This article was originally published in a journal published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues that you know, and providing a copy to your institution's administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution’s website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

http://www.elsevier.com/locate/permissionusematerial
A note on the convergence of the mean shift

Xiangru Li, Zhanyi Hu*, Fuchao Wu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, PR China

Received 17 June 2005; received in revised form 25 September 2005; accepted 5 October 2006

Abstract

Mean shift is an effective iterative algorithm widely used in computer vision community. However, to our knowledge, its convergence, a key aspect of any iterative algorithm, has not been rigorously proved up to now. In this paper, by further imposing some commonly acceptable conditions, its convergence is proved.

Keywords: Mean shift; Convergence; Local structure; Computer vision

1. Introduction

The mean shift algorithm is a simple iterative statistical method introduced by Fukunaga and Hostetler [1], which shifts each data point to the weighted average of a sample set. The theory is studied further in Refs. [2–5]. In recent years, it has been widely applied in computer vision community [3,6], such as tracking, image segmentation, discontinuity preserving, smoothing, filtering, edge detection, etc.

Let \( \{x_i, 1 \leq i \leq n\} \) be an i.i.d. (independently and identically distributed) sample data set from probability density function \( f(x), x \in \mathbb{R}^m \). If \( f(x) \) is estimated by \( \hat{f}(x) = \sum_{i=1}^{n} w_j k(\beta \|x_i - x\|^2) \), Cheng [2] gave the mean shift procedure \( \{y_j, j = 1, 2, \ldots\} \) as the weighted averages of the samples \( \{x_i, 1 \leq i \leq n\} \) to seek the mode of \( \hat{f}(x) \), where \( w_i, w_i' > 0 \), are the weights of sample \( x_i \),

\[
\begin{align*}
w_i' &= w_i k'(\beta \|x_i - y_j\|^2) x_i / \sum_{i=1}^{n} w_i k'(\beta \|x_i - y_j\|^2), \\
\sum_{i=1}^{n} w_i' &= 1.
\end{align*}
\]

\( \beta > 0, k(x) \) is the profile function defined in Ref. [2] (sometimes called window or kernel), and \( k'(x) \) the differential of \( k(x) \). Cheng [2] proved the convergence of mean shift sequence \( \{y_j, j = 1, 2, \ldots\} \) under the following two assumptions:

1. \( k(x) = e^{-x} \).
2. The idealized mode in the density surface of random variable \( x \) is

\( q(x) = e^{-\gamma \|x\|^2}, \quad \gamma < \beta. \)

However, since the true value of \( \gamma \) is unknown, it is difficult to assure the above assumption (2) of being satisfied in real application. Hence, its applicability is limited to some extent. Refs. [3,7,8] attempted to prove the convergence of mean shift sequence \( \{y_j, j = 1, 2, \ldots\} \) under the assumption
that $k(x)$ is simply a convex and monotonically decreasing profile, and $w_i = 1/n$. But, as shown in the following, the proofs are incorrect in Refs. [3,7,8].

In Refs. [7,8], the proofs essentially depend on the wrong conclusion that $\|y_{j+1} - y_j\|$ converges to zero means \{$y_j, j = 1, 2, \ldots$\} converges. Here is a counter example:

**Counter Example 1.** Let $y_j = \sum_{i=1}^j 1/i$, then

$$\|y_{j+1} - y_j\| = \frac{1}{j+1} \to 0(j \to \infty).$$

However, it is well known that \{$y_j, j = 1, 2, \ldots$\} does not converge and is not a Cauchy sequence.

In the convergence proof of mean shift sequence \{$y_j, j = 1, 2, \ldots$\} in Ref. [3], the key step is

$$\|y_{j+m} - y_{j+m-1}\|^2 + \cdots + \|y_{j+1} - y_j\|^2 \geq \|y_{j+m} - y_j\|^2.$$  \hspace{1cm} (1)

However, inequality (1) does not hold. Here is a counter example:

**Counter Example 2.** Let $m = 2$, then

$$\|y_{j+2} - y_j\|^2 = \|y_{j+2} - y_{j+1} + y_{j+1} - y_j\|^2 = \|y_{j+2} - y_{j+1}\|^2 + \|y_{j+1} - y_j\|^2 + 2(y_{j+2} - y_{j+1})T(y_{j+1} - y_j).$$

From Theorem 2 in Ref. [3], the following inequality holds:

$$(y_{j+2} - y_{j+1})T(y_{j+1} - y_j) \geq 0.$$ Hence,

$$\|y_{j+2} - y_j\|^2 \geq \|y_{j+2} - y_{j+1}\|^2 + \|y_{j+1} - y_j\|^2.$$ It is in conflict with inequality (1). Let $y_0 = 1$, $y_1 = 2$ and $y_2 = 3$, then

$$\|y_{j+2} - y_j\|^2 = \|3 - 1\|^2 = 4$$

and

$$\|y_{j+2} - y_{j+1}\|^2 + \|y_{j+1} - y_j\|^2 = 1 + 1 = 2.$$ Therefore,

$$\|y_{j+2} - y_j\|^2 \geq \|y_{j+2} - y_{j+1}\|^2 + \|y_{j+1} - y_j\|^2.$$ In addition to the convergence problem, there are two other main limitations for the current mean shift algorithm:

1. No sufficient attention has been paid to the difference and the anisotropy of the local structure around different samples. For example, as shown in Fig. 1, since the sample distribution in the neighborhood of $x_2$ is denser than that of $x_1$, the scale for $x_2$ should ideally be smaller than that for $x_1$. In addition, the sample distribution is highly anisotropic in the neighborhood of $x_2$, and we should take it into account.

2. No sufficient attention has been paid to the difference of sample contributions. As we know, the peripheral samples, often more corrupted by noise, are less reliable. Hence, different samples should be ideally treated differently. In Refs. [3,7,8], the contributions are assumed to be same for all samples. In Ref. [2], although the contribution differences are considered, the local structure is not taken into account.

In the next section, we outline some means to extend the current mean shift algorithm and alleviate these two limitations by accounting for the anisotropy of local structure around every sample, the difference of relative scale and the relative importance/reliability between samples. In addition, the convergence of the iterative points \{$y_j, j = 1, 2, \ldots$\} and its function value \{$f(y_j), j = 1, 2, \ldots$\} of the extended algorithm are rigorously proved by adding a modest constraint in Section 3. The experiments results are given to evaluate the contribution of the proposed algorithm in Section 4. The conclusion and remarks are given in Section 5.

2. Preliminaries

**Definition 1.** Function $k(x)$ is called a bounded kernel if it, on $[0, +\infty)$, satisfies:

1. $k(x) \geq 0$.
2. monotonically decreasing: $k(x_1) \geq k(x_2)$, $0 \leq x_1 \leq x_2 < +\infty$. 

Fig. 1. Different local structures around different samples. Since the sample distribution in the neighborhood of $x_2$ is denser than that of $x_1$, the scale for $x_2$ should ideally be smaller than that for $x_1$. In addition, the sample distribution is highly anisotropic in the neighborhood of $x_2$, and we should take it into account. In Refs. [2,3,7], however, the relative scale and local structure are treated identically for all samples and in every direction. In Ref. [8], only the difference of relative scale between samples is accounted for.
In this work, given a bounded kernel \( k(x) \), the density estimation of random variable \( x \) is defined as
\[
\hat{f}_{H,k}(x) = \sum_{i=1}^{n} w_i K_i(x),
\]
where
\[
K_i(x) = c_k, i, h k(\|x - x_i\|_{H_i}^2),
\]
\[
\|x - x_i\|_{H_i}^2 = (x - x_i)^T H_i (x - x_i),
\]
\[
H_i = \Sigma_k^{-1} / h^2, \quad H = \{H_i, 1 \leq i \leq n\},
\]
\[
\sum_{i=1}^{n} w_i = 1, \quad w_i > 0.
\]

\( h > 0 \) is to adjust the window size on the whole, \( c_k, i, h > 0 \) is a constant to ensure that \( K_i(x) \) is a probability density function, \( w_i \) is the weight of sample \( x_i \), \( \Sigma_k \) is a positive definite matrix to account for the local structure around \( x_i \).

Thus, the iterative procedure of the mean shift is
\[
\begin{align*}
y_{j+1} &= \left( \sum_{i=1}^{n} L_i(y_j) \right)^{-1} \sum_{i=1}^{n} L_i(y_j) x_i \quad \text{if} \quad \sum_{i=1}^{n} L_i(y_j) \text{ is invertible,} \\
y_{j+1} &= y_j \quad \text{otherwise,}
\end{align*}
\]
where \( j > 0 \), and
\[
L_i(x) = -2 w_i c_k, i, h k(\|x - x_i\|_{H_i}^2) H_i
\]
\[
= 2 w_i c_k, i, h g(\|x - x_i\|_{H_i}^2) H_i,
\]
where \( g(x) = -k'(x) \).

**Remark 1.** Based on Theorem 5.59 in Ref. [9, pp. 187–188], any bounded kernel in Definition 1 is differentiable almost everywhere, that is to say, the Lebesgue measure of the set is zero on which the kernel is not differentiable. If we redefine \( k'(x) \) on the zero-measure set, we can assume that \( k'(x) \) exists wherever \( k(x) \) exists. Furthermore, since \( k(x) \) is monotonically decreasing, we can assume \( k'(x) \leq 0 \). Therefore, the mean shift procedure can be given in the closed form as shown in (3) and (4).

**Remark 2.** In fact, it was assumed \( \sum_{i=1}^{n} = h_i I_{m \times m} \) in Ref. [8], and \( \sum_{i=1}^{n} = I_{m \times m} \) in Refs. [2,3,7]. When all components of every sample \( x_i \) are uncorrelated and identically distributed, such assumptions are reasonable. However, local structure depends on the underlying spatial intensity of the samples, and manifests strong directionality and correlation. As a result, the mean shift procedure proposed in this work has the potential to alleviate the two limitations listed in Section 1 by properly specifying matrices \( \Sigma_i \). \( \Sigma_i \)s embody the local structure around every sample \( x_i \), including scale information and the anisotropy of the local structure. We should also point out that \( \Sigma_i \) has the potential to account for local structure, however, how to turn this potentiality into reality is still an open issue at the present stage, and should be further investigated.

### 3. Convergence

By convergence of mean shift, it is meant that both \( \{\hat{f}_{H,k}(y_j)\}, j = 1, 2, \ldots \) and the iterative sequence \( \{y_j, j = 1, 2, \ldots \} \) are all convergent.

**Definition 2.** A function \( k : [0, +\infty) \to R \) is convex if there exists a bounded and continuous \( k'(x) \) satisfying
\[
k(x_2) - k(x_1) > k'(x_1)(x_2 - x_1), \quad \forall x_1 \geq 0, \quad x_2 \geq 0, \quad x_1 \neq x_2.
\]

We have the following two theorems:

**Theorem 1.** If the kernel \( k(x) \) is convex, then the sequence \( \{\hat{f}_{H,k}(y_j)\}, j = 1, 2, \ldots \} \) converges and monotonically increases.

**Theorem 2.** If the kernel \( k(x) \) is convex, and the number of critical points of \( \hat{f}_{H,k}(x) \) is finite on \( S_0 = \{y_j | \hat{f}_{H,k}(y_j) \geq \hat{f}_{H,k}(y_i)\}, \) then the iterative sequence \( \{y_j, j = 1, 2, \ldots \} \) converges.

Before giving their formal proofs, we would at first point out:

**Remark 3.** Compared with the convergence conditions in Refs. [3,7,8], the only difference is that we need here additionally that “the number of critical points of \( \hat{f}_{H,k}(x) \) is finite on \( S_0 \)”. And it is not exigent and can always be satisfied in practice as the critical points usually represent the modes or classes in real applications.

**Remark 4.** If \( w_i = 1/n \) and \( \sum_{i=1}^{n} = I_{m \times m} \), then Theorem 1 is just Theorem 1 in Ref. [3].

**Proof of Theorem 1.** From Definition 1, kernel \( k(x) \) is bounded. Therefore, from equality (2), \( \hat{f}_{H,k} \) and \( \{\hat{f}_{H,k}(y_j), j = 1, 2, \ldots \} \) are all bounded. So, it suffices to show that \( \{\hat{f}_{H,k}(y_j), j = 1, 2, \ldots \} \) is monotonically increasing in order to prove its convergence. For any \( j \geq 1 \),

1. If \( y_j = y_{j+1} \), it is evident that \( \hat{f}_{H,k}(y_{j+1}) \geq \hat{f}_{H,k}(y_j) \) from Definition 1.
(2) Otherwise, $y_j \neq y_{j+1}$. Then from the definition of mean shift procedure in (3) and (4), $\sum_{i=1}^{n} L_i(y_j)$ is invertible. Thus, from (2) and (6), there exists

$$\hat{f}_{H,k}(y_{j+1}) - \hat{f}_{H,k}(y_j)$$

$$\geq \sum_{i=1}^{n} w_i c_{i,k,h} g(\|y_j - x_i\|_{H_i}^2)\|y_j + 1 - y_j\|_{H_i}^2$$

and since $g(x) = -k'(x)$, there is

$$\hat{f}_{H,k}(y_{j+1}) - \hat{f}_{H,k}(y_j)$$

$$\geq \sum_{i=1}^{n} w_i c_{i,k,h} g(\|y_j - x_i\|_{H_i}^2)\|y_j - x_i\|_{H_i}^2$$

At the same time,

$$\|y_{j+1} - x_i\|_{H_i}^2$$

$$= \|y_{j+1} - y_j + y_j - x_i\|_{H_i}^2$$

$$= \|y_{j+1} - y_j\|_{H_i}^2 + 2(y_{j+1} - y_j)^T H_i (y_j - x_i)$$

$$+ \|y_j - x_i\|_{H_i}^2$$

and

$$\|y_j - x_i\|_{H_i}^2 - \|y_{j+1} - x_i\|_{H_i}^2$$

$$= -\|y_{j+1} - y_j\|_{H_i}^2 + 2(y_{j+1} - y_j)^T H_i (x_i - y_j).$$

Hence,

$$\hat{f}_{H,k}(y_{j+1}) - \hat{f}_{H,k}(y_j)$$

$$\geq \sum_{i=1}^{n} w_i c_{i,k,h} g(\|y_j - x_i\|_{H_i}^2)\|y_{j+1} - y_j\|_{H_i}^2$$

$$\|y_{j+1} - y_j\|_{H_i}^2$$

$$+ 2(y_{j+1} - y_j)^T H_i (x_i - y_j).$$

From (3) and (5)

$$\sum_{i=1}^{n} w_i c_{i,k,h} g(\|y_j - x_i\|_{H_i}^2)(y_{j+1} - y_j)^T H_i (x_i - y_j)$$

$$= \sum_{i=1}^{n} w_i c_{i,k,h} g(\|y_j - x_i\|_{H_i}^2)(y_{j+1} - y_j)^T H_i y_{j+1}$$

$$- \sum_{i=1}^{n} w_i c_{i,k,h} g(\|y_j - x_i\|_{H_i}^2)(y_{j+1} - y_j)^T H_i y_j$$

$$= \sum_{i=1}^{n} w_i c_{i,k,h} g(\|y_j - x_i\|_{H_i}^2)\|y_{j+1} - y_j\|_{H_i}^2.$$
Based on Definition 2, we know $k'(x)$ is a bounded function, and $\sum_{i=1}^{n}L_i(y_j)$ is also bounded from (5). Therefore,

$$\nabla \hat{f}_{H,k}(y_j) \overset{(2)}{=} \sum_{i=1}^{n} w_i \nabla K_i(y_j)$$

$$= 2 \sum_{i=1}^{n} w_i c_{k,i,h} H_i(y_j - x_i) k'(\|y_j - x_i\|^2_{H_i})$$

$$= \sum_{i=1}^{n} (y_j - x_i) L_i(y_j)$$

$$= \sum_{i=1}^{n} L_i(y_j) x_i - \sum_{i=1}^{n} L_i(y_j) y_j$$

$$= \left( \sum_{i=1}^{n} L_i(y_j) \right) \left[ \left( \sum_{i=1}^{n} L_i(y_j) \right)^{-1} \times \sum_{i=1}^{n} L_i(y_j) x_i - y_j \right]$$

$$\overset{(3)}{=} \sum_{i=1}^{n} L_i(y_j) [y_{j+1} - y_j] \rightarrow 0 (j \rightarrow \infty).$$

(11)

Because the number of critical points of $\hat{f}_{H,k}(x)$ is finite on $S_0$, without loss of generality, we assume there are $m_0$ critical points $\{x^*_k, 1 \leq k \leq m_0\}$:

$$\nabla \hat{f}_{H,k}(x^*_k) = 0, \ 1 \leq k \leq m_0$$

and

$$\nabla \hat{f}_{H,k}(x) \neq 0, \ x \in S_0 \text{ but } x \neq x^*_k, 1 \leq k \leq m_0.$$ Let

$$d_0 \overset{\Delta}{=} \min \{\|x^*_j - x^*_k\|, \ 1 \leq j \neq k \leq m_0\},$$

$$S_{k,i} = \{x | \|x - x^*_k\| < \varepsilon, \ x \in S_0\}, \ 1 \leq i \leq m_0,$$

where $0 < \varepsilon < d_0/3$. From Definition 2 and (11), we know $\nabla \hat{f}_{H,k}(x)$ is continuous and $\nabla \hat{f}_{H,k}(x) \neq 0$ on the bounded closed set $V_0 = S_0 - \bigcup_{i=1}^{m_0} S_{k,i}$. Therefore, $\min_{x \in V_0} \|\nabla \hat{f}_{H,k}(x)\| \neq 0$, and there exists $c_\varepsilon > 0$ satisfying

$$\|\nabla \hat{f}_{H,k}(x)\| > c_\varepsilon \text{ for any } x \in V_0.$$

(13)

From (10) and (11), there exists $N_\varepsilon > 0$ satisfying the following two inequalities simultaneously:

$$\|y_{j+1} - y_j\| < \varepsilon, \ j \geq N_\varepsilon,$$

$$\|\nabla \hat{f}_{H,k}(y_j)\| < c_\varepsilon, \ j \geq N_\varepsilon$$

(14) and (15) mean $\{y_j, j \geq N_\varepsilon\} \subseteq \bigcup_{i=1}^{m_0} S_{k,i}$. Let

$$x^*_1 \in S_{k,i_1}, \ x^*_2 \in S_{k,i_2}, \ 1 \leq i_1 \neq i_2 \leq m_0$$

then

$$\|x^*_1 - x^*_2\| = \|x^*_1 - x^*_1 + x^*_1 - x^*_2 + x^*_2 - x^*_2\|$$

$$\geq \|x^*_1 - x^*_2\| - \|x^*_1 - x^*_1\| - \|x^*_2 - x^*_2\|$$

$$\geq d_0 - \varepsilon - \varepsilon$$

$$= \varepsilon.$$

Therefore, (14) means that $\{y_j, j \geq N_\varepsilon\}$ can only lie inside one of the neighborhoods defined around every critical point, say, $S_{k,i_0}$, and $\|y_j - x^*_0\| < \varepsilon$. In other words, $\{y_j, j = 1, 2, \ldots\}$ converges.

4. Experiments

To evaluate the contribution of the work, we conducted clustering experiments by mean shift algorithm. In this section, we report and analyze the experimental results on simulated data (Fig. 2). For convenience of presentation, $MS(\sum_i, \{w_i\}, h)$ represent the mean shift procedure with local structure $\sum_i$, weight $\{w_i\}$ and scale $h$.

In the experiments, the local structure $\sum_i$ and weight $w_i$ are estimated by

$$\hat{w}_i = 1 / |\hat{\sum}_i|,$$

(17)

![Fig. 2. Simulated data set. Several clustering experiments will be done on it to evaluate the gains of introducing local structure $\sum_i$ and weight $w_i$ for every sample $x_i$.](image-url)
4.1. Weighted or unweighted

To evaluate the advantages of assigning a weight for every sample, we conducted experiments under $\sum_{l} = \hat{\Sigma}$ and $\sum_{l} = l_{m \times m}$. The results are presented in Fig. 3. In the two conditions, the weighted mean shift converges faster than the unweighted. Usually, samples are sparser and less reliable in peripheral area than around cluster center. As in (17), therefore, we can assign a less weight for peripheral sample and a larger one for the sample around cluster center. The weights embody the samples’ relative importance and contributions. And appropriate weights can enhance the describing power and convergence rate of mean shift.

4.2. Local structure

In this subsection, two experiments are reported which were conducted, respectively, with the improved mean shift (IMS) and the traditional one. The results are presented in

Fig. 3. Comparison between weighted mean shift and the unweighted. Abscissa ($x$-axis) is the iterative number, and ordinate ($y$-axis) is the number of clusters. The dashed line is the clustering result of unweighted mean shift, the solid is that of weighted mean shift. (a) MS($\hat{\Sigma}$, $\hat{w}_i$, 4) vs. MS($\hat{\Sigma}$, $1/n$, 4); (b) MS($l_{m \times m}$, $\hat{w}_i$, 0.3) vs. MS($l_{m \times m}$, $1/n$, 0.3).

Fig. 4. Evaluating the contribution of introducing an anisotropic local structure for every sample. It is the clustering results after six time iterative by mean shift. And the black points are the initial sample data, the red stars are the clustering. (a) MS($\hat{\Sigma}$, $1/n$, 4); (b) MS($l_{m \times m}$, $1/n$, 0.3).
In Fig. 4(a), the clustering points lie more among the dense samples, and along the direction of sample distribution. However, in Fig. 4(b), many clusters appear in the blank area. Therefore, the clusters of IMS reflect the characteristics of sample distribution more accurately. By introducing an appropriate positive matrix $\sum_i$ for every sample, the mean shift can describe problems more accurately.

5. Conclusion and discussions

In this paper, we at first extended the mean shift algorithm by introducing an arbitrary positive definite matrix to account for the difference and the anisotropy of the local structure around different samples and by assigning a weight for every sample to account for its relative importance and reliability. Most of all, we gave a rigorous convergence proof for the extended algorithm. The convergence conditions in this work can be easily satisfied and slightly different from those in Refs. [2,3,7,8]. And experimental results showed the superiority of the improved algorithm.

Acknowledgments

The authors are appreciative of the inspiring comments from the anonymous reviewers. This work was supported by the National Natural Science Foundation of China under the Grant No. 60121302, the National Key Basic Research and Development Program (2004CB318107).

References


About the Author—XIANGRU LI is a teacher at Shandong University of Science and Technology. He received his Ph.D. degree in 2006 from Institute of Automation, Chinese Academy of Sciences. His research interests are now in statistical machine learning, data mining.

About the Author—ZHANYI HU is a professor at Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree (Docteur d’Etat) in computer vision from University of Liege, Belgium, in 1993. His research interests include camera calibration and 3D reconstruction, vision guided robot navigation, and image based modeling and rendering.

About the Author—FUCHAO WU is a professor at Institute of Automation, Chinese Academy of Sciences. His research interests are now in computer vision, which include 3D reconstruction, active vision, and image based modeling and rendering.