Local Spatio-Temporal Representation using the 3D Shearlet Transform

Damiano Malafronte
Università degli Studi di Genova, Italy
Email: damiano.malafronte@dibris.unige.it

Francesca Odone
Università degli Studi di Genova, Italy
Email: francesca.odone@unige.it

Ernesto De Vito
Università degli Studi di Genova, Italy
Email: devito@dima.unige.it

***Abstract*—In this work we address the problem of analyzing video sequences and representing meaningful space-time points of interest. We base our work on the 3D shearlet transform. In particular, we exploit the relation between coefficients with similar shearings to build a local representation which turns out to be really informative to understand the local spatio-temporal characteristics of the points that we are considering.**

I. 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 [1], see for instance [2, 3, 4].

In the meanwhile, many multi-scale methods have been introduced to deal with multi-dimensional signals. Among them, shearlets [5] emerge by their ability to efficiently capture anisotropic features [6], to provide an optimal sparse representation [7, 8], to detect singularities [9, 10] and to be stable against noise and blurring [11, 12]. For for further details, implementations, and references see [13] and the website http://www.shearlab.org/.

The effectiveness of shearlets is supported by a well-established mathematical theory [14] and it is tested in many applications in image processing by providing efficient algorithms [13, 15].

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.

II. SHEARLET THEORY: AN OVERVIEW

Shearlets in arbitrary space dimensions were first introduced in [14] 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 [16] 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

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

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

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

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

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

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

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

978-1-5386-1565-2/17/$31.00

where P is a given polynomial 2D fan filter [17], 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 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 $\ell = 3$).

Our algorithm is based on the following nice property of the shearlet coefficients. As shown in [18, 19, 20] 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). This results allows 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

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

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[1] (Figure 1). 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

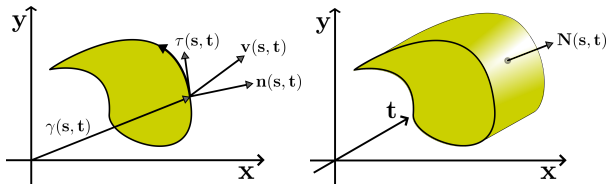
$$\Delta N(s, t_0) = \tau(s, t_0) \wedge \Delta v(s, t_0) \qquad \forall s \in [0, 1],$$

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 [16] 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 [21].



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

Fig. 1: A cartoon-like object with the main relevant geometrical and dynamical quantities (see text for details).

III. 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.

We consider a point $\hat{m}$ for the fixed scale $\hat{j}$ and the subset of shearings

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

The procedure we carry out in the discrete case is depicted in Figure 2 and consists of two parts, which we describe in the following.

1 - Reorganize the coefficients of a point neighborhood.

(a) We reorganize the information provided by $SH[f](\ell, \hat{j}, k, \hat{m})$ in three $M \times M$ matrices, each one 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 we introduce r and c, which are the discretized correspondents of the indexes k_1 and k_2 used in the continuous case. Figure 2 (a) shows the three matrices for a specific space-time point.

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

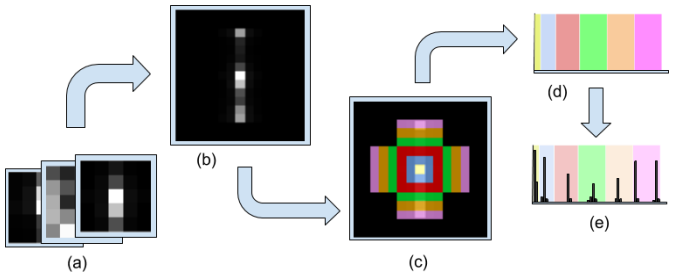


Fig. 2: The main steps of the $2D + T$ signal representation procedure: (a) we compute matrices $C_1(r, c)$, $C_2(r, c)$ and $C_3(r, c)$, (b) we create the object $\mathbf{C}$, (c-d) we map subsets of elements (i.e. shearlet coefficients) of $\mathbf{C}$ to different parts of a vector and (e) we obtain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

(b) We merge the three matrices in a single one, by recombining them relatively to the maximum shearlet coefficient (Figure 2 (b)). The obtained overall representation $\mathbf{C}$ 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 $\ell \in \{1, 2, 3\}$ and $k \in \mathbf{K}$. The matrix $\mathbf{C}$ 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 the distance of every entry of $\mathbf{C}$ with respect to the center is proportional to the distance of the corresponding angles (as defined in (4)) from the angles associated with k_{max}. 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 3).

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 *i-th* ring of values from k_{max} in $\mathbf{C}$ (as highlighted Figure 2 (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 2 (d) and (e)).

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

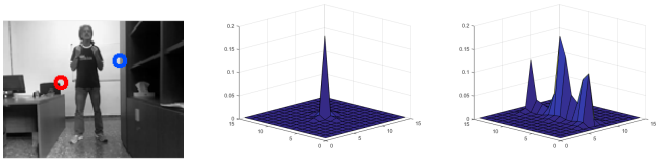
$$\mathbf{D}(\hat{m}) = \mathit{coeff}_{\bar{s}_0} \frown \mathit{coeff}_{\bar{s}_1} \frown \mathit{coeff}_{\bar{s}_2} \frown \dots;$$

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 $k_{\bar{s}_i}$.

(c) Finally we obtain the representation $\mathbf{D}(\hat{m})$ for point $\hat{m}$. The size of the descriptor is strictly dependent on the number of shearings that we are considering within our shearlet system.



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

Fig. 3: 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.

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 $k \neq s_{max}$. Figure 3 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 3 (c) is the one corresponding to the object $\mathbf{C}$ shown in Figure 2 (b)).

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.

IV. EVALUATIONS

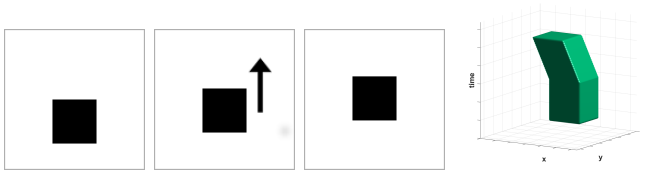
In this section we show the the effectiveness of shearlet coefficients in capturing differences among different space-time discontinuities. We consider a synthetic example and a real world video sequence.

A. 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. Figure 4 (a-c) shows a selection of meaningful frames in the synthetic sequence, while Figure 4 (d) shows the 3D shape we obtain by stacking the video frames one on top of the other.

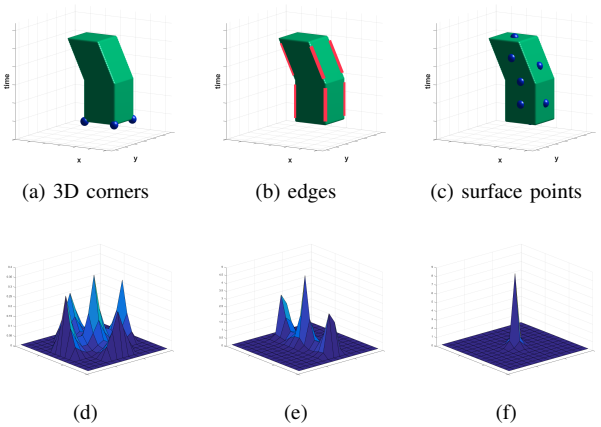
The very simple synthetic sequence contains three spatio-temporal features, which can be easily identified on the 3D shape: 3D corners, edges, and surface points. We test the shearlet-based representation introduced in the previous section on these three classes of points. These elements are highlighted in Figure 5 (a-c), while in Figure 5 (d-f) we show our representations averaged over all the points of a 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

Fig. 4: (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)

Fig. 5: Examples of points on the 3D shape considered (a-c) and corresponding average shearlet-based representation (d-f).

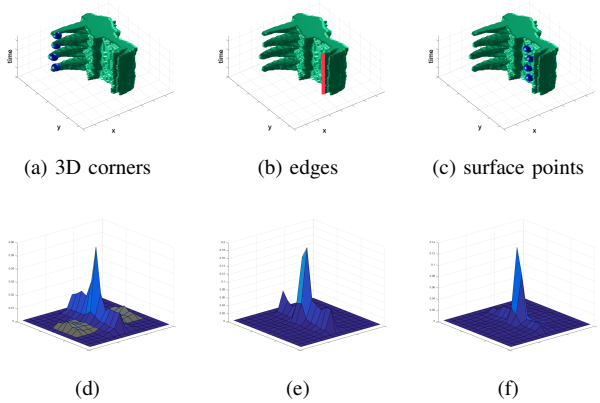
B. Real world data

We now consider a real video from the KTH dataset [22]. In the video sequence a subject is executing a *boxing* action, repeatedly moving his arms back and forth.
Figure 6 shows three meaningful frames and in (d) the 3D shape obtained by stacking the person's silhouette as the action takes place.

(a) frame 13 (b) frame 33 (c) frame 37 (d) resulting shape

Fig. 6: (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 to different spatio-temporal behavior and, for each of them, we compute our shearlet-based descriptor. The results can be appreciated in Figure 7, this time we sampled four points located on the red line in Figure 7 (b) to create the corresponding representation in Figure 7 (e), while in the two other cases the points used are only the ones shown in the corresponding pictures on the upper line.



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

(d) (e) (f)

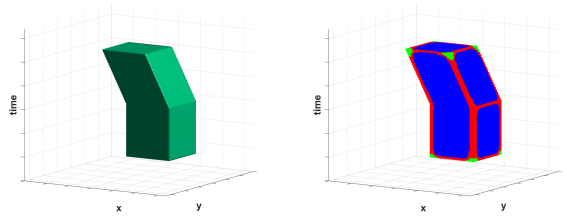
Fig. 7: Examples of points on the 3D boxing shape (a-c) and corresponding average shearlet-based representation (d-f).

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 7 (d), where the shearlet coefficients corresponding to the changes occurring on the time dimension are less pronounced (these changes are highlighted with the yellow 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 3 (c), where the changes along the temporal dimension contribute for values lower than 10^{-3}).

C. Spatio-temporal Points Classification

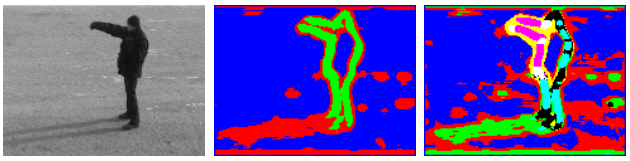
Finally we try to classify the points belonging to the two $2D+T$ signals we considered in the previous sections. To do so, we carry on two different processes in the two cases:

- for the synthetic shape, we classify each point of its surface by calculating the distance between its representation $\mathbf{D}(m)$ and the three average representations in Figure 8, then each point is colored on the basis of the representation it is most similar to.
- for the *boxing* sequence, we calculate the representation $\mathbf{D}(m)$ for all the points of f by fixing t (thus, considering a single frame within the whole sequence), then we cluster them with a K-means algorithm. The different colors in Figure 9 (b-c) represent the way all the points have been grouped together by using a different number of clusters K. It is possible to see how a greater number of clusters allows to capture a richer dynamic characterizing the movement represented in the sequence, while with a lower number of groups we can just separate points with a very low dynamic (the background, the shadow, or the inner part of the subject's body) from the stronger edges contained in the chosen frame.



(a) shape (b) classification

Fig. 8: Example of classification of the surface points of our shape: surface points (blue), edges (red) and 3D corners (green).



(a) frame (b) 3 clusters (c) 8 clusters

Fig. 9: Example of clustering of the all the points within a fixed frame of our real world sequence, we show the results (b) using 3 clusters and (c) using 8 clusters (see text for details).

V. CONCLUSION

In this paper we considered $2D + T$ signals and explored the use of 3D shearlets with the purpose of representing space-time interest points. Our analysis shows a potential for this type of representation and encourages us to address, in current work, space-time feature detection and, in perspective, action recognition. The obtained results speak in favor of the descriptor effectiveness.

REFERENCES

[1] T. Lindeberg, "Scale-space theory: A basic tool for analyzing structures at different scales," *Journal of applied statistics*, vol. 21, no. 1-2, pp. 225–270, 1994.

[2] I. Laptev, "On space-time interest points," *Int. J. Computer Vision*, vol. 64, no. 2, pp. 107–123, 2005.

[3] 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*. IEEE, 2005, pp. 65–72.

[4] T. Lindeberg, "Time-causal and time-recursive spatio-temporal receptive fields," *Journal of Mathematical Imaging and Vision*, vol. 55, no. 1, pp. 50–88, 2016.

[5] D. Labate, W.-Q. Lim, G. Kutyniok, and G. Weiss, "Sparse multidimensional representation using shearlets," in *Optics & Photonics 2005*, 2005.

[6] G. Kutyniok and D. Labate, "Resolution of the wavefront set using continuous shearlets," *Transactions of the American Mathematical Society*, vol. 361, no. 5, pp. 2719–2754, 2009.

[7] K. Guo and D. Labate, "Optimally sparse multidimensional representation using shearlets," *SIAM journal on mathematical analysis*, vol. 39, no. 1, pp. 298–318, 2007.

[8] G. Kutyniok and W.-Q. Lim, "Compactly supported shearlets are optimally sparse," *Journal of Approximation Theory*, vol. 163, no. 11, pp. 1564–1589, 2011.

[9] K. Guo, D. Labate, and W.-Q. Lim, "Edge analysis and identification using the continuous shearlet transform," *Applied and Computational Harmonic Analysis*, vol. 27, no. 1, pp. 24–46, 2009.

[10] G. Kutyniok and P. Petersen, "Classification of edges using compactly supported shearlets," *Applied and Computational Harmonic Analysis*, 2015.

[11] G. R. Easley, D. Labate, and F. Colonna, "Shearlet-based total variation diffusion for denoising," *IEEE Transactions on Image processing*, vol. 18, no. 2, pp. 260–268, 2009.

[12] Z. Chen, X. Hao, and Z. Sun, "Image denoising in shearlet domain by adaptive thresholding," *Journal of Information & Computational Science*, vol. 10, no. 12, pp. 3741–3749, 2013.

[13] G. Kutyniok and D. Labate, *Shearlets*, ser. Appl. Numer. Harmon. Anal. Birkhäuser/Springer, New York, 2012.

[14] S. Dahlke, G. Steidl, and G. Teschke, "The continuous shearlet transform in arbitrary space dimensions," *J. Fourier Anal. Appl.*, vol. 16, no. 3, pp. 340–364, 2010.

[15] M. A. Duval-Poo, F. Odone, and E. De Vito, "Edges and corners with shearlets," *IEEE Trans. Image Processing*, vol. 24, no. 11, pp. 3768–3780, 2015.

[16] G. Kutyniok, W. Lim, and R. Reisenhofer, "Shearlab 3D: Faithful digital shearlet transforms based on compactly supported shearlets," *ACM Transactions on Mathematical Software*, vol. 42, no. 1, p. 5, 2016.

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

[18] K. Guo and D. Labate, "Analysis and detection of surface discontinuities using the 3D continuous shearlet transform," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 2, pp. 231–242, 2011.

[19] ——, "Optimally sparse representations of 3D data with C^2 surface singularities using Parseval frames of shearlets," *SIAM J. Math. Anal.*, pp. 851–886, 2012.

[20] G. Kutyniok, J. Lemvig, and W. Lim, "Optimally sparse approximations of 3D functions by compactly supported shearlet frames," *SIAM J. Math. Anal.*, vol. 44, no. 4, pp. 2962–3017, 2012.

[21] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 710–732, 1992.

[22] 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*, vol. 3. IEEE, 2004, pp. 32–36.