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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

The Atlas Structure of Images

Lewis D. Griffin

Abstract— Many operations of vision require image regions to be isolated and inter-related. This is challenging when they are different in detail and extent. Practical methods of Computer Vision approach this through the tools of downsampling, pyramids, cropping and patches. In this paper we develop an ideal geometric structure for this, compatible with the existing scale space model of image measurement. Its elements are apertures which view the image like fuzzy-edged portholes of frosted glass. We establish containment and cause/effect relations between apertures, and show that these link them into cross-scale atlases. Atlases formed of Gaussian apertures are shown to be a continuous version of the image pyramid used in Computer Vision, and allow various types of image description to naturally be expressed within their framework. We show that views through Gaussian apertures are approximately equivalent to the jets of derivative of Gaussian filter responses that form part of standard Scale Space theory. This supports a view of the simple cells of mammalian V1 as implementing a system of local views of the retinal image of varying extent and resolution. As a worked example we develop a keypoint descriptor scheme that outperforms previous schemes that do not make use of learning.

Index Terms—Image Analysis, Image Representation, Image Resolution, Gaussian Derivatives, Filter Steering, Keypoints.

1 Introduction

Consider a scene (Figure 1a) containing two objects (faces) which are intrinsically similar but, because they are at different distances, manifest in the image data quite differently [1]. A vision system should have scale covariance [2] so that it can assess the similarity despite the different image appearances. For this it has to access and inter-relate image regions of different extent and level of detail. The full set of image regions of different extent and detail, and their inter-relations, has a structure something like a geographical atlas [3].

The atlas idea is familiar in computer vision. It can be implemented using an image pyramid where a stack of images, of reducing size, is formed by repeated 2×2 pixel averaging. Regions can be defined at any level of the pyramid as a set of pixels, typically square, and it is straightforward to say when a region at a fine level stands in a cause/effect relation with a region at a coarse level. A pyramid structure could be applied to Figure 1a as follows. Within the pyramid for the full image, there would be found a sub-pyramid with base (say) 512×512 covering the near face, and a sub-pyramid with base (say) 32×32 covering the far face. The coarser levels of the near-face sub-pyramid would contain very similar pixel values to the far face sub-pyramid.

Image pyramids work quite well in practice but with two problems. First that the detail changes between levels can be too large. For example if the far face extends over a 24×24 area there will not be a really good match in the near face sub-pyramid. Second that repeated 2×2 averaging only approximates the way detail disappears with increased viewing distance. Both of these problem are solved by the Scale Space framework [1, 4-7], which represents an image at different levels of detail using a continuous family of images rather than a discrete set, and uses Gaussian blurring to generate those levels rather than 2×2 averaging. Gaussian blurring correctly infers the image that would be acquired if the scene were more distant, under the reasonable assumption that the spatial sensitivity of the imaging sensors has a Gaussian form.

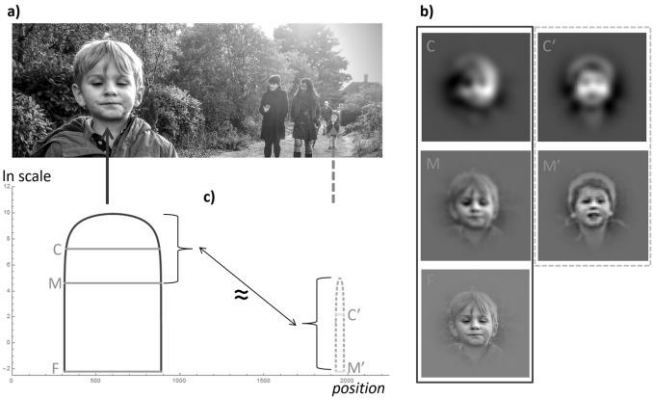

Fig 1. **The need for an atlas structure. a)** A visual system should be able to assess the similarity of the two objects (faces) indicated despite their difference in position, extent and visible detail. **b)** For that it needs to isolate and compare sub-regions of the image at different levels of scale. **c)** These sub-regions at different scales are related within the Scale Space of the image (*x* is image position, *s* is scale). The views of a region at different scales form slices of an atlas. The two faces are similar in the sense that the atlas of one is similar to a coarse segment of the atlas of the other (see section 6) i.e. C is similar to C' and M to M'.

The shift from 2×2 averaging to Gaussian blurring creates a complication when regions and their inter-relations are considered. In an image pyramid it is cut-and-dried what region of pixels at a fine level influence the pixels of a region at a coarse level, so it is straightforward to define when one region contains another, or when one region

• *L.D. Griffin is with the Dept of Computer Science, University College London, London WC1E 6BT. E-mail: l.griffin@cs.ucl.ac.uk*

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

causes another, even if the regions are at different levels of the pyramid. In scale space, however, the infinite support of Gaussian blurring kernels means that each value at a coarser scale is, in theory, dependent on the entire image at finer scales. So the definitions of containment and causation between regions, and indeed the definition of a region, are not obvious.

The aim of this paper is to present a coherent motivated system of regions and their interrelations for the scale space framework, together defining a continuous atlas structure for images (figure 1b,c). We believe that this conceptual framework can assist in the development of improved computer vision algorithms, just as scale space was influential on SIFT [8]; in support of this we present such a development for keypoint description.

We preview the paper. In section 2 we review scale space. In 3 we introduce apertures, each defined by a spatial weighting function and an associated scale, as a definition of a region. These apertures provide views of the image as through fuzzy portholes of frosted glass. In 4 we propose that one aperture should be considered to contain another if the view through the contained is stably determined by the view through the container. We formally characterize 'stable determination' in terms of reducing image norms relative-to-apertures. In 5 we define a pair of apertures to stand in a cause-effect relationship if the cause contains the effect and is as small as possible; or, equivalently, the effect is contained in the cause and is as large as possible. We discover that cause-apertures are Gaussian blurs of effect-apertures, with the amount of blur being the difference in scale between the cause and effect. Notice that the blur of the aperture from effect to cause operates in the opposite direction to the blur of the scale space. In 6 we show that the causation relation assembles apertures into 1-D families we call atlases; and we show that atlases formed of Gaussian apertures (Figure 1b), which we call Gaussian atlases (Figure 1c), are fundamental. In 7 we review measurement of image structure by derivative-of-Gaussian (DtG) filters, producing a jet of filter responses. In 8 we show that jets are to Gaussian apertures, as crops are to pyramids - they are a compact record of the view of an image through an aperture. In 9 we discuss the structure of Gaussian atlases. In section 10 the framework inspires a novel keypoint descriptor that outperforms previous non-learnt descriptors. In 11 we summarize and discuss issues arising.

1.1 Formalism

We present the theory for 1-D images, but generalisation to higher dimensions is straightforward because of the separability of the Gaussian and its derivatives [1]. Some figures show 1-D images, others 2-D. For clarity, we use Ω for the image spatial domain, rather than $\mathbb{R}$, and refer to it simply as the domain. When a variable e.g. $x \in \Omega$ is introduced we assume its type and any restrictions apply in the remainder. Functions of the domain are bolded and italicized (e.g. $\boldsymbol{G}$). $\boldsymbol{\delta}$ is the delta function at the origin; $\boldsymbol{\delta}_x$ at x. $*$ is used for convolution; and $\times$ for multiplication, where it aids readability. We use square parentheses for ordered pairs e.g. $\mathcal{A} := [\boldsymbol{A}, \alpha]$ is a generic aperture consisting of the pairing of a weighting function $\boldsymbol{A}: \Omega \to \mathbb{R}^+$ and a scale $\alpha \geq 0$.

Some frequently used notations will be:

- the 1-norm for integrals of functions i.e. $\|\boldsymbol{A}\|_1 := \int |\boldsymbol{A}|$
- $\boldsymbol{G}_s$ a 1-D Gaussian function of scale s; with its order n derivative $\boldsymbol{G}_s^{(n)}$ being called a DtG.
- $\vec{j}_s^n(\boldsymbol{I})$ the vector of DtG responses up to order n to an image $\boldsymbol{I}$, called a jet.
- $\mathcal{F} := [\boldsymbol{F}, f]$, $\mathcal{C} := [\boldsymbol{C}, c]$, $c > f$ generic fine and coarse scale apertures; with $\boldsymbol{G} := \boldsymbol{G}_{c-f}$ the Gaussian which effects the image blur to move between their *viewed* scales.
- $\mathcal{G}(w,s) := [\boldsymbol{G}_w, s]$ a Gaussian aperture of width w, for viewing the image at scale s.

We will make frequent use of inner products (IPs) which are maps from pairs of vectors (e.g. images) to a scalar value $\langle \_, \_ \rangle : V \times V \to \mathbb{R}$ that is symmetric in its arguments, linear in each, and positive-definite i.e. $\langle \vec{v}, \vec{v} \rangle \geq 0$, with equality if and only if $\vec{v} = \vec{0}$. An IP induces a norm $\|\vec{v}\|^2 := \langle \vec{v}, \vec{v} \rangle$, which measures magnitudes and so can be used to measure distances (i.e. $d(\vec{u}, \vec{v}) := \|\vec{u} - \vec{v}\|$, and angles $\cos\theta_{uv} := \|\vec{u}\|^{-1} \|\vec{v}\|^{-1} \langle \vec{u}, \vec{v} \rangle$. We use different styles of parentheses for different types of inner product (IP):

- angled, for the standard L_2 IP e.g. $\langle \boldsymbol{I}, \boldsymbol{J} \rangle := \int \boldsymbol{I}\boldsymbol{J}$
- rounded, for IPs relative to an aperture e.g. $(\boldsymbol{I}, \boldsymbol{J})_{\mathcal{A}} := \langle \boldsymbol{A}, \boldsymbol{I}_\alpha \boldsymbol{J}_\alpha \rangle$
- fences, for IPs of jets at a scale e.g. $\left| \vec{j}, \vec{k} \right|_s := \sum_{0 \leq i \leq n} (2s)^i (i!)^{-1} j_i k_i$

2 Scale Space

If the function $\boldsymbol{I}: \Omega \to \mathbb{R}$ represents the *ideal image* falling on the signal transduction surface, then its scale space is the 1-D family of functions $\boldsymbol{I}_s: \Omega \to \mathbb{R}$, $s \in \mathbb{R}^+$ generated by blurring it with Gaussian kernels ($\boldsymbol{G}_s(x) := (4\pi s)^{-1/2} e^{-x^2/(4s)}$) of increasing width $\boldsymbol{I}_s := \boldsymbol{G}_s * \boldsymbol{I}$. The ideal image is at the base of the Scale Space i.e. $\lim_{s \to 0} \boldsymbol{I}_s = \boldsymbol{I}$ because $\lim_{s \to 0} \boldsymbol{G}_s = \boldsymbol{\delta}$. The scale parameter $s \geq 0$ has dimension length-squared and is half the variance of the Gaussian. This parameterization allows compact statements of (i) scale similarity $\boldsymbol{I}_{s+t} = \boldsymbol{G}_s * \boldsymbol{I}_t$, and (ii) that scale space satisfies the heat equation $(\partial_s - \partial_{xx}) \boldsymbol{I}_s(x) = \boldsymbol{0}$. The theory of scale space was definitively expounded in [1]; earlier statements and alternative derivations are

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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

reviewed in [9, 10]; and the theory is generalized in [11-14]. Figure 2 illustrates the Scale Space of an example image, using parameterization of scale by $\ln s$ which reveals scale similarity.

Convolution by Gaussian kernels is a convenient way to express scale space and an efficient way to implement it digitally. An equivalent formulation is as the complete set of measurements of the image obtained by computing its IPs with Gaussian filters of every size at every image position (i.e. $\boldsymbol{I}_s(y) = \langle \boldsymbol{G}_s(x-y), \boldsymbol{I}(x) \rangle$). This formulation makes clear the status of scale space as a model of biological vision: individual filters correspond to individual V1 simple cell neurons; and measurements to neural responses [15-17].

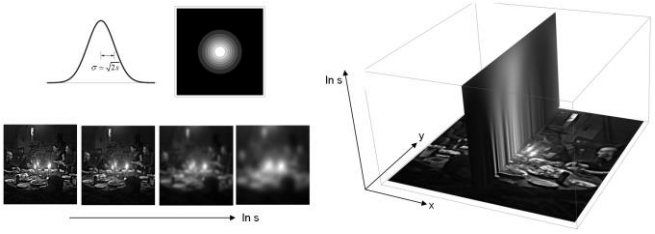

Fig 2. Illustrates Scale Space. Top-left: Gaussian kernels (1-D and 2-D) used to produce levels of scale space by convolution. Sections through the scale space of an example image: horizontal (bottom-left) and vertical (right).

3 APERTURES

We distinguish between apertures and patches. An aperture is an operator for isolating a particular image region. The fundamental operation that an aperture must support is the computation of an image IP relative to it. A patch is a record of the view of an image 'through' an aperture. They can be efficiently stored to allow computation of IPs without access to the entire image.

Figure 3 shows patches from three types of aperture. The top row are the simplest type, square crops from an image: like views through clear glass windows. Moving from top row to middle, the aperture has been changed from square to circular, and the extraction has been performed on an intermediate level of scale space: the windows have become portholes, and the glass has become frosted. Moving from middle row to bottom, the aperture has been changed to a fuzzy Gaussian weighting function: the frosted glass portholes now have a fuzzy edge, something similar being used for aesthetic reasons in modern vehicle windows.

The traditional 'crop-type' aperture can be characterized by the subset of the domain ($A \subseteq \Omega$) extracted. The high-frequency border of such apertures can result in the extracted patch changing abruptly as the aperture or image is translated. This problem has been addressed in diverse domains of signal analysis by generalizing the characterization of apertures as domain subsets, via discontinuous indicator functions (i.e. $\boldsymbol{1}_A(x) \coloneqq [x \in A]$, using the Iverson bracket), to a characterization as non-negative weighting functions [18, 19]. Let $\boldsymbol{A}: \Omega \to \mathbb{R}^+$ be a generic non-negative weighting function. Diverse forms for $\boldsymbol{A}$ have been proposed, typically continuous and bell-shape; and in computer vision methods using a scale space framework, Gaussian windows have been found effective [5, 20-23].

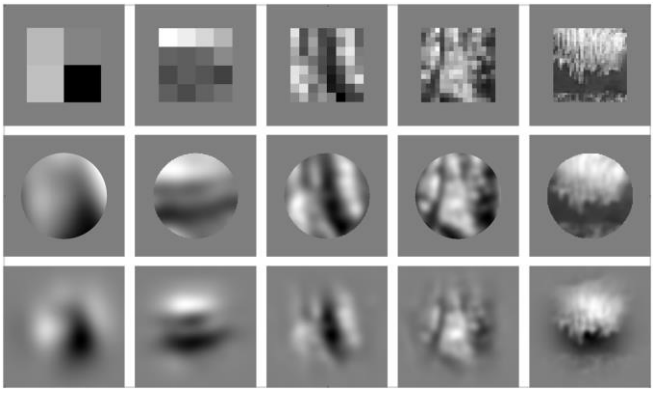

Fig 3. Illustrates different types of patch. All patches are from an image of a woodland scene. Each row is a different type of patch. Patches in the same column are roughly matched in extent and articulation which increases across columns from left to right. Top: 'Traditional' patches - cropped from an image. Middle: Aperture-based patches - using circular region apertures applied to a level of scale space. Bottom: As middle but positive weighting function apertures. The weighting functions are Gaussians, which are argued in section 6 to be fundamental.

In this work we adopt the characterisation of image apertures as a positive weighting function with an associated scale e.g. $\mathcal{A} \coloneqq [\boldsymbol{A}, \alpha]$. We will refer to apertures as coarse or fine in reference to the value of α, and large or small in reference to the extent of $\boldsymbol{A}$.

Before proceeding, we note an oddity with this characterization. Since weighting functions are not constrained, for example, to unit weight; a weighting function and a multiple of it define distinct apertures. Whereas, intuitively, they might be expected to have the same view of the image, thus define the same aperture. The advantage of our characterization is that it allows a criterion for aperture containment that performs as expected. We have been unable to find a criterion that performs as well, if we require a weighting function and a multiple of it to define the same aperture.

We define the image IP relative to $\mathcal{A}$ as $(\boldsymbol{I}, \boldsymbol{J})_{\mathcal{A}} \coloneqq \langle \boldsymbol{A}, \boldsymbol{I}_\alpha \boldsymbol{J}_\alpha \rangle$. This definition trivially satisfies the symmetry and linearity requirements for an IP, but positive-definiteness holds if and only if $\boldsymbol{A}$ is nowhere zero; if it does have zeros then the IP is improper which is harmless. The induced norm with respect to an aperture is $\|\boldsymbol{I}\|_{\mathcal{A}}^2 \coloneqq (\boldsymbol{I}, \boldsymbol{I})_{\mathcal{A}} = \langle \boldsymbol{A}, \boldsymbol{I}_\alpha^2 \rangle$, and from this distances and angles can be defined as described earlier.

4 CONTAINMENT

For image pyramids, built with 2×2 averaging, containment is clear-cut and intuitive: a region at some

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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

level of the pyramid is contained within some other region if the pixel values within the former can be perfectly computed from those within the latter. So for example, an 8×8 patch of pixels at one level contains the 4×4 patch of pixels with the same support at the next level, or the 2×2 at the level after that.

For scale space, where infinite-support Gaussian blurring is used instead of finite-support averaging, the criterion for containment is less obvious. We start by advancing an informal characterization:

If image differences can be seen as well through an aperture $\mathcal{A}$ as they can through an aperture $\mathcal{B}$ then $\mathcal{B} \subseteq \mathcal{A}$

We then adapt ideas from well-posedness and stability analysis [24, 25] to formalize this as:

$$\mathcal{B} \subseteq \mathcal{A} \quad \Leftrightarrow \quad \forall \boldsymbol{X}, \boldsymbol{Y} \;\; \|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathcal{B}} \leq \|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathcal{A}}$$

Equivalently, and more simply:

$$\mathcal{B} \subseteq \mathcal{A} \quad \Leftrightarrow \quad \forall \boldsymbol{Z} \;\; \|\boldsymbol{Z}\|_{\mathcal{B}} \leq \|\boldsymbol{Z}\|_{\mathcal{A}}$$

Summarized as:

aperture norms decrease with respect to containment.

This definition is consistent with the intuitive containment relation for image pyramids as the variance of a set of pixel values is always more than the variance of their averages across non-overlapping subsets.

To explore our norm-reduction definition we start by considering the infinite-support apertures with constant unit-value weight $\mathcal{Z}_s := [\boldsymbol{1}, s]$. The norm of an image $\boldsymbol{I}$ with respect to such is $\|\boldsymbol{I}\|^2_{\mathcal{Z}_s} = \|\boldsymbol{I}_s\|^2 = \left\langle e^{-2\omega^2 s}, \left|\hat{\boldsymbol{I}}(\omega)\right|^2 \right\rangle$, where $\hat{\boldsymbol{I}}$ is the Fourier Transform of $\boldsymbol{I}$. Since $e^{-2\omega^2 s}$ decreases with increasing s for all ω, so does the norm whatever $\boldsymbol{I}$. Hence $c \geq f \;\Leftrightarrow\; \mathcal{Z}_c \subseteq \mathcal{Z}_f$. This demonstrates that containment *is* possible in a scale space setting.

Next we consider arbitrary fine $\mathcal{F} := [\boldsymbol{F}, f]$ and coarse $\mathcal{C} := [\boldsymbol{C}, c]$ apertures; $f < c$. These apertures, which I stress are not assumed to be of gaussian form, will be used repeatedly in the remainder. They view onto the scales f and c of the scale space of an image ($\boldsymbol{I}$). For compactness, the blurring operator that changes between these scales will be written $\boldsymbol{G} := \boldsymbol{G}_{c-f}$, so $\boldsymbol{I}_c = \boldsymbol{G} * \boldsymbol{I}_f$.

Let $\boldsymbol{E}_\omega(x) := \cos(\omega x)$, $\boldsymbol{O}_\omega(x) := \sin(\omega x)$ be sinusoidal images. Given that $\boldsymbol{G}_f * \boldsymbol{E}_\omega = e^{-\frac{1}{2} f \omega^2} \boldsymbol{E}_\omega$ and similarly for $\boldsymbol{O}_\omega$, we can deduce that $\|\boldsymbol{E}_\omega\|^2_{\mathcal{F}} + \|\boldsymbol{O}_\omega\|^2_{\mathcal{F}} = e^{-f\omega^2}\|\boldsymbol{F}\|_1$ and similarly for $\mathcal{C}$. Additionally, since $f < c$, for sufficiently large $\bar{\omega}$ it must be that $e^{-f\bar{\omega}^2}\|\boldsymbol{F}\|_1 > e^{-c\bar{\omega}^2}\|\boldsymbol{C}\|_1$. Putting those together, we deduce $\|\boldsymbol{E}_{\bar{\omega}}\|^2_{\mathcal{F}} + \|\boldsymbol{O}_{\bar{\omega}}\|^2_{\mathcal{F}} > \|\boldsymbol{E}_{\bar{\omega}}\|^2_{\mathcal{C}} + \|\boldsymbol{O}_{\bar{\omega}}\|^2_{\mathcal{C}}$, which means that either $\|\boldsymbol{E}_{\bar{\omega}}\|_{\mathcal{F}} > \|\boldsymbol{E}_{\bar{\omega}}\|_{\mathcal{C}}$ or $\|\boldsymbol{O}_{\bar{\omega}}\|_{\mathcal{F}} > \|\boldsymbol{O}_{\bar{\omega}}\|_{\mathcal{C}}$. Whichever is true it demonstrates that $\mathcal{F} \not\subseteq \mathcal{C}$; and, since no restrictions were put on $\boldsymbol{F}, \boldsymbol{C}$, it follows that that a *finer aperture is never contained within a coarser* (recalling that 'fine' and 'coarse' refer to the level of scale space that the apertures view, not to the extent of the apertures).

This allows us to establish that containment is a partial order. Reflexivity and transitivity are trivial. For anti-symmetry we require that $\mathcal{A} \subseteq \mathcal{B} \;\wedge\; \mathcal{B} \subseteq \mathcal{A} \;\Rightarrow\; \mathcal{A} = \mathcal{B}$. Given that we have shown that a coarse aperture cannot contain a fine, the premise can only be true if the apertures are equal scale, and anti-symmetry is then trivially true.

Focussing on the case of a coarse aperture potentially contained within a finer, to evaluate the requirement that aperture norms decrease with respect to containment, we consider the generalized Rayleigh quotient $\|\boldsymbol{Z}\|_{\mathcal{C}} / \|\boldsymbol{Z}\|_{\mathcal{F}}$ and its maximum value $\lambda_{\mathcal{F},\mathcal{C}}$. Containment can be expressed in terms of this maximum value i.e. $\mathcal{C} \subseteq \mathcal{F} \;\Leftrightarrow\; \lambda_{\mathcal{F},\mathcal{C}} \leq 1$. Standard theory [26], and some re-arrangement, gives that $\lambda_{\mathcal{F},\mathcal{C}}$ is the largest eigenvalue of $\boldsymbol{F}^{-1} \times \left(\boldsymbol{G} * \left(\boldsymbol{C} \times \left(\boldsymbol{G} * \_\right)\right)\right)$; which, by the Perron-Frobenius Theorem [27], is paired with the unique positive eigenvector ($\boldsymbol{Y}$). Multiples of $\boldsymbol{Y}$ are the images whose norm fractionally decreases the least as one changes from $\mathcal{F}$ to $\mathcal{C}$. In summary:

$$\forall \mathcal{F}, \mathcal{C} \quad \exists \boldsymbol{Y} > 0, \lambda_{\mathcal{F},\mathcal{C}} > 0 \quad \lambda_{\mathcal{F},\mathcal{C}} \boldsymbol{F}\boldsymbol{Y} = \boldsymbol{G} * \left(\boldsymbol{C} \times \left(\boldsymbol{G} * \boldsymbol{Y}\right)\right)$$

A special case of containment is when $\lambda_{\mathcal{F},\mathcal{C}} = 1$, in which case we call the containment tight and denote it $\mathcal{C} \subseteq_T \mathcal{F}$. Any reduction of a tightly containing aperture, or any expansion of a tightly contained aperture, will break the containment i.e.

$$\mathcal{C} \subseteq_T \mathcal{F} \quad \Rightarrow \quad \forall \boldsymbol{\varepsilon} > 0 \;\; \mathcal{C} \not\subseteq [\boldsymbol{F} - \boldsymbol{\varepsilon}, f] \;\wedge\; [\boldsymbol{C} + \boldsymbol{\varepsilon}, c] \not\subseteq \mathcal{F}$$

The eigen-condition, though formally correct, is not a practical test for deciding whether a particular fine aperture contains a particular coarse, as it in general requires a challenging computation. However, we can use it to generate fine apertures that tightly contain a particular coarse by picking the image whose norm is not reduced by moving between the apertures i.e. given any $\boldsymbol{Y} > 0$ then $\mathcal{C} \subseteq_T \left[\boldsymbol{Y}^{-1} \times \left(\boldsymbol{G} * \left(\boldsymbol{C} \times \left(\boldsymbol{G} * \boldsymbol{Y}\right)\right)\right), f\right]$. Similarly, we can generate coarse apertures that are tightly contained within any fine: given any $\boldsymbol{Q} > 0$ then $\left[\boldsymbol{Q} / \left(\boldsymbol{G} * \left((\boldsymbol{G} * \boldsymbol{Q}) / \boldsymbol{F}\right)\right), c\right] \subseteq_T \mathcal{F}$. These constructions show, perhaps unexpectedly, that at each finer scale there is an infinity of apertures tightly-containing any given coarse, and at each coarser scale an infinity of apertures tightly contained within a fine (Fig 4). This does not occur with discrete pyramid structures.

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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

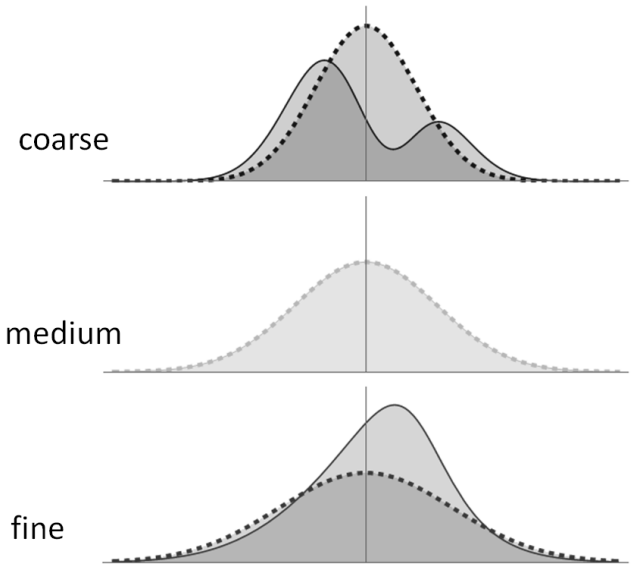

Fig 4. Illustrates tight containment of apertures. The apertures on the coarse row are associated with a strong blur; on the medium row with a medium blur; and on the fine row with a smaller blur. Both coarse apertures are tightly contained within the medium aperture, which is in turn tightly contained within both the fine apertures. Containment means that the view through the contained aperture is determined by the view through the container. The containment relation is *tight* when the containing aperture cannot be reduced in any way, nor the contained aperture be expanded in any way, without it causing the relation to fail. The dotted apertures are gaussians, the solid are not – illustrating that tight containment can hold between apertures of either type.

5. CAUSATION

Suppose a face has been viewed through a fine aperture. Recording the view will prepare the system to re-identify the person should they reappear at the same distance. To be ready to re-identify them if they reappear at greater viewing distances the system needs to record the view through some effect aperture at each coarser scale. Similarly, a vision system might make a candidate detection of some object through a coarse aperture; perhaps a bright blob has been seen that may be a face. It would then wish to examine the same region of the image through a finer cause aperture to test the detection.

Which aperture $\mathcal{F}$ of scale f should be chosen as the cause of a coarse effect aperture $\mathcal{C}$? Informally, the cause should be large enough to contain the effect, otherwise it will miss details that give rise to coarsely visible features; but it should not be larger than it needs to be, so that it views a minimum of additional structure that would need to be matched in future presentations.

We wish to determine the intersection of these two constraints – large enough, but not larger than needed. The first is easy to characterize: the cause aperture should contain the effect aperture i.e. $\mathcal{F} \supseteq \mathcal{C}$. For the second constraint we need a measure of aperture size which combines extent and amplitude, and captures how much structure an aperture can view. For this we propose the L1-norm of the weighting function. This is a simple choice that seems reasonable; for example, it is proportional to the expected squared norm of a random white noise signal $\boldsymbol{W}$ i.e.

$$\mathrm{E}\left[\left\|\boldsymbol{W}\right\|_{\mathcal{F}}^2\right] \propto \int_{x\in\Omega} \left\|\boldsymbol{\delta}_x\right\|_{\mathcal{F}}^2 = \int_{x\in\Omega} \left\langle \boldsymbol{F}, \left(\boldsymbol{G}_f * \boldsymbol{\delta}_x\right)^2 \right\rangle = \left\|\boldsymbol{F}\right\|_1 .$$

Next we observe that $\mathcal{C} \subseteq \mathcal{F} \Rightarrow \left\|\boldsymbol{C}\right\|_1 \le \left\|\boldsymbol{F}\right\|_1$ which follows easily from $\mathcal{C} \subseteq_T \mathcal{F} \Rightarrow \left\|\boldsymbol{C}\right\|_1 \le \left\|\boldsymbol{F}\right\|_1$, which we show starting from the eigen-condition for tight containment.

$$\begin{aligned}\left\|\boldsymbol{F}\right\|_1 &= \left\|\boldsymbol{Y}^{-1} \times \left(\boldsymbol{G} * \left(\boldsymbol{C} \times \left(\boldsymbol{G} * \boldsymbol{Y}\right)\right)\right)\right\|_1 \\ &= \left\langle \boldsymbol{C}, \left(\boldsymbol{G} * \boldsymbol{Y}\right)\left(\boldsymbol{G} * \boldsymbol{Y}^{-1}\right)\right\rangle \ge \left\|\boldsymbol{C}\right\|_1\end{aligned}$$

where the rightmost inequality comes from application of the Cauchy-Schwartz inequality, with intermediate step $\left\langle \boldsymbol{G} * \boldsymbol{\delta}_y, \boldsymbol{Y} \right\rangle \left\langle \boldsymbol{G} * \boldsymbol{\delta}_y, \boldsymbol{Y}^{-1} \right\rangle \ge 1$. Cauchy-Schwartz also shows that the inequality will only be an equality (the minimum $\left\|\boldsymbol{F}\right\|_1$ condition) when $c^{-1}\boldsymbol{G} * \boldsymbol{Y} = c\boldsymbol{G} * \boldsymbol{Y}^{-1}$, which is true only if $\boldsymbol{Y} = c$; and when it does the eigen-condition simplifies to $\boldsymbol{F} = \boldsymbol{G} * \boldsymbol{C}$. So in conclusion:

> the minimum of $\left\|\boldsymbol{F}\right\|_1$, subject to $\mathcal{F} \supseteq \mathcal{C}$, is uniquely achieved by $\boldsymbol{F} = \boldsymbol{G} * \boldsymbol{C}$.

This result shows that, according to the criteria we have argued for, the cause of an effect aperture is given by the blur of the effect aperture by a Gaussian of scale equal to the difference in scale between the cause and effect. Note that since the cause is at a finer scale than the effect this blurring operates in the opposite direction to that for the scale space image i.e. $\boldsymbol{F} = \boldsymbol{G} * \boldsymbol{C}$ vs. $\boldsymbol{I}_c = \boldsymbol{G} * \boldsymbol{I}_f$. In figure 5 a cause-effect pair of apertures are illustrated.

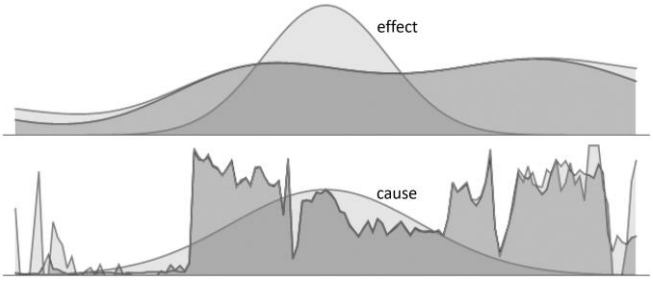

Figure 5 – Cause-effect apertures. The apertures (unimodal cruves) are in cause and effect relation. The cause aperture is the blur of the effect aperture. Two 1-D images (dark and lighter) are shown at the scales the apertures view. Observe how the image blur increases going upwards in the figure, while the aperture width increases going downwards. The two images have been chosen so that they are very similar within the view of the cause aperture; consequently they are at least as similar within the view of the upper aperture.

6. ATLASES

The causal relation between apertures is transitive; and distinct apertures have distinct causes (i.e. $\boldsymbol{C}_1 \ne \boldsymbol{C}_2 \Leftrightarrow \boldsymbol{G} * \boldsymbol{C}_1 \ne \boldsymbol{G} * \boldsymbol{C}_2$). Therefore the relation partions the set of all possible apertures into non-intersecting 1-D families which we call atlases in allusion to geographic map collections; particularly those kind which start with coarse scale maps, followed by increasingly finer scale maps of the regions covered by the coarser.

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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

Every atlas has the same structure - $\{[\boldsymbol{G}_s * \boldsymbol{Z}, c-s]\}_{0\le s\le c}$ - a closed interval family of apertures whose weighting functions are blurs of the coarsest aperture ($\boldsymbol{Z}$). $\boldsymbol{Z}$ can only be the top of an atlas if it cannot be even infinitesimally deblurred to a positive function, since that would then be the top. Thus atlases are topped by apertures whose weighting function has zero values and/or insufficiently rapid Fourier energy decay.

Of special interest are those atlases topped by delta functions at some scale *t*. Since all the finer apertures in such an atlas are gaussians, and gaussians apertures occur only in such atlases, we call them Gaussian Atlases (fig 6). Gaussian apertures have been suggested as particularly effective and natural for scale space analysis [20-23]. We denote a Gaussian aperture as $\mathcal{G}(w,s) := [\boldsymbol{G}_w, s]$ and a Gaussian atlas as $\{\mathcal{G}(t-s,s)\}_{0\le s\le t}$.

Observe that the *sum* of the scale parameters for the aperture and the blur is constant throughout the atlas. This is because the blur relation amongst the apertures of the atlas runs in the opposite direction to the ordinary blur of scale space. Consequently, the combined effect of the image blur and the windowing has the same spatial support at all levels of the atlas. This is the same as for an image pyramid when the base is a square with side length a power of 2, but without that tricky detail.

While the sum of the aperture parameters is constant across the atlas, their *ratio* is not. So while each aperture is sensitive to the same image extent, the number of degrees of freedom which it sees it with varies across the atlas, just as in a pyramid.

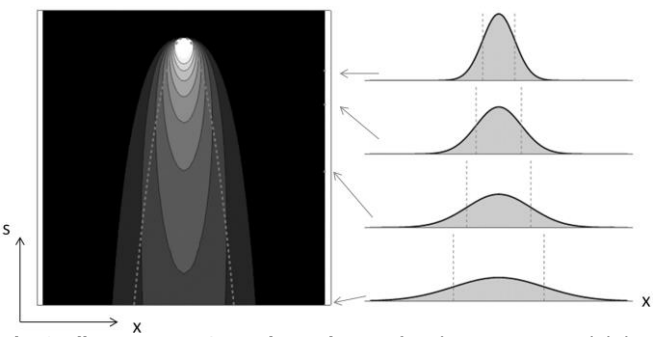

Fig 6. Illustrates a Gaussian atlas. Left**:** The aperture weighting functions of a gaussian atlas shown as a contour plot across scale space. The dashed curve marks the locus of aperture transitions between convex and concave (a convenient landmark). With increasing scale the apertures can be seen to get tighter, and the peak value increases. For improved visibility the grey-levels of the plot have been clipped at an upper threshold. Right**:** Example apertures from the atlas at a selection of scales as indicated by the arrows. Dashed lines, as at left, mark the transition between convex and concave regions.

We note that Gaussian atlases are special because:

- Amongst the atlases with a particular width (measured by aperture spatial variance) at some scale, the one with greatest extent over scale is Gaussian; so they allow tracking of a fine scale view to the coarsest scales.
- The Gaussian form for apertures is especially attractive, as it minimizes high frequency content for a fixed width (shown by the Fourier uncertainty principle), which means that views through them change as slowly as possible with their translation (useful for steerability in section 10).
- A vision system cannot be expected to directly implement all possible apertures. In such a case it may instead synthesize bespoke ones from a basis set. Since apodized IPs are linear in the aperture this synthesis is straightforward i.e.

  $$(\boldsymbol{I},\boldsymbol{J})_{[p\boldsymbol{A}+q\boldsymbol{B},s]} = p\times(\boldsymbol{I},\boldsymbol{J})_{[\boldsymbol{A},s]} + q\times(\boldsymbol{I},\boldsymbol{J})_{[\boldsymbol{B},s]}.$$

  To be uncommitted the basis set should be sufficient to generate all other apertures. The positive cone of the delta functions contains all positive functions, so these are sufficient.
- In section 8 we show that Gaussian apertures permit a simple effective analogue of the patch used in computer vision, allowing views through them to be efficiently stored and compared.

We can now give a specific answer to the puzzle problem in figure 1a. An ideal vision system would compute a separate Gaussian atlas for each point of Scale Space. One of these atlases (solid in figure 1c) views the near face, another (dashed in figure 1c) the distant. These atlases isolate the faces from the rest of the image, and coordinate views of their appearance at different scales. Apertures at matched scales of the two atlases show very similar views of the two faces. The distant face atlas matches a top portion of the near face atlas – this is the sense in which they appear similar. The lower segment of the near-face atlas shows detailed views of that face that are not available for the distant.

7. DtGs, Jets and the Hermite Transform

A scale space can also be constructed from responses to derivative of Gaussian (DtG) filters $\boldsymbol{G}_s^{(n)} := \left(\partial^n / \partial x^n\right)\boldsymbol{G}_s$, illustrated in figure 7a. Such a construction satisfies the scale similarity and heat equation properties [16, 28], but the base of the scale space is the derivative of the ideal image rather than the ideal image. Not only do DtG filters arise as the unique linear filters given the scale similarity constraint, they also provide a good fit to the receptive field profiles of V1 simple cells [16, 17, 29].

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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

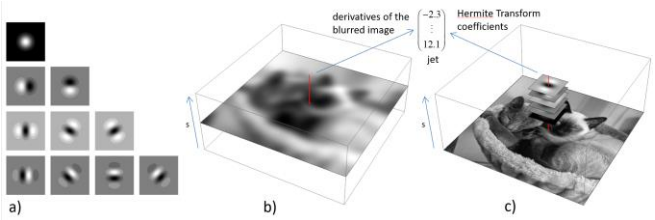

Figure 7 – The DtG model of image measurement. a) An equal-scale family of 2-D DtG filters of 0th (top row) to 3rd (bottom row) order. The family shown has, at each order, rotated copies of a single filter. This matches V1 simple cells, but the Cartesian basis derivatives typically implemented in Computer Vision are linearly equivalent [28]. **b)** The jet is equal to the point derivatives of the image blurred to the scale of the filters. **c)** The jet is equal to initial terms of the Hermite Transform of the image.

The response of a DtG filter is called a jet component $j_s^n(\boldsymbol{I}) \coloneqq \left\langle \boldsymbol{G}_s^{(n)}, \boldsymbol{I} \right\rangle$. DtGs can be considered non-infinitesimal derivative operators because jet components (with a sign change if odd order) are equal to the point derivative of the image blurred to the scale of the filter i.e. $j_s^n(\boldsymbol{I}) = (-1)^n \left(\partial^n / \partial x^n\right) \boldsymbol{I}_s \big|_{x=0}$ (Figure 7b). A jet is a vector of components from a family of DtG filters up to some order of differentiation i.e. $\vec{j}_s^n(\boldsymbol{I}) \coloneqq \left(j_s^0(\boldsymbol{I}) \quad \cdots \quad j_s^n(\boldsymbol{I}) \right)^{\mathrm{T}}$. A natural IP for jets, defined as $\left|\vec{u}, \vec{v}\right|_s \coloneqq \sum_n (2s)^n (n!)^{-1} u_n v_n$, has been derived [30, 31].

Jets can also be interpreted as truncated Hermite transforms [32, 33] (Figure 7c), the basis functions of which are the scaled Hermite polynomials, defined by $\boldsymbol{H}_w^n \coloneqq (-4w)^n \boldsymbol{G}_w^{(n)} / \boldsymbol{G}_w$. The scaled Hermite polynomials are orthogonal with respect to a Gaussian aperture with window scale w and blur scale 0 (denoted $\mathcal{G}(w,0)$ using the notation introduced in 6) i.e. $\left(\boldsymbol{H}_w^m, \boldsymbol{H}_w^n\right)_{\mathcal{G}(w,0)} = n! (8w)^n \delta_{mn}$. Jet components can be understood, not only as the IP between a DtG and the image, but also as the *aperture* IP between the image and a scaled Hermite polynomial i.e. $j_w^n(\boldsymbol{I}) = (-4w)^{-n} \left(\boldsymbol{H}_w^n, \boldsymbol{I}\right)_{\mathcal{G}(w,0)}$. Since the scaled Hermite polynomials are a complete orthogonal basis relative to $\mathcal{G}(w,0)$ [34], images can be expressed as a weighted sum of those polynomials, with the weights relating to jet components i.e. $\left\| \boldsymbol{I} - \sum_{n \geq 0} (-2w)^{-n} (n!)^{-1} j_w^n(\boldsymbol{I}) \boldsymbol{H}_w^n \right\|_{\mathcal{G}(w,0)}^2 = 0$. This linkage between jets and the Hermite Transform provides a clear justification for the jet IP previously proposed [30, 31] i.e. $\left|\vec{j}_w^\infty(\boldsymbol{I}), \vec{j}_w^\infty(\boldsymbol{J})\right|_w = (\boldsymbol{I}, \boldsymbol{J})_{\mathcal{G}(w,0)}$.

8. Jets as Patches

We have defined patches to be the view of an image through an aperture. With an image pyramid, patches are simply sub-arrays of pixel values cropped from some level of the pyramid. They are a perfectly efficient record of the view. In contrast, the aperture-based approach we have developed seems not to have such an efficient representation of views. In particular, if we consider $\boldsymbol{A}\boldsymbol{I}_\alpha$ to be the patch arising from viewing the image $\boldsymbol{I}$ through the aperture $[\boldsymbol{A}, \alpha]$ then that has the full dimensionality of a function over the image domain, and so has a huge memory footprint however small the aperture. However we will describe how jets can be considered as an alternative memory-efficient approach to patches for Gaussian apertures.

In the previous section we derived that for a Gaussian aperture with zero blur scale its IP was equal to the IP of infinite order jets measured with DtGs of scale that match the aperture. We can easily amend the formula to remove the restriction on the aperture having zero blur scale, obtaining $(\boldsymbol{I}, \boldsymbol{J})_{\mathcal{G}(w,s)} = \left|\vec{j}_{w+s}^\infty(\boldsymbol{I}), \vec{j}_{w+s}^\infty(\boldsymbol{J})\right|_w$. This shows that the infinite order jet measured with DtG filters of scale w, allows us to compute IPs for all the apertures in the Gaussian atlas $\{\mathcal{G}(w-s,s)\}_{0 \leq s \leq w}$ with height w i.e. $(\boldsymbol{I}, \boldsymbol{J})_{\mathcal{G}(w-s,s)} = \left|\vec{j}_w^\infty(\boldsymbol{I}), \vec{j}_w^\infty(\boldsymbol{J})\right|_{w-s}$.

So infinite order jets are a record of the view through a Gaussian aperture, but being infinite dimensional they are still impractical as a patch. One could just use finite order jets as patches and accept that they only approximate the view through the corresponding Gaussian aperture i.e. $\left|\vec{j}_{w+s}^n(\boldsymbol{I}), \vec{j}_{w+s}^n(\boldsymbol{J})\right|_w \approx (\boldsymbol{I}, \boldsymbol{J})_{\mathcal{G}(w,s)}$, but the quality of the approximation is variable: if s is small, w large, and n small it will be poor.

To get a better controlled approximation we advance the general statement

$$\left|\vec{j}_w^n(\boldsymbol{I}), \vec{j}_w^n(\boldsymbol{J})\right|_w \approx (\boldsymbol{I}, \boldsymbol{J})_{\mathcal{G}(w',s')}$$

i.e. the n-jet IP approximates some Gaussian aperture IP. We will argue for what the parameters (w', s') of the approximated aperture are, as a function of the jet order (n) and scale (w); and then confirm our approximation with numerical experiments.

We use two constraints to determine w', s'. First is that the approximating aperture should belong to the Gaussian atlas with top scale w – it certainly does as $n \to \infty$. This means that $w' + s' = w$. Second we try to maximize the similarity of the diagonals (relative to a delta function basis) of the two IPs. Specifically we compare $\|\boldsymbol{\delta}_x\|_{\mathcal{G}(w',s')}$ and $\left\|\vec{j}_w^n(\boldsymbol{\delta}_x)\right\|_w$. The first evaluates to $(8\pi s')^{-1/2} \boldsymbol{G}_{2w'+s'}(x)$ and the second to $(8\pi w)^{-1/2} \boldsymbol{G}_{w/2}(x) \sum_{0 \leq i \leq n} \left((8w)^i i!\right)^{-1} \left(\boldsymbol{H}_w^i(x)\right)^2$. The first is *exactly* Gaussian in form, the second *roughly*. We can make them similar by equalizing their variances: $\operatorname{var}_x\left(\|\boldsymbol{\delta}_x\|_{\mathcal{G}(w',s')}\right) = 2w' + s'$, $\operatorname{var}_x\left(\left\|\vec{j}_w^n(\boldsymbol{\delta}_x)\right\|_w\right) = \frac{4n+1}{2n+1} w$. Solving the two constraints – $w' + s' = w$ and $2w' + s' = \frac{4n+1}{2n+1} w$ –

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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

yields $w' = \frac{2n}{2n+1} w$, $s' = \frac{1}{2n+1} w$. So we arrive at the approximation:

$$\left| \vec{j}_s^n(\boldsymbol{I}), \vec{j}_s^n(\boldsymbol{J}) \right|_s \approx (\boldsymbol{I}, \boldsymbol{J})_{\mathcal{G}\left(\frac{2n}{2n+1}s, \frac{1}{2n+1}s\right)}$$

In words:

- a vision system, probing the image with DtG filters of scale s up to order n, senses the image like viewing it through Gaussian apertures of width $\frac{2n}{2n+1}s$ and blur scale $\frac{1}{2n+1}s$.

This approximation can also be read in the other direction to determine the scale and order of DtGs needed to approximate a given Gaussian aperture:

$$(\boldsymbol{I}, \boldsymbol{J})_{\mathcal{G}(w,s)} \approx \left| \vec{j}_{s+w}^{\left[\frac{w}{2s}\right]}(\boldsymbol{I}), \vec{j}_{s+w}^{\left[\frac{w}{2s}\right]}(\boldsymbol{J}) \right|_{s+w}$$

(where $[\_]$ is the rounding function). Figure 8a,b show a Gaussian aperture and an approximately equivalent DtG filter system.

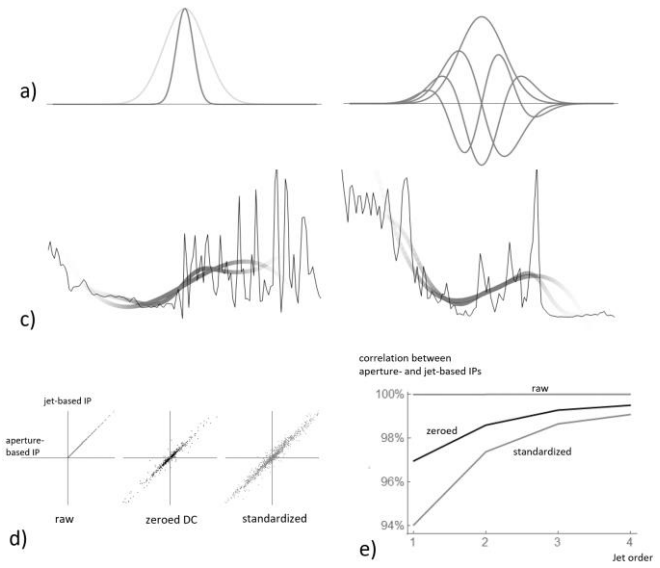

Figure 8 – Gaussian aperture approximated equivalent to a DtG family. **a)** A gaussian aperture (pale) and the blur kernel of its associated scale (darker). **b)** The system of DtG filters up to 3rd order that approximate the view through (a). **c)** Each plot shows a natural image profile (thin), the view through the Gaussian aperture shown in (a) (thick), and the view according to the jet measured by the filters in (b) (thick). **d)** For pairs of natural image profiles the IPs computed by aperture (a) or by jet (b) are highly correlated. **e)** The correlation between the jet IP and the equivalent-aperture IP increases with jet order, for natural signals.

Figure 8c show that the views of two natural signals according to the Gaussian aperture of Figure 8a and the approximating jet of Figure 8b are similar. For a more quantitative evaluation of the approximation we performed an experiment using a thousand pairs of 1-D profiles extracted at random orientations from a database of natural images. We compared the IPs between pairs of these profiles computed using either finite order jets or approximately-equivalent Gaussian apertures. To make the comparison more exacting we repeated the experiment but using image profiles with their DC component subtracted so that their mean within the aperture was a zero. To make it still more exacting we repeated the experiment but using profiles that were standardized by a linear transformation that gave them zero mean and unit standard deviation within the aperture.

As shown in Figure 8d the correlation between the jet and aperture IPs for raw profiles approached 100%. This is because the variation about the mean of natural signals is typically small compared to the mean itself; so the DC component, which varies widely from profile to profile, is the primary determinant of either type of IP. The correlation between the two types of IP is still very high for the profiles with DC component removed: for the pair of IPs illustrated in 8a,b it is 99.3%. When the profiles are standardized, equating their contrast as well as silencing their DC components, the correlation drops to a still high 98.7%. Figure 8e shows that the correlation between the IPs improves with jet order, which is not surprising given that we know that it becomes perfect as the order becomes infinite. In conclusion, the results of figure 8 suggest that the approximation $\left| \vec{j}_s^n(\boldsymbol{I}), \vec{j}_s^n(\boldsymbol{J}) \right|_s \approx (\boldsymbol{I}, \boldsymbol{J})_{\mathcal{G}\left(\frac{2n}{2n+1}s, \frac{1}{2n+1}s\right)}$ is sufficiently accurate, for natural image data, that jets can be considered as functionally equivalent to patches from Gaussian apertures.

9 Gaussian Atlases

We consider $\{\mathcal{G}(t-s,s)\}_{0 \le s \le t}$ as an example Gaussian atlas. Based on the approximation in section 8, the aperture at scale s within this atlas approximates the jet arising from measurement by a DtG filter family of scale t and order $n = \left[\frac{t-s}{2s}\right]$ i.e. coarser levels in the atlas are equivalent to increasingly truncated jets.

The width (measured as spatial standard deviation) of the aperture at scale s in the example atlas is $\sqrt{2(t-s)}$. Plotting this using a simple position and scale parameterization ($x \times s$) reveals a parabolic shape (fig 9a). A different picture results if we use a parameterization which makes the degrees of freedom of the scale space more homogeneous [5, 35, 36]; in particular, using $\left(x/\sqrt{2s}\right) \times \ln s$ reveals the atlas to have a roughly pyramidal shape (fig 9b).

We can chart the degrees of freedom of an atlas using the approximate equivalence between Gaussian apertures and jets. We have done this in figs 9a,b by marking the heights of apertures approximating different orders of jet, and sub-dividing the apertures according to the dimensionality of the jets. The rendering of the atlas in fig 9b reveals how closely it is a continuous analogue of the discrete image pyramid (fig 9c).

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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

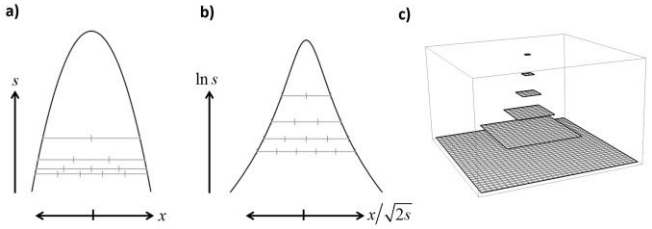

Figure 9 – Gaussian atlas and image pyramid compared. **a)** The black curves indicate the widths of the apertures in an atlas (shown dashed in figure 6). Horizontal lines indicate apertures equivalent to jets; subdivided by the degrees-of-freedom of the jet**. b)** As (a) but rendered with coordinates that make the degrees of freedom of scale space more homogeneous. **c)** An image pyramid as often used in Computer Vision.

We end with a tour through the atlas from coarse to fine, considering 2-D images, and what apertures at different scales reveal about them. At the top of the atlas is the aperture $\mathcal{G}(0,t)=[\boldsymbol{\delta},t]$ which reveals a single degree-of-freedom about the image. An IP with respect to this aperture is exactly equal to the IP with zero order jets. For natural images, such apertures provide nothing of use since local mean intensity is so dependent on illumination.

Going finer we reach the aperture $\mathcal{G}(\frac{2}{3}t,\frac{1}{3}t)$ whose IP approximates the 1st order jet IP. The 1st order jet has three degrees-of freedom, so this tell us that the aperture gives a view like a superior 3-pixel patch. First order jets provide a gradient vector in addition to mean intensity. The magnitude of the gradient is determined by local illumination, but the magnitude divided by the mean intensity is stable to intensity multiplication. It has been suggested that human vision is insensitive to 1st order structure [37], but Computer Vision has many effective descriptors that make effective use of the distribution of gradient directions over a region [8, 38].

Going finer we reach the aperture $\mathcal{G}(\frac{4}{5}t,\frac{1}{5}t)$ whose IP approximates the 2nd order jet IP. This jet has six degrees-of-freedom so the aperture gives a view like a superior 6-pixel patch. Sufficient articulation is visible through such apertures to allow local symmetry to be tested for [39] revealing around seven qualitatively distinct classes of structure. Basic Image Features are a scheme to do this directly from the equivalent 2nd order jet [40, 41], and Local Binary Patterns do something comparable based on 3×3 patches of down-sampled images [42].

Finer still we reach the aperture $\mathcal{G}(\frac{6}{7}t,\frac{1}{7}t)$ whose IP approximates the 3rd order jet IP. Such apertures reveal the image with approximately ten dimensions of articulation (the dimension of the 3rd order jet). At present there are no published schemes to classify this level of complexity based on geometry, though it seems plausible [43]. Certainly curved versus straight edges should be distinguishable, and ramps versus edges, but probably much more.

Finer still, the aperture $\mathcal{G}(\frac{8}{9}t,\frac{1}{9}t)$ approximates 4th order jets with 15 dimensions, and $\mathcal{G}(\frac{10}{11}t,\frac{1}{11}t)$ approximates 5th order jets with 21 dimensions. V1 simple cells may possibly have sufficiently articulated filters that they can compute this order of jet, but not likely higher [44]. Whether a modest codebook of *geometrically* distinct forms for such apertures is possible is unknown; modern Computer Vision systems would instead typically employ a *learnt* codebook whose bins are driven by their utility at inferring semantic labels when part of a larger recognition system [45].

As one progresses to even finer scales of the atlas the views become higher and higher dimensional. In some problem domains, verbatim recording of these views may be useful when individual rigid objects need to be recognized, but in natural images where recognition of object *class* is more important than object *identity*, and non-rigid deformation and occlusion are frequent, such records are unlikely to be worth the cost of storage. A possible alternative is to store an *incomplete* record of the view. One way to do this would be to store precisely located sub-aperture views at a restricted set of locations. For example, with a face one might use a Gaussian aperture to get exact views down to the level C in fig 1b, with nested, attached, relatively-located Gaussian apertures each focusing on an eye, a mouth etc., and going down to the level M [46]. Another possibility is to store unlocated sub-aperture views for all locations [47] - a locally-orderless representation [47-49]. When these views are quantized this is called a Bag-of-Textons representation in Computer Vision. For example in [50] gaussian-windowed local histograms of BIF classifications are used as a descriptor. There are many other examples [51].

10 Experimental evaluation

We explore the usefulness of the Gaussian aperture framework using image keypoint matching as an example. Keypoints are a common construction used in a range of Computer Vision systems [52]. They are sparse but numerous locations within an image identified by a detector with the aim of reducing the combinatorics of image-to-image matching. Once localised a dominant scale and orientation for each is computed based on local image structure; and a descriptor of the image neighbourhood, at the dominant scale aligned to the dominant orientation, is computed. Descriptors for different keypoints can then be compared with the aim of establishing matches between images from which dense correspondence can be interpolated.

Keypoint description is non-trivial because of: geometric and luminance distortions; positional, rotational and scale variability in keypoint detection; and noise. Many keypoint descriptors have been proposed (most famously SIFT [8]), and several datasets on which to compare them have been assembled. Recently, the HPatches dataset [53] has been developed to unite the various advantages of previous datasets; performance scores for baseline descriptors have been computed and a competition workshop ran at ECCV 2016. We will present results on HPatches using methods developed under the aperture framework, after first describing the steerability properties of apertures and jets.

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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

10.1 Steerability

2D DtG filter families are rotationally steerable [7, 15]. Meaning that a rotation of the family, about the filters common centre, can be computed by linear re-combination of the original family. This property transfers to the jets that the families measure. For example, the 1st order jet $\vec{j} \coloneqq \begin{pmatrix} j_{00} & j_{10} & j_{01} \end{pmatrix}^{\mathrm{T}}$ transforms under an image rotation of θ to $\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix} \vec{j}$.

Rotation steerability can be understood informally but intuitively in the Gaussian aperture framework, to wit: jets approximate a Gaussian windowed view of the Gaussian blurred image, nothing is lost or gained from the view of the image through that window as the image is rotated about the window centre, hence the rotated jet is determined by the original jet. Further using this equivalence between jets and Gaussian apertures suggests that jets should be *approximately* steerable for small translations and scalings of the image. Calculation confirms this [33]. For example, after a small translation δ_x the jet is approximately $\vec{j} + \delta_x \begin{pmatrix} j_{10} & j_{20} & j_{11} \end{pmatrix}^{\mathrm{T}}$, and after a small stretch of $1+\delta_{xx}$ in the x-direction the jet is approximately $\vec{j} - \delta_{xx} \begin{pmatrix} (2s) j_{20} & \left(j_{10} + (2s) j_{30}\right) & (2s) j_{21} \end{pmatrix}^{\mathrm{T}}$. Similarly, for a small rotation δ_θ the jet is approximately $\vec{j} + \delta_\theta \begin{pmatrix} 0 & -j_{01} & j_{10} \end{pmatrix}^{\mathrm{T}}$. Note that for rotation only the original jet components are needed, whereas translation needs components one order higher, and re-scaling two orders.

Steering of jets computed in this way is approximate because it is a linearization of the trajectory of the jets through jet space as the image is transformed, and because higher order jet terms may be needed that are not available in the original jet. In practice we can control the approximation by not translating or rescaling too far, and by assuming that any unavailable higher order terms are zero [33]. Results in the next section show that the approximation is good enough to be useful.

10.2 HPatches Results

The HPatches dataset consists of 65×65 pixel patches, organized into pairs, in disjoint training and test sets [53]. In the classification challenge, positive pairs show matching scene locations, and are classified as easy or hard dependent on the amount of between-image variation and the inaccuracy of the keypoint localization. Negative pairs show non-matching locations, either from the 'same' or 'different' scenes. From the two types of positive pair, and two types of negative pair, four separate sub-challenges are constructed, with overall performance defined as the mean over the four. In each classification sub-challenge a randomized list of 200K positive pairs and 1M negative pairs has to be ranked according to confidence of match, and the ranking is scored as average precision.

We have computed the performance scores of several novel descriptors on the HPatches classification challenge. In all cases we identically pre-processed the patches by performing a type of sphering about the mean patch. Specifically, we (i) standardized each patch to zero-mean and unit variance, (ii) computed the mean of standardized patches, (iii) divided the values of each standardized patch by their RMS deviation from the mean patch. The aim of the pre-processing was to lessen pedestal and contrast variation within positive pairs and to make the distribution of patches more uniform.

Table 1 lists scores for: a selection of baseline descriptors; the ECCV 2016 'local features' competition entries; and for our descriptors. Citations for descriptors are given in table where available. Descriptors can be classified as to whether they are purely engineered with tuning of only a small number of parameters, or whether they use supervised learning from labelled training data to tune a large number of parameters to the specifics of the challenge. Before our contribution learnt descriptors strongly outperformed unlearnt. The best learnt - *cassa-yt* - a CNN approach, scores 93.89%; the best non-learnt - *cmp-dm-1* - scores 75.04%. The novel schemes we present are non-learnt, using only a small number of parameters, tuned on a training set disjoint from the testing. Our results reduce, without eliminating, the lead that learnt schemes have over non-learnt.

TABLE 1
KEYPOINT DESCRIPTOR SCORES ON HPATCHES CLASSIFICATION CHALLENGE

Descriptor	Geometric Encoding used	Supervised Learning used	Descriptor dimension	2016 Competition	Score (%)
random			0	baseline	20.00
meanstd	patch statistics		2	baseline	46.46
SIFT [8]	SIFT		128	baseline	68.25
pyramid	downsampled patch		64		68.91
g-aperture	gaussian aperture		4225		70.99
j-aperture	DtG jets (order = 16)		153		71.22
rootsiftpca [54]	SIFT		80	best baseline w/o learning	74.28
cmp-dm-1	SIFT		128	best entry w/o learning	75.04
hannover-cl [55]		Siamese CNN	128	entry	76.71
tfeat-ratio-star [56]		CNN	128	baseline	82.87
zagreb-nm [57]		CNN	256	entry	85.07
s-j-aperture	DtG jets (order = 10)		66		85.10
cmp-ab [54]	continuous version of SIFT	+ superv. linear proj. from 567D to 100D	100	entry	86.69
m-s-j-aperture	DtG jets (order = 27)		406		86.83
cmp-dm-2	SIFT	+ details unknown	128	entry	91.41
casia-yt [58]		CNN	128	winning entry	93.89

Shaded rows are new schemes presented in this paper.

Starting with lower scores, our first descriptor (*pyramid*) is a square downsampled patch (like figure 3 top row). Tuning the scheme's parameters on the training data, we found trimming the patch to 64×64 and downsampling to an 8×8 patch was most effective, giving a score slightly higher than the raw SIFT descriptor provided as an HPatches baseline.

To compare against *pyramid* we tuned a gaussian aperture descriptor - *g-apertures* (like figure 3 bottom row) and a jet-based approximation - *j-apertures*. The parameters of the two schemes were optimized together, so that the tuned order of n=16 and DtG scale of $\sigma_{\text{filter}}=2^{4.0}$ for *j-aperture* corresponds to the tuned $\sigma_{\text{blur}}=2^{1.48}$, $\sigma_{\text{window}}=2^{3.98}$ for *g-apertures* according to equations of section 8. The two aperture

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. Citation information: DOI 10.1109/TPAMI.2017.2777856, IEEE Transactions on Pattern Analysis and Machine Intelligence

schemes perform almost identically, and marginally better than *pyramid*, in accordance with our theory.

The performance of *j-apertures* (71.22%) does not reach the performance of *cmp-dm-1* (75.04%) which is the best of previous non-learnt descriptors. The decisive difference seems to be that *j-aperture* is predicated on unperturbed positional correspondence between the two patches, whereas SIFT-based schemes, such as *cmp-dm-1*, allows for jittered correspondence. To address this we developed a steerable jet descriptor.

Our steerable jet approach is predicated on the difference between the two patches in a matching positive pair in large part arising from a linear spatial transformation. Under this assumption, if we could measure the jets in the correct different positions, orientations, etc. in the two patches they would match better than in the default position and orientation. However, because the position, orientation, etc. of jets (equivalently Gaussian apertures) is so fuzzily defined, there is no need to re-measure the jet, we can approximate 'nearby' jets by steering the original jet as illustrated in Figure 10.

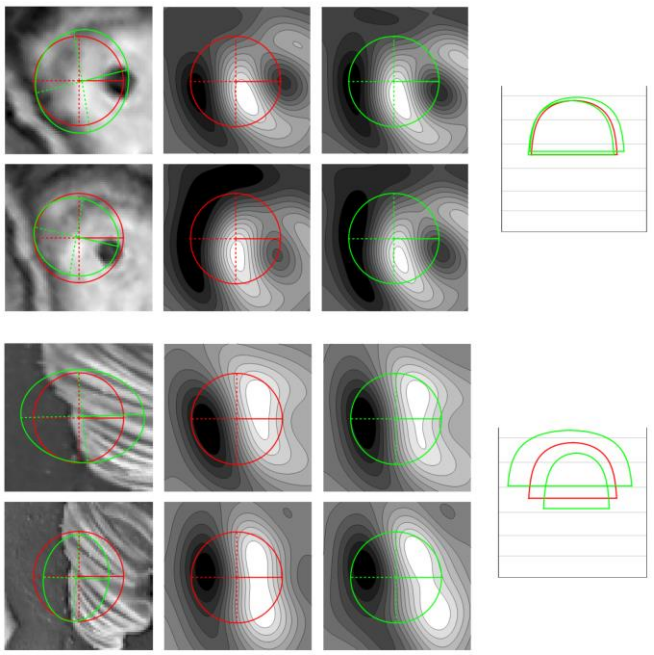

Figure 10 – Examples of the *s-j-aperture* descriptor. The top example is from the 'easy' positive pair set, the bottom from the 'hard'. Left column: patches overlaid with a red circle indicating the measured aperture, with radius $\sqrt{2}$ times the filter scale; and green ellipses indicating the steered aperture. 2nd column: the view of the patch through the measured aperture, computed from the measured jet. 3rd column: view through the steered aperture, computed from the steered jet (which was computed from the measured jet). Note how the steered views, within a pair, are more similar than the measured views. Right column: image-horizontal cross-sections of the scale space extent of the atlases arising from the measured and steered apertures. Vertical axis is log-scale; the bottom line is pixel scale, the next line a blur of $\sigma = 1$, then $\sigma = 2$, etc. These panels show the considerable overlap of the measured and steered atlases, which supports the accuracy of the steering. In all panels of the figure it is important to bear in mind that the red and green lines do not mark hard cut-off but only the start of the decaying part of the apertues.

To compute the optimal steer of the jets $\vec{j}_1, \vec{j}_2$ of the patches in a pair we compute six derivatives of each with respect to rotation (one), translation (two) and rescalings (three) of the image, as described in section 10.1. We arrange these derivatives as columns in two matrices $\mathbf{D}_1, \mathbf{D}_2$. Let the vector of parameters of the transformation applied to $\vec{j}_1$ be $\vec{\lambda} = \begin{pmatrix} \delta_\theta & \delta_x & \delta_y & \delta_{xx} & \delta_{xy} & \delta_{yy} \end{pmatrix}^{\mathrm{T}}$, steering it to $\vec{j}_1 + \vec{\lambda}^{\mathrm{T}}\mathbf{D}_1$. The inverse transformation applied to $\vec{j}_2$, assuming it is small, has parameters $-\vec{\lambda}$ steering it to $\vec{j}_2 - \vec{\lambda}^{\mathrm{T}}\mathbf{D}_2$. Solving to get the steered jets are as close as possible gives $\vec{\lambda} = (\mathbf{D}_1 + \mathbf{D}_2)^{+}(\vec{j}_1 - \vec{j}_2)$, where the + superscript denotes the pseudo-inverse using the jet IP. The distance between the optimally steered jets, computed using the jet norm, is the score for the patch pair.

This scheme works well but becomes less effective when the transformation that relates the patches is large. In particular, this occurs for a small number of patch pairs where there is a very large rotational change. We can improve the scheme by performing an exact rotation of one jet, by fixed amounts, before performing jet steering and using whichever rotation leads to the smallest jet distance. We use rotations of $\{\pm 0.8, \pm 0.4, 0\}$ radians.

This scheme - *s-j-aperture* - using a jet order of n=10 and a scale of $\sigma_{filter} = 2^{3.8}$ (both tuned on the trainin dataset) gives a score of 85.10%, a considerable improvement on the previous best (*cmp-dm-1*) for non-learnt descriptors of 75.04%, while producing a descriptor with lower dimension than anything of similar performance. We have not performed formal speed tests, but the computations are simple and non-iterative and should be competitively fast.

We have experimented with wringing every last drop of performance out of the jet steering approach. The best scheme we found (*m-s-j-aperture*) uses a higher order (n=27) jet, up to eight steering transformations performed in sequence, with two extra parameters to control the magnitude of the transformations and to choose when to stop performing them. Since this substantially increases the computation time and the descriptor dimensionality, while only slightly improving performance (86.83%), we do not advocate its use instead of *s-j-aperture*.

In conclusion, the fuzziness of Gaussian apertures allows a highly effective keypoint descriptor that performs 10 percentage points better than other non-learnt descriptors. While we acknowledge that the current best learnt descriptors perform a further 9 percentage points better, the operation of non-learnt schemes is more easily understood than learnt schemes, which may eventually lead to even better learnt schemes.

11 Summary & Discussion

We have presented a geometric structure for isolating image regions at different scales and inter-relating them: a

continuous version of the discrete image pyramid. Its fundamental element is the aperture, a positive weighting function paired with a level of scale space that it views. Such an aperture gives a view on the image as if through a fuzzy porthole of frosted glass. To organize these apertures into cross-scale structures we first defined a containment relation which holds when one aperture does not see anything that the other does not. We showed that only apertures that view finer scales can contain apertures that view coarser scales. We defined containment to be tight if the containg aperture cannot be reduced or the contained expanded. We showed, unexpectedly, that there are multiple apertures at any given fine scale that tightly contain any given coarser aperture. We simplified this complex structure of containment relations by defining a cause/effect relation to hold when there was containment, and the cause was as small as possible (or equivalently the effect was as large as possible). It transpired that causes were related to effects by a blur of the aperture equivalent to the fine to coarse scale change. We noted that it was important to appreciate that the blurring relation of aperture causation operates in the opposite direction to the blurring process of scale space images. The cause/effect relation strings apertures into 1-D cross-scale families we called atlases. We argued that preeminent within possible atlases were those composed of Gaussian apertures.

Having established the special status of Gaussian apertures, we related views through them to jets measured by DtG filters, showing that finite order jets approximate the views through equivalent apertures. We checked this approximation using computations on natural images. Finally, we showed that Gaussian atlases were like a continuous version of the image pyramid, and that various types and modes of image description can naturally be expressed in terms of them. Using keypoint description as an example, we showed how the aperture framework could inspire improved useful algorithms. We developed a keypoint descriptor that outperforms previous non-learnt descriptors, halving the lead that learnt methods have over non-learnt.

We briefly consider the biological relevance of our model. DtGs are an accepted model of Simple Cell neurons in mammalian primary visual cortex (V1) [16, 17]. As a model it fits the near linear response of these cells and accounts for the structure of their receptive fields [29]. Though it must be noted that there is still much that it does not account for [16, 59]. Since DtGs effectively compute derivatives of the blurred image, the model allows an interpretation of V1 as a multi-scale differential geometry engine [60]. This runs counter to an older interpretation, dominant in experimental Psychology, of Simple Cells as measuring local Fourier energy, an ensemble thus computing something like a patchwise Fourier Transform [61]. The framework in this paper provides theory underlying the patchwise view: it gives a picture of V1 as implementing a wide set of fuzzy-edged, frosted-glass portholes for viewing the image, using hardware that looks quite different from that.

We conclude with some ideas on how the theory of atlas structure could be developed further.

- On-demand synthesis of bespoke apertures from a Gaussian basis could be used to more precisely segment image objects. This could be efficiently done using large apertures to cover large parts of the object, and progressively smaller apertures to cover into the corners. There is neurophysiological and psychophysical evidence consistent with such a scheme [62, 63].
- Many effective local image descriptors are concatentations of a few repetitions of the same scheme operating at progressively coarser scales over wider extents [41, 64]. This gives, in effect, a local foveated view of the image. This is different from the atlases that we have described, which get tighter with increasing scale. Such foveated descriptors could correspond to views through apertures which are not iso-scale, but instead get coarser as the weighting function decays away from the centre.
- The nesting of atlases provides a means to put locally-orderless descriptions of image structure on a firm footing, possibly leading to refined schemes.

REFERENCES

[1] J. J. Koenderink, "The structure of images," *Biol Cybern,* vol. 50, pp. 363-70, 1984.

[2] T. Lindeberg, "Generalized axiomatic scale-space theory," *Advances in Imaging and Electron Physics,* vol. 178, pp. 1-96, 2013.

[3] J. Koenderink, A. van Doorn, and J. Wagemans, "The nature of the visual field, a phenomenological analysis," *Pattern Recognition Letters,* 2015.

[4] L. M. J. Florack, B. ter Haar Romeny, M. Viergever, and J. J. Koenderink, "The Gaussian scale-space paradigm and the multiscale local jet," *International Journal of Computer Vision,* vol. 18, pp. 61-75, Apr 1996.

[5] T. Lindeberg, *Scale-Space Theory in Computer Vision*: Springer, 1993.

[6] J. Weickert, S. Ishikawa, and A. Imiya, "Linear scale-space has first been proposed in Japan," *JOURNAL OF MATHEMATICAL IMAGING AND VISION,* vol. 10, pp. 237-252, 1999.

[7] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *icip*, 1995, p. 3444.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, pp. 91-110, Nov 2004.

[9] J. Weickert, S. Ishikawa, and A. Imiya, "On the history of Gaussian scale-space axiomatics," in *Gaussian Scale-Space Theory (J. Sporring, M. Nielsen, L. Florack, P. Johansen, Eds.)*, ed: Kluwer Academic Publishers, 1997, pp. 45-60.

[10] T. Lindeberg, "On the axiomatic foundations of linear scale-space," in *Gaussian scale-space theory*, ed: Springer, 1997, pp. 75-97.

[11] R. Duits, L. Florack, J. De Graaf, and B. ter Haar Romeny, "On the axioms of scale space theory," *Journal of Mathematical Imaging and Vision,* vol. 20, pp. 267-298, 2004.

[12] L. D. Griffin, "Scale-imprecision space," *Image and Vision Computing,* vol. 15, pp. 369-398, MAY 1997 1997.

[13] T. Lindeberg, "Generalized Gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space," *Journal of Mathematical Imaging and Vision,* vol. 40, pp. 36-81, 2011.

[14] L. D. Griffin, "Critical point events in affine scale-space," in *Gaussian Scale-Space Theory*. vol. 8, ed, 1997, pp. 165-180.

[15] J. J. Koenderink and A. J. van Doorn, "Representation of local geometry in the visual system," *Biol Cybern,* vol. 55, pp. 367-75, 1987.

[16] T. Lindeberg, "A computational theory of visual receptive fields," *Biological cybernetics,* vol. 107, pp. 589-635, 2013.

[17] R. A. Young, R. M. Lesperance, and W. W. Meyer, "The Gaussian Derivative model for spatial-temporal vision: I. Cortical model," *Spatial Vision,* vol. 14, pp. 261-319, 2001 2001.

[18] T. M. Lehmann, C. Gonner, and K. Spitzer, "Survey: Interpolation methods in medical image processing," *IEEE transactions on medical imaging,* vol. 18, pp. 1049-1075, 1999.

[19] H. Mostafavi, "Optimal window functions for image correlation in the presence of geometric distortion," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 27, pp. 163-169, 1979.

[20] N. Jaccard, N. Szita, and L. Griffin, "Segmentation of phase contrast microscopy images based on multi-scale local Basic Image Features histograms," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization,* pp. 1-9, 2015.

[21] L. D. Griffin, P. Elangovan, A. Mundell, and D. C. Hezel, "Improved segmentation of meteorite micro-CT images using local histograms," *Computers & Geosciences,* vol. 39, pp. 129-134, 2012.

[22] J. Gårding and T. Lindeberg, "Direct computation of shape cues using scale-adapted spatial derivative operators," *International Journal of Computer Vision,* vol. 17, pp. 163-191, 1996.

[23] O. Linde and T. Lindeberg, "Composed complex-cue histograms: An investigation of the information content in receptive field based image descriptors for object recognition," *Computer Vision and Image Understanding,* vol. 116, pp. 538-560, 2012.

[24] J. Hadamard, "Sur les problèmes aux dérivées partielles et leur signification physique," *Princeton university bulletin,* vol. 13, p. 28, 1902.

[25] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill-posed problems*. Washington, DC: Winston, 1977.

[26] R. A. Horn and C. R. Johnson, *Matrix Analysis* Cambridge University Press, 1985.

[27] S. U. Pillai, T. Suel, and S. Cha, "The Perron-Frobenius theorem: some of its applications," *Signal Processing Magazine, IEEE,* vol. 22, pp. 62-75, 2005.

[28] J. J. Koenderink and A. J. van Doorn, "Generic neighbourhood operators," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 14, pp. 597-605, Jun 1992.

[29] J. P. Jones and L. A. Palmer, "The two-dimensional spatial structure of simple receptive-fields in cat striate cortex," *Journal of Neurophysiology,* vol. 58, pp. 1187-1211, Dec 1987.

[30] L. D. Griffin, "The second order local-image-structure solid," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, pp. 1355-1366, AUG 2007 2007.

[31] M. Loog, "The jet metric," in *Scale Space and Variational Methods in Computer Vision*, ed: Springer Berlin Heidelberg, 2007, pp. 25-31.

[32] J. J. Koenderink and A. J. van Doorn, "Local Image Operators and Iconic Structure," presented at the Algebraic Frames for the Perception-Action Cycle, 1997.

[33] S. Makram-Ebeid and B. Mory, "Scale-space image analysis based on Hermite polynomials theory," in *Scale-Space Methods in Computer Vision, Proceedings*. vol. 2695, L. D. Griffin, Lillholm M., Ed., ed, 2003, pp. 57-71.

[34] L. Debnath, "On Hermite Transforms," *Mathematicki Vesnik,* vol. 1, pp. 285-292, 1964.

[35] L. Florack, "Image Structure, volume 10 of Computational Imaging and Vision Series," ed: Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

[36] T. Lindeberg, "Feature detection with automatic scale selection," *International journal of computer vision,* vol. 30, pp. 79-116, 1998.

[37] J. J. Koenderink and A. J. van Doorn, "Image processing done right," in *Computer Vision—ECCV 2002*, ed: Springer, 2002, pp. 158-172.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 886-893.

[39] L. D. Griffin and M. Lillholm, "Symmetry Sensitivities of Derivative-of-Gaussian Filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, pp. 1072-1083, 2010 2010.

[40] L. D. Griffin, M. Lillholm, M. Crosier, and J. van Sande, "Basic Image Features (BIFs) Arising from Approximate Symmetry Type," in *Proc. Conference on Scale Space and Variational Methods in Computer Vision*. vol. 5567, X.-C. Tai, K. Morken, M. Lysaker, and K.-A. Lie, Eds., ed: Springer, 2009, pp. 343-355.

[41] M. Crosier and L. D. Griffin, "Using Basic Image Features for Texture Classification," *International Journal of Computer Vision,* vol. 88, pp. 447-460, JUL 2010 2010.

[42] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 24, pp. 971-987, 2002.

[43] W. R. Ball, "On Newton's classification of cubic curves," *Proceedings of the London Mathematical Society,* vol. 1, pp. 104-143, 1890.

[44] L. D. Griffin, "Basic Colors and Image Features," *Handbook of Experimental Phenomenology: Visual Perception of Shape, Space and Appearance,* pp. 449-475, 2013.

[45] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. European Conference on Computer Vision 2006*, ed: Springer, 2006, pp. 490-503.

[46] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Transactions on computers,* pp. 67-92, 1973.

[47] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, "Locally orderless tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 1940-1947.

[48] B. van Ginneken and B. M. ter Haar Romeny, "Applications of locally orderless images," *Journal of Visual Communication and Image Representation,* vol. 11, pp. 196-208, 2000.

[49] J. J. Koenderink and A. J. Van Doorn, "The structure of locally orderless images," *International Journal of Computer Vision,* vol. 31, pp. 159-168, 1999.

[50] N. Jaccard, N. Szita, and L. Griffin, "Segmentation of phase contrast microscopy images based on multi-scale local Basic Image Features histograms," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization,* pp. 1-9, 2015.

[51] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-

features for object categorization and semantic video retrieval," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007, pp. 494-501.

[52] J. Li and N. M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing,* vol. 71, pp. 1771-1787, 2008.

[53] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "HPatches: A benchmark and evaluation of handcrafted and learned local descriptors," *arXiv preprint arXiv:1704.05939,* 2017.

[54] A. Bursuc, G. Tolias, and H. Jégou, "Kernel local descriptors with implicit rotation matching," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 595-598.

[55] L. Chen, F. Rottensteiner, and C. Heipke, "Invariant descriptor learning using a Siamese convolutional neural network," in *XXIII ISPRS Congress, Commission III 3 (2016), Nr. 3*, 2016, pp. 11-18.

[56] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *BMVC*, 2016, p. 3.

[57] N. Markuš, I. S. Pandžić, and J. Ahlberg, "Learning Local Descriptors by Optimizing the Keypoint-Correspondence Criterion," *arXiv preprint arXiv:1603.09095,* 2016.

[58] Y. Tian, B. Fan, and F. Wu, "L2-Net: Deep Learning of discriminative patch descriptor in euclidean space," presented at the CVPR '17, Hawaii, 2017.

[59] B. A. Olshausen and D. J. Field, "How close are we to understanding V1?," *Neural Computation,* vol. 17, pp. 1665-1699, Aug 2005.

[60] J. J. Koenderink, "The brain a geometry engine," *Psychol Res,* vol. 52, pp. 122-7, 1990.

[61] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *VISION RESEARCH,* vol. 20, pp. 847-856, 1980.

[62] P. R. Roelfsema, "Cortical algorithms for perceptual grouping," *Annu. Rev. Neurosci.,* vol. 29, pp. 203-227, 2006.

[63] B. B. Kimia, "On the role of medial geometry in human vision," *Journal of Physiology-Paris,* vol. 97, pp. 155-190, 2003.

[64] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, 1999, pp. 1150-1157.

Lewis D. Griffin received the BA (honors) degree in Mathematics and Philosophy from Oxford University, United Kingdom, in 1988, and the PhD degree from the University of London in 1995 for a thesis “Descriptions of Image Structure” in the area of computational vision. Following positions at Aston University (Vision Sciences) and Kings College London (Imaging Sciences) he has been at University College London (Computer Science) since 2005, where is now a Reader. His research interests include image structure, color vision, machine learning and biomedical modelling, with applications in security science, biomedicine and geoscience.