NATO Kickoff Summary and Research Plan - Oct 2000

Alfred O. Hero, Yossi Francos, Olivier Michel

November 28, 2000

1 Overview

In early October Francos, Hero and Michel met for a day to discuss a research plan for the NATO award received in August 2000. After three overview presentations by each of the participants and ensuing discussions a picture emerged of a common vehicle for joint research. The objective is to apply Rényi entropy and divergence estimation methods, developed by Hero, Michel and co-workers [1, 2, 3], to the image indexing/segmentation algorithms based on the Wold decomposition, investigated by Francos and co-workers at INRIA Sophia-Antipolis [4, 5]. Many possible ideas were discussed, some of them quite ambitious, perhaps overly so for a one year contract. Three reasonable aims that came out of the discussions which could be accomplished in a one year timeframe are: 1) use a reduced rank representation for the texture field, e.g. derived from the ML estimates of the deterministic (evanescent component) and indeterminsitic (AR component) parameters of each image; 2) replacement of the KL divergence measure with the Rényi divergence measure; 3) study of methods to determine/estimate the most appropriate parameter α of the Rényi divergence measure. A more ambitious approach was discussed based on determining distances between sets of posterior distributions generated by a space of priors. This approach needs much more development in my opinion, and should be postponed until the next phase of the project (should there be one). To accomplish the above aims we will need access to the code generated by Francos and INRIA for extracting parameters of the model (either via least squares or max likelihood).

2 Linear Ordering of Image Models Using Rényi Divergence

Consider a reference image I_o and a database of images $\{I_k\}_{k=1}^p$ which are to be indexed relative to I_R. We assume that each image is realization I generated by a j.p.d.f. $f(I|\underline{\theta})$ which depends on a vector of unknown parameters $\underline{\theta}$ lying in a specified parameter space Θ. Under this probabilistic model the k-th observed image I_k is assumed to have been generated from model $f(I|\underline{\theta}_k)$, where $\underline{\theta}_k$ is called the "true parameter" underlying I_k, $k = 1, \ldots, p$. The "true indexing" of the images is induced by the rank ordering of the divergences $\{D_\alpha(\underline{\theta}_k \| \underline{\theta}_o)\}_{k=1}^p$ where

$$D_\alpha(\underline{\theta}_k \| \underline{\theta}_o) = \frac{1}{\alpha - 1} \ln \int f^\alpha(I|\underline{\theta}_k) f^{1-\alpha}(I|\underline{\theta}_o) \, dI \tag{1}$$

is the Rényi divergence between the image densities ($\alpha \in [0, 1]$). In the limit as $\alpha \to 1$ the Rényi divergence converges to the Kullback-Liebler (KL) divergence, also called the relative Shannon entropy:

$$D_1(\underline{\theta}_k \| \underline{\theta}_o) = \int f(I|\underline{\theta}_k) \ln f(I|\underline{\theta}_k)/f(I|\underline{\theta}_o) \, dI$$

and when $\alpha = \frac{1}{2}$ the Rényi divergence is monotonically related to the Hellinger distance [6]:

$$\begin{aligned} H(\underline{\theta}_k \| \underline{\theta}_o) &= \int (\sqrt{f(I|\underline{\theta}_k)} - \sqrt{f(I|\underline{\theta}_o)})^2 \, dI \\ &= 2\left(1 - \int \sqrt{f(I|\underline{\theta}_k) f(I|\underline{\theta}_o)} \, dI\right). \end{aligned}$$

Except in the case of $\alpha = \frac{1}{2}$ the Rényi divergence is not a true distance measure between image model pairs since it is not symmetric. However, we will not be the first to abuse terminology and refer to the divergence as a "distance" when convenient.

2.1 Rényi Divergence and the Error Exponent

The general Rényi divergence is the most appropriate measure for describing the difference between a pair of image models since it is directly related to the exponential rate of decay of the minimal achievable discrimination error probability. For two values $\underline{\theta}_o$ and $\underline{\theta}_1$ of the parameter consider the problem of discriminating between two image j.p.d.f's $f(I|\underline{\theta}_o)$ and $f(I|\underline{\theta}_1)$ based on a random sample of images $I_1, \ldots, I_n$ generated from one of these j.p.d.f.'s. Let the prior probabilities of $\underline{\theta}_o$ and $\underline{\theta}_1$ be p and $1-p$, respectively. Then the probability of decision error of the minimum probability of error test of $H_0 : \underline{\theta} = \underline{\theta}_o$ vs. $H_1 : \underline{\theta} = \underline{\theta}_1$ is

$$P_e(n) = \beta(n)p + \alpha(n)(1-p)$$

where $\beta(n)$ and $\alpha(n)$ are the respective probabilities of deciding $\underline{\theta}_0$ when $\underline{\theta}_1$ (Type II error) is true and vice versa (Type I error). Assuming $p \in (0,1)$ the Rényi divergence measure can be related to the probability of error through the Chernoff bound of large deviations theory [7, Corollary 3.4.6]:

$$\liminf_{n\to\infty} \frac{1}{n} \log P_e(n) = - \sup_{\alpha \in [0,1]} \{(1-\alpha) D_\alpha(\theta_1 \| \theta_o)\}. \tag{2}$$

This limit is known as the error exponent of the optimal Bayes decision rule, which can be shown to be equal to the Neyman Pearson likelihood ratio test (LRT) implemented with zero threshold. Thus the asymptotic rate of exponential decay of the minimum error probability is obtained by finding the value of α which maximizes the log integral in the definition (1) of the Rényi divergence.

Let α_o be the maximizer on the right hand side of (2):

$$\alpha_o = \mathrm{argmax}_{\alpha \in [0,1]} \int f^\alpha(I|\underline{\theta}_1) f^{1-\alpha}(I|\underline{\theta}_o) \, dI \tag{3}$$

The implication of (2) is that it is the Rényi divergence $D_{\alpha_o}(\theta_1 \| \theta_o)$, and not the KL divergence, which most closely characterizes the detectable differences between image models. This is the theoretical motivation of the general Rényi divergence measure for image indexing which is used below.

3 Some Possible Research Directions in Indexing

In practice the image model parameters $\underline{\theta}_k$'s are unknown so that the actual relative ordering of Rényi divergences $\{D_\alpha(\underline{\theta}_k \| \underline{\theta}_o)\}_{k=1}^p$ is also unknown. The problem of indexing can be stated as follows: based on a single realization I_k of each image j.p.d.f. $f(I|\underline{\theta}_k)$ estimate the actual rank ordering of Rényi divergences $\{D_\alpha(\underline{\theta}_k | \underline{\theta}_o)\}_{k=1}^p$.

3.1 Wold Decomposition Methods

In [4, 5] an indexing method was proposed using a multivariate Gaussian j.p.d.f. model and the Wold decomposition for each of the images. This model depends on several parameters $\alpha, \beta, \{\nu_i, s_i, t_i\}_i$ which specify the evanescent component, $\{C_i, D_i, \omega_i, \nu_i\}_i$ which specify the harmonic component, and $\sigma^2, \{b(k,l)\}_{kl}$ which specify the zero mean purely indeterministic component. In [4, 5] the estimated KL divergence was used to do the indexing. For the Gaussian case a closed form expression for KL divergence $D_1(\underline{\theta}_k|\underline{\theta}_o)$ is available in terms of the parameters $\underline{\theta}_k$ and $\underline{\theta}_o$. A simple substitute-and-plug estimate of the KL divergence was proposed in [4, 5] involving estimation of $\underline{\theta}_k$ from I_k and substituting this into the expression for the KL divergence.

These comments suggest the following extension of the indexing procedure of [4, 5]:

Proposition 1 *Reimplement the indexing procedure [4, 5] with the KL divergence measure replaced by the Rényi divergence measure. The Rényi divergence measure can also be computed in closed form for the Gaussian multivariate Gaussian j.p.d.f. model under the Wold decomposition. Issues to study would include:*

1. *Choice of parameter α. There are several possible approaches here:*
   (a) *Find MLE's $\hat{\underline{\theta}}_k$ of $\underline{\theta}_k$, $k = 0, \ldots, p$, and substitute them into Rényi divergence formula. Then compute the sequence of "optimal" pairwise Rényi parameters*

   $$\hat{\alpha}_k = \mathrm{argmax}_{\alpha \in [0,1]} \int f^{\alpha}(I|\hat{\underline{\theta}}_k) f^{1-\alpha}(I|\hat{\underline{\theta}}_o)\, dI, \quad k = 1, \ldots, p$$

   One could use the mean, the median, or some other function of the $\hat{\alpha}_k$'s as the "univeral" α to be used in the rank ordering of Rényi divergences $D_{\alpha}(\hat{\underline{\theta}}_k \| \hat{\underline{\theta}}_o)$. Another possibility would be to use $\alpha = \hat{\alpha}_{k_o}$ corresponding to the closest image:

   $$k_o = \mathrm{argmin}_k D_{\hat{\alpha}_k}(\hat{\underline{\theta}}_k \| \hat{\underline{\theta}}_o)$$

   which could be motivated by the fact that in indexing it is frequently the closest image which is of most interest.

   (b) *A plausible min-max alternative would be to select α as the solution to*

   $$max_{\alpha \in [0,1]} \min_k \int f^{\alpha}(I|\hat{\underline{\theta}}_k) f^{1-\alpha}(I|\hat{\underline{\theta}}_o)\, dI$$

   which has the nice theoretical justification of being the most conservative mechanism for choosing α. Applying saddlepoint theory in a similar manner to previous work [8, 9], I would guess that the solution to the above min-max problem would be identical to the solution α of

   $$\min_{w_1, \ldots, w_p} \max_{\alpha \in [0,1]} \int \sum_{k=1}^{p} w_k f^{\alpha}(I|\hat{\underline{\theta}}_k) f^{1-\alpha}(I|\hat{\underline{\theta}}_o)\, dI$$

   where $w_k \geq 0$ and $\sum_{k=1}^{p} w_k = 1$. The set of weights w_k that attain the min-max value above can be interpreted as the "least favorable priors" which make the error exponent as small as possible.

2. *Investigate computationally efficent ways to compute the Rényi divergence for the Wold Gaussian model. This should include a study of the relative difficulty of computing the divergence (number of non-linear operations, roundoff errors, number of flops) for different values of α. The prospect of indexing very large databases makes this an important issue.*

3. *Investigate reduced rank representations obtained by identifying a reduced parameter set which contains sufficient information to accurately separate classes of images. This could be studied empirically. It could also be studied by exploring (linear) transformations of the parameters which lead to exact or approximate block diagonal Fisher information matrices (as derived by Francos). This would be feasible by investigating the structure of the SVD of the Fisher matrix. These transformed parameters could then be estimated directly, by applying the transformation to the ML or LS estimates of the full parameter vector (derived by Francos). Better performance may be obtained by applying profile likelihood methods or by maximizing a reduced likelihood function that has been marginalized over the "nuisance parameters".*

4. *The issue of influence of MLE errors on indexing could be studied systematically by using the Fisher matrix analysis to find expectation variance of the MLE and then perform a perturbation analysis on the Rényi divergence of the form* $D_\alpha(\hat{\underline{\theta}}_k \| \hat{\underline{\theta}}_o) = D_\alpha(\underline{\theta}_k \| \underline{\theta}_o) + (\nabla_{\underline{\theta}_k, \underline{\theta}_o} D_\alpha)^T (\hat{\underline{\theta}}_k - \underline{\theta}_k, \hat{\underline{\theta}}_o - \underline{\theta}_o) + HOT$ *to find how variance in MLE's translates into indexing errors.*

3.2 More Ambitious Research Directions

Proposition 2 *Several other ideas that came up during our discussions were:*

1. *Development of methods for computing the Rényi divergence between image models directly via MST. This would avoid having to make model assumptions. This would require making a reduced rank approximation to the images, e.g. by expanding on some reduced basis (wavelets, Hough, SVD, KLT, etc) in order to keep computational complexity under control.*

2. *Introduction of a prior density* $f(\underline{\theta})$ *on the parameter vector would allow Bayes optimal or min-max indexing to be performed. This would be attractive since substitution of MLE's for the* $\underline{\theta}_o, \ldots, \underline{\theta}_p$ *in the rank ordering of Rényi information is suboptimal, especially when the MLE's are inaccurate. One problem with Bayes method is that they depend rather critically on the priors which are not easily specified. A way around this is to use minmax theory to specify the worst case priors but the analysis and implementation details involved would be complicated by the fact that multidimensional integration would have to be performed over the parameter space to identify least favorable densities. Furthermore, if the quality of the MLE's is sufficient there would not be much gained here. Therefore aforementioned study of how MLE error relates to indexing error would be a prerequisite to pursuing this line of inquiry.*

These directions are more ambitious than those of the previous section and should be postponed until later, in my opinion.

Comments?

References

[1] A. Hero and O. Michel, "Robust entropy estimation strategies based on edge weighted random graphs," in *Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE)*, San Diego, CA, July 1998.

[2] A. Hero and O. Michel, "Estimation of rényi information divergence via pruned minimal spanning trees," in *IEEE Workshop on Higher Order Statistics*, Caesaria, Israel, 1999.

[3] J. G. B. Ma, A. O. Hero and O. Michel, "Image registration with minimal spanning tree algorithm," in *IEEE Int. Conf. on Image Processing*, Vancouver, CA, October 2000.

[4] R. Stoica, J. Zerubia, and J. M. Francos, “The two-dimensional wold decomposition for segmentation and indexing in image libraries,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, Seattle, 1998.

[5] R. Stoica, J. Zerubia, and J. M. Francos, “Image retrieval and indexing: A hierarchical approach in computing the distance between textured images,” in *IEEE Int. Conf. on Image Processing*, Chicago, 1998.

[6] M. Basseville, “Distance measures for signal processing and pattern recognition,” *Signal Processing*, vol. 18, pp. 349–369, 1989.

[7] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Springer-Verlag, NY, 1998.

[8] B. Baygun and A. O. Hero, “Optimal simultaneous detection and estimation under a false alarm constraint,” *IEEE Trans. on Inform. Theory*, vol. 41, no. 3, pp. 688–703, 1995.

[9] A. Hero, R. Piramuthu, J. Fessler, and S. Titus, “Minimax emission computed tomography using high-resolution anatomical side information and b-spline models,” *IEEE Trans. on Inform. Theory*, vol. IT-45, no. 3, pp. 920–938, April 1998.