ON STOCHASTIC GRADIENT DESCENT AND QUADRATIC MUTUAL INFORMATION FOR IMAGE REGISTRATION

Abhishek Singh and Narendra Ahuja

Department of Electrical & Computer Engineering
University of Illinois at Urbana-Champaign
abhishek_singh@ieee.org, n-ahuja@illinois.edu

ABSTRACT

Mutual information (MI) is quite popular as a cost function for intensity based registration of images due to its ability to handle highly non-linear relationships between intensities of the two images. More recently, quadratic mutual information (QMI) has been proposed as an alternative measure that computes Euclidean distance instead of KL divergence between the joint and the product of the marginal densities of pixel intensities. In this paper, we examine the conditions under which QMI is advantageous over the classical MI measure, for the image registration problem. We show that QMI is a better cost function to use for optimization methods such as stochastic gradient descent. We show that the QMI cost function remains much smoother than the classical MI measure on stochastic subsampling of the image data. As a consequence, QMI has a higher probability of convergence, even for larger degrees of initial misalignment of the images.

***Index Terms*—** Mutual information, image registration.

1. INTRODUCTION

Mutual information (MI) has seen considerable success as a cost function for registration applications since it was first proposed (independently) by several authors in [1, 2, 3]. As a similarity measure, it is invariant to non-uniform changes in lighting, and different image modalities. It has the ability to work under non-linear intensity relationships between images. It has been successfully employed to learn a variety of parametric transformations, as well as dense deformation fields [4, 5, 6] for aligning multimodal medical images.

Essentially, the goal of a mutual information based registration algorithm is to maximize the statistical dependence between the intensities of the two images. This dependence is measured in terms of the distance between the joint density of the intensities and the product of the marginal densities. In the classical definition of mutual information, this 'distance' is the Kullback-Leibler divergence. Computing classical mutual information therefore involves the following steps: 1) Estimating the joint and marginal densities of the intensities from the two images, commonly done using the Parzen windowing method [7, 8, 9] and, 2) Approximating expectations using the sample average (for computing the KL divergence).

Quadratic mutual information (QMI) has been proposed as an alternative measure of mutual information that uses the Euclidean distance instead of the KL divergence above [10]. It has been shown that computation of QMI can be done using a simple pairwise interaction model of samples, using Parzen windows [10].

Although QMI has been used as a registration criterion for image alignment tasks before [11, 12, 6], the reasons and conditions under which QMI outperforms classical MI for registration problems have not been analysed.

In this paper, we explore the conditions under which QMI is advantageous as a cost function as compared to classic MI, for the image registration problem. Through systematic experiments, we show that although both involve the same complexity of computation (specifically for image registration tasks), the QMI sample estimator is more robust to stochastic subsampling of the pixels as compared to classical MI. That is, the QMI estimator exhibits much smaller variance as compared to classical MI when computed across different sets of i.i.d samples. Since image data often tends to yield a large number of intensity samples (particularly 3D images), robustness to stochastic subsampling makes QMI a particularly attractive cost function for the image registration problem. Our results show that due to these properties, the use of QMI allows for faster convergence and higher probability of convergence when using stochastic gradient descent optimization, as compared to classical MI.

2. MUTUAL INFORMATION BASED REGISTRATION

2.1. Notation and Problem Formulation

Consider two p-dimensional images I_1 and I_2, defined over a discrete spatial region Ω (a bounded region of $\mathbb{Z}^p$).

To register the two images, we look for a transformation $T : \Omega \to \Omega$, that maximizes a cost function of the form,

$$\mathcal{I}(T) = \mathcal{I}\left(I_1(\mathbf{x}), I_2(T(\mathbf{x}))\right) \tag{1}$$

where $\mathcal{I}(T)$ measures the similarity between the first image $I_1(\mathbf{x})$ and the transformed second image $I_2(T(\mathbf{x}))$.

Let $i_1 = I_1(\mathbf{x})$, and $i_2 = I_2(T(\mathbf{x}))$, be d-dimensional ($d = 1$ for grayscale, 3 for RGB etc) intensities of the images I_1 and I_2 at locations $\mathbf{x}$ and $T(\mathbf{x})$ respectively. Furthermore, define $\mathbf{i} = [i_1, i_2]$ and $\mathbf{I}_T(\mathbf{x}) = [I_1(\mathbf{x}), I_2(T(\mathbf{x}))]$.

Let the joint density of i_1 and i_2 be $p(\mathbf{i}, T)$ and the marginal densities be $p(i_1)$ and $p(i_2, T)$ respectively.

The joint and marginal densities can be estimated using the Parzen windowing technique [7] as follows:

$$\hat{p}(\mathbf{i}, T) = \frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\phi_a(\mathbf{I}_T(\mathbf{x}) - \mathbf{i}) \tag{2}$$

$$\hat{p}(i_1) = \frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\phi_a(I_1(\mathbf{x}) - i_1) \tag{3}$$

$$\hat{p}(i_2, T) = \frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\phi_a(I_2(T(\mathbf{x})) - i_2) \tag{4}$$

where $\phi_a(.)$ is a unit volume kernel function with width parameter a.

2.2. Classical Mutual Information

Mutual information is conventionally defined as the Kullback-Leibler divergence between $p(\mathbf{i}, T)$ and $p(i_1)p(i_2, T)$,

$$\mathcal{I}(T) = \int_{\Re^{2d}} p(\mathbf{i}, T) \log \left(\frac{p(\mathbf{i}, T)}{p(i_1)p(i_2, T)} \right) d\mathbf{i} \tag{5}$$

Some straightforward manipulations yield,

$$\mathcal{I}(T) = E\left[\log p(\mathbf{i}, T)\right] - C_1 - E\left[\log p(i_2, T)\right], \tag{6}$$

where the entropy of the reference image I_1 is independent of the transformation T, and is therefore a constant (C_1) for optimization purposes.

We use the Parzen density estimators as described above to estimate $\mathcal{I}(T)$ yielding,

$$\begin{aligned} \hat{\mathcal{I}}(T) &= E\left[\log\left(\frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\phi_a(\mathbf{i} - \mathbf{I}_T(\mathbf{x}))\right)\right] - C_1 \\ &\quad -E\left[\log\left(\frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\phi_a(i_2 - I_2(T(\mathbf{x})))\right)\right] \end{aligned} \tag{7}$$

The next step in the estimation is the approximation of the expectations. A common approach is to use the empirical distribution (or equivalently, the strong law of large numbers), in which case the expectations can be approximated as:

$$\begin{aligned} \hat{\mathcal{I}}(T) &\approx \frac{1}{|\Omega|}\sum_{\mathbf{y}\in\Omega}\log\left(\frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\phi_a(\mathbf{I}_T(\mathbf{y}) - \mathbf{I}_T(\mathbf{x}))\right) - C_1 \\ &- \frac{1}{|\Omega|}\sum_{\mathbf{y}\in\Omega}\log\left(\frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\phi_a(I_2(T(\mathbf{y})) - I_2(T(\mathbf{x})))\right) \end{aligned} \tag{8}$$

2.3. Quadratic Mutual Information

Instead of the KL divergence, quadratic mutual information is the Euclidean distance between $p(\mathbf{i}, T)$ and $p(i_1)p(i_2, T)$ [10],

$$\mathcal{I}_{ED}(T) = \int_{\Re^{2d}} \left(p(\mathbf{i}, T) - p(i_1)p(i_2, T)\right)^2 d\mathbf{i} \tag{9}$$

This can be expanded as,

$$\begin{aligned} \mathcal{I}_{ED}(T) &= \int_{\Re^{2d}} p^2(\mathbf{i}, T) d\mathbf{i} + \int_{\Re^{2d}} p^2(i_1)p^2(i_2, T) d\mathbf{i} \\ &- 2\int_{\Re^{2d}} p(\mathbf{i}, T)p(i_1)p(i_2, T) d\mathbf{i} \end{aligned} \tag{10}$$

Let us denote the three terms in the above expression as $\mathcal{I}_{ED1}(T)$, $\mathcal{I}_{ED2}(T)$ and $\mathcal{I}_{ED3}(T)$, such that $\mathcal{I}_{ED}(T) = \mathcal{I}_{ED1}(T) + \mathcal{I}_{ED2}(T) - \mathcal{I}_{ED3}(T)$.

We now substitute Parzen density estimators of Eqns. 2, 3 and 4 into each of the three terms of $\mathcal{I}_{ED}(T)$. The first term, $\mathcal{I}_{ED1}(T)$, now becomes,

$$\hat{\mathcal{I}}_{ED1}(T) = \frac{1}{|\Omega|^2}\sum_{\mathbf{x}\in\Omega}\sum_{\mathbf{y}\in\Omega}\int_{\Re^{2d}}\phi_a(\mathbf{I}_T(\mathbf{x}) - \mathbf{i})\phi_a(\mathbf{I}_T(\mathbf{y}) - \mathbf{i}) d\mathbf{i} \tag{11}$$

We observe that the integral is now simply computing a convolution of two kernel functions. For many commonly used kernels, this convolution can be easily and analytically computed. For example, convolving two identical Gaussian kernels yields another Gaussian kernel with a scaled width parameter.

The above equation can therefore be *exactly* computed without any approximations as,

$$\hat{\mathcal{I}}_{ED1}(T) = \frac{1}{|\Omega|^2}\sum_{\mathbf{x}\in\Omega}\sum_{\mathbf{y}\in\Omega}\psi_b(\mathbf{I}_T(\mathbf{x}) - \mathbf{I}_T(\mathbf{y})) \tag{12}$$

where,

$$\psi_b(.) = \phi_a * \phi_a(.). \tag{13}$$

Such a closed form solution for the integral is not possible while estimating mutual information using the classical definition of (5).

A similar procedure follows for computing the second term $\mathcal{I}_{ED2}(T)$ to yield,

$$\begin{aligned} \hat{\mathcal{I}}_{ED2}(T) &= \frac{1}{|\Omega|^2}\sum_{\mathbf{x}\in\Omega}\sum_{\mathbf{y}\in\Omega}\psi_b\left(I_1(\mathbf{x}) - I_1(\mathbf{y})\right) \\ &\times \frac{1}{|\Omega|^2}\sum_{\mathbf{x}\in\Omega}\sum_{\mathbf{y}\in\Omega}\psi_b\left(I_2(T(\mathbf{x})) - I_2(T(\mathbf{y}))\right) \\ &= C_2 \times \frac{1}{|\Omega|^2}\sum_{\mathbf{x}\in\Omega}\sum_{\mathbf{y}\in\Omega}\psi_b(I_2(T(\mathbf{x})) - I_2(T(\mathbf{y}))), \end{aligned} \tag{14}$$

where C_2 is computed using the reference image I_1 alone and is therefore a constant while optimization over the transformation T. After some algebraic manipulations with the assumption that $\phi_a(.)$ is a separable kernel, we can exploit the kernel convolution property again to compute the third term $\mathcal{I}_{ED3}(T)$ as,

$$\begin{aligned} \hat{\mathcal{I}}_{ED3}(T) &= \frac{2}{|\Omega|^3}\sum_{\mathbf{x}\in\Omega}\sum_{\mathbf{y}\in\Omega}\sum_{\mathbf{z}\in\Omega}\Big[\psi_b(I_1(\mathbf{x}) - I_1(\mathbf{z})) \\ &\quad \times\psi_b(I_2(T(\mathbf{x})) - I_2(T(\mathbf{y})))\Big] \qquad (15) \\ &= \frac{2}{|\Omega|^2}\sum_{\mathbf{x}\in\Omega}\sum_{\mathbf{y}\in\Omega}C_3(\mathbf{x})\psi_b(I_2(T(\mathbf{x})) - I_2(T(\mathbf{y}))), \qquad (16) \end{aligned}$$

where $C_3(\mathbf{x}) = \frac{1}{|\Omega|}\sum_{\mathbf{z}\in\Omega}\psi_b(I_1(\mathbf{x}) - I_1(\mathbf{z}))$ is a function of the reference image I_1, and is independent of T.

2.4. Comparisons

From (8), the computational complexity for computing the classical MI is $\mathcal{O}(N^2)$, where $N = |\Omega|$ is the number of pixels using which MI is computed.

Fundamentally, computing QMI is an $\mathcal{O}(N^3)$ operation [10], due to the the third term I_{ED3} as seen in (15). However, for the image registration problem, QMI offers a unique advantage - since the reference image I_1 remains fixed throughout, the innermost sum in Eqn. 15 ($C_3(\mathbf{x})$) remains constant and can be precomputed and stored. Therefore, while running an iterative optimization algorithm, $C_3(\mathbf{x})$ needs to be computed just once, and all subsequent computations of QMI in every iteration have complexity $\mathcal{O}(N^2)$, which is the same as classical MI.

As described, QMI and classical MI differ in the divergence measure they employ to measure the discrepancy between the joint and the product of the marginals. It has been argued in [13] and later justified in [14, 15] that if the aim is not to calculate an absolute value of the divergence but rather to find a distribution that minimizes/maximizes the divergence, then a rather relaxed family of divergence measures (including Euclidean distance) can be employed while keeping the result of the optimization the same distribution. This suggests that if both QMI and classical MI are allowed to reach their optimum values in a registration problem, QMI cannot be expected to yield a 'better' solution than classical MI.

Where then does the advantage of QMI lie? The advantage lies in the behavior of the estimator. Estimating QMI requires only smooth Parzen windowing operations, wheres the classical MI requires approximating expectations with the empirical or sample average. The empirical average can be viewed as a Parzen windowing procedure with the width of the window approaching zero. It is well known that the estimation variance of the Parzen window method grows with decreasing window width. This only suggests that the variance in the estimation of classical MI should be higher than that of QMI. This difference should manifest itself when both QMI and classic MI are computed using stochastic subsampling of the data.

We verify this hypothesis using a set of carefully designed registration experiments using the stochastic gradient descent algorithm.

2.5. Optimization

Note that QMI and the classical MI are different cost functions and they operate at different numerical ranges of function values. For a fair comparison of optimization performance (rate of convergence etc) of these two cost functions, the choice of the stepsizes μ for both the update rules becomes critical. To make the comparison fair, we therefore propose to normalize the size of the steps, with respect to the magnitude of the gradient. Therefore we use an optimization rule of the form,

$$\hat{T}^{+} = \hat{T} + \mu \frac{\nabla \hat{\mathcal{I}}_{ED}(\hat{T}, n)}{||\nabla \hat{\mathcal{I}}_{ED}(\hat{T}, n)||_2} \tag{17}$$

where n (between 0 and $N = |\Omega|$) is the number of (randomly chosen) samples using which the gradient is computed in each step.

We call the above update scheme the normalized stochastic gradient descent (NSGD) algorithm. This is useful for comparing two different cost functions since the update rule is independent of the magnitude of the gradient (and thus independent of the scales of the cost function). The same stepsize μ can be chosen for both the update rules for a fair comparison.

We use the NSGD algorithm in our simulations, and we show that even for very small n, the QMI cost function remains significantly smoother than the classical MI cost function.

3. EXPERIMENTS

3.1. 1-D Signals

Consider a reference or a base signal as shown in Fig. 1 (left) and a test signal that is created by adding i.i.d Gaussian noise to the base signal, and distorting a part of it to simulate occlusion, as shown in Fig. 1 (right).

We now plot the QMI and MI costs, as a function of a shift or displacement parameter, between these two signals. Fig. 2 shows the cost functions obtained when the number of samples used to compute the quantities are $n = 250, 350$ and 450. We have used a Gaussian kernel $\phi_a(.)$ for density estimation for both QMI and MI. The kernel bandwidth a is computed from samples using the maximum likelihood technique [16, 17, 8], and is found to be around 0.01 for this example. Note that the plots in Fig. 2 are displayed after appropriately scaling the cost functions for a better visual comparison. The scaling does not affect registration if the NSGD algorithm is used for optimization, as discussed earlier.

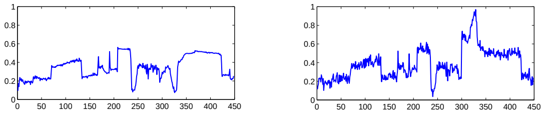

Fig. 1. *Left:* Reference signal. *Right:* Test signal created by altering the reference signal between indices 300 and 400, and then adding Gaussian noise.

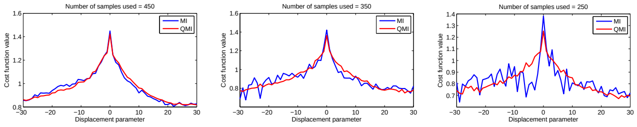

Fig. 2. Plots of the QMI and MI cost functions, while registering the signals of Fig. 1. The number of samples n used for computing the functions is different in each figure. *Left:* n = 450, *Center:* n = 350, *Right:* n = 250.

We notice that the QMI cost surface remains relatively smooth even when using just 250 samples (out of 450 total). To better quantify this difference in smoothness, we compute the variance of both the estimators, at a displacement of 10 samples (fixed). We plot this variance as a function of n, the number of samples used to compute the quantities. The (normalized) variance is computed as,

$$Var\left(\hat{\mathcal{I}}_{ED}(T,n)\right) = \frac{E\left[\left(\hat{\mathcal{I}}_{ED}(T,n) - E\left[\hat{\mathcal{I}}_{ED}(T,n)\right]\right)^2\right]}{E\left[\hat{\mathcal{I}}_{ED}(T,n)\right]^2} \tag{18}$$

where the transformation T is a shift or displacement of 10 samples in this simulation. For every n, the expectations are computed by averaging over 500 Monte Carlo trials with random choice of n samples in each trial.

Fig. 3 shows the plots of variance vs. number of samples used for computation, for three different Parzen window kernel widths a. We see that the QMI estimator remains very stable (low variance) with small n, for any kernel width.

Note that for large kernel widths, both QMI and MI show similar behavior in terms of variance. However, large kernels are unsuitable for registration since they oversmooth the density.

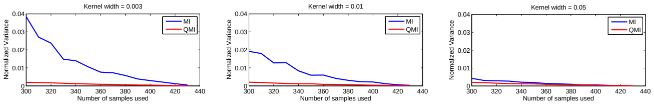

Fig. 3. Variance of the QMI and MI estimators, as a function of the number of samples n used to compute the quantities. The Parzen window kernel width a used for computing densities is different in each figure. *Left:* a = 0.003, *Center:* a = 0.01 (optimal in ML sense), *Right:* a = 0.05.

3.2. 2-D images

Fig. 4 shows the three image pairs that we use in our simulations. The *Brain* image pair consists of multimodal images of the brain, obtained using T1-weighted MRI and Proton-Density MRI respectively

[18]. The *Toys* image pair consist a test image that is obtained after illuminating the scene in the reference image with a closely placed lamp. This creates non-uniform lighting, shadows and specularities, and is therefore challenging for registration. The third image pair, *City*, consists of an aerial view of a city. The test image is a cropped out version of the reference image. This poses a registration challenge since there is significant clutter in the reference image, with very similar looking buildings on all sides.

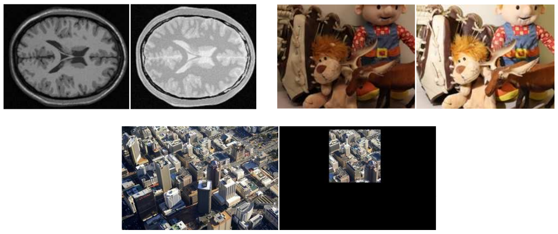

Fig. 4. Image pairs used in the statistical validation of the QMI cost function. *Left: Brain* image. *Center: Toys* image. *Right: City* image.

We create misalignments between the reference images and the test images using a linear transformation model of the form,

$$T^* = \begin{bmatrix} 1+\rho_1 & \rho_3 & 0 \\ \rho_2 & 1+\rho_4 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad (19)$$

where each $\rho_i \sim \mathcal{N}(0, \sigma^2)$ is a normally distributed random variable with mean 0 and variance σ^2. Therefore, T is a 'random' transformation matrix with the parameter σ^2. The goal of image registration is to estimate or recover this transformation matrix as best as possible. Higher values of σ^2 produce transformations that create greater misalignments. We therefore refer to the parameter σ^2 as a measure of the degree of misalignment it produces. We use these random transformation matrices to compute statistical measures of performance, such as probability of convergence vs. degree of misalignment (σ^2) or number of samples used (n).

In our first experiment with these images, we fix the degree of initial misalignment to be $\sigma^2 = 0.1$, and then generate 100 instances of transformation matrices from the above model. We apply these transformations to the three reference images. We then use the test image to estimate these transformations using the NSGD algorithm, with both the QMI cost function and the MI cost function. In order to define successful convergence, we define an error metric between the estimated transform $\hat{T}$ and T^* as the Frobenius norm:

$$\text{Error metric } = ||\hat{T} - T^*||_F \qquad (20)$$

We define convergence if the average value of the error metric over the last 10 iterations of the NSGD algorithm is below 0.005. We count the number of times (out of 100) when each of QMI and MI led to convergence. This count gives us a measure of probability of convergence for each of QMI and MI. We repeat the same procedure with different choices of n (the number of samples used for evaluating the cost function gradient). Fig. 5 shows these results of probability of convergence, for $n = 2000, 1000$ and 500. Clearly, the QMI cost function significantly outperforms the classical MI in all three cases.

For our next set of simulations, we fix the value of n to be 1000. For this value of n, we again compute the probability of convergence for both QMI and MI using 100 random transformation matrices as described above. We now compute this probability for several different values of σ^2, which controls the degree of initial misalignment. We therefore obtain curves that show how the probability of convergence varies with the degree of initial misalignment, for each of the cost functions. These curves are shown in Fig. 6.

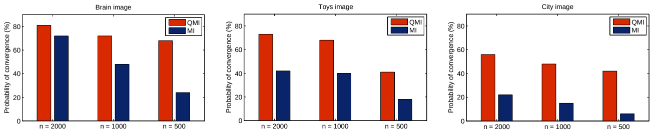

Fig. 5. Probability of convergence vs. number of samples used for computation (n), for QMI and MI cost functions. *Left: Brain* image. *Center: Toys* image. *Right: City* image.

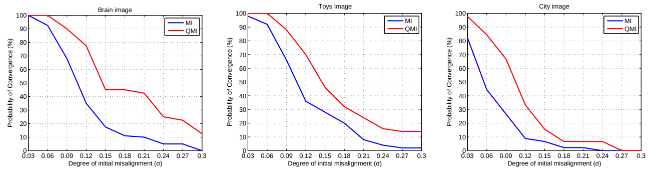

Fig. 6. Probability of convergence vs. degree of initial misalignment (σ^2), for the three image pairs. The QMI cost function maintains a significantly higher probability of convergence throughout. *Left: Brain* image. *Center: Toys* image. *Right: City* image

We can observe that for QMI, the probability of convergence remains significantly higher throughout, for all three image pairs. The QMI cost function, therefore, has a bigger range of convergence as compared to the classical MI.

4. CONCLUSION

In this paper we have investigated the conditions under which QMI is advantageous as a cost function as compared to classic MI, for the image registration problem. We have shown that the QMI cost function remains much smoother when computed with stochastic subsampling, and this leads to greater probability of convergence and greater range of convergence when optimized. Our simulations, though simple, have been chosen to clearly characterize and demonstrate this advantage. To the best of our knowledge, the comparative behavior of these two cost functions under stochastic subsampling for the image registration problem has not been investigated before. In future work, we would be looking into a more principled and theoretical justification of our observations. It would also be interesting to explore how other QMI and MI approximation schemes fit into such a comparison.

5. ACKNOWLEDGMENTS

This work was supported by US Office of Naval Research grant N00014-12-1-0259, the Joan and Lalit Bahl Fellowship and the Computational Science & Engineering Fellowship at UIUC.

6. REFERENCES

[1] P. Viola and III W.M. Wells, "Alignment by maximization of mutual information," in *IEEE ICCV*, 1995.

[2] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *Medical Imaging, IEEE Transactions on*, vol. 16, no. 2, pp. 187 –198, april 1997.

[3] Studhol C., D. L. G. Hill, and D. J. Hawkes, "Automated 3d registration of truncated mr and ct images of the head," in *BMVC*, 1995, BMVC '95, pp. 27–36.

[4] C. Chefd'hotel, G. Hermosillo, and O. Faugeras, "Flows of diffeomorphisms for multimodal image registration," in *IEEE ISBI*, 2002, pp. 753 – 756.

[5] Gerardo Hermosillo, Christophe Chefd'Hotel, and Olivier Faugeras, "Variational methods for multimodal image matching," *IJCV*, vol. 50, pp. 329–343, 2002.

[6] Abhishek Singh, Ying Zhu, and Christophe Chefd'hotel, "A variational approach for optimizing quadratic mutual information for medical image registration," in *IEEE ICASSP*, 2012.

[7] Emanuel Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. pp. 1065–1076, 1962.

[8] P Viola and W.M. Wells, "Alignment by maximization of mutual information," *IJCV*, vol. 24, pp. 137–154, 1997.

[9] William M. Wells III, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, no. 1, pp. 35 – 51, 1996.

[10] Jose C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, Springer, 2010.

[11] J. Atif, X. Ripoche, C. Coussinet, and A. Osorio, "Non rigid medical image registration based on the maximization of quadratic mutual information," in *IEEE Bioengineering Conference*, 2003.

[12] J. Atif, X. Ripoche, and A. Osorio, "Combined quadratic mutual information to a new adaptive kernel density estimator for non rigid image registration," in *SPIE Medical Imaging Conference*, 2004.

[13] J.N. Kapur, *Measures of information and their applications*, Wiley, New Delhi, 1994.

[14] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *JMLR*, 2003.

[15] F. Topsoe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. Information Theory*, 2000.

[16] Abhishek Singh and Jose Principe, "Information theoretic learning with adaptive kernels," *Signal Processing*, vol. 91, no. 2, pp. 2003–2013, 2010.

[17] Abhishek Singh and Jose Principe, "Kernel width adaptation in information theoretic cost functions," in *IEEE ICASSP*, 2010.

[18] Chris A. Cocosco, Vasken Kollokian, Remi K.-S. Kwan, G. Bruce Pike, and Alan C. Evans, "Brainweb: Online interface to a 3d mri simulated brain database," *NeuroImage*, vol. 5, pp. 425, 1997.