Uses of Information Theory in Medical Imaging

Wang Zhan, Ph.D.
Center for Imaging of Neurodegenerative Diseases
Tel: 415-221-4810x2454, Email: Wang.Zhan@ucsf.edu

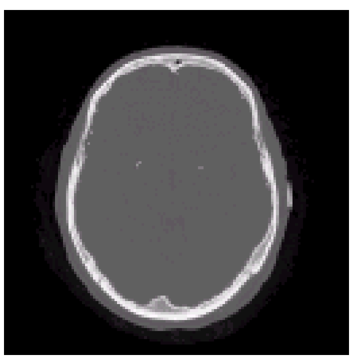
Karl Young (UCSF) and M. Farmar (MSU)

Topics

- Image Registration
  - Information theory based image registration (JPW Pluim, et al, IEEE TMI 2003)
- Feature Selection
  - Information theory based feature selection for image classification optimization (M. Farmer, MSU, 2003)
- Image Classification
  - Complexity Based Image Classification (Karl Young, USF, 2007)

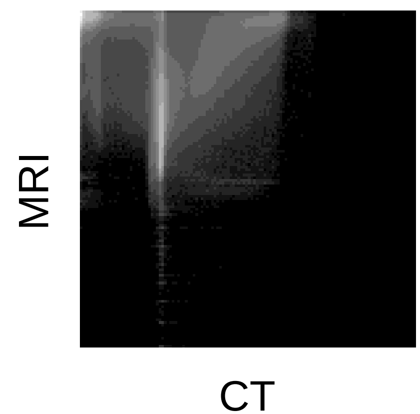
Image Registration

- Define a transform T that maps one image onto another image such that some measure of overlap is maximized (Colin’s lecture).
  - Discuss information theory as means for generating measures to be maximized over sets of transforms



CT

MRI



Three Interpretations of Entropy

- The amount of information an event provides
  - An infrequently occurring event provides more information than a frequently occurring event
- The uncertainty in the outcome of an event
  - Systems with one very common event have less entropy than systems with many equally probable events
- The dispersion in the probability distribution
  - An image of a single amplitude has a less disperse histogram than an image of many greyscales
    - the lower dispersion implies lower entropy

Measures of Information

- Hartley defined the first information measure:
  - H = n log s
  - n is the length of the message and s is the number of possible values for each symbol in the message
  - Assumes all symbols equally likely to occur
- Shannon proposed variant (Shannon's Entropy)

$$H = \sum_i p_i \cdot \log \frac{1}{p_i}$$

  - weighs the information based on the probability that an outcome will occur
  - second term shows the amount of information an event provides is inversely proportional to its prob of occurring

Alternative Definitions of Entropy

- The following generating function can be used as an abstract definition of entropy:

$$H(P) = h\left(\frac{\sum_{i=1}^{M} v_i \cdot \varphi_1(p_i)}{\sum_{i=1}^{M} v_i \cdot \varphi_2(p_i)} \right)$$

- Various definitions of these parameters provide different definitions of entropy.
  - Actually found over 20 definitions of entropy

Measure	$h(x)$	$\varphi_1(x)$	$\varphi_2(x)$	v_i
1	x	$-x\log x$	x	v
2	$(1-r)^{-1}\log x$	x^r	x	v
3	x	$-x^r\log x$	x^r	v
4	$(s-r)^{-1}\log x$	x^r	x^s	v
5	$(1/s)\arctan x$	$x^r\sin(s\log x)$	$x^r\cos(s\log x)$	v
6	$(m-r)^{-1}\log x$	x^{r-m+1}	x	v
7	$(m(m-r))^{-1}\log x$	$x^{r/m}$	x	v
8	$(1-t)^{-1}\log x$	x^{t+s-1}	x^s	v
9	$(1-s)^{-1}(x-1)$	x^s	x	v
10	$(t-1)^{-1}(x^t-1)$	$x^{1/t}$	x	v
11	$(1-s)^{-1}(e^x-1)$	$(s-1)x\log x$	x	v
12	$(1-s)^{-1}(x^{\frac{s-1}{r-1}}-1)$	x^r	x	v

Measure	$h(x)$	$\varphi_1(x)$	$\varphi_2(x)$	v_i
13	x	$-x^r \log x$	x	v
14	$(s-r)^{-1}x$	$x^r - x^s$	x	v
15	$(\sin s)^{-1}x$	$-x^r \sin(s \log x)$	x	v
16	$\left(1+\frac{1}{\lambda}\right)\log(1+\lambda) - \frac{x}{\lambda}$	$(1+\lambda x)\log(1+\lambda x)$	x	v
17	x	$-x \log\left(\frac{\sin(sx)}{2\sin(s/2)}\right)$	x	v
18	x	$-\frac{\sin(xs)}{2\sin(s/2)} \log\left(\frac{\sin(sx)}{2\sin(s/2)}\right)$	x	v
19	x	$-x \log x$	x	w_i
20	x	$-\log x$	1	v_i
21	$(1-r)^{-1}\log x$	x^{r-1}	1	v_i
22	$(1-s)^{-1}(e^x - 1)$	$(s-1)\log x$	1	v_i
23	$(1-s)^{-1}(x^{\frac{r-1}{s-1}} - 1)$	x^{r-1}	1	v_i

#	**Name**	**#**	**Name**	**#**	**Name**	**#**	**Name**
1	Shannon	7	Varma	13	Taneja	19	Belis-Guiasu,Gil
2	Renyi	8	Kapur	14	Sharma-Taneja	20	Picard
3	Aczel-Daroczy	9	Havdra-Charvat	15	Sharma-Taneja	21	Picard
4	Aczel-Daroczy	10	Arimoto	16	Ferreri	22	Picard
5	Aczel-Daroczy	11	Sharma-Mittal	17	Sant’anna-Taneja	23	Picard
6	Varma	12	Sharma-Mittal	18	Sant’anna-Taneja		

Note that only 1 and 2 satisfy simple uniqueness criteria (i.e. unique additive functionals of probability density functions)

Entropy for Image Registration

- Define estimate of joint probability distribution of images:
  - 2-D histogram where each axis designates the number of possible intensity values in corresponding image
  - each histogram cell is incremented each time a pair (I_1(x,y), I_2(x,y)) occurs in the pair of images ("co-occurrence")
  - if images are perfectly aligned then the histogram is highly focused; as the images mis-align the dispersion grows
  - recall one interpretation of entropy is as a measure of histogram dispersion

Entropy for Image Registration

- Joint entropy (entropy of 2-D histogram):

$$H(A,B) = -\sum_{i,j} p(i,j) \cdot \log[p(i,j)]$$

- Consider images “registered” for transformation that minimizes joint entropy, i.e. dispersion in the joint histogram for images is minimized

Example

Joint Entropy of 2-D Histogram for rotation of image with respect to itself of 0, 2, 5, and 10 degrees

3.82

6.79

6.98

7.15

Mutual Information for Image Registration

- Recall definition(s):
  - $I(A,B) = H(B) - H(B|A) = H(A) - H(A|B)$
    - amount that uncertainty in B (or A) is reduced when A (or B) is known.
  - $I(A,B) = H(A) + H(B) - H(A,B)$
    - maximizing is equivalent to minimizing joint entropy (last term)
- Advantage in using mutual info over joint entropy is it includes the individual input's entropy
- Works better than simply joint entropy in regions of image background (low contrast) where there will be high joint entropy but this is offset by high individual entropies as well - so the overall mutual information will be low
- Mutual information is maximized for registered images

Derivation of M. I. Definitions

$$H(A,B)=\sum_{a,b} p(a,b)\cdot\log(p(a,b)), \text{where } \ p(a,b)=p(a|b)\cdot p(b)$$

$$H(A,B)=\sum_{a,b}[p(a|b)\cdot p(b)]\cdot\log[p(a|b)\cdot p(b)]$$

$$H(A,B)=\sum_{a,b}[p(a|b)\cdot p(b)]\cdot\{\log[p(a|b)]+\log[p(b)]\}$$

$$H(A,B)=\sum_{a,b} p(a|b)\cdot\log[p(a|b)]\cdot p(b)+\sum_{a,b} p(b)\cdot\log(p(b))\cdot p(a|b)$$

$$H(A,B)=\sum_{a} p(a|b)\cdot\log[p(a|b)]\cdot\sum_{b} p(b)+\sum_{b}\sum_{a} p(a|b)\cdot p(b)\cdot\log(p(b)$$

$$H(A,B)=\sum_{a} p(a|b)\cdot\log[p(a|b)]+\sum_{b} p(b)\cdot\log(p(b))$$

$$H(A,B)=H(A|B)+H(B)$$

therefore $I(A,B)=H(A)-H(B|A)=H(A)+H(B)-H(A,B)$

Definitions of Mutual Information II

– 3) $$I(A,B) = \sum_{a,b} p(a,b) \cdot \log\left(\frac{p(a,b)}{p(a)\,p(b)} \right)$$

- This definition is related to the Kullback-Leibler distance between two distributions
- Measures the dependence of the two distributions
- In image registration I(A,B) will be maximized when the images are aligned
- In feature selection choose the features that minimize I(A,B) to ensure they are not related.

Additional Definitions of Mutual Information

- Two definitions exist for normalizing Mutual information:
  - Normalized Mutual Information (Colin – improved MR-CT, MR-PET):

$$NMI(A,B) = \frac{H(A) + H(B)}{H(A,B)}$$

  - Entropy Correlation Coefficient:

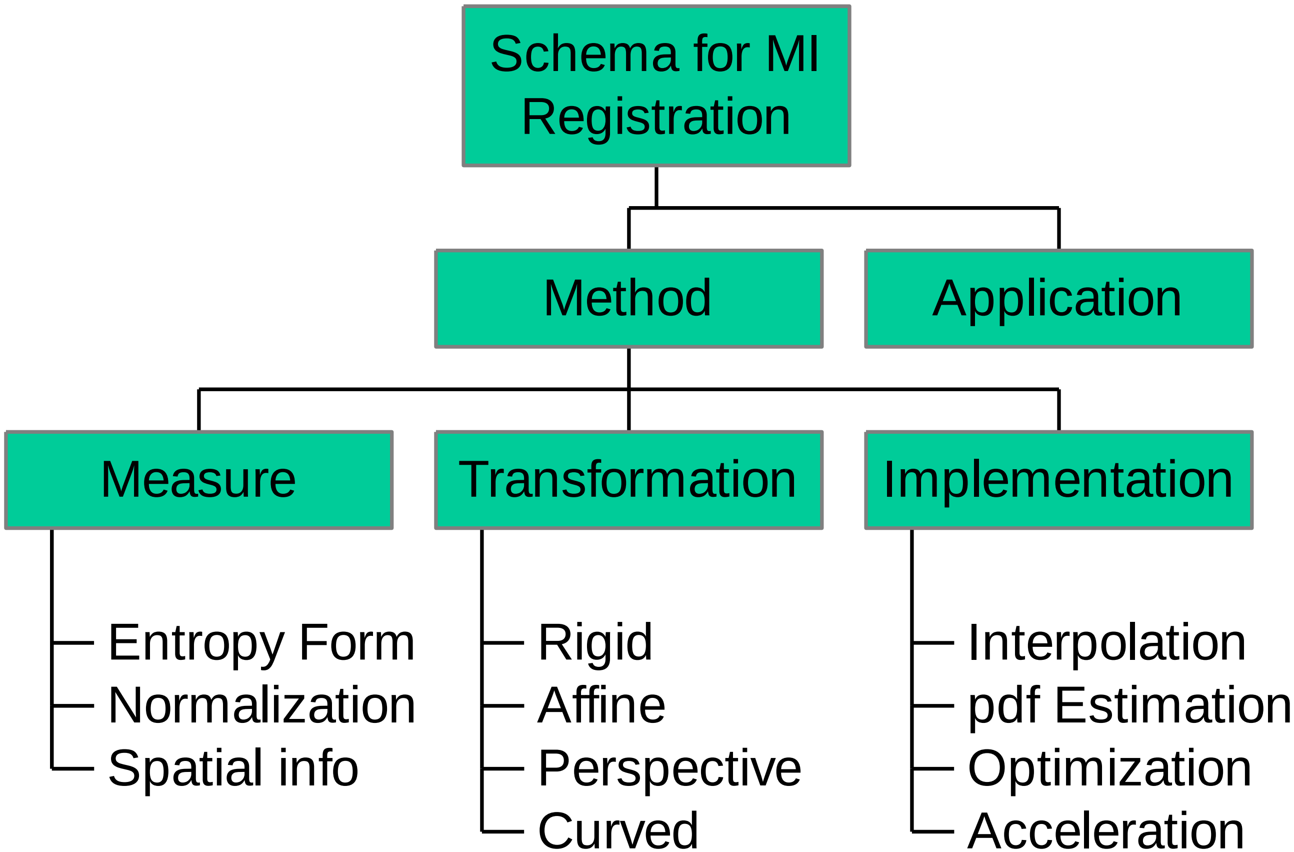
$$ECC(A,B) = 2 - \frac{2}{NMI(A,B)}$$

Properties of Mutual Information

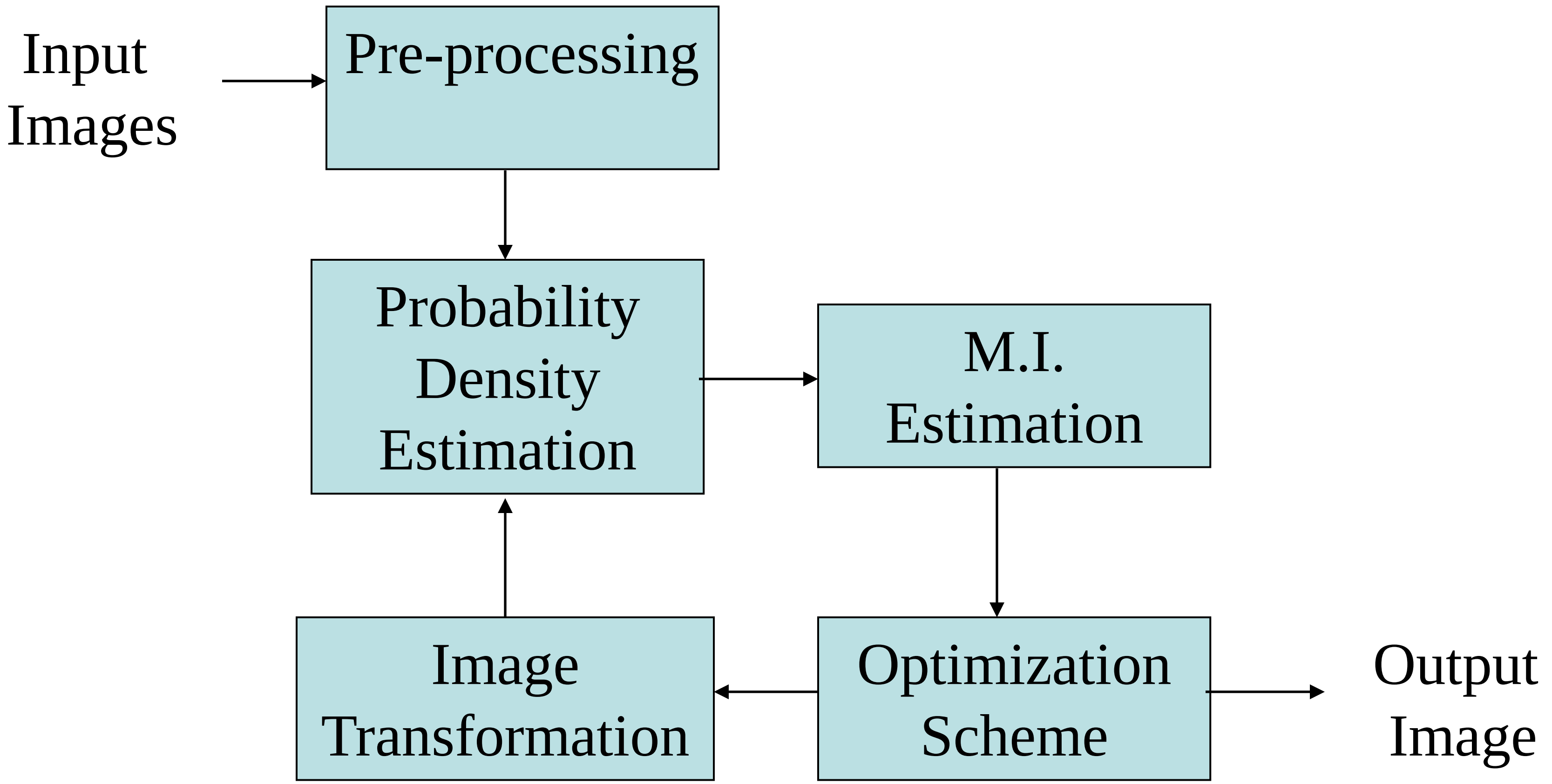
- MI is symmetric: I(A,B) = I(B,A)
- I(A,A) = H(A)
- I(A,B) <= H(A), I(A,B) <= H(B)
  - info each image contains about the other cannot be greater than the info they themselves contain
- I(A,B) >= 0
  - Cannot increase uncertainty in A by knowing B
- If A, B are independent then I(A,B) = 0
- If A, B are Gaussian then:

$$I(A,B) = -\tfrac{1}{2}\log(1-\rho^2)$$

Schema for Mutual Information Based Registration



M.I. Processing Flow for Image Registration



Probability Density Estimation

- Compute the joint histogram h(a,b) of images
  - Each entry is the number of times an intensity a in one image corresponds to an intensity b in the other
- Other method is to use Parzen Windows
  - The distribution is approximated by a weighted sum of sample points Sx and Sy
    - The weighting is a Gaussian window

$$P(x, y, Sx, Sy) = \tfrac{1}{N} \sum_{S} W(Dist(x, y; Sx, Sy))$$

M.I. Estimation

- Simply use one of the previously mentioned definitions for entropy
  - compute M.I. based on the computed distribution function

Optimization Schemes

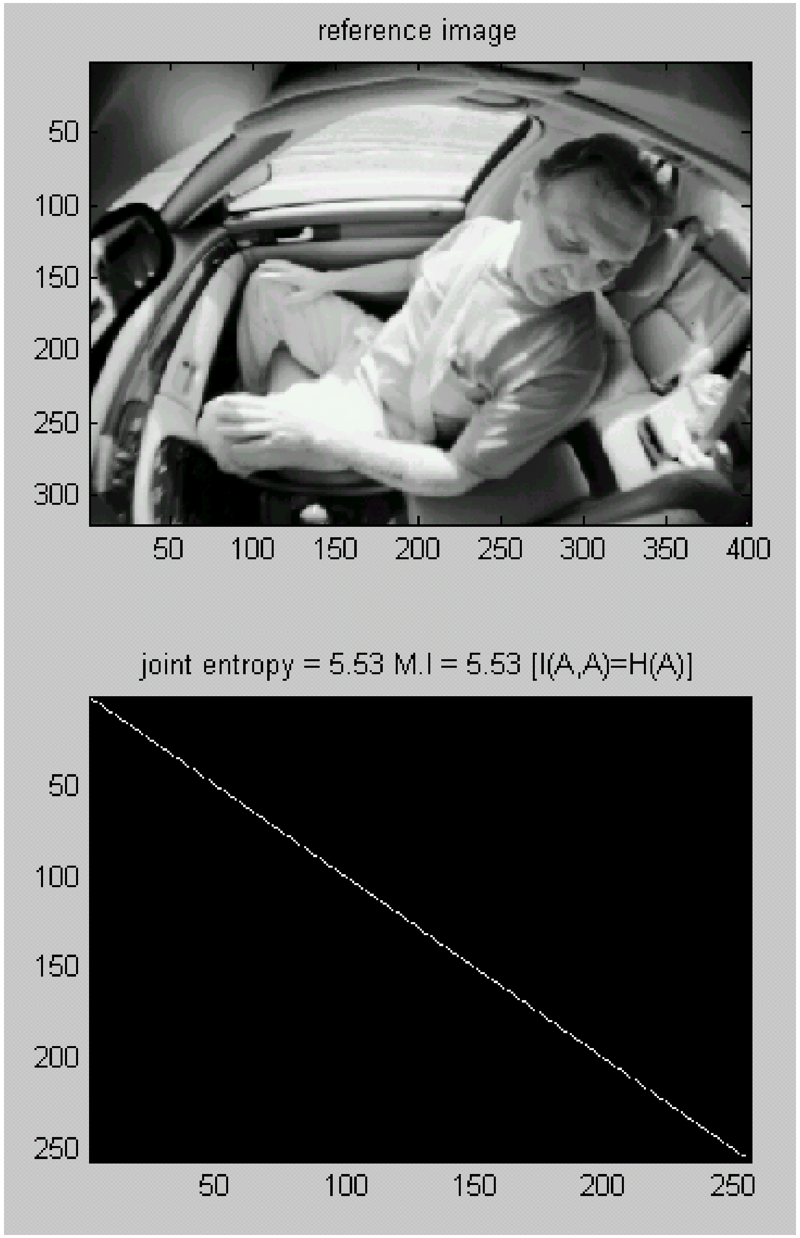
- Any classic optimization algorithm suitable
  - computes the step sizes to be fed into the Transformation processing stage.

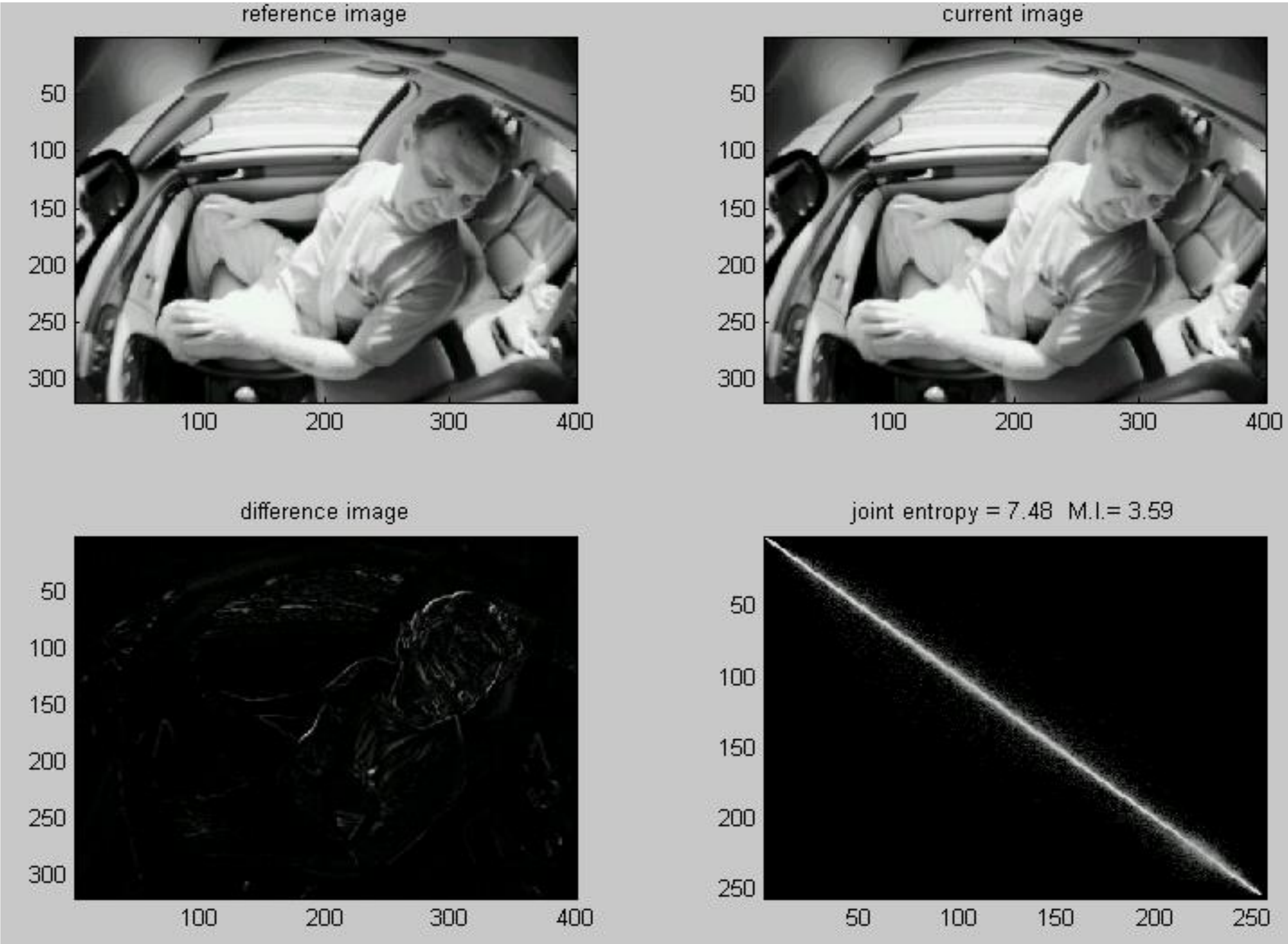
Image Transformations

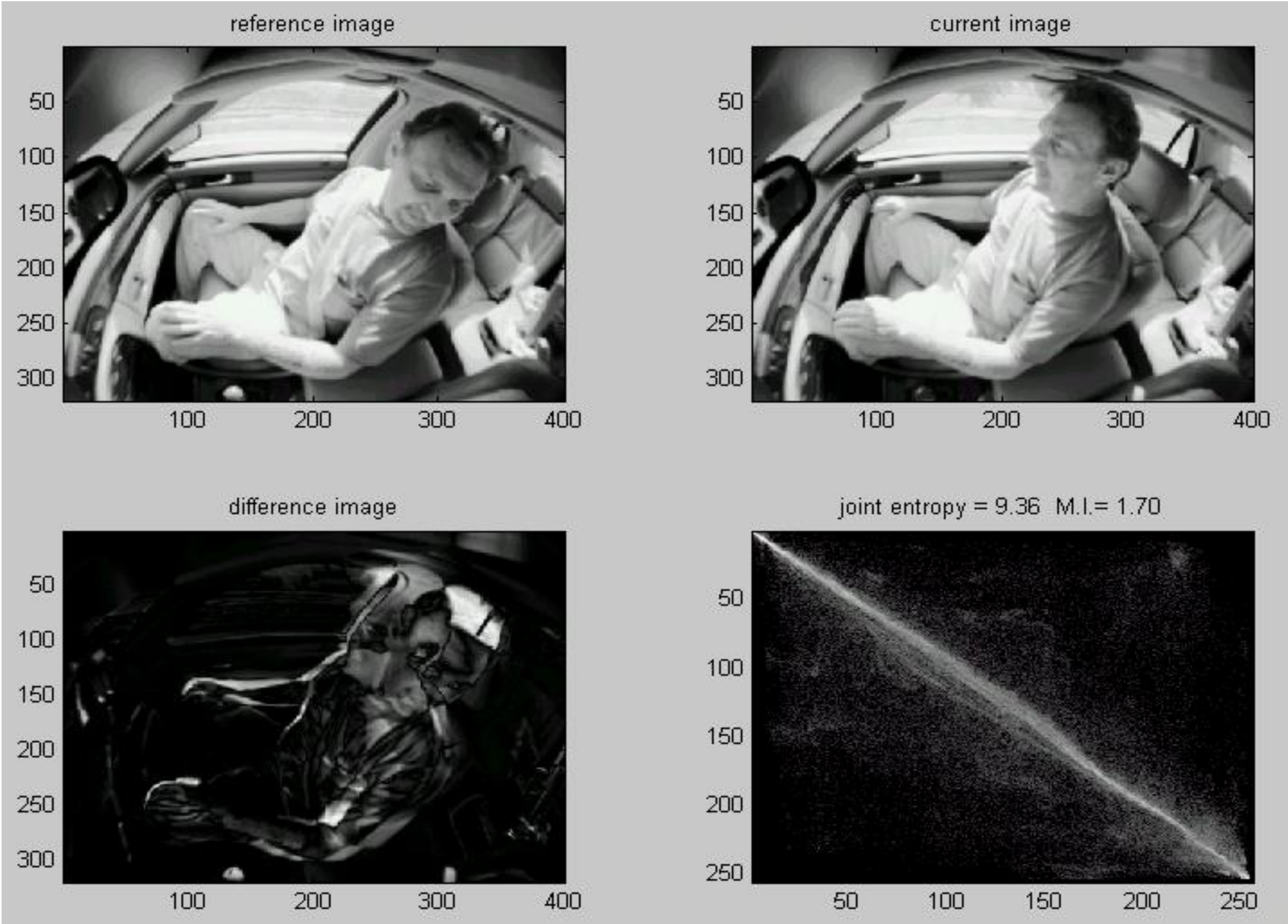
- General Affine Transformation defined by:

$$\mathbf{T}(x, y) = \begin{pmatrix} x \cdot s_{1,1} & y \cdot s_{1,2} & d_x \\ x \cdot s_{2,1} & y \cdot s_{2,2} & d_y \end{pmatrix}$$

- Special Cases:
  - S = I (identity matrix) then translation only
  - S = orthonormal then translation plus rotation
    - rotation-only when D = 0 and S orthonormal.


reference image
50
100
150
200
250
300
50
100
150
200
250
300
350
400
joint entropy = 5.53 M.I = 5.53 [I(A,A)=H(A)]
50
100
150
200
250
50
100
150
200
250


reference image
current image
difference image
joint entropy = 7.48 M.I.= 3.59
50
100
150
200
250
300
100
200
300
400


reference image
current image
difference image
joint entropy = 9.36 M.I.= 1.70
50
100
150
200
250
300
400

Mutual Information based Feature Selection

- Tested using 2-class Occupant sensing problem
  - Classes are RFIS and everything else (children, adults, etc).
  - Use edge map of imagery and compute features
    - Legendre Moments to order 36
    - Generates 703 features, we select best 51 features.
- Tested 3 filter-based methods:
  - Mann-Whitney statistic
  - Kullback-Leibler statistic
  - Mutual Information criterion
    - Tested both single M.I., and Joint M.I. (JMI)

Mutual Information based Feature Selection Method

- M.I. tests a feature's ability to separate two classes.
  – Based on definition 3) for M.I.

$$I(A,B) = \sum_{a}\sum_{b} p(a,b) \cdot \log\left(\frac{p(a,b)}{p(a)\,p(b)}\right)$$

  – Here A is the feature vector and B is the classification
    - Note that A is continuous but B is discrete
  – By maximizing the M.I. We maximize the separability of the feature
    - Note this method only tests each feature individually

Joint Mutual Information based Feature Selection Method

- Joint M.I. tests a feature's independence from all other features:

$$I(A_1, A_2, \ldots, A_N; B) = \sum_{k=1,N} I(A_k; B \mid A_{k-1}, A_{k-2}, \ldots, A_1)$$

- Two implementations proposed:
  - 1) Compute all individual M.I.s and sort from high to low
  - Test the joint M.I of current feature with others kept
    - Keep the features with the lowest JMI (implies independence)
    - Implement by selecting features that maximize:

$$I(A_j, B) - \beta \cdot \sum_k I(A_k, A_j)$$

Joint Mutual Information based Feature Selection Method

- Two methods proposed (continued):
  - 2) Select features with the smallest Euclidean distance from:
    - The feature with the maximum: $I(A_j, B)$
    - And the minimum: $\sum_k I(A_k, A_j)$

Mutual Information Feature Selection Implementation Issue

- M.I tests are very sensitive to the number of bins used for the histograms
- Two methods used:
  - Fixed Bin Number (100)
  - Variable bin number based on Gaussianity of data

$$M_{bins} = \log N + 1 + \log(1 + \kappa \cdot \sqrt{N/6})$$

  - where N is the number of points and k is the Kurtosis

$$\kappa = \frac{1}{\sigma^4 \sqrt{24N}} \cdot \sum_{k=1,N} (x_k - \bar{x})^4 - \sqrt{\frac{3N}{8}}$$

Image Classification

- Specifically: Application of Information Theory Based Complexity Measures to Classification of Neurodegenerative Disease

What Are Complexity Measures ?

- Complexity

Many strongly interacting components introduce an inherent element of uncertainty into observation of a complex (nonlinear) system

Good Reference:

W.W. Burggren, M. G. Monticino. Assessing physiological complexity. *J Exp Biol*. 208(17),3221-32 (2005).

Proposed Complexity Measures

(Time Series Based)

- Metric Entropy – measures number, and uniformity of distribution over observed patterns

  J. P. Crutchfield and N. H. Packard, Symbolic Dynamics of Noisy Chaos ,*Physica* ***7D*** (1983) 201.

- Statistical Complexity – measures number and uniformity of restrictions in correlation of observed patterns

  J. P. Crutchfield and K. Young, Inferring Statistical Complexity, *Phys Rev Lett* **63** (1989) 105.

- Excess Entropy – measures convergence rate of metric entropy

  D. P. Feldman and J. P. Crutchfield, Structural Information in Two-Dimensional Patterns: Entropy Convergence and Excess Entropy , Santa Fe Institute Working Paper 02-12-065

Proposed Complexity Measures

- **Statistical Complexity** is COMPLIMENTARY to **Kolmogorov Complexity**
- **Kolmogorov complexity** estimates complexity of algorithms – the shorter the program the less complex the algorithm
- “random” string “typically” can be generated by no short program so is “complex” in the Kolmogorov sense = entropy
- But randomness as complexity doesn’t jibe with visual assessment of images -> **Statistical Complexity**
- Yet another complimentary definition is standard **Computational Complexity** = run time

References

- J.P.W. Pluim, J.B.A. Maintz, M.A. Viergever, “Mutual Information Based Registration of Medical Images: A Survey”, IEEE Trans on Medical Imaging, Vol X No Y, 2003
- G.A. Tourassi, E.D. Frederick, M.K. Markey, and C.E. Floyd, “Application of the Mutual Information Criterion for Feature Selection in Computer-aided Diagnosis”, Medical Physics, Vol 28, No 12, Dec. 2001
- M.D. Esteban and D. Morales, “A Summary of Entropy Statistics”, Kybernetika. Vol. 31, N.4, pp. 337-346. (1995)