(An Overview of) Synergistic Reconstruction for Multimodality/Multichannel Imaging Methods

Simon R. Arridge, Matthias J. Ehrhardt, and Kris Thielemans

1Department of Computer Science, University College London, UK
2Department of Mathematical Sciences, University of Bath, UK
3Institute for Mathematical Innovation, University of Bath, UK
4Institute of Nuclear Medicine, University College London, UK

Imaging is omnipresent in modern society with imaging devices based on a zoo of physical principles, probing a specimen across different wavelengths, energies and time. Recent years have seen a change in the imaging landscape with more and more imaging devices combining that which previously was used separately. Motivated by these hardware developments, an ever increasing set of mathematical ideas is appearing regarding how data from different imaging modalities or channels can be synergistically combined in the image reconstruction process, exploiting structural and/or functional correlations between the multiple images. Here we review these developments, give pointers to important challenges and provide an outlook as to how the field may develop in the forthcoming years.
1. Introduction

Images have been an ever-present component of human civilisation for thousands of years because of their unique ability to record and represent complex information in a form directly interpretable by human intelligence. Whereas directly recorded images are ubiquitous in static and moving formats, the enormous advances in engineering, physics, mathematics and computer science in the last half-century has led to the increasing availability of indirect imaging methods. Specifically, we refer to technologies where data \( g \), belonging to some space \( Y \), is measured and the desired image \( f \), in a different space \( X \), is recovered by solving an inverse problem of the form

\[
g = A(f) + e
\]  

(1.1)

where \( A \) is the (possibly non-linear) mapping that models the generation of data from a given estimate of \( f \) and \( e \) represents noise, arising from one or several sources of error.

Such image reconstruction problems arise in many areas of science, including geophysics, non-destructive testing, atmospheric physics and, notably, in medical imaging which is the principal focus of this article. Many different physical phenomena can be measured (representing the space \( Y \)) such as electro-magnetic, acoustic, and optical waves, as well as particle counting processes, and many different physical parameters can be reconstructed (representing the space \( X \)), including density, sound speed, attenuation coefficients, molecular relaxation rates, tracer concentration, amongst many others. Modalities based on X-Rays (computed tomography) and magnetic resonance (both of which led to the award of Nobel prizes for their development) are well known examples present in every hospital.

With the success of such technologies has come the quest to push the boundaries of achievable imaging in regard to speed, resolution and additional physical parameters. Although most imaging techniques were originally envisioned as purely 2D methods, the relentless increase in computing power has made 3D imaging routine. Nevertheless, the challenge to address so called “4D” or even “5D” imaging (adding the dimension of time and/or spectral variations, or potentially both) still presents difficulties, both in terms of computation time, and, more fundamentally, in terms of adequate data acquisition within constraints such as patient tolerance and safety. We will refer to these extensions as Multichannel Imaging (MCI).

A separate, but related, extension to conventional image reconstruction modalities is the development of Multimodality Imaging (MMI). The key difference here is that the measurements usually are of different physical phenomena, and/or the recovered images represent two or more different physical parameters. The increasing interest in MMI is accelerating with the ever increasing advances in systems and reconstruction techniques [1–3].

This article is a brief overview of some recent developments in these topics with a focus on image reconstruction methods. The emphasis is on the various different incarnations of synergistic reconstruction wherein several images are recovered simultaneously from several data-sets where there are some common underlying properties that can be exploited during the reconstruction process. Joint reconstruction is often considered for data-sets acquired concurrently, i.e. sufficiently close together in time to be effectively simultaneous in comparison to temporal variation in the images. We also briefly cover the joint reconstruction of multiple images from data that was sequentially acquired, e.g. such as in dynamic imaging, follow-up studies or many multi-modality cases.

The article is organised as follows. We provide brief definitions and terminology in section 2 as well as a taxonomy of applications. Section 3 on methods for “guided reconstruction”, where a single image is reconstructed with regularisation based on one or more other images, will serve as a gentle introduction to the main body of this review, section 4, where we provide an overview of the dominant notions for synergistic reconstruction. We conclude this review in section 5 with a discussion and point out important challenges and an outlook for the future of the field. For completeness we summarise useful concepts from inverse problems and image reconstruction in an appendix section A.
2. A Taxonomy of Problems

(a) Basic Definitions

In this review we concentrate on scalar-valued images as most of the synergistic reconstruction literature has been developed in this context. Some concepts could be generalised to (geometrical) vector or tensor-based images (for instance for velocity or diffusion). We use the word “image” both for the continuous function \( f(x) \) and the discretised version where the function is obtained as a sum over basis functions \( b_i(x) \)

\[
f(x) = \sum_i f_i b_i(x)
\]  

(2.1)

where \( x \) is a coordinate in space (most often three dimensional: \( x \in \mathbb{R}^3 \)). Most authors use cuboid basis functions (“voxels”). We will introduce multi-channel images below.

A commonly used generic setting for solving problem (1.1) is to solve an associated variational problem

\[
f^* = \arg\min_{f} [D(g, A(f)) + \alpha \Psi(f)]
\]  

(2.2)

where \( D \) is the data fit functional measuring a suitable distance between the observed data \( g \) and the output of the model \( A(f) \), and \( \Psi \) is a regularisation functional. The approach (2.2) is often referred to as a variational regularization of the inverse problem (1.1). Under the Bayesian interpretation, (2.2) is the negative logarithm of the product of the likelihood and the prior, i.e.

\[
D(g, A(f)) \equiv - \log(P(g|A(f))), \quad \alpha \Psi(f) \equiv - \log(P(f)).
\]

In MCI we consider a vector of measurements \( g \in Y := \prod_{m=1}^{M} Y_m \), a vector of images \( f \in X := \prod_{p=1}^{P} X_p \) and the forward operator \( A : X \to Y \) mapping between these product spaces. Within this description we distinguish

- Multi-Channel Single-Image (MCSI), which implies reconstructing a single image from multiple channels, i.e. \( M > 1, P = 1 \).
- Single-Channel Multi-Image (SCMI), which implies reconstructing multiple images from a single channel, i.e. \( M = 1, P > 1 \).
- Multi-Channel Multi-Image (MCMI) which implies reconstructing multiple images from multiple channels, i.e. \( M > 1, P > 1 \).

In the case of MMI the various domain and range spaces are composed of different modalities and quantitative image representations, and we may write all terms in stacked form:

\[
g \equiv \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}, \quad A \equiv \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \quad f \equiv \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}
\]  

(2.3)

Each modality may be independently linear or nonlinear, well-posed, weakly ill-posed or strongly ill-posed.

MCSI implies some redundancy in the set of measurements, but is advantageous when it gives rise to a better posed inverse problem, e.g. in parallel Magnetic Resonance Imaging (MRI) [4], or in inverse scattering problems with multi-frequencies [5]. As this review is on joint reconstruction of multiple images, MCSI is not further considered, although of course it can occur as a sub-problem in a multi-image context, such as Positron Emission Tomography (PET)/MRI.

Unless requiring reference to particular case details, we will use a single notation for all the above:

\[
g = A(f).
\]  

(2.4)

(b) Guided Reconstruction

Closely related to the synergistic reconstruction problem in MCI or MMI is the possibility of using one acquired modality or channel, with a robustly reconstructed image, as a prior for subsequent,
usually less well-posed, image reconstruction problems. This constitutes a sequential process where two different acquisitions are made, image 1 is reconstructed, and then image 2 guided by image 1. Most often, the key idea is that image 1 has high resolution structural information whereas image 2 is low resolution and/or a functional image.

Although (within medical imaging) the above concept is sometimes called *anatomically-guided image reconstruction* or *structure-driven image regularisation*, we will refer to this approach as *guided image reconstruction* in order to generalise to other applications. We describe the main techniques in §3. A more detailed review of applications and methods can be found in [6]; see also [7] for PET/MRI applied to neurology.

(c) Synergistic Reconstruction

In terms of the definitions given in § 2(a) we can list a number of applications, grouped according to their similarity in regard to image reconstruction strategies:

- **SCMI:** Examples are acoustic speed and attenuation from Ultrasound (US) data [8], absorption and scattering from unscattered photons only, or from photon intensity only, in Diffuse Optical Tomography (DOT) [9], and attenuation and activity (*i.e.* tracer concentration) estimation in PET and Single Photon Emission Computed Tomography (SPECT) [10]. However, SCMI is usually a very ill-conditioned problem with unsatisfactory results, often exhibiting non-uniqueness [11]. An exception is PET with Time of Flight (TOF) where the extra information on the approximate location of the activity along a line of response is sufficient to provide uniqueness up-to some constant [12].

- **MCMI:** Multispectral Imaging. Modalities that can be classified as *multispectral imaging* include multispectral Computed Tomography (CT) [13–19], multispectral Electrical Impedance Tomography (EIT) [20,21], multispectral DOT [22], multispectral Photoacoustic Tomography (PAT) [23] and Quantitative Photoacoustic Tomography (QPAT) [24–26]. A common feature of multispectral imaging is the expression of $f(\lambda)$ as a mixture of component images $\sum_m E_m(\lambda) z_m$ (2.5) where $\lambda$ is the energy/wavelength and the spectral signatures $E_m(\lambda)$ of the components may be fully or partially known. Therefore these problems are often posed in two steps: a channel by channel reconstruction for each $\lambda$ followed by a spectral unmixing problem solving (2.5) for $z$. If the prior is defined in terms of $f$ it may introduce extra bias into the recovery of $z$. Alternatively, the prior may be directly imposed on $z$; see [21,27,28] for examples. A benefit of a one-step reconstruction procedure is that there is no propagation of errors from the channel-wise tomographic inverse problem to the spectral unmixing one. A drawback is that the full inverse problem may become nonlinear, which potentially leads to a longer computation time.

- **MCMI:** Multi-Energy imaging. An example is the reconstruction of both attenuation and tracer concentration from detection of both unscattered and scattered photons, the latter having reduced energy. This has been demonstrated in both SPECT [29–31] and PET [32–34].

- **MCMI:** Multi-Time Imaging. In dynamic/kinetic imaging the aim is to explicitly separate different temporal variations as separate images. The time-series of images can be reconstructed with e.g. a nuclear norm constraint [35]. The assumption is that the number of separate temporal components is small and could be concisely expressed using Principle Component Analysis (PCA), Independent Component Analysis (ICA) or Non-Negative Matrix Factorisation (NMF), for example. This can be extended to allow outlier representations via the *Low-Rank plus sparse* approach [36]. Another strategy...
is to combine all temporal data into a single data frame and use its reconstruction in a compressed sensing-style reconstruction [37]. Alternatively, the temporal behaviour could be expressed as an approximately known function with some random components modelled by a Kalman process [38], or it can be constrained to follow a (potentially non-linear) kinetic model

\[ f(t) = \Phi(k, t) \]  

with \( k \) parametric images, similar to (2.5); see [39–41] for reviews. However, we are not aware of any literature yet where there is a prior that couples the parametric images.

- **MCMI**: MRI imaging with multiple sequences also generate multiple images, with many different applications and sometimes overlap with the aforementioned categories, e.g. multimodal dynamical MRI [42], multiparametric MR [43], multi-contrast MRI [44–46], and MR Fingerprinting which aims to estimate multiple images corresponding to different tissue properties [47,48]; see also [49].

- **Multi-Modality Multi-Image (MMMI)**: The distinguishing aspect is that multiple modalities may have very different physical measurements. Notable examples are PET and MRI (PET/MRI) [50–52], PAT and US [53], geophysics applications with multiple data (e.g. electromagnetic waves, seismic waves, radar, DC resistivity, groundwater flow) [54–61].

- **MMMI**: Coupled Physics Imaging (CPI). These methods are so-named because the measurement involves the cross-generation of one type of wave from another [62]. Examples include PAT and Optical Coherence Tomography (OCT) [63], quasi-static elasticity imaging [64] and Acousto-Electric Tomography [65].

### 3. Guided Reconstruction

In this section we briefly describe the main methods that have been developed for guided reconstruction, as many of these ideas can and have been extended to the synergistic case.

#### (a) The Continuous Setting

An obvious way to include information from an auxiliary image is to take a regularisation strategy for one modality, \( f_1 \) say, and introduce a local dependence on \( f_2 \). For example, the form expressed in eq. (A 6) can be extended to the form

\[ \Psi(f_1) := \int_{\Omega} w(f_2(x)) \psi(|\nabla f_1(x)|) dx, \]

\[ \rightarrow \Psi'(f_1) = -\nabla \cdot \left( w(f_2(x)) \frac{\psi'(|\nabla f_1(x)|)}{|\nabla f_1(x)|} \right) \nabla f_1(x) = -\nabla \cdot (w(f_2)) \kappa(f_1(x)) \nabla f_1(x) \]  

(3.2)

The choice for the weighting \( w(f_2) \) could be quite general, and need not be strictly local. A recurring concept is to make \( w \in [0,1] \) an edge-indicator such as (A 9) computed on \( f_2 \), i.e.

\[ w(f_2(x)) = \exp \left( -|\nabla f_2(x)|^2/\epsilon_2^2 \right) \]  

(3.3)

This form of prior favours a reconstruction of \( f_1 \) where its edges are colocated with those of \( f_2 \); see figure 1(a). A more powerful approach may be also to encourage the direction of image gradients to be aligned; see figure 1(b). This is similar to the Edge Enhancement Diffusion concept outlined in (A 10)-(A 12). Again, the new concept is to control the flow of \( f_1 \) based on the gradient directions in \( f_2 \) rather than only on those of \( f_1 \) itself. For image reconstruction the methodology was introduced by Kaipio et al. [66] by defining a tensor field, \( B(f_2) \), that incorporates directional...
structure from $f_2$:

$$
\Psi(f_1) = \int_\Omega |\nabla f_1|^2_{B(f_2)} dx = \int_\Omega |U^T \nabla f_1|^2 dx
$$

(3.4)

Specifically, for two-dimensional images, the choice

$$
B = I - (1 - \beta) \hat{\nu} \hat{\nu}^T = \hat{\tau} \hat{\tau}^T + \beta \hat{\nu} \hat{\nu}^T = RK R^T = UU^T
$$

(3.5)

where $R = [\hat{\nu} \hat{\tau}] = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$ is a rotation matrix, $K = \begin{pmatrix} \beta & 0 \\ 0 & 1 \end{pmatrix}$ is an anisotropy matrix, $U = RK^{1/2}$ and $\beta \in [0, 1]$ is again an edge-indicator, leads to the diffusion flow

$$
\frac{\partial f_1}{\partial t} = -\nabla \cdot R K R^T \nabla f_1 = -\tilde{\nabla} \cdot K \tilde{\nabla} f_1.
$$

(3.6)

Here $\tilde{\nabla} = R^T \nabla = \left( \begin{array}{c} \frac{\partial}{\partial \hat{\nu}} \\ \frac{\partial}{\partial \hat{\tau}} \end{array} \right)$ is the gradient in local “gauge” coordinates. This approach can be generalised by using other local functions $\psi$ resulting in the form

$$
\Psi(f_1) := \int_\Omega \psi(|U^T \nabla f_1|) dx,
$$

(3.7)

$$
\rightarrow \Psi'(f_1) = -\nabla \cdot U \left( \frac{\psi(|U^T \nabla f_1(x)|)}{|U^T \nabla f_1(x)|} \right) U^T \nabla f_1(x)
$$

(3.8)

$$
= -\tilde{\nabla} \cdot K^{1/2} \left( \frac{\psi(|\nabla f_1(x)|)}{|\nabla f_1(x)|} \right) K^{1/2} \tilde{\nabla} f_1(x).
$$

(3.9)

Of course here we can only give a glimpse into continuous regularisation for guided reconstruction. Similar ideas exist for regularisers that do not fit (3.1) such as the total variation [67] and the total generalised variation [68]; see for instance [6,69,70].

(b) The Discrete Setting

The above description is given in the continuous setting, but can be readily discretised. Here we briefly describe some methods that are specific to the discrete case.

(i) Markov Random Field Priors

One prominent example is based on Markov Random Fields (MRFs) (A 14), where weights are made dependent on $f_2$:

$$
\Psi(f_1) = \frac{1}{p} \sum_i \sum_{j \in N(i)} w_{ij}(f_2) ||f_i - f_j||^p.
$$

(3.10)

The simplest choice is to set the weights to zero across known edges (derived from $f_2$), ideally with some blurring to accommodate imperfect edge information [71]. Leahy & Yan estimated both image values and edge-indicators by incorporating known edge information (obtained from MRI) into an MRF prior that reduced the weights across edges, while encouraging continuous edges [72].

To avoid having to determine the edges, the most popular choice nowadays to choose the weights in (3.10) is called the Bowsher prior [73]. Here for every voxel $i$, only the $n$ weights are kept non-zero which correspond to the $n$ smallest differences $||f_{2i} - f_{2j}||$.

Another choice for the weights, inspired by the kernel-method described in A(e), is to use a similarity function $K$ between “features” computed on $f_2$, potentially together with a dependence
on the distance between the $i$ and $j$ voxels to enforce locality, e.g.:

$$ w_{ij}(f_2) = K(T_i(f_2), T_j(f_2)) \exp \left( -\frac{||r_i - r_j||^2}{2\sigma_s^2} \right) \quad (3.11) $$

with $T_i(f_2)$ a feature-vector computed at voxel $i$, $r_i$ the spatial coordinate of voxel $i$, and $\sigma_s$ a tunable parameter. This has been used with the Radial Basis Function (RBF) similarity function (A.19) for multi-tracer PET by Ellis et al. [74], who sparsified the weights by keeping only the $n_{\text{max}}$ largest weights for each voxel, similar to the Bowsher prior\(^1\). Bland et al. extended this idea by further adapting the weights over iterations by including Gaussian differences of the PET of the previous update [75].

(ii) Kernel Methods

Wang & Qi have used the kernel method (see section A(e)) to represent the image as a linear combination of transformed “features” computed on the guidance [76]. The image that needs to be estimated can then be written in terms of a coefficient image $\zeta$ as

$$ f_{1,i} = \sum_j K(T_i(f_2), T_j(f_2))\zeta_j \quad (3.12) $$

with $K$ a similarity function (“kernel function”) between features $T_i(f_2)$ and $T_j(f_2)$. The kernel matrix $K(T_i(f_2), T_j(f_2))$ was sparsified to improve computational performance and stability. The reconstruction then becomes an optimisation problem in terms of $\zeta$. This approach has been used in MMI: see [76–78] (PET with MRI), [79] (SPECT with PET), [80] (fluorescence molecular tomography with CT or MRI), [81] (DOT with CT), but also for MCMI: PET dynamic imaging using static images as guide [76] or temporal features derived from the raw data [82].

One potential problem with kernel-based methods is that the kernel matrix can be too restrictive such that “unique” features in the image that is reconstructed are suppressed. This can be mitigated by limiting the spatial extent of the kernel function [83], or by adapting the kernel matrix by using features computed on both the guidance and current estimate of the image. The latter approach can also be called the “hybrid” kernel method [52].

(iii) Basis Function Selection

The basis functions (2.1) can be chosen based on the guidance, for instance increasing spatial extent at locations where $f_2$ is smooth.

An approach originating in the machine learning community is “dictionary learning”, where images are written in terms of a dictionary, obtained from some training data. However, most of the literature does not fit in the guided reconstruction category as it uses data from the same modality for learning the dictionary, and/or adapts the dictionary from a current estimate of the image. Tang et al. reconstruct PET images using a quadratic penalty encouraging similarity with the previous iterate denoised using a pre-defined dictionary. The dictionary was trained on various images, including MRI images from the same subject [84]. Tahaei et al. instead reconstruct the PET image directly as a sparse combination of the dictionary [85]. The dictionary was learned from an MRI image of the same subject, and then changed to allow different contrast and impose non-negativity. Both papers show promising results, although somewhat surprisingly Tang et al. obtained good results with a dictionary trained on a simple hollow sphere as well. Sudarshan et al. extended the method of [84], by using a coupled dictionary encoding both PET and MRI images, with the latter obtained from the same subject, adapting the dictionary at each iteration [86].

“Super-voxels” (or super-pixels) are another closely related concept for selecting basis functions. They were originally developed for segmentation where voxels are grouped based on similarity and spatial closeness. Multiple “layers” of different super-voxel realisations were later

\(^1\)Note the relation between this approach and the discretised version of (3.3)
4. Synergistic Reconstruction

This section reviews important concepts for synergistic image reconstruction. Some methods are related to the concepts for guided reconstruction (see section 3) whereas some are directly formulated for the synergistic setting.

(a) Joint MAP Estimation

In a Bayesian framework, a central concept for many synergistic image reconstruction approaches is to formulate the multi-modality inverse problem (2.4) as a joint maximum a posteriori (JMAP) estimate given by

\[
(f_1^*, f_2^*) = \arg\min_{f_1, f_2} \left[ -\log P(f_1, f_2|g_1, g_2) = -\log P(g_1, g_2|f_1, f_2) + \alpha \Psi(f_1, f_2) + \text{const} \right],
\]

where \(P(g_1, g_2|f_1, f_2)\) is the multi-modality likelihood. This is in contrast to the Bayesian framework for guided reconstruction which would suggest a conditional a posteriori (CMAP) estimate

\[
f_1^* = \arg\min_{f_1} \left[ -\log P(f_1|g_1, f_2) = -\log P(g_1|f_1) + \alpha \Psi(f_1|f_2) \right].
\]

While (4.1) is conceptually simple, it is in general difficult to specify a good multi-modality likelihood. The situation significantly simplifies when certain conditional independences are assumed (see [89] for more details), since then the likelihood factors as

\[
P(g_1, g_2|f_1, f_2) = P(g_1|f_1)P(g_2|f_2)
\]

and the JMAP (4.1) becomes

\[
(f_1^*, f_2^*) = \arg\min_{f_1, f_2} \left[ D(g_1, A_1(f_1)) + D(g_2, A_2(f_2)) + \alpha \Psi(f_1, f_2) \right].
\]

In the special case of white Gaussian distributed noise in each modality, the JMAP then reads

\[
(f_1^*, f_2^*) = \arg\min_{f_1, f_2} \left[ \frac{1}{2\sigma_1^2} \| g_1 - A_1(f_1) \|^2 + \frac{1}{2\sigma_2^2} \| g_2 - A_2(f_2) \|^2 + \alpha \Psi(f_1, f_2) \right],
\]

which defines a natural statistical scaling between the two least squares terms. Such conditional independence assumptions are used in almost all contributions based on the JMAP although often not explicitly mentioned.

We mention also that similar questions arise w.r.t. the different image channels, including scaling between terms and quantitative difference in images scales, as many (joint) priors depend on image scale. In the optimisation literature pre-scaling between different dimensioned variables is known as sphereing (referring to the ellipticity of the posterior covariance), but this is rarely made explicit in the synergistic literature.

(b) Joint Sparsity

The variational synergistic reconstruction problem (4.3) needs a regulariser \(\Psi\) which encodes the desired properties between the images \(f_1\) and \(f_2\). Choosing such a regulariser is a highly nontrivial task and a good choice will generally depend on each individual application depending on what properties \(f_1\) and \(f_2\) are expected to share.

In many applications it is desirable for \(f_1\) and \(f_2\) to have many common edges, i.e. edges are more likely to occur at the same locations in \(f_1\) and \(f_2\) than it is for edges to occur at different locations. If, in addition, both images are likely to have a small jump set, then this a
priori information can be encoded via the Joint Total Variation (JTV) [50,55,90,91]

$$\text{JTV}(f) = \sum_i \sqrt{\|\nabla_i f_1\|^2 + \|\nabla_i f_2\|^2}$$  \hspace{1cm} (4.5)$$

where $\nabla_i f_k$ denotes the spatial gradient of $f_k$ at location (e.g. voxel) $i$. This regulariser can be readily extended to an arbitrary number of images by summing over all gradient norms. An alternative but completely equivalent viewpoint is to define the joint total variation as the sum over the 2-norm of the Jacobian of the vector-valued image $f$. In the context of colour image processing, this regulariser is also called Vectorial Total Variation (VTV) [92,93].

An alternative to sparsity of the Jacobian is to consider joint sparsity of the wavelet coefficients [94]. Since wavelets are highly localised one would expect this prior to be useful in similar situations as described before for the joint total variation. Joint sparsity can also be used in other over-complete basis settings such as “super-voxels”.

The notion of joint sparsity is not limited to explicitly given transforms such as the gradient or the wavelet transform. For instance, when considering a dynamic sequence $f_1, \ldots, f_n$ it is often desirable to promote a low rank of the Casoratti matrix

$$[f_1(\cdot), \ldots, f_n(\cdot)]$$  \hspace{1cm} (4.6)$$

where we abused MATLAB notation for simplicity. Such low rank can be for instance promoted with the nuclear norm [35,36].

(c) Joint Geometric Regularisation

The idea in joint geometric regularisation is to define a prior that enforces consistent geometric structure between channels or modalities. Although similar to the motivation in § 3(a) the key point is that it is optimised with respect to all images.

The Parallel Level Sets (PLS) prior [89,95] is defined as

$$\psi^{\text{PLS}}(f_1, f_2) = \int_\Omega \varphi \left( \psi \left( \|\nabla f_1(x)\| \|\nabla f_2(x)\| \right) \right) - \psi \left( |\langle \nabla f_1(x), \nabla f_2(x) \rangle| \right) \, dx$$  \hspace{1cm} (4.7)$$

for strictly increasing functions $\varphi, \psi$. One can see that this prior achieves its smallest value if the two gradients are co-linear (or parallel) at each point, i.e. for almost all $x \in \Omega$ there exists a scaling factor $\beta \in \mathbb{R}$ such that $\nabla f_1(x) = \beta \nabla f_2(x)$ or $\nabla f_2(x) = \beta \nabla f_1(x)$. See also [96] for a comprehensive overview of this concept.

This generalised framework encompasses some previously used regularisers. For instance, if $\phi(t) = t$ and $\psi(t) = t^2$, then it measures the squared norm of the cross-product of $\nabla f_1$ and $\nabla f_2$, which has been successfully used in the geophysics literature [55,56]. The gradient of $\psi^{\text{PLS}}$ w.r.t.
features, for example using the “jet” of derivatives of two random variables \( f_1 \) and \( f_2 \) produces an anisotropic diffusive flow where the diffusivity depends on \( f_1 \) and the gradient w.r.t. \( f_2 \). However (4.8) is extremely non-convex, and minimisation of joint entropy requires careful attention such as applying multiple re-starts from different initialisations [107] which has prevented its wider uptake as a regularisation scheme. As an alternative to Shannon entropy the Burg entropy has been proposed [109] which is computationally more tractable whilst still providing comparable results.

An approach combining joint feature-space clustering with image diffusion was developed in [111]. Here a distance measure in \( P(f_1, f_2) \) defined a diffusivity that favoured intra-class smoothing above inter-class smoothing. The method was applied to enhance multichannel MRI images, but could in principle be used to regularise image reconstruction problems as well.

\( f_1 \) produces an anisotropic diffusive flow where the diffusivity depends on \( f_2 \), and the gradient w.r.t. \( f_2 \) produces an anisotropic diffusive flow where the diffusivity depends on \( f_1 \) so as to overall favour the alignment of both image gradients [89].

There are also other priors which promote parallel level sets or parallel gradients. For instance, the Total Nuclear Variation (TNV) \([13,19,97,98]\) promotes sparsity in the singular values of the matrix of gradients (the Jacobian) \( V(x) := [\nabla f_1(x), \nabla f_2(x), \ldots, \nabla f_M(x)] \) at almost every \( x \in \Omega \). One can see that the Jacobian has rank 1 if and only if all images \( f_1, \ldots, f_M \) have pairwise parallel level sets. An advantage of TNV compared to PLS is that the former is convex in the joint argument \( (f_1, \ldots, f_M) \) whereas the latter is in general nonconvex. See also [6,96] for a deeper discussion of the connections between TNV and PLS. The same idea has also been exploited for higher-order regularization such as the total generalised variation, see [99] for more details.

In principle, the concepts of joint structure could be extended to higher order geometrical features, for example using the “jet” of derivatives \( M^j : f \mapsto \{ \frac{\partial f}{\partial x_i}, \frac{\partial^2 f}{\partial x_i \partial x_j}, \ldots, \frac{\partial^j f}{\partial x_1 \partial x_2 \ldots \partial x_j} \} \), which is the basis of several approaches to analysing shape in images, including at multiple scales [100,101]. For example, one could consider the definition of curvature \( \gamma = \hat{\nu}^T H \hat{\nu} \), where \( \hat{\nu} = \frac{\nabla f}{\| \nabla f \|} \) is the local level set normal direction and \( H = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix} \) is the image Hessian [102]. Images with the same local curvature may have stronger corresponding structural similarity then those with only local normal coincidences; see figure 1 (c).

(d) Regularisation exploiting Joint Statistics

Joint entropy is a measure of randomness characterizing the joint probability density function \( P(f_1, f_2) \) of two random variables \( f_1 \) and \( f_2 \). The joint Shannon Entropy is given by:

\[
S(f_1, f_2) = - \int_\Omega P(f_1(x), f_2(x)) \log(P(f_1(x), f_2(x))) dx.
\] (4.8)

See figure 2 for an illustration. Joint Entropy and Mutual Information (i.e. the difference between joint and marginal entropies) are routinely used in image registration [103].

Setting \( \Psi(f_1, f_2) = S(f_1, f_2) \) specifies a regularisation scheme that minimises the joint entropy between \( f_1 \) and \( f_2 \). Qualitatively, entropy measures the “peakiness” of a probability distribution; i.e. the more concentrated the distribution around cluster points the lower the entropy, and minimising it will lead to images that have less uncertainty [104–109]. One drawback of the definition in (4.8) is that it depends only on pixel intensity value and does not admit any spatial correlation. This motivated Tang and Rahmim to extend the Joint Entropy concept to a multi-resolution description based on wavelets [110].

Differentiation of \( S(f_1, f_2) \) can be made computationally efficient using Parzen kernel density estimators to develop a continuous function based on the sample pixels in \( f_1, f_2 \) [106,108]. However (4.8) is extremely non-convex, and minimisation of joint entropy requires careful attention such as applying multiple re-starts from different initialisations [107] which has prevented its wider uptake as a regularisation scheme. As an alternative to Shannon entropy the Burg entropy has been proposed [109] which is computationally more tractable whilst still providing comparable results.

An approach combining joint feature-space clustering with image diffusion was developed in [111]. Here a distance measure in \( P(f_1, f_2) \) defined a diffusivity that favoured intra-class smoothing above inter-class smoothing. The method was applied to enhance multichannel MRI images, but could in principle be used to regularise image reconstruction problems as well.

(e) Recycling of Guided Reconstruction for Synergistic Reconstruction

An alternative to joint reconstruction via joint regularization (4.3) is to alternate between guided reconstructions. In its most generality let \( \mathcal{A}(A, g, f, v) \) be an algorithm that takes data \( g \), current estimate \( f \) of the solution to the inverse problem \( Af = g \) and guide image \( v \), then one can always

- \( f_1 \) produces an anisotropic diffusive flow where the diffusivity depends on \( f_2 \), and the gradient w.r.t. \( f_2 \) produces an anisotropic diffusive flow where the diffusivity depends on \( f_1 \) so as to overall favour the alignment of both image gradients [89].

There are also other priors which promote parallel level sets or parallel gradients. For instance, the Total Nuclear Variation (TNV) \([13,19,97,98]\) promotes sparsity in the singular values of the matrix of gradients (the Jacobian) \( V(x) := [\nabla f_1(x), \nabla f_2(x), \ldots, \nabla f_M(x)] \) at almost every \( x \in \Omega \). One can see that the Jacobian has rank 1 if and only if all images \( f_1, \ldots, f_M \) have pairwise parallel level sets. An advantage of TNV compared to PLS is that the former is convex in the joint argument \( (f_1, \ldots, f_M) \) whereas the latter is in general nonconvex. See also [6,96] for a deeper discussion of the connections between TNV and PLS. The same idea has also been exploited for higher-order regularization such as the total generalised variation, see [99] for more details.

In principle, the concepts of joint structure could be extended to higher order geometrical features, for example using the “jet” of derivatives \( M^j : f \mapsto \{ \frac{\partial f}{\partial x_i}, \frac{\partial^2 f}{\partial x_i \partial x_j}, \ldots, \frac{\partial^j f}{\partial x_1 \partial x_2 \ldots \partial x_j} \} \), which is the basis of several approaches to analysing shape in images, including at multiple scales [100,101]. For example, one could consider the definition of curvature \( \gamma = \hat{\nu}^T H \hat{\nu} \), where \( \hat{\nu} = \frac{\nabla f}{\| \nabla f \|} \) is the local level set normal direction and \( H = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix} \) is the image Hessian [102]. Images with the same local curvature may have stronger corresponding structural similarity then those with only local normal coincidences; see figure 1 (c).

(d) Regularisation exploiting Joint Statistics

Joint entropy is a measure of randomness characterizing the joint probability density function \( P(f_1, f_2) \) of two random variables \( f_1 \) and \( f_2 \). The joint Shannon Entropy is given by:

\[
S(f_1, f_2) = - \int_\Omega P(f_1(x), f_2(x)) \log(P(f_1(x), f_2(x))) dx.
\] (4.8)

See figure 2 for an illustration. Joint Entropy and Mutual Information (i.e. the difference between joint and marginal entropies) are routinely used in image registration [103].

Setting \( \Psi(f_1, f_2) = S(f_1, f_2) \) specifies a regularisation scheme that minimises the joint entropy between \( f_1 \) and \( f_2 \). Qualitatively, entropy measures the “peakiness” of a probability distribution; i.e. the more concentrated the distribution around cluster points the lower the entropy, and minimising it will lead to images that have less uncertainty [104–109]. One drawback of the definition in (4.8) is that it depends only on pixel intensity value and does not admit any spatial correlation. This motivated Tang and Rahmim to extend the Joint Entropy concept to a multi-resolution description based on wavelets [110].

Differentiation of \( S(f_1, f_2) \) can be made computationally efficient using Parzen kernel density estimators to develop a continuous function based on the sample pixels in \( f_1, f_2 \) [106,108]. However (4.8) is extremely non-convex, and minimisation of joint entropy requires careful attention such as applying multiple re-starts from different initialisations [107] which has prevented its wider uptake as a regularisation scheme. As an alternative to Shannon entropy the Burg entropy has been proposed [109] which is computationally more tractable whilst still providing comparable results.

An approach combining joint feature-space clustering with image diffusion was developed in [111]. Here a distance measure in \( P(f_1, f_2) \) defined a diffusivity that favoured intra-class smoothing above inter-class smoothing. The method was applied to enhance multichannel MRI images, but could in principle be used to regularise image reconstruction problems as well.

(e) Recycling of Guided Reconstruction for Synergistic Reconstruction

An alternative to joint reconstruction via joint regularization (4.3) is to alternate between guided reconstructions. In its most generality let \( \mathcal{A}(A, g, f, v) \) be an algorithm that takes data \( g \), current estimate \( f \) of the solution to the inverse problem \( Af = g \) and guide image \( v \), then one can always...
Figure 2: Joint Probability measures of multiple images. Top row: two images $f_1$, $f_2$ and their joint histogram. Bottom row: the result of applying edge-preserving diffusion to $f_1$ and $f_2$ and their new joint histogram. The joint entropy of the bottom right is lower than the top right, and the classes are more clearly separated.

create an algorithm to perform joint reconstruction by iterating

$$f_1^{k+1} = A(A_1, g_1, f_1^k, f_2^k)$$

(4.9)

$$f_2^{k+1} = A(A_2, g_2, f_2^k, f_1^{k+1})$$

(4.10)

or choose a guide image based on previous iterates, $v = B(f^k)$ and update in parallel

$$f_1^{k+1} = A(A_1, g_1, f_1^k, v)$$

(4.11)

$$f_2^{k+1} = A(A_2, g_2, f_2^k, v).$$

(4.12)

The algorithm $A$ may be derived from an algorithm which solves a guided variational reconstruction problem but it does not have to be. It is also possible to use different algorithms $A_1$ and $A_2$.

This ad hoc approach has a number of advantages. First of all, this approach is highly modular, meaning that any guided reconstruction algorithm can be recycled into a joint reconstruction algorithm. Second, most guided reconstruction algorithms are better understood and have favourable properties when compared with joint reconstruction algorithms. For example, many guided variational reconstruction problems are convex and are independent of the scaling of the guide image – two properties which joint variational approaches often lack.

That being said, it has also a number of disadvantages. Most importantly, there is no guarantee that the sequence $f^k$ will converge and if it does, how the limit can be characterised. Second, the sequence $f^k$ will depend on the actual implementation of the algorithm $A$ like number of subiterations, step size etc even if the algorithm $A$ itself is well-characterised as converging to the optimal solution of a guided variational regularisation problem.

This approach has been used for spectral CT [15] where the algorithm $A$ was solving a directional total variation regularised least squares problem and the guide image $v$ is either chosen to be a weighted average over the previous iterate or a randomly chosen image from the previous iterate. See also the next section.
(f) Optimisation-inspired Synergistic Reconstruction

As we will highlight in section 5, many joint reconstruction variational problems and algorithms have unfavourable properties. For example, when solving (4.3) with joint total variation (4.5) regularisation and grouped coordinate descent, the iterations read

\[
\begin{align*}
f_{k+1}^1 &= \arg \min_f D(g_1, A_1(f)) + \alpha \text{JTV}(f, f_k^2) \\
f_{k+1}^2 &= \arg \min_f D(g_2, A_2(f)) + \alpha \text{JTV}(f_{k+1}^1, f).
\end{align*}
\]

(4.13)

This algorithm (and the underlying variational problem) has two potential drawbacks. First, the same regularisation parameter \(\alpha\) is being used for both modalities. Second, since the regulariser compares the magnitude of the gradients of the two modalities, the reconstruction will favour the modality with the larger scale even though often they capture two very different physical phenomena which should not be compared.

A way out of these problems is to change the iterations (4.13) and introduce a weighting \(\mu_k^j > 0\) and one regularization parameter for each modality \(\alpha_j\) and iterate

\[
\begin{align*}
f_{k+1}^1 &= \arg \min_f D(g_1, A_1(f)) + \alpha_1 \text{JTV}(\mu_{k+1}^1 f_{k+1}^1, f) \\
f_{k+1}^2 &= \arg \min_f D(g_2, A_2(f)) + \alpha_2 \text{JTV}(\mu_{k+1}^2 f_{k+1}^2, f).
\end{align*}
\]

This approach has been studied for another algorithm (ADMM [112]) in the context of PET/MRI [113] and with a weighted quadratic prior (similar to (3.11)) instead of joint total variation for the same application in [114]. While this ad hoc modification potentially overcomes the aforementioned problems, it has the drawback that it is unclear how to choose the parameters \(\mu_k^j, \alpha_j\) and if the iterates \(f^k\) converge.

A related but different approach are the infimal-convolution Bregman total variation iterations [51]. Here the starting point are Bregman iterations based on the total variation regulariser [115] which given an iterate \(f_k\) and a subgradient \(p_k \in \partial \text{TV}(f_k)\) read

\[
\begin{align*}
f^{k+1} &= \arg \min_f \frac{1}{2}\|Af - g\|^2 + \alpha \text{D}_{\text{TV}}^{p_k}(f, f_k) \\
p^{k+1} &= \alpha^{-1} (p_k - \nabla^*(Af^{k+1} - g)) .
\end{align*}
\]

(4.14)

(4.15)

Bregman iterations are proven to converge to a total variation minimizing solution of \(Af = g\) (1.1), so early stopping is required in order to provide regularization; see for instance [116, chapter 6].

Bregman iterations can be extended to multiple modalities by replacing the Bregman distance \(\text{D}_{\text{TV}}^{p_k}(f, f_k)\) with a weighted sum of pairwise channel correlations

\[
\sum_{i,j=1}^M w_{ij} \text{D}_{\text{TV}}^{p_k}(f_i, f_j).
\]

(4.16)

The resulting algorithm is coined "Color Bregman iterations" [117]. Its convergence is guaranteed in some special cases; see [117] for more details. A problem with (4.16) is that it only promotes positive correlations between the edges in channels and negative correlations are suppressed. In order to circumvent this problem the Bregman distance of the total variation was replaced by the infimal convolution of Bregman distances with opposite sign subgradients. Whilst no proof of convergence for resulting iterations is known, these were shown in [51] to be competitive with the state-of-the-art for joint PET-MR reconstruction.

(g) Joint-dictionary learning methods

Methods from §3.(b)iii can be extended to the joint problem.
Sudarshan et al. recently extended their work on PET reconstruction guided by a joint-dictionary using an MRI image from the same subject [86] to the joint problem for PET and (undersampled) MRI data [118]. The proposed method uses a pre-trained joint dictionary for PET and MRI magnitude images. The MRI phase image was not included in the dictionary nor penalised, as it is usually sensitive to noise and phase wrap-around. The prior penalises the square of the differences between both reconstructed images and their denoised images (obtained by sparse coding). The method alternates between PET, MRI phase and magnitude reconstruction, and sparse coding. The method performed well on simulated data. However, a practical difficulty would be how to obtain high quality data for training the dictionary.

Song et al. proposed an algorithm for multi-contrast MRI that alternates three stages: coupled dictionary learning (from random patches of current images), coupled sparse denoising (of all patches), and $k$-space consistency (as a gradient-step of normal image reconstruction with a quadratic penalty that encourages similarity between the denoised and reconstructed image) [119]. An interesting point is that multiple dictionaries are used: a coupled dictionary for all contrasts, and independent dictionaries for each contrast. The latter are used to encode residual image features that do not fit the coupled dictionary. As with any adaptive dictionary learning problem, this problem is highly non-convex and the authors acknowledge that the suggested approach will at most converge to a local minimum, with no convergence proofs available.

5. Discussion and Outlook

In this article we have given a (somewhat personal) overview of the history and current state of synergistic reconstruction, with an emphasis on medical imaging. We have attempted to frame a common setting for commonly used methods and present some of the key concepts that arise in quite different applications. Due to space constraints as well as consistency of readability we have inevitably had to omit certain topics which we discuss here. At the same time we indicate the current challenges which will motivate the outlook for research in the next several years.

We have implicitly assumed that images have been acquired and reconstructed in a consistent coordinate system. This may not always be the case. In the case of multiple modalities, images might have different natural sampling, or, more generally, different basis functions. It is possible to insert the necessary transformations into the joint priors, although in practice most authors choose the highest sampling for all images. If data-sets are not acquired at the same time, misalignment due to motion can create difficulties with some methods more sensitive to misalignment than others [120]. This leads to the topic of joint misalignment estimation and reconstruction, with recent contributions in the guided context jointly estimating the image and the misalignment modelled either via a convolution kernel [121] or a spatial transformation [122,123]. If there is motion during the acquisition, it is possible to jointly estimate the image and the spatial transformation between motion different states [124,125]. There is considerable scope to combine this with methods from the image registration literature, such as joint penalties on the image and motion fields [126] and synergistic image registration for multi-modality data with the same underlying motion which benefits from different contrast and structures in the different modalities [127,128].

There are also several open questions in regard to algorithm choices and their implementation. Should (4.3) be solved with grouped coordinate descent or should all coordinates be updated in parallel? What are efficient algorithms to solve (4.3)? Since the computational cost of separate reconstruction of all channels/modalities scales linearly with respect to their number, one might aim to have a similar computational cost for synergistic reconstruction. It is currently unknown how existing methods theoretically scale in this regard. Another challenge is the non-convexity of many JMAP approaches, with in some cases presence of local minima. In such a case, heuristic update mechanisms are often used, for instance by using small steps in an alternating update scheme such that artefacts do not arise.

There is an ever increasing trend towards learning based methods in imaging, which is also seen in almost all branches of science and data analysis in general. See [129–131] for some
reviews on learned image reconstruction methods and appendix §A(f) for a brief discussion on the use of Neural Networks (NNs) as regularisers. There is however currently very little research on applying these techniques to the multi-channel/modality problem, possibly due to the relative scarcity of data and the difficulties in obtaining ground truth in the situations that synergistic image reconstruction tries to tackle. Future research in this area is likely to emphasise the unification of model-driven and data-driven methods.

A practical difficulty with synergistic methods, especially in multi-modality imaging, is the need for software that can handle large amounts of data, is capable to accurately compute the system models (2.3), and ideally allows easy experimentation with novel algorithms. It is therefore often necessary to combine several software packages, ideally via an overarching framework [132] or by writing interfaces to other packages such that they can be used in an optimisation library such as [133–135].

Finally, although synergistic methods hold considerable promise to expand imaging into application areas where ill-conditioning of the single channel/modality otherwise impedes sufficient image quality, it also comes with its dangers. Like any regularisation, prior information can generate bias. This can lead to cross-talk, creating structure in one image when it should only be present in the other. Addressing this will need both prior design [28] and extensive studies to validate these methods for each application.

Acknowledgements

Funding. This work was funded by the Computational Collaborative Project in Synergistic PET-MR Reconstruction, CCP PET-MR, UK EPSRC grant EP/M022587/1; its associated Software Flagship project EPSRC grant EP/P022200/1 and the Computational Collaborative Project in Synergistic Reconstruction for Biomedical Imaging, CCP SyneRBI, UK EPSRC grant EP/T026693/1. MJE acknowledges support from the EPSRC (EP/S026045/1, EP/T026693/1), the Faraday Institution (EP/T007745/1) and the Leverhulme Trust (ECF-2019-478).

Acknowledgements. We thank our many colleagues and students, who are too numerous to mention individually, for interesting and stimulating discussions over the years.

References

4. Pruessmann KP. 2006 Encoding and reconstruction in parallel MRI. *NMR in biomedicine* 19, 288–299.


40. Matthews TP, Anastasio MA. 2017 Joint reconstruction of the initial pressure and speed of sound distributions from combined photoacoustic and ultrasound tomography measurements. *Inverse Problems* 33.


92. Bresson X, Chan TF. 2008 Fast dual minimization of the vectorial total variation norm and applications to color image processing. *Inverse Problem & Imaging* 2, 455.


A. Mathematical Tools for Image Reconstruction

In this appendix we briefly outline some main concepts and methods used in image reconstruction, which we frame in the context of ill-posed inverse problems. There is an ever growing number of strategies for the design of regularisation functionals to ameliorate ill-posedness using geometrical, statistical or learned approaches. We mention some key principles here. For a general introduction to the regularization of inverse imaging problems, we refer the reader to [116,136,137].

We note that there are some fundamental differences between the Continuous Setting and the Discrete Setting which are related through (2.1). Any approach derived in the continuous setting needs to be discretised to allow a computational implementation. However, some approaches start directly from a discretisation, where the continuous functions are written in terms of basis functions, normally as a linear combination as in (2.1). Regularisation can then proceed by choosing appropriate basis functions, or by adding a penalty $\Psi(f)$ on the coefficients, or both.

(a) Aspects of Optimisation

The fundamental algorithm to optimise the variational representation (2.2) is gradient descent which, given an initial guess $f^{(k)}$, computes the iterative updates

$$f^{(k+1)} = f^{(k)} - \tau_k \left(D'(f^k) + \alpha \Psi'(f^k)\right)$$

where $\tau_k$ is a step length which can be constant or varying with the iterations. Rather than an explicit classical descent strategy, a more general approach is the Proximal Gradient Descent (PGD) which takes the form of two steps

likelihood update step $f^{(k+1/2)} = f^{(k)} - \tau_k D'(f^k)$

proximal step $f^{(k+1)} = \text{prox}_{\tau_k \alpha \Psi}(f^{(k+1/2)})$

where the proximal operator solves the auxiliary problem

$$\text{prox}_{\tau_k \alpha \Psi}(z) = \arg \min \{ \Phi_z(f) := \frac{1}{2} \| f - z \|^2 + \tau_k \alpha \Psi(f) \}$$

A possible approximate solution of the proximal step is via the evolution of a PDE induced by the form of the regularisation functional [136]

$$\frac{\partial f}{\partial t} = -\Phi_f^{(k+1/2)}(f) = f - f^{(k+1/2)} + \tau_k \alpha \Psi'(f)$$

(b) Regularisation and Image Diffusion

In the continuous setting, when following the classical approach (A 1), we require the gradient $\Psi' \equiv \frac{\partial \Phi}{\partial f}$, which corresponds to the first variation (Euler-Lagrange equation) if $\Psi$ is defined in variational form; e.g.,

$$\Psi(f) := \int_\Omega \psi(|\nabla f|)dx \quad \Rightarrow \quad \Psi'(f) = -\nabla \cdot \left( \frac{\psi'(|\nabla f|)}{\kappa} \right) \nabla f$$

Furthermore, as suggested by (A 5), interpreting the iterative steps as a time evolution suggests the interpretation of the minimisation of the prior as an image flow. In the choice given in (A 6) this is of diffusion type, since the term on the right is a second order derivative, i.e.

$$\frac{\partial f(x)}{\partial t} = \nabla \cdot (\kappa(x) \nabla f(x))$$

where $\kappa(x)$ plays the role of a spatially varying diffusivity. The local function $\psi$ in (A 6) (usually taken to be convex) admits many commonly used regularisation schemes including first order
Tikhonov ($\psi(s) = \frac{1}{2}s^2$) and total variation (TV) ($\psi(s) = s$). A particular function that we will refer to in this article is the Perona-Malik function \[138\] in one of the forms

\[
\begin{align*}
\text{PM1} & \quad \psi(t) := \frac{\epsilon^2}{2} \log \left( 1 + \frac{t^2}{\epsilon^2} \right) \quad \Rightarrow \quad \psi'(t) = \frac{\epsilon^2 t}{\epsilon^2 + t^2} \quad (A\ 8) \\
\text{PM2} & \quad \psi(t) := \frac{\epsilon^2}{2} \left[ 1 - \exp \left( -\frac{t^2}{\epsilon^2} \right) \right] \quad \Rightarrow \quad \psi'(t) = t \exp \left( -\frac{t^2}{\epsilon^2} \right) \quad (A\ 9)
\end{align*}
\]

with $\epsilon$ a threshold indicating a level below which small gradients are considered as noise. The resultant diffusivities $\kappa = \psi'(|\nabla f|)$ can be interpreted as edge-indicator functions.

We may also define a flow without it being the variation of a functional form. Weikert \[102\] proposed using a tensor

\[
\frac{\partial f}{\partial t} = \nabla \cdot (D(J_\rho(\nabla f)) \nabla f) \quad (A\ 10)
\]

where the (symmetric, positive semi-definite) structure tensor is constructed as

\[
J_\rho(\nabla f)(x) = G_\rho * (\nabla f \nabla f^T) = \begin{bmatrix} G_\rho * f_x^2 & G_\rho * f_x f_y \\ G_\rho * f_x f_y & G_\rho * f_y^2 \end{bmatrix},
\]

with eigensystem \{\eta_k, v_k\} and

\[
D = \begin{bmatrix} v_1^T & v_2^T \end{bmatrix} \begin{bmatrix} \zeta_1 & 0 \\ 0 & \zeta_2 \end{bmatrix} \begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{bmatrix}. \quad (A\ 12)
\]

Furthermore $v_1 \approx \hat{\nu}, v_2 \approx \hat{\tau}$, with equality as $\sigma, \rho \to 0$. Then the Edge Enhancing Diffusion (EED) approach is designed to reduce smoothing in the normal direction by an edge-indicator derived from (A 9)

\[
\zeta_1 = \kappa(\nabla f) = \exp (-|\nabla f|^2/\epsilon^2); \quad \zeta_2 = 1 \quad (A\ 13)
\]

(c) Markov Random Fields

A fundamental concept in the discrete setting is the Markov Random Field (MRF) \[139\] : \[
\Psi(f) = \frac{1}{p} \sum_i \sum_{j \in \mathcal{N}(i)} w_{ij} |f_i - f_j|^p \quad (A\ 14)
\]

\[
\frac{\partial \Psi}{\partial f_i} = \sum_{j \in \mathcal{N}(i)} w_{ij} |f_i - f_j|^{p-1} \quad (A\ 15)
\]

Sometimes the discrete MRF corresponds to the discretisation of a continuous functional; e.g for $p = 2$ we have a quadratic form

\[
\Psi(f) = \langle f, Lf \rangle \quad \text{with} \quad L_{ij} = \begin{cases} w_{ij} & j \neq i \\
-\sum_{j \neq i} w_{ij} & i = j \end{cases} \quad (A\ 16)
\]

Taking a 4-connected neighbourhood with uniform weights $w_{ij} = 1$ corresponds to the discretisation of the Laplacian $L \approx -\nabla^2$ which is the gradient of the first order Tikhonov prior. However, in general, it is not always possible to explain MRFs as discretisation of a continuous model.

The MRF concept extends to a global one and the concept of Non-local regularisation \[140\]. For example the Laplacian in the local mathematical sense of a second order derivative extends to the Graph Laplacian \[141–145\]. The conjunction of an MRF with kernel-based methods (see § A(e)) is related to the so-called bilateral filtering technique in image processing \[146\].
(d) Sparsity

Sparsity has been a prominent concept in regularisation for several decades, closely connected to the principles of Compressed Sensing [147]. It can be formulated in both continuous and discrete settings. The assumption is that under some (possibly invertible) transform

$$\xi = T(f) \leftrightarrow f = T^{-1}(\xi) \quad (A\ 17)$$

the transformed parameters $\xi$ have many/mostly zero or close to zero, which can be described by specifying that the zero-norm $L_0 := ||\xi||_0$, which simply counts the non-zero components in $\xi$, should be minimised. However, since optimisation of $L_0$ norm leads to a non-convex problem, it is conveniently replaced by its convex relaxation given by the $L_1$-norm. Possibilities for the transform $T$ include finite differences (gradient regularisation), wavelets, and NNs where $T$ is called an encoder and $T^{-1}$ a decoder. Methods employing sparsity also allow the possibility that the transform space is over-complete, i.e. the solution has a non-unique representation in the basis; an example is the use of a dictionary learned from example solutions [148,149].

(e) Kernel Methods

Kernel Methods compute properties such as classifiers in terms of transformed feature vectors obtained from one or more related images or training data [150]. The transformation can be a (generally non-linear) mapping to a high dimensional space. The features may be obtained from any abstraction model, such as patches, or geometric or statistical measures. The inner product in the transformed space defines a similarity measure on the features. In practice, the transformation is determined by specifying this similarity measure $K$, called the “kernel function”. Although sparsifying transforms and feature vectors are not synonymous, for simplicity of representation, we use the same notation as (A 17)

$$K(T_i(f), T_j(f)) \quad (A\ 18)$$

where $T_i$ is now interpreted as the feature vector associated with pixel $i$ and $f$ represents the image(s) on which the features are computed. A common choice for the kernel function is the Radial Basis Function (RBF)

$$K(t_1, t_2) := \exp\left\{ -\frac{||t_1 - t_2||^2}{2\sigma^2} \right\}. \quad (A\ 19)$$

The “kernel trick” consists of computing linear functions of the transformed variables in terms of $K$.

(f) Regularisation using Neural Networks

There are many techniques for combining inverse problems with artificial intelligence in general and NNs particular; see [129] for a review. Here we mention only one natural approach which is to replace the proximal operator (A 3) by a learned operator such that

$$f^{(k+1)} = F_\Theta\left(f^{(k+1/2)}\right) \quad (A\ 20)$$

where $F_\Theta : X \to X$ is a NN trained on suitable pairs of ground truth and approximately reconstructed images, and where $\Theta$ represents the weights and other parameters of the network architecture [151].

Alternatively we could consider the evolution (A 5) as an image update

$$f \mapsto f + F_\Theta\left(f^{(k+1/2)}\right) \quad (A\ 21)$$

where $F_\Theta$ now takes the form of a Residual Neural Net [152].

In neither approach is the update explicitly derived as the variation of a function which leads to difficulties in corresponding convergence guarantees, and prevents an explicit Bayesian
interpretation as the maximisation of a posterior. However, note that the Regularisation by Denoising (RED) framework provides a general technique to interpret denoising algorithms as variational methods under certain restrictions of their properties [153].