Overview of Mutual Information Registration

May 8, 2006

Motivation

- Registering images of different modalities.
- Intensity-based ("direct") methods, such as SSD and cross-correlation, assume a linear relationship between intensities.
- In using mutual information, we assume there is a functional relationship between intensities at the same location in different images. Since this functional is unknown, we resort to a generic measure on how much one image describes the other.

History and Papers

- Entropy and mutual information were introduced in Shannon's landmark 1940's paper that started the field of information theory.
- Maes, et al. at Leuven (Belgium) [3] and Wells, et al. at MIT / Brigham and Women's [6] introduced the idea of using MI for registration separately. Both of these papers should be read carefully. The first gives a good derivation of MI and describes the non-gradient based solution. The second describes a Parzen window (stochastic), gradient search, and includes an example illustrating why MI should work on multiple modalities.
- The paper by Studholme, et al. [5] contains a good motivation for the idea of mutual information and introduces "normalized mutual information".
- The book of Hajnal, Hill and Hawkes [2, Section 3.4.8] contains a similar, though less detailed, discussion and places MI in a historical context with other algorithms.

- See Duda, Hart and Stork [1, Appendix A.7] for a simplified, brief introduction to a few ideas in information theory.

Images, Intensities and Probabilities

- Consider the two images $J_A(\mathbf{x})$ and $J_B(\mathbf{x})$.
- Abusing notation (I am uncomfortable with the notation in the MI papers), we will use A and B to denote both the set of all possible intensities in the two images and to indicate the images themselves.
- We will think of intensities as samples from a random variable, which means each image forms a distribution (of intensities).
- Writing
$$h_A(a) =\mid \{\mathbf{x} | J_A(\mathbf{x}) = a\} \mid,$$
we get the histogram of intensities. Dividing by the number of pixels gives the empirical density:
$$p_A(a) = \frac{h_A(a)}{\sum_{a' \in A} h_A(a')}.$$
We do the same thing for B to obtain $p_B(b)$.
- We can form a joint histogram:
$$h_{A,B}(a,b) =\mid \{\mathbf{x} | J_A(\mathbf{x}) = a \text{ and } J_B(\mathbf{x}) = b\} \mid .$$
and compute the empirical joint density:
$$p_{A,B}(a,b) = \frac{h_{A,B}(a,b)}{\sum_{a' \in A, b' \in B} h_{A,B}(a',b')}.$$
- The intuition behind the idea of a joint density is important and it explains the idea of the linear relationship required for SSD and cross-correlation measures.
- The "marginal probabilities" can be recovered from the joint density:
$$p_A(a) = \sum_{b \in B} p_{A,B}(a,b) \qquad \text{and} \qquad p_B(b) = \sum_{a \in A} p_{A,B}(a,b).$$
Of course, these formulas hold for all densities, not just the ones described here.

- Conditional probabilites are

$$p_{A|B}(a|b) = \frac{p_{A,B}(a,b)}{p_B(b)} \qquad \text{and} \qquad p_{B|A}(b|a) = \frac{p_{A,B}(a,b)}{p_A(a)}$$

Entropy, Joint Entropy and Conditional Entropy

- The entropy of a distribution is the negative expected value of the log of the density:

$$H(A) = -\sum_{a \in A} p_A(a) \ln p_A(a).$$

- Entropy is always non-negative (because $-p \ln p$ is non-negative on the interval [0..1]).

- Entropy is maximized when p_A is uniform, and minimized when p_A is an impulse function. When p_A is a (discretized) Gaussian distribution, then $H(A)$ increases with increasing variance of the distribution.

- The joint entropy of two distributions is

$$H(A,B) = -\sum_{a \in A, b \in B} p_{A,B}(a,b) \ln p_{A,B}(a,b).$$

  Note that when p_A and p_B are independent, $H(A,B) = H(A) + H(B)$, whereas when p_A and p_B are perfectly correlated $H(A,B) = H(A) = H(B)$.

- The conditional entropy is

$$H(A|B) = -\sum_{a \in A, b \in B} p_{A,B}(a,b) \ln p_{A|B}(a|b).$$

  At first this is somewhat counter-intuitive, but the following point should make it clearer:

  - The sum is the expected value of $\ln p_{A|B}(a|b)$, just as in the other definitions of entropy. In fact, if we put $p_{A,B}(a,b)$ in each and sum over a and b, we'd get the same definitions.

  Intuitively, the conditional entropy is low when A is well-explained by B.

- Finally, note that

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

Mutual Information

- Defined in terms of entropy:

$$\begin{aligned} I(A,B) &= H(A) + H(B) - H(A,B) \\ &= \sum_{a,b} p_{A,B}(a,b) \ln \frac{p_{A,B}(a,b)}{p_A(a)p_B(b)} \\ &= H(A) - H(A|B) \\ &= H(B) - H(B|A) \end{aligned}$$

- Some properties:
  - $I(A,B) \geq 0$
  - If p_A and p_B are independent (bad, in this case) then $I(A,B) = 0$.
  - If p_A and p_B are perfectly correlated (good, in this case), then $I(A,B) = H(A) = H(B)$.
  - The second expression for $I(A,B)$ (the summation) is the Kullback-Leibler measure between two densities. In this case the densities are the joint density and what the joint density would be if the two distributions were independent.
- Intuitively, $I(A,B)$ is high when A is well-explained by B (B is well-explained by A).
- Finally, maximizing $I(A,B)$ is better than minimizing $H(A,B)$. In minimizing $H(A,B)$, all that is sought is a region of overlap between the images where there is low entropy. This could (and often is) the background region. Including $H(A)$ and $H(B)$, which increase with increasing complexity and variability in the image regions, forces the alignment into areas of both significant content as well as low joint entropy.

Mutual information as an alignment evaluation function

- Let A be the fixed image and B be the moving image.
- Let $T(B;\boldsymbol{\alpha})$ be the transformation function described by parameters $\boldsymbol{\alpha}$.

- Our goal is to find the parameters $\boldsymbol{\alpha}$ maximizing

$$I(A, T(B; \boldsymbol{\alpha})) = H(A) + H(T(B; \boldsymbol{\alpha})) - H(A, T(B; \boldsymbol{\alpha})).$$

- In order to evaluate this objective function, we must transform image B based on the parameters, re-compute the resulting histogram, densities, and entropies, and then re-evaluate.
- One subtlety is that $H(A)$ must be re-evaluated as the transformation changes because the region of overlap between the images will change.
- Even though there is no spatial information in the definition of mutual information, this works because the intensities are spatially-coherent

Algorithm 1: Non-derivative search [3]

- Powell's method, starting with searches in the directions of the individual rigid transformation parameters. Within plane parameters are manipulated first.
- Recompute marginal densities at all steps, including only the region of overlap between images, as above.
- Do NOT do trilinear histogramming. Instead, do partial-volume interpolation in the histogram. (This is justified both intuitively and empirically.)
- Expensive computation, slow convergence.

Aside: Parzen Windowing for Density Modeling

- Given a set of points $\{z_i\}$, what is the underlying density from which they are drawn?
- Could define an interval size, r, and then for any z, the density is the fraction of points such that $|z - z_i| < r/2$?
- This can be written in a functional form as

$$p(z) = \frac{1}{K} \sum_{i=1}^{K} R(z - z_i)$$

where

$$R(x) = \begin{cases} 1/r & |x| < r/2 \\ 0 & \text{otherwise} \end{cases}$$

- This idea can be extended further by replacing the above definition of R with a Gaussian, e.g.

$$R(\mathbf{x}) = G(\mathbf{x}; \Sigma) = (\sqrt{2\pi}|\Sigma|)^{-m} \exp(-1/2\mathbf{x}^T\Sigma^{-1}\mathbf{x})$$

  Here the covariance matrix Σ is crucial.

- The final importantant idea in Parzen windowing is replacing $\{z_i\}$ with a randomly sampled subset. As few as 50-100 samples are often used.

Algorithm 2: Density modeling through Parzen windows

- Parzen windows density:

$$p_A(a) = \frac{1}{N}\sum_i G(a - a_i; \Sigma)$$

  where a_i is the set of intensities of a randomly-chosen set of N points.

- A similar form holds for the joint density.

- Empirical expected value of entropy, using a second set of M randomly chosen points:

$$H(A) \approx -1\frac{1}{M}\sum_j \ln \sum_i G(a_j - a_i; \Sigma)$$

- For fixed sets $\{a_i\}$ and $\{a_j\}$ this is now a differentiable function.

- We can form the MI objective function

$$I(A, T(B; \boldsymbol{\alpha})) = H(A) + H(T(B; \boldsymbol{\alpha})) - H(A, T(B; \boldsymbol{\alpha}))$$

  using the sampling techniques described (sampling from A and B to compute the joint density), compute the derivative with respect to the parameters in $\boldsymbol{\alpha}$, and apply gradient descent.

Applications

- Originally, rigid (brain) registration: MR - CT, MR - PET, CT - PET. See [3] and [6]

- Free-form deformations: hierarchical splines. See [4]. This is very straightforward.

- 2d-3d. See the paper by Zollei, et al. [7]. When taking the derivatives with respect to pose parameters, they propagate them all of the way to the function used to form the DRR ("digitally reconstructed radiograph").

Normalized Mutual Information

- Normalized mutual information prevents an exceptional case where the images move toward extremely low overlaps [5].
- The objective function is

$$\tilde{I}(A, B) = \frac{H(A) + HB}{H(A, B)}.$$

- Like "regular" MI, this increases with increasing $H(A)$, increasing $H(B)$ and decreasing $H(A, B)$.
- This approach has been adopted by the algorithms that use Powell's method of minimization, but not (to my knowledge) by algorithms that use Parzen windowing

Other Discussion

- The Wells paper [6] provides an excellent illustrative example of entropy as an alignment measure.
- Is MI really an appropriate measure? Densities are being created where none exists. Yet, there is a functional relationship between intensities. How should this be captured generically since the function is unknown?

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2001.

[2] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes, editors. *Medical Image Registration*. CRC Press, 2001.

[3] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.

[4] D. Rueckert, I. Somoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18:712–721, 1999.

[5] C. Studholme, D. Hill, and D. Hawkes. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32:71–86, 1999.

[6] W. M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–51, 1996.

[7] L. Zollei, E. Grimson, A. Norbash, and W. Wells. 2d-3d rigid registration of x-ray fluoroscopy and CT images using mutual information and sparsely sampled histogram estimators. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 696–703, 2001.