Imaging Applications of Stochastic Minimal Graphs

Alfred Hero[†], Bing Ma[#], and Olivier Michel[*]
Dept of EECS[†], The University of Michigan Ann Arbor, MI 48109-2122, USA
Intervideo Inc.[#], Fremont CA, USA
Dept. d'Astrophysique[*], Université de Nice Sophia-Antipolis, Nice, France
hero.eecs.umich.edu, bing@intervideo.com, omichel@unice.fr

Abstract— **This paper presents an overview of some of the recent theory and application of stochastic minimal graphs in the context of entropy estimation for imaging applications. Stochastic graphs which span a set of extracted image features can be constructed to yield consistent estimators of Jensen's entropy difference for between pairs of images. Unlike traditional plug-in entropy estimates based on density estimation, stochastic graph methods provide direct estimates of these quantities. We review the stochastic graph approach to entropy estimation, compare convergence rates to that of plug-in estimators, and discuss a geo-registration application.**

I. Introduction

Let $\mathcal{I}$ be a stochastic image and let feature vectors $Z^{(1)}, \ldots, Z^{(n)}$ be extracted from this image. We focus on the case that the feature vectors are i.i.d. realizations of a random variable Z generated by a feature density $f(Z)$. This is appropriate for piecewise homogeneous images from which repeated feature vectors can be sampled from a homogeneous region of the image. Examples of such a feature vector are: the position and orientation of an edge; a vector of samples in a textured region; the output vector of a spatial innovations filter; etc. This paper is concerned with estimating the joint α-entropy (see (1)) of the feature vector density based on feature samples extracted from the images.

Entropy estimation is of interest for pattern analysis, image complexity assessment, model identification, tests of independence, and other applications where invariance to scale, translation and other invertible transformations is desired in the discriminant. It was shown earlier [7] that minimal graphs such as the minimal spanning tree (MST) could be used to come up with direct estimates of α-entropy without requiring the difficult step of density estimation. This paper expands on this approach with special emphasis on imaging applications.

The results presented here can also be applied to indexing and content-based retrieval of images using entropic measures of distance between a query image having feature density f_0 and a database of images having feature densities $\{f_i\}$. For example the α-divergence $D_\alpha(f_1 \| f_0) = (\alpha - 1)^{-1} \ln \int f_1^\alpha(z) f_0^{1-\alpha}(z) dz$ converges to the Kullback-Liebler (KL) divergence as $\alpha \to 1$, which has been proposed for registration and indexing of images [10]. When f_0 is known the α-divergence can be directly estimated by minimal graph methods similar to those presented below using the measure transformation method outlined in [6]. However, for unknown f_0 and unknown f_1 the existence of consistent minimal-graph estimators of $D_\alpha(f_1 \| f_0)$ is an open problem. This paper will be concerned with an alternative dissimilarity function, called the α-Jensen difference, which is a function of the joint entropy of Z_0 and Z_1. As will be shown below, this function can be estimated using minimal graph entropy estimation techniques and behaves similarly to the α-divergence.

II. Entropy Estimation

Let Z be a feature vector in $\mathbf{R}^d$ with j.p.d.f $f(Z)$. Assume that f has bounded support. The α-entropy, also known as Rényi entropy is defined as

$$H_\alpha(f) = \frac{1}{1-\alpha} \ln \int_{\mathcal{Z}} f^\alpha(z) dz. \tag{1}$$

This entropy function converges to the Shannon entropy $-\int f(z) \ln f(z) dz$ as $\alpha \to 1$.

Most non-parametric entropy estimation techniques are based on estimation of the density function followed by substitution of these estimates into the entropy functional (1). For example, when this plug-in technique is applied to α-entropy it yields

$$H_\alpha(\hat{f}) = \frac{1}{\alpha - 1} \ln \int \hat{f}^\alpha(z) dz \tag{2}$$

where $\hat{f}$ is an empirical estimate of the density. For the special case of estimation of Shannon entropy recent non-parametric estimation proposals have included: histogram estimation plug-in kernel density estimation plug-in and sample-spacing density estimator plug-in. The reader is referred to [3] for a comprehensive overview of previous work in non-parametric estimation of Shannon entropy. The main difficulties with non-parametric methods are due to the infinite dimension of the spaces in which the unconstrained densities lie. Specifically: density estimator performance is poor without stringent smoothness conditions; no unbiased density estimators generally exist; density estimators have high variance and are sensitive to outliers; the high dimensional integration required to evaluate the entropy might be difficult.

The problems with the above methods can be summarized by the basic observation: on the one hand parameterizing the scalar entropy functional with an infinite dimensional density function is a costly over-parameterization, while on the other hand artificially enforcing lower dimensional density parametrizations can

[1] This work was supported in part by Veridian-ERIM International, a NATO Collaborative Linkage Grant, and AFOSR MURI Grant F49620-97-0028.

0-7803-6725-1/01/$10.00 ©2001 IEEE

produce significant bias in the estimates. This observation has motivated us to develop direct methods which accurately estimate the entropy without the need for performing artificial low dimensional parameterizations or non-parametric density estimation [5], [7], [6]. These methods are based on constructing minimal graphs spanning the feature vectors in the feature space. The overall length of these minimal graphs can be used to construct a strongly consistent estimator of entropy for Lebesgue continuous densities. In particular, let $\mathcal{Z}^{(n)} = \{Z^{(1)}, \ldots, Z^{(n)}\}$ and define

$$L_n = L(\mathcal{Z}^{(n)}) = \min_{e \in \mathcal{T}} \sum_e |e|^\gamma, \qquad (3)$$

the overall length of a graph spanning n i.i.d. vectors $Z^{(i)}$ in $\mathbf{R}^d$ each with density f. Here $\gamma \in (0, d)$ is real, e are edges in a graph connecting pairs of $Z^{(i)}$'s, $|e|$ denotes Euclidean (l_2) norm of the edge, and the minimization is over some suitable subsets $\mathcal{T}$, e.g. spanning trees, of the $\binom{n}{2}$ edges of the complete graph. Examples include the minimal spanning tree (MST), Steiner tree (ST), minimal matching bipartite graph, traveling salesman problem (TSP). The asymptotic behavior of L_n over random points $\mathcal{Z}^{(n)}$ has been studied for over half a decade [2], [11]. When the graph $\mathcal{T}$ is "quasi-additive" we showed in [7] that

$$\hat{H}_\alpha(\mathcal{Z}^{(n)}) = \ln L_n / n^\alpha - \ln \beta_{L,\gamma} \qquad (4)$$

is an asymptotically unbiased and almost surely consistent estimator of the un-normalized α-entropy of f where $\alpha = (d-\gamma)/d$ and $\beta_{L,\gamma}$ is a constant bias correction depending on the graph minimization criterion, e.g. MST, ST or TSP, but independent of f. Consistency (4) also holds when the power exponent function $|e|^\gamma$ in (3) is replaced by a positive function $g(|e|)$ which locally behaves as $|e|^\gamma$ as $|e| \to 0$ [11]. The fact that (4) holds for any quasi-additive graph construction opens the possibility of many different types of consistent graph-based entropy estimation algorithms. However, among the currently known quasi-additive algorithms the MST is the fastest (with polynomial run time) and as such it has been adopted for all of our entropy estimation applications.

Optimal pruning of these minimal graphs can robustify the entropy estimator against outliers from contaminating distributions. Divergence $D_\alpha(f_1 \| f_0)$ between the observed feature density f and a reference feature density f_0 can be estimated similarly via performing a preprocessing step before implementing the minimal-graph entropy estimator. This preprocessing step applies a measure transformation on the feature space which converts the reference density to a uniform density over the unit cube as explained in [6].

As contrasted with density estimation techniques of entropy estimation minimal graph entropy estimators enjoy the following properties: they have faster asymptotic convergence rates, especially for non-smooth densities and for low dimensional feature spaces; they completely bypass the complication of chosing and fine tuning parameters such as histogram bin size, density kernel width, complexity, and adaptation speed; the α parameter in the α-entropy function is varied by varying the interpoint distance measure used to compute the weight of the minimal graph. On the other hand, the need for combinatorial optimization is a bottleneck for large number of feature samples. This has motivated the development of greedy minimal graph approximations that preserve advantages such as robustness against outliers [7].

III. Entropy Estimator Convergence Comparisons

Here we compare asymptotic convergence rates of the direct minimal-graph entropy estimator (4) and the indirect density plug-in entropy estimator (2) as a function of the number n of i.i.d. samples of Z. For proofs of the following propositions see [4]. Let $Z \in \mathbf{R}^d$ have joint Lebesgue density f. Define the class of Hölder continuous functions $\Sigma_d(\kappa, c)$ over $\mathbf{R}^d$

$$\Sigma_d(\kappa, c) = \left\{ f(x) : \|f(x) - p_x^{\lfloor \kappa \rfloor}(z)\| \leq c \, \|x - z\|^\kappa \right\}$$

where $p_x^k(z)$ is the Taylor polynomial of f of order k expanded about the point x. As κ becomes large the class $\Sigma_d(\kappa, c)$ contains functions which are increasingly non-smooth.

For the indirect estimator (2) it makes sense to consider a minimax optimal density estimation strategy which minimizes the worst case estimator mean integrated square error (MISE) over the densities lying in $\Sigma_d(\kappa, c)$ [8]. The minimax estimator can be implemented as a piecewise polynomial with bin size that decreases in n at a specified optimal rate. The resultant MISE has the fastest possible rate of convergence over all $\Sigma_d(\kappa, c)$ and the rates of convergence of the squared bias and the variance are identical.

Proposition 1: Assume that the Lebesgue density f is supported on the unit d-dimensional cube $[0, 1]^d$, $f \in \Sigma_d(\kappa, c)$ and that $\int f^{\alpha-1}(z) dz < \infty$. Then, if $\hat{f}$ is a minimax MISE density estimator

$$\sup_{f \in \Sigma_d(\kappa, c)} \left| E[H_\alpha(\hat{f})] - H_\alpha(f) \right| = n^{-\kappa/(2\kappa+d)} \, C_{\kappa,c}(1 + o(1))$$

where $C_{\kappa,c}$ is a constant dependent on f.

While not needed for the comparison performed below, it follows from Appendix A that the worst case estimator MSE for minimax estimation $\sup_{f \in \Sigma_d(\kappa, c)} \sqrt{E\left[H_\alpha(\hat{f}) - H_\alpha(f)\right]^2}$ also has rate of convergence $n^{-\kappa/(2\kappa+d)}$.

For the direct minimal-graph estimator (4) convergence rates are more difficult to establish. The convergence of quasi-additive minimal graphs has been studied for a large number of problems including minimal spanning trees, Steiner trees, and the traveling salesman problem [11] The following specifoes the convergence rate of such estimators

Proposition 2: Assume that the Lebesgue density f over $[0, 1]^d$ satisfies the property that f^ν is of bounded variation for all $\nu \in (0, 1)$ and that $\int f^{-1/d}(x) dx < \infty$. Then for $d \geq 2$

$$\left| E[\hat{H}_\alpha(\mathcal{Z}^{(n)})] - H_\alpha(f) \right| = n^{-1/d} \, K_{L,\gamma}(1 + o(1))$$

where $K_{L,\gamma}(f)$ is a constant depending on f.

Observe that as compared to Proposition 1 the convergence rate in Proposition 2 only depends on the weaker condition of bounded variation of f. A comparison between the convergence rates of the two propositions indicates that the direct estimator converges with faster asymptotic rate in n when:

$$\kappa < \frac{d}{d-2}.$$

Thus the direct estimator always has faster convergence than the indirect estimator when $d = 2$, i.e. the feature vector Z lies in the plane, or when $d > 2$ but κ is large, i.e. f is non-smooth.

A. Estimation of α-Jensen Difference

Let f_0 and f_1 be two densities and $\beta \in [0,1]$ be a mixture parameter. The α-Jensen difference is the difference between the α-entropies of the mixture $f = \beta f_0 + (1-\beta) f_1$ and the mixture of the α-entropies of f_0 and f_1 [1]:

$$\begin{aligned}\Delta H_\alpha(\beta, f_0, f_1) &\stackrel{\Delta}{=} \qquad\qquad (5)\\ &H_\alpha(\beta f_0 + (1-\beta) f_1) - [\beta H_\alpha(f_0) + (1-\beta) H_\alpha(f_1)],\end{aligned}$$

For $\alpha \in [0,1]$ the α-Jensen difference is a measure of dissimilarity between f_0 and f_1: as the α-entropy $H_\alpha(f)$ is concave in f it is clear from Jensen's inequality that $\Delta H_\alpha(\beta, f_0, f_1) \geq 0$ with equality iff $f_0 = f_1$ a.e.

The α-Jensen difference can be motivated as an index function for content-based retrieval and image registration as follows. Assume that two sets of labeled feature vectors $\mathcal{Z}_0 = \{Z_0^{(i)}\}_{i=1,\ldots,n_0}$ and $\mathcal{Z}_1 = \{Z_1^{(i)}\}_{i=1,\ldots,n_1}$ are extracted from images $\mathcal{I}_0$ and $\mathcal{I}_1$, respectively. Assume that each of these sets consist of independent realizations from densities f_0 and f_1, respectively. Define the union $\mathcal{Z} = \mathcal{Z}_0 \cup \mathcal{Z}_1$ containing $n = n_0 + n_1$ unlabeled feature vectors. Any consistent entropy estimator constructed on the unlabeled $Z^{(i)}$'s will converge to $H_\alpha(\beta f_0 + (1-\beta) f_1)$ as $n \to \infty$ where $\beta = \lim_{n\to\infty} n_0/n$. This motivates the following consistent minimal-graph estimator of Jensen difference (5) for $\beta = n_0/n$:

$$\begin{aligned}\widehat{\Delta H}_\alpha(\beta, f_0, f_1) &\stackrel{\Delta}{=} \qquad\qquad (6)\\ &\hat{H}_\alpha(\mathcal{Z}_0 \cup \mathcal{Z}_1) - \left[\beta \hat{H}_\alpha(\mathcal{Z}_0) + (1-\beta)\hat{H}_\alpha(\mathcal{Z}_1)\right],\end{aligned}$$

where $\hat{H}_\alpha(\mathcal{Z}_0 \cup \mathcal{Z}_1)$ is the minimal-graph entropy estimator (4) constructed on the n point union of both sets of feature vectors and $\hat{H}_\alpha(\mathcal{Z}_0)$, $\hat{H}_\alpha(\mathcal{Z}_1)$ are constructed on the individual sets of n_0 and n_1 feature vectors, respectively. We can similarly define the density-based estimator of Jensen difference based on entropy estimates of the form (2) constructed on $\mathcal{Z}_0 \cup \mathcal{Z}_1$, $\mathcal{Z}_0$ and $\mathcal{Z}_1$.

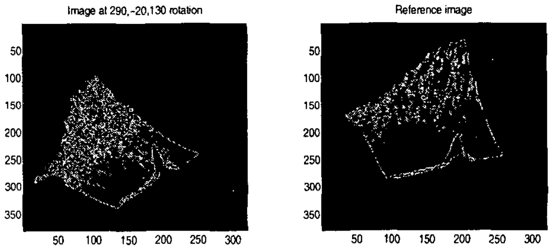
For some indexing problems the marginal entropies $\{H_\alpha(f_i)\}_{i=1}^K$ over the database are all identical so that the indexing function $\{H_\alpha(\beta f_0 + (1-\beta) f_i)\}_{i=1}^K$ is equivalent to $\{\Delta H_\alpha(\beta, f_0, f_i)\}_{i=1}^K$. The problem of registering a query image to a database of images which are generated by rigid transformations of a reference image is an important example of this simplifying situation.

IV. Geo-Registration Application

The objective is to register two types of images — a set of electro-optical (EO) images and a terrain height map. For this multisensor image registration problem, there usually exist distortions between the two types of images. The distortions are due to difference acquisition conditions of the images such as shadowing, diffraction, terrain changes over time, clouds blocking the illumination sources, seasonal variations, etc. Existence of such differences between the images to be registered requires that the registration algorithms to be robust to noise and other small perturbations in intensity values.

For this image registration problem the set of EO images are generated from the *a priori* digital elevation model (DEM) of a terrain patch (the terrain height map) at different look angles (determined by the sensor's location) and with different lighting positions.

Geo-registration of a EO reference image to DEM's in an image database is accomplished by selecting a candidate DEM image from the database and projecting it into the EO image plane of the reference image. The objective is to find the correct viewing angle such that the corresponding EO image is the best match to the EO reference image. Figure 1 shows an DEM projected into the EO image plane with viewing angles (290, -20, 130) and the reference EO image. Clearly they are not aligned.



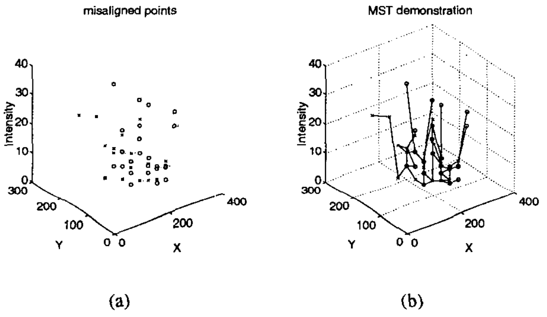
(a) (b)

Fig. 1. Misaligned EO and reference images

For matching criterion we implemented the α-Jensen difference applied to grey level features extracted from the reference images and candidate EO images derived from the DEM database. The parameter α was chosen arbitrarily as 0.5, corresponding to a MST construction minimizing the Euclidean norm in (3) without any power weighting ($\gamma = 1$). For illustration purposes we selected a very simple set of features via stratified sampling of the grey levels with centroid refinements. This sampling method produces a set of n three dimensional feature vectors $Z_i = (x_i, y_i, F(x_i, y_i))$ where $F(x, y)$ is a sample of the grey level at planar position x, y and where n is fixed in advance. The points $\{(x_i, y_i)\}_{i=1}^n$ approximate the centroids of Voronoi cells and $\{F(x_i, y_i)\}_{i=1}^{n}$ correspond to the set of n samples of the image from which we could reconstruct the original image with minimum mean square error. For more details see [9]. When the union of features from reference and target images are rendered as points in three dimensions we obtain a point cloud of features over which the MST can be constructed and the Jensen difference estimated. Since $n_1 = n_0 = n$ we have used $\beta = 1/2$ in the Jensen difference (6).

Figure 2 illustrates the MST-based registration procedure over the union of the reference and candidate image features for misaligned images, while Figure 3 shows the same for aligned images. From Figures 2(a) and 3(a) we see that for misaligned images, the representation points "x" and "o" are at larger distances, giving corresponding larger MST weight, than those for aligned images.

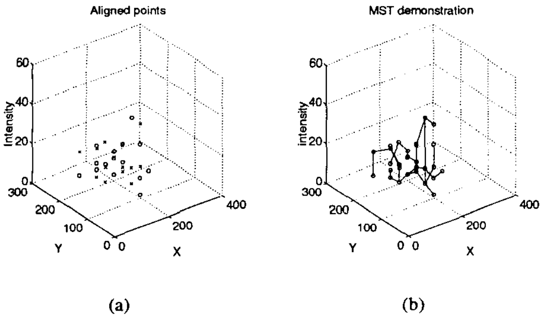
We repeat this MST construction process over the union of reference features and features derived from each of the images in the DEM database. The MST length can then be plotted in Figure 4. The x-axis stands for the image index, which corresponds to the viewing angles from the aircraft. The minimum MST length indicates the best matching of the EO image and the reference image, which corresponds to the registered pair in Figure 5.



(a)

(b)

Fig. 2. MST demonstration for misaligned images



(a)

(b)

Fig. 3. MST demonstration for aligned images. "x" denotes reference while "o" denotes a candidate image in the DEM database.

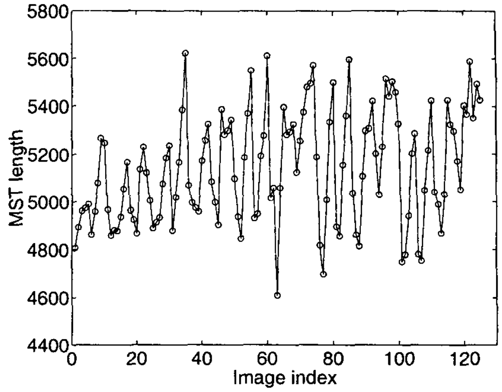
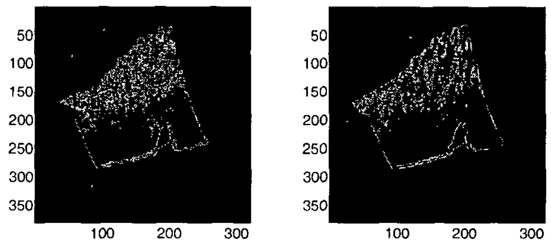


Fig. 4. MST length for different test-reference image pairs



(a) Matching image

(b) Reference image

Fig. 5. Co-registered EO-terrain maps

References

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

[2] J. Beardwood, J. H. Halton, and J. M. Hammersley, "The shortest path through many points," *Proc. Cambridge Philosophical Society*, vol. 55, pp. 299–327, 1959.

[3] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. van der Meulen, "Nonparametric entropy estimation: an overview," *Intern. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, june 1997.

[4] A. O. Hero, B. Ma, and O. Michel, "Alpha-divergence for image indexing and retrieval," Technical Report 400, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, Jan 2001. http://www.eecs.umich.edu/~hero/image_proc.html.

[5] 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.

[6] 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.

[7] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. IT-45, no. 6, pp. 1921–1939, Sept. 1999.

[8] A. P. Korostelev and A. B. Tsybakov, *Minimax theory of image reconstruction*, Springer-Verlag, New York, 1993.

[9] B. Ma, *Parametric and non-parametric approaches for multisensor data fusion*, PhD thesis, University of Michigan, Ann Arbor, MI 48109, 2001.

[10] J. A. O'Sullivan, "Divergence penalty for image registration," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, 1994.

[11] J. M. Steele, *Probability theory and combinatorial optimization*, volume 69 of *CBMF-NSF regional conferences in applied mathematics*, Society for Industrial and Applied Mathematics (SIAM), 1997.