HAL open science

Spectral Total-Variation Local Scale Signatures for Image Manipulation and Fusion

Ester Hait, Guy Gilboa

To cite this version:

Ester Hait, Guy Gilboa. Spectral Total-Variation Local Scale Signatures for Image Manipulation and Fusion. 2018. hal-01722459

HAL Id: hal-01722459

https://hal.science/hal-01722459v1

Preprint submitted on 4 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Spectral Total-Variation Local Scale Signatures for Image Manipulation and Fusion

Ester Hait and Guy Gilboa, *Member, IEEE*

***Abstract*—We propose a unified framework for isolating, comparing and differentiating objects within an image. We rely on the recently proposed total-variation transform, yielding a continuous, multi-scale, fully edge-preserving, local descriptor, referred to as spectral total-variation local scale signatures.**
We show and analyze several useful merits of this framework. Signatures are sensitive to size, local contrast and composition of structures; are invariant to translation, rotation, flip and linear illumination changes; and texture signatures are robust to the underlying structures. We prove exact conditions in the 1D case.

We propose several applications for this framework: saliency map extraction for fusion of thermal and optical images or for medical imaging, clustering of vein-like features and size-based image manipulation.

***Index Terms*—Spectral Total-Variation, Image Fusion, Image Segmentation, Edge Detection, Size Differentiation, Clustering, Saliency, Thermal Imagery, Medical Imagery.**

I. Introduction

DIFFERENTIATING objects within an image by contrast, size or structure is a fundamental image processing task. It is highly useful for various image modalities and applications, such as image clustering, enhancement and fusion. A key feature in many modalities (natural images, medical, thermal, depth etc.) are edges, or discontinuities in the data.

For this purpose, the spectral total variation transform (spectral TV) has been recently introduced as a useful edge-preserving, multi-scale decomposition tool. However, previous spectral TV-based approaches [1],[2],[3],[4] are better suited for texture-structure decomposition than for object differentiation. Other TV or spectral TV-based approaches can differentiate objects, but only by their size [5],[6],[7]. Using spectral TV for images of different modalities is challenging, due to their complex nature, multi-scaled and occluded content.

In this paper, we present a novel, unified framework for object differentiation by contrast, size or structure, including complex multi-scaled objects. We capture salient objects by exploiting the comprehensive scale and location information extracted from spectral TV, referred to as *Spectral TV Local Scale Signatures* (Fig. 1). Stemming from an edge preserving, sparse spectral transform, signatures of significant objects are sparse and strong; their locality allows good differentiation within an image. We show and analyze the essential properties of the signatures: sensitivity to size, local contrast and composition of structures; invariance to rotation, translation, flip and linear illumination change; and invariance of texture to structure.

E. Hait and G. Gilboa are with the Department of Electrical Engineering, Technion - Israel Institue of Technology, Haifa, Israel, e-mail: etyhait@campus.technion.ac.il, guy.gilboa@ee.technion.ac.il

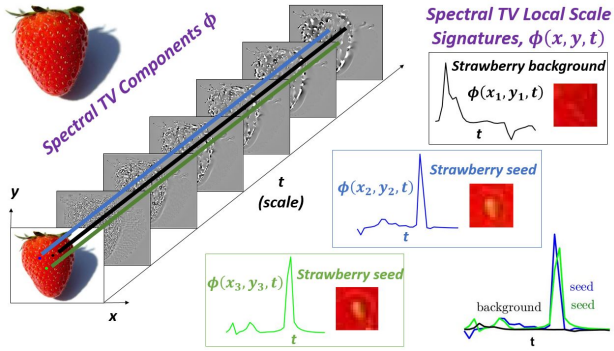


Figure 1: **Spectral TV Local Scale Signatures,** $\phi(x,y,t)$: a multi-scale spectral TV per-pixel description, sensitive to size, local contrast and composition of structures, invariant to translation, rotation, flip and linear illumination change, with texture invariance to underlying structure. Objects with common features (seeds) have similar signatures, distinct from those of different objects.

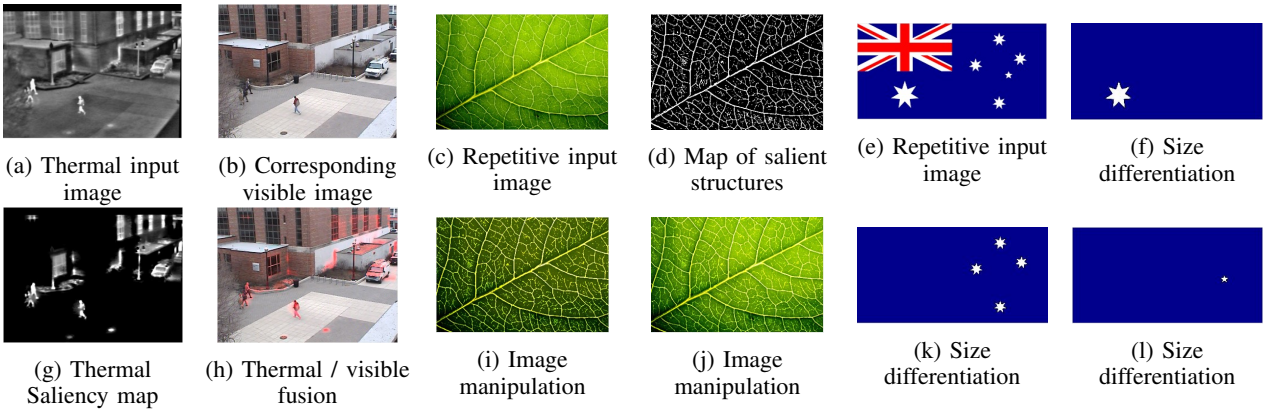
Though relying merely on a few simple cornerstones, our algorithm is applicable to various tasks (Fig. 2): fusion of thermal and visible images, or of medical images of different modalities; and clustering of vein-like repetitive structures (segmentation / edge detection) and size differentiation.

The rest of the paper is organized as follows. Sec. II briefly surveys previous image descriptors. Sec. III introduces spectral TV (Sec. III-A), and reviews and analyzes previous approaches. Sec. IV presents the novel concept of spectral TV signatures and their properties. Sec. V includes a theoretical analysis. Sec. VI presents a unified framework for object differentiation, demonstrated for synthetic images, and applied for image manipulation (VI-C) and fusion (VI-D). Sec. VII presents experimental results and comparisons to other methods . Sec. VIII concludes our work. In the Appendix we provide further analysis and suggest new visualization methods for fusion.

II. Related Work

In the past decades, high-level image understanding and processing has considerably relied on feature descriptors of different types. For example, sparse descriptors, describing characteristics of interest points only (e.g. corners) [8], [9].

Our focus here, however, is on dense descriptors, describing every image pixel using properties of the pixel and its surrounding. Various algorithms [6], [10], [11], [12] rely on



(a) Thermal input image (b) Corresponding visible image (c) Repetitive input image (d) Map of salient structures (e) Repetitive input image (f) Size differentiation

(g) Thermal Saliency map (h) Thermal / visible fusion (i) Image manipulation (j) Image manipulation (k) Size differentiation (l) Size differentiation

Figure 2: Image fusion (a-b,g-h) and image manipulation (c-f,i-l) using spectral TV local scale signatures.

fundamental properties, such as intensity and color, gradient magnitude and orientation, textures and patterns. Some of these algorithms use features for learning [13], [14]. Among these descriptors, those based on histograms of patterns [15] or of oriented gradients [16] are rotation-variant, thus suitable for texture applications. Other approaches include multi-scale decompositions using image transforms [17] or diffusion [18]; or using as features the response to a set of predefined linear filters [19], [20], [21]. Recently, convolutional neural networks (CNN) have emerged as a highly successful data-driven tool [22], [23], [24]. However, this feature learning approach is not generic and lacks solid theoretical background. To overcome this, recent CNN methods use sets of linear filters in their first convolutional layers [25], [26].

In this paper, we present a novel dense, multi-scale, edge preserving, local descriptor. Its properties are highly suited for detecting and clustering vein- and disk-like structures and for constructing saliency maps for fusion.

III. Preliminaries

A. Spectral Total Variation

The total variation (TV) functional [27] has been widely used as an image regularizer, e.g. for denoising and deconvolution [27],[28],[29],[30], decomposition and texture analysis [31],[32],[33] and fusion [34],[35]. Denoting image domain as Ω, and the gradient (understood as the distributional gradient) as ∇, the TV functional is defined as:

$$J(u) = \int_{\Omega} |\nabla u(x)| dx. \tag{1}$$

For an input image $f(x)$, the gradient descent evolution of this functional, known as TV flow [36], is defined as,

$$\begin{aligned} \frac{\partial u}{\partial t} &= \text{div}\left(\frac{\nabla u}{|\nabla u|}\right) \quad \text{in } (0,\infty)\times\Omega, \\ u(x,0) &= f(x) \quad \text{in } x \in \Omega, \end{aligned} \tag{2}$$

with Neumann boundary conditions (note that the right-hand-side of (2) should be understood as a negative subgradient of J). Performing TV flow, up to a scale T, is analogous to solving the ROF [27] optimization problem [1]:

$$\arg\min_u \left\{ J(u) + \frac{1}{2T}\|f-u\|_2^2 \right\}. \tag{3}$$

Denoting $t \in [0,\infty)$ as the **time** or **scale** parameter of the TV flow (2), and u_{tt} as the second time derivative of its solution $u(x,t)$, the spectral TV transform [37] is defined by,

$$\phi(x,t) = u_{tt}(x,t)t. \tag{4}$$

We refer to $\phi(x,t)$ as the spectral component or band at scale t (see Fig. 1, left for a visualization of ϕ at some scales). The main properties of this nonlinear transform are as follows:

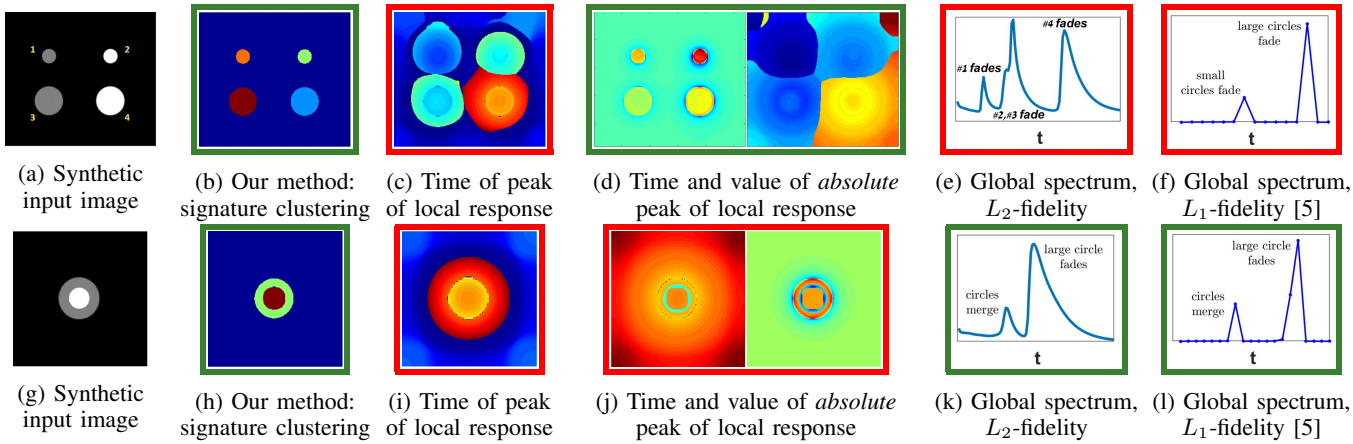
1) **Reconstruction.** Any zero mean, bounded variation function f can be reconstructed by $f(x) = \int_0^\infty \phi(x,t)dt$.
2) **Spectral representation.** We define a nonlinear eigenfunction with respect to the subdifferential of TV, $\partial J(u)$, as: $\lambda u \in \partial J(u)$. Examples of such eigenfunctions are disks, or convex sets with low curvature on the boundary, bounded by the perimeter to area ratio [36]. Then the spectral response of an eigenfunction f with eigenvalue λ is a Dirac delta at scale $1/\lambda$: $\phi(t) = \delta(t-1/\lambda)f$. Thus, a signal composed of spatially-separated eigenfunctions has a highly sparse representation (analogous to Fourier transform representation of sine functions).
3) **Orthogonality.** Under some conditions, such as in the discrete 1D case, spectral components are orthogonal to each other: $\langle \phi(t_1), \phi(t_2)\rangle = 0$, $\forall t_1 \neq t_2$ [1].
4) **Filtering.** Given a filter $H(t)$, extending the reconstruction formula, $f_H(x) = \int_0^\infty H(t)\phi(x,t)dt$, allows the design of various edge-preserving TV filters.
5) **Translation and rotation invariance.** The ϕ components inherit the properties of the TV functional and are translation and rotation invariant.

These properties apply to the multiscale representation $\phi(x_0,t)$ of each pixel x_0. We show how this characterization of pixel relations with its neighborhood can successfully serve as a generic pixel descriptor.

We later use two additional definitions. First, the spectrum $S^f(t)$ is defined (using the original definition of [37]) as:

$$S^f(t) = \|\phi(x,t)\|_{L^1} = \int_{\Omega} |\phi(x,t)| dx, \tag{5}$$

which can be seen as the L^1 amplitude of the response at each scale $t \in [0,\infty)$. Second, the residual image $f_r(x)$, generated



(a) Synthetic input image (b) Our method: signature clustering (c) Time of peak of local response (d) Time and value of *absolute* peak of local response (e) Global spectrum, L_2-fidelity (f) Global spectrum, L_1-fidelity [5]

(g) Synthetic input image (h) Our method: signature clustering (i) Time of peak of local response (j) Time and value of *absolute* peak of local response (k) Global spectrum, L_2-fidelity (l) Global spectrum, L_1-fidelity [5]

Figure 3: Synthetic analysis and comparison of our method to previous spectral TV-based Approaches: successes vs. failures. **No spectral-TV based method but ours can differentiate objects by size and contrast, as well as a composition of structures**.

after some *finite* time T, is defined as:

$$f_r(x,T) = u(x,T) - u_t(x,T) \cdot T. \tag{6}$$

Further discussions can be found in [37].

B. Previous Spectral TV-based Approaches

Most previous spectral TV-based methods perform *global* scale analysis, that is, spatially integrating spectral information for each scale. Global spectral methods using the TV-flow (or L_2-fidelity as in (3)) have been successfully used for texture extraction and manipulation [1], [2], [4]. However, differentiating objects *within* an image is challenging, as different types of objects may correspond similarly in the spectrum peak (although the full signature is different).

Another successful global approach uses an L_1-fidelity term in (3) for *contrast-invariant* multi-scale decomposition, differentiating objects only by their size [5]. However, we are interested in differentiating objects both by size and contrast. The only previous *local* approach, performing scale analysis for each spatial location, was suggested for structure-texture decomposition [3]. This method relies on the time of the response peak of each pixel to fit a separation surface between two bands of spectral layers: texture and structure. Despite its success for texture-structure decomposition, it is not suited for object differentiation. The approach of [3] is simple but limited, as it uses only the response *peak* information.

Toy Examples Analysis: We analyze two key synthetic examples to study the limitations of previous spectral TV-based methods (Fig. 3).

Objects of various sizes and contrasts (Fig. 3a) cannot all be differentiated using global spectral TV. Two objects (one double the size, but half the contrast of the other) respond simultaneously, reflected in a mutual peak (Fig. 3e) in the global spectrum (5). In addition, the contrast-invariant global L_1-*fidelity* method [5] cannot differentiate same-size, different-contrast objects. Its global spectrum (5) shows simultaneous responses for same-contrast objects (Fig. 3f). A local approach using the time of peak of local response for differentiation (following [3]) also fails, as different objects and background regions respond simultaneously (Fig. 3c). However, we can differentiate foreground from background and all objects from each other (Fig. 3d) by combining the time t_{peak} and signed value v_{peak} of the absolute peak of local response, defined as:

$$[v, t_{peak}](x) = max_t(|\phi(x,t)|), \;\; v_{peak}(x) = \phi(x, t_{peak}).$$

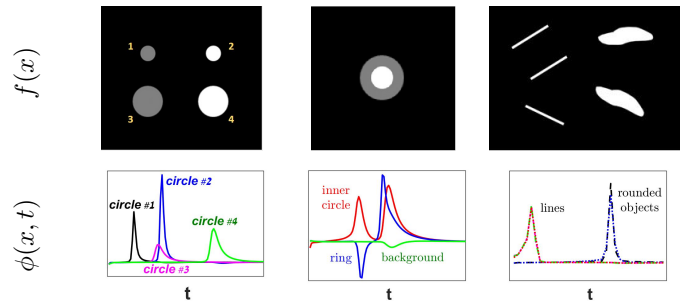
A synthetic composition of structures (Fig. 3g) simulates real multi-scaled images. We first note objects' impact on each other's behavior: objects first merge, then the merged object fades. Both global methods can approximately differentiate objects (Figs. 3k,3l), though the L_1-fidelity method is about 9 times slower. However, differentiating by the time of the peak of local response fails (Fig. 3i); and so does combining the features of the *absolute* peak of local response (Fig. 3j).
In conclusion, no previous spectral-TV based method can differentiate objects by size and contrast, as well as a composition of structures. We thus suggest to exploit more spectral TV information within a local framework.

IV. Spectral TV Local Scale Signatures

Object differentiation requires exploiting detailed, local, multi-scale information to handle different sizes, contrasts and complex structures. We thus introduce the concept of spectral TV local scale signatures. We denote the signatures of a signal $f(x)$ as $\phi_f(x,t)$, where ϕ is defined by (4). For each pixel there exists a well-defined, unique representation in the scale continuum (unlike classical pyramidal multiscale representations and wavelets), yielding a natural multi-scale pixel descriptor.

A. Properties of Spectral TV Signatures

We now summarize the main properties of the spectral TV signatures and illustrate them graphically by simple toy examples. An elaborated theoretical analysis for the 1D case is given in Sec. V.

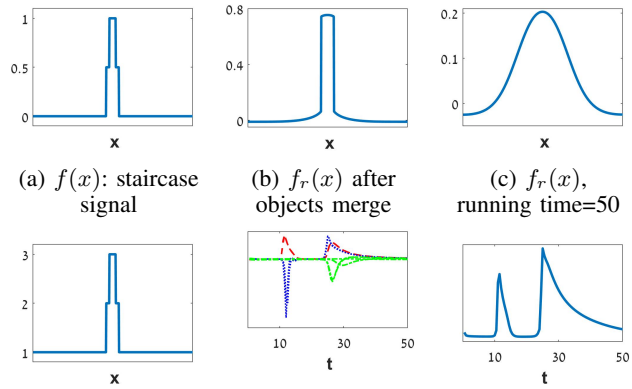


(a) Property 1: Sensitivity to Size and Local Contrast

(b) Property 2: Sensitivity to Composition of Structures

(c) Property 3: Rotation and Translation Invariance

Figure 4: 2D Demonstration of properties 1, 2, 3. Signatures are distinct due to their sensitivity to size and contrast (a), and their sensitivity to composition of structures (b). However, they are invariant to rotation and translation (c).



(a) $f(x)$: staircase signal

(b) $f_r(x)$ after objects merge

(c) $f_r(x)$, running time=50

(d) K-means clustering using $\phi(x,t)$

(e) $\phi(x,t)$ of top (red), middle (blue) and background (green) regions

(f) Global spectrum $S^f(t)$

Figure 5: Demonstration of Property 2 for the 1D staircase signal (a). Spectral TV signatures, generated during TV flow of the signal (b,c), are distinct for different regions (e). Thus, signature clustering allows differentiating the regions (d).

Sensitivity of the Signatures:

Property 1. *Sensitivity to Size and to Local Contrast*

Signatures are sensitive to change in size (spatial scaling by factor a) and change in local contrast (contrast change by factor a) by the following relations:

$$\begin{aligned} \phi_{af(x)} &= \phi_f(x, t/a), \\ \phi_{f(ax)} &= a\phi_f(ax, at). \end{aligned} \tag{7}$$

This results from spectral TV properties [37]. This allows differentiating objects by their distinct signatures (Fig. 4a).

Property 2. *Sensitivity to Composition of Structures*

Structures of comparable scales respond differently when composed together (with spatial overlay), compared to their individual responses. That is, in this case the non-linearity of the TV transform applies, allowing for some signals f, g:

$$\phi_{f+g} \neq \phi_f + \phi_g. \tag{8}$$

Composited objects thus have distinct signatures (Fig. 4b). An analytical solution (Appendix A, Proposition 1) and a demonstration (Fig. 5) are given for the staircase signal. Note, that non-linearity is marginal for structures of very different scales or for spatially distant ones.

Invariance of the Signatures:

Property 3. *Invariance to Rotation, Translation and Flip*

Signatures are rotation and translation invariant in $\mathbb{R}^n$:

$$\begin{aligned} \phi_{f(Rx)} &= \phi_f(Rx, t), \\ \phi_{f(x-a)} &= \phi_f(x-a, t), \end{aligned} \tag{9}$$

where $R(x)$ is a rotation matrix, and a is a spatial shift (Fig. 4c). This also results from spectral TV properties [37].
Since TV is invariant to the coordinate system (being rotationally invariant and sensitive only to derivatives), signatures are also invariant to flip w.r.t. an arbitrary coordinate system:

$$\phi_{f(x)}(x) = \phi_{f(-x)}(-x). \tag{10}$$

It is shown in Fig. 7 how these properties can be useful for finding similar image textures, where patch-based comparison fails.

Property 4. *Invariance of Texture to Structure*

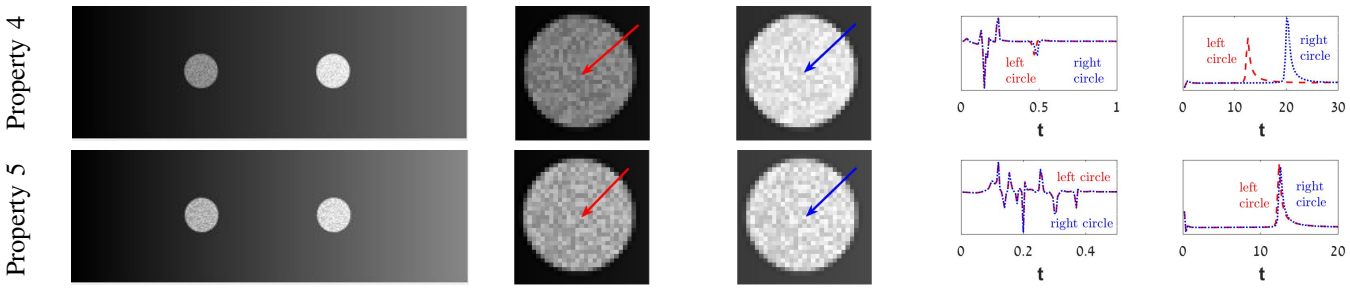
Signatures of textures (patterns) are invariant to their underlying structure, up to a certain scale, under very broad conditions. This holds for fine scales (precise conditions are given in Sec. V-C). Fig. 6 (top) shows the invariance of signatures of objects with identical textures to their different underlying global contrasts for fine scales. A 1D proof (Sec. V-C, Theorem 1) and a 1D demonstration (Fig. 9, top) are given.

Property 5. *Invariance to Linear Illumination Change*

Signatures are invariant to a linear change of illumination, which holds up to a certain scale (precise conditions are given in Sec. V-C). Fig. 6 (bottom) shows the invariance of signatures of objects with identical textures and identical global contrasts. A 1D proof (Sec. V-D, Theorem 2) and a 1D demonstration (Fig. 9, bottom) are given.

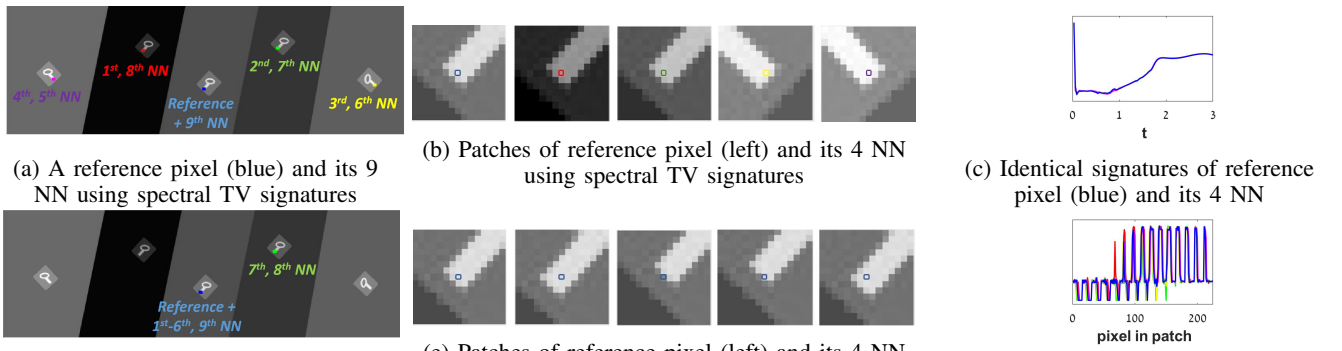
B. Implications for General Images

In Fig. 8a we show some examples where the signatures are similar and where they are distinct from each other. Comparing the signatures of different-size, same-color objects in Fig. 8a, signatures of small white stars display stronger, earlier responses, distinct from those of the larger white stripes (Fig. 8d). For thermal and medical images (Fig. 8b), signatures of objects different in size or contrast are distinct (Fig. 8c). Moreover, signatures of highly-contrasted or large objects display stronger responses than those of weakly-contrasted or small ones (Fig. 8e). Signature enhancement (see Sec. VI-D) improves the distinctness of signatures of different groups (Fig. 8f). This is useful since the salient objects in these modalities are usually highly-contrasted.



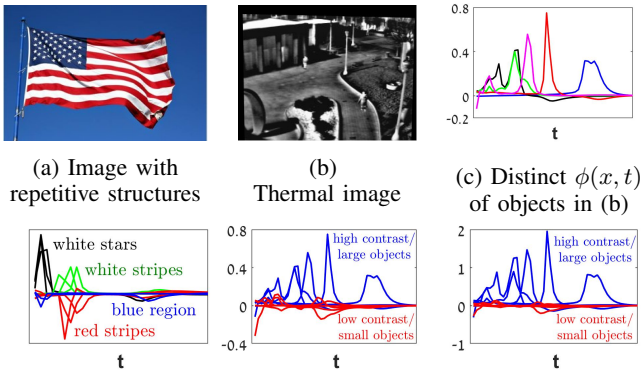
(a) $f(x)$ (b) Left circle: zoom-in (c) Right circle: zoom-in (d) $\phi(x,t)$: fine scales (e) $\phi(x,t)$: fine and coarse scales

Figure 6: 2D Demonstration of properties 4,5. Signatures of objects with **identical textures (patterns)** are invariant to their underlying structures - **different global contrasts** - for fine scales (top). Signatures of objects with **identical textures** and **identical global contrasts** are identical - thanks to their invariance to linear illumination changes (bottom).



(a) A reference pixel (blue) and its 9 NN using spectral TV signatures (b) Patches of reference pixel (left) and its 4 NN using spectral TV signatures (c) Identical signatures of reference pixel (blue) and its 4 NN (d) A reference pixel (blue) and its 9 NN using patch intensity (e) Patches of reference pixel (left) and its 4 NN using patch intensity (f) Dissimilar patch intensity descriptors of reference pixel (blue) and its 4 NN

Figure 7: 2D Demonstration of the invariance to rotation, translation and flip, and the invariance of textures to their underlying structures. Using spectral TV signatures, the 9 NN of a reference pixel (reference marked in blue) are pixels of similar textures, but of different global contrasts, rotations, translations or flips (top). Conversely, the patch intensity descriptor fails to find these texturally-similar pixels (bottom).



(a) Image with repetitive structures (b) Thermal image (c) Distinct $\phi(x,t)$ of objects in (b) (d) Distinct $\phi(x,t)$ of groups of objects in (a) (e) Non-enhanced $\phi(x,t)$ of groups of objects in (b) (f) Enhanced $\phi(x,t)$ of groups of objects in (b) (Sec. VI-D)

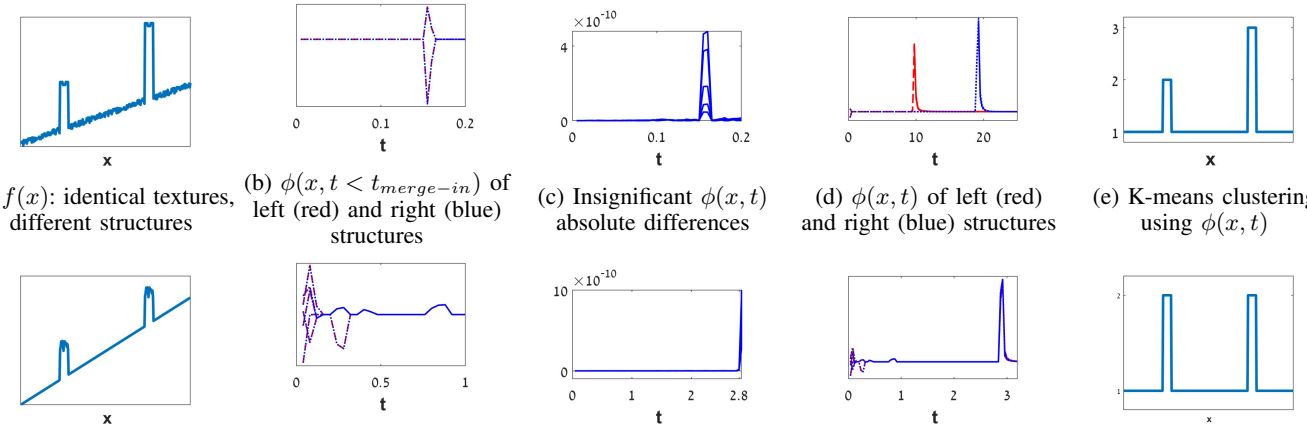
Figure 8: Distinctness of signatures of groups of objects with common features. For an image with repetitive structures (a), groups of objects have distinct signatures (d). For a thermal image (b), different objects have distinct signatures (c), and so do groups of objects (e,f).

V. Theoretical Analysis

This section presents theoretical analysis of signature properties. We formulate a sufficient condition for local patterns to merge first; and give sufficient conditions and proofs of the invariance to linear illumination change and of texture to structure. Appendix A shows an analytic solution of the separation of composited regions for the 1D staircase signal.

A. Preliminaries

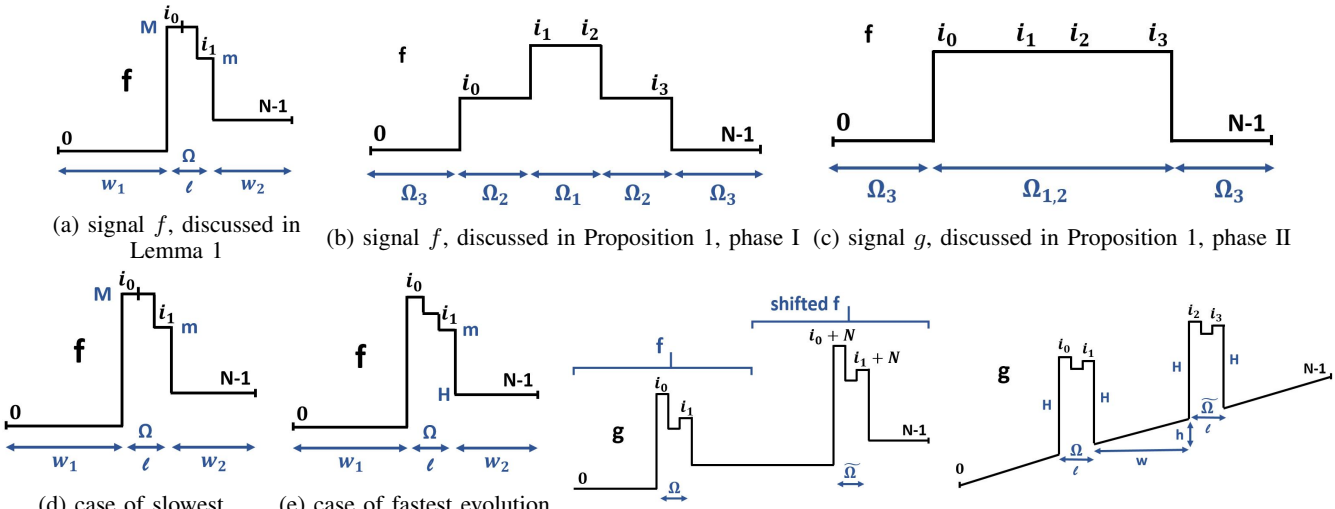
Our analysis below is based on the work of Steidl *et al.* [38], which gives an analytic solution for the TV-flow in the time continuous, spatially discrete 1D case. They shows that in a TV-flow evolution, each pixel belongs to a local constant region (all pixels in the region are connected and have the same value), which dictates its behavior. The region evolves at a certain constant speed (inversely proportional to region size), until a *merging event* occurs, that is, when two neighboring regions obtain the same value. Let $f \in \mathbb{R}^N$ be a discrete 1D input signal of size N pixels. Let $u \in \mathbb{R}^N \times [0, \infty)$ be the space-discrete realization of the TV-flow, defined by (2). We denote by $u_i(t)$ the value of u at pixel i at time t. Two main properties of this dynamic are:



(a) $f(x)$: identical textures, different structures (b) $\phi(x, t < t_{merge-in})$ of left (red) and right (blue) structures (c) Insignificant $\phi(x,t)$ absolute differences (d) $\phi(x,t)$ of left (red) and right (blue) structures (e) K-means clustering using $\phi(x,t)$

(f) $f(x)$: identical textures and structures (g) $\phi(x, t < t_{merge-in})$ of left (red) and right (blue) structures (h) Insignificant $\phi(x,t)$ absolute differences (i) $\phi(x,t)$ of left (red) and right (blue) structures (j) K-means clustering using $\phi(x,t)$

Figure 9: Demonstration of Property 4 (top) and Property 5 (bottom) for 1D signals. A signal with identical textures, but different underlying structures (a) has identical fine-scale signatures - up to $t_{merge-in}$, when local patterns merge (b,c). Signatures then become distinct (d). A signal with underlying structures of identical global contrasts (f), not only has identical signatures for $t < t_{merge-in}$ (g), but also identical signatures up to $t_{merge-struct}$, when structures merge with background simultaneously (h,i). Thus, signature clustering allows differentiating between structures (e), or structures from background (j).



(a) signal f, discussed in Lemma 1 (b) signal f, discussed in Proposition 1, phase I (c) signal g, discussed in Proposition 1, phase II

(d) case of slowest evolution of $u_{i\in\Omega}$, see Lemma 1 (e) case of fastest evolution of $u_{0<i<N-1}$, see Lemma 1 (f) signal g, discussed in Theorem 1 (g) signal g, discussed in Theorem 2

Figure 10: Signals discussed in Lemma 1 (a,d,e), Theorem 1 (f), Theorem 2 (g) and Proposition 1 (b,c).

1) There exists a finite number of merging events, $0 = t_0 < t_1 < ... < t_{n-1} < t_n$ (Proposition 4.1 (ii) in [38]).
2) Within the time intervals between merging events, $t \in [t_j, t_{j+1})$, all pixels u_i, belonging to a constant region $\{u_{i-l+1}, ..., u_{i+r}\}$ of size w_{i,t_j} with relation μ_{i,t_j} to its neighboring regions, evolve linearly (4.1 (iii) in [38]):

$$u_i(t) = u_i(t_j) + \mu_{i,t_j} \frac{2(t - t_j)}{w_{i,t_j}},$$

$$\mu_{i,t_j} = \begin{cases} 0, \ if \ \{u_{i-l}, ..., u_{i+r+1}\} \ is \ strictly \ monotonic \\ 1, \ if \ u_i \ is \ minimal \ in \ \{u_{i-l}, ..., u_{i+r+1}\} \\ -1, \ if \ u_i \ is \ maximal \ in \ \{u_{i-l}, ..., u_{i+r+1}\}. \end{cases} \quad (11)$$

B. Local Patterns Merge First

This section gives a 1D proof that regions of local patterns merge first, and only then merge with their surroundings.
Let $f : \{0, ..., N-1\} \to \mathbb{R}$ be as depicted in Fig. 10a, and let $\Omega_1 = \{0, ..., i_0 - 1\}, \Omega = \{i_0, ..., i_1\}, \Omega_2 = \{i_1 + 1, ..., N-1\}$, where Ω_1, Ω_2 are constant regions (no patterns outside Ω). We define: $w_1 \triangleq i_0$, $l \triangleq i_1 - i_0 + 1$, $w_2 \triangleq N - 1 - i_1$, and assume, without loss of generality, that $l < w_2 < w_1$ and $f[i_1 + 1] > f[i_0 - 1]$. We define: $m \triangleq \min f_{i\in\Omega}$ is attained at i_{min}, $M \triangleq \max f_{i\in\Omega}$ is attained at i_{max}, $H \triangleq f[i_1 + 1]$.
We also define the following two **critical time points**: $t_{merge-in}$, the minimal merging time of Ω, and $t_{merge-out}$, the minimal merging time of $\{i_0, ..., N-1\}$.

Lemma 1 (Local Patterns Merge First)**.** *Let f be as defined above. If $\frac{M-m}{m-H} \leq \frac{w_2}{l}$, then $t_{merge-in} < t_{merge-out}$.*

Proof. Relying on Section V-A, the TV flow dynamics of pixel i are determined by w_i, μ_i, regardless of $u_i(t=0)$. The key concept of our proof is that μ_i depends only on pixels at the immediate edge of the region. **Thus, pixel behavior is not influenced by pixels "beyond" the derivative / edge.**
We first examine the slowest possible merging of Ω, that is, the latest time of achieving equality of i_{min}, i_{max}. The slowest speed of i_{min} can be 0 (when near the boundary, case 1 of 11). But i_{max} must always decrease at a speed of at least $2/(l-1)$, through a path of total length no more than $(M-m)$. In this case (Fig. 10d), $t_{merge-in} = (M-m)\cdot(l-1)/2$. Thus,

$$t_{merge-in} < (M-m)\cdot l/2. \tag{12}$$

We now examine the fastest possible merging of Ω with Ω_2. In this case, i_{min} is at the edge of Ω (Fig. 10e) with zero speed. Thus the merging speed is bounded by the speed of Ω_2, $2/w_2$, through a path of length of at least $(m-H)$. Thus,

$$t_{merge-out} \geq (m-H)\cdot w_2/2. \tag{13}$$

From the assumption of the Lemma we have:

$$\frac{M-m}{m-H} \leq \frac{w_2}{l} \Rightarrow (M-m)\cdot l \leq (m-H)\cdot w_2. \tag{14}$$

Thus, combining (12), (13) and (14): $t_{merge-in} < t_{merge-out}$. □

C. Invariance of Texture to Structure

This section gives precise conditions for the validity of Property 4. Let $f : \{0,...,N-1\} \to \mathbb{R}$ admit the condition defined in Lemma 1. Let $g : \{0,...,2N-1\} \to \mathbb{R}$ be a concatenation of f and a translated, value-shifted, version of f (see an example in Fig. 10f). We show that texture signatures are identical for fine scales, regardless of their underlying structures. Assuming some constants $0 < C_1 < C_2$, we define:

$$g[i] = \begin{cases} f, & 0 \leq i < N \\ f + C_1, & N \leq i < i_0 + N, i_1 + N + 1 \leq i < 2N \\ f + C_2, & i_0 + N \leq i < i_1 + N + 1. \end{cases}$$

We define the regions of identical texture (up to an additive constant) as $\Omega = \{i_0,...,i_1\}$, $\widetilde{\Omega} = \{i_0+N,...,i_1+N\}$.

Theorem 1 (Invariance of Texture to Structure)**.** *Let g be as defined above. Then there exists a time $t_{merge-in}$, such that:*

$$\phi_g(i \in \Omega, t \leq t_{merge-in}) = \phi_g(i \in \widetilde{\Omega}, t \leq t_{merge-in}). \tag{15}$$

Proof. First, relying on Lemma 1: $t_{merge-in}(g_{i\in\Omega}) < t_{merge-out}(g_{i\in\Omega})$, and $t_{merge-in}(g_{i\in\widetilde{\Omega}}) < t_{merge-out}(g_{i\in\widetilde{\Omega}})$. Second, based on (11), the speed of $u(t)$ at pixel i, $\partial_t u_i$, is invariant to translation and to change by an additive constant. Since $m_{i\in\Omega} = m_{i\in\widetilde{\Omega}}$, $\mu_{i\in\Omega} = \mu_{i\in\widetilde{\Omega}}$, flow dynamics are identical in Ω and in $\widetilde{\Omega}$ until the internal merge.
Therefore: $t_{merge-in} \triangleq t_{merge-in}(g_{i\in\Omega}) = t_{merge-in}(g_{i\in\widetilde{\Omega}})$. From the definition of ϕ (4) we deduce:

$$\phi_g(i \in \Omega, t \leq t_{merge-in}) = \phi_g(i \in \widetilde{\Omega}, t \leq t_{merge-in}).$$

□

D. Invariance to Linear Illumination Change

This section gives precise conditions for the validity of Property 5. We show that signatures of complete structures, as well as their textures, are invariant to a linear change of baseline, up to a scale related to the scale of the structure (width and height), as seen in Fig. 9i. This is as opposed to the fine scales discussed in Theorem 1, as seen in Fig. 9d.
Let $f : \{0,...,N-1\} \to \mathbb{R}$ admit the condition defined in Lemma 1, and l be as defined there. Let $g : \{0,...,2N-1\} \to \mathbb{R}$ be a concatenation of f and a translated f, with a linearly-changing baseline, as depicted in Fig. 10g, as follows:

$$g[i] = \begin{cases} a\cdot i + b + f[i], & 0 \leq i < N \\ a\cdot i + b + f[i-N], & N \leq i < 2N. \end{cases}$$

We define structure regions as: $\Omega = \{i_0,...,i_1\}$, $\widetilde{\Omega} = \{i_2,...,i_3\}$, and define: $\Delta \triangleq \max g_{i\in\Omega} - \min g_{i\in\Omega}$, $w \triangleq i_2 - i_1$-1, $h \triangleq g[i_2-1] - g[i_1+1]$, such that: $a = h/w$. For simplicity, we assume (though this can be relaxed) structures are of equal heights with respect to the linear slope:

$$\begin{aligned} g[i_0] - g[i_0-1] = g[i_1] - g[i_1+1] = \\ g[i_2] - g[i_2-1] = g[i_3] - g[i_3+1] \triangleq H. \end{aligned} \tag{16}$$

We assume H is large enough and l small enough, so that local patterns merge first, as in Lemma 1. We also require the following condition:

$$\frac{h(w+1)}{2} > \frac{(H+\Delta)\cdot l}{l+1}. \tag{17}$$

Theorem 2 (Invariance to Linear Illumination Change)**.** *Let g be as defined above. Then there exists a time $t_{merge-struct}$, such that:*

$$\phi_g(i \in \Omega, t \leq t_{merge-struct}) = \phi_g(i \in \widetilde{\Omega}, t \leq t_{merge-struct}). \tag{18}$$

Proof. Following the same concept as in Theorem 1, patterns in $\Omega, \widetilde{\Omega}$ merge first and simultaneously. To determine the next merging event, we explore the dynamics of structures $\Omega, \widetilde{\Omega}$ vs. those of the linear baseline. Using (16), we will now analyze only the dynamics of the left structure and its neighborhood, as both structures behave the same.
For two neighboring pixels $i, j = i+1$, which are of different regions and of nonzero speed, the merge time according to (11) is $\frac{|g[i]-g[j]|}{2/w_i + 2/w_j}$. Thus the slowest merging of the left structure with the baseline occurs when $\partial_t u_i \neq 0$ only for $i = \{i_0,...,i_1+1\}$, after its internal evolution made u_{i_1} increase by Δ. Then, an upper bound on the merging time of the left structure is: $t_{merge-struct} \leq t_{merge}(g_{\{i_1,i_1+1\}}) = \frac{(H+\Delta)\cdot l}{2(l+1)}$.
We now calculate the merging time of the linear baseline, $t_{merge-line}$. The preceding evolution is a series of merging events t_k of regions h_k, gradually growing with speeds $v_k = 2/k$ by one pixel at a time, such that $\sum_{k=1}^{w} h_k = h/w$. Thus: $t_{merge-line} = \sum_{k=1}^{w} \frac{h_k}{v_k} = \frac{h(1+w)}{4}$.
Given (17), $t_{merge-struct} < t_{merge-line}$. Therefore, TV flow dynamics are identical for $\Omega, \widetilde{\Omega}$ until $t_{merge-struct}$. Relying on (4), we deduce that:

$$\phi_g(i \in \Omega, t \leq t_{merge-struct}) = \phi_g(i \in \widetilde{\Omega}, t \leq t_{merge-struct}).$$

□

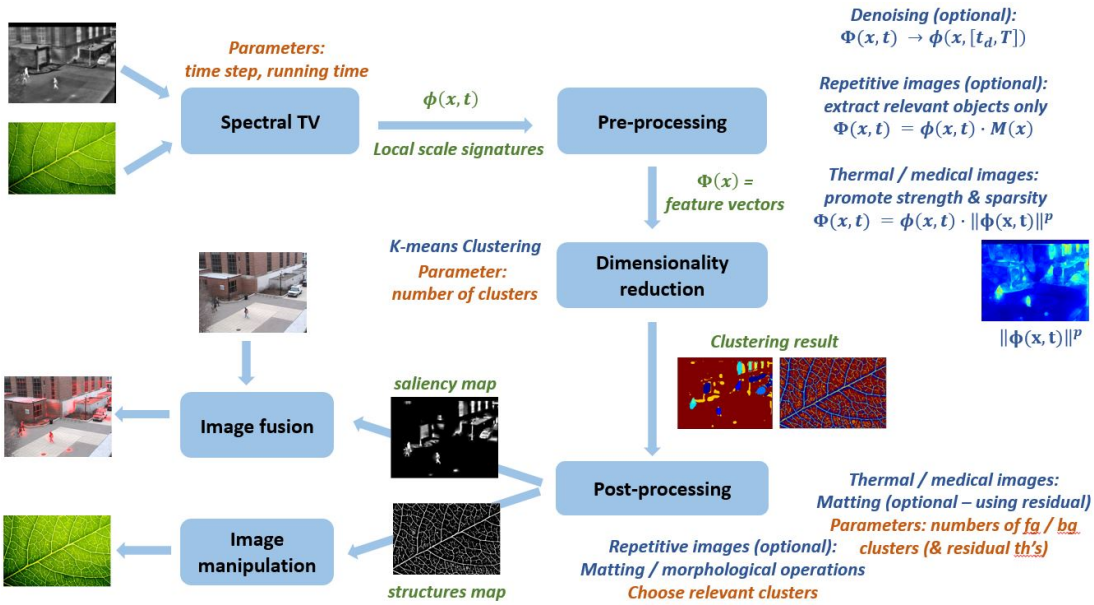


Figure 11: Image manipulation and image fusion using spectral TV local scale signatures: algorithm flowchart.

VI. Algorithm and Applications

The properties of spectral TV signatures can allow grouping together objects with shared features. To facilitate this, we perform dimensionality reduction by clustering of signatures. This allows to partition the image into a set of distinct groups with common spectral TV responses, which can serve for isolating and differentiating salient objects.

A. A Unified Framework

We present a generic unified framework with various applications for images of different modalities (Fig. 11, Algorithm 1). We first decompose the image into its TV elements, using the TV-transform[1] of (4), calculating up to the maximal scale of relevant image structures, T. We use the gray-level version of the input image (in this work, color information is not used). We then perform application-dependent preprocessing on the acquired signatures, denoted $\phi_f(x)$, generating more relevant and enhanced descriptors, denoted $\Phi_f(x)$, to be used as the feature vectors for clustering. A basic dimensionality reduction is performed, using K-means clustering [39]. Last, application-dependent postprocessing of clusters is applied when needed.

Spectral TV Feature Denoising: Since most image noise appears in the first spectral TV bands, an optional denoising step is inherited within the spectral TV scheme, simply by omitting some of the first spectral TV bands.
Denoting the minimal preserved scale (determined by the expected noise variance) as t_d, and the maximal scale calculated for the transform as T, we define denoising as:

$$\Phi_f(x) = \phi_f(x, [t_d, T]). \tag{19}$$

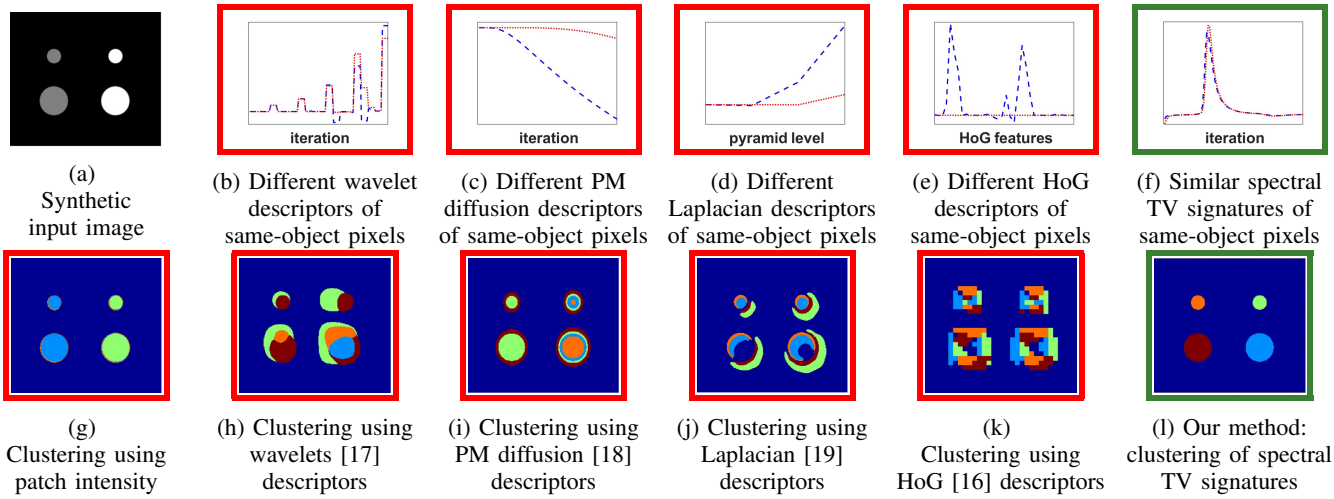
[1]See spectral TV code in http://guygilboa.eew.technion.ac.il/code/code1/

Data: Input image: natural with repetitive structures; thermal; medical; or synthetic.
Result: Image manipulation, image fusion or object differentiation.
Spectral TV transform
Preprocessing:
  case *synthetic image* **do**
    Foreground / background separation, Sec. VI-B
  case *repetitive image* **do**
    Spectral TV Denoising, Eq. 19 (optional)
    Selecting regions of interest, Eq. 20 (optional)
  case *thermal or medical image* **do**
    Spectral TV Denoising, Eq. 19 (optional)
    Signature enhancement, Eq. 21
Dimensionality reduction: K-means clustering
Postprocessing:
  case *repetitive image* **do**
    Matting or morphological operations, Sec. VI-C (optional)
  case *thermal or medical image* **do**
    Matting, Sec. VI-D (optional - using residual, Eq. 6)
Application:
  case *repetitive image* **do**
    Image manipulation
  case *thermal or medical image* **do**
    Image fusion, Appendix B

Algorithm 1: A unified framework for various modalities and applications using spectral TV local scale signatures.

B. Synthetic Images: Object Differentiation

We first show how disk-type objects are well differentiated in this framework. As we want to use simple unsupervised K-means, we first perform a rough foreground/background separation (as the background may contain several clusters). This is done by exploiting the *negativity* of dominant peaks



(a) Synthetic input image
(b) Different wavelet descriptors of same-object pixels
(c) Different PM diffusion descriptors of same-object pixels
(d) Different Laplacian descriptors of same-object pixels
(e) Different HoG descriptors of same-object pixels
(f) Similar spectral TV signatures of same-object pixels
(g) Clustering using patch intensity
(h) Clustering using wavelets [17] descriptors
(i) Clustering using PM diffusion [18] descriptors
(j) Clustering using Laplacian [19] descriptors
(k) Clustering using HoG [16] descriptors
(l) Our method: clustering of spectral TV signatures

Figure 12: Synthetic analysis and comparison of spectral TV signatures to other descriptors. Clustering (bottom row) succeeds only when relying on spectral TV signatures, since only they are the obly descriptors (top row), which guarantee similarity for same-object pixels but distinctness for different-object pixels.

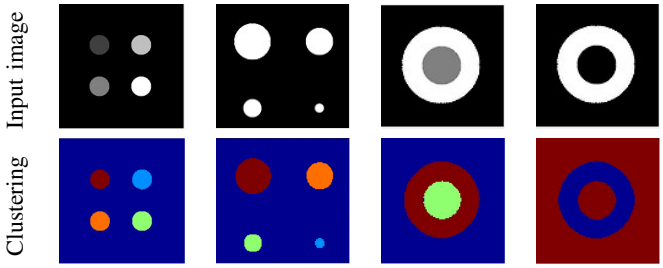


Figure 13: Synthetic images: object differentiation results.

of background signatures (for dark background). Figs. 3, 13 show clustering results of basic structures. Fig. 12 shows how clustering using our method outperforms clustering using other well-known image decompositions or K-means based on different descriptors [17],[18],[19],[16]. These examples illustrate the unique advantages of the proposed approach. The signatures are very similar for same-object pixels and very distinct compared to pixels of other object.

C. Application I: Image Manipulation

We extract a map of salient objects of desired sizes or structures for image manipulation: enhancement, attenuation or coloring of certain structures. We can either explore clusters of manually predefined pixels, or choose interesting structures after clustering. Preprocessing may include denoising (Eq. 19), or selecting regions of interest using a map $M(x)$:

$$\Phi_f(x) = \phi_f(x,t) \cdot M(x). \tag{20}$$

Postprocessing may require image matting [40] or morphological operations of relevant clusters.

D. Application II: Image Fusion

We extract a saliency map from thermal or medical (MRI-T2) images to be fused into a corresponding different-modality image of the same scene (visible or medical, respectively). This relies on the high contrast of salient objects in these modalities (hot objects, or tumors or abnormal organ structures, respectively). To improve clustering, preprocessing includes signature enhancement - "stretching" each signature according to its L_p norm (usually L_1 norm):

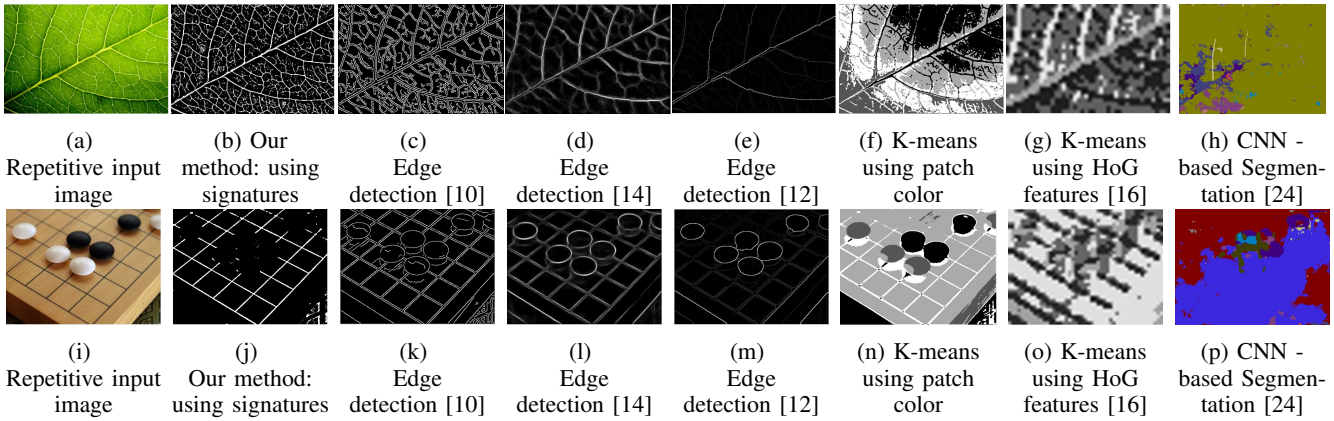
$$\Phi_f(x) = \phi(x,t) \cdot ||\phi(x,t)||^p. \tag{21}$$

Signatures of salient objects "stretch" more, thus promoting their strength and sparsity (see Fig. 8e vs. Fig. 8f). Denoising (19) is also optional. We cluster these enhanced, possibly denoised signatures, $\Phi_f(x)$. Using more clusters allows capturing smaller or narrower objects, e.g. people and lampposts. Finally, our postprocessing includes image matting [40] to generate the saliency map. The user chooses parameters K_1, K_2, where the K_1 strongest clusters (in the sense of centroid intensity) form an initial foreground map; and the K_2 weakest clusters form an initial background map. Matting then classifies all other pixels as foreground/background, and the resulting foreground is the saliency map.

The relevance of highly contrasted but large image regions is often low (depends on the application). In addition, *weak* signatures may nevertheless indicate *relevant* objects, which have not responded yet within a limited running time, or objects near image boundaries. We can handle both issues by incorporating image residual $f_r(x)$ (6) into the postprocessing matting, requiring to select thresholds for $f_r(x)$. This generates an alternative detailed map with large or near-boundary objects, while avoiding long running times. See examples in Figs. 17r, 18d. See Appendix B for novel fusion visualization methods.

VII. Experimental Results

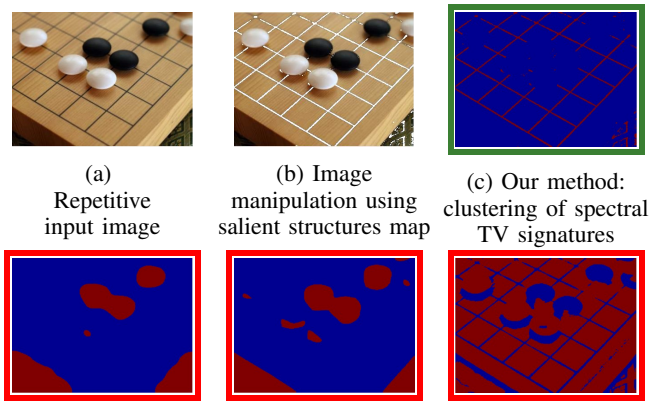
We show experimental results for various image modalities and applications, such as image fusion, image segmentation / edge detection and size differentiation, achieving comparable or superior results compared to other techniques.



(a) Repetitive input image (b) Our method: using signatures (c) Edge detection [10] (d) Edge detection [14] (e) Edge detection [12] (f) K-means using patch color (g) K-means using HoG features [16] (h) CNN - based Segmentation [24]

(i) Repetitive input image (j) Our method: using signatures (k) Edge detection [10] (l) Edge detection [14] (m) Edge detection [12] (n) K-means using patch color (o) K-means using HoG features [16] (p) CNN - based Segmentation [24]

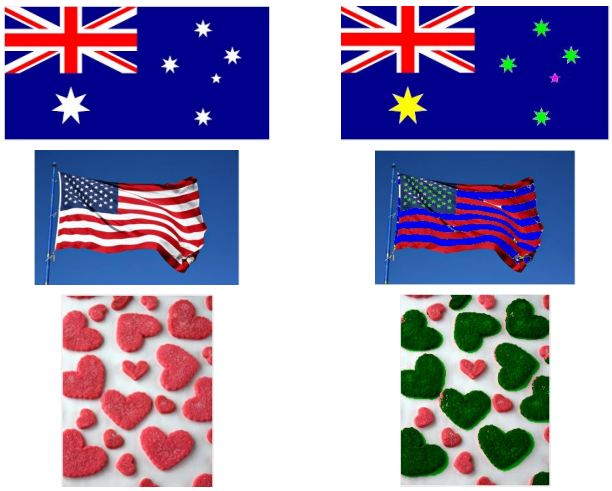
Figure 14: Results and comparisons: salient structure extraction for repetitive images.



(a) Repetitive input image (b) Image manipulation using salient structures map (c) Our method: clustering of spectral TV signatures

(d) Clustering using wavelets [17] descriptors (e) Clustering using PM diffusion [18] descriptors (f) Clustering using Laplacian [19] descriptors

Figure 15: Comparison of K-means clustering into two clusters, using spectral TV signatures vs. other descriptors.



(a) Repetitive image (b) Image manipulation

Figure 16: Image manipulation using size differentiation.

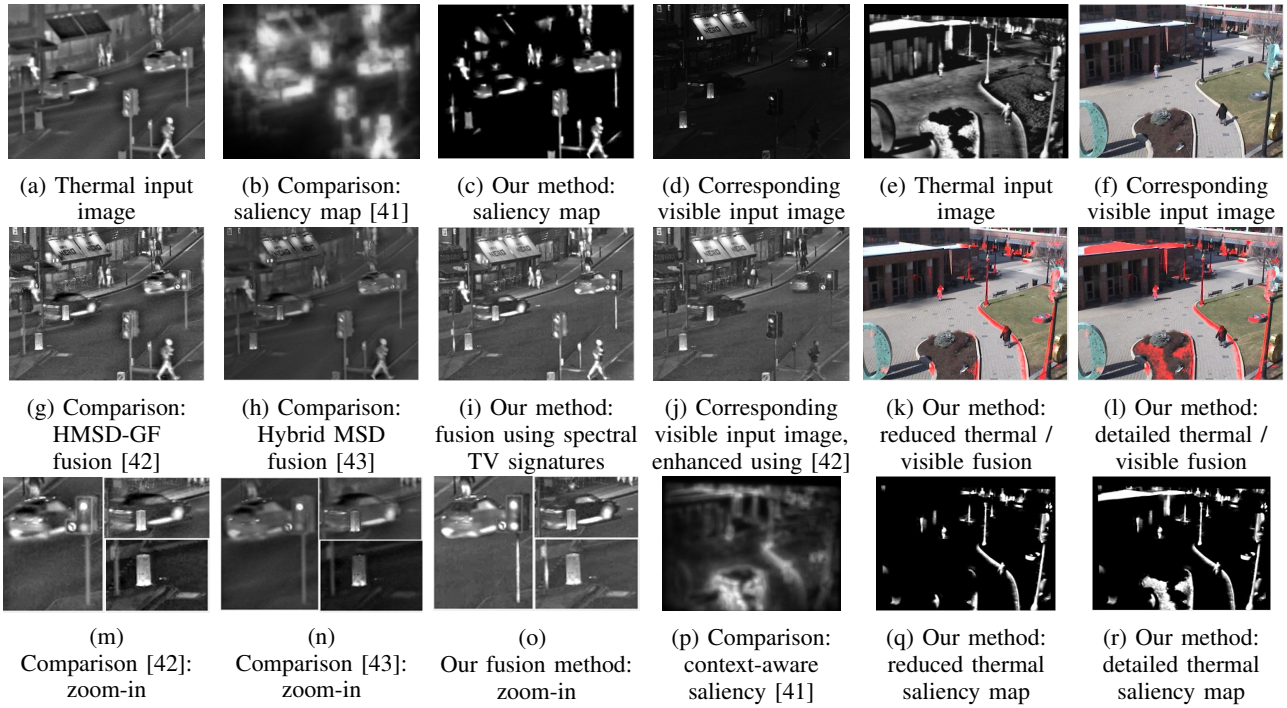
A. Image Manipulation

Fig. 14 demonstrates salient repetitive structure extraction. A comprehensive comparison clearly shows how our method outperforms other state-of-the-art methods. Our method allows to extract fine salient structures in challenging images: leaf veins with varying-illumination background (b), or thin stripes of a game-board, of the same color as other objects (j). This is thanks to the properties discussed earlier of invariance to rotation, translation and linear illumination. Conversely, other methods [10],[14],[12] fail to detect well such fine structures (c,d,e,k,l,m,n). Other methods rely on less stable features, or features sensitive to illumination changes (f) or to rotations [16] (g,o). In the case of learning-based methods, the features are trained for *semantic* segmentation [24] (h,p).
Fig. 15 demonstrates how basic K-means clustering (with $K = 2$) using spectral TV signatures (c) shows highly meaningful clustering compared to the same procedure, based on other well-known image decompositions or descriptors (d,e,f) [17],[18],[19]. Fig. 15b shows an application of signature-based stripe extraction for image manipulation.

The size sensitivity property (Property 1) allows to differentiate objects of similar colors by size. Additional image matting or morphological operations may be used to reconstruct the fine original shape-boundaries from round-shaped clusters. Image manipulation examples based on image segmentation / edge detection or size differentiation are given in Figs. 2, 16.

B. Image Fusion

Figs. 2, 17 demonstrate fusion results of thermal and visible images. Our method can better process fine details, compared to state-of-the-art fusion methods [42], [43]. In addition, our saliency extraction outperforms salient thermal detail extraction, using a well established generic saliency extraction method [41]. Note that as opposed to our work, previous thermal saliency extraction work is usually specifically designed for *human* detection [44]. We also demonstrate our two novel fusion visualization methods (Appendix B), as well as incorporating image residual into the postprocessing to extract a more elaborated saliency map.
Fig. 18 shows (left) a challenging thermal/visible image fusion example. We successfully extract only the salient details from

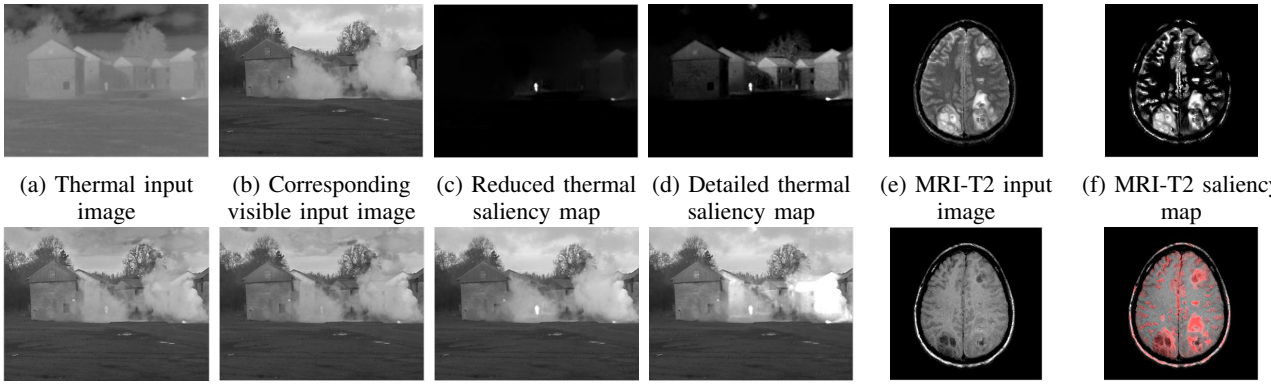


(a) Thermal input image (b) Comparison: saliency map [41] (c) Our method: saliency map (d) Corresponding visible input image (e) Thermal input image (f) Corresponding visible input image

(g) Comparison: HMSD-GF fusion [42] (h) Comparison: Hybrid MSD fusion [43] (i) Our method: fusion using spectral TV signatures (j) Corresponding visible input image, enhanced using [42] (k) Our method: reduced thermal / visible fusion (l) Our method: detailed thermal / visible fusion

(m) Comparison [42]: zoom-in (n) Comparison [43]: zoom-in (o) Our fusion method: zoom-in (p) Comparison: context-aware saliency [41] (q) Our method: reduced thermal saliency map (r) Our method: detailed thermal saliency map

Figure 17: Thermal / visible fusion results and comparisons. A saliency map (c) extracted from a thermal image (a) is fused (i,o) using salient feature injection (Appendix B) into the corresponding visible image, enhanced using [42] (j). Our fusion result outperforms state-of-the-art methods (g,h,m,n). We also show how a reduced saliency map (q) extracted from a thermal image (e) is fused (k) using temperature gradient coloring (Appendix B, using $g(S) = exp(S)$) into the corresponding visible image (f). Alternatively, incorporating image residual into the postprocessing matting generates a detailed saliency map (r) for fusion (l). Our algorithm also outperforms extracting salient thermal details using a state-of-the-art generic saliency method [41] (b,p).



(a) Thermal input image (b) Corresponding visible input image (c) Reduced thermal saliency map (d) Detailed thermal saliency map (e) MRI-T2 input image (f) MRI-T2 saliency map

(g) HMSD-GF fusion [42] (h) Hybrid MSD fusion [43] (i) Reduced thermal / visible fusion (j) Detailed thermal / visible fusion (k) Corresponding MRI-T1 input image (l) MRI-T2 / MRI-T1 fusion

Figure 18: Thermal / visible and medical examples of fusion visualization using temperature gradient coloring (Appendix B using $g(S) = exp(S)$). A reduced saliency map (c), extracted from a challenging, nearly piecewise constant thermal image (a), is fused (i) into the corresponding visible image (b). Alternatively, incorporating image residual into the postprocessing matting generates a detailed saliency map (d) for fusion (j). Our results outperform state-of-the-art methods (g,h). Our medical example shows how a saliency map (f) extracted from an MRI-T2 image (e) is fused (l) into the corresponding MRI-T1 image (k).

a nearly piecewise constant thermal image, outperforming other methods [42], [43]. We also demonstrate (right) medical image fusion, fusing regions of tumor and brain structure extracted from an MRI-T2 image into an MRI-T1 image.

VIII. Conclusion

We design an algorithm to isolate and differentiate objects of different contrasts, sizes and structures. We use the comprehensive scale and space information generated by the spectral TV transform, referred to as spectral TV local scale signatures. Given their high dimensionality and redundancy, we reduce their dimensionality to partition an image into meaningful groups. We prove some useful merits of our local framework: sensitivity to size, local contrast and composition of structures, as well as invariance to rotation, translation, flip and linear illumination change. We also provide conditions for the invariance of texture to structure. This enables to construct a unified generic framework which can serve for many image-processing tasks. We show the applicability of our approach for various cases of multi-scaled images and for different modalities.

Appendix

A. Differentiating Composited Regions by Distinct Signatures

We present an analytic solution for the 1D staircase signal behavior to demonstrate Property 2. It is shown that within each region the signatures of all pixels are identical, and distinct with respect to each other (thus can be easily clustered). Let $f : \{0, ..., N-1\} \to \mathbb{R}$ be as depicted in Fig. 10b, and let $0 < i_0 < i_1 < i_2 < i_3 < N-1$. We denote signal regions as $\Omega_1 \triangleq \{i_1, ..., i_2\}$, $\Omega_2 \triangleq \{i_0, ..., i_1 - 1\} \cup \{i_2 + 1, ..., i_3\}$, $\Omega_3 \triangleq \{0, ..., i_0 - 1\} \cup \{i_3 + 1, ..., N-1\}$, and their sizes as $m_1 \triangleq i_2 - i_1 + 1$, $m_2 \triangleq i_1 - i_0 + i_3 - i_2$, $m_3 \triangleq i_0 + N - 1 - i_3$, respectively. Let $f_{i\in\Omega_1} > f_{i\in\Omega_2} > f_{i\in\Omega_3}$. Without loss of generality, we assume that $m_3 > m_1$.

Proposition 1 (Sensitivity to Composition of Structures)**.** *Let f be as defined above. Then:*

$$\phi_f(i \in \Omega_k, t) = \varphi_k(t), \qquad k = 1, 2, 3, \tag{22}$$

such that $\varphi_k(t) \neq \varphi_l(t)$, $\forall k \neq l$, $k, l \in \{1, 2, 3\}$. Note: region signatures are identical $\forall i \in \Omega_k$. even for the ***disjoint*** *Ω_2.*

Proof. Relying on Section V-A, we analyze the TV flow of f.

1) Phase I: $t \in [0, t_1)$ (Fig. 10b): following (11):

$$u(i,t) = \begin{cases} u(i,0) + \frac{2t}{m_1} \cdot (-1), & i \in \Omega_1 \\ u(i,0) + \frac{2t}{m_2} \cdot 0, & i \in \Omega_2 \\ u(i,0) + \frac{2t}{m_2} \cdot 1, & i \in \Omega_3. \end{cases}$$

$m_3 > m_1 \to u_t(i \in \Omega_3) < u_t(i \in \Omega_1)$. Thus, at t_1 regions Ω_1, Ω_2 merge to form a new region $\Omega_{1,2}$ of size $m_{1,2} = m_1 + m_2$.

2) Phase II: $t \in [t_1, t_2)$ (Fig. 10c): following (11):

$$u(i,t) = \begin{cases} u(i,t_1) + \frac{2t}{m_{1,2}} \cdot (-1), & i \in \Omega_{1,2} \\ u(i,t_1) + \frac{2t}{m_3} \cdot 1, & i \in \Omega_3. \end{cases}$$

Thus, regions $\Omega_{1,2}$, Ω_3 merge at t_2: $u(i, t > t_2) = C$. Differentiating the TV flow of different regions twice in time:

$$u_{tt}(i,t) = \begin{cases} 2(\frac{1}{m_1} - \frac{1}{m_{1,2}})\delta(t-t_1) + \frac{2}{m_{1,2}}\delta(t-t_2), i \in \Omega_1 \\ \frac{2}{m_{1,2}}\Big(-\delta(t-t_1) + \delta(t-t_2)\Big), \quad i \in \Omega_2 \\ -\frac{2}{m_3}\delta(t-t_2), \quad i \in \Omega_3. \end{cases}$$

From (4) we obtain (for some $A, B, C > 0$):

$$\phi_f(i,t) = \begin{cases} A \cdot \delta(t-t_1) + B \cdot \delta(t-t_2)), \ i \in \Omega_1 \\ -B \cdot \delta(t-t_1) + B \cdot \delta(t-t_2)), \ i \in \Omega_2 \\ -C \cdot \delta(t-t_2), \ i \in \Omega_3. \end{cases}$$

□

B. Fusion Visualization Methods

Human observers, unlike computer systems, may prefer viewing salient information when fused into a corresponding different-modality image. We suggest two fusion visualization methods. We denote the saliency map as $S(x)$, a corresponding *registered* image as $V(x)$, and the fused image as $F(x)$. We first suggest injecting salient features directly into the corresponding image (Fig. 17i):

$$F(x) = max(V(x), S(x)).$$

This allows introducing information which only appears in $S(x)$ on top of $V(x)$. However, the typically low quality thermal information might overlap the more detailed visible information; and salient white objects will not be visualized as salient. To overcome this, we suggest the temperature gradient coloring method (Figs. 17k, 17l, 18i, 18j, 18l). $F(x)$ is a gray-level or RGB replicate of $V(x)$ ($\{R_V(x), G_V(x), B_V(x)\}$)), enhanced in locations corresponding to $S(x)$:

$$F(x) = V(x) \cdot g(S(x)),$$
$$or \qquad F(x) = \{R_V(x) \cdot g(S(x)), G_V(x), B_V(x)\},$$

then normalized to avoid clipping. $g(S)$ must be:

1) Positive: $\forall x, \ \ g(S(x)) > 0$.
2) Monotonically increasing: $\forall x, \ \ \frac{\partial g(S(x))}{\partial S(x)} > 0$.
3) Null for non-salient objects: $g(S(x) = 0) = 1$.

Some useful examples are $g(S) = 1 + S$ and $g(S) = exp(S)$. Advantages of this method are: avoiding overlapping detailed information with low-quality one; visualizing the gradient of temperatures (e.g., the hotter the object - the redder it appears); and handling salient white objects. Conversely, details which appears only in the saliency map appear weaker in the fusion.

References

[1] M. Burger, G. Gilboa, M. Moeller, L. Eckardt, and D. Cremers, "Spectral decompositions using one-homogeneous functionals," *SIAM Journal on Imaging Sciences*, vol. 9, no. 3, pp. 1374–1408, 2016.

[2] M. Benning, M. Möller, R. Z. Nossek, M. Burger, D. Cremers, G. Gilboa, and C.-B. Schönlieb, "Nonlinear spectral image fusion," in *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 2017, pp. 41–53.

[3] D. Horesh and G. Gilboa, "Separation surfaces in the spectral tv domain for texture decomposition," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4260–4270, 2016.

[4] H. Lu *et al.*, "Multisensor image fusion and enhancement in spectral total variation domain," *IEEE Transactions on Multimedia*, 2017.

[5] L. Zeune, S. A. van Gils, L. W. Terstappen, and C. Brune, "Combining contrast invariant l1 data fidelities with nonlinear spectral image decomposition," in *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 2017, pp. 80–93.

[6] T. Brox and J. Weickert, "A tv flow based local scale measure for texture discrimination," in *European Conference on Computer Vision*. Springer, 2004, pp. 578–590.

[7] D. M. Strong, J.-F. Aujol, and T. F. Chan, "Scale recognition, regularization parameter selection, and meyer's g norm in total variation regularization," *Multiscale Modeling & Simulation*, vol. 5, no. 1, pp. 273–303, 2006.

[8] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[10] J. Canny, "A computational approach to edge detection," *IEEE Trans. on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[11] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[12] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[13] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 5, pp. 530–549, 2004.

[14] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1558–1570, 2015.

[15] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005*, vol. 1. IEEE, 2005, pp. 886–893.

[17] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Transactions on image processing*, vol. 1, no. 2, pp. 205–220, 1992.

[18] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[19] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[20] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern recognition*, vol. 24, no. 12, pp. 1167–1186, 1991.

[21] C. Sagiv, N. A. Sochen, and Y. Y. Zeevi, "Integrated active contours for texture segmentation," *IEEE transactions on image processing*, vol. 15, no. 6, pp. 1633–1646, 2006.

[22] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[24] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Dec 2017.

[25] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[26] S. Mallat, "Understanding deep convolutional networks," *Phil. Trans. R. Soc. A*, vol. 374, no. 2065, p. 20150203, 2016.

[27] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[28] A. Chambolle and P.-L. Lions, "Image recovery via total variation minimization and related problems," *Numerische Mathematik*, vol. 76, no. 2, pp. 167–188, 1997.

[29] S. Osher, A. Solé, and L. Vese, "Image decomposition and restoration using total variation minimization and the h," *Multiscale Modeling & Simulation*, vol. 1, no. 3, pp. 349–370, 2003.

[30] G. Gilboa, N. Sochen, and Y. Y. Zeevi, "Estimation of optimal pde-based denoising in the snr sense," *IEEE Transactions on Image Processing*, vol. 15, no. 8, pp. 2269–2280, 2006.

[31] J.-F. Aujol, G. Gilboa, T. Chan, and S. Osher, "Structure-texture image decompositionmodeling, algorithms, and parameter selection," *International Journal of Computer Vision*, vol. 67, no. 1, pp. 111–136, 2006.

[32] L. A. Vese and S. J. Osher, "Modeling textures with total variation minimization and oscillating patterns in image processing," *Journal of scientific computing*, vol. 19, no. 1, pp. 553–572, 2003.

[33] G. Gilboa, N. Sochen, and Y. Y. Zeevi, "Variational denoising of partly textured images by spatially varying constraints," *IEEE Transactions on Image Processing*, vol. 15, no. 8, pp. 2281–2289, 2006.

[34] M. Kumar and S. Dass, "A total variation-based algorithm for pixel-level image fusion," *IEEE Transactions on Image Processing*, vol. 18, no. 9, pp. 2137–2143, 2009.

[35] Y. Ma, J. Chen, C. Chen, F. Fan, and J. Ma, "Infrared and visible image fusion using total variation model," *Neurocomputing*, vol. 202, pp. 12–19, 2016.

[36] F. Andreu, C. Ballester, V. Caselles, J. Mazón *et al.*, "Minimizing total variation flow," *Differential and integral equations*, vol. 14, no. 3, pp. 321–360, 2001.

[37] G. Gilboa, "A total variation spectral framework for scale and texture analysis," *SIAM journal on Imaging Sciences*, vol. 7, no. 4, pp. 1937–1961, 2014.

[38] G. Steidl, J. Weickert, T. Brox, P. Mrázek, and M. Welk, "On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and sides," *SIAM Journal on Numerical Analysis*, vol. 42, no. 2, pp. 686–713, 2004.

[39] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[40] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 228–242, 2008.

[41] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[42] Z. Zhou, M. Dong, X. Xie, and Z. Gao, "Fusion of infrared and visible images for night-vision context enhancement," *Applied optics*, vol. 55, no. 23, pp. 6480–6490, 2016.

[43] Z. Zhou, B. Wang, S. Li, and M. Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters," *Information Fusion*, vol. 30, pp. 15–26, 2016.

[44] J. W. Davis and V. Sharma, "Robust background-subtraction for person detection in thermal imagery." in *CVPR Workshops*, 2004, p. 128.

Ester Hait is pursuing her Ph.D. in Electrical Engineering in the Technion - Israel Institute of Technology, where she received her B.Sc. (Cum Laude) and M.Sc. in Electrical Engineering in 2014 and 2016, respectively. Her research interests include image processing and computer vision.

Guy Gilboa received his Ph.D. from the Electrical Engineering Department, Technion - Israel Institute of Technology in 2004. He was a postdoctoral fellow with UCLA and had various development and research roles with Microsoft and Philips Healthcare. Since 2013 he is a faculty member with the Electrical Engineering Department, Technion - Israel Institute of Technology. He has authored some highly cited papers on topics such as image sharpening and denoising, nonlocal operators theory, and texture analysis. He received several prizes, including the Eshkol Prize by the Israeli Ministry of Science, the Vatat Scholarship, and the Gutwirth Prize. He serves at the editorial boards of the journals IEEE SPL, JMIV and CVIU.