Preprint submitted to
SAMPLING THEORY IN SIGNAL AND IMAGE PROCESSING

DVI file produced on

Local Spatio-Temporal Representation using the 3D Shearlet Transform

Damiano Malafronte
DIBRIS, Università di Genova, Via Dodecaneso 35
Genova, 16146, Italy
damiano.malafronte@dibris.unige.it

Ernesto De Vito
DIMA, Università di Genova, Via Dodecaneso 35
Genova, 16146, Italy
devito@dima.unige.it

Francesca Odone
DIBRIS, Università di Genova, Via Dodecaneso 35
Genova, 16146, Italy
francesca.odone@unige.it

Abstract. In this work we address the problem of analyzing video sequences and of representing meaningful space-time points of interest by using the 3D shearlet transform. We introduce a local representation based on shearlet coefficients of the video, regarded as 2D+T signal. This representation turns out to be informative to understand the local spatio-temporal characteristics, which can be easily detected by an unsupervised clustering algorithm.

Keywords : shearlet transform, spatio-temporal features, video analysis

2010 AMS Mathematics Subject Classification — 42C40

1. Introduction

In this paper we consider a specific type of signal, video sequences, where spatial shapes and structures evolve over time. Usually, in video sequence analysis, the goal is to identify space-time points of interest, which may be associated with space-time discontinuities on the $2D+T$ signal. In recent years such points have been studied with a reference to space-scale theory [20], see for instance [19, 4, 21].

In the meanwhile, many multi-scale methods have been introduced to deal with multi-dimensional signals. Among them, shearlets [18] emerge by their ability to efficiently capture anisotropic features [12], to provide an optimal sparse representation [8, 16], to detect singularities [11, 17] and to be stable

against noise and blurring [6, 1]. For for further details, implementations, and references see [13] and the website[1].

The effectiveness of shearlets is supported by a well-established mathematical theory [2] and it is tested in many applications in image processing by providing efficient algorithms [13, 5].

It is natural to expect that 3D-shearlet representation will be applicable to the analysis of $2D + T$ signals even if the latter carry rather peculiar properties and behaviors. Based on this representation, we propose a procedure to analyze the shearlet coefficients of video sequences and we show that they efficiently enhance different types of local spatio-temporal points of interest.

The paper is organized as follows. In Section 2 we review the discrete shearlet transform specialized to the $2D + T$ case and we show that the directional informations encoded with the shearlets coefficients are naturally associated with spatial-temporal points of interest. Then, in Section 3, we propose a procedure to represent shearlets coefficients, in order to describe different types of space-time features. The procedure is motivated by the fact that the directional information embedded in a space-time neighborhood is more complex to treat than the purely $2D$ case. Our contribution is in deriving an efficient way to collect neighboring information. Section 4 shows an empirical evidence of our claim on a synthetic and a real case. Section 5 is left to a final discussion.

2. **Shearlet Theory: an overview**

Shearlets in arbitrary space dimensions were first introduced in [2] in the continuous realm. Here we briefly review the construction of the discrete shearlet transform of a $2D + T$ signal f by adapting the approach given in [15] for 3D signals.

Denoted by L^2 the Hilbert space of square-integrable functions $f : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{C}$ with the usual scalar product $\langle f, f' \rangle$, the discrete shearlet transform $SH[f]$ of a signal $f \in L^2$ is the sequence of coefficients

$$SH[f](\ell, j, k, m) = \langle f, \Psi_{\ell,j,k,m} \rangle$$

where $\{\Psi_{\ell,j,k,m}\}$ is a family of filters parametrized by

a) a label $\ell = 0, \dots, 3$ associated with four regions $\mathcal{P}_\ell$ in the frequency domain;
b) the scale parameter $j \in \mathbb{N}$;
c) the shearing vector $k = (k_1, k_2)$ where $k_1, k_2 = -\lceil 2^{j/2} \rceil, \dots, \lceil 2^{j/2} \rceil$;
d) the translation vector $m = (m_1, m_2, m_3) \in \mathbb{Z}^3$.

For $\ell = 0$ the filters, which do not depend on j and k, are

$$\Psi_{0,m}(x, y, t) = \varphi(x - cm_1)\varphi(y - cm_2)\varphi(t - cm_3), \tag{1}$$

[1] `http://www.shearlab.org`

where $c > 0$ is a step size and φ is a $1D$-scaling function. The system $\{\Psi_{0,m}\}_m$ takes care of the low frequency cube

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

For $\ell = 1$ the filters are defined in terms of translations and two linear transformations

$$A_{1,j} = \begin{pmatrix} 2^j & 0 & 0 \\ 0 & 2^{j/2} & 0 \\ 0 & 0 & 2^{j/2} \end{pmatrix} \qquad S_{1,k} = \begin{pmatrix} 1 & k_1 & k_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

namely the parabolic dilations and the shearings, so that

$$\Psi_{1,j,k,m}(x, y, t) = 2^j \psi_1 \left(S_{1,k} A_{1,j} \begin{pmatrix} x \\ y \\ t \end{pmatrix} - \begin{pmatrix} cm_1 \\ \hat{c}m_2 \\ \hat{c}m_3 \end{pmatrix} \right), \tag{2}$$

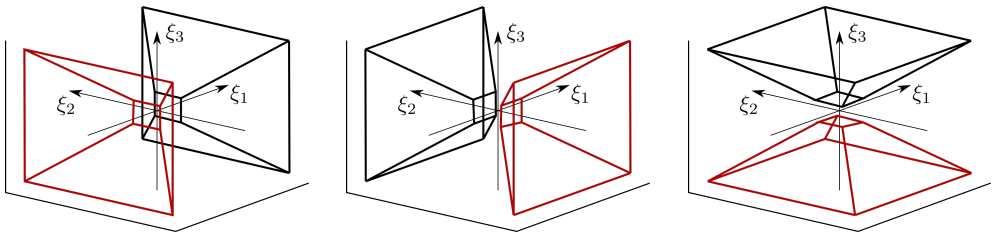
where c is as in (1) and $\widehat{c} > 0$ is another step size (in the rest of the paper we assume that $c = \hat{c} = 1$ for sake of simplicity). The Fourier transform of mother shearlet ψ_1 is of the form

$$\widehat{\psi}_1(\xi_1, \xi_2, \xi_3) = \widehat{\psi}(\xi_1)\, P(\xi_1, \xi_2) \widehat{\varphi}(\xi_2)\, P(\xi_1, \xi_3) \widehat{\varphi}(\xi_3), \tag{3}$$

where P is a given polynomial 2D fan filter [3], ψ is the 1D wavelet function associated with the scaling function φ (here $\hat{f}$ denotes the Fourier transform of a function f). Note that, according to (2), the coarsest scale corresponds to $j = 0$. The system $\{\Psi_{1,j,k,m}\}$ takes care of the high frequencies in the (two-fold) pyramid along the x-axis

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

For $\ell = 2, 3$ we have a similar definition by interchanging the role of x and y (for $\ell = 2$) and of x and t (for $\ell = 3$). Figure 1 shows a visualization of the three pyramidal partitions associated with high frequencies in the Fourier domain.



(a) (b) (c)

Figure 1. The three two-fold 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 with its negative part.

Our algorithm is based on the following nice property of the shearlet coefficients. As shown in [9, 10, 14] if f is locally regular in a neighborhood of m, then $SH[f](\ell, j, k, m)$ has a fast decay when j goes to infinity for any $\ell \neq 0$ and k. If f has a surface singularity at m with normal vector $(1, n_1, n_2) \in \mathcal{P}_1$, then $SH[f](\ell, j, k, m)$ has a fast decay for any $\ell \neq 1$ or $k \neq (\lceil 2^{j/2} n_1 \rceil, \lceil 2^{j/2} n_2 \rceil) =: k^*$, whereas if $\ell = 1$ and $k = k^*$ the shearlet coefficients have slow decay (a similar result holds if the normal direction of the surface singularity belongs to the other two pyramids). We want to point out the fact that while this holds for surface singularities, isolated point singularities are not efficiently detected. These results allow to associate to any shearing vector $k = (k_1, k_2)$ a direction (without orientation) parametrized by two angles, namely *latitude* and *longitude*, given by

$$(\cos\alpha \cos\beta, \cos\alpha \sin\beta, \sin\alpha) \qquad \alpha, \beta \in [-\frac{\pi}{2}, \frac{\pi}{2}]. \tag{4}$$

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

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

The fact that shearlets are sensitive to orientations allows to select different spatial-temporal features. For example, assume that the region of interest is a rigid body whose boundary is described at the initial time $t = 0$ by the curve

$$x = x(s) \qquad y = y(s) \qquad s \in [0, 1].$$

The evolution of the body in time describes a 3D-volume whose boundary is the surface parametrized by

$$\begin{cases} x = x(s,t) \\ y = y(s,t) \\ t = t \end{cases} \qquad s \in [0,1], t \in [0,T],$$

where for each $s \in [0, 1]$, $t \mapsto (x(s,t), y(s,t))$ is the time evolution of the point $(x(s), y(s))$ on the curve at time $t = 0$. A computation shows that the normal vector to the surface is

$$N(s,t) = n(s,t) + \tau(s,t) \wedge v(s,t), \tag{5}$$

where $\tau(s,t)$ and $n(s,t)$ are the tangent and the normal vectors (in the xy-plane) to the boundary of the body at time t in the point $(x(s,t), y(s,t))$ and $v(s,t)$ is the corresponding $2D$-velocity vector[2] (Figure 2). Hence, if the boundary has not corners, but at time $t = t_0$ there is a discontinuous change of velocity $\Delta v(s, t_0)$, then

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

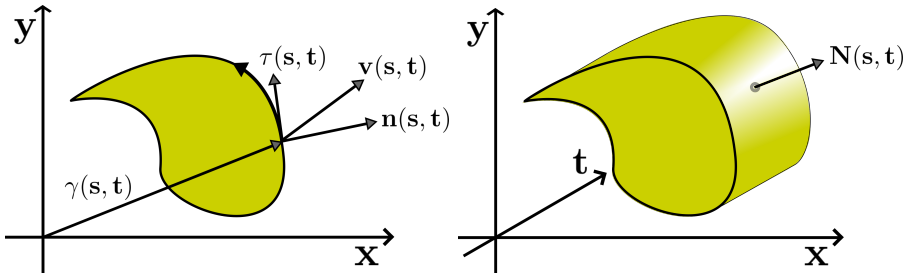
[2]In (5) the $2D$-vectors are regarded as 3D-vectors where the t-component is zero.

which has a non-zero component only along the t-axis. This behaviors looks like an edge in the plane $t = t_0$. On the contrary, if $(x(s_0), y(s_0))$ is a spatial corner of the body, but the velocity is always smooth, then

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

which has non-zero components both in the xy-plane, $\Delta n(s_0, t)$, and along the t-axis, $\Delta\tau(s_0, t) \wedge v(s_0, t)$, so that we have an edge given by the temporal trajectory of the spatial corner.

To compute the shearlet coefficients we use the digital implementation described in [15] based on the relation between the pair scaling function/wavelet (φ, ψ) and the quadrature mirror filter pair (h, g), which in our application is the filter pair introduced in [22].



(a) at a fixed time $\mathbf{t}$ (b) evolution over time

Figure 2. A cartoon-like object with the main relevant geometrical and dynamical quantities (see text for details).

3. Representation of Spatio-Temporal points

In this section we propose a representation which allows us to aggregate local spatio-temporal information provided by shearlets in order to enhance different types of discontinuities of a $2D + T$ signal.

Fixed a scale $\hat{j}$, we associate to any point $\hat{m}$ of the $2D + T$ signal the representation defined by the following three step procedure, depicted in Figure 3.

1 - Reorganize the shearlet coefficients at the point $\hat{m}$.

(i): Defined the set of discrete shearings

$$\mathbf{K} = \left\{ k = (k_1, k_2) \mid k_1, k_2 = -\lceil 2^{j/2} \rceil, \ldots, \lceil 2^{j/2} \rceil \right\}. \tag{6}$$

of size $M \times M$, we reorganize the information provided by $SH[f](\ell, \hat{j}, k, \hat{m})$ in three $M \times M$ matrices , with M equal to the cardinality of the set $\left\{ k_1 \mid k_1 = -\lceil 2^{j/2} \rceil, \ldots, \lceil 2^{j/2} \rceil \right\}$ (in Figure 3, M is equal to 5). Each matrix is associated with a pyramid ℓ, where each entry is related to a specific shearing: $C_\ell(r, c) = SH[f](\ell, \hat{j}, k_{rc}, \hat{m})$ with $\ell = 1, 2, 3$, where $r, c = 1, \ldots, M$ and k_{rc} is the corresponding shearing in $\mathbf{K}_j$ defined in

Eq. (6). Figure 3 (a) shows the three matrices for a specific space-time point.

(ii): We merge the three matrices in a single one, by recombining them relatively to the maximum absolute value shearlet coefficient (Figure 3 (b)). Fixed a scale j and the corresponding set of shearings $\mathbf{K}$, we tile the three matrices C_1, C_2, C_3 in a bigger matrix $\mathbf{C}$. Then, we shift the entries both horizontally and vertically, in such a way that the obtained $\mathbf{C}$ is centered on k_{max}, the shearing corresponding to the coefficient with the maximum value in the set $SH[f](\ell, \hat{j}, k, \hat{m})$, with $\ell = 1, 2, 3$, and $k \in \mathbf{K}$. This property is necessary to obtain a rotation invariant representation in the following steps of this pipeline, for the values in $\mathbf{C}$ are redistributed similarly when two similar spatio-temporal structures are considered, even if they are oriented differently in the space-time domain. The matrix $\mathbf{C}$ models how the shearlet coefficients vary in a neighborhood of the direction with maximum variation, and it is built so that coefficients associated with close shearings are also close in $\mathbf{C}$. We will see how different kinds of spatio-temporal elements can be associated with different kinds of local variations in $\mathbf{C}$. These different patterns can be better appreciated with a 3D visualization (see Figure 4).

2 - Compute a rotation-invariant representation

(i): 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 *i-th* ring of values from k_{max} in $\mathbf{C}$ (as highlighted Figure 3 (c)). 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 (Figure 3 (d) and (e)).

(ii): We build a vector containing the values of the coefficients corresponding to each set as follows:

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

where $\frown$ is the concatenation operator, we define $\mathit{coeff}_{\bar{s}_i}$ to be the set of coefficients associated with each shearings subset $\bar{s}_i$:

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

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

(iii): Finally, we obtain the representation $\mathbf{D}(\hat{m})$ for point $\hat{m}$. The size of the representation depends on the number M of shearings and the scale j.

3 - Derive a final reduced representation The representation $\mathbf{D}(\hat{m})$ entangles the relationships between the direction of maximum variation s_{max}

for a given point $\hat{m}$ and the directions corresponding to the other shearings $k \neq s_{max}$, organized in squared rings of increasing side, see Figure 3 (b) where the colors label the different rings. In real applications, in order to ensure stability it is often useful to have a more compact representation.

(i): To this aim, the final compact representation $\mathbf{F}(\hat{m})$ is obtained by summing up the shearlet coefficients in the same squared ring (see Figure 3 (f)). For example, the first entry of the vector $\mathbf{F}(\hat{m})$ is simply the shearlet coefficient corresponding to s_{max} (the yellow pixel in Figure 3 (b)), the second entry of $\mathbf{F}(\hat{m})$ is the sum over the eight shearlet coefficients associated with the shearings in the second ring (the blue pixels in Figure 3 (b)), and so on. We consider two instances of the representation $\mathbf{F}(\hat{m})$:

- $\mathbf{F}_i(\hat{m})$, built by only considering the representation $\mathbf{D}(\hat{m})$ at a single scale i,
- $\mathbf{F}_{i,j}(\hat{m})$, obtained by concatenating the reduced representations $\mathbf{F}_i(\hat{m})$ and $\mathbf{F}_j(\hat{m})$, at scales i and j.

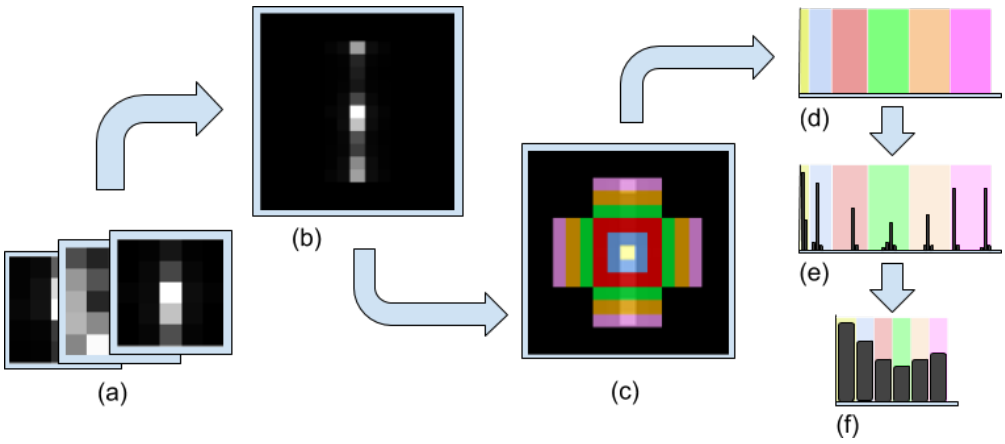
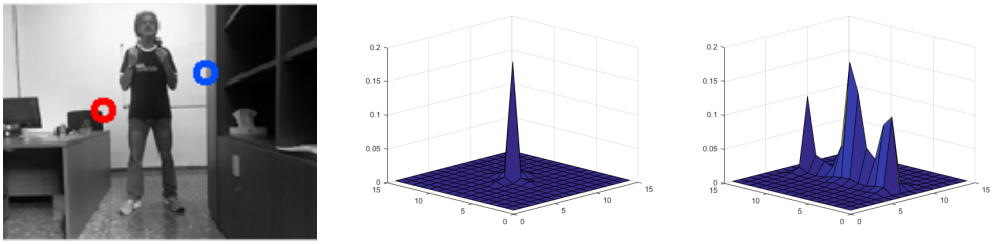


Figure 3. The main steps of the $2D+T$ signal representation procedure: (a) we compute matrices $C_1(r,c)$, $C_2(r,c)$ and $C_3(r,c)$, (b) we create the object $\mathbf{C}$, (c-d) we map subsets of elements (i.e. shearlet coefficients) of $\mathbf{C}$ to different parts of a vector, (e) we obtain the representation $\mathbf{D}(\hat{m})$ for our point, (f) we make the compact descriptor $\mathbf{F}(\hat{m})$.

Figure 4 shows a possible way to visualize the values contained in the matrix $\mathbf{C}$ for two different points, the idea is to view the object $\mathbf{C}$ as a height-map so that to have an insight about the directions in which we found the highest variations (the visualization in Figure 4 (c) is the one corresponding to the object $\mathbf{C}$ shown in Figure 3 (b)).



(a) selected points (b) edge (c) corner

Figure 4. 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.

In the next section we show how this representation can be useful to characterize each point in our signal with respect to its spatio-temporal nature.

4. **Evaluations**

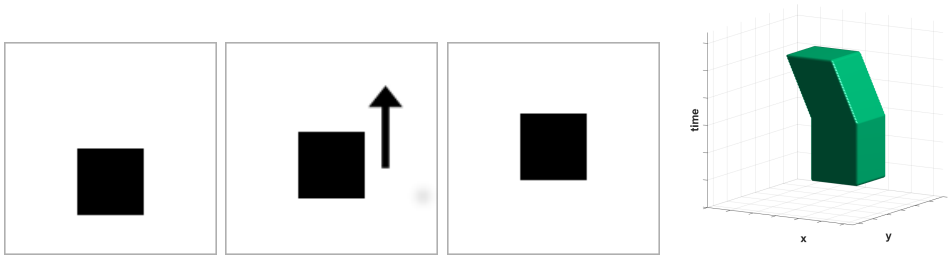
In this section we show the the effectiveness of shearlet coefficients in capturing differences among different space-time discontinuities. We consider a synthetic example and a real world video sequence.

4.1. **Synthetic data.**

The first sequence is a stationary square, which at frame 64 starts to move up with constant speed until frame 108, when the square stops to move. To avoid boundary problems, the sequence is composed of white frames before frame number 20 and after frame number 108, for a total of 128 frames. Figure 5 (a-c) shows a selection of meaningful frames in the synthetic sequence, while Figure 5 (d) shows the 3D shape we obtain by stacking the video frames one on top of the other.

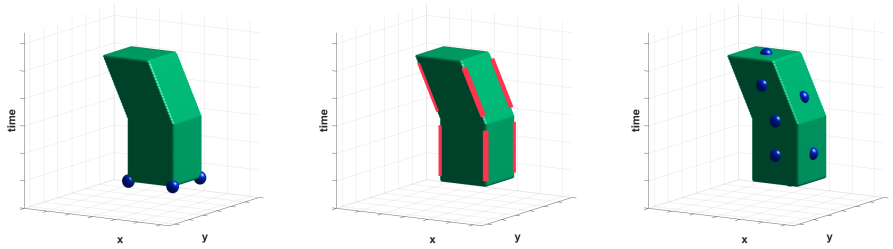
The very simple synthetic sequence contains three spatio-temporal features, which can be easily identified on the 3D shape: 3D corners, edges, and surface points. We test the shearlet-based representation introduced in the previous section on these three classes of points. We sample four 3D corners, points on spatio-temporal edges, and a set of points living on spatio-temporal surfaces. These different sets of elements are highlighted in Figure 6 (a-c), while in Figure 6 (d-f) we show our representations averaged over all the points of each specific class.

These figures show that our representation is very distinctive and easily allows to detect the kind of spatio-temporal features.

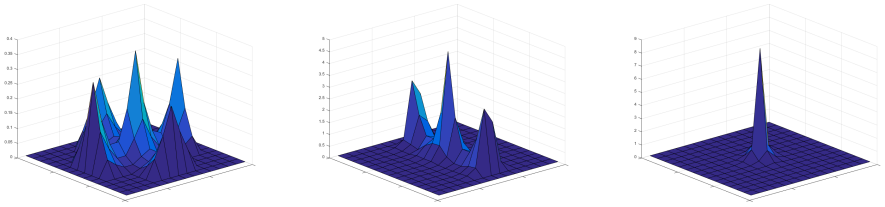


(a) 20-64 (b) 65-107 (c) 108 (d) resulting shape

Figure 5. (a-c) sample frames of the video sequence used to generate the shape taken into account in this section, and (d) the shape resulting from the behavior of the black square within the sequence.



(a) 3D corners (b) edges (c) surface points



(d) (e) (f)

Figure 6. Examples of points on the 3D shape considered (a-c) and corresponding average shearlet-based representation (d-f).

4.2. **Real world data.** We now consider a real video from the KTH dataset [23]. In the video sequence a subject is executing a *boxing* action, repeatedly moving his arms back and forth.
Figure 7 shows three meaningful frames and in (d) the 3D shape obtained by stacking the person's silhouette as the action takes place. The latter allows us to obtain a visual impression of the 3D shape associated with meaningful areas

in the image (the boxing person), although in the following we will consider the entire image, including background information.

(a) frame 30 (b) frame 32 (c) frame 37 (d) 3D shape

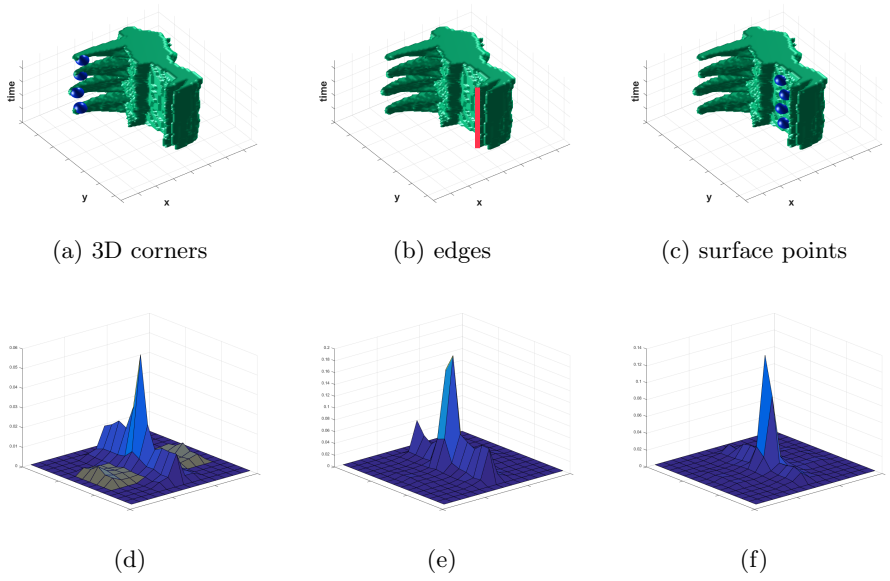
Figure 7. (a-c) sample frames of the *boxing* sequence and (d) corresponding shape generated from the movement.

As in the case of synthetic data, we select points which are associated with different spatio-temporal behavior and, for each of them, we compute our shearlet-based descriptor. The results can be appreciated in Figure 8. We sampled four points located on the red line in Figure 8 (b) to create the corresponding representation in Figure 8 (e), while in the two other cases the points used are only the ones shown in the corresponding pictures on the upper row.

While for surface points the behavior is similar both in the synthetic and in the boxing scenario, things are a little bit different in the two other cases. This is because both spatial and temporal variations in real data are less significant, and the signal discontinuities are not as strong. This can be seen in Figure 8 (d), where the shearlet coefficients corresponding to the changes occurring on the time dimension are less pronounced (these changes are highlighted with the light gray overlay). However, our representation correctly handles the cases in which there is not any temporal change, keeping the corresponding values near to zero (as in Figure 4 (c), where the changes along the temporal dimension contribute for values lower than 10^{-3}).

4.3. **Spatio-temporal Points Classification.** We now try to classify the points belonging to the two $2D + T$ signals we considered in the previous sections. To do so, we carry out two different processes in the two cases:

- for the synthetic data (square sequence) we consider the 3D shape and we classify each point of its surface by calculating the distance between its representation $\mathbf{D}(m)$ and the three average representations in Figure 9, then each point is colored on the basis of the representation it is most



(a) 3D corners (b) edges (c) surface points

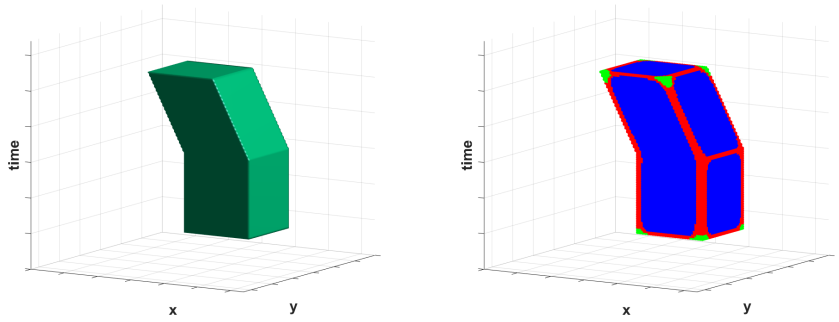
(d) (e) (f)

Figure 8. Examples of points on the 3D boxing shape (a-c) and corresponding average shearlet-based representation (d-f).

similar to.

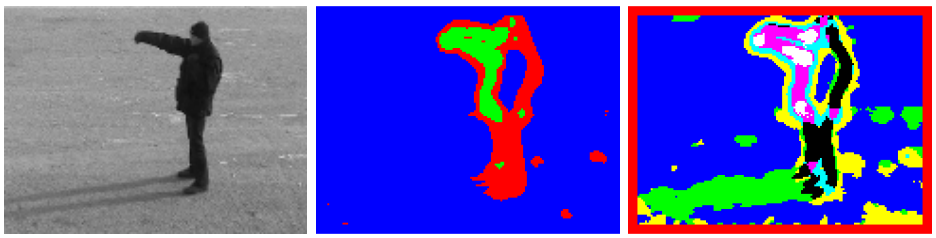
- for the real data (*boxing* sequence), we consider the image sequence and calculate the representation $\mathbf{F}(m)$ for all the points of a single frame at time t; then we cluster them with a *k-means* algorithm. The different colors in Figure 10 (b-c) represent the way the points have been grouped together by using a different number of clusters K. It is possible to see how a larger number of clusters allows us to capture a richer dynamic, characterizing the movement represented in the sequence (by separating the moving parts in different groups, and also segmenting points belonging to shadows), while with a lower number of groups separates points belonging to the still parts of the subject and points belonging to moving elements.

The effectiveness of our representations are shown in Figure 11, where we compare them with the result of clustering the raw shearlet coefficients. In all examples we fix the number of clusters to $k = 8$. By applying *k-means* to the shearlet coefficients we obtain the clusters shown in Figure 11 (a). It can be noticed how points belonging to similar primitives (the arms moving back and forth, and the moving front side of the jacket) are separated in two different groups. Also, points distributed along the back of the subject are not grouped in the same set, even if their spatio-temporal behavior is the same.



(a) shape (b) classification

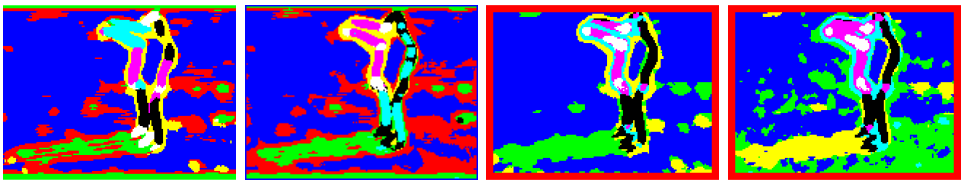
Figure 9. Example of classification of the surface points of our shape: surface points (blue), edges (red) and 3D corners (green).



(a) frame (b) $\mathbf{F}_2(\hat{m})$, k=3 (c) $\mathbf{F}_2(\hat{m})$, k=8

Figure 10. Example of clustering all the points within a fixed frame of a real world sequence; the reduced representation $\mathbf{F}_2(\hat{m})$ is fed to a clustering process considering (b) 3 clusters and (c) 8 clusters (see text for details).

Instead, if we consider our representations built on top of shearlet coefficients, we can see how the clustering process correctly separates points which are associated with different spatio-temporal primitives, by grouping together the elements which are moving in two different sets w.r.t. their spatial appearance (see the white and magenta point sets). Also the points along the back of the subject and belonging to the other straight and still edges are coherently grouped together (this is particularly effective with the reduced representation $\mathbf{F}_2(\hat{m})$, see the black-colored points in Figure 11 (c)). In this last case we have considered only the coefficients belonging to a single scale $j = 2$, if we also consider the ones belonging to a finer one $j = 3$ representing the behavior of the signal at higher frequencies, and we concatenate the two representations we obtain an even more precise separation of all the points of the previously selected frame (see Figure 11(d)).



(a) $SH[f]$ (b) $\mathbf{D}(\hat{m})$ (c) $\mathbf{F}_2(\hat{m})$ (d) $\mathbf{F}_{2,3}(\hat{m})$

Figure 11. Results of a *k-means* clustering executed with k=8 clusters: (a) using the shearlet coefficients as they are provided by the 3D Shearlet Transform, (b) exploiting our representation and considering a single scale, (c) using our compact representation and considering a single scale (d) or concatenating the information coming from two scales.

Figure 12. Frames from a *che vuoi* gesture from the ChaLearn datasets: the male depicted raises bot hands and shakes them, before lowering them again.

Figure 13 shows a similar result from a different sequence drawn from another reference benchmark dataset [23]— significant frames are shown in Figure 12. We can see that, by considering the same number of clusters $k = 8$ our reduced representation $\mathbf{F}(\hat{m})$ produces meaningful groups of points, also highlighting correctly the richer spatio-temporal behavior of the hands of the subject, which in this particular sequence are moving toward the upper part of his body. Instead if we cluster the raw shearlet coefficients, dynamic points are grouped together with background areas (see red-colored points).

Finally, Figures 14 and 15 show the clusters obtained from the $\mathbf{F}_2(\hat{m})$ and $\mathbf{F}_3(\hat{m})$ respectively, for different choices of $k = 4, 8, 12$. In this case the higher resolution of the original image, allows us to appreciate the different structures highlighted at the two different scales. It can also be noticed how the dynamic event of the two hands moving up and down is better captured at the finer scale, as expected.

5. **Conclusion**

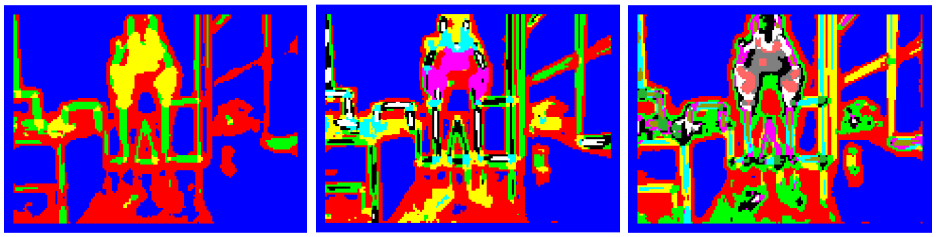
In this paper we considered dynamic video sequences regarded as $2D + T$ signals, and we explored the use of the 3D-shearlet transform with the purpose

(a) sample frame (b) $SH[f]$ (c) $\mathbf{F}_3(\hat{m})$

Figure 13. Results of the clustering process (with k=8) of a frame from the *ChaLearn* dataset [7] and (b) using directly the shearlet coefficients, (c) by means of our representation.

(a) $\mathbf{F}_2(\hat{m})$, k=4 (b) $\mathbf{F}_2(\hat{m})$, k=8 (c) $\mathbf{F}_2(\hat{m})$, k=12

Figure 14. Results of a *k-means* clustering by fixing the scale parameter $j = 2$ and changing the number of clusters created.



(a) $\mathbf{F}_3(\hat{m})$, k=4 (b) $\mathbf{F}_3(\hat{m})$, k=8 (c) $\mathbf{F}_3(\hat{m})$, k=12

Figure 15. Results of a *k-means* clustering by fixing the scale to the finest available and changing the number of clusters created.

of representing the local behavior of the signal and highlight different types of space-time interest points. We derived a procedure to represent the behavior of the $2D + T$ signal in a very compact way. To assess the effectiveness of our representation, we applied an unsupervised clustering algorithm which allowed us to segment a video frame into coherent groups of interest points. We considered both synthetic and real video sequences. In the case of real data, our

representation is also compared with similar results obtained by clustering the raw shearlet coefficients, where we notice very clearly the benefit of the procedure we propose.
The representation appears to be very promising for future applications to video analysis and action recognition.

References

[1] Z. Chen, X. Hao, and Z. Sun. Image denoising in shearlet domain by adaptive thresholding. *Journal of Information & Computational Science*, 10(12):3741–3749, 2013.

[2] S. Dahlke, G. Steidl, and G. Teschke. The continuous shearlet transform in arbitrary space dimensions. *J. Fourier Anal. Appl.*, 16(3):340–364, 2010.

[3] M. N. Do and M. Vetterli. The contourlet transform: An efficient directional multiresolution image representation. *Trans. Img. Proc.*, pages 2091–2106, 2005.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72. IEEE, 2005.

[5] M. A. Duval-Poo, F. Odone, and E. De Vito. Edges and corners with shearlets. *IEEE Trans. Image Processing*, 24(11):3768–3780, 2015.

[6] G. R. Easley, D. Labate, and F. Colonna. Shearlet-based total variation diffusion for denoising. *IEEE Transactions on Image processing*, 18(2):260–268, 2009.

[7] S. Escalera, X. Baro, J. Gonzlez, M. Bautista, M. Madadi, M. Reyes, V. Ponce-Lpez, H. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *Workshop at the European Conference on Computer Vision*, pages 459–473. Springer, 2014.

[8] K. Guo and D. Labate. Optimally sparse multidimensional representation using shearlets. *SIAM journal on mathematical analysis*, 39(1):298–318, 2007.

[9] K. Guo and D. Labate. Analysis and detection of surface discontinuities using the 3D continuous shearlet transform. *Appl. Comput. Harmon. Anal.*, 30(2):231–242, 2011.

[10] K. Guo and D. Labate. Optimally sparse representations of 3D data with C^2 surface singularities using Parseval frames of shearlets. *SIAM J. Math. Anal.*, pages 851–886, 2012.

[11] K. Guo, D. Labate, and W.-Q. Lim. Edge analysis and identification using the continuous shearlet transform. *Applied and Computational Harmonic Analysis*, 27(1):24–46, 2009.

[12] G. Kutyniok and D. Labate. Resolution of the wavefront set using continuous shearlets. *Transactions of the American Mathematical Society*, 361(5):2719–2754, 2009.

[13] G. Kutyniok and D. Labate. *Shearlets*. Appl. Numer. Harmon. Anal. Birkhäuser/Springer, New York, 2012.
[14] G. Kutyniok, J. Lemvig, and W. Lim. Optimally sparse approximations of 3D functions by compactly supported shearlet frames. *SIAM J. Math. Anal.*, 44(4):2962–3017, 2012.
[15] G. Kutyniok, W. Lim, and R. Reisenhofer. Shearlab 3D: Faithful digital shearlet transforms based on compactly supported shearlets. *ACM Transactions on Mathematical Software*, 42(1):5, 2016.
[16] G. Kutyniok and W.-Q. Lim. Compactly supported shearlets are optimally sparse. *Journal of Approximation Theory*, 163(11):1564–1589, 2011.
[17] G. Kutyniok and P. Petersen. Classification of edges using compactly supported shearlets. *Applied and Computational Harmonic Analysis*, 2015.
[18] D. Labate, W.-Q. Lim, G. Kutyniok, and G. Weiss. Sparse multidimensional representation using shearlets. In *Optics & Photonics 2005*, 2005.
[19] I. Laptev. On space-time interest points. *Int. J. Computer Vision*, 64(2):107–123, 2005.
[20] T. Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994.
[21] T. Lindeberg. Time-causal and time-recursive spatio-temporal receptive fields. *Journal of Mathematical Imaging and Vision*, 55(1):50–88, 2016.
[22] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 710–732, 1992.
[23] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

ACKNOWLEDGEMENT

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