Mean Shift is a Bound Optimization

Mark Fashing and Carlo Tomasi, *Member, IEEE*

M. Fashing and C. Tomasi are with the Department of Computer Science, Duke University, Durham, NC 27708-0129. E-mail: {mark; tomasi}@cs.duke.edu

ABSTRACT

We build on the current understanding of mean shift as an optimization procedure. We demonstrate that in the case of piecewise constant kernels mean shift is equivalent to Newton's method. Further, we prove that for all kernels the mean shift procedure is a quadratic bound maximization.

INDEX TERMS

Mean shift, bound optimization, Newton's method, adaptive gradient descent, mode seeking.

I. INTRODUCTION

Mean shift is a nonparametric, iterative procedure introduced by Fukunaga and Hostetler [1] for seeking the mode of a density function represented by a set S of samples. The procedure uses so-called *kernels*, which are decreasing functions of the distance from a given point $\mathbf{t}$ to a point $\mathbf{s}$ in S.

For every point $\mathbf{t}$ in a given set T, the sample means of all points in S weighted by a kernel at $\mathbf{t}$ are computed to form a new version of T. This computation is repeated until convergence. The resulting set T contains estimates of the modes of the density underlying set S. The procedure will be reviewed in greater detail in Section II.

Cheng [2] revisited mean shift, developing a more general formulation and demonstrating its potential uses in clustering and global optimization. Recently, the mean shift procedure has met with great popularity in the computer vision community. Applications range from image segmentation and discontinuity-preserving smoothing [3], [4] to higher level tasks like appearance-based clustering [5], [6] and blob tracking [7].

Despite the recent popularity of mean shift, few attempts have been made since Cheng [2] to understand the procedure theoretically. For example, Cheng [2] showed that mean shift is an instance of gradient ascent and also notes that, unlike naïve gradient ascent, mean shift has an adaptive step size. However, the basis of step size selection in the mean shift procedure has remained unclear. We show that in the case of piecewise constant kernels the step is exactly the Newton step and in all cases it is a step to the maximum of a quadratic bound.

Another poorly understood area is that of mean shift with an evolving sample set. Some variations on the mean shift procedure use the same set for samples and cluster centers. This

causes the sample set to evolve over time. The optimization problem solved by this variation on mean shift has yet to be characterized.

In this paper, we build on the current understanding of mean shift as an optimization procedure. Fukunaga and Hostetler [1] suggested that mean shift might be an instance of gradient ascent. Cheng [2] clarified the relationship between mean shift and optimization by introducing the concept of the *shadow kernel* and showed that mean shift is an instance of gradient ascent with an adaptive step size. We explore mean shift at a deeper level by examining not only the gradient but also the Hessian of the shadow kernel density estimate. In doing so, we establish a connection between mean shift and the Newton step, and we demonstrate that in the case of piecewise constant kernels mean shift is equivalent to Newton's method optimization. Further, we prove that for all kernels the mean shift procedure is a quadratic bound maximization (see Fig. 1), and we show that this characterization also holds for mean shift with evolving sample sets.

In Section II-A we provide a brief review of the mean shift procedure, and in Section II-B we provide a brief overview of bound optimization. In Section III-A we examine the gradient and the Hessian of the shadow kernel density estimate and establish a relationship between mean shift and Newton's method. In Section III-B we then prove that the mean shift procedure is a bound optimization. In Section IV we discuss implications and the resulting strengths and weaknesses of the mean shift procedure.

II. Background on Mean Shift and Bound Optimization

A. A Review of the Mean Shift Procedure

In this paper, we examine the generalized version of the mean shift procedure developed by Cheng [2]. We initially restrict ourselves to the case where S is a stationary set of samples. The mean shift procedure also allows S and T to be the same set, where T is the set of means. We address the case of evolving sample sets at the end of our analysis.

The concept of a *kernel* (see Def. 2) is fundamental to the mean shift procedure, and indeed mean shift is conventionally defined in terms of a kernel. However, in this paper we define mean shift in terms of a *profile* (see Def. 1). This improves the clarity of our analysis and does not stray far from the conventional definition.

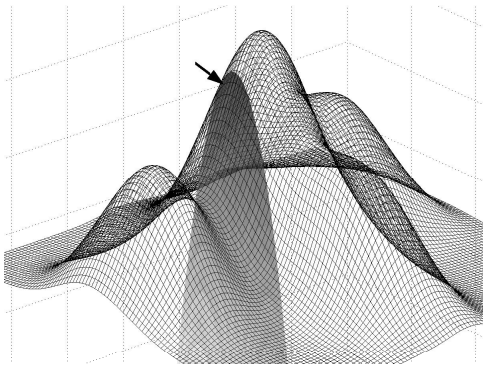


Fig. 1. Quadratic bound (solid) of the shadow density estimate (mesh). The point of tangency is marked with a black arrow. One iteration of the mean shift procedure maximizes this bound. These data were generated by drawing 100 random, normally distributed samples from each of 5 random, normally distributed means.

Definition 1: A profile k is a piecewise continuous, monotonically nonincreasing function from a nonnegative real to a nonnegative real such that the definite integral $\int_0^\infty k(r)dr < \infty$.

Definition 2: A kernel K is a function from a vector $\mathbf{x}$ in the n-dimensional real Euclidean space, X, to a nonnegative real, such that $K(\mathbf{x}) = k(\|\mathbf{x}\|^2)$, for some profile k.

We reason in terms of profiles for two reasons. First, we spend much time considering first and second derivatives of profiles. One cannot differentiate a kernel directly. Rather one must differentiate the profile of the kernel. Reasoning in terms of kernels creates an additional and unnecessary layer of indirection. Second, in Section III-B we will consider a space that is closely related to our definition of mean shift in terms of profiles.

The generalized mean shift procedure given by Cheng [2], rewritten in terms of a profile, follows.

Definition 3: Let X be an n-dimensional real Euclidean space and S a set of sample vectors in X. Let w be a weight function from a vector in X to a nonnegative real. Let the *sample mean* $\mathbf{m}$ with profile k at $\mathbf{x} \in X$ be defined such that

$$\mathbf{m}(\mathbf{x}) = \frac{\sum_{\mathbf{s}\in S} k(\|\mathbf{s}-\mathbf{x}\|^2)w(\mathbf{s})\mathbf{s}}{\sum_{\mathbf{s}\in S} k(\|\mathbf{s}-\mathbf{x}\|^2)w(\mathbf{s})}. \tag{1}$$

Let $M(T) = \{\mathbf{m}(\mathbf{t}) : \mathbf{t} \in T\}$. One iteration of mean shift is given by $T \leftarrow M(T)$. The full mean shift procedure iterates until it finds a fixed point $T = M(T)$.

Note 1: Since we will make frequent use of norms like $\|\mathbf{s}-\mathbf{x}\|^2$, from here on we will use the shorthand $r = \|\mathbf{s}-\mathbf{x}\|^2$.

In addition to generalizing mean shift, Cheng [2] demonstrated that mean shift with fixed S seeks the modes of the density estimate given by a *shadow profile*.

Definition 4: A profile h is said to be the *shadow* of a profile k iff

$$h(r) = b + c\int_r^\infty k(t)dt, \tag{2}$$

where b is a real and c is a positive real. It follows that h is a shadow of k if k is the negative of the derivative of h, possibly scaled.

Cheng [2] defined the shadow profile such that b was replaced by a piecewise constant function. This suggests that shadow profiles exist which are not continuous and therefore not continuously differentiable. We define shadow profiles such that their continuity, and therefore their differentiability, is guaranteed.

Theorem 1 (Cheng [2]): Mean shift with profile k seeks the modes of the density estimate

$$q(\mathbf{x}) = \sum_{\mathbf{s}\in S} h(r)w(\mathbf{s}), \tag{3}$$

where h is a shadow of the profile k.

Although h may be any shadow of k, the modes of the density estimate obtained using h will always be the same, since all shadows of k are equivalent up to a scale factor and a translation.

This brief overview of mean shift should give sufficient insight into the procedure to support the following analysis. For a more detailed explanation of the procedure, its definitions and constraints please refer to Fukunaga and Hostetler [1] and Cheng [2].

B. An Overview of Bound Optimization

A bound optimization algorithm can be constructed for any objective function which possesses a structure that can be exploited to find a lower bound (see Def. 5). One well known example is the EM algorithm, described as a bound optimization by Neal and Hinton [8].

Definition 5: Let $q(\mathbf{x})$ be the objective function, where $\mathbf{x}$ is in the space X. The (lower) bounding function $\rho(\mathbf{x})$ is a function such that $\rho(\mathbf{x}_0) = q(\mathbf{x}_0)$ at some point of tangency $\mathbf{x}_0 \in X$ and $\rho(\mathbf{x}) \leq q(\mathbf{x})$ for every $\mathbf{x} \in X$.

If there exists a bounding function ρ tangent to the objective function q for every $\mathbf{x}_0 \in X$ and it is significantly less computationally expensive to maximize ρ than q, then we can construct a

bound optimization algorithm for q by simply finding and maximizing the bound of q until we reach a stationary point.

III. Analysis of the Mean Shift Procedure

A. The Relationship Between Mean Shift and Newton's Method

We can derive the gradient and the Hessian with respect to $\mathbf{x}$ of the density q from its definition (see Eqn. 3):

$$\nabla q = -2\sum_{\mathbf{s}\in S} h'(r)w(\mathbf{s})(\mathbf{s}-\mathbf{x}) \tag{4}$$

$$\nabla^2 q = 2\sum_{\mathbf{s}\in S}(w(\mathbf{s})(h'(r)I+ 2h''(r)(\mathbf{s}-\mathbf{x})(\mathbf{s}-\mathbf{x})^T)). \tag{5}$$

Indeed Eqn. 4 is the gradient reported by Cheng [2].

By Def. 1, a profile k is nonnegative and the integral $\int_0^\infty k(r)dr < \infty$. It follows that $\lim_{a\to\infty} k(a) = 0$. This leads to the analytic first and second derivatives of the profile h, where h is the shadow of the profile k (see Eqn. 2);

$$h'(r) = -ck(r) \tag{6}$$

$$h''(r) = -ck'(r). \tag{7}$$

It is possible for k to have a finite number of discontinuities at which k', and therefore h'', will not be defined. For the case of piecewise constant profiles, we define k' to be zero. We argue that since the left derivative and the right derivative are both zero, it is reasonable to interpolate k' at discontinuities in k in order to preserve continuity in k'.

By substitution we can obtain the gradient and Hessian in terms of the profile k,

$$\nabla q = 2c\sum_{\mathbf{s}\in S} k(r)w(\mathbf{s})(\mathbf{s}-\mathbf{x}) \tag{8}$$

$$\nabla^2 q = -2c\sum_{\mathbf{s}\in S}(w(\mathbf{s})(k(r)I+ 2k'(r)(\mathbf{s}-\mathbf{x})(\mathbf{s}-\mathbf{x})^T)). \tag{9}$$

Theorem 2: The mean shift procedure with a piecewise constant profile k is equivalent to Newton's method applied to a density estimate using the shadow of k.

Proof: If k is piecewise constant, the Hessian of q is,

$$\nabla^2 q = -2c \sum_{\mathbf{s} \in S} k(r) w(\mathbf{s}) I. \tag{10}$$

So, if k is piecewise constant, the Newton step $\mathbf{p}$ follows,

$$\mathbf{p} = -\left(\nabla^2 q\right)^{-1} \nabla q \tag{11}$$

$$= \frac{\sum_{\mathbf{s} \in S} k(r) w(\mathbf{s})(\mathbf{s} - \mathbf{x})}{\sum_{\mathbf{s} \in S} k(r) w(\mathbf{s})}. \tag{12}$$

One iteration of Newton's method yields,

$$\mathbf{x} + \mathbf{p} = \mathbf{x} + \frac{\sum_{\mathbf{s} \in S} k(r) w(\mathbf{s})(\mathbf{s} - \mathbf{x})}{\sum_{\mathbf{s} \in S} k(r) w(\mathbf{s})} \tag{13}$$

$$= \frac{\sum_{\mathbf{s} \in S} k(r) w(\mathbf{s}) \mathbf{s}}{\sum_{\mathbf{s} \in S} k(r) w(\mathbf{s})}, \tag{14}$$

which is one step of the mean shift procedure. ■

Cheng [2] showed that mean shift is gradient ascent with an adaptive step size, but the theory behind the step sizes remained unclear. We see now that the size (and direction) of a step of mean shift with a piecewise constant profile is precisely the size (and direction) of the Newton step applied to the density q (see Eqn. 3).

B. Mean Shift as a Quadratic Bound Maximization

Given Thm. 2, we can make a general statement that applies to any profile, not just piecewise constant profiles, about the optimization problem that the mean shift procedure solves at $\mathbf{x}_0 \in X$.

Theorem 3: At $\mathbf{x}_0 \in X$, the mean shift procedure maximizes the function $\rho(\mathbf{x}) = a - c \sum_{\mathbf{s} \in S} k(r_0) w(\mathbf{s}) r$, where a is a constant and c is a positive real satisfying Eqn. 6.

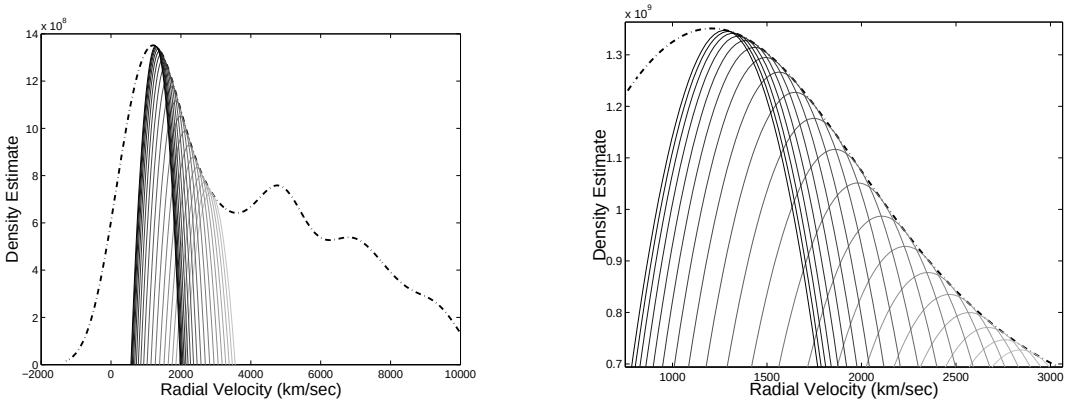
Proof: We find the gradient and Hessian of the function ρ.

$$\rho(\mathbf{x}) = a - c \sum_{\mathbf{s} \in S} k(r_0) w(\mathbf{s}) r \tag{15}$$

$$\nabla \rho = 2c \sum_{\mathbf{s} \in S} k(r_0) w(\mathbf{s})(\mathbf{s} - \mathbf{x}) \tag{16}$$

$$\nabla^2 \rho = -2c \sum_{\mathbf{s} \in S} k(r_0) w(\mathbf{s}) I \tag{17}$$

We observe that, at $\mathbf{x} = \mathbf{x_0}$ (and $r = r_0$), $\nabla q = \nabla \rho$ and $B = \nabla^2 \rho$. ρ is a quadratic, so the Newton step finds the exact maximum of ρ. ■



(a) Seeking the mode. (b) Detail.

Fig. 2. The mean shift procedure seeking the mode of a density estimate (dashed-dotted) with successive quadratic bounds (solid). The shade of the bound grows darker as mean shift nears the mode. Samples are the radial velocities of galaxies [9].

It follows from Theorem 3 that the mean shift procedure is performing a quadratic maximization (see Figs. 1 and 2). However, to prove that the mean shift procedure is a bound optimization we must show that ρ is a lower bound on q. In order to do so we will first prove a short lemma.

Lemma 1: A shadow profile is a convex function.

Proof: We observe from Eqn. 6 that $-\frac{1}{c}h'(r)$ must also be a profile and therefore the slope of shadow profile h must be monotonically nondecreasing. Therefore the value of h at the midpoint of any continuous interval in its domain cannot be larger than the average of the values of h at the ends of the interval. ■

Theorem 4: The mean shift procedure using profile k is a quadratic bound maximization over a density estimate using a continuous shadow of k.

Proof: Given Theorem 3, we need only show that for a such that $\rho(\mathbf{x}_0) = q(\mathbf{x}_0)$, $\rho(\mathbf{x}) \leq q(\mathbf{x})$, for every $\mathbf{x} \in X$. Let $\rho(\mathbf{x}_0) = q(\mathbf{x}_0)$, and therefore

$$a = \sum_{i=1}^{N} a^{(i)} = \sum_{i=1}^{N} w(\mathbf{s}^{(i)})(h(r_0^{(i)}) - h'(r_0^{(i)})r_0^{(i)}), \tag{18}$$

where N is the number of samples.

Consider ρ and q as functions of the vector

$$\mathbf{r} = \begin{bmatrix} r^{(1)} \\ \vdots \\ r^{(N)} \end{bmatrix} = \begin{bmatrix} \|\mathbf{s}^{(1)} - \mathbf{x}\|^2 \\ \vdots \\ \|\mathbf{s}^{(N)} - \mathbf{x}\|^2 \end{bmatrix}. \tag{19}$$

Observe that giving one value in $\mathbf{r}$ will restrict the other values in $\mathbf{r}$, but for the moment assume that $\mathbf{r}$ can be any vector in the N-dimensional space of positive reals, R, where N is the number of samples. In this space, $a^{(i)} + h(r^{(i)})w(\mathbf{s}^{(i)})$ is a convex function and q, the sum of convex functions, is also convex. In addition, in this space, $a^{(i)} + h'(r_0^{(i)})w(\mathbf{s}^{(i)})r^{(i)}$ is a hyperplane tangent to $h(r^{(i)})w(\mathbf{s}^{(i)})$ and ρ is a hyperplane tangent to q at $\mathbf{r}_0$ (see Fig. 3), where

$$\mathbf{r}_0 = \begin{bmatrix} r_0^{(1)} \\ \vdots \\ r_0^{(N)} \end{bmatrix} = \begin{bmatrix} \|\mathbf{s}^{(1)} - \mathbf{x}_0\|^2 \\ \vdots \\ \|\mathbf{s}^{(N)} - \mathbf{x}_0\|^2 \end{bmatrix}. \tag{20}$$

Thus the hyperplane ρ can never exceed the convex function q.

As a final note consider that giving one value $r \in \mathbf{r}$ corresponding to some sample s restricts the position of $\mathbf{x_0}$ to a hypersphere around s with radius r. This restricts the other values of $\mathbf{r}$. The functions q and ρ exist in the subspace of R in which the values of the vector $\mathbf{r}$ are spatially possible given the samples S in the space X. This proof subsumes this subspace in showing that for any $\mathbf{r} \in R$ the function ρ is less than or equal to the density estimate q. ■

Now let us consider the mean shift procedure with an evolving sample set. Observe that Theorem 3 holds regardless of whether the sample set is stationary or evolving.

Corollary 1 (to Theorem 4): Theorem 4 holds for the mean shift procedure with an evolving sample set and fixed sample weights.

Proof: Mean shift chooses the vector $\mathbf{r}$ to maximize some objective function. In the case of stationary samples, the restrictions on $\mathbf{r}$ are known *a priori*. In the case of evolving samples we do not know what subspace of R valid values of $\mathbf{r}$ live in, so we do not know the density estimate that we are maximizing *a priori*. However, we do know that, if our sample weights are fixed, the valid values of $\mathbf{r}$ will form a subspace of R. Since we know that ρ lower bounds q in the space R, we know that Theorem 4 holds for the case of evolving sample sets with fixed sample weights. ■

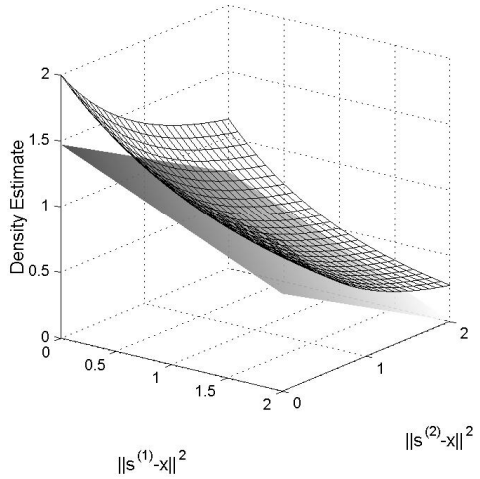


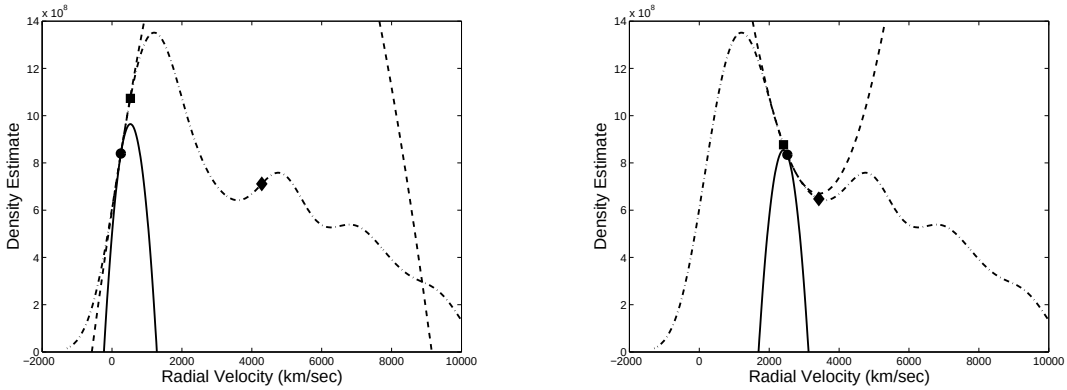
Fig. 3. With respect to $\mathbf{r}$, mean shift finds a linear bound (solid) of the shadow density estimate (mesh). An example with two samples, s_1 and s_2. The $r^{(1)}$-axis increases with the square of the distance from s_1 and the $r^{(2)}$-axis increases with the square of the distance from s_2. The point of tangency, $\mathbf{r}_0$, is at $(1, 1)$.

IV. Discussion and Conclusions

We have shown that the mean shift procedure with fixed sample weights is a quadratic bound maximization both for stationary and evolving sample sets. A number of important consequences follow directly from this fact. Most significantly, unlike Newton's method, each iteration of mean shift is guaranteed to bring us closer to a stationary point (see Fig. 4).[1] Observe that if mean shift at $\mathbf{x}_0$ yields $\mathbf{x}_1$ such that $\mathbf{x}_1 \neq \mathbf{x}_0$, then $q(\mathbf{x}_1) \geq \rho(\mathbf{x}_1) > \rho(\mathbf{x}_0) = q(\mathbf{x}_0)$. On the other hand, like Newton's method, mean shift can get stuck at a saddle point or mistake a start at a local minimum for a local maximum.

A number of less obvious consequences follow as well. Given this description of the mechanism behind the mean shift procedure, we may consider extensions to the procedure that were not possible before. For example, mean shift currently operates in a Euclidean space with Euclidean distances. We can substitute non-Euclidean norms into Eqn. 15 and maximize this bound at each

[1] As we have shown, piecewise constant profiles are a subcase in which mean shift is equivalent to Newton's method. In this subcase, the quadratic lower bound found by mean shift is equivalent to the quadratic approximation found by Newton's method. This equivalence does not hold generally.



(a) Overshooting. (b) Seeking the wrong root.

Fig. 4. The quadratic bound of the mean shift procedure (solid) and the quadratic approximation of Newton's method (dashed) seeking the mode of a density estimate (dashed-dotted). The starting point is indicated by a circle, while the sample mean is indicated by a square and the Newton step by a diamond. Notice that in both cases the Newton step actually results in a *decrease* in the value of the objective function. Samples are the radial velocities of galaxies [9].

iteration.

Another question to ask is whether or not we can speed up mean shift by tightening the bound. Naturally if we make no assumptions aside from those in Definition 1 we have no guarantees, except that a convex shadow exists, and we cannot tighten the bound. However, the mean shift procedure typically employs Gaussian and truncated Gaussian kernels (exponential and truncated exponential profiles). We can use this extra information about profile shape to tighten the bound; the difficulty is in finding a bound which is computationally easy to maximize. Such an enhancement would provide an additional speedup to existing techniques such as locality sensitive hashing [10].

Until now, the internal mechanism behind the mean shift procedure had been poorly understood. We expect that our new insights will pave the way for many new extensions to the mean shift algorithm.

APPENDIX

Please refer to the URL `http://www.cs.duke.edu/~mark/meanshift/` for supporting files, including functions to generate figures like those presented in this paper.

ACKNOWLEDGMENT

This work was supported by NSF grant IIS-0222516.

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, Aug. 1995.

[3] D. Comaniciu and P. Meer, "Mean shift analysis and applications," in *Proc. ICCV*, Sept. 1999, pp. 1197–1203.

[4] ——, "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. Ramanan and D. A. Forsyth, "Finding and tracking people from the bottom up," in *Proc. CVPR*, June 2003, pp. 467–474.

[6] ——, "Using temporal coherence to build animal models," in *Proc. ICCV*, Oct. 2003, pp. 338–346.

[7] R. Collins, "Mean-shift blob tracking through scale space," in *Proc. CVPR*, June 2003, pp. 234–240.

[8] R. M. Neal and G. E. Hinton, "Learning in graphical models," M. I. Jordan, Ed. Dordrecht: Kluwer Academic Publishers, 1998, ch. A view of the EM algorithm that justifies incremental, sparse, and other variants, pp. 355–368.

[9] G. Palumbo, G. Tanzella-Nitti, and G. Vettolani, *Catalogue of radial velocities of galaxies*. New York: Gordon and Breach, 1983.

[10] B. Georgescu, I. Shimshoni, and P. Meer, "Mean shift based clustering in high dimensions: A texture classification example," in *Proc. ICCV*, Oct. 2003, pp. 456–463.

Mark Fashing received the BS degree in computer science from The College of William and Mary, Williamsburg, Virginia, in 2001. He is a now a PhD candidate in computer science at Duke University, Durham, North Carolina. His current research interests are in computer vision and machine learning.

Carlo Tomasi received the DSc degree in electrical engineering from the University of Padova, Padova, Italy, in 1986 and the PhD degree in computer science from Carnegie Mellon University, Pittsburgh, Pennsylvania, in 1991. He is now an associate professor of computer science at Duke University, Durham, North Carolina. His research interests include computer vision, medical imaging, and applied mathematics. He is a member of the IEEE.