Journal of Mathematical Imaging and Vision (2018) 60:1008–1024
https://doi.org/10.1007/s10851-018-0791-3

Space–Time Signal Analysis and the 3D Shearlet Transform

Damiano Malafronte[1] · Ernesto De Vito[2] · Francesca Odone[1]

Received: 30 May 2017 / Accepted: 11 January 2018 / Published online: 21 February 2018

Abstract
In this work, we address the problem of analyzing video sequences by representing meaningful local space–time neighborhoods. We propose a mathematical model to describe relevant points as local singularities of a 3D signal, and we show that these local patterns can be nicely highlighted by the 3D shearlet transform, which is at the root of our work. Based on this mathematical framework, we derive an algorithm to represent space–time points which is very effective in analyzing video sequences. In particular, we show how points of the same nature have a very similar representation, allowing us to compute different space–time primitives for a video sequence in an unsupervised way.

Keywords Shearlet transform · 2D $+$ T signal analysis · Space–time local primitives

1 Introduction

Spatial local keypoints and appropriate local descriptors have been extensively considered in image processing and computer vision, and they have been successfully studied on a variety of multi-scale models [28,29]. They have then been applied to image matching or to the higher level image classification problem, often in conjunction with appropriately designed global descriptors.

In the past decade, the video processing scenario has been characterized by a growing interest toward the so-called space–time interest points which incorporate appearance as well as dynamic local information. From the pioneering work of Laptev [27], who proposed a reformulation of Harris corners [15] for the space–time, soon followed by alternative and possibly richer approaches [6,16,31,35,36], we have appreciated the power of these key points as low level building blocks for motion analysis and action recognition. An exhaustive overview of related state of the art can be found in [4], see also [37].

Space–time interest points are usually associated with the concept of points characterized by some special behavior both in space and in time (e.g., nonsmooth in both directions). Thus, the classical computational framework starts with a key point detection stage, often in conjunction with an appropriate key point descriptor [17,26,32,33]. Generally, keypoints are detected by looking for singularities both in space and in time [27]. In this paper, we argue that in the space–time domain there is a richer set of information to be exploited: different interesting local primitives can be observed and associated with an appropriate meaning in space and time. These primitives also include interesting spatial structures (spatial corners or edges) moving smoothly or smooth surfaces undergoing significant velocity changes.

The mathematical framework we consider is the one of the shearlets [25]. Among the multiresolution image representations, shearlets emerge by their ability to efficiently capture anisotropic features [18], to provide an optimal sparse representation [11,21], to detect singularities [14,22] and to be stable against noise and blurring [2,9]. For further details, implementations and references, see [20]. The effectiveness of shearlets is supported by a well-established mathematical theory [3], and it is tested in many applications in image processing by providing efficient algorithms [7,8,20]. Shearlets have seldom been applied to spatio-temporal data,

E. De Vito is member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). F. Odone is member of the Gruppo Nazionale per il Calcolo Scientifico (GNCS) of the Istituto Nazionale di Alta Matematica (INdAM).

✉ Damiano Malafronte
damiano.malafronte@dibris.unige.it

Ernesto De Vito
devito@dima.unige.it

Francesca Odone
francesca.odone@unige.it

[1] DIBRIS, Università degli Studi di Genova, Genova, Italy

[2] DIMA, Università degli Studi di Genova, Genova, Italy

with the exception of shearlet-based video denoising and inpainting [24]—see also [30], comparing shearlet-based performances on video enhancement and denoising tasks with previously existing techniques.

In this work, we exploit different properties of shearlets. In particular, we focus on the ability of shearlet coefficients to detect the wavefront set of a signal both in 2D [18] and in the 3D settings [12,23], by directly encoding meaningful directional informations, as, for example, the normal direction at each point of a surface singularity. From the computational viewpoint, we adopt 3D shearlets implemented in ShearLab (see http://www.shearlab.org/).

The contribution of the paper is twofold. On the theoretical side, we propose a toy mathematical model to describe some of the significant properties of the complex behavior of a real video sequence. We consider a rigid compact 2D region that, by moving in time, generates a $2D + T$ volume V. The spatial–temporal points are now associated with the wavefront set of the 3D “cartoon-like” signal [23]

$$f(x, y, t) = \begin{cases} 1 & (x, y, t) \in V \\ 0 & (x, y, t) \notin V. \end{cases}$$

We show that the corresponding shearlet coefficients provide a clear signature of different spatial-temporal primitives. Clearly, our model does not capture the full complexity of a real video sequence, for instance, it does not deal with occlusions, but it provides an important insight of what happens in the real world by highlighting the kind of spatio-temporal primitive each space–time point belongs to.

Motived by our theoretical framework, we propose an algorithm to represent key points highlighting their appearance and dynamic properties. First, we consider the 3D shearlet transform of a video sequence. Then, we derive a shearlet-based rotation-invariant representation of each point with respect to its space–time neighborhood at a fixed scale. This representation describes the behavior of the signal in the neighborhood and helps us discriminating among different type of points. We discuss how this representation does not vary too much on sets of known spatial and spatio-temporal key points such as edges, corners and space–time interest points [27]. We also show how to identify the main primitives in a video signal, by adopting an unsupervised approach and clustering points to obtain the most significant space–time primitives within the signal.

The real video sequences we use to discuss our findings are taken from the Chalearn (*che vuoi* [10]) and the KTH (*boxing*, *handwaving* and *walking* [32]) datasets, while synthetic data have been generated in-house.

This paper is organized as follows. Section 2 reviews shearlets on $2D + T$ signals. Section 3 introduces the concept of spatio-temporal primitives. In Sect. 4, we describe our approach to represent points in their space–time neighborhood and discuss the expressiveness of the representation on both synthetic and real data. Section 5 discusses the results we obtain when clustering points with respect to the proposed representation. Section 6 is left to a conclusive discussion.

2 The 3D Shearlet Frame

In this section, we briefly review the construction of the shearlet frame for $2D + T$ signals. We follow the presentation in [19], which is a standard reference for the proofs and other informations.

We first set the notation. We denote by L^2 the Hilbert space of functions $f : \mathbb{R}^3 \to \mathbb{C}$ such that

$$\int_{\mathbb{R}^3} |f(x, y, t)|^2 \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}t < +\infty,$$

where $\mathrm{d}x\mathrm{d}y\mathrm{d}t$ is the Lebesgue measure of $\mathbb{R}^3$, by $\|f\|$ the corresponding norm and by $\langle f, f' \rangle$ the scalar product between two functions $f, f' \in L^2$. Given an element $f \in L^2$, we denote by $\widehat{f}$ its Fourier transform, i.e.,

$$\widehat{f}(\xi_1, \xi_2, \xi_3) = \int_{\mathbb{R}^3} f(x, y, t) \mathrm{e}^{-2\pi i(\xi_1 x + \xi_2 y + \xi_3 t)} \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}t,$$

provided that f is integrable, too.

We recall that a frame for L^2 is a family $\{\psi_i\}_{i \in I}$ of functions such that each ψ_i is in L^2 and

$$A \|f\|^2 \leq \sum_{i \in I} |\langle f, \psi_i \rangle|^2 \leq B \|f\|^2 \quad \forall f \in L^2,$$

where A, B are positive constants, called frame bounds. The shearlet frame $\mathcal{F}_{\mathrm{SH}}$ is defined in terms of four different subfamilies labeled by the index $\ell = 0, \ldots, 3$ as it follows.

The first family

$$\mathcal{F}_{\mathrm{SH},0} = \left\{ \varphi_m \mid m \in \mathbb{Z}^3 \right\},$$

associated with the index $\ell = 0$ takes care of the low frequencies cube

$$\mathcal{P}_0 = \left\{ (\xi_1, \xi_2, \xi_3) \in \widehat{\mathbb{R}}^3 \mid |\xi_1| \leq 1, |\xi_2| \leq 1, |\xi_3| \leq 1 \right\}$$

and it is given by

$$\varphi_m(x, y, t) = \varphi(x - cm_1, y - cm_2, t - cm_3),$$

where $m = (m_1, m_2, m_3) \in \mathbb{Z}^3$ labels the translations, $c > 0$ is a step size, and

$$\varphi(x, y, t) = \phi_1(x)\phi_1(y)\phi_1(t),$$

where ϕ_1 is a 1D-scaling function.

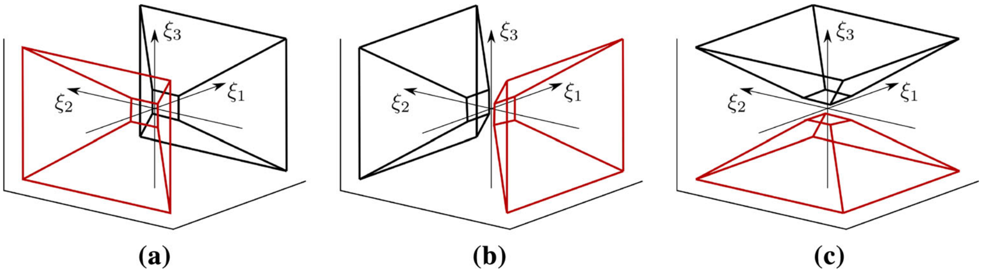


Fig. 1 Three pyramids $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$, with displayed in black the area belonging to the positive part of the corresponding symmetry axis and in red the one related to its negative part (Color figure online)

The other three families are associated with the high frequency domain. Each of them corresponds to the pyramid whose symmetry axis is one of the cartesian axes ξ_1, ξ_2, ξ_3 in the Fourier domain, see Fig. 1. For example, for $\ell = 1$ the pyramid is

$$\mathcal{P}_1 = \left\{ (\xi_1, \xi_2, \xi_3) \in \widehat{\mathbb{R}}^3 \mid |\xi_1| > 1, \left|\frac{\xi_2}{\xi_1}\right| \leq 1, \left|\frac{\xi_3}{\xi_1}\right| \leq 1 \right\},$$

and similarly for the other two pyramids.

Fixed $\ell = 1, 2, 3$, each

$$\mathcal{F}_{\mathrm{SH},\ell} = \left\{ \psi_{\ell,j,k,m} \mid j \in \mathbb{N}, k \in \mathbf{K}_j, m \in \mathbb{Z}^3 \right\},$$

where

$$\mathbf{K}_j = \left\{ k = (k_1, k_2) \in \mathbb{Z}^2, \max\{ |k_1|, |k_2| \} \leq \lceil 2^{j/2} \rceil \right\}, \quad (1)$$

is defined in terms of parabolic dilations

$$A_{1,j} = \begin{pmatrix} 2^j & 0 & 0 \\ 0 & 2^{j/2} & 0 \\ 0 & 0 & 2^{j/2} \end{pmatrix}, \quad A_{2,j} = \begin{pmatrix} 2^{j/2} & 0 & 0 \\ 0 & 2^j & 0 \\ 0 & 0 & 2^{j/2} \end{pmatrix},$$

$$A_{3,j} = \begin{pmatrix} 2^{j/2} & 0 & 0 \\ 0 & 2^{j/2} & 0 \\ 0 & 0 & 2^j \end{pmatrix},$$

where the index j refers to the dyadic scale (note that $j = 0$ corresponds to the coarsest scale), and shearings

$$S_{1,k} = \begin{pmatrix} 1 & k_1 & k_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad S_{2,k} = \begin{pmatrix} 1 & 0 & 0 \\ k_1 & 1 & k_2 \\ 0 & 0 & 1 \end{pmatrix},$$

$$S_{3,k} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ k_1 & k_2 & 1 \end{pmatrix},$$

where the index $k = (k_1, k_2) \in \mathbf{K}_j$ controls the shearing and runs over the indexes $\mathbf{K}_j$ defined in (1). Explicitly, the functions $\psi_{\ell,j,k,m}$ are given by

$$\psi_{\ell,j,k,m}(x, y, t) = 2^j \psi_\ell \left(S_{\ell,k} A_{\ell,j} \begin{pmatrix} x - c_1 m_1 \\ y - c_2 m_2 \\ t - c_3 m_3 \end{pmatrix} \right), \quad (2)$$

where for $\ell = 1$, $c_1 = c$ and $c_2 = c_3 = \widehat{c}$, where $\widehat{c}$ is another step size (for $\ell = 2, 3$ the values of c_1, c_2, c_3 are interchanged accordingly) and the parameter $m = (m_1, m_2, m_3) \in \mathbb{Z}^3$ labels the translations, as for the family $\mathcal{F}_{\mathrm{SH},0}$. Following [24], the generating function ψ_1 is of the form

$$\widehat{\psi}_1(\xi_1, \xi_2, \xi_3) = \widehat{\psi}_1(\xi_1) \left(P\left(\frac{\xi_1}{2}, \xi_2\right) \widehat{\phi}_1(\xi_2) \right) \times \left(P\left(\frac{\xi_1}{2}, \xi_3\right) \widehat{\phi}_1(\xi_3) \right), \quad (3)$$

where P is suitable polynomial 2D Fan filter [5], ψ_1 is the 1D wavelet function associated with the scaling function ϕ_1 defining the family $\{\varphi_m\}$. Similar equations hold for $\ell = 2, 3$ by interchanging the role of ξ_1, ξ_2 and ξ_3. We observe that to obtain a frame it is necessary to assume some technical condition on the smoothness of ϕ_1 and on the vanishing momenta of ψ_1, see [20].

The shearlet transform of a signal $f \in L^2$ is given by

$$\mathrm{SH}[f](\ell, j, k, m) = \begin{cases} \langle f, \varphi_m \rangle & \text{if } \ell = 0 \\ \langle f, \psi_{\ell,j,k,m} \rangle & \text{if } \ell = 1, 2, 3, \end{cases}$$

where $j \in \mathbb{N}$, $k \in \mathbf{K}_j$, $m \in \mathbb{Z}^3$. We stress the fact that, as shown in (1), the number of shearing parameters $\mathbf{K}_j$ depends on j. In the experiments, we use the digital implementation described in [24], which is based on the well-known relation between the pair (ϕ_1, ψ_1) and the quadrature mirror filter pair (h, g), i.e.,

$$\phi_1(x) = \sqrt{2} \sum_{n \in \mathbb{Z}} h(n) \phi_1(2x - n) \quad (4)$$

$$\psi_1(x) = \sqrt{2} \sum_{n \in \mathbb{Z}} g(n) \phi_1(2x - n). \quad (5)$$

where h is a 1D low-pass filter and g is the corresponding high-pass filter.

Furthermore, a maximum number J of scales is considered and it assumed that the signal f at the finest scale is given by

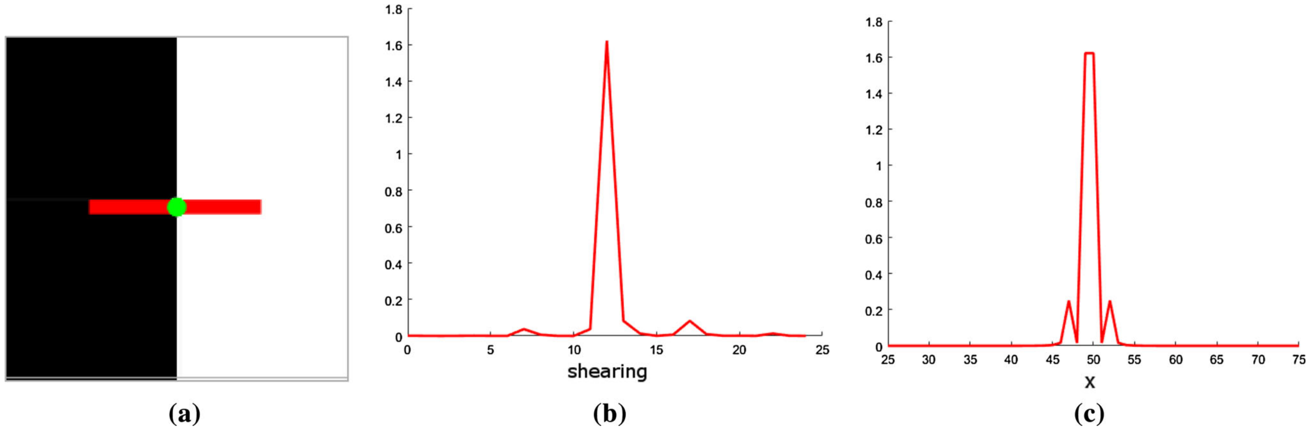


Fig. 2 Coefficients analysis on a 3D surface (see text). **a** A section of a surface parallel to the yt plane. **b** A plot representing the coefficients varying at different shearings, on the x axis there are the indexes corresponding to all the shearings in $\mathbf{K}_j$ for the pyramid $\mathcal{P}_1$ where the central peak corresponds to the shearing vector $k = (0, 0)$. **c** The coefficients decay for neighboring points along the surface normal (red line) (Color figure online)

$$f(x, y, t) = \sum_{m \in Z^3} f_{J,m}\, 2^{3J/2} \phi_1(2^J x - cm_1) \\ \times \phi_1(2^J y - cm_2)\phi_1(2^J t - cm_3).$$

so that $f_{J,m} \simeq f(cm_1 2^{-J}, cm_2 2^{-J}, cm_3 2^{-J})$ since ϕ_1 is well localized around the origin. The digital shearlet transform depends on the number of scales $J + 1$, the directional Fan filter P in (3) and the low-pass filter h associated with the scaling function ϕ_1 by (4).

Our algorithm is based on the following nice property of the shearlet coefficients. As shown in [12,13,23] if the signal f is locally regular in a neighborhood of m, then $\mathrm{SH}[f](\ell, j, k, m)$ has a fast decay when j goes to infinity for any $\ell \neq 0$ and $k \in \mathbf{K}_j$. Suppose now that f has a surface singularity at cm with normal vector $(1, n_1, n_2) \in \mathcal{P}_1$ and set $k^* = (\lceil 2^{j/2} n_1 \rceil, \lceil 2^{j/2} n_2 \rceil)$. If $\ell = 2, 3$, then $\mathrm{SH}[f](\ell, j, k, m)$ has a fast decay for any $k \in K_j$, whereas if $\ell = 1$ we have the same good behavior only if $k \neq k^*$, whereas if $k = k^*$ the shearlet coefficients have a slow decay (a similar result holds if the normal direction of the surface singularity belongs to the other two pyramids). This behavior of the shearlet coefficients allows to associate to any shearing vector $k = (k_1, k_2)$ a direction (without orientation) parametrized by two angles, *latitude* and *longitude*, α and β. Thus, the direction associated with k is given by

$$(\cos\alpha\cos\beta, \cos\alpha\sin\beta, \sin\alpha) \quad \alpha, \beta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]. \qquad (6)$$

The correspondence explicitly depends on ℓ and, for the first pyramid, it is given by

$$\tan\alpha = \frac{2^{-j/2} k_2}{\sqrt{1 + 2^{-j} k_1^2}} \quad \tan\beta = 2^{-j/2} k_1 \quad \alpha, \beta \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right].$$

The above formula shows that the ability to resolve different directions strongly depends on the number of available shearings in $\mathbf{K}_j$. In particular, at coarsest scales we detect the normal direction of singularity surfaces at a low resolution.

Through a simple example, we illustrate the above behavior. We consider a black cube, and we fix a point of a side of the cube parallel to the yt plane. We compute the shearlet coefficients moving along the normal direction outside the cube. The behavior is shown in Fig. 2 where in the first column we show a xt-section of the cube at a given t. In the second column, we plot the value of the shearlet coefficients at the point on the surface in the first pyramid $\mathcal{P}_1$. We show the coefficients associated with the grid of directions represented by the shearings in $\mathbf{K}_j$, unrolling them along the x axis. In this example, $k_1, k_2 \in \{-2, 1, 0, 1, 2\}$ and the value 12 in Fig. 2b corresponds to $k_1 = k_2 = 0$, as expected. The coefficients of the other pyramids contain negligible values $\sim 10^{-16}$. In the third column, we fix the shearing corresponding to the peak and see how the coefficients evolve by moving along the normal direction corresponding to the red line in Fig. 2a. The coefficients decay as we move away from the discontinuity, giving us an empirical evidence of the appropriateness of 3D shearlets in localizing interest points.

Figure 3 shows a similar analysis on a 3D edge produced by two surfaces, one parallel to plane xt and the other parallel to plane yt. In this case, we identify two significant peaks in two different pyramids (the main peaks in Fig. 2b, e.

Within every pyramid ($\mathcal{P}_1$ for Fig. 3a–c and $\mathcal{P}_3$ for Fig. 3d–f), we see a behavior similar to the case of the 3D surface (Fig. 2b). However, the secondary peaks have higher values, for the spatio-temporal neighborhood around the point has a richer behavior. These peaks are also due to the fact that we visualize two-dimensional information (the shearlet

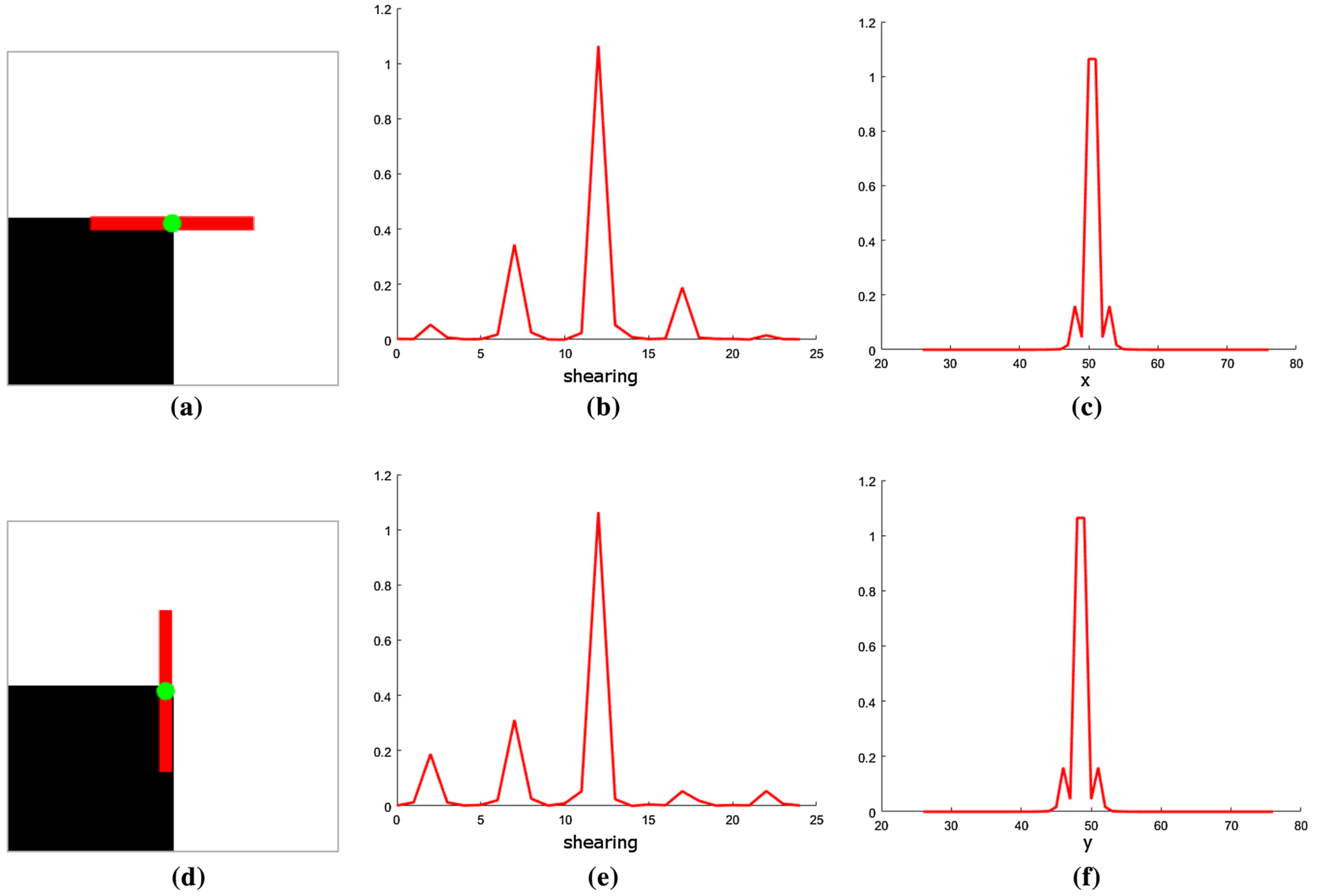


Fig. 3 Coefficients analysis on a 3D edge (see text). **a**, **d** A section of the edge parallel to xy plane, where we highlight the two normal vectors. **b**, **e** The coefficients varying at different shearings in the two meaningful pyramids $\mathcal{P}_1$ and $\mathcal{P}_3$ where both central peaks correspond to the shearing vector $k = (0, 0)$. **c**, **f** The decay of coefficients for neighboring points along the corresponding normal directions

coefficients associated with a 2D grid of directions) as a 1D function; thus, they appear to be distant on the 1D unrolled function.

The plots we show in this section have been obtained thanks to the a priori information we have on the normal direction which is in general not available in real data. This issue will be addressed in the following sections, where we identify a representation procedure applicable in the general case.

3 Spatio-Temporal Primitives

Clearly, a video is a temporal sequence of 2D spatial images and it can be regarded as a $2D + T$ signal that fits the above theoretical framework.

In this context, 2D spatial discontinuities in an image, such as edges and corners, generate different space–time behaviors as the image evolve in time. Moreover, the temporal evolution of a given point in the image is continuous, but may undergo a loss of regularity in correspondence of velocity changes. Therefore, if we analyze the behavior of the signal in space–time, we may observe different types of primitives (see also Fig. 4):

- *Spatio-temporal surfaces*, caused by 2D edges with a smooth velocity spanning surfaces in space–time.
- *Spatio-temporal edges* either caused by 2D corners moving smoothly or by 2D edges undergoing a velocity change. These two primitives could be discriminated by detecting the orientation of the 3D edge, see Fig. 4b, c.
- *Spatio-temporal corners or vertices* caused by 2D corners undergoing a velocity change.

These spatio-temporal primitives are easily associated with classical 3D features: surfaces, edges and vertices and can be analyzed by adapting 3D signal representation models. It should be observed, though, that $2D + T$ features have a very specific nature that characterizes them beyond their three-dimensional structure. For instance, we could further

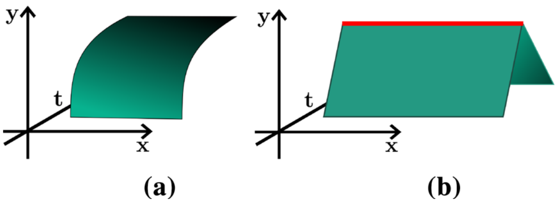


Fig. 4 Spatio-temporal primitives which can take place in the space–time domain, by considering how the image in the background of each one of these moves over time: **a** A 2D edge moving smoothly spawns a spatio-temporal surface, **b** A 2D edge undergoing a velocity change, thus producing a 3D edge, **c** A 2D corner moving smoothly also producing a 3D edge, **d** A 2D corner undergoing a velocity change providing a 3D vertex

cluster these primitives in still and moving entities (corresponding to different orientations in the 2D + T space). Also, the third component (time) has a different intrinsic scale, and very precise constraints since spatial features do not disappear all of a sudden and time can only proceed forward. In the reminder of the paper, we refer to 2D edges when considering image discontinuities and 3D or spatio-temporal edges when discussing the behavior in space–time. As for corners, we will refer to 2D corners in space and to vertices or 3D corners in space–time.

We now observe that thanks to the sensitivity to singularity and orientation of shearlets we may identify different spatial-temporal primitives. To better understand the relationship between coefficients and primitives, we start by considering a toy model for a space region evolving over time. We assume that the region of interest is a rigid planar body $\mathcal{C}$ moving in the time interval $[0, T]$. We further assume that the boundary of $\mathcal{C}$ can be parametrized at the initial time $t = 0$ by the simple closed curve

$$\gamma(s) = x(s)\mathbf{i} + y(s)\mathbf{j} \quad s \in [0, L],$$

where L is the length of the boundary, s is the arc length oriented and the curve is oriented so that the interior of the body is on the left side, see Fig. 5. We denote by $\mathbf{i}$ and $\mathbf{j}$ to be the canonical unit vectors of the x-axis and y-axis, respectively. Since the body is rigid, the time evolution of each point $\gamma(s)$ is given by

$$\gamma(s,t) = r(t) + R(t)(\gamma(s) - r(0)) = x(s,t)\,\mathbf{i} + y(s,t)\,\mathbf{j},$$

where $r(t)$ is the time evolution of the center of mass of the body and $R(t)$ is the time-dependent rotation around the center of mass. The evolution of the body in time describes a 3D volume whose boundary is the surface parametrized by

$$\sigma(s,t) = x(s,t)\,\mathbf{i} + y(s,t)\,\mathbf{j} + t\,\mathbf{k} \quad s \in [0, L], t \in [0, T],$$

where $\mathbf{k}$ is the canonical unit vector of the t-axis.

We now compute the normal vector to the surface at spatial–temporal point $\sigma(s,t)$

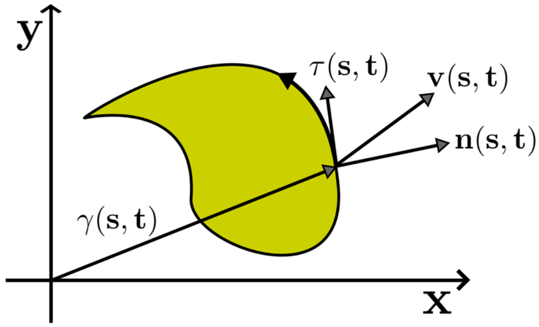

y
$\tau(\mathbf{s},\mathbf{t})$
$\mathbf{v}(\mathbf{s},\mathbf{t})$
$\mathbf{n}(\mathbf{s},\mathbf{t})$
$\gamma(\mathbf{s},\mathbf{t})$
x

Fig. 5 A body at time t with the main relevant geometrical and dynamical quantities

$$\begin{aligned} N(s,t) &= \frac{\partial\sigma}{\partial s}(s,t) \times \frac{\partial\sigma}{\partial t}(s,t) = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial x}{\partial s}(s,t) & \frac{\partial y}{\partial s}(s,t) & 0 \\ \frac{\partial x}{\partial t}(s,t) & \frac{\partial y}{\partial t}(s,t) & 1 \end{bmatrix} \\ &= n(s,t) + \tau(s,t) \times v(s,t) \end{aligned}$$

where

$$\begin{aligned} \tau(s,t) &= \frac{\partial x}{\partial s}(s,t)\mathbf{i} + \frac{\partial y}{\partial s}(s,t)\mathbf{j} \\ n(s,t) &= \frac{\partial y}{\partial s}(s,t)\mathbf{i} - \frac{\partial x}{\partial s}(s,t)\mathbf{j} \\ v(s,t) &= \frac{\partial x}{\partial t}(s,t)\mathbf{i} + \frac{\partial y}{\partial t}(s,t)\mathbf{j} \end{aligned}$$

are the tangent and normal external unit vectors to the boundary of $\mathcal{C}$ at spatial point $(x(s,t), y(s,t))$ and $v(s,t)$ is the corresponding velocity, where all of them are regarded as 3D vectors. Since s is the arc length, the tangent vector $\tau(s,t)$ has norm 1 and $n(s,t)$ corresponds to the external normal unit vector since it is obtained by clockwise rotating the tangent vector $\tau(s,t)$ by $\pi/2$, see Fig. 5.

Let us consider the following four basic setups or behaviors:

1. The boundary is smooth, so that both $\tau(s,t)$ and $n(s,t)$ are smooth, and the velocity is always smooth. Then, the surface parametrized by σ is everywhere smooth and in

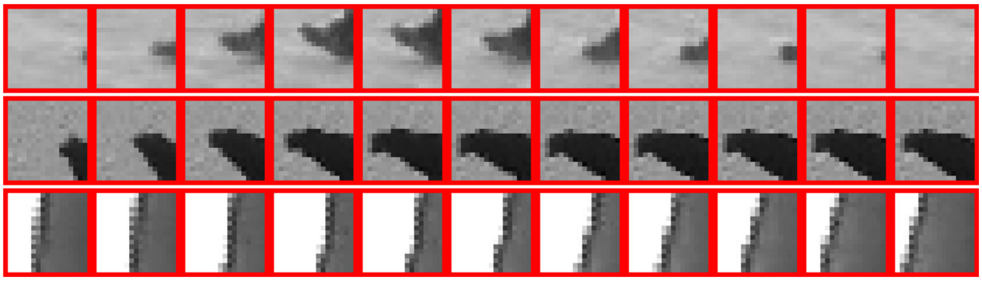


Fig. 6 Space–time features in real data. Top: the tip of a foot changing direction at the end of a step produces a spatio-temporal corner; middle: the tip of a fist in the extension phase of a punching action produces a spatio-temporal edge; bottom: the side of an arm translating as a person is walking leads to a spatio-temporal surface

each point there is a tangent plane whose normal vector is given by $N(s,t)$, (see Fig. 4a); if the velocity is zero, then the normal vector N is simply given by n. Here we expect a single coefficient to have an high value, exactly the one directed along the surface normal.

2. The boundary is smooth, so that both $\tau(s,t)$ and $n(s,t)$ are smooth, but the velocity at time $t = t_0$ is not regular. Hence, the two surfaces

$$\{\sigma(s,t) \mid s \in [0,L], t \in [0,t_0]\} \quad \text{and}$$
$$\{\sigma(s,t) \mid s \in [0,L], t \in [t_0,T]\}$$

create a 3D edge in the plane $t = t_0$ and $N(s,t)$ is discontinuous at $t = t_0$ for all $s \in [0,L]$ with sharp variation given by

$$\Delta N(s,t_0) = \tau(s,t_0) \times \Delta v(s,t_0) \quad \forall s \in [0,1],$$

where Δf is the jump of f (with respect the second variable) at t_0, i.e.,

$$\Delta f(s,t_0) = \lim_{t\to t_0^+} f(s,t) - \lim_{t\to t_0^-} f(s,t),$$

and $\Delta N(s,t_0)$ has a nonzero component only along the t-axis and lives on the 3D edge (see Fig. 4b). In this case, the shearlet coefficients would include two maximum values associated with the two surfaces.

3. The velocity is smooth, but $(x(s_0), y(s_0))$ is a 2D corner of the boundary; then, the two surfaces

$$\{\sigma(s,t) \mid s \in [0,s_0], t \in [0,T]\} \quad \text{and}$$
$$\{\sigma(s,t) \mid s \in [s_0,L], t \in [0,T]\}$$

create a 3D edge parametrized by the temporal evolution of the 2D corner $(x(s_0), y(s_0))$. Hence, $N(s,t)$ is discontinuous at s_0 for all $t \in [0,T]$ with sharp variation given by

$$\Delta N(s_0,t) = \Delta n(s_0,t) + \Delta\tau(s_0,t) \times v(s_0,t) \qquad \forall t \in [0,T],$$

where $\Delta N(s_0,t)$ is the jump of N (with respect the first variable) at s_0 and it has two contributions: the former is in the xy-plane and the latter along the t-axis. As above the vector $\Delta N(s_0,t)$ lives on the 3D edge (see Fig. 4c). Again, the shearlet coefficients would include two maximum values associated with the two surfaces.

4. The boundary has a 2D corner at point $(x(s_0), y(s_0))$, and there is a change of velocity at time $t = t_0$ lighter in the direction or in the speed. At the spatial–temporal point $(x(s_0,t_0), y(s_0,t_0), t_0)$, there is a vertex, which is the junction of the four surfaces

$$\mathcal{S}_1 = \{\sigma(s,t) \mid s \in [0,s_0], t \in [0,t_0]\}$$
$$\mathcal{S}_2 = \{\sigma(s,t) \mid s \in [s_0,L], t \in [0,t_0]\}$$
$$\mathcal{S}_3 = \{\sigma(s,t) \mid s \in [0,s_0], t \in [t_0,T]\}$$
$$\mathcal{S}_4 = \{\sigma(s,t) \mid s \in [s_0,L], t \in [t_0,T]\},$$

where $\mathcal{S}_1$ has a 3D edge in common with $\mathcal{S}_2$ and it has a 3D edge in common with $\mathcal{S}_3$ (and a similar relation for the other three surfaces). At the vertex, there are four normal vectors (see Fig. 4d).

This toy model may be adapted to real data, as we will see in the next sections. We start by observing examples of different local behaviors within video sequences. In Fig. 6 (top), we may observe the evolution of the tip of a foot changing direction at the end of a step; this behavior produces a spatio-temporal corner or vertex. In the center of the figure, we analyze the tip of a fist in the extension phase of a punching action, producing a spatio-temporal (or 3D) edge. Finally, at the bottom, we may observe the side of an arm translating as a person is walking, producing a spatio-temporal surface.

4 Enhancing Space–Time Features with Shearlets

In this section, we propose a method to represent local spatio-temporal information provided by shearlets in order

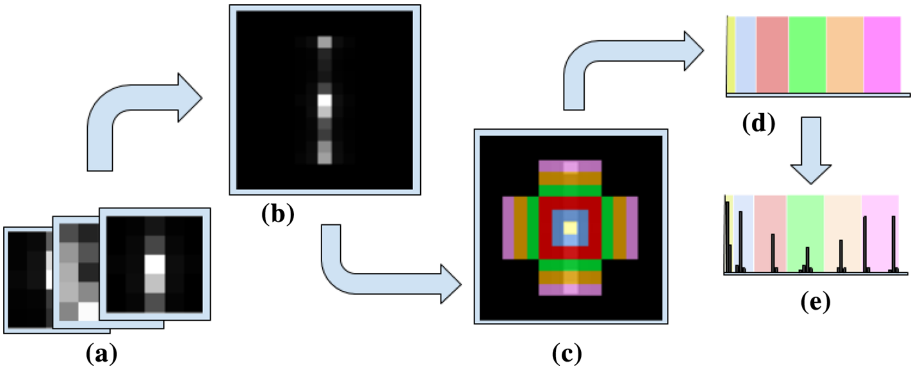


Fig. 7 Main steps of the $2D+T$ signal representation procedure. For each space–time point $\hat{m}$: **a** We compute matrices C_1, C_2 and C_3, **b** We create the object **C** which includes the space–time coefficients of the point neighborhood, (**c**, **d**) We map subsets of elements (i.e., shearlet coefficients) of **C** to different parts of a vector and (**e**) We obtain the representation for our point (Color figure online)

to enhance different types of discontinuities of a $2D+T$ signal.

4.1 The Method

We consider a spatial temporal point $\hat{m} = (\hat{x}, \hat{y}, \hat{t})$ for the fixed scale $\hat{j}$ and the subset of shearings

$$\mathbf{K} = \left\{ k = (k_1, k_2) \mid k_1, k_2 = -\lceil 2^{\hat{j}/2} \rceil, \ldots, \lceil 2^{\hat{j}/2} \rceil \right\},$$

where $M = 2\lceil 2^{\hat{j}/2} \rceil + 1$ is the cardinality of **K**, where we suppressed the dependence on $\hat{j}$ from **K** and M. The procedure we carry out in the discrete case is depicted in Fig. 7 and consists of two parts, which we describe in the following. In the first part, we merge the coefficients obtained from the different pyramids; in the second one, we derive a representation for the point neighborhood considered. This representation should be meaningful of a specific space–time primitive.

4.1.1 Reorganize the Coefficients of a Point Neighborhood

(a) We reorganize the information provided by $\mathrm{SH}[f](\ell, \hat{j}, k, \hat{m})$ in three $M \times M$ matrices, each one associated with a pyramid $\ell = 1, 2, 3$, where each entry is related to a specific shearing: $C_\ell(r, c) = \mathrm{SH}[f](\ell, \hat{j}, k_{rc}, \hat{m})$ with $\ell = 1, 2, 3$, where $r, c = 1, \ldots, M$ and k_{rc} is the corresponding shearing in $\mathbf{K}_j$ defined in (1). As usual in this kind on analysis, we discard the informations related to the shearlet coefficients in the low frequency pyramid $\ell = 0$ since they are related to the smoothness of the signal. Figure 7a shows the three matrices for a specific space–time point.

(b) We merge the three matrices in a single one, by recombining them relatively to the maximum shearlet coefficient (the central element of the column depicted in Fig. 7b). For a given scale j and a fixed set of shearings **K**, the central element of **C** corresponds to $k_{\max}$, the shearing corresponding to the coefficient with the maximum value in the set $\mathrm{SH}[f](\ell, \hat{j}, k, \hat{m})$, with $\ell \in \{1, 2, 3\}$ and $k \in \mathbf{K}_j$. The eight values of **C** around the center (the blue ring in Fig. 7c) correspond to the value associated with the first 8-neighborhoods of $k_{\max}$. These shearing can be in one of the three cones, and hence, the corresponding values are the entries of one of the three matrices C_1, C_2 and C_3. This tiling procedure is repeated to cover to full index set $\mathbf{K}_j$. This property is needed to obtain a rotation-invariant representation in the next steps of this pipeline, since the values in **C** are redistributed similarly when considering two similar spatio-temporal primitives, even if they are oriented differently in the space–time domain. The matrix **C** models how the shearlet coefficients vary in a neighborhood of the direction where there is the maximum variation, and it is built in a way so that coefficients which are referred to shearings which are close one to the other end up being close in **C**. We will see how different kinds of spatio-temporal elements can be associated with different kinds of local variations in **C**.

4.1.2 Compute a Compact Rotation-Invariant Representation

(a) We group the available shearings in subsets $\bar{s}_i$, according to the following rule: $\bar{s}_0 = \{k_{\max}\}$ and $\bar{s}_i$ will contain the shearings in the ith ring of values from $k_{\max}$ in **C** (as highlighted in Fig. 7c). We extract the values corresponding to the coefficients for $\bar{s}_1$ (by looking at the 8-neighborhood of $k_{\max}$), then we consider the adjacent outer ring (that is, the 24-neighborhood without its 8-neighborhood) to have the coefficients corresponding to $\bar{s}_2$, and so on (Fig. 7d, e). By construction, the elements of C are grouped in subsets, each of them associated with a ring, and the first and last element of each subset are closed each other. For the subsets $\bar{s}_i$ for $i > 2$ not all the coefficients are selected, this is due to the way the object **C** is built. Selecting all elements would introduce redundancy in the representation; hence, only some parts of them are considered to build it.

(b) We build a vector concatenating the values of the coefficients corresponding to each set as it follows. We first define $\text{coeff}_{\bar{s}_i}$ to be the set of coefficients associated with each shearings subset $\bar{s}_i$:

$$\text{coeff}_{\bar{s}_0} = \text{SH}[f](\ell_{k_{\max}}, \hat{j}, k_{\max}, \hat{m})$$
$$\text{coeff}_{\bar{s}_i} = \left\{ \text{SH}[f](\ell_{\bar{s}_i}, \hat{j}, k_{\bar{s}_i}, \hat{m}), k_{\bar{s}_i} \in \bar{s}_i \right\},$$

where $\ell_{k_{\max}}$ is the pyramid associated with the shearing $k_{\max}$ and where $\ell_{\bar{s}_i}$ represents the pyramid associated with each shearing $k_{\bar{s}_i}$. Then, we set

$$\mathbf{D}(\hat{m}) = \text{coeff}_{\bar{s}_0} {}^\frown \text{coeff}_{\bar{s}_1} {}^\frown \text{coeff}_{\bar{s}_2} {}^\frown \ldots;$$

where $^\frown$ denotes the concatenation between vectors. The size of the representation is strictly dependent on the number M of shearings, and it depends on the chosen scale, as we introduced previously.

At this point, the object $\mathbf{D}(\hat{m})$ entangles the relations between the direction of maximum variation $s_{\max}$ for a given point $\hat{m}$ and the directions corresponding to the other shearings $k \neq s_{\max}$.

4.2 Expressiveness of Coefficients

We analyze the space–time neighborhood coefficients $\mathbf{C}$ for different types of points. First, we consider a simple synthetic sequence, with a dark square on a white background. At the beginning of the sequence, the square is still; then, at frame 64 it starts translating up with constant speed until frame 108, when the square stops again until the end of the sequence. To avoid boundary problems, the sequence is composed of white frames before frame number 20 and after frame number 108. Figure 8a–c shows a selection of meaningful frames in the synthetic sequence, while Fig. 8d–f shows the volume we may obtain by stacking the video frames (and in particular the square silhouette) one on top of the other. In this synthetic example, we easily identify three types spatio-temporal features, clearly visible on the 3D shape: surface points, 3D edges and vertices; in (d–f) we show manually selected points. Figure 8g–i shows average $\mathbf{C}$ computed on space–time point neighborhood of all the marked points of a given type. In spite of averaging, the 3D visualization we present highlights the neighborhood structure and allows us to show how $\mathbf{C}$ allows us to distinguish between different kinds of spatio-temporal structures. This speaks in favor of the expressiveness of 3D shearlet coefficients for the local space–time analysis we are considering.

At this point, an observation is in order. In the case of surfaces, we identify only one meaningful peak around which we reorganize the other (negligible) contributions. Instead, in the case of 3D edges and 3D corners, $\mathbf{C}$ presents a more peaks than expected. In the case of 3D edges, we would expect two peaks, but in the construction of $\mathbf{C}$, the second peak is replicated, due to the complexity of the point and the periodicity of the matrix C on each subset associated with the different rings. A similar behavior is already observed in Fig. 3.

Furthermore, with respect to the theory, the 3D vertex in Fig. 8f corresponds to the intersection of three surfaces, instead of four. This is due to the fact that we are dealing with a synthetic image with blank frames below the frame 20. The 3D vertices at frame 64 are at the intersection of four surfaces, as expected; however, two of them are coplanar, so that we have only three distinct normal directions.

Figure 9 shows that the space–time neighborhood coefficients $\mathbf{C}$ have a similar behavior in real data. It highlights two points of a real image sequence, an edge (in blue) and a corner (in red). The behavior of the neighborhood coefficients is coherent with what previously discussed.

As a further evidence, we analyze the average $\mathbf{C}$ over sets of key points automatically detected by well-known algorithms in image processing and computer vision. We consider two spatial features, edges [1] and corners [34] and a space–time feature, STIP [27].

Edges Figure 10 shows the average coefficients of all edge points obtained by the Canny detector applied to a 2D frame extracted from video sequence. It is worth noting that since the our algorithm also detects corner points and moving edges, the 3D visualization also includes small lateral peaks.

Corners Figure 11 shows the behavior of corner points, automatically detected by the classical Harris algorithm. In this case, we report the visualization for the subset of still and moving corners, which are more distinctive as expected, since our representation takes into account space–time information, while Harris corner detector does not.

STIP Figure 12 shows the average descriptor for the points detected as Laptev STIPs on a different image frame. It is well known that STIP detector identifies very few points, meaningful both in space and in time. The choice of this specific image frame has been done considering the limitations of the detection algorithm, which performs particularly well only in the presence of very sharp space–time variations. This is clearly identified by the behavior of the neighborhood coefficients; indeed, we observe peaks both in space and in time directions.

5 Identifying Coherent Groups of Points

So far we have discussed the behavior of 3D shearlet coefficients in the space–time neighborhood of a point or a set

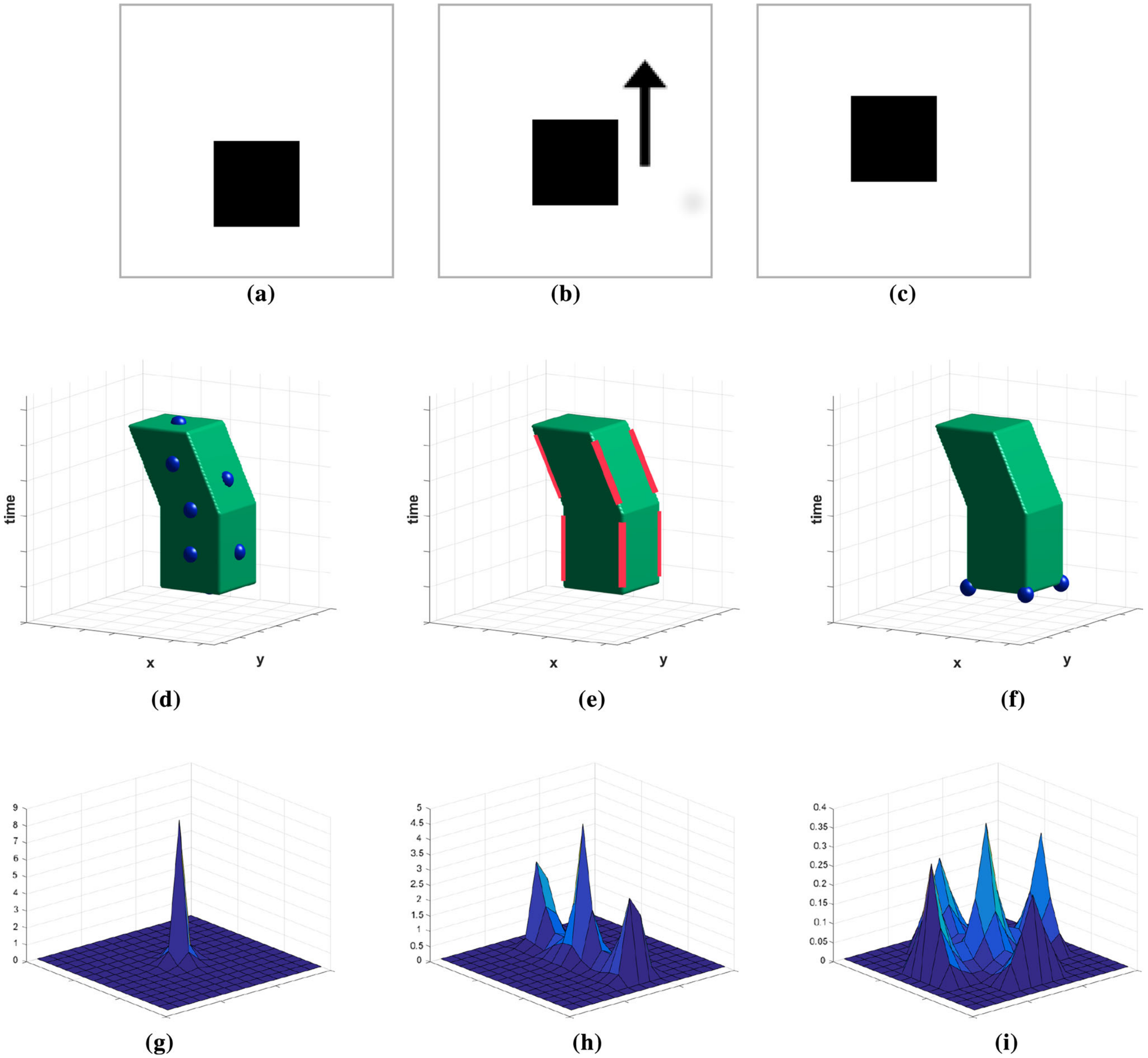


Fig. 8 **a**–**c** Sample frames of the synthetic video sequence. **d**–**f** Manually selected points on the 2D + T surface (**g**–**i**) and corresponding average **C**

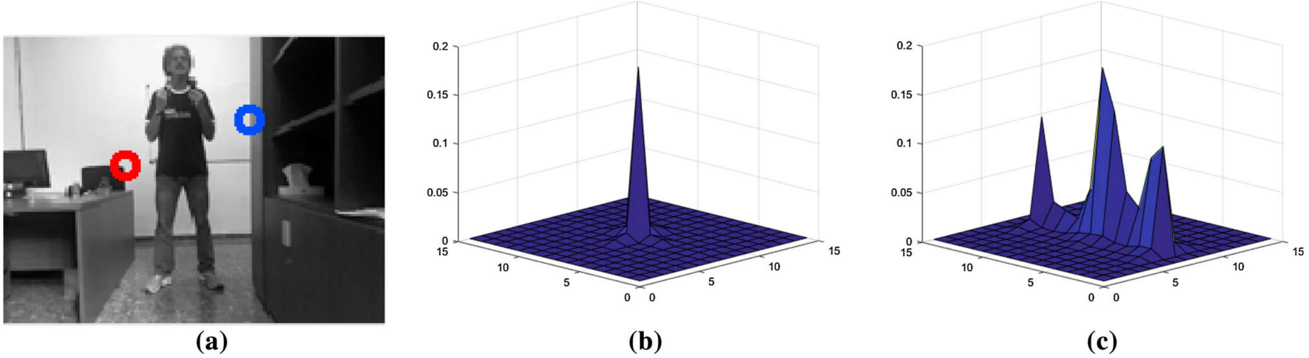
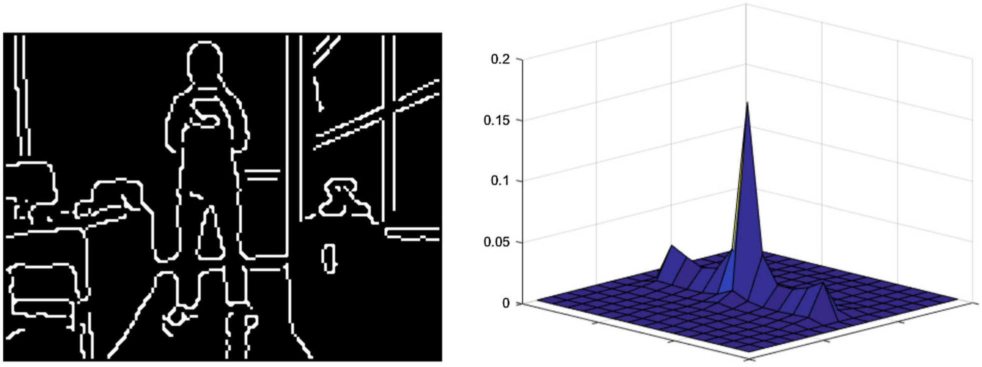
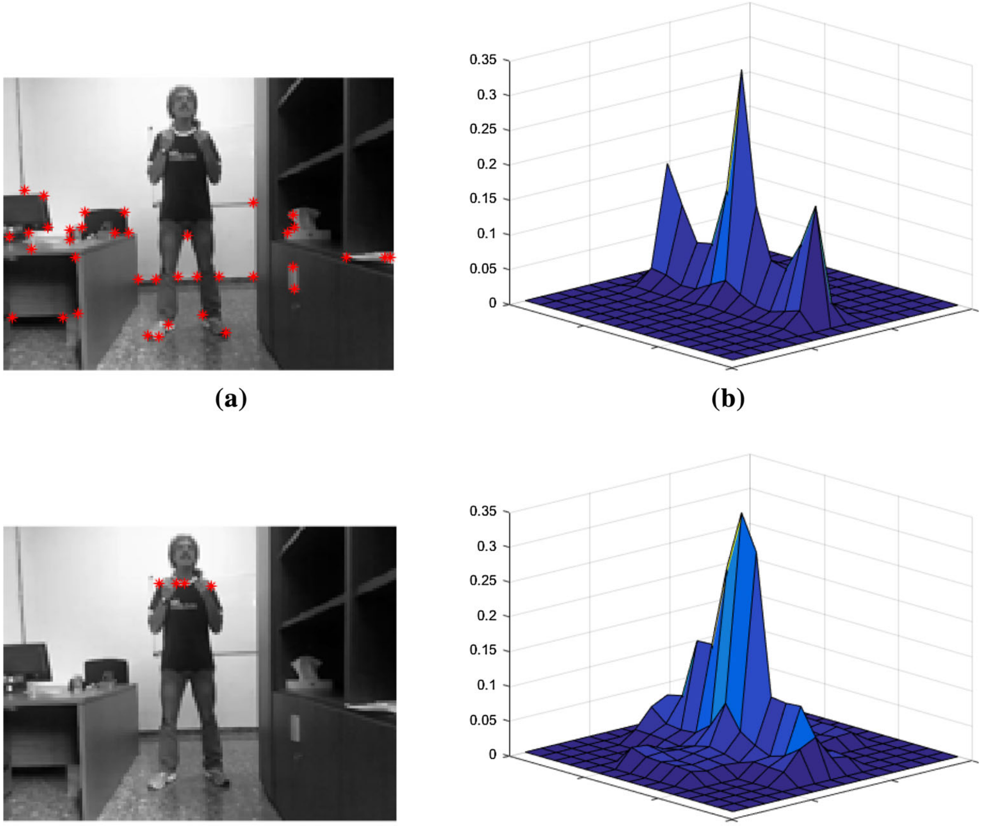


Fig. 9 Example of visualization in 3D of the result of the process, for these example we selected a static spatial Edge (the blue circle) and a static spatial corner (the red circle), which are characterized by two different behaviors of change. **a** Selected points, **b** Edge and **c** Corner (Color figure online)



(a) (b)

Fig. 10 **a** Frame points automatically extracted by Canny edge detector and **b** A 3D visualization of **C** Averaged on all the edge points



(a) (b)

(c) (d)

Fig. 11 Harris corners. **a** Still Harris corners **b** and the shape visualization of their average descriptor. **c** Moving Harris corners (**d**) and the shape visualization of their average descriptor

of previously detected points. Here we discuss how we can group sets of points by similarity, with the goal of identifying automatically different types of space–time primitives.

We fix a frame in a video sequence, we compute the shearlet coefficients of a suitable temporal neighborhood of the frame, and we apply our algorithm to assign the local representation **D** to each point of the given frame. Hence, we cluster the points with a k-means algorithm in p clusters and we consider the clusters centroids as an unsupervised estimate of our space–time primitives of the video frame.

Figure 13 shows the results obtained for different choices of p. The sequence is acquired by a still camera and represents a subject boxing in the air. The frame we selected to present the results represents the exact moment in which the subject is inverting the direction of movement of his arm—

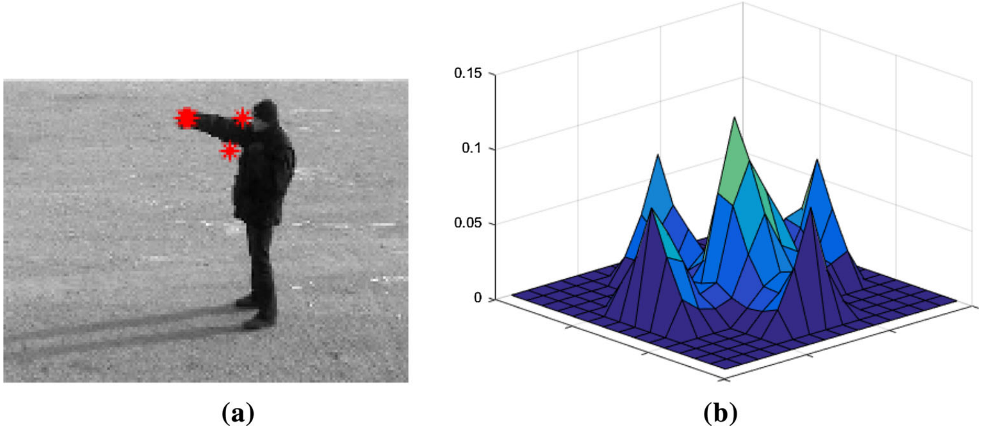


Fig. 12 Laptev STIPs and a 3D visualization of **C** Averaged on all the edge points

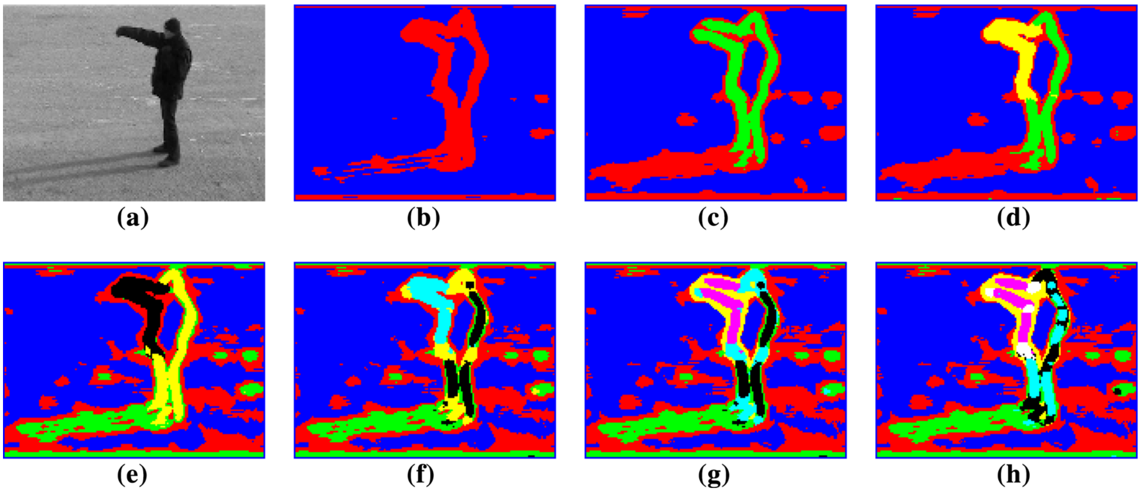


Fig. 13 Clusters of space–time primitives for different choices of p (best seen in pdf). **a** Frame, **b** $p = 2$, **c** $p = 3$, **d** $p = 4$, **e** $p = 5$, **f** $p = 6$, **g** $p = 7$ and **h** $p = 8$

as shown in Fig. 12. Let us briefly comment the results for different choices of p, which highlight space–time points at different granularities:

- $p = 2$: the first partition obtained creates two groups, a set of points containing almost all the points in the sequence without a significant local change neither in space nor in time (background points and those belonging to the inner part of the body of the subject) and another one containing points which are undergoing some spatio-temporal change.
- $p = 3$: the clustering process better separates the points belonging to the background and those related to the shape of the subject, without additionally differentiating these points. Background is divided in two parts, depending on the texture.
- $p = 4$: the additional cluster allows us to separate points that belong to spatio-temporal elements with a higher dynamics, for example, the arm of the subject boxing in the air.
- $p = 5$: a new cluster does not provide significant changes.
- $p = 6$: different elements are now separated in a very nice way, the edges belonging to the arm are grouped in a separate cluster w.r.t. the edges belonging to the back and the legs, also, it is possible to see how points which look like spatial corners are grouped together (in the yellow cluster), without any differentiation regarding their spatio-temporal behavior.
- $p = 7$: no additional information.
- $p = 8$: the points colored in white represent the last cluster added within this trial, we can see how these elements could correspond to spatial corners with particular dynamics (the fist is inverting direction, the corners joining the arm to the head and to the chest undergo some changes, and the front tip of the jacket is moving while the

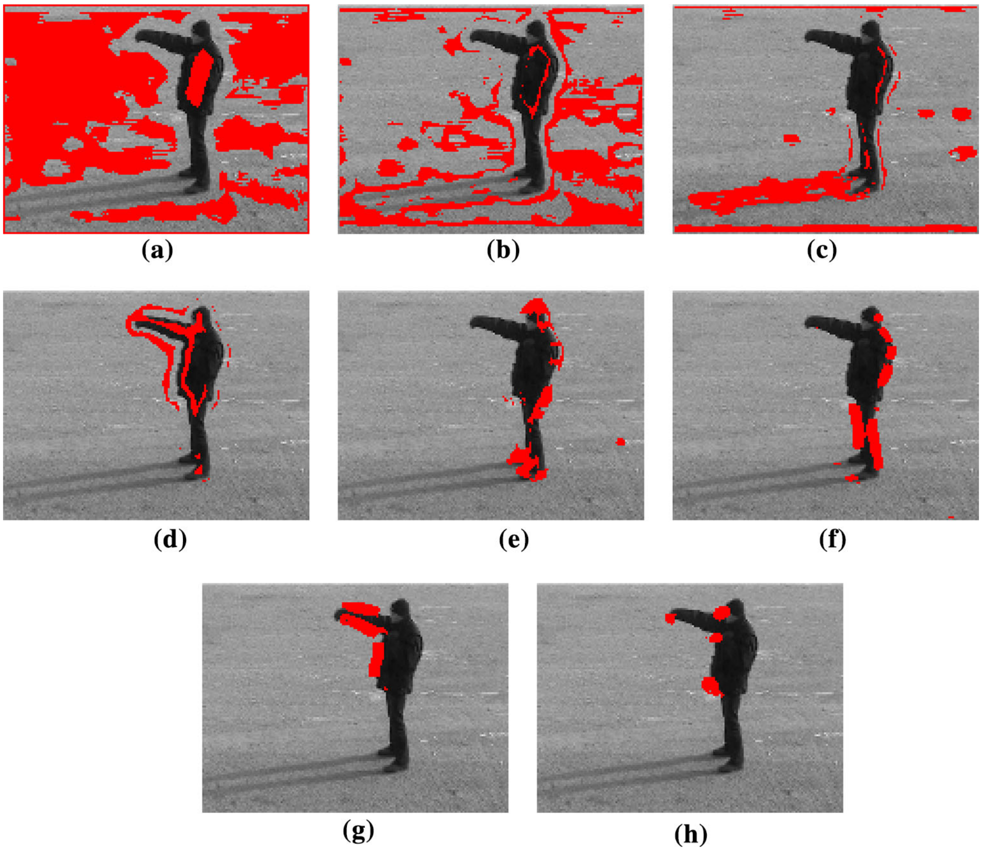


Fig. 14 Points belonging to the different eight clusters calculated on a frame of the *boxing* sequence (see text). **a** Cluster # 1 from $p = 8$, **b** Cluster # 2 from $p = 8$, **c** Cluster # 3 from $p = 8$, **d** Cluster # 4 from $p = 8$, **e** Cluster # 5 from $p = 8$, **f** Cluster # 6 from $p = 8$, **g** Cluster # 7 from $p = 8$ and **h** Cluster # 8 from $p = 8$

subject is punching). These points are also highlighted in Fig. 14h, and the corresponding average **C** is highlighted in Fig. 15h. Their similarity with the STIP points in Fig. 12 is apparent.

This result highlights many nice properties of our descriptor: the separations of all the points of the image frame into different sets, with respect to their spatio-temporal behavior, is obtained thanks to a space–time continuity of the representation inherited by the shearlet transform; as p grows we may identify an interesting nested structure; even in an entirely unsupervised approach most of the points clusters automatically detected can be associated with known feature points, such as edges or corners.

As a last observation, we discuss whether the estimated space–time clusters are persistent among different video frames and different video. The intuition is that the answer should be negative since the estimated space–time primitives are learnt by a short temporal observation, and thus, different primitives may be present or not. To this purpose, we compare sets of primitives estimated on different frames and compare them through the Euclidean distance, building similarity matrices. Note that, in every matrix, the entries of the two sets have been reordered so that to keep the values corresponding to the best similarity obtained along the diagonal, and that the assignment of the entries of the two centroids sets has been carried on by means of the Hungarian algorithm. Figure 16 shows the self-similarity within a set of space–time primitives. We consider this example as a baseline observation, showing how the primitives are somewhat redundant (this is visible by the block structure of the matrix that shows how different primitives are similar to one another). If we compare centroids obtained at different frames of the same sequence (Fig. 17), we observe again a very similar dominant diagonal, possibly due to the fact we are observing a periodic action. If we compare video frames from different type of actions, we obtain noisier similarity matrices. Figure 18 compares a boxing frame with a handwaving frame; in this case, the dominant diagonal is still present, showing that each primitive has at least a counterpart on the other frame. In fact, the two actions, even if they are quite different, have many things in common: they are upper body actions,

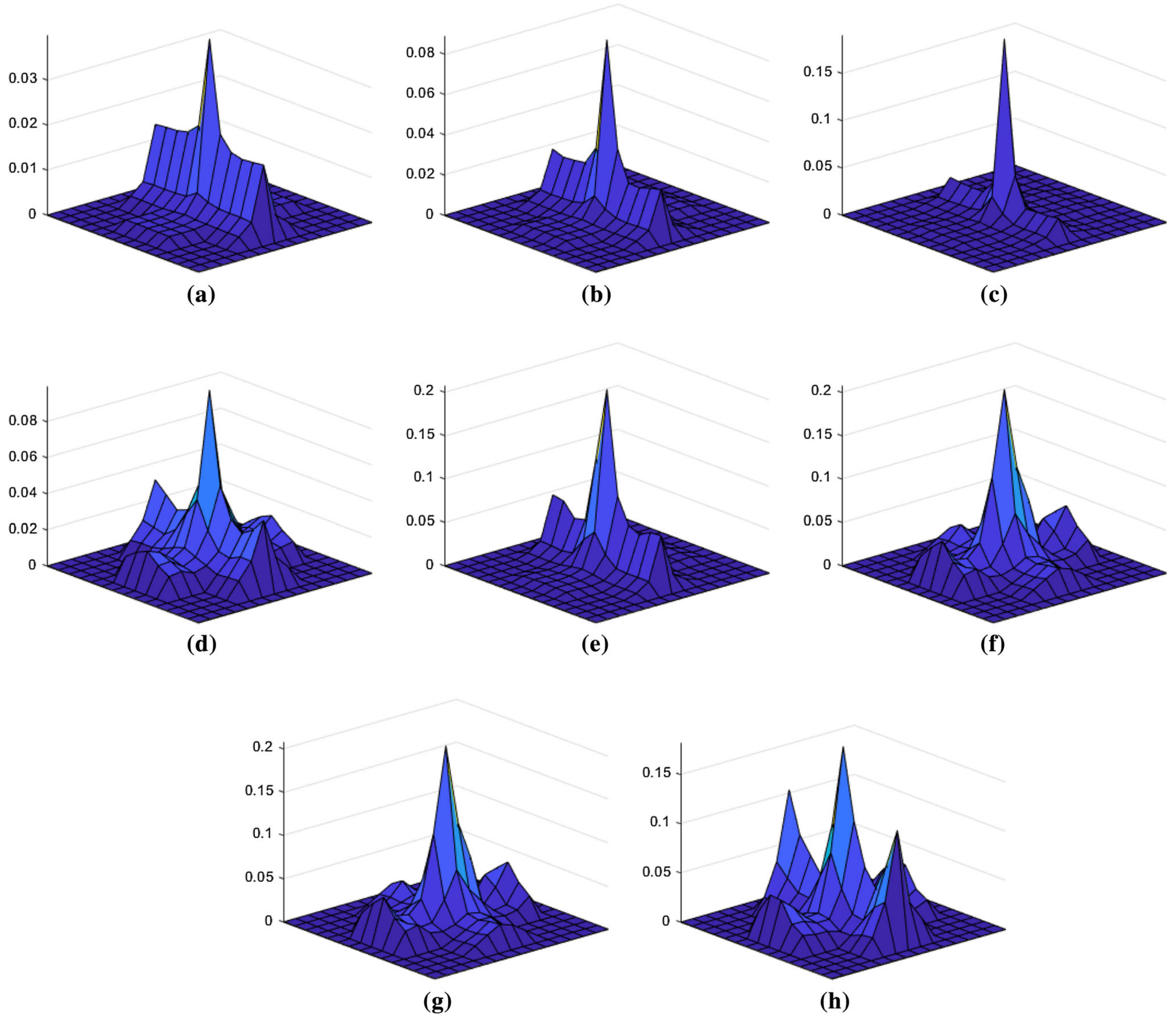


Fig. 15 3D visualization of the **C** objects related to the centroids of the clusters shown in Fig. 14a–g. **a C** for cluster # 1, **b C** for cluster # 2, **c C** for cluster # 3, **d C** for cluster # 4, **e C** for cluster # 5, **f C** for cluster # 6, **g C** for cluster # 7 and **h C** for cluster # 8

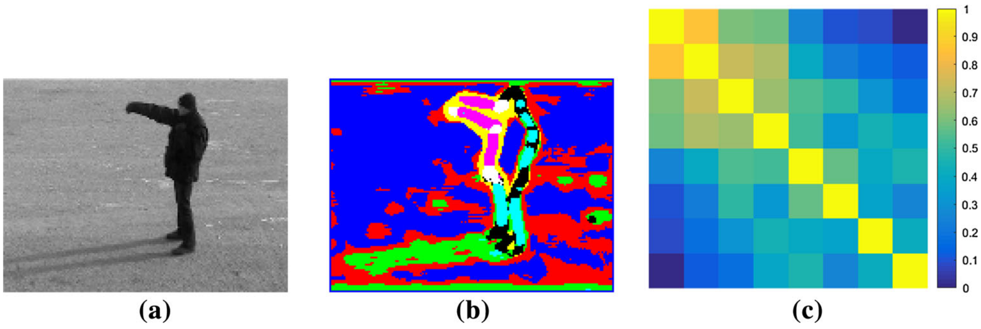


Fig. 16 Self-similarity matrix for a video frame of the *boxing* sequence. **a** Frame, **b** Clusters and **c** Self-similarity

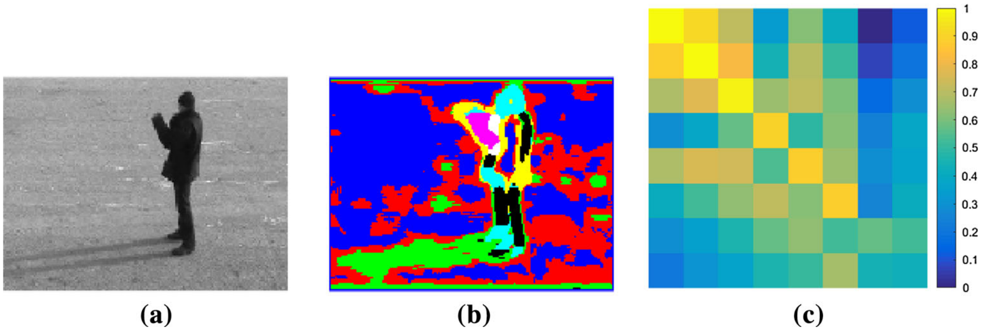


Fig. 17 Similarity matrix between two video frames of the *boxing* sequence (the reference frame is shown in Fig. 16). **a** Frame, **b** Clusters and **c** Similarity

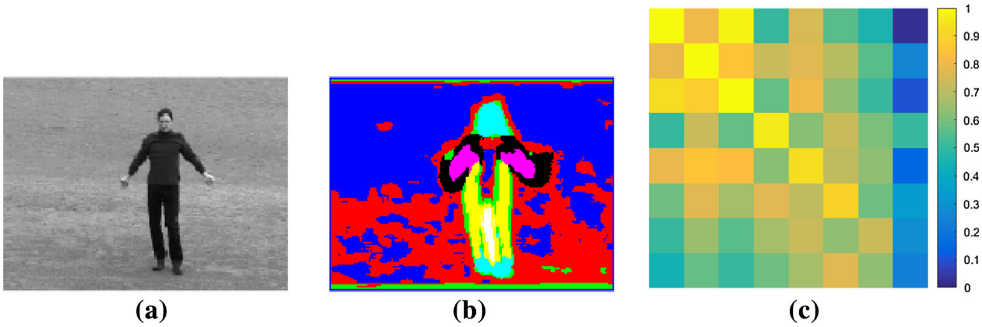


Fig. 18 Similarity matrix between a video frame of the *boxing* sequence (Fig. 16) and a frame of the *handwaving* sequence. **a** Frame, **b** Clusters and **c** Similarity

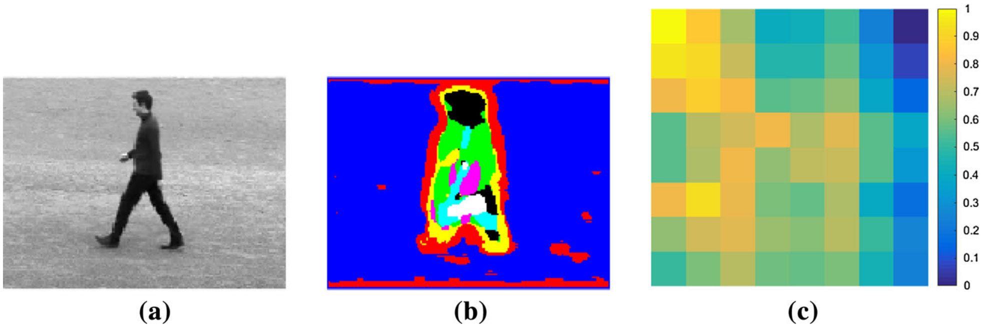


Fig. 19 Similarity matrix between a video frame of the *boxing* sequence (Fig. 16) and a frame of the *walking* sequence. **a** Frame, **b** Clusters and **c** Similarity

with abrupt changes of direction and are executed at a similar pace; thus, we expected them to share at least a subset of very similar spatio-temporal primitives. Finally, Fig. 19 compares the boxing with a walking frame, two very different types of dynamics, as confirmed by the noisy similarity matrix we obtain.

6 Conclusions

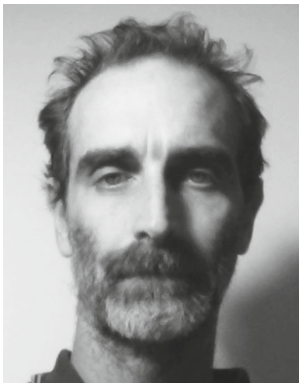
In this paper, we discussed how to analyze space–time signals, or more specifically video sequences, in the framework of shearlets. The goal of our work was to evaluate the behavior of the signal in a space–time local neighborhood. Starting from a theoretical analysis, followed by toy as well as real examples, we discussed what are the typical patterns one may find in space–time signals. Then, we derived a point representation based on signal coefficients and show that it appears to be stable on set of points of the same nature, while also meaningfully highlighting their spatio-temporal behavior. Based on this property, we derived an unsupervised approach to identify different space–time primitives of a video frame. This primitives are the centroids of space–time points clusters obtained by the k-means algorithm. Our analysis shortens the gap between theory and algorithms and allows us to derive a computational model which may be applied to motion analysis and action recognition.

In this paper, we considered one frame at a time with its temporal neighborhood. We are currently investigating how to integrate the analysis at the level of the entire video. We conclude by observing that shearlets may lead to a perfect scale invariant representation. On 2D signals, this has been clearly demonstrated in the theory and exploited in practice in [8]. Furthermore, it would be of interest to exploit the multi-scale property of the shearlet coefficients to detect spatial-temporal patterns at different scales. This requires a representation with a large number of different scales and, at the present, this poses some implementation problems, whose solution will be the objective of future work.

References

1. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 679–698 (1986)
2. Chen, Z., Hao, X., Sun, Z.: Image denoising in shearlet domain by adaptive thresholding. J. Inf. Comput. Sci. **10**(12), 3741–3749 (2013)
3. Dahlke, S., Steidl, G., Teschke, G.: The continuous shearlet transform in arbitrary space dimensions. J. Fourier Anal. Appl. **16**(3), 340–364 (2010)
4. Dawn, D.D., Shaikh, S.H.: A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector. Vis. Comput. **32**(3), 289–306 (2016)
5. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans. Image Process. **14**(12), 2091–2106 (2005)
6. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (2005)
7. Duval-Poo, M.A., Odone, F., De Vito, E.: Edges and corners with shearlets. IEEE Trans. Image Process. **24**(11), 3768–3780 (2015)
8. Duval-Poo, M.A., Noceti, N., Odone, F., De Vito, E.: Scale invariant and noise robust interest points with shearlets. IEEE Trans. Image Process. **26**(6), 2853–2867 (2017)
9. Easley, G.R., Labate, D., Colonna, F.: Shearlet-based total variation diffusion for denoising. IEEE Trans. Image Process. **18**(2), 260–268 (2009)
10. Escalera, S., Baro, X., Gonzalez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce-Lopez, V., Escalante, H.J., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: dataset and results. In: ECCV Workshops (2014)
11. Guo, K., Labate, D.: Optimally sparse multidimensional representation using shearlets. SIAM J. Math. Anal. **39**(1), 298–318 (2007)
12. Guo, K., Labate, D.: Analysis and detection of surface discontinuities using the 3D continuous shearlet transform. Appl. Comput. Harmon. Anal. **30**(2), 231–242 (2011)
13. Guo, K., Labate, D.: Optimally sparse representations of 3D data with C^2 surface singularities using Parseval frames of shearlets. SIAM J. Math. Anal. **44**, 851–886 (2012)
14. Guo, K., Labate, D., Lim, W.Q.: Edge analysis and identification using the continuous shearlet transform. Appl. Comput. Harmonic Anal. **27**(1), 24–46 (2009)
15. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, vol. 15, pp. 10–5244. Manchester, UK (1988)
16. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: Proceedings of the tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1–Volume 01, ICCV '05, pp. 166–173, Washington, DC, USA. IEEE Computer Society (2005)
17. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: In BMVC+08 (2008)
18. Kutyniok, G., Labate, D.: Resolution of the wavefront set using continuous shearlets. Trans. Am. Math. Soc. **361**(5), 2719–2754 (2009)
19. Kutyniok, G., Labate, D.: Introduction to Shearlets. Springer, Berlin (2012)
20. Kutyniok, G., Labate, D.: Shearlets. Appl. Numer. Harmon. Anal. Birkhäuser/Springer, New York (2012)
21. Kutyniok, G., Lim, W.Q.: Compactly supported shearlets are optimally sparse. J. Approx. Theory **163**(11), 1564–1589 (2011)
22. Kutyniok, G., Petersen, P.: Classification of edges using compactly supported shearlets. Appl. Comput. Harmonic Anal. **42**(2), 245–293 (2017)
23. Kutyniok, G., Lemvig, J., Lim, W.Q.: Optimally sparse approximations of 3D functions by compactly supported shearlet frames. SIAM J. Math. Anal. **44**(4), 2962–3017 (2012)
24. Kutyniok, G., Lim, W.Q., Reisenhofer, R.: Shearlab 3d: faithful digital shearlet transforms based on compactly supported shearlets. ACM Trans. Math. Softw. **42**, 5:1–5:42 (2016)
25. Labate, D., Lim, W.Q., Kutyniok, G., Weiss, G.: Sparse multidimensional representation using shearlets. In: Optics and Photonics (2005)
26. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8. IEEE (2008)
27. Laptev, I.: On space–time interest points. Int. J. Comput. Vis. **64**(2), 107–123 (2005)
28. Lindeberg, T.: Scale-Space Theory in Computer Vision, vol. 256. Springer, Berlin (1993)
29. Mallat, S., Hwang, W.L.: Singularity detection and processing with wavelets. IEEE Trans. Inf. Theory **38**(2), 617–643 (1992)
30. Negi, P.S., Labate, D.: 3D discrete shearlet transform and video processing. IEEE Trans. Image Process. **21**(6), 2944–2954 (2012)
31. Oikonomopoulos, A., Patras, I., Pantic, M.: Spatiotemporal salient points for visual recognition of human actions. IEEE Trans. Syst. Man Cybern. Part B **36**(3), 710–719 (2006)
32. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: 17th International Conference on Proceedings of the Pattern Recognition (ICPR'04) Volume 3–Volume 03 (2004)
33. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM International Conference on Multimedia, MM '07, pp. 357–360, New York, NY, USA. ACM (2007)
34. Shi, J., Tomasi, C.: Good features to track. In: 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94, pp. 593–600. IEEE (1994)
35. Willems, G., Tuytelaars, T., Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: Proceedings of the 10th European Conference on Computer Vision: Part II, ECCV '08 (2008)
36. Wong, S.F., Cipolla, R.: Extracting spatiotemporal interest points using global information. In: ICCV, pp. 1–8. IEEE Computer Society (2007)
37. Zhang, Z., Tao, D.: Slow feature analysis for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3), 436–450 (2012)

Damiano Malafronte obtained his B.Sc. in Computer Science, in 2012, and his M.Sc. in Computer Science, in 2014, both from the University of Genoa (Italy). He is currently completing his Ph.D. in Computer Science under the supervision of Prof. Francesca Odone and Prof. Ernesto De Vito, researching on the use of the Shearlet Transform for the analysis of spatio-temporal signals. His research interests are in the fields of signal and image processing, computer vision and human–machine interaction.



Ernesto De Vito received a Laurea degree on Physics, summa cum laude, in 1991, and a Ph.D. in Physics in 1995 both from the University of Genova (Italy). He visited the "Laboratoire de Physique Theorique", Universite de Sophia Antipolis, Nice (France), in 1996 as a post-doc. In 1997–2007, he was Assistant Professor at the University of Modena (Italy) in the Department of Mathematics. In 2007, he moved to the Department of Mathematics at University of Genova (Italy). He is now Full Professor in the same department. He published over 50 papers on international journals and conference proceedings on machine learning, harmonic analysis and quantum mechanics. His current research interests include the mathematical aspects of machine learning theory and the study of reproducing formulas both in the continuous and discrete frameworks for signal analysis.

Francesca Odone is an Associate Professor in Computer Science at the University of Genova, Italy, where she leads the Computational Vision group of her Department. She received a Laurea degree in Information Sciences and a Ph.D. in Computer Science both from the University of Genova. For over 2 years, in 1999–2000, she was a visiting student at Heriot-Watt University (Edinburgh, UK) with a EU Marie Curie research grant. In 2002–2005, she was a researcher at the Italian National Institute for Solid State Physics. Her research interests are in the fields of Computer Vision and Machine Learning. In particular, most of her research activity in recent years has been devoted to finding good visual representations, able to capture the complexity of a problem, while allowing for the design of systems with the ability to perform their visual tasks in real time. Over the years, she published over 100 papers on these research topics, on international conferences and journals.