Support Vector Machine Density Estimator as a Generalized Parzen Windows Estimator for Mutual Information Based Image Registration

Sudhakar Chelikani[1], Kailasnath Purushothaman[1], and James S. Duncan[2]

[1] Departments of Diagnostic Radiology,
[2] Departments of Electrical Engineering and Diagnostic Radiology, Yale University, New Haven, CT 06520-8042

Abstract. Mutual Information is perhaps the most widely used multi-modality image registration method. A crucial step in mutual information is the estimation of the probability density function (pdf). In most cases, the Parzen window estimator is employed for this purpose which results in an excessive computational cost. In this paper we demonstrate that replacing the Parzen density estimator with a Support Vector Machine (SVM) density estimation will result in a significant reduction of the computational time. We verified this by registering 2D portal images to DRRs (digitally reconstructed radiographs) projected from 3D CT volumetric data.

1 Introduction

One of the prime motivations behind using *Mutual Information (MI)* in image analysis is the ability to perform multi-modal registration. Since its inception as a similarity metric or measure in 1995 [1,2], it has been successfully applied in many applications concerned with 2D-3D and 3D-3D registration. In principle, mutual information measures the statistical dependence between two images. The metric operates solely on the pixel intensities without reference to any extraneous properties of the object that is imaged, or the imaging system itself. Consequently, it is quite robust when applied across imaging modalities.

At the core of the metric's algorithm is the evaluation of the probability densities of the data sets that are being registered. In most imaging modalities, the probability density functions (pdf) are not easily predicted and need to be estimated. The various ways of predicting or estimating pdfs are grouped under *parametric*, *semi-parametric* and *non-parametric* methods. A Gaussian is a standard example of a parametric pdf, while a finite mixture model such as a *mixture of Gaussians* is an example of a semi-parametric pdf [3]. However, in many cases, non-parametric methods are the models of choice because no prior assumptions about the forms of the pdfs are required. Examples are histograms, K-nearest neighbour method, and kernel based estimators such as the *Parzen Windows* [4]. Viola *et al.* use the Parzen windows density estimator because it provides a very smooth estimate [2]. This is achieved by using the full data set

as input to compute subsequent estimates. Employing large data samples results in an accurate estimate, but with the penalty of a higher computational expense [5].

An alternate method of density estimation using the *Support Vector Method (SVM)* was proposed by Vapnik *et al.* [6,7]. The SVM technique expresses the density estimation problem as a linear operator equation and uses the regression approach with a regularization term to converge to the solution. It can be classified under kernel based estimators. We integrated the SVM estimator into the mutual information registration method to reduce the computational cost of using the Parzen windows estimator. The SVM method is sparser in comparison to the Parzen windows method because many of the data elements are not used in the estimate. The compromise between data sparsity and estimation accuracy is controlled by the use of a regularization term which is used in addition to the kernel width.

2   Mathematical Background

Mutual information has multiple interpretations in standard literature. The *Kullback-Leibler* distance measure interprets the mutual information between two random variables X and Y as the relative entropy between the joint probability distribution of the variables and the product of their marginal distributions. This in essence is an estimate of their statistical dependence. The metric can be expressed as follows:

$$
\begin{aligned}
I\left(X,Y\right) &= D\left(p\left(X,Y\right) \parallel p\left(X\right)p\left(Y\right)\right) \\
&= E_{X,Y}\left[\log \frac{p\left(X,Y\right)}{p\left(X\right)p\left(Y\right)}\right] \\
&= \sum_{x\in X}\sum_{y\in Y} p\left(x,y\right)\log\frac{p\left(x,y\right)}{p\left(x\right)p\left(y\right)}
\end{aligned}
\tag{1}
$$

An alternate formulation is to relate mutual information to the *entropy* of the random variables under consideration. The entropy of a random variable is the expected value of the negative log probability.

$$
H\left(X\right) = E_X\left[-\log p\left(X\right)\right] = -\sum_{x\in X} p\left(x\right)*\log p\left(x\right) \tag{2}
$$

Similarly, the joint entropy between X,Y can be expressed as:

$$
H\left(X,Y\right) = E_{X,Y}\left[-\log p\left(X,Y\right)\right] = -\sum_{x\in X}\sum_{y\in Y} p\left(x,y\right)*\log p\left(x,y\right) \tag{3}
$$

Consequently,
$$
I\left(X,Y\right) = H\left(X\right) + H\left(Y\right) - H\left(X,Y\right) \tag{4}
$$

In the context of registering 2D images to a 3D volume, such as 2D portal, X-Ray fluoroscopy etc. to a 3D CT volumetric data set, the mutual information metric can be formulated in the following manner. Let X represent a set of sample points, and $U(X)$ be the 2D portal image, and $V(T(X))$ be the transformed CT-derived 2D DRR (digitally reconstructed radiograph). The objective function is then written as:

$$
\begin{aligned}
I\left(U\left(X\right), V\left(T\left(X\right)\right)\right) &= H\left(U\left(X\right)\right) + H\left(V\left(T\left(X\right)\right)\right) - H\left(U\left(X\right) V\left(T\left(X\right)\right)\right) \\
&= E_{U,V}\left[\log p\left(U\left(x\right), V\left(T\left(x\right)\right)\right)\right] - E_U\left[\log p\left(U\left(x\right)\right)\right] \\
&\quad - E_V\left[\log p\left(V\left(T\left(x\right)\right)\right)\right] \quad \forall x \in X, \forall y \in Y \qquad (5)
\end{aligned}
$$

The computational burden in mutual information registration lies in accurately estimating the joint and marginal pdfs in equation 5.

2.1   Parzen Window Density Estimation

This method of density estimation was first introduced by Parzen [8]. The probability density is estimated as a linear combination of a set of symmetric kernels, or window functions, that are centered at the sample points. The Parzen method assumes that it is modeling a continuous pdf $p(x)$. A window function, for instance a Gaussian: $g_\psi(u)$ with variance ψ is chosen and we obtain the approximation $p_n(x,a)$ to the density $p(x)$ at the sample point x given the sample set a of size n. It is expressed as:

$$
p_n\left(x, a\right) = \frac{1}{n} \sum_{x_a \in a} g_\psi\left(x - x_a\right) \qquad (6)
$$

This interpolates between the sample points and each point contributes an amount proportional to its distance from x. The parameter that needs to be estimated is the variance. If it is too wide, the estimate will be too coarse and if it is too narrow, it will have too much variability. Consequently, the variance chosen should be optimal, and the computation tends to be intensive if the sample set is large. Another constraint when using the Parzen method is that the sample set used to represent the density should be different from the set used to compute the density approximation. If N_a and N_b be the two sets, then at every point in set N_b we need to calculate the linear combination of N_a window functions, making the total $N_a * N_b$ computations. This computation becomes prohibitively large with increasing N_a and N_b.

2.2   Support Vector Method of Density Estimation

We follow the method suggest by Vapnik *et al.* [6,7] in estimating the density. Let the distribution function corresponding to the density $p(x)$ be $F(x)$. Then,

$$
F\left(x\right) = \int_{-\infty}^{x} p\left(t\right)\, dt \qquad (7)
$$

However, in reality, the distribution function is unknown and only the set of *independent and indentically distributed (iid)* data is provided. The function $F(x)$ is approximated with the empirical function $F_l(x)$ [7]:

$$F_l(x) = \frac{1}{l} \sum_{i=1}^{l} \theta(x - x_i) \tag{8}$$

where $\theta(x)$ is the step-function.

Solving equations 7 with the aproximation in 8 is an ill-posed problem. This implies that small deviations in F_l can cause large deviations in the solution for $p(x)$. In a linear operator representation, we are solving $Ap = F$ with the approximation F_l. One of the standard methods for formulating solutions to such problems was proposed by Tikhonov [9]. A regularizing function $\psi(p)$ is introduced, and the solution is formulated as follows [6]:

$$\psi(p) = \parallel p \parallel^2 \tag{9}$$

This functional is minimized subject to the constraint

$$\max_i \left| F_l(x) - \int_{-\infty}^{x} p(t)\, dt \right|_{x=x_i} = \epsilon_l \tag{10}$$

where ϵ_l is one of the regularization parameters to control accuracy. We now look for a solution of the form

$$p(x) = \sum_{i=1}^{l} \beta_i \mathcal{K}_l(x, x_i) \tag{11}$$

where $\mathcal{K}_l(x, x_i) = \frac{1}{l}\mathcal{K}\left(\frac{x-x_i}{l}\right)$. This a generalized form of the Parzen estimator, where each β_i is a non-uniform weighting coefficient. This is equivalent to minimizing

$$\psi(p) = \parallel p \parallel^2 = \sum_{i,=1}^{l} \beta_i\, \beta_j \mathcal{K}_l(xi, x_j) \tag{12}$$

subject to the constraints

$$\max_l \left| F_l(x) - \sum_{j=1}^{l} \beta_j \int_{-\infty}^{x} K_h(x_i, t)\, dt \right|_{x=x_i} = \epsilon_l$$

$$\beta_i \geq 0, \quad \sum_{i=1}^{l} \beta_i = 1 \tag{13}$$

This is now a standard optimization problem and is solved without much computational expense. Only a few of the β_i are non-zero and the x_i corresponding to those β_i are called the *support vectors* [6,7]. The chosen kernel function will correspondingly have the smallest number of support vectors. Consequently, the support vector method of density estimation is analogous to the Parzen windows method, but with the additional choice of the regularization term, and a sparser data set, and with a faster convergence.

3   Results

To test the Parzen windows density estimator and the support vector estimator, we first employed both methods to compute the density contours of an arbitrary data set. The data set consists of 500 points sampled from a set of three isotropic Gaussian functions. The Parzen method employed the full data set in estimating the densities and required $6*10^4$ computations. The support vector method used only 80 data points chosen as support vectors and converged to a similar density estimate in 16000 iterations. The computation time for the latter was about one third that of the former. Figure 1(a) is a plot of the density iso-contours (contours on which the probability density is constant) using the Parzen method and figure 1(b) is plot of the same using the support vector method. The figures show the data points used for density computation in both cases.
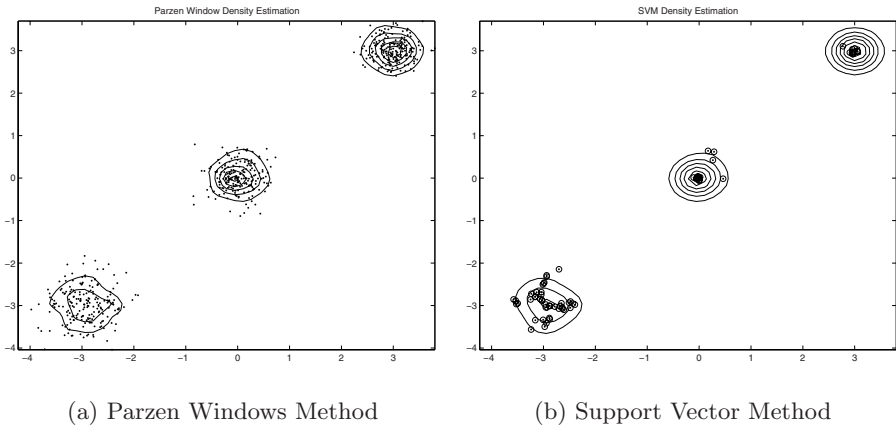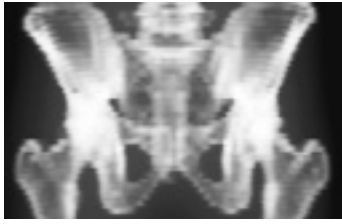

(a) Parzen Windows Method   (b) Support Vector Method

Fig. 1. Comparison of the Parzen and Support Vector methods in density estimation. Figures show contours of constant probability density. The Support vector method used a sparser sample set and computed the same iso-contours in one third the time.

Subsequent to the previous simulation, we incorporated both density estimation techniques into a mutual information code and used it to test the registration of anterior-posterior (AP) images of the pelvis derived from a phantom. One of was a portal image acquired using a high energy (6MeV) radiation beam. This image was registered to two DRRs projected from a CT volume of the phantom. Prior to projecting the DRR, the CT volume was translated by 5 pixels (1 pixel = 1.1mm) along the x-axis in one case, and rotated by 5 degrees about the AP axis in the other. The DRRs thus acquired were registered to the portal image. Figure 2 shows the portal image and a sample DRR image that has been rotated about the AP axis. Figures 3 and 4 are plots of the MI values for the translational and rotational misalignment experiments. The MI function is maximum

(a) AP Portal Image   (b) Misaligned DRR Image

Fig. 2. Portal and DRR images of a pelvic phantom. The latter was rotated by 5° to test the registration algorithm with the Parzen and Support Vector density estimators.

in both instances, and for both density estimation methods, close to the correct transformation values. We implemented both algorithms on a Pentium III, 450 MHz computer and the computation times were 25 min and 9 min respectively for the Parzen and SVM estimators respectively. In estimating the density, the SVM method used approximately 25% to 30% of the data while the Parzen extimator used almost all the available data.
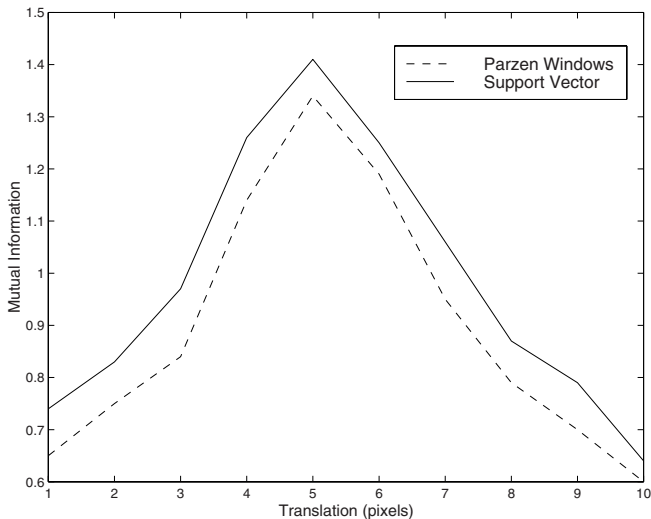

Fig. 3. Plot of the MI function for registration with misalignment of 5 pixels along the x-axis,using the Parzen Windows and the Support Vector methods.
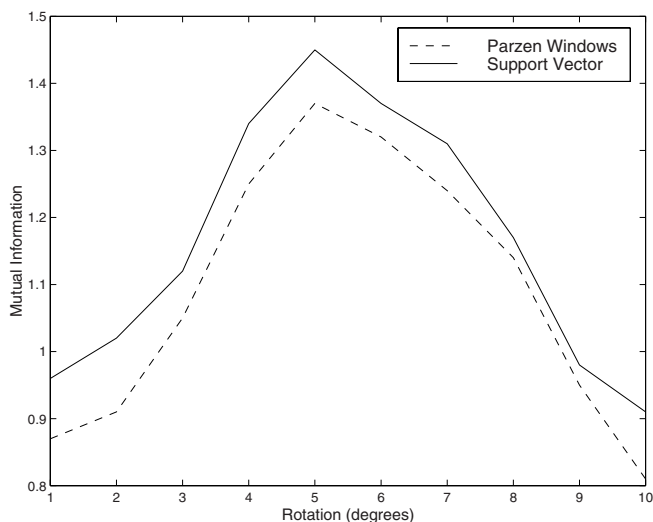
Fig. 4. Plot of the MI function for registration with misalignment of 5° about the AP axis, using the Parzen Windows and the Support Vector methods.

4   Conclusion

The support vector method provided approximately the same density estimation as the popular Parzen windows method with less computational expense. This was verified by incorporating SVM as part of a mutual information registration algorithm. The computation for the Parzen method scaled as $\mathcal{O}\left(N^2\right)$, while the SVM estimator scales as $\mathcal{O}\left(kN^2\right)$, with $k < 1$. The SVM method used about one third the data and was about 2.5 times faster than the Parzen windows method. However, these performance characteristics should be taken only as potential indicators and not as true computational measures. We are in the process of testing the SVM algorithm on real data. The true test lies in how the SVM technique will perform when registering real clinical data. The complexity of such data will pose real challenges to the judicious selection of support vectors. If a corresponding speedup is achieved with clinical data also, it will result in a significant saving of time when performing 2D to 3D and 3D to 3D registration. This will be particularly useful when doing real time registration.

References

1. Collignon, A. and Maes, F. and Delaere, D. and Vandermeulen, P. and Suetens, P. and Marchal, G.: Automated Multi-Modality Image Registration based on Information Theory. *Info. Proc. in Med. Imaging (IPMI)* (1998) 263–74
2. Viola, P. and Wells, W.: Alignment by Maximization of Mutual Information. *Int. J. Comp. Vision* **24** (1997) 137–54

3. McLachlan, G.J. and Peel, D.: Finite Mixture Models. Wiley series in Probability and Mathematical Statistics John Wiley & Sons New York (2000)
4. Izenman, A.J.: Recent Developements in Nonparametric Density Estimation. *J. Am. Stat. Assoc.* **86** (1991) 205–24
5. Thévenaz, P. and Ruttimann, U.E. and Unser M.: A Pyramid Approach to Subpixel Registration based on Intensity. *IEEE Trans. Image Processing* **7** (1998) 27-41
6. Vapnik, N. and Mukherjee, S.: Support Vector Method for Multivariate Density Estimation. CBCL Paper #170AI Memo #1653, Massachusetts Institute of Technology, Cambridge, MA (1999)
7. Weston, J. and Gammerman, A. and Stitson, M. and Vapnik, V. and Vovk, V. and Watkins, C.: Density Estimation using Support Vector Machines. Technical report CSD-TR-97-23, Royal Holloway College, UK (1997)
8. Parzen, E.: On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **33** (1962) 1065–74
9. Tikhonov, A. N.: Solution of Incorrectly Formulated Problems and the Regularization Method. *Soviet Math. Dokl.* **4** (1963) 1035–38