Density Difference Detection with Application to Exploratory Visualization

Marko Rak, Tim König, Johannes Steffen, Dirk Joachim Lehmann, and Klaus-Dietz Tönnies

Department of Simulation and Graphics
Otto von Guericke University, Magdeburg, Germany

Abstract. Identifying differences among the distribution of samples of different observations is an important issue in many research fields. We provide a general framework to detect these difference spots in d-dimensional feature space. Such spots occur not only at various locations, they may also come in various shapes and multiple sizes, even at the same location. We address these challenges by a scale-space representation of the density function difference of the observations in feature space. Using three classification scenarios from UCI Machine Learning Repository we show that interest spots carry valuable information about a data set. To this end, we establish a simple decision rule on top of our framework. Results indicate state-of-the-art performance, underpinning the importance of the information that is carried by the detected spots. Furthermore, we outline that the output of our framework can be used to guide exploratory visualization of high-dimensional feature spaces.

Keywords: Density Difference, Kernel Density Estimation, Scale Space, Dendrogram, Blob Detection, Affine Shape Adaption, Exploratory Visualization, Orthographic Star Coordinates

1 Introduction

Sooner or later a large portion of pattern recognition tasks come down to the question *What makes X different from Y?* Some scenarios of that kind are:

> Detection of forged money based on image-derived features: *What makes some sort of forgery different from genuine money?*
> Comparison of medical data of healthy and non-healthy subjects for disease detection: *What makes the healthy different from the non-healthy?*
> Comparison of document data sets for text retrieval purposes: *What makes this set of documents different from another set?*

Apart from this, spotting differences in two or more observations is of interest in fields of computational biology, chemistry or physics. Looking at it from a general perspective, such questions generalize to

> *What makes samples of group X different from the samples of group Y?*

This question usually arises when we deal with grouped samples in some feature space. For humans, answering such questions tends to become more challenging with increasing number of groups, samples and feature space dimensions, up to the point where we miss the forest for the trees. This complexity is not an issue to automatic approaches, which, on the other hand, tend to either overfit or underfit patterns in the data. Therefore, semi-automatic approaches are needed to generate a number of interest spots which are to be looked at in more detail.

We address this issue by a scale-space difference detection framework. Our approach relies on the density difference of group samples in feature space. This enables us to identify spots where one group dominates the other. We draw on kernel density estimators to represent arbitrary density functions. Embedding this into a scale-space representation, we are able to detect spots of different sizes and shapes in feature space in an efficient manner. Our framework:

- applies to d-dimensional feature spaces
- is able to reflect arbitrary density functions
- selects optimal spot locations, sizes and shapes
- is robust to outliers and measurement errors
- produces human-interpretable results

Please note that large portions of the subsequent content were already covered in our previous work [16]. Within the current work we go into detail on a second spot detector (complementing the one used previously), provide an extended evaluation and show how the output of our framework can be used to guide the exploratory visualization of high-dimensional feature spaces. The latter may be seen as an intermediate step prior to applying other means of data analysis to the identified interest spots.

Our presentation is structured as follows. We outline the key foundations of our framework in Section 2. The specific parts of our framework are detailed in Section 3, while Section 4 outlines our contribution to exploratory visualization. Section 5 comprises our results on several data sets from UCI Machine Learning Repository. In Section 6, we close with a summary of our work, our most important results and an outline of future work.

2 Theoretical Foundations

Searching for differences between the sample distribution of two groups of observations g and h, we, quite naturally, seek for spots where the density function $f^g(\mathbf{x})$ of group g dominates the density function $f^h(\mathbf{x})$ of group h, or vice versa. Hence, we try to find positive-/negative-valued spots of the density difference

$$f^{g-h}(\mathbf{x}) = f^g(\mathbf{x}) - f^h(\mathbf{x}) \tag{1}$$

w.r.t. the underlying feature space $\mathbb{R}^d$ with $\mathbf{x} \in \mathbb{R}^d$. Such spots may come in various shapes and sizes. A difference detection framework should be able to deal with these degrees of freedom. Additionally, it must be robust to various sources of error, e.g. from measurement, quantization and outliers.

We propose to superimpose a scale-space representation to the density difference $f^{g-h}(\mathbf{x})$ to achieve the above-mentioned properties. Scale-space frameworks have been shown to robustly handle a wide range of detection tasks for various types of structures, e.g. text strings [23], persons and animals [8] in natural scenes, neuron membranes in electron microscopy imaging [20] or microaneurysms in digital fundus images [2]. In each of these tasks the function of interest is represented through a grid of values, allowing for an explicit evaluation of the scale-space. However, an explicit grid-based approach becomes intractable for higher-dimensional feature spaces.

In what follows, we show how a scale-space represenation of $f^{g-h}(\mathbf{x})$ can be obtained from kernel density estimates of $f^g(\mathbf{x})$ and $f^h(\mathbf{x})$ in an implicit fashion, expressing the problem by scale-space kernel density estimators. Note that by the usage of kernel density estimates our work is limited to feature spaces with dense filling. We close with a brief discussion on how this can be used to compare observations among more than two groups.

2.1 Scale Space Representation

First, we establish a family $l^{g-h}(\mathbf{x};t)$ of smoothed versions of the densitiy difference $l^{g-h}(\mathbf{x})$. Scale parameter $t \geq 0$ defines the amount of smoothing that is applied to $l^{g-h}(\mathbf{x})$ via convolution with kernel $k_t(\mathbf{x})$ of bandwidth t as stated in

$$l^{g-h}(\mathbf{x};t) = k_t(\mathbf{x}) * f^{g-h}(\mathbf{x}). \tag{2}$$

For a given scale t, spots having a size of about $2\sqrt{t}$ will be highlighted, while smaller ones will be smoothed out. This leads to an efficient spot detection scheme, which will be discussed in Section 3. Let

$$l^g(\mathbf{x};t) = k_t(\mathbf{x}) * f^g(\mathbf{x}) \tag{3}$$

$$l^h(\mathbf{x};t) = k_t(\mathbf{x}) * f^h(\mathbf{x}) \tag{4}$$

be the scale-space representations of the group densities $f^g(\mathbf{x})$ and $f^h(\mathbf{x})$. Looking at Equation 2 more closely, we can rewrite $l^{g-h}(\mathbf{x};t)$ equivalently in terms of $l^g(\mathbf{x};t)$ and $l^h(\mathbf{x};t)$ via Equation 3 and 4. This reads

$$\begin{aligned} l^{g-h}(\mathbf{x};t) &= k_t(\mathbf{x}) * f^{g-h}(\mathbf{x}) && (5)\\ &= k_t(\mathbf{x}) * \left[f^g(\mathbf{x}) - f^h(\mathbf{x})\right] && (6)\\ &= k_t(\mathbf{x}) * f^g(\mathbf{x}) - k_t(\mathbf{x}) * f^h(\mathbf{x}) && (7)\\ &= l^g(\mathbf{x};t) - l^h(\mathbf{x};t). && (8) \end{aligned}$$

The simple yet powerful relation between the left and the right-hand side of Equation 8 will allow us to evaluate the scale-space representation $l^{g-h}(\mathbf{x})$ implicitly, i.e. using only kernel functions. Of major importance is the choice of the smoothing kernel $k_t(\mathbf{x})$. According to scale-space axioms, $k_t(\mathbf{x})$ should suffice

a number of properties, resulting in the uniform Gaussian kernel of Equation 9 as the unique choice, cf. [3, 24].

$$\phi_t(\mathbf{x}) = \frac{1}{\sqrt{(2\pi t)^d}} \exp\left(-\frac{1}{2t}\mathbf{x}^{\mathrm{T}}\mathbf{x}\right) \tag{9}$$

2.2 Kernel Density Estimation

In kernel density estimation, the group density $f^g(\mathbf{x})$ is estimated from its n^g samples by means of a kernel function $K_{\mathbf{B}^g}(\mathbf{x})$. Let $\mathbf{x}_i^g \in \mathbb{R}^{d\times 1}$ with $i = 1, \ldots, n^g$ being the group samples. Then, the group density estimate is given by

$$\hat{f}^g(\mathbf{x}) = \frac{1}{n^g} \sum_{i=1}^{n^g} K_{\mathbf{B}^g}\left(\mathbf{x} - \mathbf{x}_i^g\right). \tag{10}$$

Parameter $\mathbf{B}^g \in \mathbb{R}^{d\times d}$ is a symmetric positive-definite matrix, which controls the sample influence to the density estimate. Informally speaking, $K_{\mathbf{B}^g}(\mathbf{x})$ applies a smoothing with bandwidth $\mathbf{B}^g$ to the "spiky sample relief" in feature space.

Plugging kernel density estimator $\hat{f}^g(\mathbf{x})$ into the scale-space representation $l^g(\mathbf{x}; t)$ defines the scale-space kernel density estimator $\hat{l}^g(\mathbf{x}; t)$ to be

$$\hat{l}^g(\mathbf{x}; t) = k_t(\mathbf{x}) * \hat{f}^g(\mathbf{x}). \tag{11}$$

Inserting Equation 10 into the above, we can trace down the definition of the scale-space density estimator $\hat{l}^g(\mathbf{x}; t)$ to the sample level via transformation

$$\begin{aligned}
\hat{l}^g(\mathbf{x}; t) &= k_t(\mathbf{x}) * \hat{f}^g(\mathbf{x}) && (12)\\
&= k_t(\mathbf{x}) * \left[\frac{1}{n^g} \sum_{i=1}^{n^g} K_{\mathbf{B}^g}\left(\mathbf{x} - \mathbf{x}_i^g\right)\right] && (13)\\
&= \frac{1}{n^g} \sum_{i=1}^{n^g} \left(k_t * K_{\mathbf{B}^g}\right)\left(\mathbf{x} - \mathbf{x}_i^g\right). && (14)
\end{aligned}$$

Though arbitrary kernels can be used, we choose $K_{\mathbf{B}}(\mathbf{x})$ to be a Gaussian kernel $\Phi_{\mathbf{B}}(\mathbf{x})$ due to its convenient algebraic properties. This (potentially non-uniform) kernel is defined as

$$\Phi_{\mathbf{B}}(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\mathbf{B})}} \exp\left(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{x}\right). \tag{15}$$

Using the above, the right-hand side of Equation 14 simplifies further because of the Gaussian's cascade convolution property. Eventually, the scale-space kernel density estimator $\hat{l}^g(\mathbf{x}; t)$ is given by Equation 16, where $\mathbf{I} \in \mathbb{R}^{d\times d}$ is the identity.

$$\hat{l}^g(\mathbf{x}; t) = \frac{1}{n^g} \sum_{i=1}^{n^g} \Phi_{t\mathbf{I}+\mathbf{B}^g}\left(\mathbf{x} - \mathbf{x}_i^g\right) \tag{16}$$

Using this estimator, the scale-space representation $l^g(\mathbf{x};t)$ of group density $f^g(\mathbf{x})$ and analogously that of group h can be estimated for any $(\mathbf{x};t)$ in an implicit fashion. Consequently, this allows us to estimate the scale-space representation $l^{g-h}(\mathbf{x};t)$ of the density difference $f^{g-h}(\mathbf{x})$ via Equation 7 by means of kernel functions only.

2.3 Bandwidth Selection

When regarding bandwidth selection in such a scale-space representation, we see that the impact of different choices for bandwidth matrix $\mathbf{B}$ vanishes as scale t increases. This can be seen when comparing matrices $t\mathbf{I}+\mathbf{0}$ and $t\mathbf{I}+\mathbf{B}$ where $\mathbf{0}$ represents the zero matrix, i.e. no bandwidth selection at all. We observe that relative differences between them become neglectable once $\|t\mathbf{I}\| \gg \|\mathbf{B}\|$. This is especially true for large sample sizes, because the bandwidth will then tend towards zero for any reasonable bandwidth selector anyway. Hence, we may actually consider setting $\mathbf{B}$ to $\mathbf{0}$ for certain problems, as we typically search for differences that fall above some lower bound for t.

Literature bares extensive work on bandwidth matrix selection, for example, based on plug-in estimators [6, 21] or biased, unbiased and smoothed cross-validation estimators [7, 19]. All of these integrate well with our framework. However, in view of the argument above, we propose to compromise between a full bandwidth optimization and having no bandwidth at all. We define $\mathbf{B}^g = b^g\mathbf{I}$ and use an unbiased least-squares cross-validation to set up the bandwidth estimate for group g. For Gaussian kernels, this leads to the optimization of 17, cf. [7], which we achieved by golden section search over b^g.

$$\arg\min_{\mathbf{B}^g} \frac{1}{n^g\sqrt{\det(4\pi\mathbf{B}^g)}} + \frac{1}{n^g(n^g-1)} \sum_{i=1}^{n^g}\sum_{\substack{j=1\\ j\neq i}}^{n^g} \left(\Phi_{2\mathbf{B}^g} - 2\Phi_{\mathbf{B}^g}\right)\left(\mathbf{x}_i^g - \mathbf{x}_j^g\right) \quad (17)$$

2.4 Multiple Groups

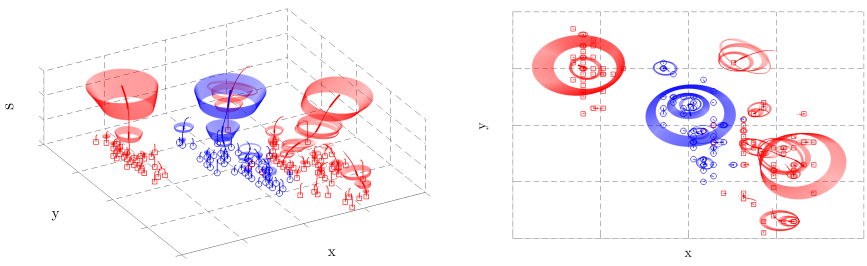
If differences among more than two groups shall be detected, we can reduce the comparison to a number of two-group problems. We can consider two typical use cases, namely *one group vs. another* and *one group vs. rest*. Which of the two is more suitable depends on the specific task at hand. Let us illustrate this using two medical scenarios. Assume we have a number of groups which represent patients having different diseases that are hard to discriminate in differential diagnosis. Then we may consider the second use case, to generate clues on markers that make one disease different from the others. In contrast, if these groups represent stages of a disease, potentially including a healthy control group, then we may consider the first use case, comparing only subsequent stages to give clues on markers of the disease's progress.

3 Detection Framework

To identify the positve-/negative-valued spots of a density difference, we apply the concept of blob detection, which is well-known in computer vision, to the scale-space representation derived in Section 2. In scale-space blob detection, some blobness criterion is applied to the scale-space representation, seeking for local optima of the function of interest w.r.t. space and scale. This directly leads to an efficient detection scheme that identifies a spot's location and size. The latter corresponds to the detection scale.

In a grid-representable problem we can evaluate blobness densely over the scale-space grid and identify interesting spots directly using the grid neighborhood. This is intractable here, which is why we rely on a more refined three-stage approach. First, we trace the local spatial optima of the density difference through scales of the scale-space representation. Second, we identify the interesting spots by evaluating their blobness along the dendrogram of optima that was obtained during the first stage. Having selected spots and therefore knowing their locations and sizes, we finally calculate an elliptical shape estimate for each spot in a third stage.

Spots obtained in this fashion characterize elliptical regions in feature space as outlined in Figure 1. The representation of such regions, i.e. location, size and shape, as well as its strength, i.e. its scale-space density difference value, are easily interpretable by humans, which allows to look at them in more detail using some other method. The elliptical nature of the identified regions is also a limitation of our work, because non-elliptical regions may only be approximated by elliptical ones. We now give a detailed description of the three stages.



(a) "Isometric" View (b) Top View

Fig. 1. Detection results for a two-group (red/blue) problem in two-dimensional feature space (xy-plane) with augmented scale dimension s; Red squares and blue circles visualize the samples of each group; Red/blue paths outline the dendrogram of scale-space density difference optima for the red/blue group dominating the other group; Interesting spots of each dendrogram are printed thick; Red/blue ellipses characterize the shape for each of the interest spots

3.1 Scale Tracing

Assume we are given an equidistant scale sampling, containing non-negative scales $t_1, \ldots, t_n$ in increasing order and we search for spots where group g dominates h. More precisely, we search for the non-negatively valued maxima of $l^{g-h}(\mathbf{x}; t_{i-1})$. The opposite case, i.e. group h dominates g, is equivalent.

Let us further assume that we know the spatial local maxima of the density difference $l^{g-h}(\mathbf{x}; t_{i-1})$ for a certain scale t_{i-1} and we want to estimate those of the current scale t_i. This can be done taking the previous local maxima as initial points and optimizing each w.r.t. $l^{g-h}(\mathbf{x}; t_i)$. In the first scale, we take the samples of group g themselves. As some maxima may be converged to the same location, we merge them together, feeding unique locations as initials into the next scale t_{i+1} only. We also drop any negatively-valued locations as these are not of interest to our task. They will not become of interest for any higher scale either, because local extrema will not enhance as scale increases, cf. [13]. Since derivatives are simple to evaluate for Gaussian kernels, we can use Newton's method for spatial optimization. We can assemble gradient $\frac{\partial}{\partial \mathbf{x}} l^{g-h}(\mathbf{x}; t)$ and Hessian $\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}} l^{g-h}(\mathbf{x}; t)$ sample-wise using

$$\frac{\partial}{\partial \mathbf{x}} \Phi_{\mathbf{B}}(\mathbf{x}) = -\Phi_{\mathbf{B}}(\mathbf{x})\, \mathbf{B}^{-1}\mathbf{x} \qquad \text{and} \tag{18}$$

$$\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}} \Phi_{\mathbf{B}}(\mathbf{x}) = \Phi_{\mathbf{B}}(\mathbf{x}) \left(\mathbf{B}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1} - \mathbf{B}^{-1} \right). \tag{19}$$

Iterating this process through all scales, we form a discret dendrogram of the maxima over scales. A dendrogram branching means that a maxima formed from two (or more) maxima from the preceding scale.

3.2 Spot Detection

The maxima of interest are derived from a scale-normalized blobness criterion $c_\gamma(\mathbf{x}; t)$. Two main criteria, namely the determinant of the Hessian [5] given in Equation 20[1] and the trace of the Hessian [13] given in Equation 22 have been discussed in literature. In contrast to our previous work [16], we do not focus on a single criterion. Instead, we will later investigate both in comparison.

$$c_\gamma^{\det}(\mathbf{x}; t) = t^{\gamma d} \underbrace{(-1)^d \det\left(\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}} l^{g-h}(\mathbf{x}; t) \right)}_{c^{\det}(\mathbf{x}; t)} \tag{20}$$

$$= t^{\gamma d} \qquad c^{\det}(\mathbf{x}; t) \tag{21}$$

$$c_\gamma^{\mathrm{tr}}(\mathbf{x}; t) = t^{\gamma d} \underbrace{(-1) \mathrm{tr}\left(\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}} l^{g-h}(\mathbf{x}; t) \right)}_{c^{\mathrm{tr}}(\mathbf{x}; t)} \tag{22}$$

$$= t^{\gamma d} \qquad c^{\mathrm{tr}}(\mathbf{x}; t) \tag{23}$$

[1] $(-1)^d$ leads to a consistent criterion for even and odd dimensions.

Because the maxima are already spatially optimal, we can search for spots that maximize $c_\gamma(\mathbf{x};t)$ w.r.t. the dendrogram neighborhood only. Note that we do not require the superscript because the remained is independent of the choice of the blobness criterion. Parameter $\gamma \geq 0$ can be used to introduce a size bias, shifting the detected spot towards smaller or larger scales. The definition of γ highly depends on the type of spot that we are looking for, cf. [12]. This is impractical when we seek for spots of, for example, small and large skewness or extreme kurtosis at the same time.

Addressing the parameter issue, we search for all spots that maximize $c_\gamma(\mathbf{x};t)$ locally w.r.t. some $\gamma \in [0,\infty)$. Some dendrogram spot s with scale-space coordinates $(\mathbf{x}_s;t_s)$ is locally maximal if there exists a γ-interval such that its blobness $c_\gamma(\mathbf{x}_s;t_s)$ is larger than that of every spot in its dendrogram neighborhood $\mathcal{N}(s)$. This leads to a number of inequalities, which can be written as

$$t_s^{\gamma d} c(\mathbf{x}_s;t_s) \underset{\forall n\in\mathcal{N}(s)}{>} t_n^{\gamma d} c(\mathbf{x}_n;t_n) \qquad \text{or} \tag{24}$$

$$\gamma\, d \log\frac{t_s}{t_n} \underset{\forall n\in\mathcal{N}(s)}{>} \log\frac{c(\mathbf{x}_n;t_n)}{c(\mathbf{x}_s;t_s)}. \tag{25}$$

The latter can be solved easily for the γ-interval, if any. We can now identify our interest spots by looking for the maxima along the dendrogram that locally maximize the width of the γ-interval. More precisely, let $w_\gamma(\mathbf{x}_s;t_s)$ be the width of the γ-interval for dendrogram spot s, then s is of interest if the dendrogram Laplacian of $w_\gamma(\mathbf{x};t)$ is negative at $(\mathbf{x}_s;t_s)$, or equivalently, if

$$w_\gamma(\mathbf{x}_s;t_s) > \frac{1}{|\mathcal{N}(s)|}\sum_{n\in\mathcal{N}(s)} w_\gamma(\mathbf{x}_n;t_n). \tag{26}$$

Intuitively, a spot is of interest if its γ-interval width is above neighborhood average. This is the only assumption we can make without imposing limitations on the results. Interest spots indentified in this way will be dendrogram segments, each ranging over a number of consecutive scales.

3.3 Shape Adaption

Shape estimation can be done in an iterative manner for each interest spot. The iteration alternatingly updates the current shape estimate based on a measure of anisotropy around the spot and then corrects the bandwidth of the scale-space smoothing kernel according to this estimate, eventually reaching a fixed point. The second moment matrix of the function of interest is typically used as an anisotropy measure, e.g. in [14] and [15]. Since it requires spatial integration of the scale-space representation around the interest spot, this measure is not feasible here.

We adapted the Hessian-based approach of [10] to d-dimensional problems. The aim is to make the scale-space representation isotropic around the interest spot, iteratively moving any anisotropy into the symmetric positive-definite

shape matrix $\mathbf{S} \in \mathbb{R}^{d \times d}$ of the smoothing kernel's bandwidth $t\mathbf{S}$. Thus, we lift the problem into a generalized representation $l^{g-h}(\mathbf{x}; t\mathbf{S})$ of anisotropic scale-space kernels, which requires us to replace the definition of $\phi_t(\mathbf{x})$ by that of $\Phi_{\mathbf{B}}(\mathbf{x})$.

Starting with the isotropic $\mathbf{S}_1 = \mathbf{I}$, we decompose the current Hessian via

$$\frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}} l^{g-h}(\,\cdot\,; t\mathbf{S}_i) = \mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{T}} \tag{27}$$

into its eigenvectors in columns of $\mathbf{V}$ and eigenvalues on the diagonal of $\mathbf{D}^2$. We then normalize the latter to unit determinant via

$$\mathbf{D} = \sqrt[d]{\det(\mathbf{D})}\mathbf{D} \tag{28}$$

to get a relative measure of anisotropy for each of the eigenvector directions. Finally, we move the anisotropy into the shape estimate via

$$\mathbf{S}_{i+1} = \left(\mathbf{V}^{\mathrm{T}}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}\right)\mathbf{S}_i\left(\mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^{\mathrm{T}}\right) \tag{29}$$

and start all over again. Iteration terminates when isotropy is reached. More precisely: when the ratio of minimal and maximal eigenvalue of the Hessian approaches one, which usually happens within a few iterations.

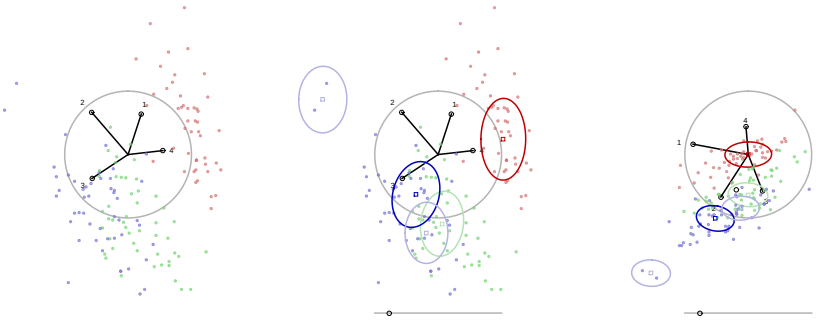
4 Exploratory Visualization

As mentioned introductory, exploratory visualization may be a reasonable intermediate step prior to directly applying other means of data analysis to the interest spots. There are plenty of visualization techniques that aim at identification of interesting patterns in the distribution of samples in high-dimensional feature spaces. For this work, we focus on a recent in-house development namely orthographic star coordinates [11]. We next give a short introduction to the topic and discuss how outputs of our framework can be used to guide the visual exploration process.

4.1 Star Coordinate Visualization

Star coordinate visualizations make use of projections from d-dimensional feature spaces to a two-dimensional projection plane which is then visualized. Such projections are characterized by a projection matrix $\mathbf{P} \in \mathbb{R}^{2 \times d}$ the columns of which can be interpreted as d points in two-dimensional space. Modifying these so-called anchor points is equivalent to manipulation of the projection plane itself, which the star coordinate visualization exploits by an interactive interface like that shown in Figure 2.

In general, star coordinates allow for arbitrary projections thus potentially introducing arbitrary distortions to the visualization of the high-dimensional content. This is not desirable for various reasons, therefore [11] proposed to restrict the interaction to orthographic projections. Orthography is directly related



(a) Original View (b) Augmented View (c) Focused View

Fig. 2. Exploratory visualization of a three-group (red/green/blue) problem in 4-dimensional feature space by orthographic star coordinates; Original orthographic star coordinates (left) augmented with output of our framework (middle) and focused on a particular interest spot (right); Moveable anchor points are connected to the origin by thick black line segments; A slider for scale selection is located at the bottom of the interface; The remaining visual content is discussed in the text

to d-dimensional rotation, enforcing this property thus provides an intuitive way to "rotate" the high-dimensional content in front of a user's viewpoint. This directly targets the human's ability to interpret spatial relations from a steerable sequence of projections which is pretty much what we do with two-dimensional visualizations of three-dimensional content on a daily basis.

4.2 Preserving Orthography

Regarding orthography, we have to address two main issues. First, how to recover an orthographic projection when starting from an arbitrary projection. Second, how to reinforce orthography during interactive anchor movement. A sufficient condition for orthography of some anchor point constellation $\mathbf{P}_o \in \mathbb{R}^{2\times d}$ is that

$$\mathbf{P}_o{\mathbf{P}_o}^{\mathrm{T}} = \mathbf{I}, \tag{30}$$

whereby $\mathbf{I} \in \mathbb{R}^{2\times 2}$ is the identity matrix, cf. [11]. Therefore, given an arbitrary non-orthographic $\mathbf{P}$ we may seek to make $\mathbf{P}\mathbf{P}^{\mathrm{T}} \in \mathbb{R}^{2\times 2}$ identity. Since the latter Gramian matrix is almost certainly positive-definite in practice,[2] we can obtain it's Cholesky factor $\mathbf{L} \in \mathbb{R}^{2\times 2}$ and manipulate the decomposition as follows

$$\mathbf{L}\mathbf{L}^{\mathrm{T}} = \mathbf{P}\mathbf{P}^{\mathrm{T}} \tag{31}$$

$$\mathbf{I} = \underbrace{\mathbf{L}^{-1}\mathbf{P}}\,\underbrace{\mathbf{P}^{\mathrm{T}}\mathbf{L}^{-\mathrm{T}}} \tag{32}$$

$$\mathbf{I} = \quad \mathbf{P}_o \qquad {\mathbf{P}_o}^{\mathrm{T}} \tag{33}$$

[2] Rare semi-definite cases are avoided by regularization $\mathbf{P}\mathbf{P}^{\mathrm{T}} + \epsilon\mathbf{I}$ for some small ϵ.

with $\mathbf{P}_o$ being the recovered orthographic projection.[3] Regarding the second issue, we can simply take the steps just outlined, continuously reinforcing orthography during interactive movement of particular anchors. Note how the anchor points of the given non-orthographic $\mathbf{P}$ are all transformed in the same manner by the (inverse of the) Cholesky factor $\mathbf{L}$ to obtain the orthographic anchor points $\mathbf{P}_o$. This avoids any experience of "arbitrariness" during user interaction.

4.3 Guiding Explorations

As already discussed in [11], there are certain open questions associated to star coordinate visualizations. This includes suitable anchor point constellations, centers of "rotation", i.e. the choice of the origin in d-dimensional feature space prior to projection, as well as a reasonable zoom into the data after projection. Otherwise put, we need to know where to look at and how. The interest spots detected by our framework can be used to address these issues, thereby also providing an interactive mechanism to switch among potentially interesting structures.

As show in Figure 2, we have augmented the star coordinate visualization by a scale selection slider, letting the user choose the size (scale) of structures he/she is interested in. Based on his/her selection, the visualization is overlayed with the output of our framework that corresponds to the selected scale. Specifically, we transparently visualize the locations of maxima that were found during scale tracing (see Section 3.1) and their respective shapes, which were estimated during shape adaption (see Section 3.3). In case a maximum was found interesting (see Section 3.2), it's location and shape is highlighted opaquely instead.

When interactively selecting a maxima, the visualization is changed to put focus on the selection. Specifically, the origin of the d-dimensional feature space is shifted to the maxima's location thereby making it the center of "rotation". The user can then change the zoom to a multiple of the maxima's scale by keyboard bindings if desired. By another binding, he/she may also align the projection plane with the two most significant axes of the shape estimate to get a reasonable initial constellation of anchor points. To this end, the unit eigenvectors that correspond to the two largest eigenvalues of the shape estimate are used to fill the rows of the projection matrix.

We combined the above with a binding that resets the visualization to just before focusing a selection which allows to rapidly explore several potentially interesting spots before the user eventually moves on to differently sized structures. Changing the scale selection slider steadily, the course of locations and shapes of the maxima gives an impression on how the data is structured from coarse to fine without missing any highlighted interest spot.

5 Experiments

We next demonstrate that interest spots carry valuable information about a data set. Due to the lack of data sets that match our particular detection task a ground

[3] This formulation is another view on the Gram-Schmidt process used in [11].

truth comparison is impossible. Certainly, artificially constructed problems are an exception. However, the generalizability of results is at least questionable for such problems. Therefore, we chose to benchmark our approach indirectly via a number of classification tasks. The rational is that results that are comparable to those of well-established classifiers should underpin the importance of the identified interest spots.

Next we show how to use these interest spots for classification using a simple decision rule and detail the data sets that were used. We then investigate parameters of our approach and discuss the results of the classification tasks in comparison to decision trees, Fisher's linear discriminant analysis, k-nearest neighbors with optimized k and support vector machines with linear and cubic kernels. All experiments were performed via leave-one-out cross-validation.

5.1 Decision Rule

To perform classification we establish a simple decision rule based on interest spots that were detected using the *one group vs. rest* use case. Therefore, we define a group likelihood criterion as follows. For each group g, having the set of interest spots $\mathcal{I}^g$, we define

$$p^g(\mathbf{x}) = \max_{s \in \mathcal{I}^g} l^{g-h}\left(\mathbf{x}_s; t_s \mathbf{S}_s\right) \cdot \exp\left(-\frac{1}{2}\left(\mathbf{x}-\mathbf{x}_s\right)^{\mathrm{T}}\left(t_s \mathbf{S}_s\right)^{-1}\left(\mathbf{x}-\mathbf{x}_s\right)\right). \quad (34)$$

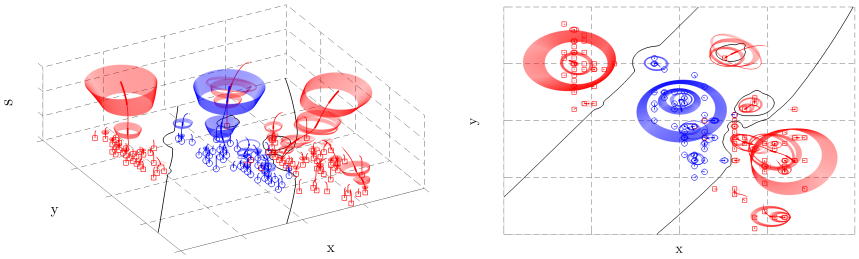
This is a quite natural trade-off, where the first factor favors spots s with high density difference, while the latter factor favors spots with small Mahalanobis distance to the location $\mathbf{x}$ that is investigated. We may also think of $p_g(\mathbf{x})$ as an exponential approximation of the scale-space density difference using interesting spots only. Given this, our decision rule simply takes the group that maximizes the group likelihood for the location of interest $\mathbf{x}$. Figure 3 illustrate the decision boundary obtained from this rule.

5.2 Data Sets

We carried out our experiments on three classification data sets taken from UCI Machine Learning Repository. A brief summary of them is given in Table 1. In the first task, we distinguish between benign and malign breast cancer based on manually graded cytological charateristics, cf. [22]. In the second task, we distinguish between genuine and forged money based on wavelet-transform-derived features from photographs of banknote-like specimen, cf. [9]. In the third task, we differentiate among normal, spondylolisthetic and disc-herniated vertebral columns based on biomechanical attributes derived from shape and orientation of the pelvis and the lumbar vertebral column, cf. [4].

5.3 Parameter Investigation

Before detailing classification results, we investigate two aspects of our approach. Firstly, we inspect the importance of bandwidth selection, benchmarking no



(a) "Isometric" View (b) Top View

Fig. 3. Feature space decision boundaries (black plane curves) obtained from group likelihood criterion for the two-dimensional two-group problem of Figure 1 using $c_\gamma^{\det}$ for spot detection; Red squares and blue circles visualize the samples of each group; Red/blue paths outline the dendrogram of scale-space density difference optima for the red/blue group dominating the other group; Interesting spots of each dendrogram are printed thick; Red/blue ellipses characterize the shape for each of the interest spots

Table 1. Data sets from UCI Machine Learning Repository

	Breast Cancer	Banknote Authentication	Vertebral Column
Groups	benign/malign	genuine/forged	normal/spondylo./herniated
Samples	444/239	762/610	100/150/60
Dimensions	10	4	6

kernel density bandwidth against the least-squares cross-validation technique that we use. Secondly, we determine the influence of the scale sampling rate. For the latter we space $n+1$ scales for various n equidistantly from zero to

$$t_n = F_{\chi^2}^{-1}(1-\epsilon|d) \max_g \left(\sqrt[d]{\det(\Sigma_g)} \right), \tag{35}$$

where $F_{\chi^2}^{-1}(\cdot\,|d)$ is the cumulative inverse-χ^2 distribution with d degrees of freedom and Σ_g is the covariance matrix of group g. Intuitively, t_n captures the extent of the group with largest variance up to a small ϵ, i.e. here $1.5 \cdot 10^{-8}$.

To investigate the two aspects, we compare classification accuracies with and without bandwidth selection as well as sampling rates ranging from $n = 100$ to $n = 300$ in steps of 25. From the results, which are given in Table 2, we observe that bandwidth selection is almost neglectable for the Breast Cancer (BC) and the Banknote Authentication (BA) data set no matter which criterion is used for spot detection. However, the impact is substantial throughout all scale sampling rates for the Vertebral Column (VC) data set for both criteria. This may be due to the comparably small number of samples per group for this data set.

Regarding the second aspect, we observe that for both criteria the BA and VC data set classification accuracy increases only slightly when the scale sampling rate rises. On the BC data set, accuracy remains stable, except for the lower rates when $c_\gamma^{\det}$ is used for spot detection. There is no such drop for the

Table 2. Classification accuracy of our decision rule in $\lfloor\%\rfloor$ for data sets of Table 1 for both detectors with/without bandwidth selection

	Scale sampling rate n								
c_γ^{det}-based decision rule	100	125	150	175	200	225	250	270	300
Breast Cancer	65/65	97/97	97/97	95/95	97/97	95/95	97/97	96/96	97/97
Banknote Authentication	96/94	96/96	96/96	98/98	98/98	98/98	98/98	98/98	99/99
Vertebral Column	87/82	88/83	88/84	88/83	88/85	88/85	88/86	88/86	88/87
	Scale sampling rate n								
c_γ^{tr}-based decision rule	100	125	150	175	200	225	250	270	300
Breast Cancer	96/96	97/97	96/96	96/96	96/96	96/96	96/96	96/96	96/96
Banknote Authentication	95/95	97/97	98/98	97/97	98/98	98/98	98/98	98/99	99/99
Vertebral Column	89/81	89/80	89/80	89/83	89/83	89/84	89/83	89/84	89/83

c_γ^{tr}-derived results, indicating a higher sensitivity of the latter for sparser samplings. Apart from that, the differences between the results of both criteria are minor for all data sets and sampling rates. From the results we conclude that bandwidth selection is a necessary part for interest spot detection. We further recommend $n \geq 200$, because accuracy is saturated at this point for all data sets independently of the choice of the spot detection criterion. For the remaining experiments we used bandwidth selection and a sampling rate of $n = 200$.

5.4 Classification Results

A comparison of classification accuracies of our decision rule against the aforementioned classifiers is given in Table 3. For the BC data set we observe that except for the support vector machine (SVM) with cubic kernel all approaches were highly accurate, scoring between 94% and 97% with our c_γ^{det}-based decision rule being topmost and the c_γ^{tr}-derived results being only slightly worse. Even more similar to each other are results for the BA data set, where all approaches score between 97% and 99%, with ours lying in the middle of this range. Results are most diverse for the VC data sets. Here, the SVM with cubic kernel again performs significantly worse than the rest, which all score between 80% and 85%, while our $c_\gamma^{\mathrm{det}}/c_\gamma^{\mathrm{tr}}$-based decision rules peak at 88% and 89% respectively. Other research showed similar scores on the given data sets. For example the artificial neural networks based on pareto-differential evolution in [1] obtained 98% accuracy for the BC data set, while [18] achieved 83% to 85% accuracy on the VC data set with SVMs with different kernels. These results suggest that our interest points carry information about a data set that are similarly important than the information carried by the well-established classifiers.

Confusion tables for our approach are given in Table 4 for all data sets. As can be seen, our $c_\gamma^{\mathrm{det}}/c_\gamma^{\mathrm{tr}}$-based decision rules gave balanced inter-group results for the BC and the BA data set. We obtained only small inaccuracies for the recall of the benign (96%/96%) and genuine (97%/96%) groups as well as for the precision of the malign (94%/93%) and forged (96%/95%) groups. Results for the VC data set were more diverse. Here, a number of samples with disc herniation were

Table 3. Classification accuracies of different classifiers in $\lfloor\%\rfloor$ for data sets of Table 1

	Breast Cancer	Banknote Authen.	Verteral Column
decision tree	94	98	82
k-nearest neighbors	97	99	80
Fisher's discriminant	96	97	80
linear/cubic kernel SVM	96/90	99/98	85/74
$c_\gamma^{\mathrm{det}}/c_\gamma^{\mathrm{tr}}$-based decision rule	97/96	98/98	88/89

mistaken for being normal, lowering the recall of the herniated group (86%/86%) noticeably. However, more severe inter-group imbalances were caused by the normal samples, which were relatively often mistaken for being spondylolisthetic or herniated discs. Thus, recall for the normal group (76%/80%) and precision for the herniated group (74%/76%) decreased significantly. The latter is to some degree caused by a handful of strong outliers from the normal group that fall into either of the other groups, which can already be seen from the group likelihood plots in Figure 4. This finding was made by others as well, cf. [17].

Table 4. Confusion table for predicted/actual groups of our $c_\gamma^{\mathrm{det}}/c_\gamma^{\mathrm{tr}}$-based decision rule for data sets of Table 1

(a) Breast Cancer

act. / pred.	benign	malign	precision
benign	429/429	4/6	99/98
malign	15/15	235/233	94/93
recall	96/96	98/97	$\lfloor\%\rfloor$

(b) Banknote Authentication

act. / pred.	genuine	forged	precision
genuine	742/736	0/0	100/100
forged	20/26	610/610	96/95
recall	97/96	100/100	$\lfloor\%\rfloor$

(c) Vertebral Column

act. / pred.	normal	spondylo.	herniated	precision
normal	76/80	1/1	6/7	91/90
spondylo.	10/8	145/145	2/1	92/94
herniated	14/12	4/4	52/52	74/76
recall	76/80	96/96	86/86	$\lfloor\%\rfloor$

The other classifiers performed similarly balanced on the BA and BC data set. Major differences occured on the VC data set only. A precision/recall comparison of all classifiers on the VC data set is given in Table 5. We observe that the precision of the normal and the herniated group are significantly lower (gap $>$ 12%) than that of the spondylolisthetic group for all classifiers except for our decision rules, for which at least the normal group is predicted with a similar precision by both rules. Regarding the recall we note an even more unbalanced behavior. Here, a strict ordering from spondylolisthetic over normal to herniated disks occurs. The differences of the recall of spondylolisthetic and normal are

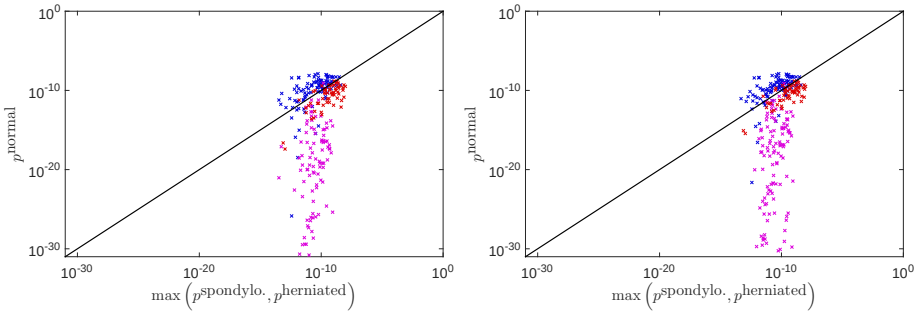


Fig. 4. Sample group likelihoods and decision boundary (black diagonal line) for the Vertebral Column data set of Table 1 using c_γ^{det} (left) and c_γ^{tr} (right) for spot detection; Normal, spondylolisthetic and herniated discs in blue, magenta and red, respectively

significant (gap > 16 %) and those between normal and herniated are even larger (gap > 18 %) among all classifiers that we compared against. The recalls for our decision rules are distributed differently, ordering the herniated before the normal group. Also the magnitude of differences is less significant (gaps $\approx 10\%/6\%$) for both decision rules. Results of the comparison indicate that the information that is carried by our interest points tends to be more balanced among groups than the information carried by the well-established classifiers that we compared against. The final question which interest spot detection criterion (c_γ^{det} or c_γ^{tr}) should be recommended cannot be answered satisfactorily based solely on our evaluation, because result differ only insignificantly. Yet, we advocate c_γ^{det} since it has been shown to provide better scale selection properties under affine transformation of the feature space, cf. [13].

Table 5. Classification precision and recall of different classifiers in $\lfloor\%\rfloor$ for the Vertebral Column data set of Table 1

	normal group		spondylo. group		herniated group	
	precision	recall	precision	recall	precision	recall
decision tree	69	83	97	95	68	50
k-nearest neighbors	70	74	96	96	58	55
Fisher's discriminant	70	80	87	92	74	48
linear/cubic kernel SVM	76/59	85/82	97/90	96/91	72/52	61/18
$c_\gamma^{\text{det}}/c_\gamma^{\text{tr}}$-based decision rule	91/90	76/80	92/94	96/96	74/76	86/86

6 Conclusion

We proposed a detection framework that is able to identify differences among the sample distributions of different observations. Potential applications are mani-

fold, touching fields such as medicine, biology, chemistry and physics. Our approach bases on the density function difference of the observations in feature space, seeking to identify spots where one observation dominates the other. Superimposing a scale-space framework to the density difference, we are able to detect interest spots of various locations, size and shapes in an efficient manner.

Our framework is intended for semi-automatic processing, providing human-interpretable interest spots for further investigation of some kind. We outlined how the output of our framework can be used to guide exploratory visualization of high-dimensional feature spaces as an intermediate step prior to other means of data analysis. Furthermore, we showed that the detected interest spots carry valuable information about a data set on a number of classification tasks from the UCI Machine Learning Repository. To this end, we established a simple decision rule on top of our framework. Results indicate state-of-the-art performance of our approach, which underpins the importance of the information that is carried by the detected interest spots.

In the future, we plan to extend our work to support repetitive features such as angles, which currently is a limitation of our approach. Modifying our notion of distance, we would then be able to cope with problems defined on, e.g. a sphere or torus. Future work may also include the migration of other types of scale-space detectors to density difference problems. This includes the notion of ridges, valleys and zero-crossings, leading to richer sources of information.

Acknowledgements This research was partially funded by the project "Visual Analytics in Public Health" (TO 166/13-2) of the German Research Foundation.

References

1. Abbass, H.A.: An evolutionary artificial neural networks approach for breast cancer diagnosis. Artificial Intelligence in Medicine 25, 265–281 (2002)
2. Adal, K.M., Sidibe, D., Ali, S., Chaum, E., Karnowski, T.P., Meriaudeau, F.: Automated detection of microaneurysms using scale-adapted blob analysis and semi-supervised learning. Computer Methods and Programs in Biomedicine 114, 1–10 (2014)
3. Babaud, J., Witkin, A.P., Baudin, M., Duda, R.O.: Uniqueness of the Gaussian kernel for scale-space filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence 8, 26–33 (1986)
4. Berthonnaud, E., Dimnet, J., Roussouly, P., Labelle, H.: Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters. Journal of Spinal Disorders and Techniques 18, 40–47 (2005)
5. Bretzner, L., Lindeberg, T.: Feature tracking with automatic selection of spatial scales. Computer Vision and Image Understanding 71, 385–392 (1998)
6. Duong, T., Hazelton, M.L.: Plug-in bandwidth matrices for bivariate kernel density estimation. Journal of Nonparametric Statistics 15, 17–30 (2003)
7. Duong, T., Hazelton, M.L.: Cross-validation bandwidth matrices for multivariate kernel density estimation. Scandinavian Journal of Statistics 32, 485–506 (2005)

8. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645 (2010)
9. Glock, S., Gillich, E., Schaede, J., Lohweg, V.: Feature extraction algorithm for banknote textures based on incomplete shift invariant wavelet packet transform. In: Annual Pattern Recognition Symposium. vol. 5748, pp. 422–431 (2009)
10. Lakemond, R., Sridharan, S., Fookes, C.: Hessian-based affine adaptation of salient local image features. Journal of Mathematical Imaging and Vision 44, 150–167 (2012)
11. Lehmann, D.J., Theisel, H.: Orthographic star coordinates. IEEE Transactions on Visualization and Computer Graphics 19, 2615–2624 (2013)
12. Lindeberg, T.: Edge detection and ridge detection with automatic scale selection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 465–470 (1996)
13. Lindeberg, T.: Feature detection with automatic scale selection. International Journal of Computer Vision 30, 79–116 (1998)
14. Lindeberg, T., Garding, J.: Shape-adapted smoothing in estimation of 3-d depth cues from affine distortions of local 2-d brightness structure. In: European Conference on Computer Vision. pp. 389–400 (1994)
15. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision 60, 63–86 (2004)
16. Rak, M., König, T., Tönnies, K.D.: Spotting differences among observations. In: International Conference on Pattern Recognition Applications and Methods. pp. 5–13 (2015)
17. Rocha Neto, A.R., Barreto, G.A.: On the application of ensembles of classifiers to the diagnosis of pathologies of the vertebral column: A comparative analysis. IEEE Latin America Transactions 7, 487–496 (2009)
18. Rocha Neto, A.R., Sousa, R., Barreto, G.A., Cardoso, J.S.: Diagnostic of pathology on the vertebral column with embedded reject option. In: Pattern Recognition and Image Analysis, vol. 6669, pp. 588–595. Springer Berlin Heidelberg (2011)
19. Sain, S.R., Baggerly, K.A., Scott, D.W.: Cross-validation of multivariate densities. Journal of the American Statistical Association 89, 807–817 (1992)
20. Seyedhosseini, M., Kumar, R., Jurrus, E., Giuly, R., Ellisman, M., Pfister, H., Tasdizen, T.: Detection of neuron membranes in electron microscopy images using multi-scale context and Radon-like features. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 670–677 (2011)
21. Wand, M.P., Jones, M.C.: Multivariate plug-in bandwidth selection. Computational Statistics 9, 97–116 (1994)
22. Wolberg, W., Mangasarian, O.: Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In: National Academy of Sciences. pp. 9193–9196 (1990)
23. Yi, C., Tian, Y.: Text string detection from natural scenes by structure-based partition and grouping. IEEE Transactions on Image Processing 20, 2594–2605 (2011)
24. Yuille, A.L., Poggio, T.A.: Scaling theorems for zero crossings. IEEE Transactions on Pattern Analysis and Machine Intelligence 8, 15–25 (1986)