HIGH-DIMENSIONAL MUTUAL INFORMATION ESTIMATION FOR IMAGE REGISTRATION

Jan Kybic

Center for Machine Perception, CTU, Prague, Czech Republic
kybic@fel.cvut.cz

ABSTRACT

We present a new algorithm for mutual information estimation for image registration based on the nearest neighbor entropy estimator of Kozachenko and Leonenko. We modify the algorithm to be numerically robust and computationally efficient, with optimal asymptotic complexity $O(N_{\text{pixels}} d_{\text{dim}})$. We propose two MI-based criteria exploiting the high-dimensionality of the feature space and show their effectiveness in determining the correct alignment even in difficult cases when classical criteria fail.

1. INTRODUCTION

Mutual information (MI) [1, 2] is then the registration criterion of choice for multimodal image registration, for its ability to explore the statistical dependency between images not known *a priori*. However, the predominantly used histogram-based estimator makes the MI criterion usable only for features of very small dimensionality d, in most cases scalar pixel intensities are used.

We shall present an alternative entropy estimator, sufficiently accurate and computationally efficient even for much higher feature dimensions d, making feasible to use more powerful higher dimensionality features. Two examples of such high-dimensional MI criteria will be demonstrated, one based on color and other on pixel neighborhoods.

2. IMAGE REGISTRATION

Given two images $\mathcal{F}$ and $\mathcal{G}'$, and a family of geometrical transformations[1], we search for a geometrical transformation T such that a warped test image $\mathcal{G} = \mathcal{G}' \circ \mathsf{T}$ is as similar as possible to a reference image $\mathcal{F}$ in the sense of maximizing an image similarity criterion J.

[1] such as translations, affine transformations, or general warpings

2.1. SSD and MI criteria

One of the simplest criteria is the (negative) sum of the squared differences (SSD):

$$J_{\text{SSD}}(\mathcal{F}, \mathcal{G}) = -\sum_{i \in \text{pixels}} \|\mathbf{f}(\mathbf{x}_i) - \mathbf{g}(\mathbf{x}_i)\|^2 \quad (1)$$

where $\mathbf{f}(\mathbf{x}_i)$ and $\mathbf{g}(\mathbf{x}_i)$ are intensity values of images $\mathcal{F}$ and $\mathcal{G} \circ \mathsf{T}$ at pixel locations $\mathbf{x}_i$. They can be scalars or vectors, e.g. for color images. The SSD criterion is simple, fast, and optimal for i.i.d. Gaussian noise corruption.

If the dependence between the image intensities is *a priori* unknown, MI criterion is normally used:

$$J_{\text{MI}}(\mathcal{F}, \mathcal{G}) = I(\mathsf{F}, \mathsf{G}) = H(\mathsf{F}) + H(\mathsf{G}) - H(\mathsf{F}, \mathsf{G}) \quad (2)$$

where $H()$ stands for a differential entropy[2] of random variables F, G; pixel values $f_i = f(\mathbf{x}_i)$, $g_i = g(\mathbf{x}_i)$ are assumed to be realizations of F, G: $f_i \sim \mathsf{F}$, $g_i \sim \mathsf{G}$, $(f_i, g_i) \sim (\mathsf{F}, \mathsf{G})$.

2.2. Higher dimensionality MI criteria

We propose to use more general and more powerful d-dimensional feature vectors instead of using just simple scalar pixel values ($d = 1$) as in (2). Each sample vector $\mathbf{f}_i$ will correspond to one spatial location $\mathbf{x}_i$. Let us present two examples:

- The 3D color component vector (for color images)

$$\mathbf{f}_i^{\text{Co}} = (f^R(\mathbf{x}_i), f^G(\mathbf{x}_i), f^B(\mathbf{x}_i)) \quad (3)$$

leads to a color MI *(CoMI)* criterion $J_{\text{CoMI}} = I(\mathsf{F}^{\text{Co}}, \mathsf{G}^{\text{Co}})$. It adapts automatically to any relation of the image colors.

- The neighborhood criterion (presented here in the 2D case and for grey-level images) $J_{\text{NbMI}} = I(\mathsf{F}^{\text{Nb}}, \mathsf{G}^{\text{Nb}})$, forms feature vectors of dimension $d = (2h + 1)^2$ from pixel values in the neighborhood of a current location:

$$\mathbf{f}_i^{\text{Nb}} = \left(f(x - \Delta_x, y - \Delta_y) \right)_{|\Delta_x| \leq h, \, |\Delta_y| \leq h} \quad (4)$$

This criterion learns correspondences between image details such as peaks, ridges and transitions, generalizing the work [3] (that only uses $d = 2$) by adding spatial relationships.

[2] $H(\mathsf{F}) = -\int p_{\mathsf{F}}(\xi) \log p_{\mathsf{F}}(\xi) \, \mathrm{d}\xi$

3. ENTROPY ESTIMATION

Mutual information criteria (2,3,4) are evaluated using a suitable entropy estimator. It needs to be fast, of at most weakly superlinear complexity in both N and d, as we might want to evaluate it for 100-dimensional features and images with 10^7 pixels at each iteration of the optimizer.

3.1. Histogram and graph estimators

A histogram estimator works well in small dimensions ($d = 1, 2$). For higher d however, the number of bins increases exponentially, raising the computational costs, causing quantization errors and increasing variance, making the estimator unusable [4]. Parzen windowing helps but is excessively costly for higher d.

Area of Voronoi diagram cells [5], minimum spanning tree length [6], or pairwise interaction through a kernel estimator [7] can be used to estimate Shannon or Rényi entropy; all approaches are slow for $d > 2$, with $O(dN^2)$ at best.

4. PROPOSED MI ESTIMATOR

Kozachenko and Leonenko (KL) [8, 4, 9] proposed to estimate the entropy $H(\mathsf{F})$ from N samples $\mathbf{F} = \{\mathbf{f}_i\}_{i=1}^N$ using distances $\varrho_i = \min_{j \neq i} \|\mathbf{f}_i - \mathbf{f}_j\|$ from samples $\mathbf{f}_i$ to their nearest neighbors (NN):

$$H_{\mathsf{KL}}^{(N)}(\mathbf{F}) = \frac{d}{N} \sum_{i=0}^N \log \varrho_i + \log \frac{(N-1)\pi^{d/2}}{\Gamma(1+d/2)} + \gamma \quad (5)$$

where $\gamma \approx 0.577$ is the Euler constant. The estimator is consistent with bias $\langle H_{\mathsf{KL}}^{(N)}(\mathbf{F}) \rangle - H(\mathsf{F}) \sim O(N^{-1/2})$.

4.1. Robustifying the KL estimator

The KL estimator (5) diverges if two samples are equal, e.g. due to quantization. This should not happen for smooth densities, however, in practice we need to treat such cases. Our approach is to replace the distance ϱ by $\max(\varrho, \varepsilon)$. This corresponds to assuming quantization and calculating the upper bound for H and leads to a KLD estimator:

$$H_{\mathsf{KLD}}^{(N)}(\mathbf{F}) = \frac{d}{N} \sum_{i=0}^N \max(\varrho, \varepsilon) + \log \frac{(N-1)\pi^{d/2}}{\Gamma(1+d/2)} + \gamma$$

4.2. Batch KL estimator

A range of algorithms exist for the all-NN search needed by the KL estimator (5) — from a brute force $O(dN^2)$ to asymptotically optimal $O(dN \log N)$ solution [10]. Unfortunately, all known implementations are currently too slow for our purposes for high N and d.

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

Table 1. Evaluation time in seconds of the proposed KLBD MI estimator in dimension d with N samples and $M = 20$.

We sidestep the issue by taking an approximative approach: We randomly (and possibly non-exhaustively) divide the N samples into groups $\mathbf{F}_1, \ldots, \mathbf{F}_{\lfloor N/M \rfloor}$, so that each group contains M samples. The estimate is then a mean of the H_{KLD} estimator on the groups:

$$H_{\mathsf{KLBD}}^{(N)}(\mathbf{F}) = \frac{1}{\lfloor N/M \rfloor} \sum_{i=1}^{\lfloor N/M \rfloor} H_{\mathsf{KLD}}^{(M)}(\mathbf{F}_i) \quad (6)$$

This is the estimator that we use in the experimental part for evaluating (2,3,4). The batch size M controls the trade-off between its bias $O(M^{-1/2})$ and its speed $O(MNd)$ (using the brute force NN search). The experiments show that the estimator performs surprisingly well even for very small batch sizes $M = 20 \sim 100$. It is simple and fast.

5. EXPERIMENTS

The first experiment (Table 1) measures the speed of evaluating a MI between two sets of samples using the KLBD entropy estimator (6) with $M = 20$, including all overheads, on a 1.4 GHz Pentium machine. The execution times are very reasonable and correspond to the predicted M-fold slowdown with respect to evaluating the SSD (1).

Figure 1 shows the bias and variance of the KLD, KLBD and histogram MI estimators for two 1D Gaussian random variables with varying dependency and $N = 10^4$. We observe that all KL-based estimators have about the same variance, larger than the histogram estimator but still acceptable. The bias of the KLBD estimator decreases with M, for $M \geq 50$ it is already better than that of the histogram and it essentially vanishes for the KLD estimate.

We evaluate the new proposed MI-based image registration criteria *CoMI* and *NbMI* on real images by a rotation experiment: Starting from two perfectly aligned images of size 512^2 pixels, we rotate one of them by $\pm 10°$ (using linear interpolation), crop the images to a fixed size to avoid the influence of the background and evaluate the criterion. A good criterion should have its maximum for a rotation angle 0 and decrease smoothly away from it.

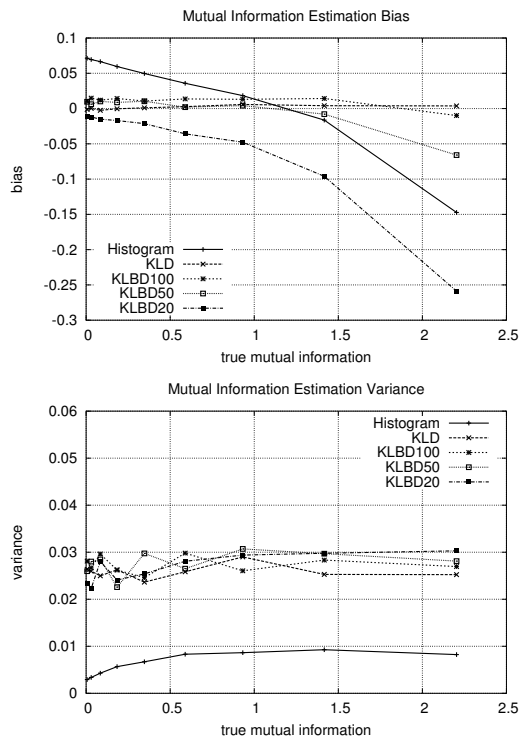First of all we compare the CoMI (3), the vector SSD (1), and standard scalar gray-scale histogram-based

Fig. 1. Biases and variances of several mutual information estimators (histogram, KLD, and KLBD with $M = 20, 50$ and 100) for two linearly dependent Gaussian random sources from $N = 10^4$ samples. 100 experiments were performed.

MI criteria when registering a color Mandrill image with itself (Fig. 2), to verify that all criteria work well in this simple case. We then modify the colors in one of the images (by increasing the saturation and brightness and rotating the colormap) and add some i.i.d. Gaussian noise to individual color components. This confuses the SSD criterion beyond usability and the standard MI is only slightly better, while the CoMI still provides correct and almost undisturbed results. We also observe that the uncertainty due to the stochastic character of the estimator (one standard deviation shown) is below the level of changes we need to detect for a registration accuracy that can be realistically expected, (i.e. around 1 pixel $\sim 0.2°$).

We perform the same kind of experiment comparing the SSD and MI criteria and the NbMI criterion (4) with $h = 2$, leading to $d = 25$ dimensional features. First we register a gray-scale Lena image with itself, to find that all criteria work well. Then we register a low-pass version of the image with a smoothed Sobel-detected edges from the same image. The SSD criterion is useless in this case, showing only irrelevant oscillations due to the global orientation of the edges, while the NbMI criterion identifies the correct alignment flawlessly.

6. CONCLUSIONS

We have developed a fast and robust approximative version of a little known binless Kozachenko-Leonenko NN entropy estimator. This has allowed us to extend MI-based image registration criteria to higher-dimensionality feature vectors. Experiments have shown that such criteria are very general and versatile thanks to their ability to adapt to the statistics of the features used, that need not to be known *a priori*. We have presented two examples: a color based criterion with 3D features; and a neighborhood criterion with 25D features. Mutual information could not have been previously calculated for such features. The criteria perform adequately and in robust manner even when registering difficult images of very different nature, for which classical methods fail.

References

[1] Paul Viola and William M. Wells III, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, , no. 2, pp. 137–154, 1997.

[2] J.P.W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[3] D. Rueckert, M. J. Clarkson, D. L. G. Hill, and D. J. Hawkes, "Non-rigid registration using higher-order mutual information," in *Proceedings of SPIE Medical Imaging 2000: Image Processing*, 2000, pp. 438–447.

[4] Jonathan D. Victor, "Binless strategies for estimation of information from neural data," *Physical Review E*, vol. 66, no. 5, pp. 051903(15), Nov. 2002.

[5] Erik G. Miller, "A new class of entropy estimators for multidimensional densities," in *Proceedings of ICASSP2003*, 2003.

[6] O. Michel A. O. Hero, B. Ma and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Proc. Magazine*, vol. 19, no. 5, pp. 85–95, sep 2002.

[7] Kari Torkkola and William M. Campbell, "Mutual information in learning feature transformations," in *Proc. 17th International Conf. on Machine Learning*. 2000, pp. 1015–1022, Morgan Kaufmann, San Francisco, CA.

[8] L. F. Kozachenko and N. N. Leonenko, "On statistical estimation of entropy of random vector," *Probl. Inf. Trans.*, vol. 23, no. 9, 1987, (in Russian).

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

[10] Paul B. Callahan and S. Rao Kosaraju, "A decomposition of multi-dimensional point-sets with applications to k-nearest-neighbors and n-body potential fields," in *Proceedings 24th Annual AMC Symposium on the Theory of Computing*, 1992, pp. 546–556.
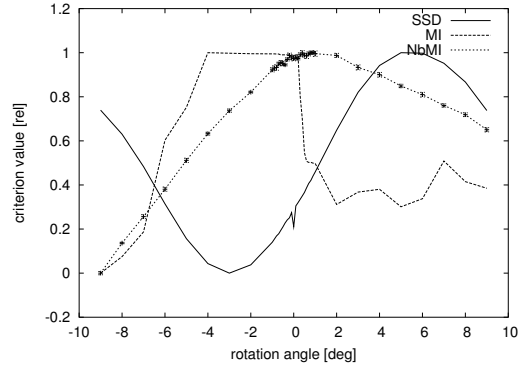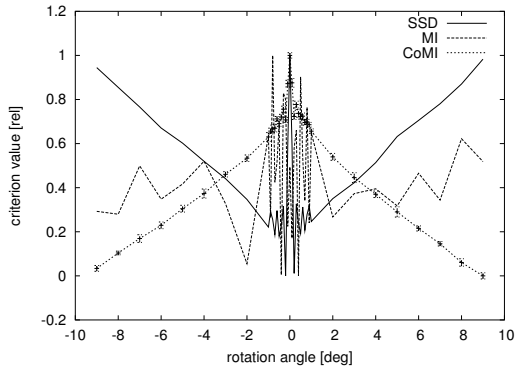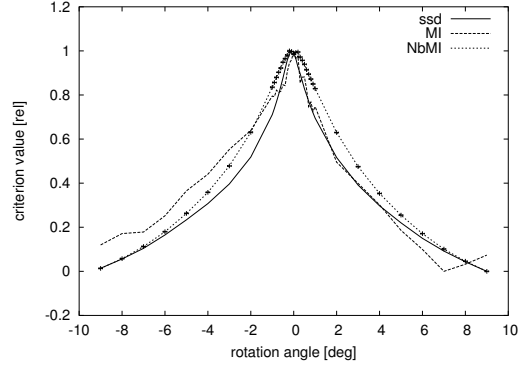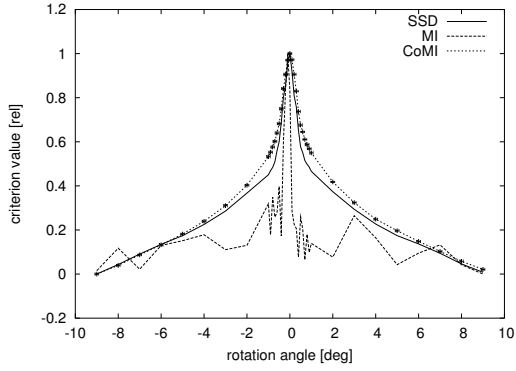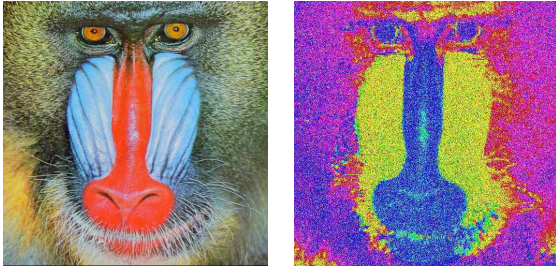
Fig. 2. *SSD*, *MI* and *CoMI (with $M = 20$) criteria all work well when registering a Mandrill color image (top left)* with itself. We show the criteria (rescaled to $[0, 1]$) as a function of a rotation angle *(top graph)*. One standard deviation is shown for the CoMI criterion, since the estimator is stochastic. When registering the original image with its color-modified and noisy version *(top right)*, the *SSD* and *MI* criteria break down, while the proposed *CoMI* criterion still gives correct results *(bottom graph)*.

Fig. 3. The correct rotation angle can be determined from any of the *SSD*, *MI* and *NbMI* criteria when registering a B&W *Lena* image *(top left)* with itself. We show the criteria (rescaled to $[0, 1]$) as a function of a rotation angle *(top graph)*. One standard deviation is shown for the NbMI criterion, since the estimator is stochastic. When registering a blurred version of the original image with a blurred version of its edges *(top right)*, the *SSD* and *MI* are clearly inadequate, while the proposed *NbMI* criterion allows for the correct angle to be reliably detected *(bottom)*.