Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 1756–1762
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

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

$$y_{j+1} = \sum_{i=1}^{n} w_i^j x_i$$

to seek the mode of $\hat{f}(x)$, where w_i, $w_i^j > 0$, are the weights of sample x_i,

$$w_i^j = w_i k'(\beta\|x_i - y_j\|^2) x_i \Big/ \sum_{i=1}^{n} w_i k'(\beta\|x_i - y_j\|^2),$$

$$\sum_{i=1}^{n} w_i^j = 1.$$

$\beta > 0$, $k(x)$ 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) $k(x) = \mathrm{e}^{-x}$.
(2) The idealized mode in the density surface of random variable x is

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

However, since the true value of γ 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

* Corresponding author. Tel.: +86 10 62616540; fax: +86 10 62551993.
E-mail addresses: xrli@nlpr.ia.ac.cn (X. Li), huzy@nlpr.ia.ac.cn (Z. Hu), fcwu@nlpr.ia.ac.cn (F. Wu).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.10.016

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

$$\|y_{j+m} - y_{j+m-1}\|^2 + \cdots + \|y_{j+1} - y_j\|^2 \geqslant \|y_{j+m} - y_j\|^2. \quad (1)$$

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

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

$$\begin{aligned}\|y_{j+2} - y_j\|^2 &= \|y_{j+2} - y_{j+1} + y_{j+1} - y_j\|^2 \\ &= \|y_{j+2} - y_{j+1}\|^2 + \|y_{j+1} - y_j\|^2 \\ &\quad + 2(y_{j+2} - y_{j+1})^{\mathrm{T}}(y_{j+1} - y_j).\end{aligned}$$

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

$$(y_{j+2} - y_{j+1})^{\mathrm{T}}(y_{j+1} - y_j) \geqslant 0.$$

Hence,

$$\|y_{j+2} - y_j\|^2 \geqslant \|y_{j+2} - y_{j+1}\|^2 + \|y_{j+1} - y_j\|^2.$$

It is in conflict with inequality (1). Let $y_j = 1$, $y_{j+1} = 2$ and $y_{j+2} = 3$, 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,

$$\|y_{j+2} - y_j\|^2 \geqslant \|y_{j+2} - y_{j+1}\|^2 + \|y_{j+1} - y_j\|^2.$$

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. 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.

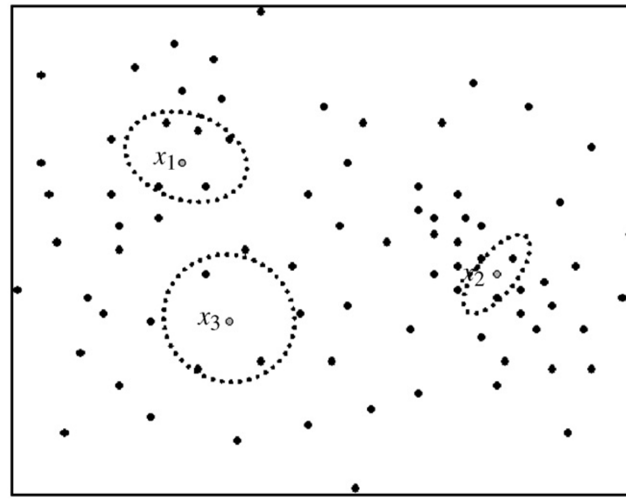

Fig. 1. Different local structures around different samples. Since the sample distribution in the neighborhood of x_2 is denser than that of x_1, the scale for x_2 should ideally be smaller than that for x_1. In addition, the sample distribution is highly anisotropic in the neighborhood of x_2, and we should take it into account.

(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 $\{\hat{f}(y_j), j = 1, 2, \ldots\}$ 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) $k(x) \geqslant 0$.
(2) monotonically decreasing: $k(x_1) \geqslant k(x_2)$, $0 \leqslant x_1 \leqslant x_2 < +\infty$.