Structured Data : Learning, Prediction, Dependency, Testing

Paper Review : High dimensional mutual information estimation for image registration

Alexandre BRUCKERT
Elliot HOFMAN
Gregoire MAGENDIE

Contents

I - Problem description

Image registration is a common signal processing problem, whose applications can be used to solve a lot of tasks : automatic registration in medical imaging, movement detection in videos, morphing generation with images, ... Intuitively, the problem can be seen as defining and maximizing a similarity metric between two images. The paper focuses on this approach, proposes a new metric and compares with the other existing strategies.

Traditional methods used for image registration consist in the estimation of the mutual information between the two images. Given some images X and Y, the idea is to compute $I(X, Y) = H(X) + H(Y) - H(X, Y)$, where H is the differential entropy $H(X) = -\int p_X(\xi) \log p_X(\xi) d\xi$. Existing algorithms estimate those quantities with histogram methods, thus implying that the data is first set into bins to estimate the frequency probabilities in the above formula. If those methods work well when working in small dimension (typically $d = 1, 2$), they become quite prohibitive as the dimension of the data grows up, because the number of bins needed to separate the samples increases exponentially. Some solutions have been proposed to tackle this issue, but all of them remain very costly when $d > 2$ ($\mathcal{O}(dN^2)$).

As a result, the mutual information estimation is usually made upon 1-dimensional samples (scalars), pixel gray-scale intensity for example. The intuition of the author of the paper Ḧigh dimensional mutual information estimation for image registrationï̈s that some performance would be gained if one could incorporate more features in the samples (the three RGB intensities for instance). To be able to efficiently compute the mutual information, the author presents a new entropy estimator, based on the nearest neighbor entropy estimator of Kozachenko and Leonenko. He modifies the algorithm so that the new version is numerically robust and computationally efficient. The asymptotic complexity $\mathcal{O}(dN_{\text{pixel}})$ allows for mutual information estimation with high dimensional data.

A remarkable benefit from this approach is that it makes image registration able to be computed with complex feature maps rather than just 1 or 2 attributes per pixel. The results show that the new entropy estimator is effective in determining the correct alignment even in difficult cases in which the classical criteria fails.

Before we dive deeper into the theory, let's sum up the problem. Formally, given two images F and G and a family of geometrical transformations (such as translations, rotations, affine transformations, ...), the purpose is to find the transformation T such that $T(G)$ is as close as possible to F, according to some image similarity criterion.

II - Proposed approach

The mutual information is computed with a modified version of the Kozachenko and Leonenko (KL) entropy estimator, which relies on nearest neighbors between the samples $F = \{f_i\}_{i=1}^N$:

$$H_{\mathrm{KL}}^{(N)}(F) = \frac{d}{N}\sum_{i=1}^{N}\log(q_i) + \log\Big(\frac{(N-1)\pi^{d/2}}{\Gamma(1+\frac{d}{2})}\Big) + \gamma$$

where $q_i = \min_{j\neq i} ||f_i - f_j||$ and $\gamma \approx 0.577$ is the Euler-Mascheroni constant.

We present in the following sections the justifications to obtain this formula, and how the entropy estimator from the paper is derived from this KL estimator.

(a) - Kozachenko and Leonenko entropy estimator

Let X be the vector of the samples drawn from the image, $X = [x_1, \dots, x_N]$ with $x_i \in \mathbb{R}^d$ (for example, x_i is the 3-dimensional vector of the RGB values at pixel i). Denoting by μ the (unknown) density of X, the Shannon entropy is defined as :

$$H(X) = -\int \mu(x)\log\mu(x)dx$$

Suppose we have an unbiased estimator $\widehat{\log\mu(x)}$, then

$$\widehat{H}(X) = -\frac{1}{N}\sum_{i=1}^{N}\widehat{\log\mu(x_i)}$$

would be an unbiased estimator of $H(X)$.

In order to estimate $\widehat{\log\mu(x_i)}$, we consider the probability distribution $P_k(\epsilon)$ for the distance between x_i and its k-th nearest neighbor. The quantity $P_k(\epsilon)d\epsilon$ is equal to the chance that there is one point within distance $r \in [\frac{\epsilon}{2}, \frac{\epsilon}{2} + \frac{d\epsilon}{2}]$ from x_i, that there are $k-1$ other points at smaller distances, and that the remaining $N-k-1$ points have larger distances from x_i. Let $p_i(\epsilon) = \int_{||x-x_i||<\epsilon/2}\mu(x)dx$ be the the mass of the ϵ-ball centered at x_i. Using the above statement and the trinomial formula, we can calculate :

$$P_k(\epsilon)d\epsilon = \frac{(N-1)!}{1!(k-1)!(N-k-1)!} \times \frac{dp_i(\epsilon)}{d\epsilon}\epsilon \times p_i^{k-1} \times (1-p_i)^{N-k-1}$$

which allows to obtain $\mathbb{E}[\log p_i] = \int_0^\infty \log p_i(\epsilon)P_k(\epsilon)d\epsilon = \psi(k) - \psi(N)$,
where $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)}$ is the digamma function.

If we assume that μ is constant on the ϵ-ball, we can set $p_i(\epsilon) \approx c_d \epsilon^d \mu(x_i)$ with c_d the volume of the d-dimensional unit ball. Using those two equations together gives $\log \mu(x_i) \approx \psi(k) - \psi(N) - d\mathbb{E}[\log \epsilon] - \log c_d$, which finally leads to the KL estimator :

$$\widehat{H}(X) = \psi(N) - \psi(k) + \log c_d + \frac{d}{N} \sum_{i=1}^{N} \log \epsilon(i)$$

where $\epsilon(i)$ is the distance from x_i to its k-th nearest neighbor.

The formula for $H_{\mathrm{KL}}^{(N)}(F)$ is obtained with $k = 1$, taking into account that for the ℓ_2 norm $c_d = \frac{\pi^{d/2}}{\Gamma(1+\frac{d}{2})}$ and that $\psi(N) = H_{N-1} \sim \log(N-1)$, $\psi(1) = -\gamma$.

In a nutshell, the KL estimator's main idea is to estimate the density μ of the samples by computing the k-th nearest distances rather than computing the frequencies of all the samples (which is what the histogram-based methods do). The only step which is costly in this algorithm is the computation of the nearest neighbors, which can be done in $\mathcal{O}(dN^2)$ time using a brute force method or in $\mathcal{O}(dN \log N)$ using a more optimized solution. This is still faster than an histogram method, and the bias (caused by the assumption that the density μ is locally constant on the ϵ-balls) decreases as $\mathcal{O}(\frac{1}{\sqrt{N}})$

(b) - Degenerancies and Batch-KL estimator

A drawback of the previous KL estimator is that the formula will diverge if the data contains two identical samples due to the 0 in the log. A simple way to deal with such degenerancies is to add some random noise with very low magnitude to the data as a preprocessing. The author of the paper proposes an other solution with a threshold :

$$H_{\mathrm{KLD}}^{(N)}(F) = \frac{d}{N} \sum_{i=1}^{N} \log(\max(\epsilon, q_i)) + \log\Big(\frac{(N-1)\pi^{d/2}}{\Gamma(1+\frac{d}{2})}\Big) + \gamma$$

This will ensure that the KL estimator can never diverge.

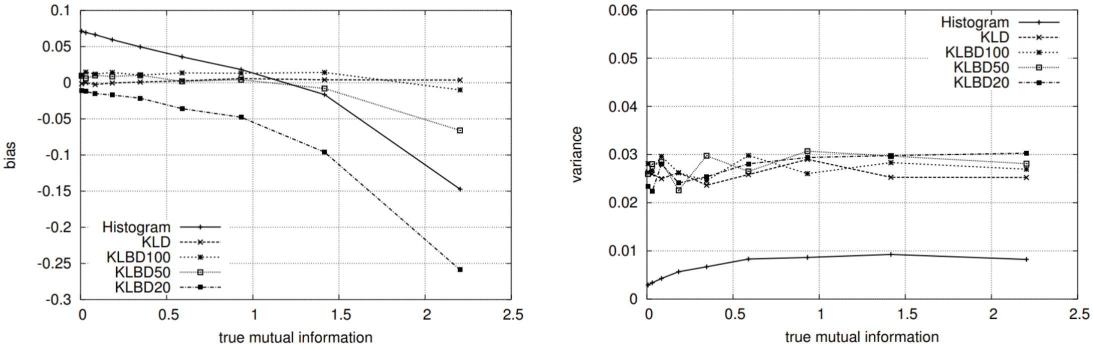
The time complexity (due to the nearest distances computation) is still too high and prohibitive for the image registration tasks that the author considers (images of dimensions 600*600 lead to $N = 360\ 000$ samples, and feature maps could increase the dimensionality of the samples up to $d = 100$). To sidestep this issue, he proposes to randomly divide the N samples into groups $F_1, \ldots, F_{\lfloor N/M \rfloor}$ so that each group contains M elements. This leads to the final entropy estimator :

$$H_{\mathrm{KLBD}}^{N}(F) = \frac{1}{\lfloor N/M \rfloor} \sum_{i=1}^{\lfloor N/M \rfloor} H_{\mathrm{KLD}}^{(M)}(F_i)$$

With this approach, assuming the brute force strategy to compute the nearest distances, the time complexity will grow as $\mathcal{O}(\lfloor N/M \rfloor dM^2)$, that is $\mathcal{O}(dNM)$ (and $\mathcal{O}(dN \log M)$ with a more optimal nearest neighbors approach such as BallTree or KDTree). The batch size M controls the trade-off between the estimator's bias $\mathcal{O}(\frac{1}{\sqrt{M}})$ and its speed $\mathcal{O}(dNM)$.

III - Discussion

Before he simulates some image registration testing tasks, the author of the paper evaluates the goodness of the KL estimator to check his theoretical (asymptotic) consistency and verify that the estimation is faster than the existing histogram methods. Two 1D Gaussian random variables are generated with varying level of correlation, and the estimators Histogram, KLD and KLBD-M with different values of M are compared.



All the KL estimators seem to have roughly the same variance, which is higher than the variance of the Histogram estimator but still quite low. With $M \geqslant 50$ the bias of KLBD-M is already lower than the Histogram estimator bias, and the estimator KLD almost reaches a bias of 0. A good compromise between the bias and the time complexity has been evaluated to be $M = 20$. The author shows the execution time for different values of N and d, M being fixed to 20.

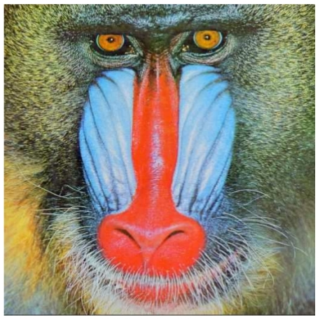
d	$N = 10^3$	$N = 10^4$	$N = 10^5$
1	0.004	0.04	0.51
3	0.005	0.06	0.67
10	0.010	0.13	1.40
30	0.023	0.29	3.51
100	0.078	0.86	10.69

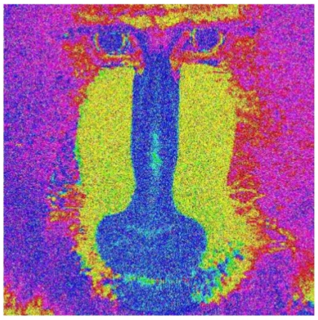
For image registration performance evaluation, two scenarios are considered. In the first example, a 512 × 512 RGB image of a mandrill is modified (increased saturation and brightness, inverted colormap and some iid. Gaussian noise is added) and rotated from -10 to 10 degrees. At each degree several similarity metrics are calculated between the modified image and its original version, and scaled into [0, 1]. A good criterion should have its maximum for a rotation angle 0 and decrease smoothly away from it.

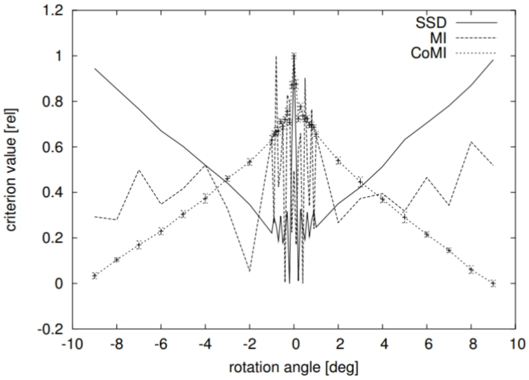
The SSD metric corresponds to the negative sum of squared differences computed on gray-scale pixel intensities ($J_{\mathrm{SSD}}(F,G) = -\sum_{i \in \mathrm{pixels}} ||f(x_i) - g(x_i)||^2$).

The MI metric corresponds to the mutual information computed on the frequency probabilities of the gray-scale pixel intensities ($J_{\mathrm{MI}}(F,G) = H(F) + H(G) - H(F,G)$, with $H(F) = -\sum_{i \in \mathrm{set(samples)}} p_i \log p_i$).

The CoMI metric corresponds to the KLBD-M estimation computed on the pixels with the 3-dimensional feature map $f_i^{\mathrm{Co}} = (f_R(x_i), f_G(x_i), f_B(x_i))$ (thus making use of the three channels of the image, unlike the two other metrics).

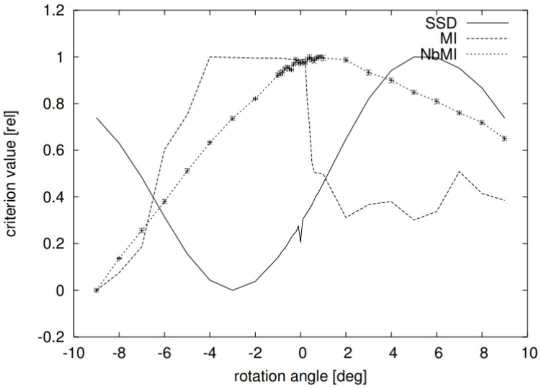






The only metric that constitutes a good criterion is the CoMI estimator, the SSD and MI metrics get confused by the noisy and color-modified version.

In the second example, a 512×512 grayscale image of Lena is used, and the metrics are compared in order to chose a good criterion to register a blurred version of Lena with a blurred version of its edges. This time, NbMI corresponds to the KLBD-M estimation computed on the pixels with the 25-dimensional feature map $f_i^{\mathrm{Nb}} = (f(x - \Delta_x, y - \Delta_y))_{|\Delta_x| \leqslant h, |\Delta_y| \leqslant h}$, $h = 2$.



Once again, the only metric that is able to perform image registration is the NbMI estimator.

The proposed KLBD estimator is a fast and robust version of the binless KL entropy estimator. The ability to estimate the mutual information between two signals although their samples live in a higher-dimensionality feature space allows to define better criterions based on several features rather than just 1 or 2 attributes. The two examples introduced a 3D color based feature map and a 25D neighborhood feature map : mutual information could not have been calculated for such features with the traditional methods.