Proc. VIIth Digital Image Computing: Techniques and Applications, Sun C., Talbot H., Ourselin S. and Adriaansen T. (Eds.), 10-12 Dec. 2003, Sydney

False-Peaks-Avoiding Mean Shift Method for Unsupervised Peak-Valley Sliding Image Segmentation

Hanzi Wang and David Suter

Department of. Electrical. and Computer Systems Engineering
Monash University, Clayton 3800, Victoria, Australia
{hanzi,wang; d.suter}@eng.monash.edu.au

Abstract. The mean shift (MS) algorithm is sensitive to local peaks. In this paper, we show both empirically and analytically that when using sample data, the reconstructed PDF may have false peaks. We show how the occurrence of the false peaks is related to the bandwidth h of the kernel density estimator, using examples of gray-level image segmentation. It is well known that in MS-based approaches, the choice of h is important: we provide a quantitative relationship between false peaks and h. For the gray-level image segmentation problem, we provide a complete unsupervised peak-valley sliding algorithm for gray-level image segmentation.

1 Introduction

The mean shift method is popular for a wide variety of applications such as video tracking [9], image filtering [5], clustering [6] and image segmentation [4, 7]. In essence, it is a local (and thereby somewhat robust) form of mode seeking. It is local because it operates on a window and it also achieves a degree of scale selectivity since it works with a smoothed estimate of the underlying density function. In the most commonly used form [10, 7], the window size and the smoothing are directly related to a quantity h that is the "bandwidth" choice (for the kernel density estimator).

Although many authors have remarked that the value h needs to be chosen with care, the general impression given is that the results are not that sensitive to the choice of h and that one generally takes a pragmatic "hit and miss" affair. In this paper we illustrate that there are two issues affected by the setting of h: the rather disastrous appearance of false peaks (where the application of the mean shift process will fail) and the choice of scale (affecting the significance of actual peaks in the underlying density – at large scales the density is very smoothed and local peaks are disregarded or merged). The latter behavior is much more benign and, indeed, as it performs a type of controlled scale-space analysis, can be used to advantage.

This paper provides an important warning about the sensitivity of the mean shift to false peak noise due to quantization. In this paper, we choose, for simplicity, the problem of histogram based gray level image segmentation. We show that one can rather simply predict values of h that will be problematic; and thereby, in this setting,

we provide a means for a completely automated approach. This negates the need for the setting of a value for any parameter, including h.

2 Density Gradient Estimation and Mean Shift

There are several nonparametric methods available for probability density estimation [12]. Kernel estimation is one of the most popular techniques. Given a set of n data points $\{x_i\}_{i=1,\ldots,n}$ in a d-dimensional Euclidian space R^d, the multivariate kernel density estimator with kernel K and window radius/band-width h is ([12], p.76):

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

K(x) should satisfy some conditions ([14], p.95). The Epanechnikov kernel ([12], p.76) is a kernel that yields minimum mean integrated square error (MISE):

$$K_e(\mathrm{X}) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-\|\mathrm{X}\|^2) & if\ \|\mathrm{X}\| < 1 \\ 0 & otherwise \end{cases} \tag{2}$$

where c_d is the volume of the unit d-dimensional sphere, e.g., $c_1=2$, $c_2=\pi$, $c_3=4\pi/3$.

The estimate of the density gradient can be defined as the gradient of the kernel density estimate (1)

$$\hat{\nabla} f(x) \equiv \nabla \hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} \nabla K\left(\frac{x - x_i}{h}\right) \tag{3}$$

According to (3), the density gradient estimate of the Epanechnikov kernel can be written as:

$$\hat{\nabla} f(x) = \frac{n_x}{n(h^d c_d)}\frac{d+2}{h^2}\left(\frac{1}{n_x}\sum_{x_i \in S_h(x)}[x_i - x]\right) \tag{4}$$

where the region $S_h(x)$ is a hypersphere of the radius h, having the volume $h^d c_d$, centered at x, and containing n_x data points.

The mean shift vector $M_h(x)$ is defined as:

$$\mathrm{M_h}(x) \equiv \frac{1}{n_x}\sum_{x_i \in S_h(x)}[x_i - x] = \frac{1}{n_x}\sum_{x_i \in S_h(x)} x_i - x \tag{5}$$

Equation (4) can be rewritten as:

$$\mathrm{M_h}(x) \equiv \frac{h^2}{d+2}\frac{\hat{\nabla} f(x)}{\hat{f}(x)} \tag{6}$$

Equation (6) firstly appeared in [10]. Equations (5) and (6) show that the mean shift vector is the difference between the local mean and the center of the window, and the mean shift is an unsupervised nonparametric estimator of density gradient. Applying the mean shift leads to the steepest ascent with a varying step size that is the magnitude of the gradient [3]. The converged centers correspond to modes (or centers of the regions of high concentration) of data. A proof of the convergence can be found in [6, 7]. The mean shift method has been widely exploited and applied in low level computer vision tasks [3, 4, 11, 13, 5, 7].

During histogram analysis, valleys between modes need to be found to set thresholds [2]. In [8], the Gaussian kernel is employed in a method to detect saddle

points (i.e., valleys in our method). However, the saddle detection method needs a pre-partitioned cluster and its complementary cluster set. In the next subsection, we will provide a simple method using the Epanechnikov kernel.

2.1 Mean Shift Valley (MSV) Algorithm

We note that the direction opposite to the mean shift vector will always points toward to a local minimum density, thus, we define the mean shift valley vector:

$$\mathrm{MV_h}(x) = -\mathrm{M_h}(x) = x - \frac{1}{n_x} \sum_{x_i \in S_h(x)} x_i \tag{7}$$

Replace $\mathrm{M_h}(x)$ in (6) by $\mathrm{MV_h}(x)$, we can obtain:

$$\mathrm{MV_h}(x) \equiv -\frac{h^2}{d+2} \frac{\hat{\nabla} f(x)}{\hat{f}(x)} \tag{8}$$

In practice, the step-size given by the above analysis may lead to oscillation, particularly when finding valleys. To derive a recipe for avoiding the oscillations we use an adjustable step-size $0 < p \le 1$:

$$\mathrm{y_{k+1}} = \mathrm{y_k} + p \cdot MV_h(y_k) \tag{9}$$

If the step at $\mathrm{y_k}$ is so that $\mathrm{MV_h(y_k)^T\ MV_h(y_{k+1})} > 0$ then the oscillation is avoided.
The mean shift valley algorithm can be described as:

1. Choose the bandwidth, h; set p =1; and initialize the location of the window
2. Compute the shift step vector $\mathrm{MV_h(y_k)}$.
3. Compute y_{k+1} by equation (9) and $\mathrm{MV_h}(y_{k+1})$.
4. If $\mathrm{MV_h(y_k)^T\ MV_h(y_{k+1})} > 0$, go to step 5; Otherwise, we let $p=p/2$. Repeat step 3 and 4 until $\mathrm{MV_h(y_k)^T\ MV_h(y_{k+1})} > 0$;
5. Translate the search window by $p \cdot MV_h(y_k)$.
6. Repeat step 3 to step 5 until convergence.

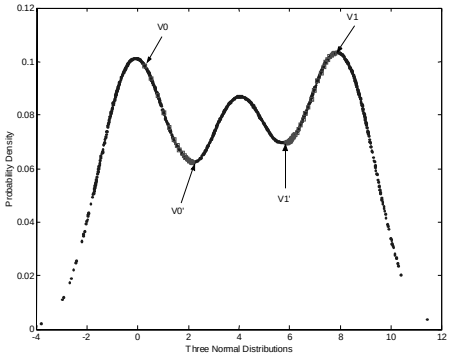


Figure 1. An example of the application of the mean shift valley method.

To illustrate the mean shift valley method, three normal modes (mode 1 includes 600 data points, mode 2 includes 500 data points, and mode 3 includes 600 data points)

with total 1700 data points were generated in Figure 1. We selected two initial points: V0 (0.3) and V1 (7.8). The search window radius was chosen as 2.0. The mean shift valley method automatically found the local minimum densities (converged points). Precisely, V0' was located at 2.1831, and V1' was at 5.8898. The centers (V0' and V1') of the converged windows correspond to the local minimum probability densities. If we use V0' and V1' as two density thresholds, the whole data can be decomposed into three modes. Table 1 gives the obtained parameters.

	Mode 1		Mode 2		Mode 3	
	Mean	Number	Mean	Number	Mean	Number
Generated Data	0	600	4	500	8	600
Estimated Parameters	-0.0736	603	4.0419	488	7.9592	609

Table 1. Applying the mean shift valley method to decompose data.

There is one exceptional case: when there are no local valleys (e.g., uni-modal), the mean shift valley method is divergent. This can easily be avoided by terminating when no samples fall within the window.

3 Gray-Level Image Histogram and Mean Shift

If we are segmenting a gray-level image, the mean-shift equations can be rewritten as functions on the image intensity histogram:

$$\hat{f}(x) = \frac{d+2}{2nh^{d+2}c_d} \sum_{t_i \in S_h(x)} H(t_i)\left(h^2 - \|x - t_i\|^2\right) \tag{10}$$

where $H(t_i)$ be the histogram on gray level t_i (t_i is an integer and $0 \le t_i \le 255$).

The kernel density function in equation (10) is related to discrete gray levels $\{t_i \mid t_i \in S_h(x)\}$ and the corresponding histogram $\{H(t_i) \mid t_i \in S_h(x)\}$.

Likewise:

$$\begin{aligned}\hat{\nabla} f(x) &= \frac{1}{n(h^d c_d)} \frac{d+2}{h^2} \left[\sum_{t_i \in S_h(x)} H(t_i)(t_i - x) \right] \\ &= \frac{\sum_{t_i \in S_h(x)} H(t_i)}{n(h^d c_d)} \frac{d+2}{h^2} \left(\frac{\sum_{t_i \in S_h(x)} H(t_i) t_i}{\sum_{t_i \in S_h(x)} H(t_i)} - x \right)\end{aligned} \tag{11}$$

The last term in (11) is called the sample mean shift $M_h(x)$ in discrete gray level space:

$$M_h(x) = \frac{\sum_{t_i \in S_h(x)} H(t_i) t_i}{\sum_{t_i \in S_h(x)} H(t_i)} - x \tag{12}$$

Equation (12) is derived from the Epanechnikov kernel. (Note: reference [15] used a Gaussian kernel - see equation (15) and (17) in that paper).

4 The False Peak Noise

In implementing the mean shift approach in this setting, we found, to our surprise, in some cases there are a lot of peaks appearing between two consecutive gray levels near a local maximum density (see Figure 2(a) and (b)). We call these peaks the false peaks. These false peaks will seriously affect the performance of the mean shift method, i.e. the mean shift is very sensitive to these noise peaks and the mean shift loop will stop at these false peaks instead of a real local maximum density.

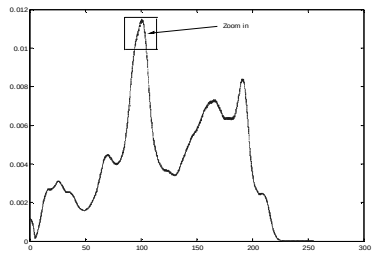
Here we analytically determine the conditions leading to this problem. For simplicity, we choose a one-dimensional setting. Let $\hat{f}(t_k)$ be the kernel density estimate at gray level t_k; let $0 < \delta x < 1$; d=1; and c_d=2. Using equation (10) we have

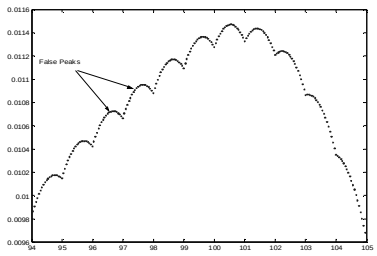
$$\hat{f}(t_k + \delta x) = \frac{3}{4nh^3} \sum_{t_i \in S_h(t_k + \delta x)} H(t_i)\left(h^2 - \left\|t_k + \delta x - t_i\right\|^2\right) \tag{13}$$

If h is an integer (h>0) and t_k+h<255, and considering t_i has to be a series of consecutive unsigned integer, we have $\{t_i \mid t_i \in S_h(t_k + \delta x)\} = \{t_i \mid t_i \in S_h(t_k)\} \Upsilon \{t_i \mid t_i = t_k + h\}$

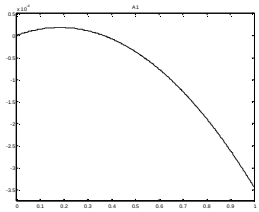
The equation (13) can be rewritten as:

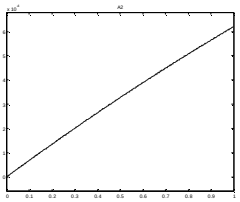
$$\begin{aligned}
\hat{f}(t_k + \delta x) &= \frac{3}{4nh^3}\left[\sum_{t_i \in S_h(t_k)} H(t_i)\left(h^2 - \left\|t_k + \delta x - t_i\right\|^2\right) + H(t_k + h)\left(h^2 - \left\|h - \delta x\right\|^2\right)\right] \\
&= \hat{f}(t_k) + \frac{3}{4nh^3}\left[\sum_{t_i \in S_h(t_k)} H(t_i)\left(-\left\|\delta x\right\|^2 + 2\delta x\left(\frac{\sum_{t_i \in S_h(t_k)} H(t_i) t_i}{\sum_{t_i \in S_h(t_k)} H(t_i)} - t_k\right)\right) + H(t_k + h)\left(-\left\|\delta x\right\|^2 + 2h\,\delta x\right)\right] \\
&= \hat{f}(t_k) + \frac{3}{4nh^3} \sum_{t_i \in S_h(t_k)} H(t_i)\left(-\left\|\delta x\right\|^2 + 2\delta x M_h(t_k)\right) + \frac{3}{4nh^3} H(t_k + h)\left(-\left\|\delta x\right\|^2 + 2h\,\delta x\right)
\end{aligned} \tag{14}$$

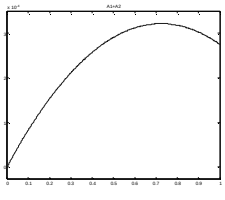




(a) (b)







(c) (d) (e)

Figure 2. False peak noise. (a) Original probability density distribution with h equal to 5; (b) Zoom in a part of (a). Many false peaks introduced by A1+A2 in Eq. (15); (c)-(e) A1, A2, and A1+A2 in Eq. (15) with t_k=95.

We let:

$$A2 = \frac{3}{4nh^3} H(t_k + h)(-\|\delta x\|^2 + 2h\delta x) \tag{15a}$$

When h>> δx , *A2* can be approximated as a linear equation (see Figure 2(d)).

Equation (14) can be rewritten as:

$$\hat{f}(t_k + \delta x) = \hat{f}(t_k) + A1 + A2 \tag{15b}$$

Now we calculate the differential of $\hat{f}(t_k + \delta x)$:

$$\frac{d\hat{f}(t_k + \delta x)}{d(\delta x)} = \frac{3}{4nh^3} \sum_{t_i \in S_h(t_k)} H(t_i)(-2\delta x + 2M_h(t_k)) + \frac{3}{4nh^3} H(t_k + h)(-2\delta x + 2h)$$

$$= \frac{3}{4nh^3}\left[-2\delta x\left(\sum_{t_i \in S_h(t_k)} H(t_i) + H(t_k + h)\right) + 2\left(\sum_{t_i \in S_h(t_k)} H(t_i)M_h(t_k) + H(t_k + h)h\right)\right] \tag{16}$$

Let (16) equal zero, we obtain:

$$\delta x = \frac{\sum_{t_i \in S_h(t_k)} H(t_i)M_h(t_k) + H(t_k + h)h}{\sum_{t_i \in S_h(t_k)} H(t_i) + H(t_k + h)} \tag{17}$$

Substituting equation (12) into equation (17), and if $0 < \delta x < 1$, i.e. if:

$$\frac{\sum_{t_i \in S_h(t_k)} H(t_i)(t_k - t_i)}{H(t_k + h)} < h < \frac{\sum_{t_i \in S_h(t_k)} H(t_i)(t_k - t_i) + \sum_{t_i \in S_h(t_k)} H(t_i) + H(t_k + h)}{H(t_k + h)} \tag{18}$$

there will be a false peak appearing between two consecutive gray level, t_k and t_k+1.

In Figure 2, when we apply the mean shift with initial location at 95, we find it stopped at 95.7244, instead of the real local maximum density at 101. From (17), we obtained δx =0.7244, i.e. there is a false peak between 95 and 96.

We let L be the left item in the in equation (18) and R be the right item of (18); let $x_{MS(tk)}$ be the convergent point, obtained by the mean shift method with initial point at t_k, corresponding to the local peak. Thus if the condition:L<h<R is satisfied, we can predict that there will be a false peak between t_k and t_k+1 (see Table 2).

h	L	R	δx	t_k	$x_{MS(tk)}$	False peak between t_k and t_k+1
5	-1.45	7.45	0.72	95	95.72	yes
6	-2.84	6.99	0.89	95	95.89	yes
7	-5.76	6.06	1.08	95	96.96	no
7	-5.10	7.45	0.96	96	96.96	yes
8	-9.68	3.95	1.30	95	97.94	no
8	-7.99	6.19	1.13	96	97.94	no
8	-7.42	8.96	0.94	97	97.94	yes

Table 2. False peaks prediction

The above analysis suggests that one could devise an approach that adaptively adjusts h depending upon whether false peaks are predicted. If a false peak is detected, we can use the following adjustment to avoid the influence of the false peak:

$$y_{k+1} = y_k + \text{ceil}(M_h(y_k)) \text{ for MS step} \quad (19a)$$

$$y_{k+1} = y_k + \text{floor}(MV_h(y_k)) \text{ for MSV step} \quad (19b)$$

5 An Unsupervised Algorithm for Image Segmentation

Consider the peaks {P(i)} and valleys{V(i)}. V(0)=0 and V(n)=255. $V(0) \le P(1) < V(1) < .... < P(n) \le V(n)$. The algorithm is described as follows:

(1) Initialise the bandwidth h and the location of search window.
(2) Apply the MS method to obtain peak P_k with the initial window location V_{k-1}+1.
(3) Apply the MSV method to obtain valley V_k with initial window location P_k+1.
(4) Repeat step (2) and (3) until P_k or V_k is equal to or larger than 255. The questions remains as to how many of these peaks are significant. We post-process by step (5).
(5) Validate peaks and valleys
  (5a) Remove peaks too small compared with the largest.
  (5b) Remove the smaller of two consecutive peaks if too close.
  (5c) Calculate the normalized contrast [1] for a valley and two neighboring peaks:

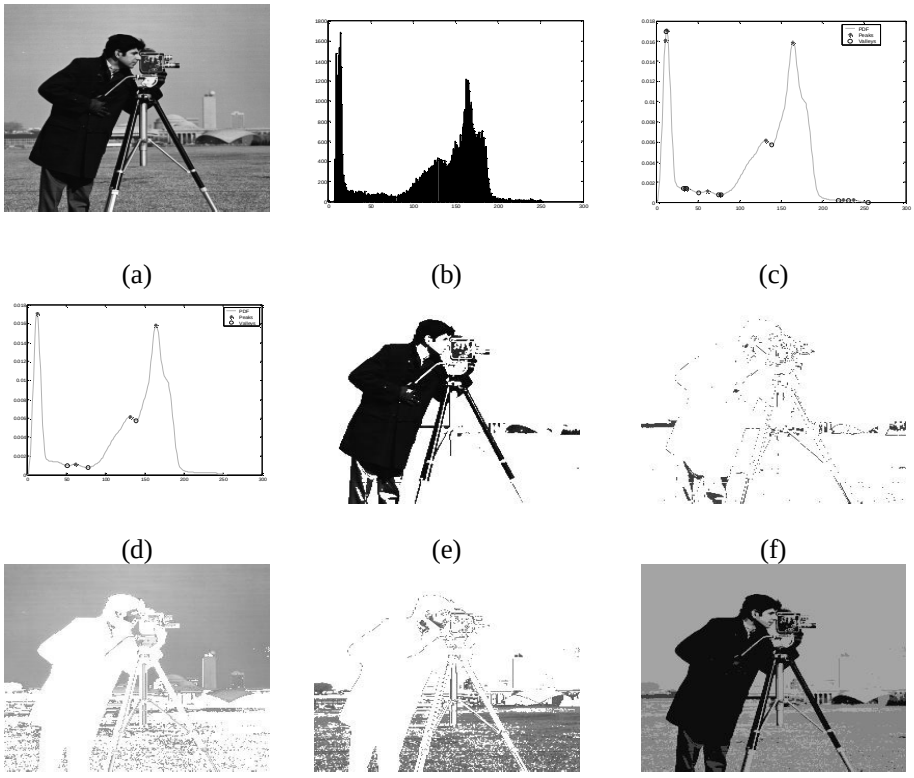
$$\text{Normalized Contrast} = \frac{Contrast}{Height} \quad (20)$$

where the contrast is the difference between the smaller peak and the valley. Remove the smaller one of the two peaks if this is small.

After step 5(a)-5(c), we obtain several significant peaks {PS(1),…PS(k)}. The valleys then are chosen as the minimum of the valleys between two consecutive significant peaks. Thus we have k-1 valleys {VS(1),…VS(k-1)}.
(6) Using the obtained valleys finally obtain k segmented images by {[0, VS(1)], [VS(1), VS(2)], … [VS(k), 255]}.

6 Experimental Results

In this section, we will use several examples to show the performance of the proposed method in segmenting images. Figure 3 demonstrates the segmentation procedures of the proposed method. Figure 3 (c)/(d) shows the obtained peaks and valleys before/after validation . Before the validation, there are ten peaks and ten valleys obtained. Near a local plateau, there will be some insignificant peaks and valleys. After applying step 5 in section 5, we finally obtained three validated valleys and thus we have four segmented images Figure 3 (e-h). The final result is shown in (i). Figure 4 hows another experiment on a x-ray medical image. From Figure 4, we can see that the x-ray image has been successfully segmented: the background (Figure 4 (c)), the bone (Figure 4 (d)), and the tissues (Figure 4 (e)) were extracted separately.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3. The segmentation results of the proposed method (h=7). (a) original image of the cameraman; (b) gray-level histogram; (c) peaks and valleys of $\hat{f}(x)$ before merging; (d) final peaks and valleys; (e)-(h) the resulting segmented images; (i) the final segmented image.

The computational speed of the proposed algorithm is efficient: about 0.27 second using MATLAB code on an AMD 800MHz personal computer.

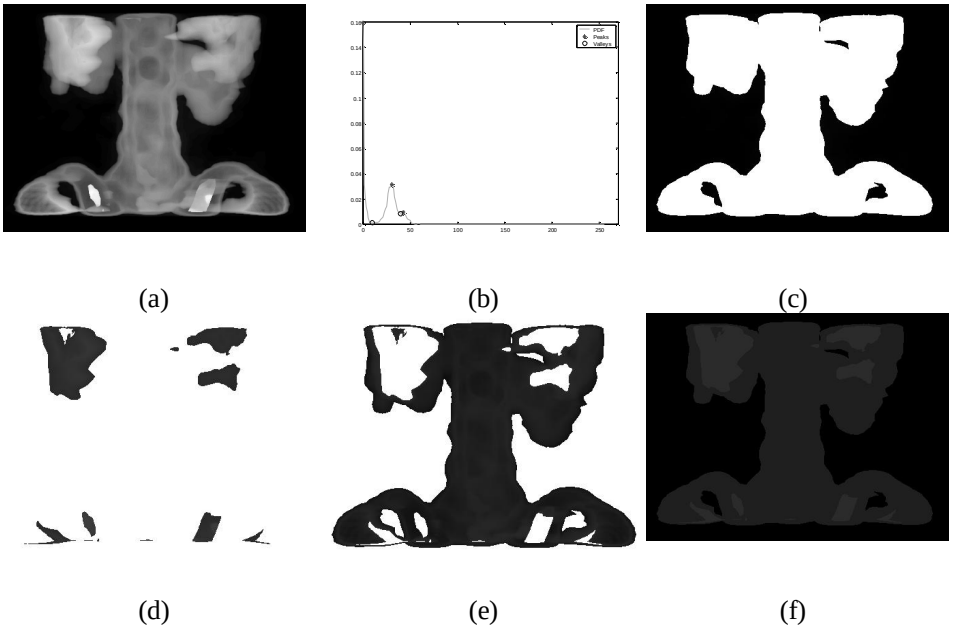


Figure 4. The application of the proposed method on medical images (h=2). (a) the original x-ray image; (b) the final peaks and valleys after validation; (c)-(e) the resulting segmented images; (i) the final segmented image.

7 Conclusion

The mean shift (MS) method is well known and popular: yet its sensitivity to false local peaks is virtually unrecognized. In this paper, we analyze the influence of false peak noise on the MS and MSV method, in particular, we show how the occurrence of the false peaks is related to the bandwidth of the kernel density estimator. We then provide an algorithm to avoid the false peak problem in image segmentation.

As Comaniciu, etc. [8] pointed out, because the MS iterations (and the saddle point iterations) converge to a point with zero gradient, a test for local maximum (or local minimum) is needed by checking on either the eigenvalues of the Hessian matrix of second derivatives, or on the stability of the converged point through perturbing the converged point by a random vector of small norm. Due to its ability to predict false peaks, the false peak noise theory provides a third possible way to test the local maximum (peaks) and local minimum (valleys).

A novel unsupervised peak-valley sliding algorithm for image segmentation is also presented in this paper. We use the MS method to find peaks and the MSV method to find valleys. The peaks and valleys are alternatively found one by one. After validating the obtained peaks and valleys, we use the validated valleys as density thresholds to segment the image.

Generally speaking, if h is large, the details will be smoothed and the image will be under-segmented; on the other hand, if h is chosen too small, there will be a lot of noise (including peak noise and valley noise) and the image will be over-segmented. However, the theory of false peak shows there are exceptional cases existing: false peaks appear only when equation (18) is satisfied.

References:

1. Albiol, A., L. Torrest, and E.J. Delp. An Unsupervised Color Image Segmentation Algorithm for Face Detection Applications. in Proceedings of International Conference on Image Processing. 2001: p. 681-684.
2. Cheng, H.D. and Y. Sun, A Hierarchical Approach to Color Image Segmentation Using Homogeneity. IEEE Trans. Image Processing, 2000. **9**(12): p. 2071-2082.
3. Cheng, Y., Mean Shift, Mode Seeking, and Clustering. IEEE Trans. Pattern Analysis and Machine Intelligence, 1995. **17**(8): p. 790-799.
4. Comaniciu, D. and P. Meer, Robust Analysis of Feature Spaces: Color Image Segmentation. in Proceedings of 1997 IEEE Conference on Computer Vision and Pattern Recognition, San Juan, PR, 1997: p. 750-755.
5. Comaniciu, D. and P. Meer, Mean Shift Analysis and Applications. in Proceedings 7th International Conference on Computer Vision, Kerkyra, Greece, 1999: p. 1197-1203.
6. Comaniciu, D. and P. Meer, Distribution Free Decomposition of Multivariate Data. Pattern Analysis and Applications, 1999. **2**: p. 22-30.
7. Comaniciu, D. and P. Meer, Mean Shift: A Robust Approach towards Feature Space Analysis. IEEE Trans. Pattern Analysis and Machine Intelligence, 2002. **24**(5): p. 603-619.
8. Comaniciu, D., V. Ramesh, and A.D. Bue. Multivariate Saddle Point Detection for Statistical Clustering. in Europe Conference on Computer Vision (ECCV). 2002. Copenhagen, Danmark: p. 561-576.
9. Comaniciu, D., V. Ramesh, and P. Meer. Real-Time Tracking of Non-Rigid Objects Using Mean Shift. in Proceedings of 2000 IEEE Conference on Computer Vision and Pattern Recognition. 2000: p. 142-149.
10. Fukunaga, K. and L.D. Hostetler, The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. IEEE Trans. Info. Theory, 1975. **IT-21**: p. 32-40.
11. Keselman, Y. and E. Micheli-Tzanakou. Extraction and Characterization of Regions of Interest in Biomedical Images. in Proceedings. IEEE International Conference on Information Technology Applications in Biomedicine. 1998: p. 87-90.
12. Silverman, B.W., Density Estimation for Statistics and Data Analysis. 1986, London: Chapman and Hall.
13. Wakahara, T. and K. Ogura. Extended Mean Shift in Handwriting Clustering. in Proceedings, Fourteenth International Conference on Pattern Recognition. 1998: p. 384-388.
14. Wand, M.P. and M. Jones, Kernel Smoothing. Chapman & Hall, 1995.
15. Yang, X. and J. Liu, Unsupervised Texture Segmentation with One-Step Mean Shift and Boundary Markov Random Fields. Pattern Recognition Letters, 2001. **22**: p. 1073-1081.