the range of different appearances of an object and this representation should be congruous with human perception of similarity. Texton have

*This work has been supported in part by the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China (GRF Project no. 610109 and 610210), and the National Natural Science Foundation of China (NSFC) (project No.61005011/F030403).

been proven effective in categorizing materials [16] as well as generic object classes [17]. Here we use textons [13] for describing human textual and color perception. To establish a metric for region similarity and dissimilarity, we construct a histogram for each superpixel region to quantitatively indicate the proportion of contribution from a specific texton. The texton thus plays a similar role as a "codeword".

In computer vision, the problem of segmentation and perceptual grouping remains challenging despite years of extensive study. The essence of segmentation can be regarded as clustering with elaborately designed pixel features and inter-pixel distance measures that tries to approximate humans visual perception of similarity. In the feature space, the cluster shape of features belonging to an image segment is often irregular. The fact that mode seeking methods can perform arbitrary shaped clustering makes it superior than many traditional clustering algorithms assuming regular shaped clusters in terms of segmentation performance. Despite the considerable literatures on mode seeking, we observe that many emphasize their applications in image segmentation while potentially posing them as low level preprocessing oriented [1] [3, 4, 5, 6]. Due to the pixel-wise operation, there has not been much fundamental improvement in terms of segmentation quality, as illustrated in Fig.1(b) and Fig.1(c). This tends to generate inferior segmentation when dealing with complex images, while image scenes often do contain abundant artificial or natural textural information. The concept of histogram based mode seeking have been introduced in mean shift tracking [2] [18, 19, 20], yet few have explored its application in image segmentation.

Our method combines the advantage of both mode seeking clustering and superpixel textual content which is far more informative than pixel-wise color. Instead of operating with pixels, we propose region-wise operations and we formulate the mode seeking problem into a constrained optimization problem for each kernel shift step. Region-wise operation allows one to investigate and design features much more versatile and powerful. Our method thus possesses the potential to outperform the segmentations produced by traditional mode seeking methods where simple pixel-wise features are not able to adequately describe the visual similarity. Such scheme also considerably alleviates the computational power required. Without loss of generality, suppose the complexity of a mode seeking algorithm is $O(N^2)$ where N is the total number of pixels. Consider the superpixelized image with N' regions (or superpixels). If $N = 100N'$, then for the same mode seeking algorithm the complexity has been reduced to 1/10000 of the original complexity. In practice, the overall algorithm complexity might not be considered such ideally. It does not hinder us, however, to show the potential of complexity reduction by region-wise operation.

2. Related Works

There exist considerable previous literatures related to our method concerning the aspects of texton representation and mode seeking. In review of these methods, most of them can be categorized into the following categories:

2.1. Relation with Texton Segmentation

In [13], Malik et al. proposed a image segmentation method based on normalized cuts with contour and texture analysis. A 40-D filter bank is used to convolve with the input image and to produce the response image. They also construct texton histogram for overlapped dense regions and χ^2 is adopted as the distance metric between two histograms. K-means is used to generate textons, turning the voting for histogram construction into hard decisions. Works in [14, 15] adopted similar strategies for texture similarity analysis. Different from their method, our method adopts EM soft clustering, which, in comparison with k-means, models the distribution much better since k-means only assumes spherical, uniform cluster shapes. Accordingly, we observe a boost in segmentation performance using textons generated by EM. In addition, the posterior probabilities of belonging to textons returned by EM enable one to adopt soft voting. Histograms constructed by soft clustering tend to reflect region similarity more accurately and the performance are less dependent on the number of textons.

2.2. Relation with Non-Euclidean Mode Seeking

Mode seeking provides a versatile tool for feature space analysis by finding local density maxima (or modes) in the feature space. In mode seeking clustering, data belonging to the same cluster fall within the same density attraction basin where the attraction force points to the direction that mostly increases the the estimated density. The feature space is partitioned by several clusters or basins with density maxima being the cluster centroids (the lowest points of basins).

Mean shift is regarded as one of the most canonical mode seeking algorithms with numerous real applications in computer vision. First proposed in [7] in 1975 and generalized in [8] in 1995, the method has not received wide attention until the publication of [1] in 2002. The method only assumes Mahalanobis distance metric where Euclidean distance is a special case.

Several works tried to introduce more versatile distance metrics into mode seeking. Zhao et al. [19] proposed a differentiable Earth Mover's Distance (EMD) that can be used as a distance metric for mean shift tracking. Leichter [20] proposed an alternative trackers that employ cross-bin metrics based on Mean Shift (MS) iterations. Both methods, however, only aim at tracking problem.

There have been interesting efforts that generalize mean shift to non-linear manifolds and intrinsically model curved

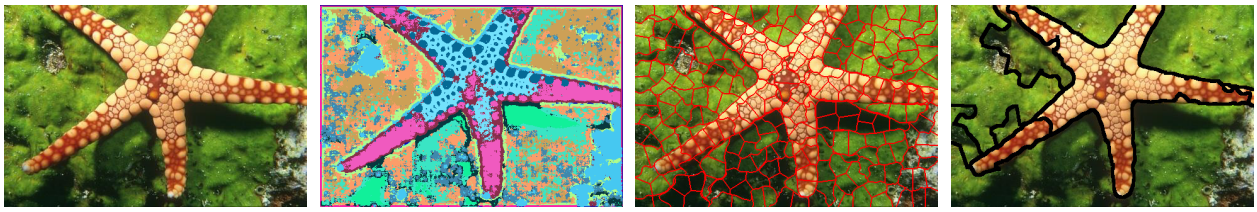


Figure 2. Algorithmic flow of the proposed method. Columns 1 to 4 respectively correspond to the original image, texton map (each pixel assigned to the most probable texton), superpixelized image and the final segmentation result. The histogram bandwidth and spatial bandwidth are respectively set to 1.2 and 60.

mean shift space [28, 29]. In contrast, we enforce the structure of the mean histogram directly as an explicit constraint. While intrinsic formulation is of great theory interest, our primary objective is to effectively perform mode seeking given the problem setting for certain task.

Sheikh et al. [3] proposed medoid shift, a mode seeking method that is able to adopt arbitrary, non-differentiable distance metrics. The method essentially transforms the mode seeking problem into a finite point searching problem. The shifted kernel location can appear at limited locations where there are data, thus only pair-wise data distance is needed and no metric differentiability is required. As is reported by [4], however, medoid shift is prone to over-fragmentation when data is sparse. On the other hand, the computation complexity of medoid shift increases significantly with respect to the increase of data size. Only simple, small scale (with respect to image size and number) image segmentation experiments were tested in [3]. Our method does not find approximate shifting locations but seeks an exact, optimal location for each kernel shift step. The proposed method thus works better with relatively sparse data, while its computational complexity increases relatively slower with the increase of data size.

3. The Proposed Image Segmentation Method

Our segmentation method consists of three major steps to perform segmentation. For any input image, the algorithm automatically decides the segment number with no human interaction. An algorithmic flow is illustrated in Fig. 2.

3.1. Representation by Textons

Using raw pixel-wise features is not be robust to noise and is difficult to extract invariant properties from the images. We convolve the image with a set of 17 filters (filter bank) to generate 17 response images, constructing a compact pixel-level image representation. In detail, we adopt a bank of 17 filters of size 15×15 which is composed of Gaussians with 3 different scales (1, 2, 4) applied to LAB channels, Lapacians of Gaussians with 4 different scales (1, 2, 4, 8) and the derivatives of Gaussians with two different scales (2, 4) for each axis (x and y). The filter bank we adopted is exactly the same as that adopted by [15, 17].

The obtained 17-D response image pixels are to be clustered to generate textons. Unlike popular texton generation schemes which commonly use k-means as the clustering method, we adopt K cluster Expectation-Maximization to softly cluster response image pixels and generate K textons. Since k-means only assumes spherical cluster shape which can be far from real data distribution, its texton representation and the region texton statics are far inferior than EM. Using EM we are also able to obtain the K posterior probabilities of belonging to the K textons for each pixel.

3.2. Superpixelization and Local Bag of Textons

To reduce computational complexity, we use the method proposed by X. Ren et al. [14, 21, 22] to generate superpixelized images[1]. The parameters N_sp, N_sp2 and N_ev corresponding to the number of superpixels coarse/fine and the number of eigenvalues are first respectively set to 200, 1000 and 40. By this set of parameters we are able to obtain coarsely and finely superpixelized images with more than 200 and 1000 superpixels respectively.

For each coarse superpixel and fine superpixel, we softly vote its texton frequency by averaging posterior texton probabilities over all member pixels and construct a histogram for each superpixel. We call this method "Bag of textons" since there is no constraint on the textons sequence. And just like BoW where frequency of words characterizes the document type, the frequency of textons here characterizes the region appearance and defines the similarities between any two regions.

The set of coarse superpixels are the basic units we want to cluster to generate final segmentations, while the set of fine superpixels serve as mode seeking samples for pdf estimation. For each coarse superpixel, a histogram kernel and a spatial kernel are initialized with respect to the superpixel histogram and superpixel spatial location. Mode seeking is then performed for each coarse superpixel based on samples (fine superpixel histograms and spatial locations). The

[1]The corresponding Matlab code is kindly available at http://www.cs.sfu.ca/∼mori/research/superpixels/

advantage of such strategy is that larger coarse superpixels speeds up the algorithm and contain more region information, while the larger number of fine superpixels give adequate sampling support to estimate a better pdf.

3.3. Proposed Convex Shift Algorithm

For traditional mean shift algorithm, suppose $\mathbf{x}^r$ and $\mathbf{x}^s$ respectively represents the d dimensional feature space vector and 2 dimensional spatial coordinate of an image pixel. For mean shift based image segmentation, one adopts the following multivariate kernel density estimator:

$$\hat{f}_{h_r,h_s}(\mathbf{x}^r,\mathbf{x}^s) = \frac{C}{Nh_r^d h_s^2}\sum_{i=1}^{N} k(\|\frac{\mathbf{x}^r-\mathbf{x}_i^r}{h_r}\|^2)k(\|\frac{\mathbf{x}^s-\mathbf{x}_i^s}{h_s}\|^2). \quad (1)$$

where the function $k(x)$ is the *profile* of a kernel and C is a normalization constant that makes the above multivariate kernel integrates to one. $h_r > 0$ and $h_s > 0$ are the smoothing parameters called the bandwidth. Taking the derivative of $\hat{f}(\mathbf{x}^r,\mathbf{x}^s)$ with respect to $\mathbf{x}^r$ and defining the new kernel profile $g(x) = -k'(x)$, one has:

$$\begin{aligned}
&\frac{\partial f_{h_r,h_s}(\mathbf{x}^r,\mathbf{x}^s)}{\partial \mathbf{x}^r} \\
&= \frac{2C}{Nh_r^{d+2}h_s^2}\sum_{i=1}^{N}(\mathbf{x}_i^r-\mathbf{x}^r)g(\|\frac{\mathbf{x}^r-\mathbf{x}_i^r}{h_r}\|^2)k(\|\frac{\mathbf{x}^s-\mathbf{x}_i^s}{h_s}\|^2) \\
&= \frac{2C}{Nh_r^{d+2}h_s^2}[\sum_{i=1}^{N} g(\|\frac{\mathbf{x}^r-\mathbf{x}_i^r}{h_r}\|^2)k(\|\frac{\mathbf{x}^s-\mathbf{x}_i^s}{h_s}\|^2)] \\
&[\frac{\sum_{i=1}^{N}\mathbf{x}_i^r g(\|\frac{\mathbf{x}^r-\mathbf{x}_i^r}{h_r}\|^2)k(\|\frac{\mathbf{x}^s-\mathbf{x}_i^s}{h_s}\|^2)}{\sum_{i=1}^{N} g(\|\frac{\mathbf{x}^r-\mathbf{x}_i^r}{h_r}\|^2)k(\|\frac{\mathbf{x}^s-\mathbf{x}_i^s}{h_s}\|^2)} - \mathbf{x}^r]
\end{aligned} \quad (2)$$

The last term of equation (2) is the *mean shift* for the feature space kernel.

$$\mathbf{m}_{h_r,h_s}(\mathbf{x}^r) = \frac{\sum_{i=1}^{N}\mathbf{x}_i^r g(\|\frac{\mathbf{x}^r-\mathbf{x}_i^r}{h_r}\|^2)k(\|\frac{\mathbf{x}^s-\mathbf{x}_i^s}{h_s}\|^2)}{\sum_{i=1}^{N} g(\|\frac{\mathbf{x}^r-\mathbf{x}_i^r}{h_r}\|^2)k(\|\frac{\mathbf{x}^s-\mathbf{x}_i^s}{h_s}\|^2)} - \mathbf{x}^r \quad (3)$$

The mean shift vector for the spatial kernel can be obtained similarly.

Since we use histograms to model region statistics, we adopt K-L divergence to measure the distance between two histograms:

$$d_{KL}(H,K) = \sum_{p=1}^{d} h_p \log\frac{h_p}{k_p}.$$

where d is the histogram dimension, $H = [h_1, h_2, ..., h_d]^\top$ and $K = [k_1, k_2, ..., k_d]^\top$ are two histograms with the constraints $\sum_{i=1}^{d} h_p = \sum_{i=1}^{d} k_p = 1$. Suppose the histogram kernel center is denoted as $\mathbf{x}^h = [x_{(1)}^h, x_{(2)}^h, ..., x_{(d)}^h]^\top$ and the histogram of the ith sample (fine superpixel) is denoted as $\mathbf{x}_i^h = [x_{i,(1)}^h, x_{i,(2)}^h, ..., x_{i,(d)}^h]^\top$. Plugging in the K-L divergence distance measure, we have the following density estimator:

$$\hat{f}_{h_h,h_s}(\mathbf{x}^h,\mathbf{x}^s) = \frac{C}{Nh_h^d h_s^2}\sum_{i=1}^{N} k(\frac{d_{KL}(\mathbf{x}^h,\mathbf{x}_i^h)}{h_h^2})k(\|\frac{\mathbf{x}^s-\mathbf{x}_i^s}{h_s}\|^2). \quad (4)$$

The mode seeking problem thus becomes increasing the estimated density subject to the sum of histogram bins in each color channel equals to 1, which is a constrained gradient ascent problem. For histogram kernel, we introduce linear relaxation using K-L divergence kernel with a linear profile, while for spatial kernel, the normal kernel is adopted. Notice that the K-L divergence kernel is meaningful only when the histogram structure is preserved. The density estimator thus becomes:

$$\hat{f}_{h_h,h_s}(\mathbf{x}^h,\mathbf{x}^s) = \frac{C}{Nh_h^d h_s^2}\sum_{i=1}^{N} k_{KL}(\frac{d_{KL}(\mathbf{x}^h,\mathbf{x}_i^h)}{h_h^2})k_N(\|\frac{\mathbf{x}^s-\mathbf{x}_i^s}{h_s}\|^2). \quad (5)$$

$$k_{KL}(x) = \begin{cases} 1-x & 0\le x\le 1 \\ 0 & x>1 \end{cases}. \quad (6)$$

$$k_N(x) = \exp(-\frac{1}{2}x) \quad (7)$$

To solve the problem, rewrite (12) into the following form:

$$\begin{aligned}
\mathbf{x}^{r(l+1)} &= \frac{\sum_{i=1}^{N}\mathbf{x}_i^r g(\|\frac{\mathbf{x}^{r(l)}-\mathbf{x}_i^r}{h_r}\|^2)k(\|\frac{\mathbf{x}^{s(l)}-\mathbf{x}_i^s}{h_s}\|^2)}{\sum_{i=1}^{N} g(\|\frac{\mathbf{x}^{r(l)}-\mathbf{x}_i^r}{h_r}\|^2)k(\|\frac{\mathbf{x}^{s(l)}-\mathbf{x}_i^s}{h_s}\|^2)} \\
&= \arg\min_{\mathbf{x}^r}\sum_{i=1}^{N}\|\mathbf{x}_i^r-\mathbf{x}^r\|^2 \\
&\quad g(\|\frac{\mathbf{x}^{r(l)}-\mathbf{x}_i^r}{h_r}\|^2)k(\|\frac{\mathbf{x}^{s(l)}-\mathbf{x}_i^s}{h_s}\|^2).
\end{aligned} \quad (8)$$

where $\mathbf{x}^{r(l)}$ denotes the color space kernel location in the lth iteration. Recall the K-L divergence kernel we introduced. The linear kernel profile yields:

$$g_{KL}(x) = \begin{cases} 1 & 0\le x\le 1 \\ 0 & \text{otherwise} \end{cases}. \quad (9)$$

Plugging the kernel profile $g_{KL}(x)$ and the K-L divergence measure in equation (8), and changing the feature space kernel into the histogram kernel, the histogram kernel shift can be formulated as solving the following convex problem:

$$\begin{aligned}
\min_{\mathbf{x}^h} \quad & \sum_{i,\mathbf{x}_i^h\in\mathbf{S}^{(l)}} k_N(\|\frac{\mathbf{x}^{s(l)}-\mathbf{x}_i^s}{h_s}\|^2)\sum_{p=1}^{d} x_{(p)}^h \log\frac{x_{(p)}^h}{x_{i,(p)}^h} \\
s.t. \quad & \mathbf{x}^h \succeq 0 \\
& ||\mathbf{x}^h(1:K)||_1 = 1
\end{aligned}$$

where $\mathbf{S}^{(l)} = \{\mathbf{x}_i^h | d_{KL}(\mathbf{x}^{h(l)}, \mathbf{x}_i^h) \leq h_h^2\}$. One can verify the equivalence between solving the above problem and increasing the density estimator in equation (11). We will verify this property in the next subsection. For the spatial kernel, the strategy for calculating the spatial kernel shift is identical to that in mean shift:

$$\mathbf{m}_{h_r,h_s}(\mathbf{x}^s) = \frac{\sum_{i=1}^N \mathbf{x}_i^s g_N(\|\frac{\mathbf{x}^s - \mathbf{x}_i^s}{h_s}\|^2) k_{KL}(\frac{d_{KL}(\mathbf{x}^h, \mathbf{x}_i^h)}{h_h^2})}{\sum_{i=1}^N g_N(\|\frac{\mathbf{x}^r - \mathbf{x}_i^r}{h_s}\|^2) k_{KL}(\frac{d_{KL}(\mathbf{x}^h, \mathbf{x}_i^h)}{h_h^2})} - \mathbf{x}^s \quad (10)$$

The segmentation algorithm can be described as follows:
1. For each coarse superpixel, initialize its histogram kernel and spatial kernel according to the region color statistics and the mean spatial coordinate of the contained pixels.
2. Recursively shift the histogram kernel by solving the above convex problem using a convex solver and shift the spatial kernel according equation (10) until convergence.
3. Group the set of coarse superpixels that share similar histogram kernel locations.

3.4. Algorithm Convergence

Definition 3.1 *For any sequential step l and $l+1$ and the corresponding histogram kernel location $\mathbf{x}^h(l)$ and $\mathbf{x}^h(l+1)$, the transitory density estimation is defined as:*

$$\hat{ft}_{h_h,h_s}(\mathbf{x}^{h(l+1)}, \mathbf{x}^s) = C_1 - C_2 \sum_{i,\mathbf{x}_i^h \in \mathbf{S}^{(l)}} k_N(\|\frac{\mathbf{x}^{s(l)} - \mathbf{x}_i^s}{h_s}\|^2) \sum_{p=1}^d x_{(p)}^{h(l+1)} \log \frac{x_{(p)}^{h(l+1)}}{x_{i,(p)}^h} \quad (11)$$

where $C_1 = \frac{C}{N h_h^d h_s^2} \sum_{i=1}^N k_N(\|\frac{\mathbf{x}^s - \mathbf{x}_i^s}{h_s}\|^2)$, $C_2 = \frac{C}{N h_h^d h_s^2 h_h^2}$.

Lemma 3.1 $\hat{ft}_{h_h,h_s}(\mathbf{x}^{h(l+1)}, \mathbf{x}^s) \geq \hat{f}_{h_h,h_s}(\mathbf{x}^{h(l)}, \mathbf{x}^s)$

Proof: According to equation (4), we have:

$$\hat{f}_{h_h,h_s}(\mathbf{x}^{h(l)}, \mathbf{x}^s) = C_1 - C_2 \sum_{i,\mathbf{x}_i^h \in \mathbf{S}^{(l)}} k_N(\|\frac{\mathbf{x}^{s(l)} - \mathbf{x}_i^s}{h_s}\|^2) \sum_{p=1}^d x_{(p)}^{h(l)} \log \frac{x_{(p)}^{h(l)}}{x_{i,(p)}^h} \quad (12)$$

According to convex shift which is in the form of constrained minimization, $\mathbf{x}^{h(l+1)}$ is obtained through minimizing $\sum_{i,\mathbf{x}_i^h \in \mathbf{S}^{(l)}} k_N(\|\frac{\mathbf{x}^{s(l)} - \mathbf{x}_i^s}{h_s}\|^2) \sum_{p=1}^d x_{(p)}^h \log \frac{x_{(p)}^h}{x_{i,(p)}^h}$ over $\mathbf{x}^h$, we thus directly proved Lemma 3.1.

Lemma 3.2 $\hat{f}_{h_h,h_s}(\mathbf{x}^{h(l+1)}, \mathbf{x}^s) \geq = \hat{ft}_{h_h,h_s}(\mathbf{x}^{h(l+1)}, \mathbf{x}^s)$

Proof: Since some of the samples belonging to the lth kernel may go out of the range of $l+1$th kernel, these samples contributes negative values to $\hat{ft}_{h_h,h_s}(\mathbf{x}^{h(l+1)}, \mathbf{x}^s)$ while their contribution to $\hat{f}_{h_h,h_s}(\mathbf{x}^{h(l+1)}, \mathbf{x}^s)$ is 0. In addition, new samples may come within the range of the $l+1$th kernel, which contributes nonnegative values to $\hat{f}_{h_h,h_s}(\mathbf{x}^{h(l+1)}, \mathbf{x}^s)$. Thus, we have the above lemma.

Theory 3.1 *The estimated density monotonically increases with each convex shift step and the algorithm converges.*

Proof: Spatial kernel is independent with histogram kernel and spatial kernel shift also increases the estimated density [1]. According to Lemma 3.1 and Lemma 3.2, the estimated density thus monotonically increases with iterative histogram and spatial kernel shift. Since the estimated density is upper bounded, the algorithm is guaranteed to converge.

4. Experimental Results

We perform segmentation test on a number of natural images selected from the Berkeley Segmentation Dataset. For convex shift, a simple postprocessing is used to eliminate single superpixels by merging them into the most similar neighboring regions. The bandwidth parameters h_h, h_s are respectively set to 1.2 and 60. Our segmentation results are compared with segmentations obtained by quick shift and mean shift. We also compare our method with state of the art segmentation methods such as the Fusion of Clustering Results (FCR) method [23], the Probabilistic Rand Index Fusion (PRIF) method [26] and *gPb*-owt-ucm [30]. We use the VLFeat Matlab package [29] to implement quick shift. The parameters $ratio$, $kernelsize$ and $maxdist$ are respectively set to 0.5, 12 and 30, which is observed to be the best trade off to avoid both oversmoothing and oversegmentation. For the majority of mean shift experiments, we set h_s, h_r and minimum region size M respectively to be 8, 7 and 100 - the set of segmentation parameters adopted in [1]. Smaller bandwidth parameters are chosen for image 4, 19, 23, 24 in order to prevent serious over-merging. We adopt a unified UCM threshold for *gPb*-owt-ucm and tune it to visually optimize its segmentation. The comparison of segmentation results is illustrated in Fig.3 and Fig.4. Experimental results indicate the superiority of using the proposed method, especially on those images being more complex and textured. Under the scheme of "Bag of textons", our method significantly outperforms quick shift and mean shift for incorporating abundant textual information. Our method also slightly outperforms FCR and PRIF - which are well-designed state of the art segmentation methods - and is comparable with *gPb*-owt-ucm. We observe that *gPb*-owt-ucm indeed is very powerful but does suffer from over-merging through weak boundary and over-segmentation caused by strong intra-region variation (common problems with contour finding methods). Notice that in contrast to *gPb*-owt-

Figure 3. Comparison of segmentation results obtained by different methods. Row 1 to 7 respectively correspond to original images and results produced by quick shift, mean shift, FCR, PRIF, *gPb*-owt-ucm and the proposed method.

Figure 4. Comparison of segmentation results obtained by different methods. For every 7 rows, row 1 to 7 respectively correspond to the original images and results produced by quick shift, mean shift, FCR, PRIF, *gPb*-owt-ucm and the proposed method.

ucm, we have not even elaborately design the spatial constraint and local discontinuity rule to obtain better segmentations. Previous works such as [24, 25] allow one to plug in spatial consistency information on mode seeking in a way far better than current scheme. With better spatial consistency information, a further boost of the segmentation quality is expected.

5. Conclusions and Future Works

We have proposed a mode seeking based algorithm that can effectively segment natural color images. Compared with traditional mode seeking based segmentation method, the method tends to produce excellent segmentations that are much more semantically meaningful and perceptually congruous with human perception of similarity on complex, textured images. In addition, our method is comparable with state of the art segmentation methods. Processing of the proposed method is parallelable like traditional mode seeking methods. Thus parallel implementation is expected to significantly speed up the algorithm's processing speed. Our future work will include detailed report with more comprehensive evaluations and further improvements.

References

[1] D. Comaniciu and P. Meer. "Mean shift: A robust approach toward feature space analysis." *IEEE Trans. PAMI*, 2002.

[2] A. Yilmaz, "Object tracking by asymmetric kernel mean shift with automatic scale and orientation selection." In *CVPR*, 2007.

[3] Y. A. Sheikh, E. A. Khan and T. Kanade. "Mode-seeking by Medoidshifts." In *ICCV*, 2007.

[4] A. Vedaldi and S. Soatto. "Quick shift and kernel methods for mode seeking." In *ECCV*, 2008.

[5] A. Vedaldi and S. Soatto. "Really quick shift: Image segmentation on a GPU." In *Workshop on Computer Vision using GPUs, held with ECCV*, 2010.

[6] K. Zhang, J. T. Kwok and M. Tang. "Accelerated convergence using dynamic mean shift." In *ECCV*, 2006.

[7] K. Fukunaga and L. Hostetler. "The estimation of the gradient of a density function with application in pattern recognition." *IEEE Trans. Info. Theory*, 1975.

[8] Y. Cheng. "Mean shift, mode seeking and clustering." *IEEE Trans. PAMI*, 1995.

[9] D. Lewis. "Naive (Bayes) at forty: The independence assumption in information retrieval". In *ECML*, 1998.

[10] FF. Li and P. Perona. "A Bayesian hierarchical model for learning natural scene categories". In *CVPR*, 2005.

[11] FF. Li, R. Fergus and A. Torralba. "Recognizing and learning object categories". In *CVPR 2007 short course*, 2007.

[12] G. Qiu. "Indexing chromatic and achromatic patterns for content-based colour image retrieval". *Pattern Recognition*, 2002.

[13] J. Malik, S. Belongie, T. Leung and J. Shi. "Contour and texture analysis for image segmentation". *IJCV*, 2001.

[14] X. Ren and J. Malik. "Learning a classification model for segmentation". *ICCV*, 2003.

[15] J. Shotton. "TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context". *IJCV*, 2007.

[16] M. Varma and A. Zisserman. "A statistical approach to texture classification from single images". *IJCV*, 2005.

[17] J. Winn, A. Criminisi and T. Minka. "Categorization by learned universal visual dictionary". In *ICCV*, 2005.

[18] D. Comaniciu, V. Ramesh and P. Meer. "Real-time tracking of non-rigid objects using mean shift". In *CVPR*, 2000.

[19] Q. Zhao, Z. Yang, H. Tao. "Differential Earth Movers Distance with its applications to visual tracking". *IEEE Trans. PAMI*, 2010.

[20] I. Leichter. "Mean shift trackers with cross-bin metrics". Accpted to *IEEE Trans. PAMI*, 2011.

[21] G. Mori, X. Ren, A. Efros, and J. Malik. "Recovering human body configurations: Combining segmentation and recognition". In *CVPR*, 2004.

[22] G. Mori. "Guiding model search using segmentation". In *ICCV*, 2005.

[23] M. Mignotte. "Segmentation by fusion of histogram-based k-means clusters in different color spaces". *IEEE Trans. Image Proc.*, 2008.

[24] Z. Yu, O. Au, K. Tang and C. Xu. "Nonparametric density estimation on a graph: Learning framework, fast approximation and application in image segmentation". In *CVPR*, 2011.

[25] H. Liu and S. Yan. "Robust graph mode seeking by graph shift". In *ICML*, 2010.

[26] M. Mignotte. "A label field fusion Bayesian model and its penalized maximum rand estimator for image segmentation". *IEEE Trans. on Image Proc.*, 2010.

[27] R. Subbarao and P. Meer. "Nonlinear mean shift for clustering over analytic manifolds". In *CVPR*, 2006.

[28] H.E. Cetingul and R. Vidal. "Intrinsic mean shift for clustering on Stiefel and Grassmann manifolds". In *CVPR*, 2009.

[29] A. Vedaldi and B. Fulkerson. "VLFeat - an open and portable library of computer vision algorithms". http://www.vlfeat.org/, 2008.

[30] P. Arbelaez, M. Maire, C. Fowlkes and J. Malik. "Contour Detection and Hierarchical Image Segmentation". *IEEE Trans. PAMI*, 2010.