Effective measurement selection in truncated Kernel density estimator

[Voronoi Mean Shift algorithm for truncated kernels]

Ji Won Yoon
Statistics Department,
Trinity College Dublin
Ireland
yoonj@tcd.ie

Hyoung-joo Lee
Department of Engineering Science
University of Oxford
United Kingdom
imhjlee@gmail.com

Hyoungshick Kim
Computer Laboratory
University of Cambridge
United Kingdom
hk331@cam.ac.uk

ABSTRACT

The Gating/Truncation technique is adapted to choose relatively significant measurements rather than all measurements to speed up mean shift algorithm which is one of the well-known clustering algorithms in the field of computer vision. The conventional mean shift algorithm can be sensitive to selecting measurements since the measurements are truncated with a Gaussian window of a fixed size. In particular when a small gating window is selected, it cannot properly cluster data points located far from major clusters and thus it generates unwanted, small clusters. We present a robust gating technique for truncated mean shift algorithm based on a geometric structure called Voronoi diagram of a given data set. Unlike conventional gating/truncation techniques our proposed truncation technique can provide non-linear truncation windows with variable sizes constructed by using the Voronoi diagram to effectively identify outlier points in clusters. We also demonstrate the feasibility of this technique by applying it on synthetic and real-world image data sets. The experimental results show that the proposed truncation technique provides a more robust clustering result compared to the conventional truncation techniques. The proposed algorithm can be effectively applied to denoising of images by removing background noise.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Miscellaneous; I.5.3 [**Clustering**]: Miscellaneous

General Terms

Machine Intelligence, Data mining, Image processing and vision

ICUIMC '11, February 21-23, 2011, Seoul, Korea

Keywords

Clustering, Mean shift, Truncated Gaussian kernel, Voronoi diagram, Image processing

1. INTRODUCTION

Consider a set of n data points $\{\mathbf{x}_i\}_{i=1}^n$ in the d-dimensional space $\mathbb{R}^d$. A multivariate kernel density estimator [Scott, 1992], based on the Parzen window technique, with a kernel $K(\cdot)$ and a bandwidth h, is given by

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

Since it is time consuming to consider n measurements, the mean shift (MS) algorithm is commonly used with truncated kernels with the gating technique. In this paper, the truncated Gaussian kernel is considered for the gating scheme:

$$K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) = \begin{cases} c_{k,d}\exp\left(-\frac{1}{2}\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2\right), & \text{if } \|\frac{\mathbf{x}-\mathbf{x}_i}{h}\| \le 1, \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where the term $c_{k,d}$ is a normalisation constant so that the density integrates to one. Modes of the density function are located at stationary points where the gradient of the density function is zero, i.e. $\nabla f(\mathbf{x}) = 0$.

The mean shift (MS) algorithm [Fukunaga, 1990, Fukunaga and Hostetler, 1975, Comaniciu and Meer, 2002] has been used for finding stationary points. Given a starting point, the MS procedure is iteratively implemented based on a MS vector that is calculated using a gradient estimate. It has been shown that the procedure is guaranteed to converge to a stationary point. A region that converges to the same stationary point defines the *basin of attraction*, in which the data points form one cluster. In this sense, the MS algorithm is a non-parametric statistical clustering method. As neither does it require prior knowledge of the number of clusters nor constrains the shape of the clusters, it has been widely used in image processing and computer vision applications.

Returning to the gating technique, only τ measurements that have significant effects on the mean are selected where $n >> \tau$. The simple and conventional approach to select the small number of measurements is designed with a bandwidth. To provide a simple understanding we use a fixed bandwidth based MS rather than adaptive MS in this paper, but it is not limited to the fixed bandwidth based MS.

If we use the Gaussian Kernel, then the standard deviation is generally regarded as a bandwidth. Thus, let r be the radius of the truncation window then it can be formed by the bandwidth: $r = u(h)$ for any linear or non-linear function $u(\cdot)$. For instance, for $r = h, 2h, 3h$, the confidence intervals are 68%, 95%, and 99% respectively. In other words, if $r = k$ for Gaussian kernel, then the truncated MS will select the only measurements which are with 68% confidence. The rest measurements are ignored so teh truncated MS will lose the 32% confidence and a lot of information. In addition, in case the significant measures are located at outliers (outside of the gate), the MS will suffer from more information loss. This is why measurement selection scheme is important for the MS with truncation.

The rest of this paper is structured as follows. We describe background of the MS algorithm in Section 2, then introduce our proposed algorithm (VMS) by extending the MS algorithm with Voronoi diagram of a given data set in Section 3. In Sections 4, we experimentally evaluate the performance of the VMS against the MS on synthetic and real-world data sets, respectively. Finally, we discuss several issues and conclude this paper in Section 5.

2. MEAN SHIFT ALGORITHM

The mean shift (MS) algorithm, based on the Parzen window technique, is a non-parametric statistical method, widely applied to image processing and computer vision. Given n data points $\{\mathbf{x}_i\}_{i=1}^n$ in the d-dimensional space $\mathbb{R}^d$, the multivariate kernel density estimator with a kernel $K(\cdot)$ and a symmetric positive definite $d \times d$ bandwidth matrix $\mathbf{H}$ is given by

$$\hat{f}_{\mathbf{H},K}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i), \tag{3}$$

where the kernel is defined as

$$K_{\mathbf{H}}(\mathbf{x}) = \frac{1}{|\mathbf{H}|^{1/2}} K(\mathbf{H}^{-1/2}\mathbf{x}). \tag{4}$$

If we assume independency and isotropy between dimensions, we have $\mathbf{H} = h\mathbf{I}_d$, where $\mathbf{I}_d$ is a d-dimensional identity matrix. Then, the density estimator of Eq. (3) becomes identical to Eq. (1).

Although the mean squared error between the true density and its estimate is a known measure of a kernel density estimator, only an asymptotic approximation of this measure (AMISE; asymptotic mean integrated squared error) can be computed in practice. As the number of data points $n \to \infty$, the bandwidth $h \to 0$ at a rate slower than n^{-1}. To minimise the AMISE measure, we can use one simple, radially symmetric kernel which is called the truncated Gaussian kernel [Scott, 1992] in Eq. (2). Its profile can be defined as

$$K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) = c_{k,d}\, k\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right). \tag{5}$$

From Eqs. (2) and (5), we have

$$k\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right) = \begin{cases} \exp\left(-\frac{1}{2}\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2\right), & \text{if } \left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\| \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Replacing the kernel in Eq. (1) by its profile in Eq. (5), we have the density estimator [Bradski, 1998] by

$$\hat{f}_{h,K}(\mathbf{x}) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right). \tag{7}$$

A natural estimator of the gradient of f is the gradient of $\hat{f}_{h,K}(\mathbf{x})$

$$\begin{aligned} \hat{\nabla} f_{h,K}(\mathbf{x}) &= \frac{2c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n}(\mathbf{x} - \mathbf{x}_i)k'\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right) \\ &= \frac{2c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n}(\mathbf{x}_i - \mathbf{x})g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right) \\ &= \frac{2c_{k,d}}{nh^{d+2}}\left[\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)\right] \\ &\quad \times \left[\frac{\sum_{i=1}^{n}\mathbf{x}_i g\left(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2\right)}{\sum_{i=1}^{n} g\left(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2\right)} - \mathbf{x}\right], \end{aligned} \tag{8}$$

where we denoted $g(x) = -k'(x)$ and $\sum_{i=1}^{n} g\left(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2\right)$ is assumed to be positive. Eq. (8) has two significant terms [Comaniciu and Meer, 2002]. The first term of the product in Eq. (8) is proportional to the density estimate at $\mathbf{x}$ computed with a profile $g(\cdot)$. The second term is the mean shift, the difference between the weighted mean and $\mathbf{x}$, and is defined as

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{\sum_{i=1}^{n}\mathbf{x}_i g\left(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2\right)}{\sum_{i=1}^{n} g\left(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2\right)} - \mathbf{x}. \tag{9}$$

From Eqs. (8-9), we can obtain the following equation:

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{1}{2}h^2\frac{c_{g,d}}{c_{k,d}}\frac{\hat{\nabla} f_{h,K}(\mathbf{x})}{\hat{f}_{h,G}(\mathbf{x})}. \tag{10}$$

It is now clear that the MS vector computed with a kernel G is proportional to the normalised density gradient estimate obtained with a kernel K. This also shows that the MS vector always moves toward the direction of maximum increase in the density function [Fukunaga, 1990, Fukunaga and Hostetler, 1975, Comaniciu and Meer, 2002].

Let $\mu_i^t, t = 1, 2, \ldots$ be a sequence of successive locations of the kernel K, starting from each data point, i.e. $\mu_i^1 = \mathbf{x}_i$. Then, for $t = 2, 3, \ldots$, the current location of the kernel is updated based on the previous location and the MS vector:

$$\mu_i^t = \mu_i^{t-1} + \mathbf{m}_{h,G}(\mu_i^{t-1}). \tag{11}$$

This MS procedure is guaranteed to converge at a nearby stationary point where the gradient estimate is zero. A region that converges to the same stationary point is called the basin of attraction. A cluster is defined by those data points in the same basin of attraction.

In general, one of key issues in mean shift algorithm is to select the bandwidth h optimally. Much attention has been paid to optimal bandwidth selection [Fukunaga, 1990, Silverman, 1986] and adaptive MS algorithms, which associate different bandwidths to different data points [Comaniciu et al., 2001, Georgescu et al., 2003]. However, this is rather out of the interest of this paper since our paper propose not bandwidth optimation algorithms but a new measurement selection scheme given a particular bandwidth. In other words, sSince this paper focusses only on the measurement selection scheme for the truncated kernel of mean shift, we use the simple mean shift algorithm with a fixed bandwidth for clear comparison. Note that our measurement selection scheme can be also embedded in the adaptive mean shift algorithm after slightly modifying it.

3. MEAN SHIFT CLUSTERING USING VORONOI DIAGRAMS

Performance of the MS algorithm totally depends on a particular selection of the radius of the truncated kernel. However, too small a radius will lead to too many small clusters. In addition, the kernel becomes more sensitive to noise in such a small window size. Oppositely, too large a radius brings time consuming process to the calculation. In order to alleviate the sensitivity to radius size, we propose a modified mean shift algorithm, the Voronoi mean shift (VMS). It inherits all properties of the conventional MS algorithm but it provides better measurement selection schemes when MS adopts the truncated Gaussian kernel. In this section, a new kernel based on a Voronoi diagram is introduced, and then the proposed algorithm is introduced.

3.1 Voronoi Kernel

The conventional MS algorithm has a discontinuity problem; data points outside a window defined by the bandwidth are not considered in calculating the MS vector since $g(\cdot) = 0$. This discontinuity may result in misdirected MS vectors and undesirable local optima when a small radius is used, especially in a sparse region where there are a small number of data points. Therefore, it is considered vulnerable to outliers and tends to result in many small clusters which contain only a few data points. While we can circumvent this problem by finding an optimal radius, such an optimisation task is difficult to solve and time-consuming .

We propose to use a Voronoi diagram to alleviate the discontinuity problem. We first define the Voronoi diagram as follows: Let $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ be a set of n data points (called sites). We define the Voronoi diagram of $\mathbf{x}$ as the subdivision of the space into n regions, one for each site in $\mathbf{x}$, with the property that a point s lies in the region corresponding a site $\mathbf{x}_i$ if and only if

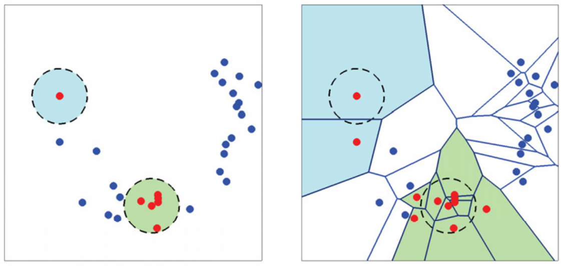
$$\|\mathbf{x}_i - s\| \leq \|\mathbf{x}_j - s\|, \forall j \neq i. \tag{12}$$

The region in the Voronoi diagram corresponding to a site $\mathbf{x}_i$ is denoted $V(\mathbf{x}_i)$; we call it the *Voronoi region* of $\mathbf{x}_i$. The basic properties of the Voronoi diagram are introduced well in [de Berg et al.,]. In our proposed method, a window is defined as Voronoi regions that are located within *or* overlapped with a hypersphere that are centred at the current location of the kernel and whose radius is the bandwidth (see Fig. 1(b)). Denoting such a window for a set of data points $\mathbf{x}$ by $\mathcal{S}(\mathbf{x})$, the resulting *Voronoi kernel* is now written as

$$K\left(\frac{\mathbf{x}-\mathbf{x}_i}{r}\right) = \begin{cases} c_{k,d} \exp\left(-\frac{1}{2}\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2\right), & \text{if } \mathbf{x}_i \in \mathcal{S}(\mathbf{x}) \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Note that the kernel is no longer radially symmetric kernel. We expect that the Voronoi kernel in effect plays a role of adapting the bandwidth with respect to the density of data. In other words, a small window is defined in a sparse region while a larger window in a dense region.

Fig. 1 illustrates the difference between the conventional MS and the VMS. The MS (Fig. 1(a)) defines a window as a circle with the radius being the bandwidth. The window would include a sufficient number of data points in a dense region (the green circle). In a sparse region (the cyan circle), however, only one data point is located within the window. In this case, the MS vector would be zero and in turn the algorithm would converge trivially to the data point. As a result, the data point itself would constitute a single cluster. While increasing the bandwidth would reduce sensitivity to outliers, it may also introduce the over-smoothing or blurring effect.



(a) Conventional kernel (b) Voronoi kernel

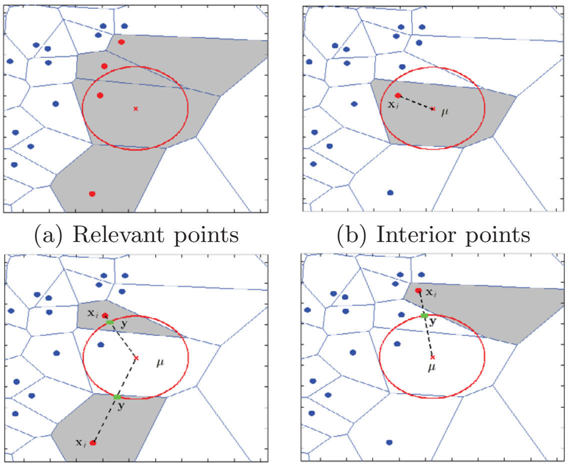
Figure 1: Windows for the conventional kernel and the Voronoi kernel given an identical bandwidth: Cyan and green shades represent the windows. Red points are selected by each kernel while blue points are not.

The VMS can alleviate this limitation. As the window is defined by the Voronoi regions, neighbouring points near an outlier can be included in the window as indicated by the cyan region in Fig. 1(b). It is equivalent to using an increased bandwidth in a sparse region. In a dense region (the green region), on the other hand, only the data points just outside the sphere are included in the window. In this case, the bandwidth is kept essentially the same as in the conventional MS.

3.2 Voronoi Mean Shift Algorithm

Finding which points are located within the Voronoi window is not a trivial task. There are two types of relevant points: interior points (points inside the hypersphere) and outer points (points outside the hypersphere but inside the relevant Voronoi regions) in Fig. 2. Two steps regarding finding relevant points are involved in the proposed algorithms: searching for interior points and searching for outer points (see Fig. 2(a-d)). By definition, the distance between an interior point $\mathbf{x}_i$ and a centre μ should be shorter than the given bandwidth such that $||\frac{\mathbf{x}_i-\mu}{\mathbf{r}}|| < 1$ where $\mathbf{r}$ is the radius of the truncated windows and we simply use $\mathbf{r} = \mathbf{h} = h\mathbf{I}$ in this paper. We can efficiently find interior points using a range search tree. For the d-dimensional space, this process runs in $O(\log^{d-1} n + m)$ time and $O(n \log^{d-1} n)$ space where m is the size of the interior points and n is the overall points [de Berg et al.,]. However, finding the outer points is not trivial. The most intuitive approach is to check whether each Voronoi region intersects the hypersphere. In a low-dimensional space ($d = 2$ or 3), this process can be implemented easily. However, in a higher-dimensional space ($d > 3$), the equation to check the intersection is complicated. Therefore we propose two practical approximation techniques.

The first technique is to consider the proximate points to



(a) Relevant points (b) Interior points

(c) Outer points (case 1) (d) Outer points (case 2)

Figure 2: Finding relevant points.

the hypersphere only. Even if the Voronoi region of an extremely far outer point may intersect with the hypersphere, it is not desirable to consider this point for clustering. Therefore we consider the only points within twice the size of the given radius as candidates. The threshold for choosing candidates can be adapted depending on applications.

The second technique is to ignore "the case 2" which is shown in Fig. 2(d). As we rarely observe such a case in practice, considering it is not practically meaningful. When we do not consider this case, the checking process for the outer point can be not only easily but also efficiently implemented even for a high-dimensional space. Let μ, $\mathbf{x}_i$ and $\mathbf{y}$ represent a centre, an outer point and a crossing point between the centre and the outer point as shown in Fig. 2 (c). With a radius, r, we can specify the crossing point as

$$\mathbf{y} = \mu + r \cdot \frac{\mathbf{x}_i - \mu}{\|\mathbf{x}_i - \mu\|}. \quad (14)$$

Then, we can just check whether a point $\mathbf{x}_i$ is a relevant outer point for a centre μ by evaluating the crossing point $\mathbf{y}$ is located within the Voronoi region of $\mathbf{x}_i$.

Given a data set $\{\mathbf{x}_i\}_{i=1}^n$ and a radius r, the VMS algorithm is as follows. An initial centre of the kernel is set for each data point $\mathbf{x}_i$. Relevant points are found, as shown in Fig. 2, for the centre, which is updated based on the MS vector computed using the Voronoi kernel. This procedure is iterated until convergence where a stationary point $\tilde{\mu}_i$ is found for each data point. After all the stationary points are found, we have clusters such that

$$\mathcal{C}_k = \{\mathbf{x}_i : \tilde{\mu}_i = \mathbf{c}_k\}, k = 1, 2, \ldots, K, \quad (15)$$

where $\{\mathbf{c}_k\}_{k=1}^K$ is a set of unique stationary points and K is the number of such points. The VMS is outlined in Algorithm 1. While we present a case of radially symmetric radius for brevity, it is straightforward to extend it to a case where different radius are used for different dimensions, i.e. $\mathbf{r} = [r_1, r_2, \ldots, r_d]$.

Algorithm 1 Voronoi mean shift

INPUT: A data set $\{\mathbf{x}_i\}_{i=1}^n$ and a radius r

FOR $i = 1$ to n **DO**
  $t = 1$
  Initialise $\mu_i^t = \mathbf{x}_i$
  REPEAT
    Search for relevant points for μ_i^t //Fig. 2
    Evaluate the Voronoi kernel for the points //Eq. (13)
    Calculate the mean shift vector //Eq. (9)
    Update the centre of the kernel //Eq. (11)
    $t \leftarrow t + 1$
  UNTIL convergence
  Set the stationary point to be $\tilde{\mu}_i = \mu_i^t$
END FOR
OUTPUT: Stationary points $\{\tilde{\mu}_i\}_{i=1}^n$ and clusters $\{\mathcal{C}_k\}_{k=1}^K$ //Eq. (15)

3.3 Comparison of KL information

Kullback-Leibler information between models p and q is defined for continuous functions as the integral

$$I(p, q) = \int p(\mathbf{x}) \log \left(\frac{p(\mathbf{x})}{q(\mathbf{x}|\theta)} \right) d\mathbf{x}, \quad (16)$$

where log denotes the natural logarithm. The notation $I(p, q)$ denotes the 'Information lost when q is used to approximate p' [Burnham and Anderson, 2002]. Now, suppose that f and f_a denotes the underlying ground truth and the approximated distribution of $f(\mathbf{x})$ by using a method for $a \in \{MS, VMS\}$. We applied importance sampling to address KL information using trial function $g(\mathbf{x})$ which is uniform distribution. The KL information for both mean shift clustering is written as

$$\begin{aligned} I(f, f_a) &= \int f(\mathbf{x}) \log \left(\frac{f(\mathbf{x})}{f_a(\mathbf{x})} \right) d\mathbf{x} \\ &= E_f \left[\log \left(\frac{f(\mathbf{x})}{f_a(\mathbf{x})} \right) \right] \\ &= \int \frac{f(\mathbf{x})}{g(\mathbf{x})} \log \left(\frac{f(\mathbf{x})}{f_a(\mathbf{x})} \right) g(\mathbf{x}) d\mathbf{x} \\ &\approx \sum_{s=1}^{S} w_s \log \left(\frac{f(\mathbf{x}_s)}{f_a(\mathbf{x}_s)} \right) \end{aligned} \quad (17)$$

where $w_s = \frac{f(\mathbf{x}_s)}{g(\mathbf{x}_s)}$ and $\mathbf{x}_s \sim g(\mathbf{x})$.

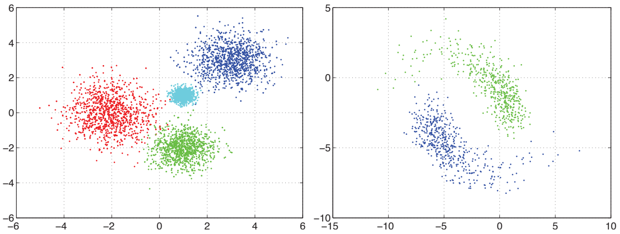
4. EXPERIMENTAL ANALYSIS

Without loss of generality, we used a radius r which is identical to a bandwidth h ($r = h$) for our experiments.

4.1 Clustering Synthetic Data Sets

We compared the conventional MS and the VMS on two synthetic data sets. The first data set consists of 500 points from four Gaussian distributions. The second data set contains 500 points from two banana-shaped clusters with $n = 500$. Both data sets are depicted in Fig. 3.

Figs. 4 and 5 show the clustering results of the MS and the VMS algorithms on the two data sets. The various bandwidths were used: from $h = 0.1$ to 1.9 incremented by 0.2 for Gaussians and from $h = 1.0$ to 3.0 by 0.2 for Bananas. For both data sets, the MS, being sensitive to outliers, identified many small clusters as it was trapped in local optima,



(a) Gaussians (b) Bananas

Figure 3: Two synthetic data sets: data points are generated (a) from four Gaussian distributions and (b) from two banana-shaped clusters.

in particular with smaller bandwidths used. On the other hand, the VMS produced much robust clustering results. It is because outliers (points far from major clusters) could be included into nearby clusters using the Voronoi kernel and clusters with a small number of points were removed. Fig. 6 shows that the VMS identified closer numbers of clusters to the true numbers than the conventional MS.

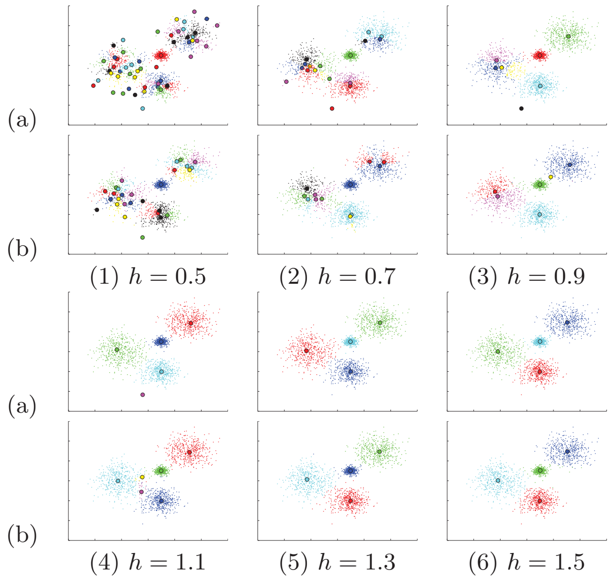


Figure 4: Clustering of the Gaussians data set by (a) the MS and (b) the VMS.

Fig. 7 shows the trajectories of the centres of the kernels. With the conventional MS, there are many trajectories where the crosses prematurely converged to points in sparse regions. On the other hand, with the VMS, most trajectories reached the major clusters in dense regions. For example, for the Gaussian data set, the conventional MS failed to assign the points around $(-0.3, -4.2)$, $(-4, -1.5)$ and $(-4, -1.4)$ into the major clusters. However, all of them were assigned to the major clusters by the VMS.

Fig. 8 compares the Kullback-Leibler (KL) divergence obtained by the two algorithms. The KL divergence between the true f and an estimated $\hat{f}$ probability densities is defined as

$$\mathrm{KL}(f, \hat{f}) = \int f(\mathbf{x}) \log \frac{f(\mathbf{x})}{\hat{f}(\mathbf{x})} d\mathbf{x}. \tag{18}$$

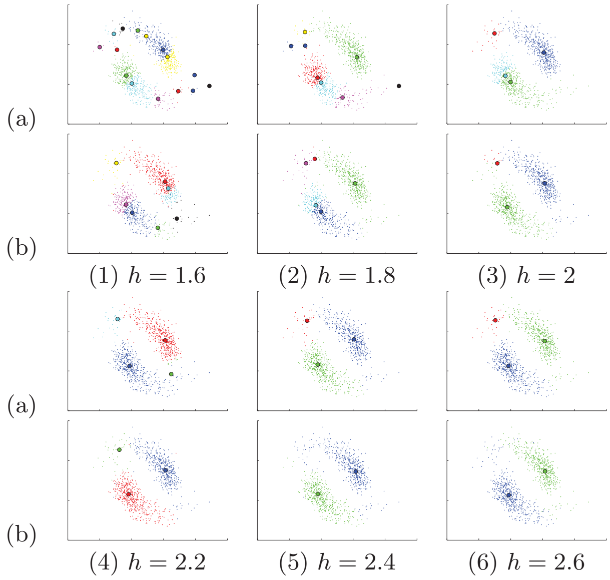
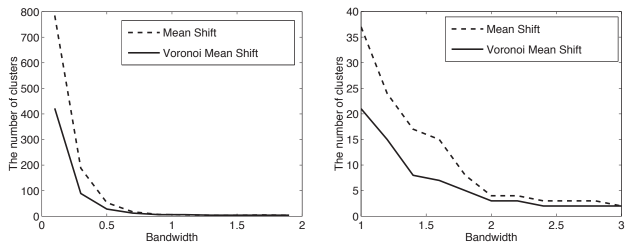


Figure 5: Clustering of the Bananas data set by (a) the MS and (b) the VMS.



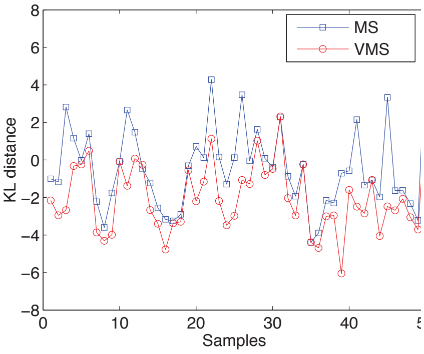
(a) Gaussians (b) Bananas

Figure 6: The numbers of clusters against the bandwidths: dashed and solid curves represent the trajectories of the numbers of clusters identified by the MS and VMS, respectively.

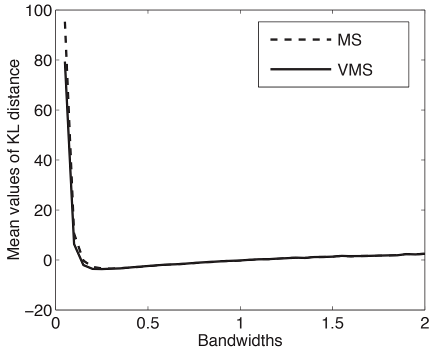
It implies 'information lost when $\hat{f}$ is used to approximate f' [Burnham and Anderson, 2002]. Fig. 8(a) shows the KL divergences of 50 random realisations with $h = 0.1$. The VMS yielded lower KL divergences (closer estimates of the probability density) for 44 out of 50. We also compared the average KL divergences by varying the bandwidths from 0 to 2 as shown in Fig. 8(b). The VMS yielded KL divergences no worse than the MS with any bandwidth used. It also shows that the optimal bandwidth is around $h = 0.2$ and that the VMS is more robust than the MS in bandwidth selection. In particular, their differences were larger with smaller bandwidths ($h \leq 0.2$) (see Fig. 8(c)).

4.2 Denoising of images

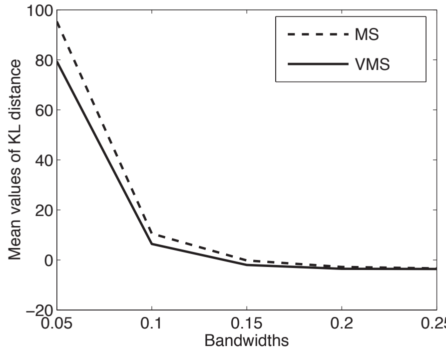
Applying the MS algorithms to images requires a preprocessing step. In general, the spaces $L*u*v$ and $L*a*b$ are used for image segmentation and filtering, which are designed to best approximate perceptually uniform colour spaces [Comaniciu and Meer, 2002]. We used the $L*u*v$ colour space in this paper, which is converted non-linearly from RGB colour values. In addition, there are two domains in an image: range and spatial domains. The colour level



(a) KL information

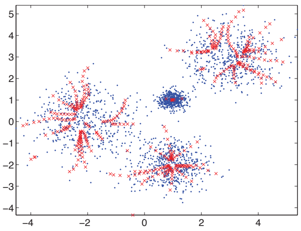


(b) The mean of KL information

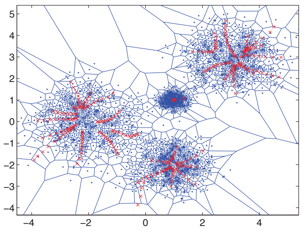


(c) The mean of KL information (full)

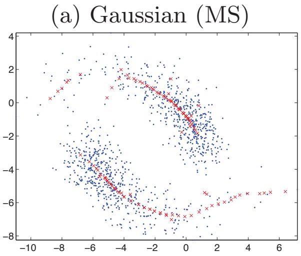
Figure 8: (a) KL divergences obtained by the MS and the VMS for $h = 0.1$, (b) the average KL divergences against the bandwidths and (c) the average KL divergences for smaller bandwidths.



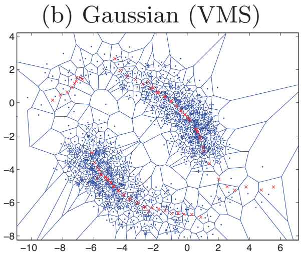
(a) Gaussian (MS)



(b) Gaussian (VMS)



(c) Banana (MS)



(d) Banana (VMS)

Figure 7: Trajectories of the centres of the kernels with $h = 0.7$ for Gaussians and $h = 2.2$ for Bananas: blue dots and red crosses represent the data points and the trajectories, respectively.

or spectral information is represented in the range domain while the locations of the image lattice are represented in the spatial domain. Thus, the dimension of the each sample point $\mathbf{x}_i$ for a colour image is five ($d = 5$): three for the range domain and two for the spatial domain.

The proposed algorithm was also applied to denoising. A noisy image (standard deviation $\sigma = 20$ for white noise) is filtered by the two MS algorithms with bandwidths $h_r = 32$ and $h_s = 2$. The raw and noisy images are shown in Fig. 9. Fig. 10 shows the images filtered by the MS and the VMS. The MS, sensitive to noise, could not reduce a large part of noise; we can see a large number of speckles, especially on the lawn and the road. On the other hand, the VMS produced a smoother image where noise was effectively reduced. The conventional algorithm identified a lot of small clusters containing five pixels or less. It implies that noisy pixels comprised independent clusters and that noise was not reduced effectively. In Fig. 11, the pixels corresponding to such clusters are marked white. The VMS, though missing some parts, reconstructed the image generally well. However, the MS did not, as a large part of Fig. 11(a) is covered by white patches. The numbers of clusters are 3477 and 1347 for the MS and the VMS, respectively.

(a) Original Image

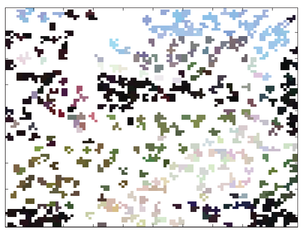
(b) Noisy image

Figure 9: (a) Original and (b) Noisy images.

(a) MS

(b) VMS

Figure 10: Filtered images by (a) the MS and (b) the VMS.



(a) MS

(b) VMS

Figure 11: Filtered images excluding clusters with less than 5 pixels by (a) the MS and (b) the VMS.

5. CONCLUSIONS

We proposed a nonlinear gating windowing scheme based on a geometric structure called Voronoi diagram of data points for truncated mean shift algorithms. Our motivation is to effectively identify outlier points in clusters by avoiding a linear increase of truncation window.

With a smaller radius used, the MS algorithm can be speed up since the smaller number of measurements are considered for calculation. However, if the radius decreases, the confidence intervals are also decreasing. This paper proposes a measurement selection scheme based on a Voronoi diagram of data points for a truncated mean shift (MS) algorithms. We call the truncated MS with the proposed measurement selection scheme by Voronoi mean shift (VMS) The VMS selects effectively significant measurements with embedding a relatively small radius of the truncation window.

Our approach has several advantages. Firstly, it can effectively assign data points into one of major clusters even with a still small gating size (the radius of truncation windows). Thus, the proposed algorithm is much robust in selecting the significant and effective measurements even though it does not increase the gating size a lot. Second, the proposed algorithm resulted in significantly lower the KL divergences, implying that the target density was estimated more accurately. In our experiments on synthetic and real-world data sets, we showed that the VMS outperformed the MS. In particular, we showed the feasibility of the proposed algorithm by applying it for denoising noisy images.

While the VMS algorithm has been shown to be promising, some issues and future research directions should be addressed.

- **Time complexity:** The VMS algorithm has extra processing compared to the conventional one. Firstly, we need to build a Voronoi diagram of data points. Secondly, extra processing time is required to find relevant points. While both operations are relatively simple and cheap in polynomial time, we are aiming at reducing complexity by adopting more efficient algorithms.

- **Blurring effect:** The window defined by the Voronoi kernel is always equal to or larger than that by the conventional kernel, which never select "outer" points. This may result in blurring or over-smoothing effects. In order to avoid such unwanted effects, we may need to consider discarding outer points that are too far outside of a window in a dense region.

6. REFERENCES

[Bradski, 1998] Bradski, G. R. (1998). Computer vision face tracking for use in a perceptual user interface. In *Proceedings of IEEE Workshop on Applications of Computer Vision*, pages 214–219.

[Burnham and Anderson, 2002] Burnham, K. P. and Anderson, D. (2002). *Model Selection and Multi-Model Inference*. Springer.

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

[Comaniciu et al., 2001] Comaniciu, D., Ramesh, V., and Meer, P. (2001). The variable bandwidth mean shift and data-driven scale selection. In *Proceedings of 8th International Conference on Computer Vision*, volume 1, pages 438–445.

[de Berg et al.,] de Berg, M., Cheong, O., van Kreveld, M., and Overmars, M. *Computational Geometry: Algorithms and Applications*. Springer, Berlin, 3rd ed. edition.

[Fukunaga, 1990] Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press, second edition.

[Fukunaga and Hostetler, 1975] Fukunaga, K. and 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.

[Georgescu et al., 2003] Georgescu, B., Shimshoni, I., and Meer, P. (2003). Mean shift based clustering in high dimensions: a texture classification example. In *Proceedings of 9th IEEE International Conference on Computer Vision*, volume 1, pages 456–463.

[Scott, 1992] Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley-Interscience.

[Silverman, 1986] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC.