Detecting Spatio-temporally Interest Points using the Shearlet Transform

Damiano Malafronte, Francesca Odone, Ernesto De Vito

Università degli Studi di Genova, Genova GE, IT,
damiano.malafronte@dibris.unige.it,
francesca.odone@unige.it, devito@dima.unige.it

Abstract. In this paper we address the problem of detecting spatio-temporal interest points in video sequences and we introduce a novel detection algorithm based on the three-dimensional shearlet transform. By evaluating our method on different application scenarios, we show we are able to extract meaningful spatio-temporal features from video sequences of human movements, including full body movements selected from benchmark datasets of human actions and human-machine interaction sequences where the goal is to segment drawing activities in smaller action primitives.

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

1 Introduction

The analysis of dynamic events by space-time interest point detection has been addressed for over a decade primarily in the field of action recognition. The concept of space-time local features has been first formulated in [14], where the connections with scale-space have also been highlighted, and application to action recognition proposed [19]. Applications to action and behavior recognition can be also found in [3, 22] where the authors also propose significant improvements on the computational cost and the overall descriptiveness of the procedure. An evaluation of the different approaches has been presented in [21] where the authors provide an analysis which is based on the joint evaluation of detectors and descriptors. A different approach has been taken in [10] and later in [20] with the goal of learning instead than engineering space-time features and descriptors.

In this work we discuss a work in progress where we propose the adoption of a well founded theoretical framework, the 3D shearlet transform [1], as a starting point for feature detection on a 2D+T signal. It is known that wavelets provide an optimal multi-scale representation only for 1D-signals. Shearlets are a multi-scale system of filters encoding directional informations and extending the main properties of wavelets to multivariate functions, as for example:

a) optimal sparse representation for functions with singularities along curves and surfaces;
b) both a continuous and a discrete representation with a well established theory;

c) shearlets with compact support either in the space domain or in the frequency domain;
d) a large class of mother shearlets devised for specific applications (as denoising, inpainting, edge/corner detection to name a few);
e) many discrete fast implementations with freely available codes.

We refer to the book [11] and references therein for details and updated bibliography. In the recent past shearlets have been applied with success to multi-scale feature detection in images [23, 5]. The extension to the 3D case [1, 6, 7, 12] makes it possible to address shape/volume analysis and video analysis problems. Regarding volume analysis the shearlets have been applied to the reconstruction and the analysis of medical images volumetric medical imaging [8, 4]; as for video analysis it is worth mentioning an application to video denoising and inpainting [17] and to video saliency analysis [15].

Specifically, we propose a feature detection algorithm based on the Shearlab3D implementation of three dimensional shearlets [13]. Our pipeline is very simple and the whole procedure is embedded in the computation of an interest measure derived from the good properties of the shearlet coefficients, which allows us to enhance local discontinuities in space-time. Figure 1 shows the main steps of the algorithm on an example frame.

(a) frame (b) interest measure (c) candidate points (d) refined points

Fig. 1: A summary of the detection pipeline we propose. (a) a frame I_t from the original video (from ChaLearn dataset); (b) interest measure IM derived from 3D shearlet coefficients enhancing interesting elements on I_t; (c) candidate local features surviving a non-maxima suppression on a space-time local neighborhood; (d) the detected meaningful points obtained by hard thresholding.

We discuss the quality of the detected points in the context of a variety of possible different applications, including action classification, salient frames extraction, and the detection of view-invariant keypoints in human gestures performed in a human-machine interaction (HMI) setting. The preliminary results speak in favor of the accuracy and meaning of the detected points.

The remainder of the paper is organized as follows. In Section 2 we review the theory of 2D+T shearlets, in Section 3 we describe our space-time interest point detection algorithm. In Section 4 we report an analysis of the detection results on a set of different possible scenarios. Finally, Section 5 is left to the conclusions.

2 Shearlets: an overview

Following [13] we briefly review the construction of the shearlet transform of a $2D+T$ signal f, first introduced in [1]. 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 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 do not depend on j and k and are given by

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

where $c > 0$ is a step size, φ is a $1D$-scaling function and 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 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. Indeed

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

where c is as in (1) and $\widehat{c}$ is another step size (in the rest of the paper we assume that $c = \hat{c} = 1$ for sake of simplicity), and the mother shearlet is

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

where P is a given polynomial 2D Fan filter [2], 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

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

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 detection algorithm is based on the following nice property of the shearlet coefficients. As shown in [6, 7, 12] 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). Based on the above result, we can uniquely associate a direction (without orientation) to any shearing vector $k = (k_1, k_2)$ and conversely. The correspondence explicitly depends on ℓ (due to the discretization the number of shearings increases with the scale, so that there is a formal dependence on j).

To compute the shearlet coefficients we use the digital implementation described in [13] based on the relation between the pair scaling function/wavelet (φ, ψ) and and the quadrature mirror filter pair (h, g), which in our application is the filter pair introduced in [16].

3 Feature detection method

In this section we show how to detect spatio-temporally interest point only by means of the information provided by the shearlet transform $SH[f]$ where f is an image sequence $f(x, y, t)$.

As recalled at the end of the previous section, points that belong to a surface singularity are characterized by a slow decay of the corresponding shearlet coefficients, as the scale parameter j grows and the shearing parameter k (and the pyramid label ℓ) corresponds to the the normal vector to the surface. A similar behavior holds true for singularities along the boundary of the surface, where two or more shearings can be meaningful [9]. Hence, we expect that the points of interest of a video are associated with high values of the shearlet coefficients and different spatial/temporal features can be extracted by looking to different pyramid labels ℓ.

These observations suggest to extend to video signals the edge detector introduced in [16] for wavelets and in [23, 5] for shearlets.

To this purpose, we first define an *interest measure* IM representing a response function calculated for each point $m = (x, y, t)$ in our signal. At a fixed scale j:

$$IM_j[f](m) = \prod_{\ell=1}^{3} \sum_{k=(k_1,k_2)} | SH[f](\ell, j, k, m) |$$

Our detection algorithm is based on the use of the measure IM as a feature enhancement process. The space-time feature detection procedure is summarized in four steps, shown in an example in Figure 1 :

a) We compute $IM_j[f]$ for $j = 1, 2$ — we control the computational cost of the procedure, by limiting the number of scales. We skip the scale $j = 0$ as it does not enhance properly the meaningful information in the signal.
b) Then, we define an overall interest measure by multiplying the values calculated for $IM_j[f]$ when we consider only the two finest scales $j = 1, 2$ namely

$$IM[f](m) = IM_2[f](m) \cdot IM_1[f](m)$$

Since we only have three scales, the analysis across scales in [16] is not meaningful. We observed that the points of interest produce high values in both the scales $j = 1, 2$ and this remark is at the root of the above definition.
c) We perform a non-maxima suppression in a spatio-temporal window N_m of size $w \times w \times w$ by setting to 0 non-maxima coefficients.
d) Finally, we detect meaningful points m on the signal by means of a thresholding step $IM[f](m) > \tau$.

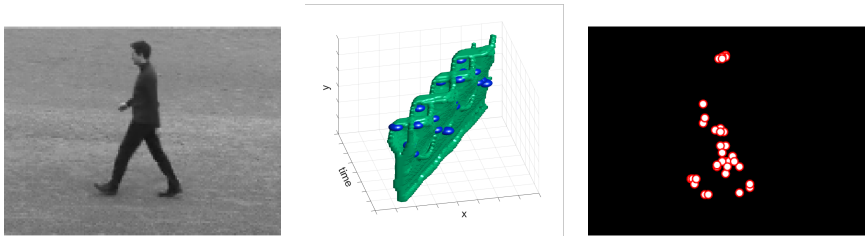
Notice that the IM measure is shown for a fixed t, then it includes values that appear to be high relative to all values in t (e.g. the areas corresponding to the elbows). Those points do not appear to survive the non-maxima suppression procedure (they are not highlighted in figure (c)) as they are not maxima with respect to the temporal direction (they will be marked as candidates in some neighboring time instant).

4 Evaluation

In this section we discuss the potential of our approach to feature detection on a variety of different applications. In what follows the neighbourhood size w is set to 9, and threshold τ is chosen on an appropriate validation set.

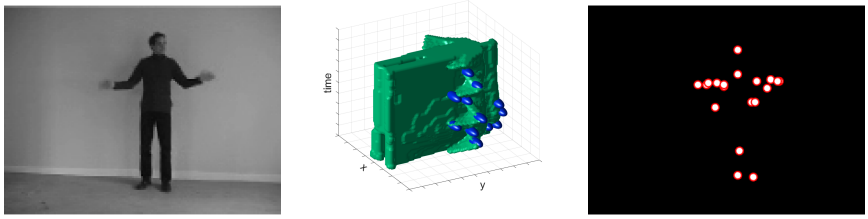
Detecting features in action sequences. We start by showing examples of the extracted features in human action sequences. Figure 2 shows the results on a walking person, in two different visualization modalities: a 3D shape of the person silhouette evolving in time where the detected features are marked in blue; a map where the positions of detected points across the whole sequence are merged. It can be noticed how all meaningful points (in particular all points corresponding to a change in direction of the foot) have been nicely detected. Similarly, Figure 3 shows an example of a different human action, a handwaving, where most features are detected on the tip of the hands.

Salient frames extraction. The space-time interest points we are detecting correspond to a "special" point in the scene (a corner) as it is undergoing some significant velocity change. The presence of these points is a cue of some interesting movement going on in the sequence. Their presence can be used as a guideline on the importance of a given frame in a video summarization process.



(a) (b) (c)

Fig. 2: A *walking* action (a) observed as (b) feature detection on a 2D+T surface (where we flipped the surface upside down to better show the points detected) and (c) summarized on a reference time instant (detected features are translated w.r.t the body centroid).



(a) (b) (c)

Fig. 3: A *handwaving* action (a) observed as (b) feature detection on a 2D+T surface (in this case there is a subset of features which are not visible, the ones lying within the surface and corresponding to the "claps") and (c) summarized on a reference time instant.

We evaluate the number of space-time interest points detected in each frame and select the most meaningful frames as the one containing a large number of those points. While doing so we also apply a non maxima suppression on the temporal neighborhood to avoid the selection of frames too close in time. Figure 4 shows examples of the number of detected interest points across time in two sequences we considered, the *walking* (from KTH) and the *che vuoi* (from ChaLearn - italian lexicon) ones.

For the sake of the experiment, we select three frames in the sequences with the highest number of points detected. Figure 5 shows the most meaningful frames of a walking sequence, corresponding to the beginning of a new stride in the walk executed in the sequence. Figure 6 shows the three most meaningful frames of the *cosa vuoi* sequence, where a male subject is executing a gesture in which he raises both his hands, shakes them, and then moves them back in

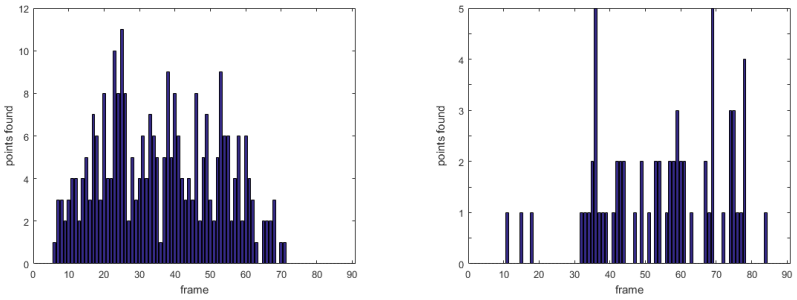


Fig. 4: Distribution of interest points found over time (a) in the *walking* and (b) in the *che vuoi* sequences.

the starting position. Similarly to the previous case, the three frames identified highlight very peculiar elements of the acquired action.

Fig. 5: Salient frames selected for the *walking* sequence (KTH) .

Fig. 6: Salient frames selected for the *che vuoi* gesture sequence (ChaLearn).

Detecting gesture primitives in HMI. We conclude with a reference to a human-machine interaction (HMI) problem. An artificial agent is observing a human performing a set of predefined planar activities (drawing different shapes). Each activity must be divided into smaller action primitives, similarly to [18].

Figures 7 and 8 show candidate frames corresponding to extrema of action primitives (where the hand features are undergoing a major velocity change): the former shows the results on a sequence of repeated line drawing actions performed on a frontal transparent surface (artificial agent view), the latter the crucial points of the action of drawing a rectangle on a table (human view). In both cases the points where the pen is changing direction have been detected.

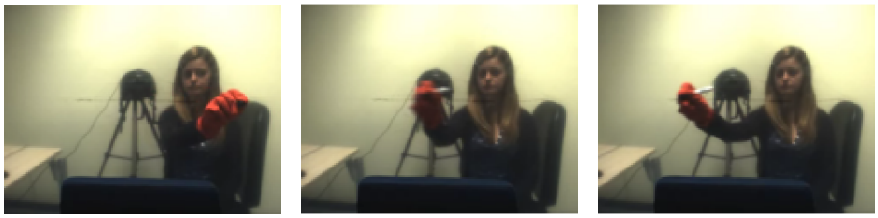


Fig. 7: Frames corresponding to a change in action primitive on the *drawing line* sequence.

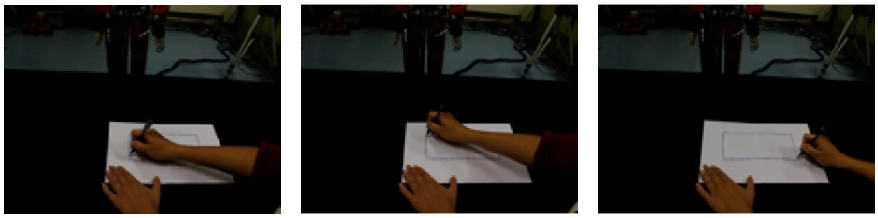


Fig. 8: Frames corresponding to a change in action primitive on the *drawing rectangle* sequence.

5 Conclusions

In this paper we presented a space-time interest point detector based on 3D shearlets. The method is grounded on a sound mathematical theory and appears to be very promising for different video analysis applications. We are currently working on developing a complete video analysis pipeline (including noise removal, detection, and description of local features) entirely based on the same transformation, where we are aiming at exploiting at best its multiscale and multidirectional properties.

Acknowledgements. The authors would like to thank Alessia Vignolo for providing the drawing data used in the experiments.

Bibliography

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

[2] M. N. Do and M. Vetterli. The contourlet transform: An efficient directional multiresolution image representation. *Trans. Img. Proc.*, pages 2091–2106, 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*, pages 65–72. IEEE, 2005.

[4] C. Duan, S. Wang, X.G. Wang, and Q.H. Huang. MRI volume fusion based on 3D shearlet decompositions. *Journal of Biomedical Imaging*, 2014:4, 2014.

[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] 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.

[7] 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.*, (2):851–886, 2012.

[8] K. Guo and D. Labate. Optimal recovery of 3d x-ray tomographic data via shearlet decomposition. *Advances in Computational Mathematics*, 39 (2):227–255, 2013.

[9] R. Houska and D. Labate. Detection of boundary curves on the piecewise smooth boundary surface of three dimensional solids. *Appl. Comput. Harmon. Anal.*, 40(1):137–171, 2016.

[10] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

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

[12] G. Kutyniok, J. Lemvig, and W.Q. Lim. Optimally sparse approximations of 3D functions by compactly supported shearlet frames. *SIAM J. Math. Anal.*, 44(4):2962–3017, 2012.

[13] G. Kutyniok, W.Q. Lim, and R. Reisenhofer. Shearlab 3D: Faithful digital shearlet transforms based on compactly supported shearlets. *ACM Transactions on Mathematical Software*, 42(1):5, 2016.

[14] I. Laptev. On space-time interest points. *Int. J. Computer Vision*, 64(2): 107–123, 2005.

[15] B. Lei, Z. Xiongwei, Z. Yunfei, and L. Yang. Video saliency detection using 3D shearlet transform. *Multimedia Tools Appl.*, 75(13):7761–7778, 2016.

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

[17] P. S. Negi and D. Labate. 3D discrete shearlet transform and video processing. *IEEE Trans. Image Processing*, 21:2944–2954, 2012.

[18] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226, 2002.

[19] 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.

[20] G.W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *Proceedings of the 11th European Conference on Computer Vision: Part VI*, 2010.

[21] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference*, pages 124–1. BMVA Press, 2009.

[22] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European conference on computer vision*, pages 650–663. Springer, 2008.

[23] S. Yi, D. Labate, G.R. Easley, and H. Krim. A shearlet approach to edge analysis and detection. *IEEE Trans. Image Process.*, 18(5):929–941, 2009.