Mean Shift: Construction and Convergence Proof

Konstantinos G. Derpanis

November 26, 2006

In most low-level computer vision problems, very little information (if any) is known about the true underlying probability density function, such as its shape, number of mixture components, etc.. Due to this lack of knowledge, parametric approaches are less relevant, rather one has to rely on non-parametric methods. In this note we consider the construction and convergence proof of the non-parametric *mean shift* method which was developed by Fukunaga and Hostetler (Fukunaga & Hostetler, 1975), later adapted by Cheng (Cheng, 1995) for the purpose of image analysis and more recently popularized in the computer vision literature by Comaniciu and Meer (Comaniciu & Meer, 2002). The mean shift procedure represents a simple iterative non-parametric procedure for density mode seeking. For a general treatment of density estimation the reader is referred to (Silverman, 1998).

1 Mean shift construction

The main idea behind the mean shift procedure is to treat the points in the d-dimensional feature space as an empirical probability density function, where the densest regions in the feature space correspond to the local maxima or modes of the underlying distribution. For a set of locations in the space, one performs a gradient ascent procedure (i.e., mean shift) on the local estimated density (i.e., Parzen window), moves the window of analysis and repeats the procedure until convergence. Data points associated (at least approximately) with the stationary points are considered members of the same cluster.

Let $\{\mathbf{x}_i\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $i, \ldots, n$, be a set of points in Euclidean space. The kernel density is defined as

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

where $K(\mathbf{x})$ is termed the kernel and h defines the radial extend of the window (see Fig. 1).

Assuming that the kernel $K(\mathbf{x})$ is differentiable, an estimate of the density gradient can be defined as the gradient of the kernel density estimate (1):

$$\hat{\nabla} f(\mathbf{x}) \equiv \nabla \hat{f}(\mathbf{x}) = \frac{1}{h^d} \sum_{i=1}^{n} \nabla K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right). \tag{2}$$

Conditions on the kernel and window radius to guarantee asymptotic unbiasedness and consistency are given in (Fukunaga & Hostetler, 1975).

Using an Epanechnikov kernel (see Fig. 2)

$$K_E(\mathbf{x}) = \begin{cases} \frac{1}{2} c_d^{-1} (d+2)(1 - \|\mathbf{x}\|^2), & \text{if } \|\mathbf{x}\|^2 < 1 \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

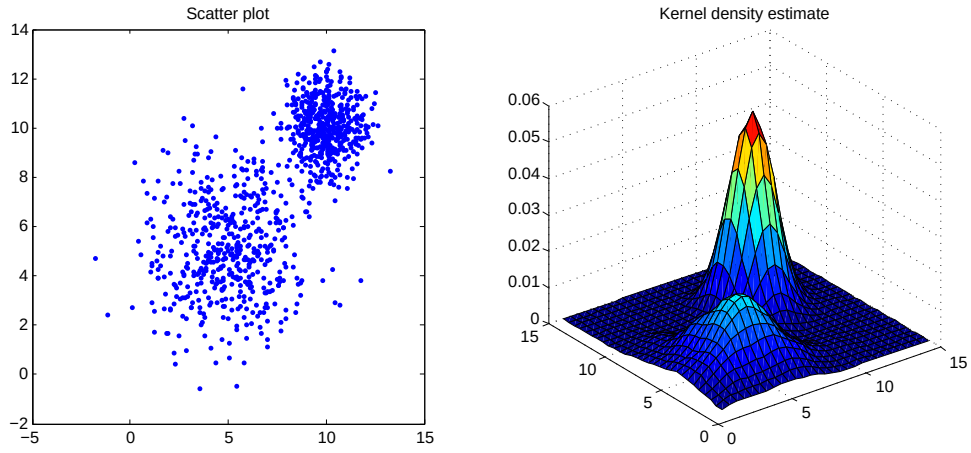

Figure 1: Kernel density estimation. (left) scatter plot of 1000 sample points from a multi-modal distribution and (right) kernel density estimate using the Epanechnikov kernel.

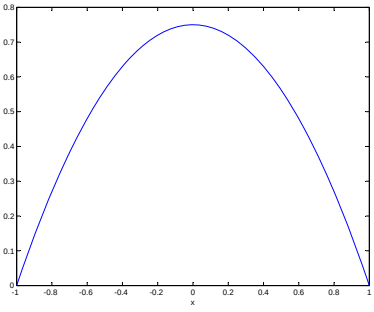

Figure 2: The one-dimensional Epanechnikov kernel.

where c_d is the volume of the d-dimensional sphere, the density gradient estimate is obtained from taking the gradient of (1):

$$\hat{\nabla} f(\mathbf{x}) \equiv \nabla \hat{f}(\mathbf{x}) = \frac{1}{n(h^d c_d)} \frac{d+2}{h^2} \left(\frac{1}{n_{\mathbf{x}}} \sum_{\mathbf{x} \in S(\mathbf{x};h)} (\mathbf{x}_i - \mathbf{x}) \right), \tag{4}$$

where $S(\mathbf{x}; h)$ denote the hypersphere of radius h containing $n_{\mathbf{x}}$ data points with volume $h^d c_d$.

The last term in (4) is termed the (sample) mean shift:

$$M(\mathbf{x}) \equiv \frac{1}{n_{\mathbf{x}}} \sum_{\mathbf{x}_i \in S(\mathbf{x};h)} (\mathbf{x}_i - \mathbf{x}) \tag{5}$$

$$= \left(\frac{1}{n_{\mathbf{x}}} \sum_{\mathbf{x}_i \in S(\mathbf{x};h)} \mathbf{x}_i \right) - \mathbf{x}. \tag{6}$$

The mean shift can be seen as simply computing the arithmetic mean of the points in the current window.

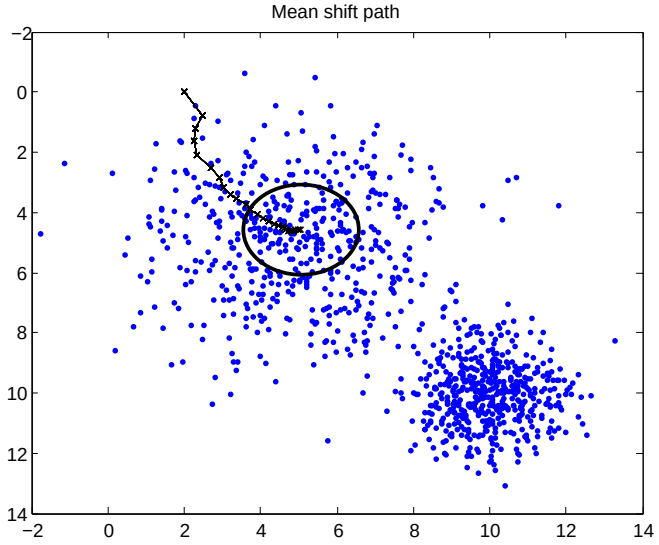

Figure 3: The path defined by successive mean shift computations culminating to a local density maxima.

Isolating the mean shift term in (4), yields,

$$M(\mathbf{x}) = \frac{h^2}{d+2}\frac{\hat{\nabla} f(\mathbf{x})}{\hat{f}(\mathbf{x})}. \tag{7}$$

One can interpret the mean shift vector (7) as a rescaled gradient having the direction of the gradient of the density. The rescaling has the effect of adapting the step size with respect to the local density, the lower the density the larger the step size, while the higher the local density the smaller the step size (Cheng, 1995). Given that the mean shift vector points in the direction of the maximum increase in density, suggests that the iterative translation of the data window by the mean shift vector defines a path leading to a local density maximum or mode (see Fig. 3). However, without a proof of convergence the simple reliance on the gradient direction is not sufficient since the procedure could result in repeated overshooting of the mode.

2 Convergence proof

In this section the proof of convergence of the Epanechnikov-based mean shift procedure is presented; the proof is adapted from (Comaniciu & Meer, 2002).

Theorem: *Let* $\hat{f} = \{\hat{f}(\mathbf{y}_j)\}$, $j = 1, 2, 3, \ldots$ *be the sequence of Epanechnikov-based density estimates computed along the path of points* $\mathbf{y}_j$ *defined by successive application of the mean shift procedure. The sequence is guaranteed to converge.*

Proof: To show that the mean shift is truly convergence, it is sufficient to demonstrate that $\hat{f}(\mathbf{x})$ is strictly monotonically increasing along the mean shift path, $\mathbf{y}_i$, or equivalently the difference between successive (non-convergence) density estimates is strictly positive, formally,

$$\mathbf{y}_i \neq \mathbf{y}_{i+1} \implies \hat{f}(\mathbf{y}_i) < \hat{f}(\mathbf{y}_{i+1}). \tag{8}$$

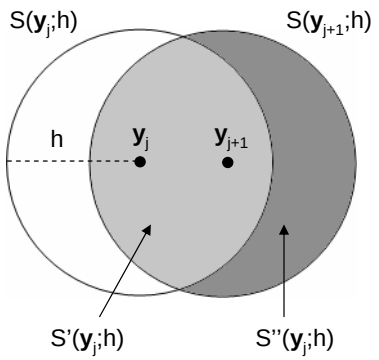

Figure 4: Successive mean shift windows.

The Epanechnikov-based density estimate is given by,

$$\hat{f}(\mathbf{y}_j) = \frac{1}{nh^d} \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} K_E\left(\frac{\mathbf{y}_j - \mathbf{x}_i}{h}\right) \tag{9}$$

$$= \frac{d+2}{2nh^d c_d} \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \left(1 - \frac{\|\mathbf{y}_j - \mathbf{x}_i\|^2}{h^2}\right), \tag{10}$$

and without loss of generality we can assume that the origin is located at y_j:

$$\hat{f}(\mathbf{y}_j) = \frac{d+2}{2nh^d c_d} \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \left(1 - \frac{\|\mathbf{x}_i\|^2}{h^2}\right). \tag{11}$$

Let $S'(\mathbf{y}_j;h) = S(\mathbf{y}_j;h) - S''(\mathbf{y}_j;h)$ and $S''(\mathbf{y}_j;h) = S(\mathbf{y}_j;h) \bigcap S'(\mathbf{y}_{j+1};h)$ be the d-dimensional windows illustrated in Fig. 4 with $n_j = n'_j + n''_j$, n'_j and n''_j be the number of points falling within the respective windows.

Since the Epanechnikov kernel is non-negative we have

$$\hat{f}(\mathbf{y}_{j+1}) \geq \frac{1}{nh^d} \sum_{\mathbf{x}_i \in S''(\mathbf{y}_j;h)} K_E\left(\frac{\mathbf{y}_{j+1} - \mathbf{x}_i}{h}\right) \tag{12}$$

$$= \frac{d+2}{2nh_d c_d} \sum_{\mathbf{x}_i \in S''(\mathbf{y}_j;h)} \left(1 - \frac{\|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2}{h^2}\right) \tag{13}$$

Next, based on the different sets defined above we have the following inequality

$$\hat{f}(\mathbf{y}_{j+1}) - \hat{f}(\mathbf{y}_j) = \frac{d+2}{2nh^d c_d} \left(\sum_{\mathbf{x}_i \in S(\mathbf{y}_{j+1};h)} \left(1 - \frac{\|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2}{h^2}\right) - \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \left(1 - \frac{\|\mathbf{x}_i\|^2}{h^2}\right) \right) \tag{14}$$

$$\geq \frac{d+2}{2nh^d c_d} \left(\sum_{\mathbf{x}_i \in S''(\mathbf{y}_j;h)} \left(1 - \frac{\|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2}{h^2}\right) - \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \left(1 - \frac{\|\mathbf{x}_i\|^2}{h^2}\right) \right) \tag{15}$$

$$= \frac{d+2}{2nh^d c_d h^2} \left(\sum_{\mathbf{x}_i \in S''(\mathbf{y}_j;h)} \left(h^2 - \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2\right) - \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \left(h^2 - \|\mathbf{x}_i\|^2\right) \right) \tag{16}$$

$$= \frac{d+2}{2nh^d c_d h^2} \left(\sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \|\mathbf{x}_i\|^2 - \sum_{\mathbf{x}_i \in S''(\mathbf{y}_j;h)} \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2 + \sum_{\mathbf{x}_i \in S''(\mathbf{y}_j;h)} h^2 - \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} h^2 \right) \tag{17}$$

$$= \frac{d+2}{2nh^d c_d h^2} \left(\sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \|\mathbf{x}_i\|^2 - \sum_{\mathbf{x}_i \in S''(\mathbf{y}_j;h)} \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2 + h^2(n''_j - n_j) \right). \tag{18}$$

Since $-n'_j = n''_j - n_j$, (18) can be rewritten as,

$$\hat{f}(\mathbf{y}_{j+1}) - \hat{f}(\mathbf{y}_j) \geq \frac{d+2}{2nh^d c_d h^2} \left(\sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \|\mathbf{x}_i\|^2 - \sum_{\mathbf{x}_i \in S''(\mathbf{y}_j;h)} \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2 - h^2 n'_j \right). \tag{19}$$

By construction, $\|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2 \geq h^2$ for all $\mathbf{x}_i \in S'(\mathbf{y}_j;h)$ which in turn implies that

$$\sum_{\mathbf{x}_i \in S'(\mathbf{y}_j;h)} \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2 \geq n'_j h^2. \tag{20}$$

Finally, substituting (6) and (20) into (18) yields,

$$\hat{f}(\mathbf{y}_{j+1}) - \hat{f}(\mathbf{y}_j) \geq \frac{d+2}{2nh^d c_d h^2} \left(\sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \|\mathbf{x}_i\|^2 - \sum_{\mathbf{x}_i \in S''(\mathbf{y}_j;h)} \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2 - \sum_{\mathbf{x}_i \in S'(\mathbf{y}_j;h)} \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2 \right) \tag{21}$$

$$= \frac{d+2}{2nh^d c_d h^2} \left(\sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \|\mathbf{x}_i\|^2 - \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \|\mathbf{y}_{j+1} - \mathbf{x}_i\|^2 \right) \tag{22}$$

$$= \frac{d+2}{2nh^d c_d h^2} \left(\sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} (\mathbf{y}_{j+1}^\top \mathbf{y}_{j+1} - 2\mathbf{y}_{j+1}^\top \mathbf{x} + \mathbf{x}_i^\top \mathbf{x}_i) \right) \tag{23}$$

$$= \frac{d+2}{2nh^d c_d h^2} \left(- \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} (\mathbf{y}_{j+1}^\top \mathbf{y}_{j+1} - 2\mathbf{y}_{j+1}^\top \mathbf{x}_i) \right) \tag{24}$$

$$= \frac{d+2}{2nh^d c_d h^2} \left(- \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} (\|\mathbf{y}_{j+1}\|^2 - 2\mathbf{y}_{j+1}^\top \mathbf{x}_i) \right) \tag{25}$$

$$= \frac{d+2}{2nh^d c_d h^2} \left(2\mathbf{y}_{j+1}^\top \sum_{\mathbf{x}_i \in S(\mathbf{y}_j;h)} \mathbf{x}_i - n_j \|\mathbf{y}_{j+1}\|^2 \right) \tag{26}$$

$$= \frac{d+2}{2nh^d c_d h^2} \left(2n_j \mathbf{y}_{j+1}^\top \mathbf{y}_{j+1} - n_j \|\mathbf{y}_{j+1}\|^2 \right) \tag{27}$$

$$= \frac{d+2}{2nh^d c_d h^2} \left(2n_j \|\mathbf{y}_{j+1}\|^2 - n_j \|\mathbf{y}_{j+1}\|^2 \right) \tag{28}$$

$$= \frac{d+2}{2nh^d c_d h^2} n_j \|\mathbf{y}_{j+1}\|^2. \tag{29}$$

Since (29) is strictly positive and hence monotonically increasing (unless $\mathbf{y}_j = \mathbf{y}_{j+1}$) and the domain is bounded, the mean shift sequence is guaranteed to converge.

References

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17(8)*, 790–799.

Comaniciu, D. & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24(5)*, 603–619.

Fukunaga, K. & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, *21(1)*, 32–40.

Silverman, B. (1998). *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall/CRC.