Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Segmenting Brain MRI using Adaptive Mean Shift

Juan Ramón Jiménez-Alaniz, Mauricio Pohl-Alfaro, Verónica Medina-Bañuelos, and Oscar Yáñez-Suárez

***Abstract*— To delineate arbitrarily shaped clusters in a complex multimodal feature space, such as the brain MRI intensity space, often requires kernel estimation techniques with locally adaptive bandwidths, such as the adaptive mean shift procedure. Proper selection of the kernel bandwidth is a critical step for a better quality in the clustering. This paper presents a solution for the bandwidth selection, which is completely nonparametric and is based on the sample point estimator to yield a spatial pattern of local bandwidths. The method was applied to synthetic brain images, showing a high performance even in the presence of varying noise level and bias.**

I. INTRODUCTION

Statistical segmentation techniques usually define a parametric model representing the tissue, assuming particular distribution forms on the selected feature space. This assumption can introduce artifacts implied by the density model choice. On the other hand, the nonparametric methods do not have embedded assumptions; the motivation to use a nonparametric approach is to let the data guide a search for the function which fits them best without the restrictions imposed by a parametric model.

A clustering technique which does not require prior knowledge of the number of clusters, and does not constrain their shape, is built upon the mean shift. This is an iterative technique which estimates the modes of the multivariate underlying distribution of the feature space; this modes are considered the centers of the densest regions in the space. The Mean Shift algorithm has been successfully used for image segmentation, particularly in brain MR images [1].

One of the limitations of the mean shift procedure is that it involves the specification of a parameter named the kernel width or *bandwidth*. While results obtained appear satisfactory, when the local characteristics of the feature space differs significantly across data, it is difficult to find an optimal global bandwidth for the mean shift procedure. For this reason, in this paper a locally adaptive bandwidth is employed. The technique estimates a bandwidth for each data point, based in a rough estimation, and then estimates the modes. The solution is tested with synthetic brain data.

The paper is organized as follows. The limitations of the fixed bandwidth density estimation are reviewed, and the sample point estimator is introduced in Section 2. Section 3 presents the criterion for bandwidth selection and the adaptive mean shift procedure. In Section 4, the variable bandwidth procedure is applied to clustering feature spaces. Finally, conclusions are presented in Section 5.

J. R. Jiménez-Alaniz, Mauricio Pohl-Alfaro, V. Medina-Bañuelos, O. Yáñez-Suárez are with the Neuroimaging Laboratory, Department of Electrical Engineering, Universidad Autónoma Metropolitana - Iztapalapa, Av. San Rafael Atlixco 186, Col. Vicentina, México, D. F., 09340, México. E-mail: {jajr, vera, yaso}@xanum.uam.mx

II. BANDWIDTH ESTIMATION

A. Fixed bandwidth density estimation

The multivariate kernel density estimator [2, p. 76] with kernel *K* and window width *h* is defined by

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

The kernel function $K(\mathbf{x})$ is a function defined for d-dimensional vectors $\mathbf{X}_i$, $i = 1, ..., n$ that are the given multivariate data set whose underlying density f is unknown and should be estimated. The kernel is taken to be a radially symmetric, non-negative function centered at zero and integrating to one, for example the multivariate Epanechnikov kernel

$$K(\mathbf{x}) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-\mathbf{x}^\mathbf{T}\mathbf{x}) & \text{if } \mathbf{x}^\mathbf{T}\mathbf{x} < \mathbf{1} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where c_d is the volume of the unit d-dimensional sphere.

The terminology *fixed bandwidth* is due to the fact that *h* is held constant across $\mathbf{x} \in R^d$. The use of a single smoothing parameter *h* in (1) implies that the version of the kernel placed on each data point $\mathbf{x}$ is scaled equally in all directions. As a result, the fixed bandwidth procedure estimates the density by taking the average of identically scaled kernels.

The most widely used way of placing a measure on the global accuracy of $\hat{f}$ as an estimator of f is the mean integrated square error (MISE) defined by

$$\begin{aligned} MISE(\hat{f}) &= \int E\{\hat{f}(\mathbf{x}) - f(\mathbf{x})\}^2 d\mathbf{x} \\ &= \int \{E\hat{f}(\mathbf{x}) - f(\mathbf{x})\}^2 d\mathbf{x} + \int \text{var } \hat{f}(\mathbf{x}) d\mathbf{x} \\ &= \left\{\text{bias}\left(\hat{f}(\mathbf{x})\right)\right\}^2 + \text{var}\left(\hat{f}(\mathbf{x})\right) \end{aligned} \tag{3}$$

Using the multidimensional form of Taylor's theorem, the bias and the variance are approximated by [2, p. 85]

$$\text{bias}(\mathbf{x}) \approx \frac{1}{2}h^2\alpha\nabla^2 f(\mathbf{x}) \tag{4}$$

$$\text{var } \hat{f}(x) \approx n^{-1}h^{-4}\beta f(\mathbf{x}) \tag{5}$$

where the constants $\alpha = \int t_1^2 K(\mathbf{t})d\mathbf{t}$ y $\beta = \int K(\mathbf{t})^2 d\mathbf{t}$ depend on the kernel. Combining (4) and (5) gives the approximate mean integrated square error

$$\frac{1}{4}h^4\alpha^2 \int \left\{\nabla^2 f(\mathbf{x})\right\}^2 d\mathbf{x} + n^{-1}h^{-d}\beta \tag{6}$$

Hence the approximately optimal fixed bandwidth, in the sense of minimizing MISE, is given by

$$h_{opt}^{d+4} = d\beta\alpha^{-2}\left\{\int(\nabla^2 f)^2\right\}^{-1} n^{-1} \tag{7}$$

Comparing approximations (4) and (5) demonstrates one of the fundamental problems of density estimation. If, in an attempt to eliminate the bias, a very small value of *h* is used, then the variance will become large. On the other hand, choosing a large value of *h* will reduce the random variation as quantified by the variance, at the expense of introducing systematic error, or bias, into the estimation. Therefore, the choice of bandwidth parameter implies a tradeoff between random and systematic error.

The value h_{opt} can be substituted back into (6) to give the approximate minimum possible MISE, and hence to guide the choice of kernel. The Epanechnikov kernel is optimum in the sense of minimizing the smallest MISE achievable [2, p. 86]. Nevertheless, the resulting bandwidth formula is of little practical use, since the appropriate value of *h* depends on the unknown density being estimated.

B. Variable bandwidth density estimation

Data-driven bandwidth selection for multivariate data is a complex problem, largely unanswered by the current techniques. The most often used method for local bandwidth adaptation takes the bandwidth proportional to the inverse of the square root of a first approximation of the local density.

The bandwidth *h* can be varied according to each data point, i.e., $h = h(\mathbf{X}_i)$. For each data point $\mathbf{X}_i$, one can obtain the *sample point* density estimator

$$\hat{f}_{sp}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(\mathbf{X}_i)^d}K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h(\mathbf{X}_i)}\right) \tag{8}$$

for which the estimate of f at $\mathbf{x}$ is the average of differently scaled kernels centered at each data point. The sample point estimators are themselves densities, being non-negative and integrating to one. Their most attractive property is that the particular choice of $h(\mathbf{X}_i)$, taken to be reciprocal of the square root of $f(\mathbf{X}_i)$

$$h(\mathbf{X}_i) = h_0\left[\frac{\lambda}{f(\mathbf{X}_i)}\right]^{1/2} \tag{9}$$

reduces considerably the bias. Here, h_0 represents a fixed bandwidth and λ is a proportionality constant.

Since $f(\mathbf{X}_i)$ is unknown it has to be estimated from the data in a first stage. An initial estimate $\tilde{f}$ (called pilot) is used to get a rough idea of the density f; this estimate yields a pattern of bandwidths corresponding to the data. The general strategy used will be as follows:

1) Find a pilot estimate $\tilde{f}(\mathbf{x})$ that satisfies $\tilde{f}(\mathbf{X}_i) > 0$ for all *i*.
2) Define bandwidth factor λ by

$$\log\lambda = n^{-1}\sum\log\tilde{f}(\mathbf{X}_i) \tag{10}$$

In the first step, the construction of the pilot estimate requires the use of another density estimation method. However, the method is insensitive to the fine detail of the pilot estimate, and therefore any convenient estimate can be used. There is no need for the pilot estimate to have any particular smoothness properties; therefore a natural kernel to use in the multivariate case is the Epanechnikov kernel. Using $\tilde{f}$ instead of f in (9), the properties of the sample point estimators remain unchanged [3].

The final estimate is however influenced by the choice of the proportionality constant λ, which divides the range of density values into low and high densities. When the local density is low, i.e., $\hat{f}(\mathbf{X}_i) < \lambda$, $h(\mathbf{X}_i)$ increases relative to h_0 implying more smoothing for the point $\mathbf{X}_i$. For data points that verify $\hat{f}(\mathbf{X}_i) > \lambda$, the bandwidth becomes narrower.

III. ADAPTIVE MEAN SHIFT

The sample point estimator can define an adaptive estimator of the density's normalized gradient, which associates to each data point a differently scaled kernel. The adaptive estimator is an iterative procedure that converges to a local mode of the underlying density.

In the same way as it is introduced in [4], the profile of a kernel *K* is a function $k : [0,\infty) \to R$ such that $K(\mathbf{x}) = k\left(\|\mathbf{x}\|^2\right)$, and if $h_i \equiv h(\mathbf{x}_i)$, the sample point estimator can be written as

$$\hat{f}_K(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i^d}k\left(\left\|\frac{\mathbf{x}-\mathbf{X}_i}{h_i}\right\|^2\right) \tag{11}$$

where the subscript *K* indicates that the estimator depends of kernel *K*.

Taking the estimate of the density gradient as the gradient of the density estimate

$$\begin{aligned}\hat{\nabla}f_K(\mathbf{x}) &\equiv \nabla\hat{f}_K(\mathbf{x}) = \tfrac{2}{n}\textstyle\sum_{i=1}^{n}\frac{\mathbf{x}-\mathbf{X}_i}{h_i^d}k'\left(\left\|\frac{\mathbf{x}-\mathbf{X}_i}{h_i}\right\|^2\right)\\ &= \tfrac{2}{n}\textstyle\sum_{i=1}^{n}\frac{\mathbf{X}_i-\mathbf{x}}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{x}-\mathbf{X}_i}{h_i}\right\|^2\right)\\ &= \tfrac{2}{n}\left[\textstyle\sum_{i=1}^{n}\frac{1}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{x}-\mathbf{X}_i}{h_i}\right\|^2\right)\right]\times\\ &\quad\left[\frac{\sum_{i=1}^{n}\frac{\mathbf{X}_i}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{x}-\mathbf{X}_i}{h_i}\right\|^2\right)}{\sum_{i=1}^{n}\frac{1}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{x}-\mathbf{X}_i}{h_i}\right\|^2\right)}-\mathbf{x}\right]\end{aligned} \tag{12}$$

where $g(x) = -k'(x)$, and it is assumed that the derivative of profile k exists for all $x \in [0,\infty)$, except for a finite set of points.

The last part of (12) contains the adaptive mean shift vector

$$M(\mathbf{x}) = \frac{\sum_{i=1}^{n}\frac{\mathbf{X}_i}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{x}-\mathbf{X}_i}{h_i}\right\|^2\right)}{\sum_{i=1}^{n}\frac{1}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{x}-\mathbf{X}_i}{h_i}\right\|^2\right)}-\mathbf{x} \tag{13}$$

A kernel *G* can be defined as

$$G(\mathbf{x}) = Cg(\|\mathbf{x}\|^2) \tag{14}$$

where *C* is a normalization constant. Then, by employing (9), the term that multiplies the mean shift vector in (12) can be written as

$$\frac{2}{n}\left[\sum_{i=1}^{n}\frac{1}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{x}-\mathbf{X}_i}{h_i}\right\|^2\right)\right]=\frac{2}{C}\left[\frac{\sum_{i=1}^{n}\tilde{f}(\mathbf{X}_i)}{n\lambda h_0^2}\right]\hat{f}_G(\mathbf{x}) \tag{15}$$

where

$$\hat{f}_G(\mathbf{x})\equiv C\frac{\sum_{i=1}^{n}\tilde{f}(\mathbf{X}_i)\frac{1}{h_i^d}g\left(\left\|\frac{\mathbf{x}-\mathbf{X}_i}{h_i}\right\|^2\right)}{\sum_{i=1}^{n}\tilde{f}(\mathbf{X}_i)} \tag{16}$$

is the density estimate of the data points weighted by the pilot density values computed with kernel *G*. Using (12), (13) and (15), the mean shift vector becomes

$$M(\mathbf{x})=\frac{\lambda}{n^{-1}\sum_{i=1}^{n}\tilde{f}(\mathbf{X}_i)}\frac{h_0^2}{2/C}\frac{\hat{\nabla}f_K(\mathbf{x})}{\hat{f}_G(\mathbf{x})} \tag{17}$$

Expression (17) shows that the adaptive mean shift is an estimator of the normalized gradient of the underlying density. The mean shift procedure is defined recursively by computing the mean shift vector $M(\mathbf{x})$ and translating the center of kernel *G* by $M(\mathbf{x})$, this procedure leads to a stationary point of the underlying density.

If $\{\{y_j\}_{j=1,2,\ldots}$ denotes the sequence of successive locations of the kernel *G*, where

$$\mathbf{y}_{j+1}=\frac{\sum_{i=1}^{n}\frac{\mathbf{X}_i}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{y}_j-\mathbf{X}_i}{h_i}\right\|^2\right)}{\sum_{i=1}^{n}\frac{1}{h_i^{d+2}}g\left(\left\|\frac{\mathbf{y}_j-\mathbf{X}_i}{h_i}\right\|^2\right)},\quad j=1,2,\ldots \tag{18}$$

is the weighted mean at $\mathbf{y}_j$ computed with kernel *G* and $\mathbf{y}_1$is the center of the initial kernel. The density estimates computed with kernel *K* in the points (18) are

$$\hat{f}_K=\left\{\hat{f}_K(j)\right\}_{j=1,2,\ldots}\equiv\left\{\hat{f}_K(\mathbf{y}_j)\right\}_{j=1,2,\ldots} \tag{19}$$

It is shown in [4] that if the kernel *K* has a convex and monotonic decreasing profile and the kernel G is defined according to $g(x)=-k'(x)$ and (14), the sequences (18) and (19) are convergent. This means that the adaptive mean shift procedure converges at a nearby point where the estimation has zero gradient, and since the modes of the density are points of zero gradient, then the convergence point is a mode candidate.

The mean shift vector points in the direction of the probability density function gradient, and since it is aligned with the local gradient estimate in $\mathbf{x}$, it can define a path leading $\mathbf{x}$ to a stationary point (mode) of the estimated probability density. The procedure computes $M(\mathbf{x})$ for each data point, shifts the kernel centers by these quantities, and iterates until the magnitudes of the shifts are less than a given threshold or a certain number of iterations is attained.

A. Applying mean shift

Multimodal and arbitrarily shaped clusters are the defining properties of a real feature space. The quality of the mean shift (MS) procedure to move toward the mode makes it the ideal computational module to analyze such spaces. Therefore, arbitrarily structured feature spaces can be analyzed by MS, the data points converging to their modes, automatically delineate clusters of arbitrary shapes.

The discontinuity preserving smoothing in an image is one application of MS. Smoothing through replacing the pixel by the mode obtained with MS, remove the noise and preserve discontinuities reducing the amount of smoothing near abrupt changes in the local structure, i.e., edges.

An improved mean shift estimate can be obtained by weighting each data by a function of its edge confidence [5], that represent the confidence of truly being in the presence of and edge between adjacent clusters, so that data that lie close to an edge (edge confidence close to one) are less influential in the determination of the new cluster center.

The weighted MS procedure is applied for the data points in the joint spatial-range domain. For both domains, an Euclidian metric is used to control the quality of the segmentation, which is dependent on the radii $h_s(\mathbf{X}_i)$ and $h_r(\mathbf{X}_i)$, corresponding to the bandwidths of the kernel estimate in the spatial and range domains. After the MS procedure is applied to every data, those points that are sufficiently close in the joint domain are fused to obtain the homogeneous region in the image. The filtered image is drawn from the output of the MS procedure, by replacing each pixel with the gray level (range domain) of the 3-dimensional mode it is associated to [1].

The iterative MS procedure for mode detection based on variable bandwidth is summarized in the sequel.

Adaptive mean shift filtering algorithm

Let $\mathbf{X}_i$and $\mathbf{z}_i$, $i=1,\ldots,n$, be the *d*-dimensional input and filtered image pixels,

1) Derive a pilot estimate $\tilde{f}$ using a fixed Epanechnikov kernel estimate, with bandwidth chosen arbitrarily.
2) Compute the bandwidth factor: $\log\lambda = n^{-1}\sum\log\tilde{f}(\mathbf{X}_i)$.
3) For each pixel $\mathbf{X}_i$ compute its adaptive bandwidth $h(\mathbf{X}_i)=h_0\left[\lambda/f(\mathbf{X}_i)\right]^{1/2}$
4) Initialize $j=1$ and $\mathbf{y}_{i,1}=\mathbf{X}_i$
5) Compute $\mathbf{y}_{i,j+1}$ according to (18) until convergence, $\mathbf{y}_{i,c}$
6) Assign $z_i=(\mathbf{X}_i^s,\mathbf{y}_{i,c}^r)$

The superscripts s and r denote the spatial and range components of a vector, respectively. The assignment specifies that the filtered data at the spatial location $\mathbf{X}_i^s$ will have the range component of the point of convergence $\mathbf{y}_{i,c}^r$.

IV. RESULTS

The performance of the adaptive MS filtering algorithm was assessed with synthetic images. The images were taken from the BrainWeb: Simulated Brain Database, and consist of T1 weighted images with a 181×217 size, 1 mm^3 voxel resolution, and a variety of noise levels and levels of intensity non-uniformity.

The pilot estimate $\tilde{f}$ was derived using a fixed bandwidth procedure. The fixed bandwidths, h_s and h_r were selected subjectively, and intentionally wrong (a larger bandwidth) in

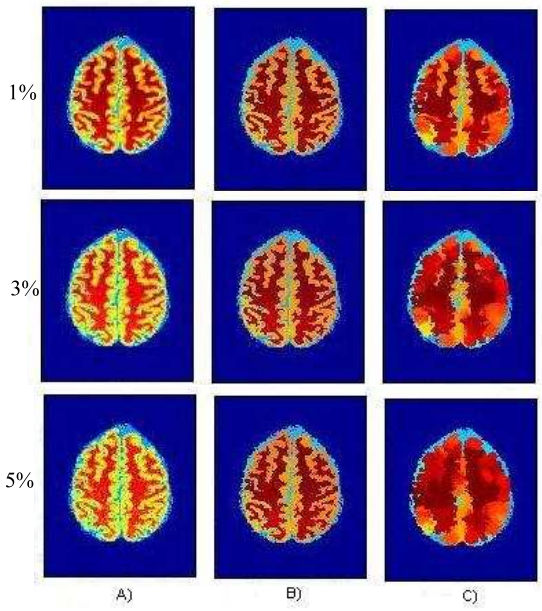

Fig. 1. Image with constant 20% level of intensity non-uniformity. The first, second and third row with 1%, 3% and 5% of noise level, respectively. A) Original image, B) Filtered with adaptive MS, C) Filtered with fixed bandwidth $(h_s, h_r) = (15, 30)$.

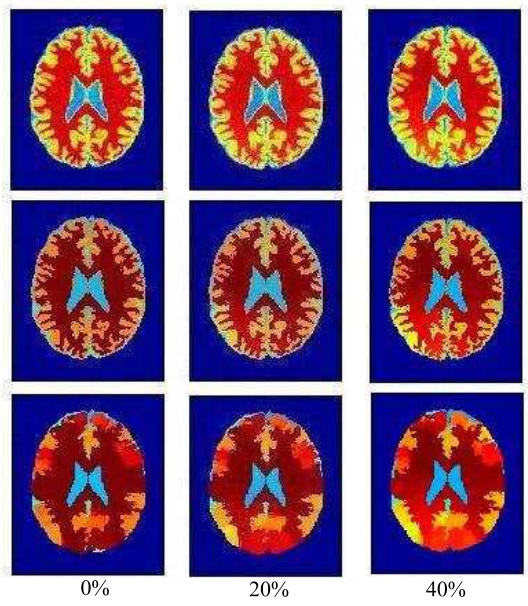

Fig. 2. Image with constant 3% noise level. The first, second and third column with 0%, 20% and 40% of level of intensity non-uniformity, respectively. Top row, original image. Middle row, filtered with adaptive MS. Bottom row, filtered with fixed bandwidth $(h_s, h_r) = (15, 30)$.

order to show that the initial fixed bandwidth has a small impact in the adaptive MS filtering.

Fig. 1, shows the same slice of three different stacks, all slices hold the same level of non-uniformity, however, the noise level is changed in each slice. Fig. 2, also shows the same slice of three distinct stacks, but the noise level is constant now, and the non-uniformity is changed in each slice. It can be seen from Fig. 1 and Fig. 2, that the fixed bandwidth, $(h_s, h_r) = (15, 30)$, has a very bad effect on the output of the MS procedure. If the same radii values are used to derive the pilot estimate, and compute the bandwidth factor, the adaptive MS filtering improves the output, recovering the homogeneous regions of the image through discontinuity preserving smoothing.

A synthetic volume was also obtained with a $181 \times 217 \times 131$ size, 3% noise and 20% intensity non-uniformity. For this volume, the ground truth is availale in order to measure the similarity (calculated by the average Tanimoto coefficient [6] for the whole stack) between the segmented image and the ground truth. The similarity indeces for fixed bandwidth (manually selected for better results) were: 0.993 for background, 0.622 for CSF, 0.737 for gray matter and 0.805 for white matter; and for adaptive bandwidth: 0.997 for background, 0.768 for CSF, 0.860 for gray matter and 0.915 for white matter.

V. CONCLUSIONS

The results of image filtering and image segmentation obtained when a statistical estimation technique is used, are influenced mainly by the kernel bandwidth, because the filtering and segmentation quality depends heavily on the bandwidth employed. A wrong bandwidth selection affects the estimation performance, by under/oversmoothing the peaks of the density.

The adaptive mean shift approach has shown than it can override this problem by means of a two-stage procedure, allowing that the bandwidth selection has a low impact in the quality of the results.

If the objective of variable bandwidth kernel estimation is to improve the performance of kernel estimators by adapting the kernel bandwidth to the local data statistics, using a fixed bandwidth is not effective when data exhibits multiscale patterns, for that reason the investigation about techniques that allow to determine the right bandwidth for this patterns must continue.

VI. ACKNOWLEDGEMENTS

This work was financially supported by CONACyT Grant number 42309 and LAFMI.

REFERENCES

[1] J. R. Jiménez-Alaniz, V. Medina-Bañuelos, and O. Yáñez-Suárez, "Data-driven brain mri segmentation supported on edge confidence and a priori tissue information," *IEEE Trans. Med. Imag.*, vol. 25, pp. 74–83, January 2006.

[2] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. No. 26 in Monographs on Statistics and Applied Probability, Chapman and Hall/CRC, 1986.

[3] P. Hall, T. C. Hu, and J. S. Marron, "Improved variable window kernel estimates of probability densities," *The Annals of Statistics*, vol. 23, pp. 1–10, February 1995.

[4] D. Comaniciu, V. Ramesh, and P. Meer, "The variable bandwidth mean shift and data-driven scale selection," in *IEEE Int. Conf. Computer Vision (ICCV'01)*, vol. 1, (Vancouver, Canada), pp. 438–445, 2001.

[5] C. M. Christoudias, B. Georgescu, and P. Meer, "Synergism in low level vision," in *16th International Conference on Pattern Recognition*, (Quebec City, Canada), pp. 150–155, August 2002. vol. IV.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. 605 Third avenue, New York, NY: John Wiley & Sons, Inc., second ed., 2001.