MEAN-SHIFT ANALYSIS USING QUASI-NEWTON METHODS

Changjiang Yang, Ramani Duraiswami, Daniel DeMenthon and Larry Davis *

Perceptual Interfaces & Reality Laboratory
University of Maryland, College Park, MD 20742
{yangcj,ramani,daniel,lsd}@umiacs.umd.edu

ABSTRACT

Mean-shift analysis is a general nonparametric clustering technique based on density estimation for the analysis of complex feature spaces. The algorithm consists of a simple iterative procedure that shifts each of the feature points to the nearest stationary point along the gradient directions of the estimated density function. It has been successfully applied to many applications such as segmentation and tracking. However, despite its promising performance, there are applications for which the algorithm converges too slowly to be practical. We propose and implement an improved version of the mean-shift algorithm using quasi-Newton methods to achieve higher convergence rates. Another benefit of our algorithm is its ability to achieve clustering even for very complex and irregular feature-space topography. Experimental results demonstrate the efficiency and effectiveness of our algorithm.

1. INTRODUCTION

Mean-shift analysis is a relatively new but important clustering approach originally invented by Fukunaga and Hostetler [1] which they called a "valley-seeking procedure". In spite of its excellent performance, it had been nearly forgotten until Cheng [2] extended it and introduced it to the image analysis community. Recently Comaniciu and Meer [3, 4] successfully applied it to image segmentation and tracking. DeMenthon [5] employed it for spatio-temporal segmentation of video sequences in a 7D feature space.

Mean-shift essentially is a feature-based analysis of data points, which requires a nonparametric estimator of the gradient of the density gradient in feature space. Advantages of feature-space methods are the global representation of the original data and the excellent tolerance to noise [6]. When a density function in feature space has peaks and valleys, it is desirable to divide data points into clusters according to the valleys of the point densities, because such boundaries in feature space are mapped back to much more natural segmentation boundaries.

*Support of NSF award 9987944 and Department of Defense contract MDA 9049-6C-1250 is gratefully acknowledged.

The mean-shift procedure consists of two steps: the estimation of the gradient of the density function, and the utilization of the results to form clusters. The gradient of the density function is estimated by a nonparametric density estimator [6]. Then starting from each sample point, the mean-shift procedure iteratively finds a path along the gradient direction away from the valleys and towards the nearest peak.

The standard mean-shift procedure utilizes the steepest ascent method to seek the stationary points of the density function. It is well known that steepest ascent method is very inefficient at solving most problems [7], especially in the case of complex and irregular density functions. In this paper, we propose to replace the steepest ascent method with the well known quasi-Newton methods which approximate the Hessian matrix from the gradient [7, 8]. There are several advantages to this approach. First, an approximation of the Hessian matrix can be found using only gradient information. Second, the method converges superlinearly. Also, the computational overhead is relatively small. Finally, the algorithm converges for cases where the steepest ascent fails.

The paper is organized as follows. Section 2 describes the standard mean-shift algorithm using the steepest ascent method. The quasi-Newton method is discussed in Section 3. The proposed mean shift algorithm using quasi-Newton method is introduced and analyzed in Section 4. Section 5 presents segmentation results using mean-shift algorithms.

2. MEAN-SHIFT ANALYSIS

Given n data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in the d-dimensional space R^d, the *kernel density estimator* with kernel function $K(\mathbf{x})$ and a window bandwidth h, is given by [6, 9, 10]

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

where the d-variate kernel $K(\mathbf{x})$ is nonnegative and integrates to one. A widely used class of kernels are the radially symmetric kernels

$$K(\mathbf{x}) = c_{k,d} k(\|\mathbf{x}\|^2), \tag{2}$$

where the function $k(x)$ is called the *profile* of the kernel, and normalization constant $c_{k,d}$ makes $K(\mathbf{x})$ integrate to one. The density estimator (1) can be rewritten as

$$\hat{f}_{h,k}(\mathbf{x}) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} k\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right), \tag{3}$$

where $c_{k,d}$ is the normalization constant. Two commonly used kernels are the Epanechnikov kernel

$$K_E(\mathbf{x}) = \begin{cases} \frac{1}{2} c_d^{-1} (d+2)(1 - \|\mathbf{x}\|^2) & 0 \le \|\mathbf{x}\| \le 1 \\ 0 & \|\mathbf{x}\| > 1, \end{cases} \tag{4}$$

and the multivariate Gaussian kernel

$$K_N(\mathbf{x}) = (2\pi)^{-d/2} e^{-\frac{1}{2}\|\mathbf{x}\|^2}. \tag{5}$$

The standard mean shift algorithm is a steepest ascent procedure which requires estimation of the density gradient:

$$\begin{aligned} \nabla \hat{f}_{h,K}(\mathbf{x}) &= \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{x}) g\left(\left\| \frac{\mathbf{x} - \mathbf{x}_i}{h} \right\|^2 \right) \\ &= c_{k,g} \hat{f}_{h,G}(\mathbf{x}) \left[\frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \right\|^2 \right)} - \mathbf{x} \right], \end{aligned} \tag{6}$$

where $g(x) = -k'(x)$ which can in turn be used as profile to define a kernel $G(\mathbf{x})$. The kernel $K(\mathbf{x})$ is called the shadow of $G(\mathbf{x})$ [2]. $\hat{f}_{h,G}(\mathbf{x})$ is the density estimation with the kernel G. $c_{k,g}$ is the normalization coefficient. The last term is the *mean shift*

$$\mathbf{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{x}-\mathbf{x}_i}{h} \right\|^2 \right)} - \mathbf{x}, \tag{7}$$

which is proportional to the normalized density gradient and always points toward the steepest ascent direction of the density function. The standard mean shift algorithm iteratively performs

- computation of the mean shift vector $\mathbf{m}(\mathbf{x}^k)$,
- updating the current position $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{m}(\mathbf{x}^k)$,

until reaching the stationary point which is the candidate cluster center.

3. THE QUASI-NEWTON METHODS

The steepest-ascent method used in the mean-shift algorithm often converges in a zigzag fashion which makes it a very inefficient method in spite of its asymptotically global convergence, i.e., it needs a very large number of steps to achieve convergence. To obtain a superlinearly convergent method it is necessary to approximate the Newton step asymptotically [11]. However, the Newton step requires computation of the Hessian matrix of the density which is computationally expensive. Quasi-Newton methods avoid this by using the observed behaviors of function and gradient to approximate the Hessian matrix. Due to their remarkable robustness and efficiency, they may be the most widely used methods for nonlinear optimization. They are implemented in all major subroutine libraries and have been used to solve a wide variety of practical problems [7, 12].

There are many quasi-Newton methods, but the BFGS method is generally considered to be the most effective [11]. The BFGS method is the same as the steepest ascent method except that the shift in the gradient direction (6) is replaced by a shift along

$$\mathbf{d}^k = B_k^{-1} \nabla \hat{f}(\mathbf{x}^k) \tag{8}$$

at the k-th iteration. The Hessian approximation at step k, B_k, is updated by

$$B_{k+1} = B_k - \frac{B_k \mathbf{s}_k \mathbf{s}_k^T B_k}{\mathbf{s}_k^T B_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^T}{\mathbf{y}_k^T \mathbf{s}_k}, \tag{9}$$

where

$$\mathbf{y}_k = \nabla \hat{f}(\mathbf{x}^{k+1}) - \nabla \hat{f}(\mathbf{x}^k), \qquad \mathbf{s}_k = \mathbf{x}^{k+1} - \mathbf{x}^k. \tag{10}$$

Initially, B_0 can be set to any symmetric positive definite matrix, for example, the identity matrix I. At step k, the current position is updated by

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \mathbf{d}^k \tag{11}$$

where α_k is the stepsize. This should be compared with the update rule in Equation (7), in the standard algorithm.

The global convergence of BFGS method was proved by Powell [13]. Under appropriate assumptions, the BFGS can be proved to converge superlinearly with rate $r = \frac{1}{2}(1 + \sqrt{5}) \approx 1.618$ [7].

4. BFGS-MEAN-SHIFT ANALYSIS

Our proposed mean shift analysis using quasi-Newton methods is an iterative procedure. Each iteration requires the performance of the following steps:

1. Compute the gradient of the density function using formula (6).
2. Update the approximation of Hessian matrix B_{k+1} from B_k using formula (9).
3. Compute the search direction $\mathbf{d}^k$ using formula (8).
4. (Optional) Find the stepsize α_k using line search algorithms [7], otherwise, set it to 1.
5. Update the current position using formula (11).

In the above steps, the extra computation brought by BFGS method is in the second and third steps. The computational complexity of step 2 is $O(d^2)$ for d-dimensional data. The computational complexity of step 3 is $O(d^3)$. However, more elegant implementation can reduce it to $O(d^2)$ arithmetic operations [14]. If the dimensionality of the data is low, the computational and memory overhead from BFGS method is very small compared with the estimation of the density gradient. In the higher dimensional case, limited memory quasi-Newton methods (L-BFGS) can be adopted to reduce the computation and memory cost [15]. The most computationally expensive step is the first step which estimates the gradient of the density function using nonparametric kernel density estimation. For N points in feature space this is potentially an $O(N^2)$ step. Recently DeMenthon [5] applied a range search method that prunes far feature regions sorted in a precomputed binary tree structure, to make this efficient. Alternatively, Elgammal et al [16] showed that the computational complexity of kernel density estimation can be reduced to linear order from quadratic one using fast multipole methods.

5. EXPERIMENTAL RESULTS

The the famous Rosenbrock's function is a good example to show the efficiencies of the steepest ascent method and the BFGS method [8]:

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2, \qquad (12)$$

The negative of this function $-f$ is used for the maximization. The two algorithms are started at point $[-1.9, 2]$. As in [8], the steepest ascent method continually zigzags along the ridge of the function for 200 iterations and still make no much progress towards the solution. The BFGS method is able to follow the shape of the ridge and converges to the peak after 20 iterations (Figure 1).

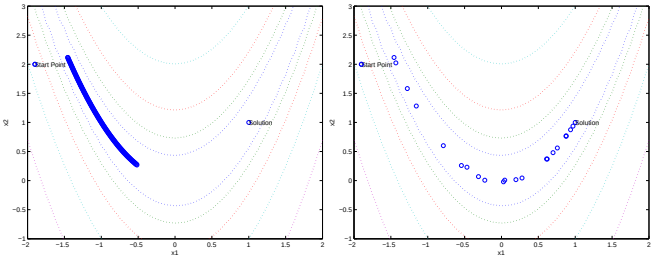

Fig. 1. Solution paths on Rosenbrock's function using the steepest ascent method (*left*), the BFGS method (*right*).

Our second experiment shows the performance of the standard mean-shift algorithm and the BFGS-mean-shift algorithm by simulations. We generate 2000 data points among which 1000 points are chosen from the 2D normal distribution with mean $[0, -2]$, and covariance matrix $[16\ \ 0;\ 0\ \ 1]$, and 1000 points are chosen from the 2D normal distribution with mean $[0, 2]$, and covariance matrix $[16\ \ 0;\ 0\ \ 1]$ (as shown in Figure 2a). The standard mean-shift and the BFGS-mean-shift are applied to the data set and the data clustered. The "+" plots the final positions of the points after 20 iterations for the standard mean-shift and 15 iterations for the BFGS-mean-shift (so that in complexity terms the comparison is fair). We can find that the BFGS-mean-shift achieves a much more compact clustering result. The k-means algorithm [6] correctly finds the clusters from the result of BFGS-mean-shift, but fails on the result of the standard mean-shift.

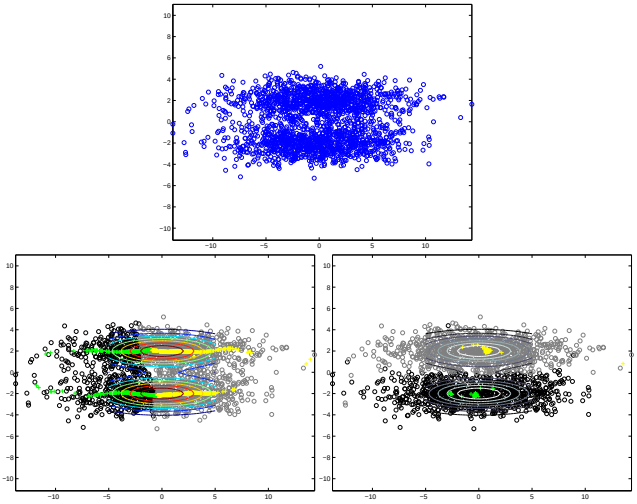

Fig. 2. Synthetic data. (*Top*) The data points generated by two normal distributions. (*Bottom left*) Segmented using the standard mean-shift. (*Bottom right*) Segmented using the BFGS-mean-shift.

The third experiment performs mean-shift segmentation on color image. The *house* image (Figure 3a) is first mapped into $L^*u^*v^*$ color space where mean-shift algorithms are applied to form clusters. To speed up the mean-shift, we applied the k-center algorithm [17] to subdivide the space into 100 sample sets and applied mean-shift algorithm on the sample sets. We applied k-means algorithm to the results of the mean-shift. After 15 iterations of the mean-shift algorithms, we can find that the BFGS-mean-shift obtained better segmentation results as shown in Figure 3.

6. CONCLUSIONS

We have described a improved mean-shift algorithm using the quasi-Newton methods. The proposed method utilizes the curvature information of the density function to guide the search process. The quasi-Newton method speeds up the mean-shift algorithm with little extra computational and memory cost. The postprocessing of results to achieve better clusters can be reduced because our method forms much

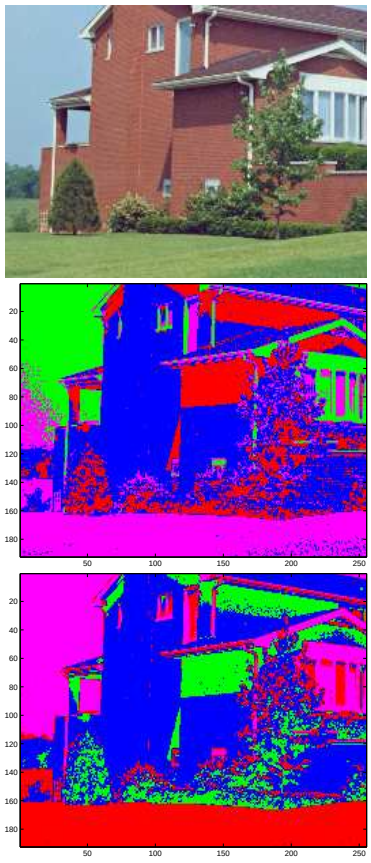

Fig. 3. *House* image. (*Top*) Original image. (*Center*) Segmented using the standard mean-shift. (*Bottom*) Segmented using the BFGS-mean-shift.

more compact clustering results. Future work will be carried out to carefully analyze the stability of the algorithm on more complicated data sets, and to further speed up the algorithm in higher dimensional space with many feature points.

Acknowledgments

We would like to thank Prof. Dianne O'Leary for valuable constructive suggestions on this work.

7. REFERENCES

[1] K. Fukunaga and L. D. 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.

[2] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, August 1995.

[3] D. Comaniciu and P. Meer, "Mean shift analysis and applications," in *Proc. Int'l Conf. Computer Vision*, 1999, pp. 1197–1203.

[4] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603 – 619, May 2002.

[5] D. DeMenthon, "Spatio-temporal segmentation of video by hierarchical mean shift analysis," in *Statistical Methods in Video Processing Workshop*, Copenhagen, Denmark, 2002.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, 2000.

[7] S. G. Nash and A. Sofer, *Linear and Nonlinear Programming*, McGraw-Hill, New York, 1996.

[8] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, San Diego, 1981.

[9] D. W. Scott, *Multivariate Density Estimation: Theory, Practical, and Visualization*, Wiley, New York, 1992.

[10] M. P. Wand and M. C. Jones, *Kernel Smoothing*, Chapman & Hall, London, 1995.

[11] J. Nocedal, "Theory of algorithms for unconstrained optimization," *Acta Numerica*, vol. 1, pp. 199–242, 1992.

[12] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Verlag, 1999.

[13] M.J.D. Powell, "Some global convergence properties of a variable metric algorithm for minimization without exact line searches," in *Nonlinear Programming*, R.W. Cottle and C.E. Lemke, Eds., pp. 53–72. AMS, Providence, RI, 1976.

[14] P. Gill and W. Murray, "Quasi-newton methods for unconstrained optimization," *J. Inst. Maths. Applics.*, vol. 9, pp. 91–108, 1972.

[15] D. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Programming*, vol. 45, pp. 503–528, 1989.

[16] A. Elgammal, R. Duraiswami, and L. Davis, "Efficient non-parametric adaptive color modeling using fast gauss transform," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Kauai, Hawaii, 2001.

[17] M. Bern and D. Eppstein, "Approximation algorithms for geometric problems," in *Approximation Algorithms for NP-Hard Problems*, D. Hochbaum, Ed. PWS Publishing Company, Boston, 1997.