Non-Parametric Windows-based Estimation of Probability Density Function in Vector Space

Mitchell Chen[1]
mitchell@robots.ox.ac.uk
Niranjan Joshi[1]
njoshi@robots.ox.ac.uk
Emma Helm[2]
emmajhelm@hotmail.com
Sir Michael Brady[1]
jmb@robots.ox.ac.uk

[1] Wolfson Medical Vision Laboratory
University of Oxford
Oxford, United Kingdom
OX1 3PJ

[2] Churchill Hospital
Oxford, United Kingdom
OX3 7LJ

Abstract

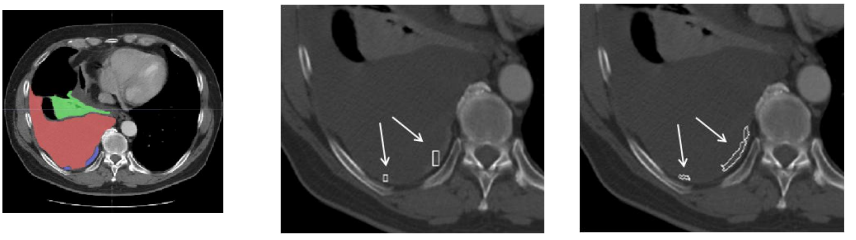
In this paper we extend the theory of non-parametric windows estimator to the vector space, aiming to establish a more generic probability density estimator that can be used in building an effective automatic image segmentation algorithm. We have verified our theoretical advancement, through two different experiments in medical imaging, and demonstrated the superior performance and benefits of this method compare to the traditional histogram estimator.

1 Introduction

Probability density functions (PDF) are central to many advanced segmentation and registration techniques. A number of PDF estimation methods have been developed and applied to image analysis. PDF estimation for medical applications increasingly uses non-parametric (NP) methods because for most medical applications, it is neither correct nor sufficient to assume a particular parametric form; because image noise is typically not Gaussian; anatomical structures are complex and variable; and the presence of various imaging artefacts. For these reasons, only NP methods are feasible for use in the field of medical image analysis. In this paper, we will focus mainly on the method of PDF estimation by histograms; and the novel approach by NP windows (NPW) [2], [4]. A third NP method, kernel density estimator (KDE), has been introduced and discussed more extensively in [3]. Histogram estimators are conceptually simple and computationally fast but require a large sample size to produce an accurate estimate. Moreover, they suffer from the binning and choice of origin problems. The kernel density estimator solves these and gives a better convergence rate. However, determining the optimal bandwidth remains challenging as even the latest cross-validation-based algorithms can be computationally demanding [2]. We have previously demonstrated [1] the advantages and use of NPW for segmenting malignant pleural

The authors would like to acknowledge funding support from GE Healthcare and Microsoft Research.

mesothelioma (MPM)[1] based on intensity values on thoracic CT scans. It was found that scalar NPW outperforms the histogram estimator in its smoothness. This method also offers advantages over KDE in terms of its computational requirement (10^3 faster).



(a) Manual Segmentation (b) Initialisation (c) Segmented Tumour

Figure 1: Preliminary level sets-based tumour segmentation using PDF estimates

In [1] we have made observations on PDFs, and showed that PDF-based segmentation for MPM is feasible, as supported by the semi-automatic segmentation results (given in Fig. 1, using level sets segmentation based on Battacharya measures). In a follow-up study involving a group of 35 data samples, the algorithm performed with a good degree of accuracy in cases where tumour was surrounded by effusion or aerated lung, with a mean difference in aerated lung of 6% (+/- 2% std.dev.) compared to radiologist derived areas. However, the algorithm was less successful at segmenting tumour (25% mean difference and +/- 15% std.dev.) from atelectatic lung or diaphragm. In fact, we note that for most complex medical segmentation problems, image intensity alone is not sufficient to give accurate and reliable results. This necessitates the need to further investigate the application of the NPW estimator in automatic image segmentation. A good starting point is to examine ways in which clinical manual segmentations are typically accomplished. We note that in addition to image pixel intensities, texture; tissue heterogeneity; and general knowledge on human anatomy are often used in identifying a tissue's boundaries in an image scan. These additional measures may potentially support the development of a better segmentation algorithm. Our goal is to establish an NPW-based estimator for vector-valued data (*n-tuple* where n is the size of the vector) where two or more image properties are associated with each pixel that initially had only a greyscale intensity measure. As most of these other quantities are derived from hence dependent on the intensity values, it is not sufficient to simply define the n-tuple joint distribution as the product of their marginal distributions. In order to incorporate these properties into our algorithm, we will need to extend the founding theories of NP windows onto the vector domain. In this paper, we present the newly developed theories and their derivations in Section 2. Experimental validation of our method is described and shown in Section 3, followed by a discussion of the results and possible future works, which is given in Section 4.

2 Methodology

We begin with a 2-tuple vector $F_{y_1,y_2}(x)$ where for each x there are two associated quantities. This can be a combination of any two arbitrary pieces of information, y_1 and y_2 given in an image sample. For instance, in an optic flow map, they can be the $u(x,y)$ and $v(x,y)$

[1] a form of lung tumour

Conditions	Case
$a_1, a_2 \neq 0$	$\frac{a_2}{a_1} y_1 + b_2 - \frac{a_2}{a_1} b_1 = y_2$
$a_2 = 0, a_1 \neq 0$	$y_2 = b_2$
$a_1 = 0, a_2 \neq 0$	$y_1 = b_1$
$a_1, a_2 = 0$	a point at (b_1, b_2)

Table 1: Specifying NPW boundaries

components of the flow. Alternatively, for this project, they could be the intensity and texture measures in a greyscale CT scan. For simplicity, a linear relation $y = ax + b$ is assumed for the data contributing to a component NP window. We have $y_1 = a_1 x + b_1$ and $y_2 = a_2 x + b_2$, giving two sets of parameters (a_1, b_1) and (a_2, b_2). In vector notation, which we will use throughout this section:

$$\vec{y} = \vec{a}x + \vec{b} \tag{1}$$

$$\text{where } \vec{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \vec{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \text{ and } \vec{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \text{ for } 0 \leq x \leq 1.$$

Assuming a uniform distribution for $x : F_x(x)$ and use i as the indexer to elements in the vectors such that $i = \{1, 2\}$.

$$x = \frac{y_i - b_i}{a_i} : F_x(x) = 1; \tag{2}$$

The joint distribution $F_{y_1, y_2}(x)$ or $F_{\vec{y}}(\vec{y})$ is then given by:

$$F_{y_i}(y_i) = \frac{1}{|dy_i/dx|} F_x(x) = \frac{1}{|dy_i/dx|} F_x\left(\frac{y_i - b_i}{a_i}\right) \tag{3}$$

$$\text{such that } \left[\frac{d\vec{y}}{dx}\right] = \begin{bmatrix} \frac{dy_1}{dx} \\ \frac{dy_2}{dx} \end{bmatrix} \text{ for } b_1 \leq y_1 \leq a_1 + b_1,\ b_2 \leq y_2 \leq a_2 + b_2.$$

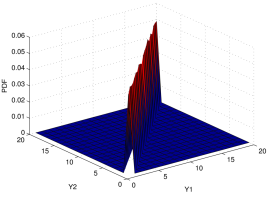
The modulus in this case is the diagonal length of a right triangle formed by a_1 and a_2, so,

$$F_{y_i}(y_i) = \frac{1}{\sqrt{a_1^2 + a_2^2}} F_x\left(\frac{y_i - b_i}{a_i}\right) = \frac{1}{\sqrt{a_1^2 + a_2^2}} \tag{4}$$

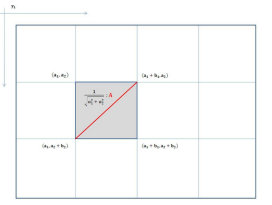
Therefore the 1-D NPW estimation for a 2-tuple vector can be found as:

$$F_{\vec{y}}(\vec{y}) = \begin{cases} \dfrac{1}{\sqrt{a_1^2 + a_2^2}} & \text{for region A and } \vec{a} \neq 0 \quad \text{(5a)} \\ 1 & \text{when } \vec{a} = 0 \quad \text{(5b)} \end{cases}$$

Note from a histogram estimate of a 2-tuple vector signal, A is simply the diagonal line crossing the region defined by the component NP window. More specifically NPW boundaries A can be written analytically, as given in Table 1.



(a) Histogram Estimation for 2-tuple vector

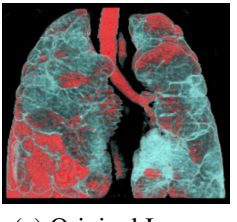


(b) Regional boundaries

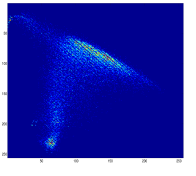
Figure 2: b) illustrates NPW estimator for a 2-tuple vector, range A is a diagonal crossing the region highlighted in grey. Shown here is one of the seven possible cases, i.e $a_1, a_2 > 0$ Note this is only the idealised scenario where the diagonal connects the corners of a defined area, detailed binning operations are necessary in the algorithm implementation

3 Experimental Results

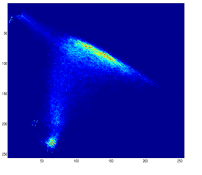
To validate our implementation of the 1-D 2-tuple NPW, we have estimated the averaged joint distributions (estimating the scanline PDFs followed by computing their algebraic mean) in two notable medical applications. The first is an estimation of the two colour channels (red and green) of a coloured CT scan of the lung (used for diagnosing emphysema, a lung disease characterised by abnormal enlargement airspaces distal to terminal bronchioles, shown in Fig. 3). The purpose is to assess the functionality of our implementation and compare results to ground truth, which in this case, is the 1-D 2-tuple histogram estimator. We then applied the algorithm to estimate the joint distribution of scanline pairs in a thoracic CT image (Fig. 4). We first considered a pair of two adjoining scanlines and then two remotely separate scanlines, all taken from the same image slice. All PDFs shown in the figures are normalised.



(a) Original Image



(b) Histogram

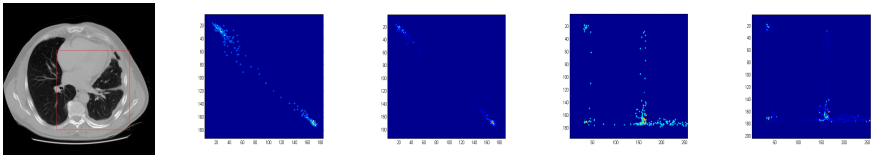


(c) NPW

Figure 3: Exp. I: Lung CT for diagnosing emphysema, performed at the same time as coronary artery CT, giving values for channels R,G; b) and c) show the peak compositions in these channels that make up the dominant colours in the scan.

4 Discussion

To evaluate the accuracy of NPW, L-2 norm defined by $L_2 = \sqrt{\Sigma_i (u_{His}(i,j) - u_{NPW}(i,j))^2}$ is used; where $u_{His}(i,j)$ and $u_{NPW}(i,j)$ are histogram and NPW estimations, respectively. The processes are also timed in order to assess the computational efficiency of our method. (Table 2) It should be noted that the NPW estimator showed a consistent high level of accuracy and good computational efficiency compared to the histogram estimator for both



(a) (b) (c) (d) (e)

Figure 4: Exp. II: a) Thoracic CT slice of a MPM patient; Region of interest outlined in red b),c) Histogram estimate of adjoining and separate scanlines, respectively; d), e)NPW estimate of the same scanline pairs)

Experiment	Time-Hist(s)	Time-NPW (s)	L-2
1	0.008395	0.008396	7.68e-3
2	0.007440	0.007480	6.61e-3

Table 2: Performance of NPW

experiments. The smoothing effect of NPW over histogram is also clearly observed in both cases. In the first experiment, we observe two peaks which correspond to the two dominant colours in the scan. Also note the scattering effect in the distant scanline case in Fig. 4, which complies with our prediction that attenuation gradually changes across the scanned region. The reduction of this effect indicates a greater degree of correlation hence giving light to scanline registration.

In this paper, we have derived and implemented the theories of NPW estimation for 1D 2-tuple vector signals. The immediate next step is the extension and implementation of NPW for 1D N-tuple vectors followed by the 2D N-tuple case. The latter would enable us to apply the vector-spaced NPW method to a wider range of applications. This includes a good use of the theories in the field of multi-modial registration where both image intensity and entropy are involved. Additionally, it is possible as future work to apply the method to estimate the joint distributions of image intensities with other key image quantities such as texture and entropy. Image texture is mostly image technique-dependent and is hard to accurately quantify. Tissue heterogeneity can, for example, be measured by information-theoretic entropy $H = -\sum_i P(i) log P(i)$ where $P(i)$ is the probability at value i. Higher entropy values suggest a more heterogeneous intensity distribution and vice versa.

References

[1] M. Chen, E. Helm, N. Joshi, M. Brady, and F. Gleeson. Feasibility evaluation of thoracic ct automatic segmentation for malignant pleural mesothelioma treatment. *MIUA*, pages 214–218, 2009.

[2] N. Joshi. Non-parametric probability density function estimation for medical images. *D.Phil Thesis, University of Oxford*, 2007.

[3] E. Parzen. On the estimation of a probability density function and mode. *The annals of mathematical statistics*, pages 1065–1076, 1962.

[4] A. Rajwade, A. Banerjee, and A. Rangarajan. A new method of probability density estimation with application to mutual information based image registration. *CVPR*, pages 1769–1776, 2006.