14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

ENHANCED SPATIAL-RANGE MEAN SHIFT COLOR IMAGE SEGMENTATION BY USING CONVERGENCE FREQUENCY AND POSITION

Nuan Song, Irene Y. H. Gu, Zhongping Cao, Mats Viberg

Department of Signals and Systems, Chalmers Univ. of Technology, Gothenburg, 41296, Sweden
email: {nuan,zhongpin}@student.chalmers.se, irenegu@chalmers.se, viberg@chalmers.se

ABSTRACT

Mean shift is robust for image segmentation through local mode seeking. However, like most segmentation schemes it suffers from over-segmentation due to the lack of semantic information. This paper proposes an enhanced spatial-range mean shift segmentation approach, where over-segmented regions are reduced by exploiting the positions and frequencies at which mean shift filters converge. Based on our observation that edges are related to spatial positions with low mean shift convergence frequencies, merging of over-segmented regions can be guided away from the perceptually important image edges. Simulations have been performed and results have shown that the proposed scheme is able to reduce the over-segmentation while maintaining sharp region boundaries for semantically important objects.

1. INTRODUCTION

Mean shift for local mode seeking and clustering was initially proposed by Fukunaga and Cheng [10,1], followed by some major development and extensions in [2]. Since then, many new studies and development have been reported on mean shift theories and applications to edge-preserving nonlinear image smoothing and segmentation [3,4,5]. One attraction of mean shift is the statistical basis and its association with the density estimate. Mean shift directly estimates the local modes (maxima) without the requirement of actually estimating the pdf. Mean shift segmentation of images is based on the fact that pixels in the same region share some similar modes. Depending on the selected features, regions with different types of similarity (e.g. intensities, colors, or texture attributes) can be estimated. By including both spatial position and range as features, mean shift takes into account both the geometrical closeness and the photometric similarity of image during image filtering and segmentation. Recent studies have shown that mean shift is related to nonlinear diffusion and bilateral filtering [6,7]. Comparing with a bilateral filter, a mean shift filter is more robust since the filter uses different sets of data during the mean shift iterations due to the changes of kernel spatial positions. Despite all these attractive properties, mean shift segmented images suffer from over-segmentation due to the lack of semantic information, a common phenomenon in most image segmentation methods based on low level image features. It also requires a careful selection of bandwidth parameters. A fixed bandwidth value for the entire image may not be suitable depending on the image property. Also, mean shift segmentation usually requires a post-processing for merging small regions. Motivated by the above, we proposed an enhanced spatial-range mean shift by combining mean shift convergence frequencies and positions. From the associations of converging frequencies and edges/homogeneous areas, merging of over-segmented regions can be guided away from the perceptually important image edges. This can lead to reduced over-segmentation, but still retaining semantic meaningful regions. Further, the selection of bandwidths in mean shift may become less sensitive by allowing an initial over-segmentation before a refined processing.

2. SYSTEM DESCRIPTION

The proposed scheme combines spatial-range mean shift with converging frequencies and positions for a refined image segmentation. As shown in Fig.1, the segmentation scheme consists of four main blocks. First, a spatial-range mean shift filter is applied to the feature vectors extracted from the image **I(s)**, **s** is the pixel position (block-1). The frequencies and positions to which mean shift filtered pixels converge are recorded in a map (block-2). Pre-segmentation is performed based on clustering similar modes in the filtered image (block-3). Refined segmentation uses the map which guides merging of over-segmented regions away from image edges (block-4).

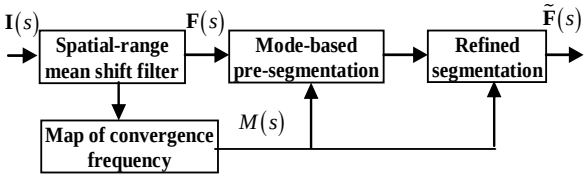

Fig.1. Block diagram of the proposed segmentation scheme

3. SPATIAL-RANGE MEAN SHIFT FILTERING

In this section, we briefly review the spatial-range mean shift filter that is used as the basis of image segmentation in this paper. Let a given set of L-dimensional feature vectors $S = \{ \mathbf{x}_i,\ i = 1, \cdots, n \}$ be given for estimating the kernel density of a feature vector **x**,

$$\hat{p}_K(\mathbf{x}) = \frac{c_k}{nh^L}\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right) \quad (1)$$

where $K(\mathbf{x}) = c_k k(\|\mathbf{x}\|^2)$ is a radially symmetric kernel with L_2 distance measure, and c_k is a normalization constant. The local modes (maxima) of $\hat{p}_K(\mathbf{x})$ can be obtained by setting $\nabla\hat{p}_K(\mathbf{x}) = 0$, leading to $\frac{1}{2}h^2 c\frac{\nabla\hat{p}_K(\mathbf{x})}{\hat{p}_G(\mathbf{x})} = m_G(\mathbf{x})$, where K in (1) is the shadow kernel of G, $G(\mathbf{x}) = c_g g(\|\mathbf{x}\|^2)$, and $g(x) = -k'(x)$, and $c = \frac{c_g}{c_k}$ is a constant, and $m_G(\mathbf{x})$ is the mean shift defined as:

$$m_G(\mathbf{x}) = \frac{\sum_{i=1}^{n}\mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x} \quad (2)$$

For a spatial-range mean shift filter, the feature vector is defined as $\mathbf{x} = [\mathbf{x}^d \quad \mathbf{x}^r]^T$, where the 1st component vector is the *domain feature* defined as the spatial position of pixel $\mathbf{x}^d = \mathbf{S} = [s_x \quad s_y]^T$, and the 2nd component feature is the *range feature* defined as the function $\mathbf{f}(\mathbf{x}^d)$ of domain feature, frequently set as the image intensity $\mathbf{x}^r = \mathbf{I}(\mathbf{x}^d)$. For a Gaussian kernel, $g(\mathbf{x})$ in (2) becomes,

$$g(\|\mathbf{x}\|^2) = \exp\left(-\frac{\|\mathbf{x}^d\|^2}{2\sigma_d^2}\right)\exp\left(-\frac{\|\mathbf{I}(\mathbf{x}^d)\|^2}{2\sigma_r^2}\right) \quad (3)$$

Or, $g(\mathbf{x}) = g_d(\mathbf{x}^d)g_r(\mathbf{I}(\mathbf{x}^d))$, where $h_d = 2\sigma_d^2$, $h_r = 2\sigma_r^2$ are the spatial and range kernel bandwidths, respectively. We only consider Gaussian kernels in this paper. This is because a Gaussian kernel is known to yield a better segmentation after convergence, as compared to using a Epanechnikov kernel [2]. Further, the shadow kernel of Gaussian is also a Gaussian kernel. Such a mean shift filter can be used as a nonlinear edge-preserving smoothing: when the differences of pixel intensities are small, the mean shift filter acts as a lowpass filter in a local image region. However, if the intensity differences are large (e.g. around edges), then the range filter kernel is close to zero value, hence no filtering is actually applied to these pixels. In such a way, a joint spatial-range mean shift filter takes into account the geometrical closeness as well as the photometric similarity in an image.

4. ENHANCED SEGMENTATION

4.1 Frequency and position where mean shift converge

The basic idea is to exploit the information (from the mean shift filter) related to image edges and homogeneous areas to guide the merging of over-segmented image regions. Since the mean shift vector is proportional to the normalized density gradient and is pointing towards the direction of maximum increase in the density (or, towards the local mode), a large number of mean shift filtered pixels is expect to converge near to (but not on) the sharp edges. Conversely, a low frequency of convergence to a pixel may correspond to a point in a flat area. This phenomenon can be confirmed by simulation results shown in Fig.1. The map **M**(**s**) in the figure shows how frequent the mean shift converges towards each pixel. For example, **M**(**s**)=0 means zero pixels converge to s, which implies discontinuities (edges) in s; **M**(**s**)=1 means that only one pixel converges to s, implies homogenous areas; while **M**(**s**)>i means that i pixels converge to s, a mode candidate for a homogenous area surrounded by image edges. In the map image, a dark pixel indicates a high convergence number. One can observe that pixels with high frequency of convergence (black) are indeed concentrated next to image edges, points with zero number convergence (white) are associated with image edges, and the main parts are the homogeneous areas which have median frequency of convergence (gray). Based on the above observations, enhancing the segmentation from mean shift can be obtained by exploring positions and frequencies of mean shift convergence. This will lead to a reduced number of over-segmented regions, resulting regions more close to semantic meaningful objects indicated by edge curves.

Fig.1. Map image **M**(**s**) shows the mean shift convergence frequency at each pixel position. (white: indicate edges, black: indicate nearer to edges; gray: indicate smooth areas. Left: from image "peppers", right: from "swimming lady").

4.2 Pre-Segmentation using Spatial-Range Mean Shift

To obtain pre-segmented image, a joint spatial-range mean shift filter, with pre-specified bandwidths h_d and h_r, is applied. The algorithm is summarized in Table 1. Meanwhile, *the frequency of convergence map* **M**(**s**) is generated during the mean shift filter process as follows: assuming that mean shift filtering of $\mathbf{x}_i = [\mathbf{s}_i \quad \mathbf{F}(\mathbf{s}_i)]^T$ converges to $\mathbf{y}_{i,c} = [\mathbf{y}_{i,c}^d \quad \mathbf{y}_{i,c}^r]^T$, where $\mathbf{y}_{i,c}^d$ is the converged position for $\mathbf{s}_i$. If $\mathbf{y}_{i,c}^d = \mathbf{s}_k$ (i.e., the converged domain feature is equal to a pixel position coordinator $\mathbf{S}_k$) then $\mathbf{M}(\mathbf{s}_k)$ value is increased by one, i.e, $\mathbf{M}(\mathbf{s}_k) \leftarrow \mathbf{M}(\mathbf{s}_k) + 1$. This process continues for all pixels in

the image. This step for forming up M(s) can be added in between Steps 5 and 6 of Table 1.

Afterwards, pre-segmentation is applied to the filtered image **F**(**s**), based on merging connected pixels into a region whose range difference is smaller than a range bandwidth. However, the segmentation is only applied to those filtered image pixels **F**(**s**) satisfying **M**(**s**)>1. That is, a set of *connected* pixels belonging to one candidate region satisfy,

$$S=\left\{\mathbf{s}_l,\mathbf{s}_m \mid \left\|\mathbf{F}(\mathbf{s}_l)-\mathbf{F}(\mathbf{s}_m)\right\|\le h_{r_2};\mathbf{M}(\mathbf{s}_l)>1,\mathbf{M}(\mathbf{s}_m)>1\right\} \quad (4)$$

The i-th candidate region is described by 2 components, the mode value and the set of pixels, as below

$$R_i=\left\{\bar{F}_i=\frac{1}{n_i}\sum_{\mathbf{s}_k\in S}\mathbf{F}(\mathbf{s}_k),\ \{\mathbf{s}_k\}\right\} \quad (5)$$

where n_i is the total number of pixels in the region i. Segmented candidate region boundaries are then drawn, where the corresponding boundary map is assigned as:

$$\mathbf{B}(\mathbf{s})=\begin{cases} i & \text{if } \mathbf{s} \text{ is a boundary point of region } i \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

In such a way, an initial over-segmented image is generated.

Initialize: F(s)= I(s).

1. Given $\mathbf{s}_i$, form $\mathbf{x}_i=[\mathbf{s}_i \ \ \mathbf{F}(\mathbf{s}_i)]^T$, set j=1 and $\mathbf{y}_{i,1}\leftarrow\mathbf{x}_i$.

2. Form a new set $S=\{\mathbf{x}_k,\ \mathrm{k}=1,\cdots,n\}$ from the image centered at $\mathbf{s}_i$ and within domain bandwidth h_d.
Compute the new center:

$$\mathbf{y}_{i,j+1}=\frac{\sum_{k=1}^{m}\mathbf{x}_k\, g\left(\left\|\frac{\mathbf{y}_{i,j}^s-\mathbf{s}_k}{h_d}\right\|^2\right) g\left(\left\|\frac{\mathbf{y}_{i,j}^r-\mathbf{F}(\mathbf{s}_k)}{h_r}\right\|^2\right)}{\sum_{k=1}^{m} g\left(\left\|\frac{\mathbf{y}_{i,j}^s-\mathbf{s}_k}{h_d}\right\|^2\right) g\left(\left\|\frac{\mathbf{y}_{i,j}^r-\mathbf{F}(\mathbf{s}_k)}{h_r}\right\|^2\right)}$$

3. Compute the mean shift $\mathbf{m}_G(\mathbf{y}_{i,j})=\mathbf{y}_{i,j+1}-\mathbf{y}_{i,j}$

4. If $\left\|\mathbf{m}_G(\mathbf{y}_{i,j})\right\|<\varepsilon$ (e.g. $\varepsilon=0.01$), go to Step 5,
Otherwise: $j\leftarrow j+1$, and repeat Steps 2-3.

5. Set the converged value $\mathbf{y}_{i,c}\leftarrow\mathbf{y}_{i,j+1}$, assign filtered value as converged range vector $\mathbf{F}(\mathbf{s}_i)\leftarrow\mathbf{y}_{i,c}^r$.

6. Repeat Steps 1-5 until all pixels in image converges.

Table 1. Algorithm for Joint spatial-range mean shift filtering

4.3. Refined-segmentation based on the frequencies and Positions where mean shift filters converge

For reducing the number of over-segmentation, a typical way is to merge small regions as post processing [2,3]. The disadvantage is that some small regions (such as eyes, lips) that are perceptually important may be merged to nearby regions (e.g. face). Also, over-segmentation may create some split regions, each having sufficiently large size. In our proposed method, the map of convergence frequency **M**(**s**) is explored when merging regions. Since the convergence frequencies indicate whether pixels are close to edges, it can be used to refine segments by merging those spurious regions which do not have image edges around.

For each pixel **s** on the candidate region boundary, **B**(**s**)>0, we compute the total number of **M**(**s**)=0 over a small window W,

$$\mathbf{v}(\mathbf{s}_i)=\sum_{s_i\in W} c(\mathbf{M}(s_i)),\quad \text{for all } \mathbf{B}(\mathbf{s}_i)>0 \quad (8)$$

Where *c(x)=1* if *x=0*, otherwise *c(x)=0*. Here using a small window in computing $\mathbf{v}(\mathbf{s}_i)$ is aimed at introducing some resilience to edge noise. If $\mathbf{v}(\mathbf{s}_i)\ge n_0$ (n_0 is a small number) then the pixel $\mathbf{s}_i$ is considered as being located near an edge which should be retained as a region boundary during candidate region merging process, otherwise it is a boundary point of a spurious region which should be merged, i.e.,

$$\begin{cases} \mathbf{B}_0(\mathbf{s}_i)=\mathbf{B}(\mathbf{s}_i), & \text{if } \mathbf{v}(\mathbf{s}_i)<n_0 \Rightarrow \text{spurious boundary} \\ \mathbf{B}_1(\mathbf{s}_i)=\mathbf{B}(\mathbf{s}_i), & \text{if } \mathbf{v}(\mathbf{s}_i)\ge n_0 \Rightarrow \text{near edge} \end{cases} \quad (8)$$

where $\mathbf{B}_0(\mathbf{s})>0$ contains the boundaries of spurious regions, its (non-zero) value denotes the candidate region index. For each spurious region, it is merged to one of the neighbor regions with which it shares the longest common border and their range difference is below the threshold T_r. That is, for a spurious region i, and its neighboring regions j, j=1,2…, the length of their common border $len_{bd}(i,j)$ is computed. Then the neighboring region j^* having the longest common border is picked up,

$$j^*=\arg\max_j\left\{len_{bd}(i,j)\right\} \quad (9)$$

1. Spatial-range mean shift filtering:
For each pixel **s** in the original image **I**(**s**), do:
1.1. Apply spatial-range mean shift **x**=[**s** **I**(**s**)] to generate filtered **F**(**s**) (see Section 4.2);
1.2. Generate a converging frequency map **M**(**s**) (see Section 4.2) .
End;
2. Pre-segmentation:
For each pixel **s**, **M**(**s**)>1, do
2.1. Generate candidate regions using (4) and (5)
2.2 Draw region boundaries in the map image **B**(**s**).
End;
3. Refined-segmentation:
3.1. Split boundary map **B**(**s**) to $\mathbf{B}_0(\mathbf{s})$ and $\mathbf{B}_1(\mathbf{s})$ using (8).
3.2. For each region indicated by $\mathbf{B}_0(\mathbf{s})>0$, merge a region to its neighbour region (see Section 4.3, (8) and (9)).
3.3. Update mode and draw new boundary for the region.
3.4. Repeat Steps 3.2 and 3.3 until all $\mathbf{B}_0(\mathbf{s})=0$.

Table 2. Pseudo algorithm of the proposed method.

Regions i and j* are then merged if they satisfy

$$\left\| \overline{F}_i - \overline{F}_{j^*} \right\| < T_r \quad (10)$$

Otherwise, check the neighbor region with the 2nd longest common boarder that satisfies (11), until either a merge can be done, or all neighbor regions are exhausted. Finally, the mode for the merged region is assigned as the averaging mode of these two regions weighted by their sizes. The merging process is continued until all spurious regions indicated by $\mathbf{B}_0(\mathbf{s}) > 0$ are merged. Table 2 summarizes the pseudo algorithm of the proposed segmentation method.

5. SIMULATIONS AND RESULTS

Simulations were conducted for a variety of 2D color images using the proposed method. Fig.2 includes five segmented images with different complexity. The bandwidth parameters used in the simulations are described in columns 2-4 of Table 3. One can see from the results (the 2nd row in Fig.2) that the convergence frequency maps clearly indicated image edges and areas next to edges, as well as smooth areas. One can also see from the results (the 4th row in Fig.2) that there is a significant reduction in spurious regions after refined segmentation, meanwhile sharp image edges are maintained. Further, the refined segmented regions are more related to semantically meaningful objects. It is also noticed that the final segmentation results are not very sensitive to the initial over-segmentation. Since the conventional mean shift segmentation is known to be sensitive to the kernel bandwidth selection [11,12], the proposed method means less sensitive and less demand to tuning the bandwidth parameter values. Also, in the proposed scheme, there is no post-processing for (blindly) merging small regions to their large neighbouring regions as in [2,4].

For further measuring the performance of the segmented images, two objective performance measures were applied as a partial indication of performance. One is the uniformity measure [8] defined as

$$U = 1 - \sum_j \left(P_j \ \sigma_j^2 / \sigma_{\max}^2 \right) \quad (11)$$

where $\sigma_{\max}^2$ is the maximum variance for all regions, σ_j^2 and P_j are the variance and weighting factor associated with the region j. A larger U value (maximum value 1.0) indicates a more homogeneous segmented image. Another measure is the evaluation function [9] defined as

$$E = \sqrt{N} \sum_{j=1}^{N} \left(e_j^2 / \sqrt{N_j} \right) \quad (12)$$

where e_j^2 is the sum of Euclidean distance between the original and segmented image pixels in the region j, N_j is the number of pixels in the j-th region, and N is the total number of regions. A smaller E indicates a better performance. Table.3 (columns 5-7) shows the performance measured from using these 2 criteria. One can observe that the uniformity measure is close to 1.0, indicating that the segmented images are rather homogenous (although this measure does not penalize small region sizes). The results of evaluation function E have indicated an improved performance after the refined segmentation.

6. CONCLUSIONS

A novel statistical-based segmentation method is proposed by combining joint spatial-range mean shift and edge-guided merging of over-segmented regions through the use of frequency of mean shift convergence and their positions. Our simulations results and preliminary performance evaluation have shown that the proposed method has provided enhanced results with reduced over-segmentation meanwhile retaining sharp image edges. Further study will be conducted on evaluations and comparisons.

REFERENCES

[1] Y. Cheng, “Mean shift, mode seeking, and clustering", IEEE Trans. PAMI, Vol. 17, No. 8, pp. 790-799, 1995.
[2] D. Comaniciu, P.Meer, “Mean shift: a robust approach toward feature space analysis”. IEEE Trans. PAMI, 603-619, 2002.
[3] D. DeMenthon, R. Megret, “Spatial-Temporal Segmentation of Video by Hierarchical Mean Shift Analysis”, in Proc. of Statistical Methods in Video Processing Workshop, Denmark, 2002.
[4] I.Y.H.Gu and V. Gui, Chapter 6: "Joint space-time-range mean shift-based image and video segmentation", in Advances in Image and Video Segmentation, edited by Y-J. Zhang. Idea Group Publishing, 2006.
[5] J.Wang, B.Thiesson, Y. Xu, M. Cohen, “Image and Video Segmentation by Anisotropic Kernel Mean Shift”, in Proc. of ECCV Conf, 2004.
[6] D. Barash, “A Fundamental Relationship between Bilateral Filtering, Adaptive Smoothing and the Nonlinear Diffusion Equation”, IEEE Trans. PAMI, Vol.24, No.6, pp.844-847, 2002.
[7] C.Tomasi, R.Manduchi, “Bilateral Filtering for Gray and Color Images”, Proc. IEEE International Conf. ICCV, India, 1998.
[8] J. Liu and Y.H. Yang, ‘Multi-resolution color image segmentation’. IEEE trans. PAMI, Vol.16, pp.689-700, 1994.
[9] MD Levine, AM Nazif, "dynamic measurement of computer generated image segmentation," IEEE Trans. PAMI, vol.7, pp.155–164,1985.
[10] K. Fukunaga, L. Hostetler, “The estimation of the gradient of a density function, with applications in pattern recognition”, IEEE Trans. Information Theory, Vol. 21, pp.32-40, 1975.
[11] D. Comaniciu, “An algorithm for data-driven bandwidth selection”, IEEE Trans. PAMI, Vol.25, No.2, pp.281-288, 2003.
[12] D. DeMenthon, R. Megret, “The variable bandwidth mean shift and data-driven scale selection”, Proc. IEEE 8th Int. Conf. on Computer Vision, Canada, pp.438-445, 2001.

Image	h_r	h_d	h_{r_2}	E function		Uniformity U
				Pre-Segment.	Refined-segment	
Swim lady	0.15	5	0.13	422.5	294.7	0.9928
Peppers	0.2	5	0.13	2480.1	1775.7	0.9851
Tree	0.1	5	0.15	2266.2	1989.5	0.9919
house	0.1	5	0.15	1129.7	1068.0	0.9864
hall	0.1	5	0.12	809.3	784.3	0.9814

Table 3. Parameters used for images in Fig.2 (columns 2-4) and segmentation performance measures (column 5-6).

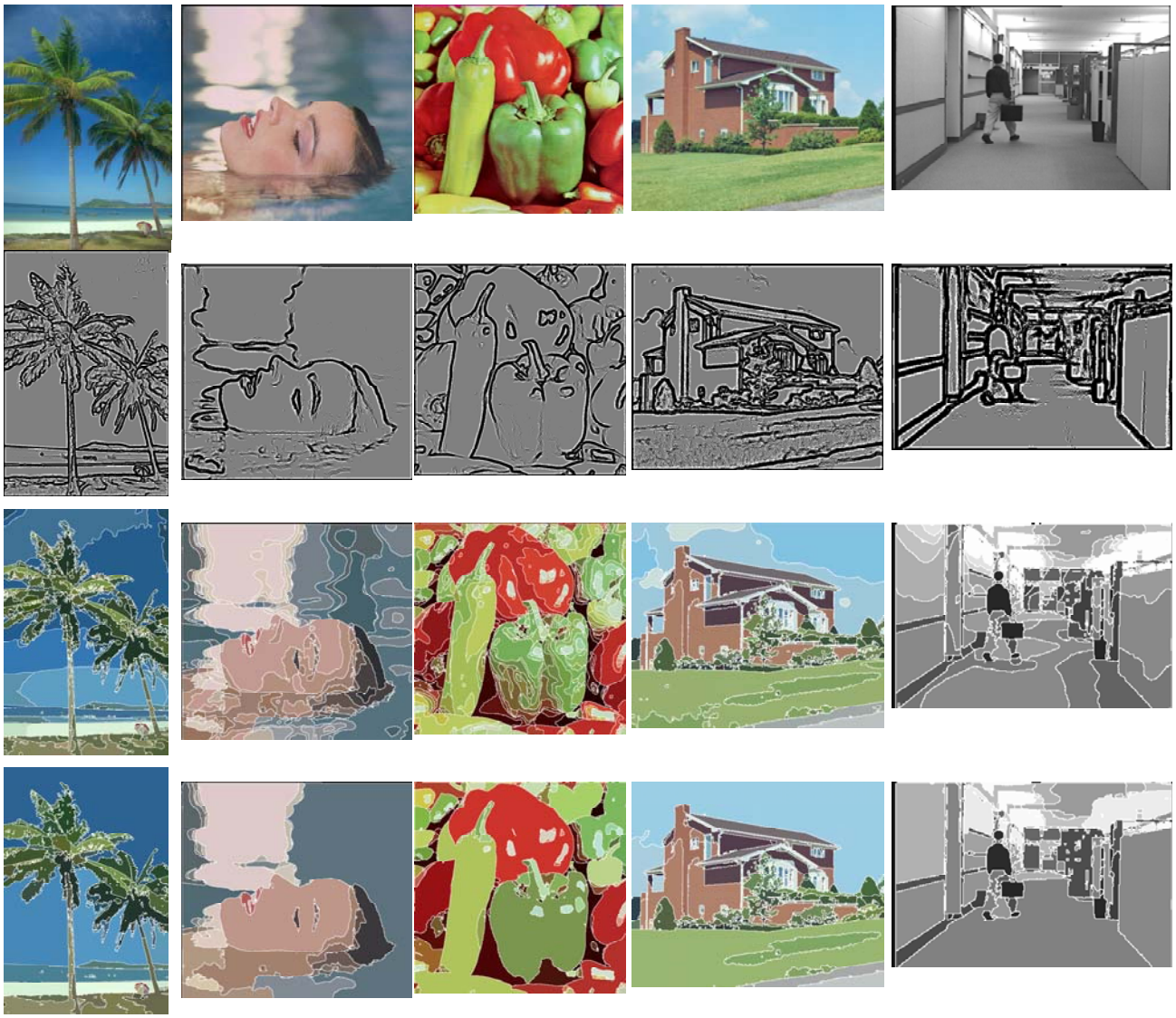

Fig.2. Segmentation results from the proposed method. 1st row: original images (left to right: tree, swimming lady, peppers, house, hall) 2nd row: the map images M(s); 3rd row: pre-segmented results; 4th row: refined segmented results.