Cumulative Residual Entropy, A New Measure of Information & its Application to Image Alignment

F. Wang [1], B. C. Vemuri[1], M. Rao[2] and Y. Chen[2]

[1]*Dept. of CISE,* [2]*Dept. of Mathematics*
University of Florida, *Gainesville, Fl. 32611*

Abstract

In this paper we use the cumulative distribution of a random variable to define the information content in it and use it to develop a novel measure of information that parallels Shannon entropy, which we dub cumulative residual entropy (CRE). The key features of CRE may be summarized as, (1) its definition is valid in both the continuous and discrete domains, (2) it is mathematically more general than the Shannon entropy and (3) its computation from sample data is easy and these computations converge asymptotically to the true values. We define the cross-CRE (CCRE) between two random variables and apply it to solve the uni- & multi-modal image alignment problem for parameterized (rigid, affine and projective) transformations. The key strengths of the CCRE over using the now popular mutual information method (based on Shannon's entropy) are that the former has significantly larger noise immunity and a much larger convergence range over the field of parameterized transformations. These strengths of CCRE are demonstrated via experiments on synthesized and real image data.

1. Introduction

Entropy is a central concept in the field of Information Theory and was originally introduced by Shannon in his seminal paper [13], in the context of communication theory. Since then, this concept and variants thereof have been extensively utilized in numerous applications of science and engineering. To date, one of the most widely benefiting application has been for data compression and transmission. Shannon's definition of entropy originated from the discrete domain and its continuous counterpart called the *differential entropy* is not a direct consequence of the definition in the discrete case. It is well known that the Shannon definition of Entropy in the discrete case does not converge to the continuous definition [6]. Moreover, the definition in the discrete case, which states that the entropy $H(X)$ in a random variable, X, is $H(X) = -\sum_x p(x)log(p(x)$ is based on the density of the random variable $p(X)$, which in general may or may not exist [6] and if does exist, needs to be estimated. The estimates however converge to the true density only under some conditions. Several alternative measures have been defined in literature [10, 1, 7] to overcome some of these drawbacks. In this regard, all of the methods either simply replace the summation with an integral or use the directed divergence from the uniform distribution. The use of directed divergence. For more details, we refer the reader to [7]. However, this approach is not a direct solution to the problem i.e., uses a comparative/relative measure. In this paper, we present a new measure of information in a random variable that will overcome the aforementioned drawbacks of the Shannon entropy and has very general properties as a consequence. This new measure is a *fundamental departure* from all the existing measures of entropy in that it is based on the probability distribution of a random variable rather than its density function. We will also present some interesting properties of this measure and then state some theorems which are proved elsewhere [4]. Following this, we will define a new matching criterion – based on our information theoretic measure – for application to the image alignment problem and compare it to methods that use the Shannon entropy in defining a match measure.

1.1 Previous Work on Image Alignment

In the context of the image alignment problem, information theoretic measures for comparing image pairs differing by an unknown coordinate transformation have been popular since the pioneering works of Viola & Wells [17] and Collignon et.al., [5]. There are numerous methods in literature for solving the image alignment problem. Broadly speaking, these can be categorized as feature-based and direct methods. We will briefly review the direct methods and refer the reader to a recent survey [9] for others.

Sum of squared differences (SSD) has been a popular technique for image alignment [15, 16]. Variants of the original formulation have been able to cope with the deviations from the image brightness constancy assumption [8]. Other matching criteria use of statistical information in the image e.g., correlation ratio [11]. Image alignment is

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

achieved by optimizing these criteria over a set of parameterized coordinate transformations. The statistical techniques can cope with image pairs that are not necessarily from the same imaging modality.

Another direct approach is based on the concept of maximizing mutual information (MI) – defined using the Shannon entropy – reported in Viola and Wells [17], Collignon et al., [5] and Studholme et al., [14]. Reported registration experiments in these works are quite impressive for the case of rigid motion. In [14], Studholme et.al., presented a normalized MI scheme for matching multi-modal image pairs misaligned by a rigid motion. Normalized MI was shown to be able to cope with image pairs not having the same field of view (FOV), an important and practical problem. Most of the effort in the recent past has been spent on coping with non-rigid deformations between the source and target multi-modal data sets [12, 3].

2 Cumulative Residual Entropy: A new measure of information

In this section we define our new information theoretic measure and derive some properties/theorems. *We do not delve into the proofs but refer the reader to a more comprehensive mathematical* – **unpublished** *technical – report [4].*

The *key idea* in our definition is to use the cumulative distribution in place of the density function in Shannon's definition of entropy. The distribution function is more regular because it is defined in an integral form unlike the density function, which is computed as the derivative of the distribution. Moreover, in practice what is of interest and/or measurable is the distribution function. For example, if the random variable describes the life span of a light bulb, then the event of interest is not whether the life span equals t, but whether it exceeds t. Our definition also preserves the well established principle that the logarithm of the probability of an event should represent the information content in the event. We dub this measure as *cumulative residual entropy henceforth abbreviated CRE*.

Definition: Let $\overline{X}$ be a random vector in $\mathcal{R}^N$, we define the CRE of $\overline{X}$, by :

$$\mathcal{E}(\overline{X}) = -\int_{\mathcal{R}_+^N} P(|\overline{X}| > \overline{\lambda}) \log P(|\overline{X}| > \overline{\lambda}) \mathrm{d}\overline{\lambda} \tag{1}$$

Where $\overline{X} = (X_1, X_2, ..., X_N)$, $\overline{\lambda} = (\lambda_1, ...\lambda_N)$ and $|\overline{X}| > \overline{\lambda}$ means $|X_i| > \lambda_i$ and $\mathcal{R}_+^N = \Big(\overline{X} \in \mathcal{R}^N; X_i \geq 0\Big)$. CRE is easily computed for various distributions (in some cases numerically). For example, in the case of the In the case of the exponential distribution with mean $1/\lambda$ and density function: $p(x) = \lambda e^{-\lambda x}$, the CRE computes to, $\mathcal{E}(X) = 1/\lambda$. For the case of the Gaussian distribution, the expression for $\mathcal{E}(X)$ will involve the error function *erf*.

Proposition 1 *$\mathcal{E}(\overline{X}) < \infty$ if for all i and some $p > N, E[|\overline{X}_i|^p] < \infty$; where E is the expectation operator.*

Proposition 2 *If X_i are independent, then*

$$\mathcal{E}(\overline{X}) = \sum_i \Big(\prod_{i \neq j} E(|X_j|)\Big)\mathcal{E}(X_i)$$

Proposition 3 *(Weak Convergence). Let the random vectors $\overline{X}_k$ converge in distribution to the random vector $\overline{X}$; by this we mean*

$$\lim_{k\to\infty} E[\varphi(\overline{X}_k)] = E[\varphi(\overline{X})] \tag{2}$$

for all bounded continuous function ϕ on $\mathcal{R}^N$, if all the $\overline{X}_k$ are bounded in L^p for some $p > N$, then

$$\lim_{k\to\infty} \mathcal{E}(\overline{X}_k) = \mathcal{E}(\overline{X}) \tag{3}$$

This is an important proposition and plays a key role in justifying the computation of CRE from the samples. In contrast with the differential entropy, which can not be computed from sample data, CRE can be easily computed from the samples. Note that, in order to compute the differential entropy, one needs to first estimate the probability density function and then use this estimate to estimate the differential entropy. Thus, there are at-least two levels of estimation processes involved here and this leads to several restrictions under which one may be able to show the convergence of these estimates to the differential entropy. In contrast, no density estimates are required for computing the CRE from samples and therefore no such restrictions apply.

Definition: Given random vectors $\overline{X}$ and $\overline{Y} \in \mathcal{R}^N$, we define the conditional CRE $\mathcal{E}(\overline{X}|\overline{Y})$ by :

$$\mathcal{E}(\overline{X}|\overline{Y}) = -\int_{\mathcal{R}_+^N} P(|\overline{X}| > x|\overline{Y}) \log P(|\overline{X}| > x|\overline{Y}) \mathrm{d}x \tag{4}$$

Proposition 4 *For any $\overline{X}$ and $\overline{Y}$*

$$E[\mathcal{E}(\overline{X}|\overline{Y})] \leq \mathcal{E}(\overline{X}) \tag{5}$$

Equality holds iff *$\overline{X}$ is independent of $\overline{Y}$. This is analogous to the Shannon entropy case. Essentially, it states that conditioning reduces CRE.*

Definition: The continuous version of the Shannon entropy called the differential entropy [6] $\mathcal{H}(X)$ of a random variable X with density f is defined as

$$\mathcal{H}(X) = -E[\log f] = -\int f(x) \log f(x) \mathrm{d}x$$

The following proposition gives a lower bound on $\mathcal{E}(X)$ in terms of the differential entropy $\mathcal{H}(X)$.

Proposition 5 *Let $X \geq 0$ have density f, then,*

$$\mathcal{E}(X) \geq C.\exp(\mathcal{H}(X)), \qquad (6)$$
$$C = \exp(\int_0^1 log(x|\log x|)\mathrm{d}x)$$

Definition: The mutual information $I(X,Y)$ of two continuous random variables X and Y using Shannon entropy is defined as :

$$I(X,Y) = \mathcal{H}(X) - E[\mathcal{H}(X/Y)] \qquad (7)$$

This measure for the discrete random variable case is now widely employed in assessing the misalignment between a pair of uni- or a pair of multi-modality image data sets.

We now define a quantity called cross-CRE (CCRE) given by

$$\mathcal{C}(X,Y) = \mathcal{E}(X) - E[\mathcal{E}(Y/X)] \qquad (8)$$

Note that $I(X,Y)$ is symmetric but $\mathcal{C}(X,Y)$ need not be. We define the symmetrized version of $\mathcal{C}$ as,

$$\tilde{\mathcal{C}}(X,Y) = \frac{1}{2}\Big(\mathcal{E}(X) - E[\mathcal{E}(Y/X)]\Big) + \frac{1}{2}\Big(\mathcal{E}(Y) - E[\mathcal{E}(X/Y)]\Big) \qquad (9)$$

¿From Proposition 4, we know that $\tilde{\mathcal{C}}$ is non-negative. In our experiments, we found that the non-symmetric CCRE given by $\mathcal{C}$ was sufficient to yield the desired results. We empirically show the superior performance of CCRE over MI and normalized-MI under low signal to noise ratio (SNR) conditions and also depict its larger capture range with regards to the convergence to the optimal parameterized transformation.

2.1 Estimating Empirical CRE

In order to compute CRE of an image, we use the histogram of an image to estimate the $P(X > \lambda)$ where X corresponds to the image intensity which is considered as a random variable. Note that as a consequence of proposition 3, empirical CRE computation based on the samples will converge in the limit to the true value. *This is not the case for the Shannon entropy computed using histograms to estimate the probability density functions, as is usually done in current literature.* In the case of CRE, we have,

$$\begin{aligned}\mathcal{E}(X) &= -\int_0^\infty P(X > \lambda)\log P(X > \lambda)\mathrm{d}\lambda \\ &= -\sum_\lambda P(X > \lambda)\log P(X > \lambda) \qquad (10)\end{aligned}$$

Hence, using a histogram to compute the CRE is well defined and justified theoretically.

Note that estimating $\mathcal{E}(X/Y)$ is done using the joint histogram and then marginalizing it with respect to the conditioned variable.

3 The Alignment Problem

The alignment problem is defined as: Given a pair of images $f(x,y)$ and $r(x',y')$, where $(x',y')^t = T\ (x,y)^t$ where T is the matrix corresponding to the unknown parameterized transformation to be determined, define a match metric $M(f(x,y), r(x',y'))$ and maximize/minimize M over all T. In our case, the matching criterion M is defined by CCRE. The class of transformations that we consider are, rigid motions, affine motions and projective transformations.

To show the marked contrast in the range of values taken by $\mathcal{C}$ and I, we compare their ranges in Figure 1 for a pair of registered MR&CT images over a range of rigid motions applied to one of the two given pair of registered images. The second row of Figure 1 shows the zoom in view (at the location of indicated by the arrow) of the range plot between 1° and 2°. Note the significant difference in the range of values of $\mathcal{C}$ and I. It is also evident that CRE is much smoother than the other two measures. As evident from the experiments described later, this characteristic of CCRE will prove to be very useful in demonstrating a large range of convergence and noise immunity for a given optimization procedure over MI defined using the Shannon Entropy. This we believe is a significant strength of our approach. Not only it is easier to find the optimum, but also from numerical computation point of view, it can depict improved tolerance to numerical roundoff errors.

4 Experiment Results

In this section we demonstrate alignment by maximization of CCRE for a variety of transformations. The performance of the CCRE was evaluated for each set. The first experiment (**with 30 image pairs**) was done for synthetic motions, where we compare the estimated alignment with the ground-truth alignments. The second experiment (two pairs of data sets) is done on the real data image pair. In all of the following experiments, bi-linear interpolation was used when needed for non-integral indexing into the image.

4.1 Synthetic Motion Experiments

In this section, we demonstrate the robustness property of CCRE and hence justifying the use of CCRE over MI and NMI (normalized-MI) in the alignment problem. This is demonstrated via experiments depicting superior performance in matching under noisy inputs and larger capture range in the estimation of the motion parameters.

4.1.1 Rigid Motion

In order to compare the robustness property of CCRE versus MI and NMI, we designed a series of experiments as

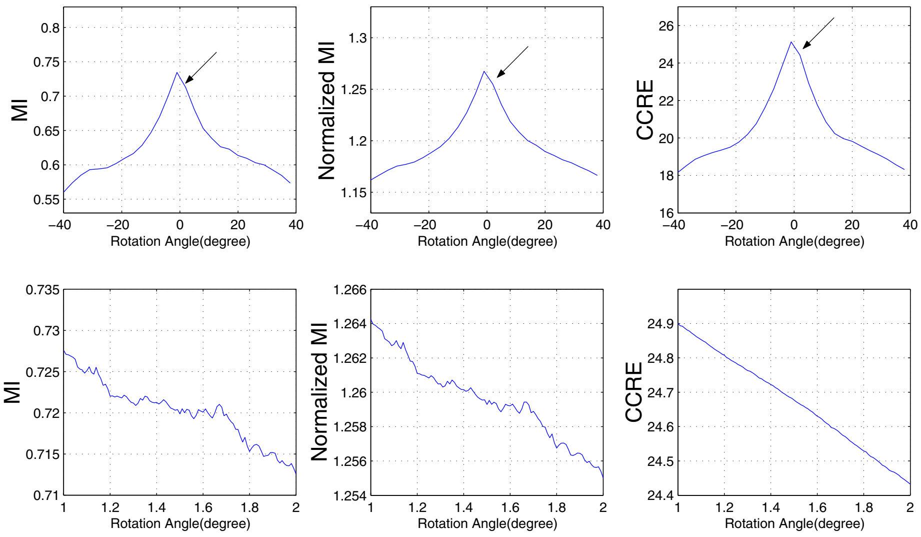

Figure 1: Comparison of the magnitude of $\mathcal{C}$ and I over a range of rotations, for a pair of MR & CT images.

follows: with the MR & CT image pair as our data, we choose the MR image as the source, the target image is obtained by applying a known rigid transformation to the CT image. The source and target image pair along with the result of estimated transformation using CCRE applied to the source with an overlay of the target edge map are shown in Figure 2. The registration is quite accurate as evident visually. Quantitative assessment of accuracy of the registration is presented subsequently.

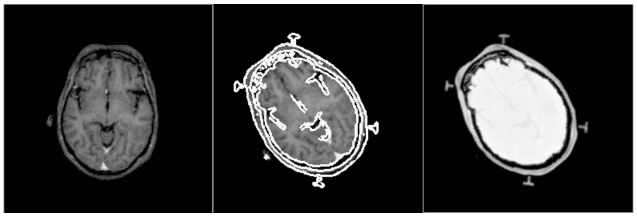

Figure 2: Rigid motion example, Left:The MR (source) image, Right: Synthetically transformed (with a rigid motion) CT (source) image. Middle: Overlay of the target edge map on the transformed source image obtained by applying the CCRE estimate of the rigid motion.

Next, we applied CCRE together with other MI algorithms to estimate motion parameters, with **30** randomly generated rigid transformations. These are normally distributed around the values of $(0^{\circ}, 5pixel, 5pixel)$, with standard deviations of (8°, $3pixel$ $3pixel$) for rotation and translation in x and y respectively. Table 1 shows the statistics of errors resulting from the 3 different methods. In each cell, the leftmost value is the rotation angle (in degrees), while the right two values show the translations in x and y directions. Of the **30** trials, the MI algorithm failed 3 times while CCRE and Normalized MI both failed only once ("failed" here means that the optimization algorithm – sequential quadratic programming (SQP) – primarily diverged). If we only count the cases which gave reasonable results, as shown in the first (for CCRE), second (for traditional MI) and third (for normalized MI) rows, CCRE and the traditional MI have comparable performances, all being very accurate. Thus, in terms of accuracy, CCRE and NMI are comparable and are both better than MI.

	mean			standard deviation		
1	0.057°	0.456	0.286	0.022°	0.236	0.079
2	0.165°	0.645	0.478	0.067°	0.271	0.204
3	0.122°	0.397	0.466	0.040°	0.093	0.077

Table 1: Comparison of estimation errors for rigid motion between CCRE, MI and normalized MI.

In the second experiment, we compare the robustness of the three methods (CCRE, MI and normalized MI) in the presence of noise. Still selecting the aerial image from the previous experiment. as our source image, we generate the target image by applying a fixed synthetic motion. We conduct this experiment by varying the amount of Gaussian noise added and then for each instance of the added noise, we register the two images using the three techniques. We expect all schemes are going to fail at some level of noise. By comparing the noise magnitude of the failure point, we can show the degree to which these methods are tolerant. We choose the fixed motion to be 10° rotation, and 5 pixel translation in both x and y direction. The numerical schemes we used to implement these registrations are all based on sequential quadratic programming (SQP) technique. Table 2 show the registration results for the three schemes. From the table, we observe that the MI

σ	True Motion	CCRE	MI	NMI
13	5 6 6	4.997 6.002 5.997	5.008 5.987 6.004	5.003 6.007 6.022
	5 7 7	4.995 7.004 7.012	0.087 6.988 7.018	5.384 7.995 6.541
	10 10 10	10.015 9.985 9.972	$FAIL$	$0\ -18.748\ -21.041$
	20 10 10	20.002 9.975 9.990		$FAIL$
	30 13 13	31.950 14.037 12.974		
	35 14 14	$FAIL$		

Table 3: Comparison of the convergence range of the rigid registration between CCRE and other MI schemes for fixed noise variance.

fails when the standard deviation of the noise is increased to 15. It is slightly better for normalized MI, which fails at 19, while CCRE is tolerant until 60, *a significant difference when compared to the traditional MI and the normalized MI methods*. This experiment conclusively depicts that CCRE has more noise immunity than both MI and the normalized MI.

σ	CCRE	MI	NMI
10	9.998 5.016 4.996	9.993 4.999 5.007	10.002 5.256 5.235
15	9.998 5.077 5.005	$0\ 6.003\ -3.000$	10.132 5.046 5.998
19	9.998 5.006 5.001	FAIL	$0\ -15.890\ 19.222$
30	9.998 5.256 5.235		FAIL
59	10.027 5.124 4.995		
60	$0\ -3.003\ 0$		
61	FAIL		

Table 2: Comparison of the registration results between CCRE and other MI algorithms for a fixed synthetic motion. The true motion is $(10°, 5, 5)$

Next, we fix the variance of noise and vary the magnitude of the synthetic motion until all of them fail. With this experiment, we can compare the convergence range for each registration scheme. From Table 3, we find that the convergence range of MI and normalized MI is estimated at $(5°, 6, 6)$ and $(9°, 10, 10)$ respectively, while our CCRE-based algorithm has a much larger capture range at $(32°, 13, 13)$. It is evident from this experiment that the capture range for reaching the optimum is significantly larger for CCRE when compared with MI and NMI in the presence of noise. Note that in all the cases, the same numerical optimization scheme – SQP – was used.

4.1.2 Affine Motion

The affine motion experiment was designed as follows: in all the experiments, we misalign MR T1&T2 image pair by a known affine transformation, then try to align them using our CCRE measure . The MR T1&T2 images are originally aligned. For the purpose of comparison, we separate the affine motion into three parts, rotation, translation and scaling. The experiments are similar to the ones for the rigid motion (table 3), we registered the source and target images while varying the synthetic affine motion until the methods fail to find the motion. Each motion parameter is evaluated independently, Table 4 summarizes the results of applying our CCRE algorithm as well as the other MI schemes. The values shown are the maximum capture range (from a zero initial guess) for each parameter in each algorithm. As evident, our algorithm has a significantly larger convergence range.

algorithm	Rotation	Translation	Scaling
CCRE	39°	30	3.2
MI	18°	15	2.2
Normalized MI	21°	14	2.6

Table 4: Convergence range of different algorithms for affine motion.

The last test for the affine motion involves varying the amount of Gaussian noise while fixing the synthetic affine motion. Table 5 depicts the noise variance which causes each algorithm to fail. Again, observe superior performance of CCRE over the other MI-based methods.

algorithm	noise Standard Deviation(σ)
CCRE	19
MI	6
Normalized MI	5

Table 5: Comparison of the registration results for a fixed affine motion, $(1.4772, -0.2605, 5.0000, 0.2605, 1.4772, 5.0000)$ and varying noise levels

4.2 Real Data Experiments

In this section, we demonstrate the algorithm performance for a pair aerial images taken over time. The transformation between the two images is assumed to be a projective transformation. Our data is approximated by a planar surface in motion viewed through a pinhole camera. This motion can be described as 2D projective transformation.

$$\begin{aligned} u(x,y) &= \frac{a_0x + a_1y + a_2}{a_6x + a_7y + 1} - x \\ v(x,y) &= \frac{a_3x + a_4y + a_5}{a_6x + a_7y + 1} - y \end{aligned} \tag{11}$$

This projective transformation requires us to estimate eight parameters for each image pair. **For brevity**, only one registration result is shown in Figure 3. Here, the source and target images are shown in the top row, and the lower left image is the overlay of the transformed source with the source

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

edge map (showing the change in the source due to the applied transformation), while the lower right image shows the overlay with the target edge map showing the registration. As evident, the registration is visually quite accurate.

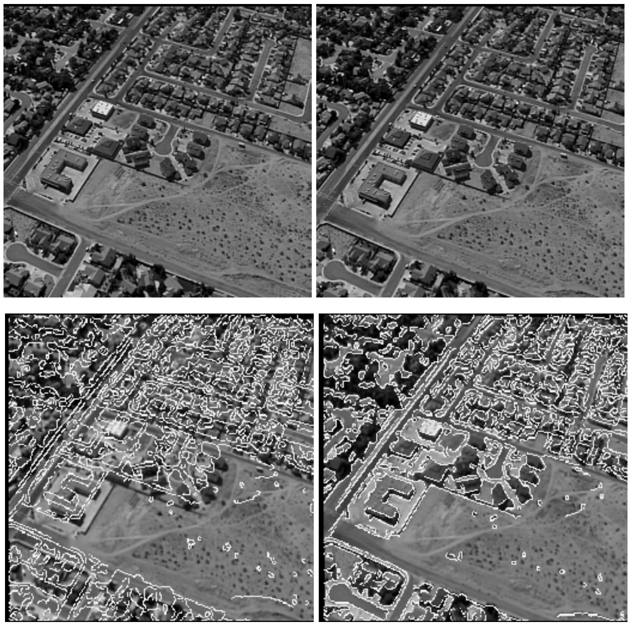

Figure 3: Registration results for the projective transformation.

5 Summary

In this paper, we presented a novel measure of information that we dub cumulative residual entropy (CRE). This measure has several advantages over the well known Shannon entropy whose definition is based on probability density functions which are hard to estimate accurately. In contrast, CRE can be easily computed from the sample data and these computations asymptotically converge to the true value. Unlike Shannon entropy, the same CRE definition is valid for both discrete and continuous domains.

We defined the cross-CRE denoted by CCRE and applied it to estimate the parameterized misalignments between image pairs and tested it on synthetic as well as real data sets from uni-modal (single imaging) source and multi-modality (MR T1 and T2 weighted) imaging sources. Comparisons were made between CCRE and MI and normalized MI both of which were defined using the Shannon entropy. Experiments depicted significantly better performance of CCRE over the other MI-based methods currently used in literature.

Acknowledgments

This research was supported in part by the NIH grant RO1 NS42075. Authors would like to thank Dr. Frank Bova of the UF Neurosurgery Dept. for the MR-CT data sets and Dr. Wen Masters of the ONR for the aerial images.

References

[1] J. Aczel and Z. 'Daroczy, On measures of information and their characterization, Academic Press, New York, 1975.

[2] Simulated brain database, available online at: www.bic.mni.mcgill.ca/brainweb/

[3] C. Chefd'Hotel, G. Hermosillo and O. Faugeras, “A variational approach to multi-modal image matching,” in IEEE Workshop on VLSM, pp. 21-28, 2001.

[4] M.Rao, Y.Chen, B.C.Vemuri and F.Wang, “Cumulative residual entropy, a new measure of information,” Tech. Rep., Institute of Fundamental Theory, Dept. of Mathematics, October 2002.

[5] A. Collignon, et. al., ”Automated multimodality image registration using information theory,” *Proc. IPMI*, pp. 263-274,1995.

[6] Thomas M. Cover, Joy A. Thomas, *Elements of Information Theory,* John Wiley and Sons, 1991.

[7] J. N. Kapur, “On the basis of relationship between measures of entropy and directed divergence,” Proc. of the Nat. Acad. Sci., 58 A(3), 375-387.

[8] S. H. Lai and M. fang, “Robust and efficient image alignment with spatially varying illumination models,” in IEEE CVPR 1999, pp. 167-172.

[9] J.B. Maintz and M. A. Viergever,“A Survey of Medical Image Registration,” *MedIA* Vol. 2, pp. 1-36,1998.

[10] A. Renyi, “On measures of entropy and information,” selected papers of Alfred Renyi, Vol. 2, 1961.

[11] A. Roche et. al., “The correlation ratio as new similarity metric for multi-modal image registration, in MICCAI'98.

[12] D. Ruckert, et.al., “ Non-rigid registration of breast MRI using MI,” in MICCAI98.

[13] C. E. Shannon, “A mathematical theory of communication,” *Bell Sys. Tech. Journ.,* vol. 27, pp. 379-423 and 623-656, 1948.

[14] C. Studholme, et.al., “An overlap invariant entropy measure of 3D medical image alignment,” *Pattern Recognition*, Vol. 32, pp. 71-86,1999

[15] R. Szeliski,J. Coughlan, ”Spline-based image registration,” *IJCV*, 22(3), p.199-218, 1997.

[16] B. C. Vemuri, et.al.,“An efficient motion estimator with application to medical image registration”, *MedIA*, 2(1), pp. 79-98, 1998 .

[17] P. A. Viola and W. M. Wells, ”Alignment by maximization of mutual information,” in *Fifth ICCV*, pp. 16-23, 1995.