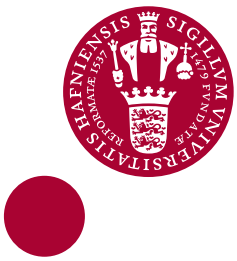


Locally orderless registration

Darkner, Sune; Sporring, Jon

Published in:
I E E E Transactions on Pattern Analysis and Machine Intelligence

DOI:
10.1109/TPAMI.2012.238

Publication date:
2013

Document version
Peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

Locally Orderless Registration

Sune Darkner and Jon Sporring

***Abstract*—This paper presents a unifying approach for calculating a wide range of popular, but seemingly very different, similarity measures. Our domain is the registration of n-dimensional images sampled on a regular grid, and our approach is well suited for gradient-based optimization algorithms. Our approach is based on local intensity histograms and built upon the technique of Locally Orderless Images. Histograms by Locally Orderless Images are well posed and offer explicit control over the 3 inherent and unavoidable scales: the spatial resolution, intensity levels, and spatial extent of local histograms. Through Locally Orderless Images, we offer new insight into the relations between these scales. We demonstrate our unification by developing a Locally Orderless Registration algorithm for two quite different similarity measures, namely, Normalized Mutual Information and Sum of Squared Differences, and we compare these variations both theoretically, and empirically. Finally, using our algorithm, we explain the empirically observed differences between two popular joint density estimation techniques used in registration: Parzen Windows and Generalized Partial Volume.**

***Index Terms*—Similarity Measure, Registration, Normalized Mutual Information, Sum of Squared Differences, Density Estimation, Local Histogram, Scale-Space, Locally Orderless Images.**

I. Introduction

IMAGE similarity measures are crucial components in image registration, and Mutual Information (MI) [1], [2] and Normalized Mutual Information (NMI) [3] are considered state of the art for image registration. MI and NMI are particularly useful for registering Magnetic Resonance Images (MRI) to MRI, and for multi-modal image registration in general. MI and NMI are entropy-based measures and hence rely on probability distributions. Probability distributions are most often approximated by discrete histograms, which pose a challenge to gradient-based optimization schemes. The most common estimation techniques are: the Parzen Window (PW) [2] and the Generalized Partial Volume (GPV) [4], [5]. Empirical comparisons have previously been presented [6], and, recently, we investigated their theoretical connection [7].

In this paper, we present Locally Orderless Registration (LOR). LOR is a framework for performing N-dimensional image registration, and it includes a common framework for a wide range of image similarity measures such as Correlation Ratio, MI, NMI, Huber Norm etc.. The framework is based on local histograms, and we use the technique of Locally Orderless Images (LOI) [8], [9], which makes the 3 natural and unavoidable scale parameters available for image registration, namely: the measurement scale – the effective resolution of the initial image; the intensity or value scale - the effective number of bins in the histogram, and the integration scale - the effective local spatial extent of local histograms. These 3 scales are implemented by smoothing with Gaussian kernels, which imposes what may be the simplest analytical structure on the local histograms. Nevertheless, these scales interact in a nontrivial manner, and we explore their relation theoretically by the local intensity moments, as well as on a simple local image model. We perform extensive empirical investigations on the influence of the scales on the density estimates, as well as NMI, through GPV and PW. To enhance the interpretability and the usability of our results, we summarize and extend our earlier theoretical work [7], where LOR is used to compare PW and GPV, and we demonstrate, both theoretically and empirically, that GPV is asymmetric, and therefore the less-preferred choice of the two. Finally, we present a unifying algorithm for PW and GPV for various measures, in addition to analytical and empirical investigations of its computational complexity. Timing results on our algorithm show that NMI is almost as fast as Sum of Squared Differences (SSD), and that (non-massively) multi-threaded implementation has only 13% overhead when compared to the theoretical computational speed.

Darkner and Sporring are with the Department of computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, Denmark, Email: {darkner,sporring}@diku.dk

Manuscript submitted to IEEE TPAMI April 2012, Major revision August 2012, Minor revision September 2012

A. Previous work

The use of MI for image registration was originally proposed by [1], [2]. An extensive overview was given in [10]. NMI was introduced as a more robust alternative, especially designed for multi-modal image registration [3]. The first implementations relied on Powell's method [4], hill climbing [3], and similar methods without gradients, which were accurate but slow. A GPU speed-up was suggested in [11]. Today, state-of-the-art implementations are gradient-based methods and group in two algorithm types. The first type is based on PW [2] and relies on the fact that the marginal and joint histograms are made continuous by using different kernels, e.g., Gaussian or B-splines [12]. The second type is based on GPV, where the distribution is sampled from the image directly [4]. Analytical derivatives of this method were presented in [13] and a generalization using B-splines was presented in [5]. A variational method relating to LOI [9] for MI (and other measures) was presented in [14]. GPV and PW were compared numerically in [6], concluding that PW is precise and GPV has a larger convergence radius. MI and NMI are notorious for their local minima and difficulty of implementation, and the choice of interpolation scheme greatly influences the smoothness of the objective function. Some investigations into this can be found in [15], [16]. An alternative approach is the Conditional Mutual Information [17]. In [7] we investigated PW and GPV for NMI, using differential calculus in a thorough step-by-step presentation. The derivations were an alternative to

Digital Object Indentifier 10.1109/TPAMI.2012.238

the variational approach in [14], and our approach revealed much faster algorithms, which allowed for a direct comparison between PW and GPV. [14] allowed for a local variant of MI, which was implemented in [18]. Furthermore, a density estimation alternative through a computational complex estimation scheme was suggested in [19], but is, however, unsuited for fast gradient-based optimization schemes.

The remainder of this article is organized as follows: in Section II the general registration framework is described. In Section III we revisit LOI as a basis for analyzing relations between scales for local histograms and discuss both GPV and PW. In Section IV we provide a theoretical comparison between GPV and PW, and in Section V we augment the theoretical comparison with empirical demonstrations of the asymmetry of GPV. In Section VI, we discuss empirical relations between scales. In Section VII we present a fast algorithm for computing PW and GPV for a large range of similarity measures, and in Section VIII we summarize our findings and conclude on our work.

II. Image registration

Image registration is the process of transforming one image $\tilde{I} : \Omega \to \Gamma$, where $\Omega \subseteq \mathbb{R}^N$ and $\Gamma \subseteq \mathbb{R}$, w.r.t. a reference image $J : \Omega \to \Gamma$ such that some functional $\mathcal{F}(\tilde{I}, J)$ is minimized. We consider the diffeomorphic transformation of M parameters, $\boldsymbol{\phi} : \Omega \times \mathbb{R}^M \to \Omega$, and for brevity we write $I = \tilde{I} \circ \boldsymbol{\phi}$. We consider functionals, $\mathcal{F}$, of the form,

$$\mathcal{F} = \mathcal{M}(I, J) + \mathcal{S}(\boldsymbol{\phi}), \tag{1}$$

where $\mathcal{M}$ is a (dis-)similarity measure and $\mathcal{S}$ is a regularization term. Typical forms of $\mathcal{S}$ are elasticity [20], fluid deformations [21], and the recent Kernel Bundle LDDMM [22]. This article focus solely on $\mathcal{M}$.

A. The Similarity Measure

Many similarity measures are of the form,

$$\mathcal{M}_\Omega = \int_\Omega F\big(\boldsymbol{x}, I(\boldsymbol{x}), J(\boldsymbol{x})\big)\, d\boldsymbol{x}, \tag{2}$$

where the loss-function, F, is integrated over the spatial domain. Popular choices of loss-functions are monomials, $F(I(\boldsymbol{x}), J((\boldsymbol{x}))) = (I(\boldsymbol{x}) - J(\boldsymbol{x}))^q$ for $q > 0$. Other similarity measures have the form of,

$$\mathcal{M}_\Gamma = \int_{\Gamma^2} F\big(\boldsymbol{x}, i, j, h_{I,J}(i,j)\big)\, di\, dj, \tag{3}$$

where $h_{I,J} : \Gamma^2 \to \mathbb{R}_+$ is the joint histogram of image I and J with intensity variables i and j. A popular choice is Mutual Information (MI) [23], $\mathcal{M}_{\text{MI}} = \mathcal{H}_I + \mathcal{H}_J - \mathcal{H}_{I,J}$, where $\mathcal{H}$ is the (joint) entropy, in which case $F = p(i,j) \ln p(i,j) - p(i) \ln p(i) - p(j) \ln p(j)$. The natural logarithm is often used for convenience, and the distribution p is the normalized (joint) histogram $p(i,j) = h(i,j) / \int_{\Gamma^2} h(i,j)\, di\, dj$, such that $p(i) = \int_\Gamma p(i,j)\, dj$, and $p(j) = \int_\Gamma p(i,j)\, di$.

A seemingly major difference between (2) and (3) is the integration domain. However, we will show that by reordering the integral by the distribution of I and J values, we may rewrite (2) in terms of local histograms $h(\boldsymbol{x}, i, j)$. This has several advantages: 1) it creates a common form for both classes of similarity measures; 2) the histogram perspective makes the 3 fundamental scales of images – measure, intensity, and integration – available for similarity measures on the form of (2); 3) the loss-function F for q-norms and similar becomes linear w.r.t. the transformation parameters; and 4) with the use of smooth kernels, the derivatives w.r.t. space and intensity are trivial, and thus are readily available for gradient descent schemes. Nevertheless, there is a minor disadvantage: in the limit of infinitely closely sampled images, the histograms have poles corresponding to image values, where the spatial gradient of the image is zero. This is a theoretical problem for similarity measures on the form of (3), which our approach carries over to measures on the form of (2). However, in practice this is of little importance, since we consider generic images, i.e., images whose structure is stable w.r.t. negligible noise, and for such images, the set of areas with zero gradients are singular points with measure zero, i.e., constant patches are non-generic in real images. We will assume that the poles in the histograms likewise have measure zero, which is supported by our observations, but which we leave to be proven in subsequent work.

Our approach for calculating similarity measures for a wide range of loss-functions, F, linear as well as non-linear, has the following form:

$$\mathcal{M} = \int_{\Omega \times \Gamma^2} F\big(\boldsymbol{x}, i, j, h_{I,J}(\boldsymbol{x}, i, j)\big)\, d\boldsymbol{x}\, di\, dj. \tag{4}$$

Most functionals in the literature are position-independent, which will be our focus as well. Henceforth, we will concentrate on two specializations of (4): $\mathcal{M}_{\text{lin}}$ or $\mathcal{M}_{\text{nlin}}$. The similarity measure $\mathcal{M}_{\text{lin}}$ uses the position-independent, linear loss-functions,

$$\mathcal{M}_{\text{lin}} = \int_{\Gamma^2} F(i,j) h_{I,J}(i,j)\, di\, dj. \tag{5}$$

This measure includes (2) with any position independent loss-function, such as monomials; it is linear w.r.t. F and h, and the transformation parameters only influence h. To understand the relation between (2) and (5), consider $\int I^2(\boldsymbol{x})\, d\boldsymbol{x}$. By introducing a discretization of intensities, $i_1 < i_2 < i_3 \ldots$, we approximate the integral as $\sum_n i_n^2 \mu_n$, where μ_n is the area of $\{\boldsymbol{x} | i_n \leq I(\boldsymbol{x}) < i_{n+1}\}$. In the limit of $\Delta_n = i_{n+1} - i_n \to 0$, this area is equal to the integral of $1/|\nabla I|$ along the isophote i_n in its arc-length parameter, but more importantly, it is well approximated by $h_n \Delta_n$, where h_n is the length of the isophote i_n. Hence, we take the limit and write $\int I^2(\boldsymbol{x})\, d\boldsymbol{x} \simeq \sum_n i_n^2 \mu_n \simeq \int i^2 h(i)\, di$.

The similarity measure, $\mathcal{M}_{\text{nlin}}$, uses the position-independent, non-linear loss-function,

$$\mathcal{M}_{\text{nlin}} = \int_{\Gamma^2} F\big(h_{I,J}(i,j)\big)\, di\, dj, \tag{6}$$

where F now denotes some non-linear functional, and this form includes MI. As will be shown later, the added complexity from linear to non-linear measures has little influence on computation time.

In this paper we will consider Normalized Mutual Information (NMI) [3], which has proven to be very powerful for the registration of medical images in general, and the Sum of Squared Differences (SSD), as a representative of a simple similarity measure.

III. Density estimation

A common algorithm for estimating the histogram of an image is counting: given an image I, a set of isophotes, $I(\boldsymbol{x}) = i_n$, $m > n \Rightarrow i_m > i_n$, bin-widths $\Delta i_n > 0$, and an indicator function,

$$P_n(i) = \begin{cases} 1, & \text{if } i_n \leq i < i_n + \Delta i_n, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The histogram may then be found as,

$$h(n) = \int_\Omega P_n(I(\boldsymbol{x}))\, d\boldsymbol{x}, \tag{8}$$

or as a sum using a suitable discretization of Ω. The bin-widths act as scale parameters, in the sense that increasing Δi_n results in a histogram with less detail. This can be stated precisely: select a discrete set of sample points and bin-widths such that $\Delta i_n = i_{n+1} - i_n$, and consider 2 neighboring histogram values, $h(n)$ and $h(n+1)$. In this case, the sum, $h'(n) = h(n) + h(n+1)$, is equivalent to evaluating the integral with a modified indicator function,

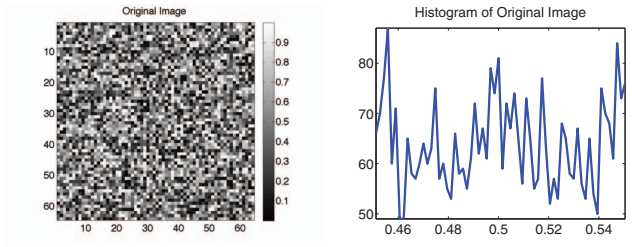
$$P'_n(i) = \begin{cases} 1, & \text{if } i_n \leq i < i_{n+1} + \Delta i_{n+1} = i_n + \Delta i'_n \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where $\Delta i'_n = \Delta i_n + \Delta i_{n+1}$. By induction it becomes clear that filtering $h(n)$ with a Boxcar function (0-order b-spline) of height 1 and width m is equivalent to increasing the extent of the indicator function as $\Delta i'_n = \sum_{k=0}^{m-1} \Delta i_{n+k}$. Thus, increasing Δi is equivalent to smoothing the histogram with a Boxcar function.

In general, the interesting scales of i are not provided by the data, and therefore the only option is to study all scales, that is, all discretizations of intensity. Along with the scale-space on the spatial parameter $\boldsymbol{x}$, this leads to a scale-space theory for space and intensity known as Imprecision Space [8]. In the general case, histograms are local. Since the scale of the region of interest is not generally given, we are also required to study all scales. This scale we denote the integration scale. The Boxcar function is often not the optimal filter for many data analysis applications, since its Fourier transformation contains zero crossings. A better, possibly most conservative, alternative is the Gaussian filter, which leads to the technique of Locally Orderless Images (LOI) [9] to be reviewed in the following section.

A. Estimating local histograms

According to LOI, a local histogram is obtained as follows: First, a (possibly deformed) image I is smoothed with the kernel K, a soft isophote i is extracted using kernel P, and finally the isophote mass is calculated in a neighborhood of a point $\boldsymbol{x}$ with kernel W. Formally,



(a) (b)

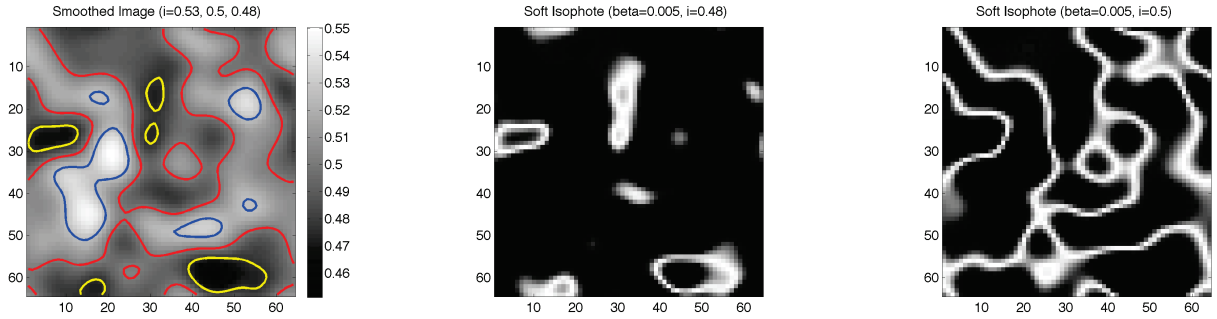
Fig. 1. (a) A random image and (b) its histogram.

$$h_I(i, \boldsymbol{x}, \boldsymbol{\Phi}, \alpha, \beta, \sigma) = P(I(\boldsymbol{x}, \boldsymbol{\Phi}, \sigma) - i, \beta) * W(\boldsymbol{x}, \alpha), \tag{10}$$

$$I(\boldsymbol{x}, \boldsymbol{\Phi}, \sigma) = I(\boldsymbol{x}, \boldsymbol{\Phi}) * K(\boldsymbol{x}, \sigma), \tag{11}$$

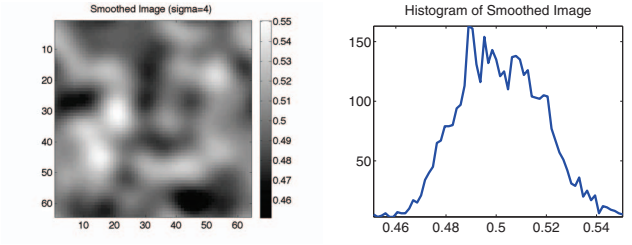
where $P : \mathbb{R} \times \mathbb{R}_+ \to [0, 1]$, is an intensity measurement of scale β and is often called the Parzen Window (PW), $K : \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}_+$ is a spatial measurement kernel of scale σ, $W : \mathbb{R}^N \times \mathbb{R}_+ \to \mathbb{R}_+$ is an integration window of integration scale α, $\cdot * \cdot$ is the convolution operator taken w.r.t. the variable $\boldsymbol{x}$, and $\boldsymbol{\Phi} \in \mathbb{R}^M$ denotes the parameters for the transformation. We will further assume that $\int K(\boldsymbol{x}, \sigma)\, d\boldsymbol{x} = \int W(\boldsymbol{x}, \alpha)\, d\boldsymbol{x} = 1$. The histogram h_J is defined in a similar way, independently of $\boldsymbol{\Phi}$. In [9] it is proposed to use $P(i, \beta) = e^{-i^2/(2\beta^2)}$, $K(\boldsymbol{x}, \sigma) = e^{-\boldsymbol{x}^T\boldsymbol{x}/(2\sigma^2)}/(2\pi\sigma^2)^{N/2}$, and $W(\boldsymbol{x}, \alpha) = e^{-\boldsymbol{x}^T\boldsymbol{x}/(2\alpha^2)}/(2\pi\alpha^2)^{N/2}$, which implies the structure of the heat diffusion in all 3 scale parameters and is considered the simplest structure imposable for studying data by all scales. In typical registration scenarios, such as registering CT and MR images, intensity and spatial scale are of quite different natures. The spatial scales can often be related to a common frame of reference, but this is often difficult for intensity scales.

In the following we will give a tutorial on how local histograms are calculated in a step by step manner, and provide intuition on the 3 scale parameters. Consider a random image and its histogram as calculated by the Matlab `hist` function, shown in Fig. 1. In terms of local histogram parameters, this corresponds to: $\alpha = \infty$, $\sigma = 0$, and $\beta = 1/\sqrt{12}$, the standard deviation of a Boxcar function of width 1. To estimate a local histogram we go through 3 steps: the first step is to smooth the original image with kernel K. The kernel K controls the image scale, σ. This is illustrated in Fig. 2 and corresponds to $\alpha = \infty$, $\sigma > 0$, and $\beta = \Delta i/\sqrt{12}$, where Δi is the original intensity scale. Since smoothing an image implies a monotonic contraction of image intensity around the mean value, we expect that the histogram is likewise contracted, when increasing σ. This is confirmed by the experiment illustrated in Fig. 2(b). The second step is to calculate the soft isophote i with kernel P: The kernel P controls intensity scale, β. This is illustrated in Fig. 3 and corresponds to $\alpha = 0$, $\sigma > 0$, and $\beta > 0$. Fig. 3(b) and 3(c) show the spread of 2 fixed isophotes for the chosen P. For a fixed position $\boldsymbol{x}$, the image contains the value of the local histogram at $\boldsymbol{x}$. Hence, the stack of images for all isophotes gives all the local histograms. The



(a) (b) (c)

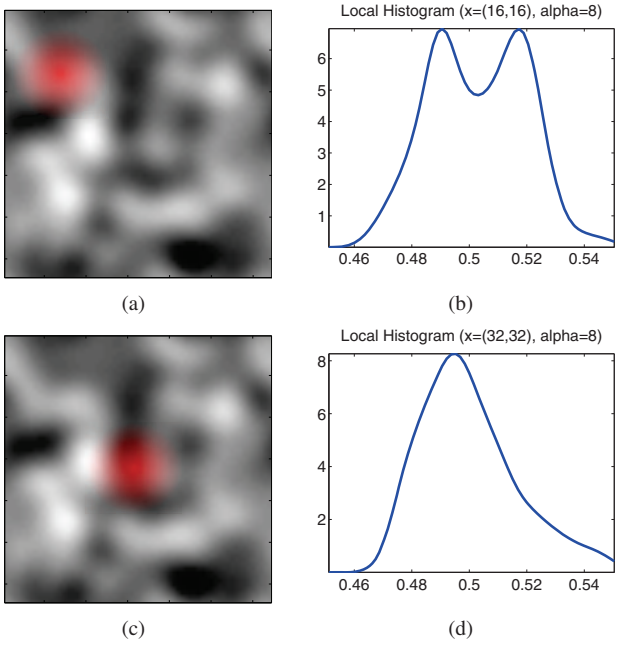
Fig. 3. Measuring isophotes in Fig. 2. Images (a) 3 isophote lines as produced by Matlab's `contour` function; (b) and (c) the yellow and red isophotes as extracted with a kernel P using $i = 0.48$ and $i = 0.50$ and in both instances $\beta = 0.005$.



(a) (b)

Fig. 2. (a) The image in Fig. 1(a), smoothed with $\sigma = 4$, and (b) its histogram.

spread of a soft isophote depends on the image geometry at $I(\boldsymbol{x}, \sigma) = i$: The spread will be large, where the gradient magnitude is small, and small, where the gradient magnitude is large. In general the width β acts as the bin-width in the histogram, and varying β corresponds to varying the degree of smoothing of the histogram. The last step is to calculate the local isophote area near $\boldsymbol{x}$ with kernel W: The kernel W controls the locality of the local histogram, α, illustrated in Fig. 4. Note that the histograms change quite significantly depending on the position of the kernel W.



(a) (b) (c) (d)

Fig. 4. Examples of local histograms generated by Locally Orderless Images in neighbourhoods as indicated by the red overlays.

B. Some relations between scales

The relation between α and σ may be stated in terms of the histogram's raw and central moments. The raw and central moments of order $n \geq 0$ of the histogram h at position $\boldsymbol{x}$ are given as,

$$\mu'_n = \int_{-\infty}^{\infty} i^n h(i, \boldsymbol{x})/k(\boldsymbol{x})\, di, \tag{12}$$

$$\mu_n = \int_{-\infty}^{\infty} (i - \mu(\boldsymbol{x}))^n h(i, \boldsymbol{x})/k(\boldsymbol{x})\, di, \tag{13}$$

where $k(\boldsymbol{x}) = \int_{-\infty}^{\infty} h(i, \boldsymbol{x})\, di$ and $\mu(\boldsymbol{x}) = \mu'_1$ is the mean value. In the following, we will evaluate these moments. We will use $L = I * K$ as a convenient shorthand

- *Normalization constant k:* Convolution is linear, thus $k(\boldsymbol{x}) = \int_{-\infty}^{\infty} P(L(\boldsymbol{x}) - i) * W(\boldsymbol{x})\, di = \left(\int_{-\infty}^{\infty} P(L(\boldsymbol{x}) - i)\, di\right) * W(\boldsymbol{x})$. The value $L(\boldsymbol{x})$ is constant w.r.t. the integration in i, and since the integral in i is over the entire domain, we conclude that it is independent of finite translations $L(\boldsymbol{x})$, and hence, independent of L, and therefore, of $\boldsymbol{x}$. Finally, since W has unit integral, we conclude that

$$k = \int_{-\infty}^{\infty} P(i)\, di \tag{14}$$

independently of $\boldsymbol{x}$. In the case of a Gaussian Parzen window with variance β^2, then $k_{\text{Gauss}} = \beta\sqrt{2\pi}$.

- *Mean value μ:* If the Parzen window, P, is centered at zero, i.e., $\int P(i) i\, di = 0$, then $P(L(\boldsymbol{x}) - i)$ is centered at $L(\boldsymbol{x})$, i.e., $\int P(L(\boldsymbol{x}) - i) i\, di = L(\boldsymbol{x}) \int P(i)\, di = L(\boldsymbol{x})k$. Using the linearity of convolution with W, and expanding L, we find that

$$\mu = \int_{-\infty}^{\infty} ih(i, \boldsymbol{x})/k\, di = k^{-1} L(\boldsymbol{x}) * W(\boldsymbol{x}) \tag{15}$$

$$= k^{-1} I(\boldsymbol{x}) * K(\boldsymbol{x}) * W(\boldsymbol{x}) \tag{16}$$

$$= k^{-1} I(\boldsymbol{x}) * W'(\boldsymbol{x}), \tag{17}$$

where $W'(\boldsymbol{x}) = K(\boldsymbol{x}) * W(\boldsymbol{x})$. In case of Gaussian K of variance σ^2 and W of variance α^2, then $W'(\boldsymbol{x})$ is a Gaussian of variance $\sigma^2 + \alpha^2$.

- *Raw moments μ'_n:* Expanding h and using the linearity of convolution, we find that the raw moments may be written as,

$$\mu'_n = \int_{-\infty}^{\infty} i^n P(L(\boldsymbol{x}) - i)/k * W(\boldsymbol{x})\, di \tag{18}$$

$$= \left(\int_{-\infty}^{\infty} i^n P(L(\boldsymbol{x}) - i)/k\, di \right) * W(\boldsymbol{x}) \tag{19}$$

$$= \eta'_n * W(\boldsymbol{x}), \tag{20}$$

where η'_n is the n'th raw moment of a random variable distributed as P/k and with mean value L. A useful relation between the scales σ and α may be derived by considering the relation between the raw and central moments of P/k: Consider the general case of the raw moments of a given statistical variable X, with mean $E(X) = 0$, and E as the expectation operator. If we construct another variable $Y = X + \bar{y}$ for some constant $\bar{y}$, then the n'th raw moment of Y is $E(Y^n) = E\big((X + \bar{y})^n\big) = E\big(\sum_{j=0}^{n} \binom{n}{j} X^j \bar{y}^{n-j}\big) = \sum_{j=0}^{n} \binom{n}{j} E\big(X^j\big) \bar{y}^{n-j}$, where $E(X^j)$ are the central moments of Y. In our case, we may consider P/k the distribution of a random variable with raw and central moments η'_n and η_n, which for the above reasons are related as,

$$\eta'_n = \sum_{j=0}^{n} \binom{n}{j} \eta_j (\eta'_1)^{n-j}. \tag{21}$$

For Parzen windows centered at zero, we have that $\eta'_1 = L(\boldsymbol{x})$, and η_j is independent of L. Thus we conclude, that for Parzen windows centered at zero, μ'_n is a linear combination of the terms $L(\boldsymbol{x})^{n-j} * W(\boldsymbol{x}),\ j = 0 \ldots n$. Hence, the relation between σ and α is non-linear for $n > 1$, and for Gaussian K and W the relation behaves locally in L-values as a pseudo-linear scale-space [24]. Finally, for a Gaussian Parzen window with variance β^2, the n'th central moment is $(n-1)!!\beta^n$ for even n and 0 otherwise, where $n!! = n(n-2)(n-4)\ldots$ is the double factorial function. Examples of raw moments, when using a Gaussian as the Parzen window, are given in Table I.

- *Central moments μ_n:* The central moments of h may be constructed from its raw moments, since

$$\mu_n = \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} \mu'_j \mu^{n-j}. \tag{22}$$

Examples of central moments, when using a Gaussian as Parzen window, are given in Table I.

To gain intuition on the relation between β and α, consider an image, which in the neighborhood of the point $\boldsymbol{x}_0$, is linear with gradient $\nabla I(x)$. The image in the neighborhood of $\boldsymbol{x}_0$ is then given as

$$I(\boldsymbol{x}) \simeq I(\boldsymbol{x}_0) + (\boldsymbol{x} - \boldsymbol{x}_0) \cdot \nabla I(\boldsymbol{x}_0), \tag{23}$$

and the isophotes near $I(\boldsymbol{x}_0)$ are all lines perpendicular to the gradient. The image in the neighborhood around $\boldsymbol{x}_0$ is invariant w.r.t. smoothing with symmetric and normalized kernels, hence σ has no influence on the local histograms for small values of σ. However, the interplay between β and α is nontrivial: the soft isophotes are constant in the perpendicular direction of the gradient. Hence, we may consider this a one-dimensional problem along the axis of the gradient x, for instance, and $I(x) \simeq ax + b$, where $a = |\nabla I(\boldsymbol{x}_0)|$, $ax = (\boldsymbol{x} - \boldsymbol{x}_0) \cdot \nabla I(\boldsymbol{x}_0)$, and $b = I(\boldsymbol{x}_0)$. The soft isophote b using a Gaussian P is $P(ax, \beta) = P(x, \beta/a)$, and convolution with a Gaussian integration kernel $W(x, \alpha)$ yields another Gaussian

$$P(ax, \beta) * W(x, \alpha) = P(x, \sqrt{\beta^2/a^2 + \alpha^2}), \tag{24}$$

due to the semi-group properties of Gaussian convolution.

In general, varying β and varying σ yields different results, since the width of a soft isophote in a point is proportional to the gradient in the point, while the extent of the local average is irrespective of the gradient in the point. In addition, near the symmetry set [25], the soft isophote will exhibit ridge-like behavior.

C. Estimating local densities

The local density distributions are obtained by normalizing to unity,

$$p_I(i|\boldsymbol{x}, \boldsymbol{\Phi}, \alpha, \beta, \sigma) \simeq \frac{h_I(i, \boldsymbol{x}, \boldsymbol{\Phi}, \alpha, \beta, \sigma)}{\int_\Gamma h_I(j, \boldsymbol{x}, \boldsymbol{\Phi}, \alpha, \beta, \sigma) dj}, \tag{25}$$

$$p_I(i|\boldsymbol{\Phi}, \alpha, \beta, \sigma) = \frac{1}{|\Omega|} \int_\Omega p_I(i|\boldsymbol{x}, \boldsymbol{\Phi}, \alpha, \beta, \sigma)\, d\boldsymbol{x}, \tag{26}$$

and where we have assumed (conditional) independence and uniformity, such that $p_I(i, \boldsymbol{x}|\boldsymbol{\Phi}, \alpha, \beta, \sigma) = p_I(i|\boldsymbol{x}, \boldsymbol{\Phi}, \alpha, \beta, \sigma)/|\Omega|$. The density p_J is defined in a similar manner. As in [14], we extend the concept to the joint distributions as follows:

$$h_{I,J}(i, j, \boldsymbol{x}, \boldsymbol{\Phi}, \alpha, \beta, \sigma) = (P(I(\boldsymbol{x}, \boldsymbol{\Phi}, \sigma) - i, \beta) P(J(\boldsymbol{x}, \sigma) - j, \beta)) * W(\boldsymbol{x}, \alpha), \tag{27}$$

$$p_{I,J}(i, j|\boldsymbol{x}, \boldsymbol{\Phi}, \alpha, \beta, \sigma) \simeq \frac{h_{I,J}(i, j, \boldsymbol{\Phi}, \boldsymbol{x}, \alpha, \beta, \sigma)}{\int_{\Gamma^2} h_{I,J}(k, l, \boldsymbol{x}, \alpha, \beta, \sigma)\, dk\, dl}, \tag{28}$$

$$p_{I,J}(i, j|\boldsymbol{\Phi}, \alpha, \beta, \sigma) = \frac{1}{|\Omega|} \int_\Omega p_{I,J}(i, j|\boldsymbol{\Phi}, \boldsymbol{x}, \alpha, \beta, \sigma)\, d\boldsymbol{x}, \tag{29}$$

where we also have assumed (conditional) independence and uniformity such that $p_{I,J}(i, j, \boldsymbol{x}|\boldsymbol{\Phi}, \alpha, \beta, \sigma) = p_{I,J}(i, j|\boldsymbol{x}, \boldsymbol{\Phi}, \alpha, \beta, \sigma)/|\Omega|$.

IV. Theoretical comparison of PW and GPV density estimation

LOI is the cornerstone for understanding the difference between the PW and GPV density estimators. In the following we will show, how these methods are related to our approach and to each other. The histogram update for the two schemes is illustrated in Fig. 5. We will now briefly introduce the two density estimation techniques.

The PW approach to estimating the joint density was originally proposed along with MI in [2], and is often used in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

n	η_n	μ'_n	μ_n
0	1	1	1
1	0	$L(\boldsymbol{x}) * W(\boldsymbol{x})$	0
2	β^2	$\left(L(\boldsymbol{x})^2 + \beta^2\right) * W(\boldsymbol{x})$	$-(\mu'_1)^2 + \mu'_2$
3	0	$\left(L(\boldsymbol{x})^3 + 3\beta^2 L(\boldsymbol{x})\right) * W(\boldsymbol{x})$	$2(\mu'_1)^3 - 3\mu'_1\mu'_2 + \mu'_3$
4	$3\beta^4$	$\left(L(\boldsymbol{x})^4 + 6\beta^2 L(\boldsymbol{x})^2 + 3\beta^4\right) * W(\boldsymbol{x})$	$-3(\mu'_1)^4 + 6(\mu'_1)^2\mu'_2 - 4\mu'_1\mu'_3 + \mu'_4$
5	0	$\left(L(\boldsymbol{x})^5 + 10\beta^2 L(\boldsymbol{x})^3 + 15\beta^4 L(\boldsymbol{x})\right) * W(\boldsymbol{x})$	$4(\mu'_1)^5 - 10(\mu'_1)^3\mu'_2 + 10(\mu'_1)^2\mu'_3 - 5\mu'_1\mu'_4 + \mu'_5$

TABLE I
EXAMPLES OF RAW AND CENTRAL MOMENTS μ'_n AND μ_n OF ORDER n, WHEN THE PARZEN WINDOW HAS CENTRAL MOMENTS η_j, $j = 0 \ldots n$, AS DOES A GAUSSIAN OF ZERO MEAN AND VARIANCE β^2.

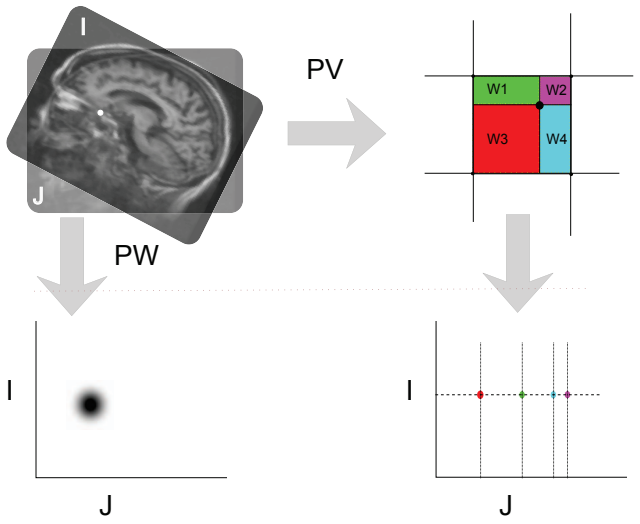


Fig. 5. The histogram update of the Parzen Window (PW) and the partial volume (PV) for a 2-dimensional example. The top left shows two images, where one has been transformed w.r.t. the other. Considering the white-spot: The bottom left shows the corresponding PW update, the top right shows the weight calculations for GPV, which results in the 4 updates illustrated in the bottom right.

the literature. Given the transformation, PW estimates the joint intensity histogram by summing the number of co-occurrences of intensities over space,

$$h_{\text{PW}}(i,j) = \frac{1}{N}\sum_{n=1}^{N} P\left(\begin{bmatrix} I(\boldsymbol{x}_n) \\ J(\boldsymbol{x}_n) \end{bmatrix} - \begin{bmatrix} i \\ j \end{bmatrix}\right) \tag{30}$$

where P is a distribution, typically of the Gaussian type. This is illustrated in the figure as the step from the upper to the lower left. Note that this requires an interpolation; typically grid points of $J(\boldsymbol{x})$ are used, and values of $\tilde{I}(\tilde{\boldsymbol{x}})$ are found by interpolation, where $\tilde{\boldsymbol{x}} = \boldsymbol{\phi}^{-1}(\boldsymbol{x})$.

Shortly after the introduction of PW, Partial Volume was introduced in [4] and extended to GPV in [5]. The algorithm is most easily explained by an example in 2-dimensions: An expanded view of the top right graph in Fig. 5 is given in Fig. 6. In the figure there are shown 9 grid points in J's coordinate system, $\boldsymbol{x}_{i,j}$, $i,j \in \{1,2,3\}$. Assume that the mapping is such that 4 neighboring grid points of I happen to land between grid points of J, as depicted by the circles, $\boldsymbol{\phi}(\tilde{\boldsymbol{x}}_{m,n})$, $m,n \in \{1,2\}$. In that case, each mapped point defines 4 rectangles, for example the areas $w^{11}_{r,s}$ $r,s \in \{1,2\}$. Now consider the mapping $\boldsymbol{\phi}(\tilde{\boldsymbol{x}}_{11})$. For

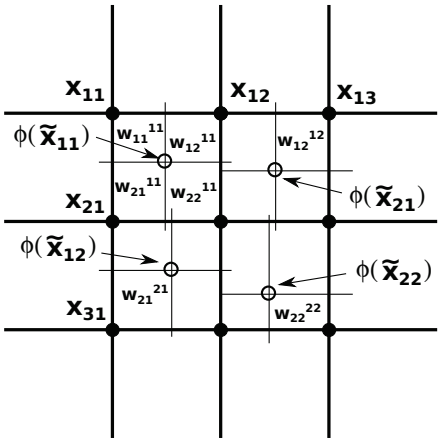


Fig. 6. 2-dimensional example of GPV variables: $\boldsymbol{x}_{i,j}$, $i,j \in \{1,2,3\}$ are neighboring grid points of J, $\boldsymbol{\phi}(\tilde{\boldsymbol{x}}_{m,n})$, $m,n \in \{1,2\}$ are assumed mappings of neighboring grid points of I, and $w^{11}_{r,s}$ $r,s \in \{1,2\}$ are areas defined by the mapping $\boldsymbol{\phi}$.

this point the histogram is updated as,

$$h(\tilde{I}(\tilde{\boldsymbol{x}}_{11}), J(\boldsymbol{x}_{11})) \mathrel{+}= w^{11}_{22}, \tag{31}$$
$$h(\tilde{I}(\tilde{\boldsymbol{x}}_{11}), J(\boldsymbol{x}_{12})) \mathrel{+}= w^{11}_{21}, \tag{32}$$
$$h(\tilde{I}(\tilde{\boldsymbol{x}}_{11}), J(\boldsymbol{x}_{21})) \mathrel{+}= w^{11}_{12}, \tag{33}$$
$$h(\tilde{I}(\tilde{\boldsymbol{x}}_{11}), J(\boldsymbol{x}_{22})) \mathrel{+}= w^{11}_{11}, \tag{34}$$

which implies that the point in the histogram corresponding to the pair $\boldsymbol{x}$ and $\boldsymbol{\phi}(\tilde{\boldsymbol{x}})$ that are closest, gets the highest increment. The rhs. corresponds to the updating of the histogram along a line, as depicted on the bottom right of Fig. 5.

A variant of the algorithm is obtained if we switch view point: Each grid point $J(\boldsymbol{x})$ is included in a number of updates in our example, and here we may instead consider the summation for a fixed grid point in J, e.g., the center point $\boldsymbol{x}_{22}$ in Fig. 6. The update for this becomes,

$$h(\tilde{I}(\tilde{\boldsymbol{x}}_{11}), J(\boldsymbol{x}_{22})) \mathrel{+}= w^{11}_{11} \tag{35}$$
$$h(\tilde{I}(\tilde{\boldsymbol{x}}_{12}), J(\boldsymbol{x}_{22})) \mathrel{+}= w^{21}_{21} \tag{36}$$
$$h(\tilde{I}(\tilde{\boldsymbol{x}}_{22}), J(\boldsymbol{x}_{22})) \mathrel{+}= w^{22}_{22} \tag{37}$$
$$h(\tilde{I}(\tilde{\boldsymbol{x}}_{21}), J(\boldsymbol{x}_{22})) \mathrel{+}= w^{12}_{12}. \tag{38}$$

GPV extends Partial volume by replacing the areas, w, with the values of a smoothing kernel, W, such that the updates in our example are performed as,

$$h(\tilde{I}(\tilde{\boldsymbol{x}}_{mn}), J(\boldsymbol{x}_{22})) \mathrel{+}= W(\boldsymbol{\phi}(\tilde{\boldsymbol{x}}_{mn})). \tag{39}$$

In general, this is slightly different from the original GPV algorithm, since cases, such as those where 4 points $\phi(\tilde{\boldsymbol{x}})$ are mapped into the same square, are handled differently. However, we consider the two algorithms to be similar approximations of histogram updates as the intersection between isophotes from J and soft isophotes from $\tilde{I}$.

A. The PW is a special case of Locally Orderless Images

We will now show that PW is a special case of LOI. Consider (10)–(11) and let $\alpha \to \infty$. In that case, the window h_I simplifies as,

$$h_I(i, \boldsymbol{x}, \boldsymbol{\Phi}, \alpha, \beta, \sigma) \to \text{const.} \int_\Omega P(I(\boldsymbol{\psi}, \boldsymbol{\Phi}, \sigma) - i, \beta)\, d\boldsymbol{\psi}, \tag{40}$$

$$p_I(i|\boldsymbol{\Phi}, \alpha, \beta, \sigma) \to \frac{\int_\Omega P(I(\boldsymbol{\psi}, \boldsymbol{\Phi}, \sigma) - i, \beta)\, d\boldsymbol{\psi}}{\int_\Gamma \int_\Omega P(I(\boldsymbol{\psi}, \boldsymbol{\Phi}, \sigma) - j, \beta)\, d\boldsymbol{\psi}\, dj}. \tag{41}$$

Choosing

$$P(i, \beta) = e^{-i^2/(2\beta^2)}, \tag{42}$$

we find that

$$\int_\Gamma \int_\Omega P(I(\boldsymbol{\psi}, \boldsymbol{\Phi}, \sigma) - j, \beta)\, d\boldsymbol{\psi}\, dj = |\Omega| \sqrt{2\pi\beta^2}, \tag{43}$$

and

$$p_I(i|\boldsymbol{\Phi}, \alpha, \beta, \sigma) \to \frac{1}{|\Omega|\sqrt{2\pi\beta^2}} \int_\Omega e^{-(I(\boldsymbol{x}, \boldsymbol{\Phi}, \sigma) - i)^2/(2\beta^2)}\, d\boldsymbol{x}. \tag{44}$$

Likewise, we have

$$p_{I,J}(i, j|\boldsymbol{\Phi}, \alpha, \beta, \sigma) \to \frac{\int_\Omega e^{-(I(\boldsymbol{x}, \boldsymbol{\Phi}, \sigma) - i)^2 + (J(\boldsymbol{x}, \sigma) - j)^2/(2\beta^2)}\, d\boldsymbol{x}}{|\Omega| 2\pi\beta^2}. \tag{45}$$

This is precisely the PW method using a Gaussian kernel with infinite support given in (30). Similar results are obtained for any integrable Parzen window, $P(i, \beta)$. The PW can be interpreted as a globally orderless image, as W extends globally.

As a side note, since both (44) and (45) obey the diffusion equation w.r.t. $\beta^2/2$, we may use Green's theorem and write,

$$p_I(i|\sqrt{\beta_0^2 + \beta^2}) = p_I(i|\beta_0) * G(i, \beta), \tag{46}$$

$$p_{I,J}(i, j|\sqrt{\beta_0^2 + \beta^2}) = p_{I,J}(i, j|\beta_0) * G([i, j]^T, \beta), \tag{47}$$

for the quick computation of a range of PW sizes, where G is a Gaussian kernel with standard deviation β. Further, $\alpha \to 0$ in MI for 2D images reduces to $-\log(\angle(\nabla I, \nabla J))$ [26], i.e., the angle between the gradients of the images at x, which is similar to the Normalized Gradient Fields proposed in [16].

B. GPV is an approximation of Locally Orderless Images

GPV may be derived from the joint histograms as follows. First, calculate the joint histogram,

$$h_{I,J}(i, j, \boldsymbol{x}, \alpha, \beta, \sigma) = \int_\Omega P(I(\boldsymbol{\psi}, \sigma) - i, \beta) P(J(\boldsymbol{\psi}, \sigma) - j, \beta) W(\boldsymbol{x} - \boldsymbol{\psi}, \alpha)\, d\boldsymbol{\psi} \tag{48}$$

$$\approx P(J(\boldsymbol{x}, \sigma) - j, \beta) \int_\Omega P(I(\boldsymbol{\psi}, \sigma) - i, \beta) W(\boldsymbol{x} - \boldsymbol{\psi}, \alpha)\, d\boldsymbol{\psi} \tag{49}$$

$$= P(J(\boldsymbol{x}, \sigma) - j, \beta)\, [P(I(\boldsymbol{x}, \sigma) - i, \beta) * W(\boldsymbol{x}, \alpha)] \tag{50}$$

Then set P to a Boxcar function,

$$P(i, \beta) = \begin{cases} 1 & \text{if } -\frac{\beta}{2} \le i < \frac{\beta}{2} \\ 0 & \text{otherwise} \end{cases} \tag{51}$$

where β is chosen such that $I(\boldsymbol{\psi}, \boldsymbol{\Phi}, \sigma)$ is mapped into non-coinciding isophote curves. The motivation for this is that all isophotes can be evaluated simultaneously at $\boldsymbol{x}$ and can be thought of as an 0-order b-spline PW. Thus, our formulation is the intersection between isophotes in J with soft isophotes in I, as discussed below (39). When integrating over the entire domain $\boldsymbol{\Omega}$, the GPV scheme is obtained. Thus GPV uses small local histograms, which are integrated to form the globally orderless image as in the PW approach. This introduces an asymmetry for $\alpha > 0$ in the joint densities, making registration results inconsistent w.r.t. inversion. This asymmetry has a direct influence on the marginal densities, giving 3 different estimates of the marginal density: estimated from the histogram of a single image, or as the integral of either of the two joint histograms. That is, ignoring the scale parameters, the histograms, say, of J are given as,

$$h(j) = \int_\Omega P(J(\boldsymbol{x}) - j)\, dx, \tag{52}$$

and the corresponding marginal in the GPV approximation is found either as,

$$\tilde{h}(j) = \int_\Omega \int_\Gamma P(J(\boldsymbol{x}) - j)[P(I(\boldsymbol{x}) - i) * W(\boldsymbol{x})]\, di\, dx \tag{53}$$

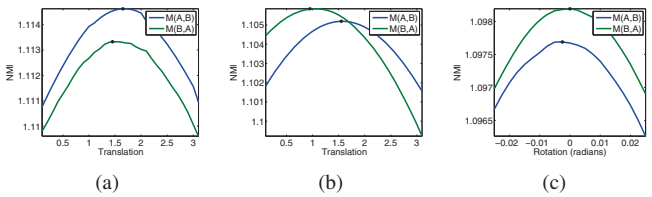
$$= \int_\Omega P(J(\boldsymbol{x}) - j) \int_\Gamma P(I(\boldsymbol{x}) - i) * W(\boldsymbol{x})\, di\, dx, \tag{54}$$

or as

$$h'(j) = \int_\Omega \int_\Gamma P(I(\boldsymbol{x}) - i)[P(J(\boldsymbol{x}) - j) * W(\boldsymbol{x})]\, di\, dx \tag{55}$$

$$= \int_\Omega \int_\Gamma P(I(\boldsymbol{x}) - i)\, di\; P(J(\boldsymbol{x}) - j) * W(\boldsymbol{x})\, dx. \tag{56}$$

The difference between these three estimates depends on the gradient of $I(\boldsymbol{x})$, and due to the scale of W, the gradient will differ for the two estimates based on the joint histograms. The asymmetry in GPV causes $\mathcal{M}(A, B) \neq \mathcal{M}(B, A)$. In the limit of $\alpha \to 0$, and when using identical kernels and parameters as Parzen windows for I and J, then GPV is symmetric, but, unfortunately, at the limit differentiability is lost, and



(a) (b) (c)

Fig. 7. GPV using NMI is asymmetric and has different optima, when comparing $\mathcal{M}(A,B)$ and $\mathcal{M}(B,A)$. Images compared are (a) two 3-dimensional Gaussians of standard deviation 5 and 11 under translation, (b) baseline and follow-up of patient number 16 from the OASIS collection [27] under translation, and (c) image 16 under rotation around the center. The optimum on each curve is denoted by a star.

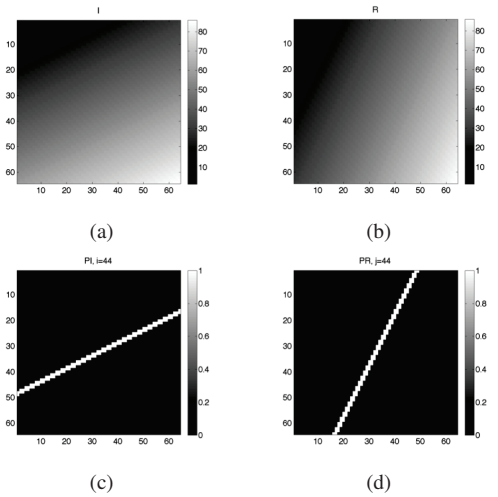
gradient-based optimizations schemes have to be abandoned. The consequence of the asymmetry in the estimate of the joint distribution will be investigated further in the following section.

V. Empirical investigations into the asymmetry in GPV

The asymmetry of GPV, i.e., $\mathcal{M}(A,B) \neq \mathcal{M}(B,A)$, has been analyzed in the previous section, and we will demonstrate that the asymmetry has not only theoretical, but also practical implications. We start by illustrating the asymmetry of GPV used for NMI. Fig. 7(a) shows $\mathcal{M}(A \circ \boldsymbol{\phi}, B)$ and $\mathcal{M}(B, A \circ \boldsymbol{\phi})$ for two 3-dimensional images of spatial Gaussian, with a standard deviation of 5 and 11, and centered in the middle of the image, sized $256 \times 256 \times 128$. We apply a translational motion, $\boldsymbol{\phi}$, one image w.r.t. the other along a fixed axis, and due to the symmetry of the Gaussians, the points of optima are nearly identical. However, on real medical images this is not the case: In Fig. 7(b), we have plotted the cost functional $\mathcal{M}(A \circ \boldsymbol{\phi}, B)$ and $\mathcal{M}(B, A \circ \boldsymbol{\phi})$ for a linear translation of two images, baseline and follow-up, of patient 16 from the OASIS collection [27]. The points of optima are clearly different. The asymmetry is also visible for rotational motion: Fig. 7(c) shows the asymmetry w.r.t. the rotation of patient 16 around the image center. The pattern is less pronounced, but it should be noted that even a small rotation around the center has a large and increasing displacement effect away from the center.

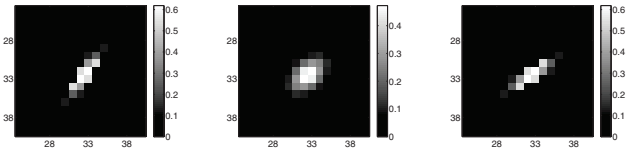
To further study the asymmetry of GPV using NMI empirically, we have constructed two images with a constant gradient, the same magnitude but different direction for each as shown in Fig. 8. We focus on a single isophote, $I(x,y) = I_0$ and $J(x,y) = J_0$, extracted using a Boxcar function. These are shown in Fig. 8(c) and 8(d). The value of the joint histogram for these intensities (I_0, J_0) is depicted in Fig. 9 as a function of space and using various estimation techniques. Fig. 9(a) shows the joint histogram's values when comparing Fig. 8(c) to Fig. 8(d) using GPV, i.e., where $I(x,y) = I_0$ is smoothed and intersected with $J(x,y) = J_0$ as $M(J,I)$ in GPV, and Fig. 9(c) shows the opposite case, $M(I,J)$. For reference, Fig. 9(b) shows the LOI estimate of the intersection of isophote I_0 and J_0. As can be observed, the spatial distribution of intensities is oriented according to the non-smoothed isophote.

Curvature adds further asymmetry, since the mass of the isophote moves in the direction of the center of the osculating



(a) (b) (c) (d)

Fig. 8. Two artificially generated images (a) and (b) with the same gradient magnitude, but different directions and the corresponding single isophotes (c) and (d) extracted using a Boxcar function.



(a) (b) (c)

Fig. 9. The GPV approximation is asymmetric. Image (a) is $\mathcal{M}(A,B)$, and (c) is $\mathcal{M}(B,A)$. Image (b) depict what (a) and (c) are approximating.

circle, when smoothed spatially. Thus, unless the two images curve in the exact same manner, the asymmetric smoothing of the GPV method will introduce further asymmetry in the similarity measure. This asymmetry is illustrated in Fig. 10, where an isophote is first extracted using a Boxcar function. The extracted isophote is then smoothed spatially, yielding the image in Fig. 10(a). This is compared with an isophote extracted as a soft isophote, as shown in Fig. 10(b). It can be seen that the images differ, especially where isophotes have high curvature. To substantiate this qualitative conclusion, we have conducted the following experiment: For a fixed image, an image of a given isophote is extracted using the 2 different methods: 1) PW as a soft isophote with fixed width $\beta_{\text{PW}} = 0.005$, and 2) GPV as an isophote extracted using a Boxcar with varying width β_{GPV} followed by spatial smoothing with a Gaussian of varying width α. Thus, for a fixed image with PW isophote width β_{PW}, we have searched for the values of β_{GPV} and α such that they minimize the Sum of Squared Differences between the two isophote images shown in Fig. 10(c). Notice in particular, that the difference between the two images of the isophotes is greatest near high curvature of the original isophote. In Fig. 10(d) is shown the result of finding optimal α for a given isophote extracted using the two methods and for varying β's. We conclude that there does not seem to be a simple relation across β's.

To empirically evaluate the degree of asymmetry as a function of α, we have conducted the following experiment: For 10 baseline and follow-up images from [27], we have

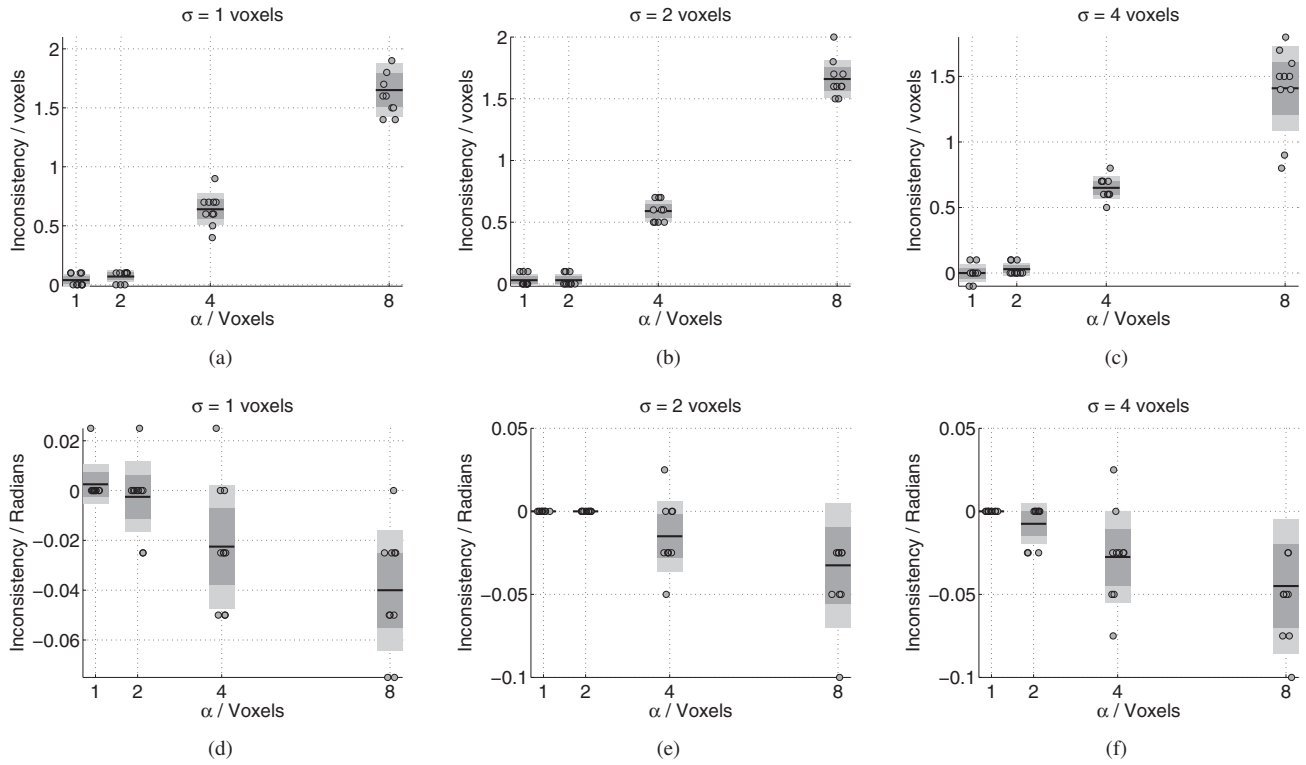


Fig. 11. GPV using NMI gives inconsistent optimization results for a simple, artificial translation (a-c), and rotation (d-f), and the inconsistency depends linearly on α but not on σ. For each boxplot, the circles represent individual measurement with slight noise added in the horizontal direction for legibility, the black line denotes the mean, the dark and light gray areas denote the 50% and 75% fractiles.

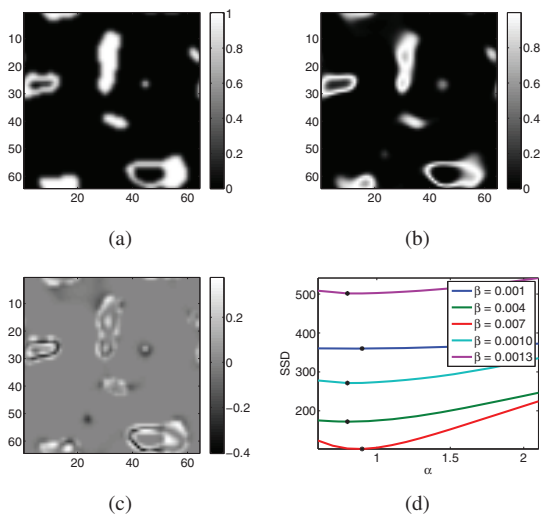


Fig. 10. The difference between smoothing Boxcar isophotes and soft isophotes appears near points of high isophote curvature. (a) The GPV isophote using $\beta_{\text{GPV}} = 0.0013$, smoothed with W using $\alpha = 0.9$. (b) The PW isophote using $\beta_{\text{PW}} = 0.005$, (c) the signed difference of (a) and (b), and (d) the SSD or MISE for a range of α and β_{GPV} using a start to denote optimum for each curve.

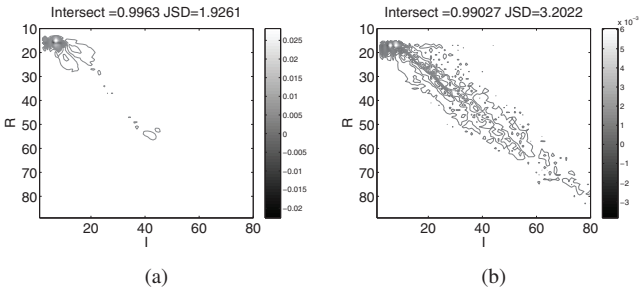


Fig. 12. Examples of the asymmetry of GPV in the estimated densities: We have subtracted the joint density distribution estimated in $\mathcal{M}(A, B)$ from the estimated density distribution in $\mathcal{M}(B, A)$. The figures show the difference between densities at (a) $\sigma = 1$, which is found to have a Jensen-Shannon divergence of 1.9, and at (b) $\sigma = 4$, which has a Jensen-Shannon divergence of 3.2.

rigidly registered the baseline and follow-up pair using NMI and GPV with a very small α, and then for a range of αs measured the spatial asymmetry in the similarity measure along the x-axis, caused by the increase in α. This is repeated for a range of σ values. The result is illustrated in Fig. 11. The experiment reveals that smoothing of the image does not eliminate the problem, and as our investigations show, asymmetry persists over all image scales. The asymmetry can also be observed in the joint density estimates. In Fig. 12 is shown the difference between the joint density used to evaluate $\mathcal{M}(B, A)$ and $\mathcal{M}(B, A)$ for 2 different values of α. The difference is seen to be non-negligible for both scales, and thus cannot be ignored.

To summarize, the GPV is asymmetric, and the degree of asymmetry increases proportionally to the curvature of the isophotes, as well as to α. The asymmetry cannot be alleviated using image smoothing, and we conclude that GPV does not offer inverse consistent registration.

VI. Empirical comparison of PW and GPV by Scales

The main difference between GPV and PW is the explicit modeling of the intensity coherence: where PW enforces coherence by Gaussian smoothing, GPV does not. In the following and using NMI, we will empirically evaluate and compare PW and GPV in terms of scales, i.e., the influences of the different kernels on the similarity measure, NMI, and the estimated joint density distribution, to provide intuition about the influence of different scales on NMI. Two types of algorithms for GPV and PW have been implemented: A fast cubic uniform B-spline approach (hereafter referred to as B-spline), which is described and analyzed in the next section, and a version based on Gaussian kernels. For a direct comparison of B-splines and Gaussians we have estimated the variance of a B-spline to be $\sigma \approx 0.6$. This allows us to investigate the effect of tuning the standard deviations of each of the kernels for both PW and GPV. We note here that some computational restrictions imposed on GPV are due to computational complexity, thus a Gaussian with local support has been used, i.e., very small values are truncated. We have performed intra-subject registration using rigid registration on a series of T1-weighted MRI of the human brain for 10 different subjects [27]. For each subject a follow-up image is registered rigidly to the baseline, such that the pair of the two volumes are aligned for a given set of scales. For a given direction (x-axis) we have translated one of the two with ± 1.5 voxels in steps of 0.1 voxel, and calculated the NMI similarity. This has been repeated for a wide range of kernels in the different spaces, i.e., different σ, β, and α including our fast B-spline-based algorithm for 10 different subjects.

A. Spatial scale, σ

When registering images, most algorithms exploit the scale-space of the images by smoothing the image with the kernel K. The idea is to capture large-scale structures of the images, so as to get closer to the optima before switching scales, in order to capture the structure at a finer scale. The actual influence on the different similarity measures has only been vaguely investigated in the literature. In spite of this uncertainty, smoothing the images is an often-used technique, and it has been empirically shown to yield good results, e.g., in [28]. We have examined the effect of image smoothing on NMI, and the results can be seen in Fig. 13 for PW and GPV respectively. Furthermore, Fig. 14 shows the estimated joint probability distribution for both PW and GPV. As can be seen, the distribution is more concentrated in a smaller area and NMI increases, when σ is large. The figures indicate that PW in general has a more pronounced peak than GPV for NMI, and that the optima is not shifted much over scales for this particular set of T1-weighted MRI of brains and using NMI.

B. Intensity scale, β

The intensity scale controls the resolution in the intensity domain, and as PW is a smoothing kernel in the intensity domain, the entropy is increased [29] proportionally to β. The

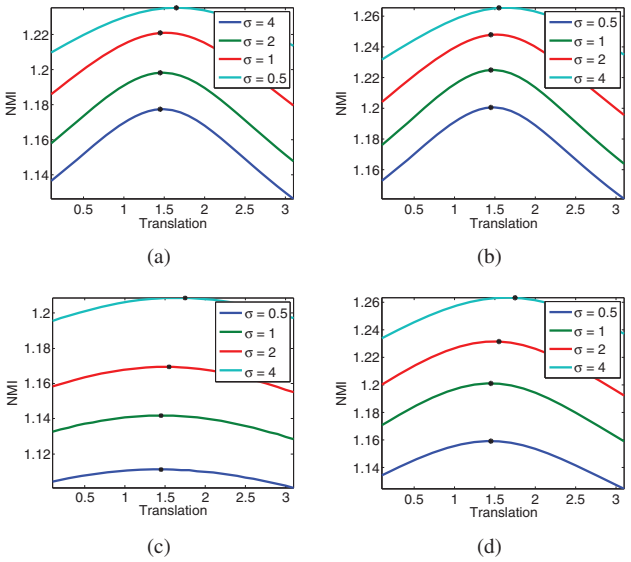


Fig. 13. The effect of image smoothing on the objective function (NMI) using the different density estimation schemes: (a) The PW using a Gaussian kernel $\beta = 0.6$, (b) PW using a cubic b-spline, and (c) GPV using a Gaussian $\alpha = 0.6$, (d) GPV using a cubic B-spline.

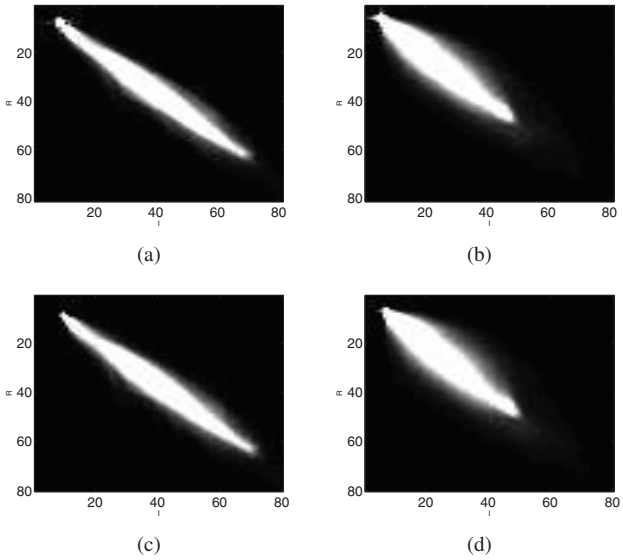
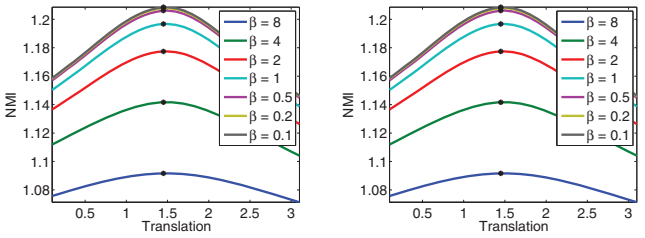


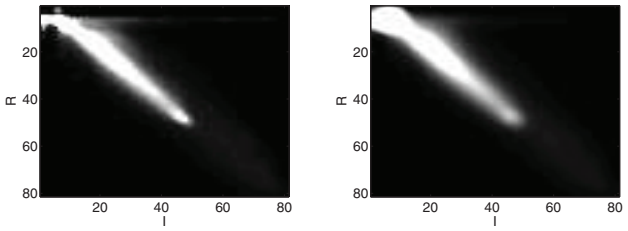
Fig. 14. The effect of image smoothing on the joint density using different estimation schemes: (a) & (b) The PW using $\beta = 0.3$, (a) $\sigma = 0.5$, and (b) $\sigma = 2$; (c) & (d) GPV using $\alpha = 0.6$, $\beta = 0.3$, and (c) $\sigma = 0.5$ and (d) $\sigma = 2$.

smoothing disperses the densities within the joint density, thus decreasing the overall NMI scores, as can be seen in Fig. 15. The effect of β on the joint density is illustrated in Fig. 16. As expected, the joint histogram becomes smoother as β is increased. The consequence of increasing β is that small scale changes in the image become negligible (see Section III-B), whereas large changes are preserved, i.e., putting more emphasis on large gradients with increasing β. We have not included GPV in this experiment; however, GPV also has an intensity scale, i.e., the width of its Boxcar function.



(a) (b)

Fig. 15. The similarity measure PW and NMI as a function of translation (a), and rotation (b), for a number of β values. The optimum on each curve is denoted by a star.

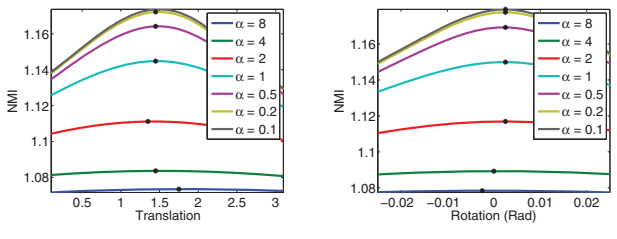


(a) (b)

Fig. 16. The effect of image smoothing with PW on the joint density estimate: the PW using $\sigma = 1$, and (a) $\beta = 0.5$, and (b) $\beta = 2$.

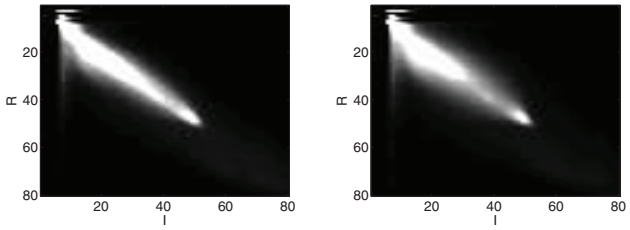
C. Integration scale, α

The kernel W can be used to describe local density estimates such as local MI or NMI [14], where each local histogram has its own NMI functional as in (4). PW is the special case of LOI, where $\alpha \to \infty$, and is thus a global density estimate, whereas GPV is an integration of local densities to become global. GPV uses a Boxcar function for P and smoothes the isophotes with W, as illustrated in Fig. 9(a) and 9(c). The effect of varying α on NMI using GPV is shown in Fig. 17. It is seen that NMI decreases and becomes more dispersed as α is increased. Comparison with Fig. 15 reveals that the effect of α on GPV is similar to the effect of β on PW: it reduces the function value due to the dispersion effect. Our theoretical investigation has revealed that smoothing is performed asymmetrically for GPV, and this is illustrated in Fig. 18, where we see horizontal dispersion but no particular vertical dispersion visible in the upper left corner. Previous empirical investigations [6] using the same B-spline kernel as PW β and partial volume α, reported that PW is more



(a) (b)

Fig. 17. The effect of varying α on the NMI functional using GPV with $\sigma = 0.2$ as a function of translation (a), and rotation (b). The optimum on each curve is denoted by a star.



(a) GPV (b) GPV

Fig. 18. The effect of the integration scale on the joint density estimate for GPV and NMI using $\sigma = 1$: (a) $\alpha = 0.5$ and (b) $\alpha = 2$.

precise, and that GPV has a larger convergence radius. From our experiments it is obvious that this difference is merely a consequence of the additional smoothing introduced by W as discussed in Section III-B. This is supported by Fig. 13: As can be seen, the PW is significantly more peaked than the GPV, which appears superficially to be a smoothed version of PW.

VII. Fast Implementations

We use a quasi-Newton gradient descent algorithm for optimizing (1). This results in a very fast and general algorithm that with only a few changes, works for many different loss-functions.

In order to use quasi-Newton methods for optimization, we need to derive the gradient of (1) w.r.t. the parameters of the uniform cubic B-spline, $\mathbf{\Phi}$. We use the notation of differentials, $\mathrm{d}g(x) = \mathrm{D}g(x)\,\mathrm{d}x$, and D is the partial derivative operator. Note that $\mathrm{d}\boldsymbol{x}$ is a vector of differentials, not the hypercube of its elements $d\boldsymbol{x}$, as in the case of integration. Further, we will only write up non-zero terms that depend on $\mathrm{d}\mathbf{\Phi}$. The differential of (1) is,

$$\mathrm{d}\mathcal{E} = \mathrm{d}\mathcal{M} + \mathrm{d}\mathcal{S}, \tag{57}$$

where arguments have been omitted for brevity. Ignoring the regularization term, we focus on the differential of the similarity measures. For (5), the differential is found to be,

$$\mathrm{d}\mathcal{M}_{\text{lin}} = \int_{\Gamma^2} F(i,j)\mathrm{d}h_{I,J}\,di\,dj, \tag{58}$$

under the mild Leibnitz integration rule, and where

$$\mathrm{d}h = \mathrm{d}\big(P(I(\boldsymbol{x})-i)P(J(\boldsymbol{x})-j) * W(\boldsymbol{x})\big) \tag{59}$$

$$= \big(\mathrm{D}P(I(\boldsymbol{x})-i)\mathrm{d}I\big)P(J(\boldsymbol{x})-j) * W(\boldsymbol{x}), \tag{60}$$

avoiding irrelevant arguments for brevity. In contrast, the differential of (2) is $\mathrm{d}\mathcal{M}_\Omega = \int_\Omega \mathrm{D}F(\boldsymbol{x}, I(\boldsymbol{x}), J(\boldsymbol{x}))\mathrm{d}I(\boldsymbol{x})\,d\boldsymbol{x}$, where smoothness typically is imposed on F and/or I. In comparison, our formulation (5) naturally allows for the added smoothing in intensity and integration space, and replaces technical difficulties in evaluating $\mathrm{D}F$ with $\mathrm{D}h$. One advantage is thus that it becomes easier to compare loss-functions directly. For (4) the differential is found to be,

$$\mathrm{d}\mathcal{M} = \int_{\Gamma^2} \mathrm{D}F(\boldsymbol{x}, h_{I,J}(\boldsymbol{x},i,j))\,\mathrm{d}h_{I,J}(\boldsymbol{x},i,j)\,d\boldsymbol{x}\,di\,dj, \tag{61}$$

using the mild Leibnitz integration rule. As shown in Section VII, the form of (61) suggests only a slight computational overhead as compared to (58). The derivatives for a range of F's are given in [30].

Using Leibnitz integration rule, the differentials of the distributions are given as

$$\mathrm{d}p_I(i,\boldsymbol{\Phi}) = \frac{1}{|\Omega|}\int_\Omega \mathrm{d}p_I(i|\boldsymbol{x},\boldsymbol{\Phi})\,d\boldsymbol{x}, \tag{62}$$

$$\mathrm{d}p_I(i|\boldsymbol{x},\boldsymbol{\Phi}) \simeq \frac{\mathrm{d}h_I(i,\boldsymbol{x},\boldsymbol{\Phi})}{\int_\Gamma h_I(j,\boldsymbol{x},\boldsymbol{\Phi})dj} - \frac{h_I(i,\boldsymbol{x},\boldsymbol{\Phi})\int_\Gamma \mathrm{d}h_I(j,\boldsymbol{x},\boldsymbol{\Phi})dj}{\left(\int_\Gamma h_I(j,\boldsymbol{x},\boldsymbol{\Phi})dj\right)^2}, \tag{63}$$

$$\mathrm{d}h_I(i,\boldsymbol{x},\boldsymbol{\Phi}) = \left(\mathrm{d}P(I(\boldsymbol{x},\boldsymbol{\Phi},\sigma)-i,\beta) * W(\boldsymbol{x},\alpha)\right), \tag{64}$$

where irrelevant arguments have been omitted for brevity. Likewise, we have:

$$\mathrm{d}p_{I,J}(i,j) = \frac{1}{|\Omega|}\int_\Omega \mathrm{d}p_{I,J}(i,j|\boldsymbol{x})\,d\boldsymbol{x}, \tag{65}$$

$$\mathrm{d}p_{I,J}(i,j|\boldsymbol{x}) \simeq \frac{\mathrm{d}h_{I,J}(i,j,\boldsymbol{x})}{\int_{\Gamma^2} h_{I,J}(k,l,\boldsymbol{x})\,dk\,dl} - \frac{h_{I,J}(i,j,\boldsymbol{x})\int_{\Gamma^2}\mathrm{d}h_{I,J}(k,l,\boldsymbol{x})\,dk\,dl}{\left(\int_{\Gamma^2} h_{I,J}(k,l,\boldsymbol{x})\,dk\,dl\right)^2}, \tag{66}$$

$$\mathrm{d}h_{I,J}(i,j,\boldsymbol{x}) = \left(\mathrm{d}P(I(\boldsymbol{\psi},\boldsymbol{\Phi},\sigma)-i,\beta)\,P(J(\boldsymbol{\psi},\sigma)-j,\beta)\right) * W(\boldsymbol{x}-\boldsymbol{\psi},\alpha). \tag{67}$$

In the context of Locally Orderless Images (LOI), GPV can be derived as follows:

$$\mathrm{d}h_I = \mathrm{d}\left(P(I(\boldsymbol{x},\boldsymbol{\Phi},\sigma)-i,\beta) * W(\boldsymbol{x},\alpha)\right) \tag{68}$$

$$= P(I(\tilde{\boldsymbol{x}},\boldsymbol{\Phi},\sigma)-i,\beta) * \left(\mathrm{D}_{\boldsymbol{x}} W(\boldsymbol{x},\alpha)\right), \tag{69}$$

and the differential w.r.t. $\boldsymbol{x}$ is found to be,

$$\mathrm{d}h_{I,J}(i,j,\boldsymbol{x},\alpha,\beta,\sigma) = P(J(\boldsymbol{x},\sigma)-j,\beta)\Big(\left(P(I(\boldsymbol{\phi}(\tilde{\boldsymbol{x}}),\sigma)-i,\beta)\right) * \left(\mathrm{D}_{\boldsymbol{x}} W(\boldsymbol{x},\alpha)\right)\Big). \tag{70}$$

In Fig. 19 is shown the pseudocode for the Sum of Squared Differences, using a spatial integration (SSD), the Parzen window approximation of the general sum of p-norms (PNORM), and the Parzen window and Generalized partial volume approximation of Normalized Mutual Information (PW and GPV). Binary code interfacing to Matlab is available [31]. The code assumes 3D images, cubic B-splines for all kernels, and M bins in the histograms. We assume that today's processors have equal processing time, e.g., of sum, log, sin etc. From the pseudocode in Fig. 19 and its notes, we see that PW and GPV have almost identical computational complexity. Results by actual implementations may vary, but in general the computation of NMI, using either GPV or PW, appears to be about as complex as SSD using B-splines. W.r.t. memory, GPV requires $192 \times N \times 8$ bytes of memory to obtain the speed, where the PW only requires $8 \times N \times 8$ bytes (on 64-bit, double precision).

```
# Given 2 images, I and J, and the determinant of the
# transformation, det, as a function of space,
# calculate PW for NMI and PNorm, GPV for NMI and
# SSD, based on N image evaluation points, and
# M marginal and M^2 joint histogram bins. Flops are
# based on cubic B-splines

FOR N evaluation points
  calculate image spline coeff.
  (60 flops)
  IF(SSD || PW || PNorm)
    calculate derivative of image spline coeff.
    (48 flops)
  FOR 64 combinations of image spline coeff.
    IF(SSD || PW || PNorm)
      update image at evaluation point
      (4 flops)
      update image gradient at evaluation point
      (12 flops)
    IF(GPV)
      update histograms
      (4 flops)
  IF(SSD)
    update residual
    (2 flops)
  IF(PW || PNorm)
    calculate histogram spline coeff.
    (20 flops)
    FOR 16 histogram spline coeff.
      IF(PNorm)
        compute P-norm
        update residual
        update derivative
        (5 flops)
      ELSE
        update histograms
        (2 flops)
IF(PW || GPV)
  calculate NMI and derivative on histograms
  (9*M^2+6M flops)
  FOR N evaluation points
    IF(GPV)
      calculate derivative of image spline coeff.
      (48 flops)
      FOR 64 combinations of image spline coeff.
        update derivative of histogram
        (16 flops)
    IF(PW)
      FOR 16 histogram spline coeff.
        update derivative of histogram
        (9 flops)
  update derivatives
  (3 flops)
# Total flop usage:
# SSD:     1134N flops
# PW:      1331N +9M^2 +6M flops
# PNorm:   1379N flops
# GPV:     1383N +9M^2 +6M flops
```

Fig. 19. Pseudocode for SSD, NMI using PW and GPV, and P-Norm using PW.

# samples	Similarity measure	SSD	PW
1000000	Avg. execution time (in sec)	1.21	1.63
	Relative exec. time to SSD	1	1.34
	Theoretic relative exec. time to SSD	1	1.17
	Overhead	1	1.13

TABLE II
THE TABLE SHOWS THE AVERAGE EXECUTION TIME ACROSS 100 FUNCTION EVALUATIONS OF SSD VS PW-NMI FOR 1000000 POINTS USING 256 BINS.

To substantiate our theoretical computation we have performed some empirical experiments. First we note that the overhead of GPV and PW is in general small. The histogram calculations will only dominate in the special case of a small number of samples and many histogram bins. We have compared computational complexity empirically for PW and GPV registration and SSD. We use cubic B-spline for K, P, and W, and histograms with 256 bins for marginal histograms and 256^2 for the joint histograms. We perform the computations on a laptop with i7-core Q820 (Quad-core) operating at 1.7 GHz and 12 GB shared memory. All similarity measures have been implemented in parallel using the Intel Threading Building Blocks library. As the code runs multi-threaded, we believe that most of the 13% overhead seen in Table II comes from the threads, which are initialized twice as many times in PW as in SSD. Our results are valid for the general algorithm but not for massive parallelism e.g. on GPU. However, it clearly demonstrates that NMI and MI no longer should be considered as severe bottlenecks in image registration. Furthermore, thread blocking can cause further latency during histogram update, thus the estimated times for single threaded implementation are very close to our estimate for large N.

A. A non-rigid registration example

To show that the computational framework is capable of performing registration, we have included a small example from the OASIS longitudinal database. We registered, in 3D, a baseline with a follow-up using the Parzen window NMI as described in this paper with 128×128 bins in joint histogram. We use a uniform B-splines deformation representation with a hyper elastic prior [20] with a node distance of 10 voxels. The results are seen in Fig. 20. The sample density is every second voxel in each direction.

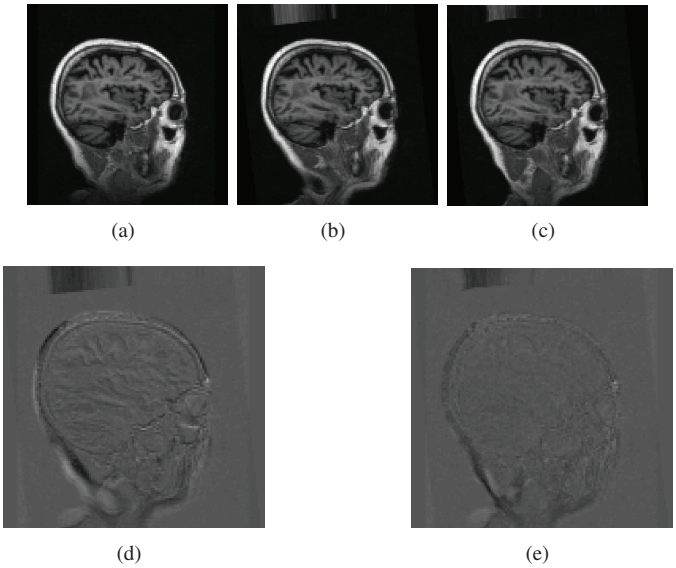


Fig. 20. Non-rigid registration in 3D using our framework with NMI and a cubic B-spline deformation model. Sagittal sections of (a) the baseline image and the follow-up image (b) rigidly registered and (c) non-rigidly registered. (d) and (e) show the difference between (a) and (b) and (a) and (c) respectively.

VIII. CONCLUSION

We have introduced Locally Orderless Registration, a framework that encompasses most of the currently used similarity measures. Our framework allows us to divide a wide range of similarity measures into 3 categories from simple global linear measures, such as the P-norm or Huber norm over non-linear global measures, such as Correlation Coefficient, Mutual Information and Normalized Mutual Information to position dependent schemes, such as Correlation Ratio and spatially encoded Mutual Information. All of these measures, or any combination thereof, are formulated in a scale-space over measurement, intensity, and integration space, offering the flexibility to easily create application-specific similarity measures in a smooth formulation well suited for gradient-based schemes. We have presented a thorough analysis of the scales in the different spaces both theoretically, through the moments of the density distribution and a simple local image model, and through rigorous empirical experiments.

We have extended our previous work [7] on the difference between Parzen window and the Generalized partial volume. Our analysis clearly shows that Generalized partial volume is an asymmetric density estimator not suited for problems that require inverse consistency. We have shown that depending on the smoothing, this error can become larger than a single voxel. Generalized partial volume achieves its computational speed by making an approximation to the local histogram and by using 0-order B-spline as the Parzen estimator. In [6] it is reported that the Parzen window is more accurate than Generalized partial volume for kernels W with $\alpha > 0$, and we show that this is due to the difference in smoothing, and not to the properties of the two density estimators. Worse still, Generalized partial volume measures the dissimilarity of the images at two different scales, and thus the effect becomes more pronounced with increased α - histograms of larger areas.

We have given an efficient implementation of LOR, and our theoretical as well as empirical analysis of the computational and storage complexity demonstrate that the Parzen window is more attractive for intensity-based registration.

We believe that the choice of density estimator should be based on the particular application. Generalized partial volume may be preferred for cases, where intensities in the two images are incoherent. However, if intensity images are to be registered, and computational efficiency or inverse consistency is a desired property, then our analysis reveals that the Parzen window is a far more attractive density estimator in comparison to Generalized partial volume.

ACKNOWLEDGMENTS

Sune Darkner would like to thank the Oticon foundation for supporting his work through the project grant "A Fast and Personalized Biomechanical Model".

REFERENCES

[1] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal, "Automated multi-modality image registration based on information theory," *Information Processing in Medical Imaging*, pp. 263–274, 1995.

[2] W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, vol. 1, no. 1, pp. 35–51, 1996.

[3] C. Studholme, D. Hill, and D. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognition*, vol. 32, no. 1, pp. 71–86, 1999.

[4] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *Medical Imaging, IEEE Transactions on*, vol. 16, no. 2, pp. 187–198, 1997.

[5] H. Chen and P. Varshney, "Mutual information-based CT-MR brain image registration using generalized partial volume joint histogram estimation," *Medical Imaging, IEEE Transactions on*, vol. 22, no. 9, pp. 1111–1119, 2003.

[6] D. Loeckx, F. Maes, D. Vandermeulen, and P. Suetens, "Comparison between parzen window interpolation and generalised partial volume estimation for nonrigid image registration using mutual information," *Biomedical Image Registration*, pp. 206–213, 2006.

[7] S. Darkner and J. Sporring, "Generalized partial volume: An inferior density estimator to parzen windows for normalized mutual information," in *Information Processing in Medical Imaging*, ser. LNCS, G. Szakely and H. Hahn, Eds., vol. 6801. Springer, 2011, pp. 436–447.

[8] L. Griffin, "Scale-imprecision space," *Image and Vision Computing*, 1997.

[9] J. Koenderink and A. Van Doorn, "The structure of locally orderless images," *International Journal of Computer Vision*, vol. 31, no. 2, pp. 159–168, 1999.

[10] J. Pluim, J. Maintz, and M. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE transactions on medical imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[11] M. Modat, G. Ridgway, Z. Taylor, M. Lehmann, J. Barnes, D. Hawkes, N. Fox, and S. Ourselin, "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine*, 2009.

[12] P. Thevenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *Image Processing, IEEE Transactions on*, vol. 9, no. 12, pp. 2083–2099, 2000.

[13] F. Maes, D. Vandermeulen, and P. Suetens, "Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information," *Medical Image Analysis*, vol. 3, no. 4, pp. 373–386, 1999.

[14] G. Hermosillo, C. Chefd'Hotel, and O. Faugeras, "Variational methods for multimodal image matching," *International Journal of Computer Vision*, vol. 50, no. 3, pp. 329–343, 2002.

[15] J. Pluim, J. Antoine Maintz, and M. Viergever, "Interpolation artefacts in mutual information-based image registration," *Computer vision and image understanding*, vol. 77, no. 2, pp. 211–232, 2000.

[16] E. Haber and J. Modersitzki, "Intensity gradient based registration and fusion of multi-modal images," *Methods of information in medicine*, vol. 46, no. 3, pp. 292–299, 2007.

[17] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, and P. Suetens, "Nonrigid image registration using conditional mutual information," in *Proceedings of the 20th international conference on Information processing in medical imaging*. Springer-Verlag, 2007, pp. 725–737.

[18] X. Zhuang, S. Arridge, D. Hawkes, and S. Ourselin, "A nonrigid registration framework using spatially encoded mutual information and free-form deformations," *IEEE Transactions on Medical Imaging*, 2011.

[19] A. Rajwade, A. Banerjee, and A. Rangarajan, "Probability density estimation using isocontours and isosurfaces: applications to information-theoretic image registration," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 3, pp. 475–491, 2009.

[20] S. Darkner, M. Hansen, R. Larsen, and M. Hansen, "Efficient hyperelastic regularization for registration," *Image Analysis*, pp. 295–305, 2011.

[21] G. Christensen, R. Rabbitt, and M. Miller, "Deformable templates using large deformation kinematics," *Image Processing, IEEE Transactions on*, vol. 5, no. 10, pp. 1435–1447, 1996.

[22] S. Sommer, M. Nielsen, F. Lauze, and X. Pennec, "A Multi-Scale kernel bundle for LDDMM: towards sparse deformation description across space and scales," in *IPMI 2011*. Springer, 2011.

[23] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, October 1948.

[24] L. Florack, R. Maas, and W. Niessen, "Pseudo-linear scale-space theory," *International Journal of Computer Vision*, vol. 31, no. 2/3, pp. 247–259, 1999.

[25] J. Bruce and P. Giblin, *Curves and Singularities*. Cambridge University Press, 1992.

[26] L. Griffin, "Histograms of infinitesimal neighbourhoods," in *Scale-Space and Morphology in Computer Vision*, ser. Lecture Notes in Computer Science, M. Kerckhove, Ed. Springer Berlin / Heidelberg, 2006.

[27] D. Marcus, T. Wang, J. Parker, J. Csernansky, J. Morris, and R. Buckner, "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[28] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes., "Non-rigid registration using free-form deformations: Application to breast mr images." *IEEE Transactions on Medical Imaging*, vol. 18(8), pp. 712–721, 1999.

[29] J. Sporring and J. Weickert, "Information measures in scale-spaces," *IEEE Trans. on Information Theory*, 1999.

[30] J. Sporring and S. Darkner, "Jacobians for lebesgue registration for a range of similarity measures," Department of Computer Science, University of Copenhagen, Tech. Rep. 4, 2011.

[31] S. Darkner and J. Sporring, "Locally orderless registration code : Fast normalized mutual information and p-norm for matlab," http://curis.ku.dk/ws/files/38136594/LOI.zip, 2012.

Sune Darkner received his Master's Degree in Applied Mathematics in 2004 and founded a software company building databases for the telecommunication industry. In collaboration with Oticon A/S, a large hearing aid manufacturer, he received his Industrial Ph.D. degree in "Shape and Deformation Analysis of the Human Ear Canal" in 2009, from the Department of Informatics and Mathematical Modelling, at the Technical University of Denmark (DTU). After holding a position at an energy company as data analyst, he rejoined the Department of Informatics and Mathematical Modelling at DTU in 2009 as a post doc. He currently holds a position as Assistant Professor in image analysis at the Department of Computer Science, University of Copenhagen. Research interests include image registration and classification, optimization and regularization, and computational physics.

Jon Sporring received his Master and Ph.d. degrees from the Department of Computer Science, University of Copenhagen, Denmark in 1995 and 1998, respectively. He completed part of his doctoral studies in the U.S., at the IBM Research Center at Almaden, California, USA. On completing his Ph.D., he was a visiting researcher at the Computer Vision and Robotics Lab at the Foundation for Research & Technology – Hellas, Greece, and as an Assistant Research Professor at 3D-Lab, School of Dentistry, University of Copenhagen. Since 2003 he has been employed as Associate Professor at the Department of Computer Science, University of Copenhagen. From 2007-2012 he served as Vice-Chair for Research at Department of Computer Science, and from 2008-2009 he was a part-time Senior Researcher at Nordic Bioscience. In 2012-2013 he was a visiting professor at the School of Computer Science, McGill University in Montreal, Canada. His primary research fields are Computer Science, particularly image processing, computer graphics, information theory, and pattern recognition.