Dipartimento di Informatica, Bioingegneria,
Robotica ed Ingegneria dei Sistemi

Spatio-Temporal Video Analysis and the 3D Shearlet Transform

by

Damiano Malafronte

DIBRIS, Università di Genova
Via Opera Pia, 13 16145 Genova, Italy http://www.dibris.unige.it/

Università degli Studi di Genova

Dipartimento di Informatica, Bioingegneria,

Robotica ed Ingegneria dei Sistemi

Ph.D. Thesis in Computer Science and Systems Engineering
Computer Science Curriculum

Spatio-Temporal Video Analysis and the 3D Shearlet Transform

by

Damiano Malafronte

May, 2018

Dottorato di Ricerca in Informatica ed Ingegneria dei Sistemi
Indirizzo Informatica
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università degli Studi di Genova

DIBRIS, Univ. di Genova
Via Opera Pia, 13
I-16145 Genova, Italy
http://www.dibris.unige.it/

Ph.D. Thesis in Computer Science and Systems Engineering
Computer Science Curriculum
(S.S.D. INF/01)

Submitted by Damiano Malafronte
DIBRIS, Univ. di Genova

Date of submission: February 2018

Title: Spatio-Temporal Video Analysis and the 3D Shearlet Transform

Advisors:

Francesca Odone
Dipartimento di Informatica, Bioingegneria, Robotica ed Ingegneria dei Sistemi
Università di Genova
francesca.odone@unige.it

Ernesto De Vito
Dipartimento di Matematica
Università di Genova
devito@dima.unige.it

Ext. Reviewers:

Gerd Teschke
Neubrandeburg University of Applied Science
teschke@zib.de

Raffaella Lanzarotti
Department of Computer Science
Università degli Studi di Milano
lanzarotti@di.unimi.it

Abstract

The automatic analysis of the content of a video sequence has captured the attention of the computer vision community for a very long time. Indeed, video understanding, which needs to incorporate both semantic and dynamic cues, may be trivial for humans, but it turned out to be a very complex task for a machine. Over the years the signal processing, computer vision, and machine learning communities contributed with algorithms that are today effective building blocks of more and more complex systems. In the meanwhile, theoretical analysis has gained a better understanding of this multifaceted type of data. Indeed, video sequences are not only high dimensional data, but they are also very peculiar, as they include spatial as well as temporal information which should be treated differently, but are both important to the overall process. The work of this thesis builds a new bridge between signal processing theory, and computer vision applications. It considers a novel approach to multi resolution signal processing, the so-called Shearlet Transform, as a reference framework for representing meaningful space-time local information in a video signal. The Shearlet Transform has been shown effective in analyzing multi-dimensional signals, ranging from images to x-ray tomographic data. As a tool for signal denoising, has also been applied to video data. However, to the best of our knowledge, the Shearlet Transform has never been employed to design video analysis algorithms. In this thesis, our broad objective is to explore the capabilities of the Shearlet Transform to extract information from 2D+T-dimensional data. We exploit the properties of the Shearlet decomposition to redesign a variety of classical video processing techniques (including space-time interest point detection and normal flow estimation) and to develop novel methods to better understand the local behavior of video sequences. We provide experimental evidence on the potential of our approach on synthetic as well as real data drawn from publicly available benchmark datasets. The results we obtain show the potential of our approach and encourages further investigations in the near future.

Contents

Introduction

The Shearlet Transform [LLKG05, KL12, DDMGL15] is a multiresolution analysis framework possessing several properties which make it suitable for the analysis of multidimensional signals. Amongst these, we recall its ability to characterize anisotropic structures, together with a straightforward way to capture and encode signal singularities. These properties have been exploited by various authors in the image processing domain, see for instance [YLEK08, GLL09, YLEK09a, KL10, EL12, DODV15, DPNODV17], where the Shearlet Transform has been successfully employed to enhance and detect different low-level singularities within 2D digital images.

In this thesis we explore the possibility of using the Shearlet Transform in a video analysis scenario. Part of the results that we present within this document can be found in [MODV17a, MODV17b, MGV$^+$17, MODVar]. We tackle a variety of computer vision tasks, all related to the local analysis of 2D+T-dimensional signals. From the theoretical standpoint, the Shearlet framework is very appropriate for the analysis of a signal in a local neighborhood at different scales. Moreover, our work shows that it is possible to extract very different types of information from a video sequence by exploiting the result of a single decomposition of the signal, namely the Shearlet coefficients.

Within this general goal, the specific objectives of this thesis are:

- to assess the capabilities of the 3D Shearlet Transform as a **tool to analyze 2D+T signals**, in terms of its ability to describe the local spatio-temporal neighborhood of a space-time point.
- to model the **spatio-temporal local structures** which may arise in the space-time domain, through the development of a simple taxonomy which may inspire future research.

- to understand the amount of **information carried by the Shearlet coefficients** calculated on a video sequence, so to develop novel Shearlet-based approaches to carry out video analysis tasks.

State of the art

The rise of Shearlets traces its roots back to the early 2000s, when the use of wavelets in signal analysis and computer vision had proved almost optimal for one-dimensional signals in many ways, and the mathematics behind classical wavelets had reached a high degree of elaboration. While considering signals in dimension two and above, wavelet systems act worse in encoding and characterizing anisotropic singularities, over the years this situation has led to the development of a new class of representations. Among these, a few representatives of such approaches are directional wavelets [AM96], ridgelets [CD99], curvelets [CD04], wavelets with composite dilations [GLL$^+$04], contourlets [DV05b], Shearlets [LLKW05], reproducing groups of the symplectic group [CDMNT06], Gabor ridge functions [GS08] – and more.

The reason why Shearlets stand out is due to the several properties they possess: they provide optimal sparse representations, are sensitive in characterizing singularities while still being stable against noise in the signal and they provide optimal sparse representation. These are just a few of all the properties that characterizes Shearlets, see [KL12, DDMGL15] for an overview and a complete list of references. From the purely mathematical perspective, their construction is based on the well-established theory of square-integrable representations (see, for example, [FÖ5]), just as wavelets are, and because of this many powerful mathematical tools are available [DST10, DHST15].

From the application standpoint we mention previous works exploiting the properties of the 2D Shearlet Transform to develop approaches for edge detection [YLEK09b, KP15], image inpainting [KKL13], image denoising [ELC09, CHS13], image separation [KL10] and anomaly detection [GPLC14]. Shearlets have also been employed in the biomedical imaging domain, to developed advanced compressed sensing techniques [PKG15, MMF$^+$17] and for phase retrieval purposes [PLPS16]. As for applications based on the 3D Shearlet Transform, we mention, for example, video denoising [NL12a, GPLC14], but to our knowledge there are a few previous attempts to exploit a proper construction of the 3D Shearlet Transform to enhance, highlight or describe three-dimensional structures [GL11].

One of the objectives of our work is to define a taxonomy of the spatio-temporal structures which might arise in the spatio-temporal domain. Previous work on space-time local analysis include some early approaches to describe locally the behavior of points lying on two-dimensional surfaces [KvD92][BBC13], other works have also considered an element within a 2D+T sequence as a 3D object, by considering its evolution over time, and defined a set of indexes on it [BGS$^+$05] so that to carry on an action recognition task. We will introduce in this document a work where the author has been able to define the nature of spatio-temporal corners [Lap05], but that approach only gives sparse labels, giving a name to a few points within the signal.

The methods that will be presented in this thesis are of different nature. One of them is related to a dense analysis of the video sequence searching for points which can be considered more interesting that the others. This is often a fundamental step in a video analysis pipeline, and several works have tried to do this in different ways. In the first case, spatial points are *tracked* over frames and their behavior is considered meaningful for a given task in case they possess an interesting behavior (like in [WKSL11][WS13]); secondly, points are tagged interesting by considering their actual spatio-temporal neighborhood, by looking at the signal as a three-dimensional function varying along the x, y and t axis. We start by considering the seminal work by Laptev et al [Lap05] which set the basis for the understanding of the meaning of spatio-temporally interesting points in the same way we intend them in our work. Laptev defined what a spatio-temporal corner is, in a way that we explore more in depth in the following sections, by extending existing algorithms and theory developed and applied previously in the image processing field [Lin96]. Dollàr et al [DRCB05] have developed the so called *cuboid* detector, a method to detect elements not only representing spatio-temporal corners but also space-time points around which the signal changes repetitively within a given span of time. An even different technique has been developed by Willems et al [WTG08], by exploiting the idea of integral video (as defined in [KSH05]) and by redefining a saliency measure based on the determinant of the Hessian of each given spatio-temporal point within the video sequence. There have been other tentatives to develop spatio-temporal interest points detectors in the years following this work [WC07, CHMG12]. However, in the following chapters we take inspiration from the idea conceived by Laptev, and we develop our own approach exploiting the information available from the Shearlet Transform to carry out our analysis.

The last set of works that we consider is related to the estimation of the motion which characterizes the subject in the scene. On this respect there is a vast literature one could explore, starting from seminal work such as the one carried out by Lukas and Kanade [LK+81], Horn and Schunck [HS81], Farneback [Far00] or a set of other approaches relying on the use of variational methods [BBPW04, BWF+05, BBM09, WPZ+09]. For a broader overview of the several approaches developed in this direction in the past, we refer to [FBK15]. More recent ones have considered the use of deep learning [WRHS13, IMS+17], achieving state-of-art results and improving the speed at which the result of the calculation is provided.

Most of the approaches mentioned above have developed methods focused on considering a single kind of information, derived from the input signal. Our approach relies on the use of the Shearlet Transform to extract different types of information from a video sequence, so to combine them to obtain a description of how elements within the scene are evolving over time. More similar to our methods are the ones introduced in [DNL09, SZ14], which try to improve the results achieved in an action recognition task by combining different kind of inputs (appearance/semantic and dynamic).

Thesis contributions

We ground our work on the result obtained in previous approaches in the 2D signal processing scenario [YLEK09c, DODV15, KP15]. Those works strengthened the basis for a more in-depth use of the Shearlet Transform in image and video processing scenarios.

From the results obtained within those works, we consider the applicability of the 3D Shearlet Transform to analyze 2D+T data, that is when our signal is characterized by two spatial dimensions and a temporal one, thus we focus on the analysis of video sequences.

Our contributions are:

- the **definition of the mathematical ground** needed to support our analysis so that to show how it is possible to derive a better low-level understanding of the content within video sequences by exploiting the properties of the 3D Shearlet Transform [MODV17b, MODVar].

- the **development of a framework**[1] for the analysis of video signals.
- the **reproduction of a few existing techniques** developed in the past in the field, by means only of the information carried by the Shearlet Transform, like the detection of spatio-temporal interest points [MODV17a] or the estimation of the apparent motion happening in the scene.
- a set of **experimental evaluations** to validate preliminarily our approach [MGV+17].

Thesis structure

This document is structured in the following way:

- In Chapter 1 we introduce the Shearlet Transform, the framework on which we base our work.
- In Chapter 2 we recap quickly which spatial singularities have been described the most in the past, we then proceed with one of our contributions, the description and the modeling of primitives in the 2D+T-dimensional case, also showing how to exploit the information brought by the Shearlet Transform to analyze such a kind of signals.
- In Chapter 3 we present the Shearlet-based framework that we have developed, devising different video processing methods to exploit the information carried by the Shearlet Transform, showing a set of experimental results for the techniques that have been developed.
- Finally, we conclude this document with our final considerations and the possible future developments of our work, while also remarking the limitations which have characterized our approach within this research.

[1] which implementation is publicly available online,
`https://github.com/damianomal/Shearlet-Framework`

Chapter 1

Shearlets

In this Chapter we introduce the Shearlet Transform, the main tool we base our work on, together to the formulas and the definitions needed to characterize its properties. By doing so, we set the ground for a better understanding of the techniques and the algorithms that we have developed, and which will be described in the following chapters.

1.1 Introduction

In this thesis, and for all the properties it owns, we decide to use the Shearlet framework and we investigate its applicability in a video processing scenario. To better understand all its fashions and to justify our choice, this chapter gives a quick overview of the needed mathematical theory, of how to construct a Shearlet system and of the corresponding Shearlet Transform, the main tool which we exploit to develop our work.

Though Shearlets were first introduced for 2D-signals [LLKG05], they were extended to arbitrary space dimensions in the seminal paper [DST10]. For 3D signals a digital implementation can be found in [KLR16], which is the main reference for this chapter.

While the notation that we introduce in the following sections is related to the one used in the above-mentioned paper, we make a few choices for the sake of clarity within our work.

1.2 Notation and Definitions

We start by reviewing the construction of the Shearlet frame for 3D signals. We denote by L^2 the Hilbert space of functions $f : \mathbb{R}^3 \to \mathbb{C}$ such that

$$\int_{\mathbb{R}^3} |f(x, y, z)|^2 \, dx \, dy \, dz < +\infty,$$

where $dx\, dy\, dz$ is the Lebesgue measure of $\mathbb{R}^3$, by $\|f\|$ the corresponding norm and $\langle f, f' \rangle$ is 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.*

$$\widehat{f}(\xi_1, \xi_2, \xi_3) = \int_{\mathbb{R}^3} f(x, y, z) e^{-2\pi i(\xi_1 x + \xi_2 y + \xi_3 z)} \, dx \, dy \, dz$$

provided that f is integrable, too. Conversely, we define the inverse Fourier transform of a signal $\hat{g}$ as

$$g(x_1, x_2, x_3) = \int_{\mathbb{R}^3} \hat{g}(\xi_1, \xi_2, \xi_3) e^{2\pi i(\xi_1 x + \xi_2 y + \xi_3 z)} \, d\xi_1 \, d\xi_2 \, d\xi_3$$

We also recall the definition of frame, which provides us with an analytic setting to generalize the idea of orthonormal bases, since Shearlet systems do not form bases but require an extension of this concept.

A frame for L^2 is a countable family $\mathcal{F} = \{\phi_i\}_{i \in I}$ such that each ϕ_i is in L^2 and

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

where A, B are constants so that $0 < A \leq B < \infty$, called *lower* and *upper frame bound*. We are interested in this definition as we consider the Shearlet frame as an analysis tool, since we try to study the associated frame coefficients $\langle f, \phi_i \rangle$ to extract information from our starting signal.

One application of frames is the analysis of elements in a Hilbert space, which can be carried on by the *analysis operator*, given by

$$T : L^2 \to \ell_2(I), \quad T(f) = (\langle f, \phi_i \rangle)_{i \in I}$$

Even if we do not consider it within this work, it is relevant to report here what is called the *frame reconstruction formula*, defined by

$$f = \sum_{i \in I} \langle f, \phi_i \rangle S^{-1} \phi_i \qquad \forall f \in L^2,$$

where $Sf = \sum_{i \in I} \langle f, \phi_i \rangle \phi_i$ is the *frame operator* associated with the frame $(\phi_i)_{i \in I}$.

1.3 Discrete Shearlet Transform

The Shearlet Transform is based on three geometrical transformations: dilations, translations and shearings. These operators are used together to generate waveforms with anisotropic supports and different orientations, while preserving the integer lattice (see [KLR16] for details).

We introduce the notation for the two-dimensional case[1], then we move to the three-dimensional setting. A Shearlet system consists of a given generating function on which a set of transformations is applied, such as a parabolic scaling matrix $A_{\ell,j}$, defined by

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

and a orientation-changing transformation, namely a *shearing* $S_{\ell,k}$, defined

[1] here the notation introduced in Section 1.2 still holds, with trivial modifications to adapt it to $\mathbb{R}^2$

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

plus a translation, which rules its position. The matrices defined above depend also on an index ℓ, since we consider a specific case where the Shearlet system we build is referred to as *cone-adapted*, for the frequency domain is partitioned into cone-shaped partitions (as shown in Figure 1.1).

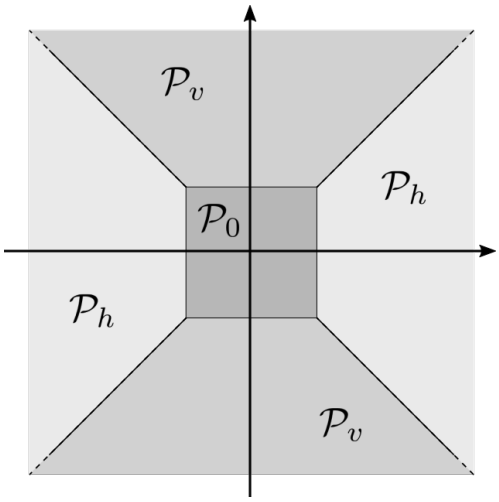


Figure 1.1: Cone-like partition of the Fourier domain.

The Shearlet system $\mathcal{F}$ consists of three families

$$\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_h \cup \mathcal{F}_v$$

where $\mathcal{F}_0 = \{\phi_m : m \in \mathbb{Z}^2\}$ is associated with the low frequency region

$$\mathcal{P}_0 = \{(\xi_1, \xi_2) \in \widehat{\mathbb{R}}^2 \mid |\xi_1| \leq 1, |\xi_2| \leq 1\}$$

and is given by the functions

$$\phi_m(x, y) = \phi(x - cm_1, y - cm_2).$$

where $m = (m_1, m_2) \in \mathbb{Z}^2$ represents the translations, $c > 0$ a step size. The other two subfamilies $\mathcal{F}_h$ and $\mathcal{F}_v$ corresponds each to a different pair of cone-like partitions as shown in Figure 1.1, and they take care, respectively, of the two sets of high-frequency regions

$$\mathcal{P}_h = \{(\xi_1, \xi_2) \in \widehat{\mathbb{R}}^2 \mid |\xi_1| > 1, |\frac{\xi_2}{\xi_1}| \leq 1\} \tag{1.1}$$

$$\mathcal{P}_v = \{(\xi_1, \xi_2) \in \widehat{\mathbb{R}}^2 \mid |\xi_2| > 1, |\frac{\xi_1}{\xi_2}| \leq 1\} \tag{1.2}$$

These two partitions are associated with a set of functions $\psi_{\ell,j,k,m} \in L^2(\mathbb{R}^2)$ which are explicitly defined as

$$\psi_{\ell,j,k,m}(x, y) = 2^{\frac{3}{4}j}\psi(S_{\ell,k}A_{\ell,j}\left(\begin{smallmatrix} x-c_1m_1 \\ y-c_2m_2 \end{smallmatrix}\right)),$$

where $c_\ell = c$ and the other parameter $c_\alpha = \widehat{c}$ with $\alpha \neq \ell$, $\widehat{c}$ is another step size, $|k| \leq \lceil 2^{j/2} \rceil$ is the shearing parameter, $m = (m_1, m_2) \in \mathbb{Z}^2$ determines the translation and matrices $S_{\ell,k}$ and $A_{\ell,j}$ are the ones introduced in the opening of this section.

In the definitions above, the function ϕ is chosen to have compact frequency support near the origin, so that the function ϕ_m will be associated with the low frequency region $\mathcal{P}_0$ in Figure 1.1. Similarly, the functions ψ are chosen so that the family of functions $\psi_{1,j,k,m}$ is associated with the two horizontal cones denoted $\mathcal{P}_h$, and the system $\psi_{2,j,k,m}$ is associated with the cones $\mathcal{P}_v$.

The corresponding Shearlet Transform of a signal $f \in L^2(\mathbb{R}^2)$ is the mapping

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

Where $\psi \in L^2(\mathbb{R}^2)$ is defined through its Fourier representation as

$$\hat{\psi}(\xi) = \hat{\psi}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\hat{\psi}_2(\tfrac{\xi_2}{\xi_1}), \tag{1.3}$$

and $\psi_1 \in L^2(\mathbb{R}^2)$ is a discrete wavelet, satisfying the discrete Calderón condition

$$\sum_{j\in\mathbb{Z}} |\hat{\psi}_1(2^{-j}\xi)|^2 = 1 \text{ for a.e. } \xi \in \mathbb{R},$$

with $\hat{\psi}_1 \in C^\infty(\mathbb{R})$ and supp $\hat{\psi}_1 \subseteq [-\frac{1}{2}, -\frac{1}{16}] \cup [\frac{1}{16}, \frac{1}{2}]$, and $\hat{\psi}_2 \in L^2(\mathbb{R})$ is a bump function for which

$$\sum_{k=-1}^{1} |\hat{\psi}_2(\xi + k)|^2 = 1 \text{ for a.e. } \xi \in [-1, 1],$$

where $\hat{\psi}_2 \in C^\infty(\mathbb{R})$ and supp $\hat{\psi}_2 \subseteq [-1, 1]$. Within this setup, ψ is called a *classical shearlet*.

We consider the implementation of the Shearlet Transform as introduced in [KLR16], by noting that the family of functions ϕ_m is associated with the low-frequency component of the signal and is strictly dependent only on the parameter m, thus it is sufficient to focus on the creation of Shearlets in $\psi_{\ell,j,k,m}$.

Following [KLR16], the generator function ψ is chosen such as

$$\hat{\psi}(\xi) = P(\xi_1/2, \xi_2)\hat{\psi}^{\text{sep}}(\xi) \tag{1.4}$$

where P is suitable polynomial 2D Fan filter [DV05a, DCZD06], and where to achieve better numerical results ψ is chosen to be *non-separable*. More precisely, with a suitable choice of the 2D fan filter P, it holds that

$$P(\xi_1/2, \xi_2)\hat{\phi}_1(\xi_2) \approx \hat{\psi}_2(\tfrac{\xi_2}{\xi_1}) \tag{1.5}$$

and

$$P(\xi_1/2, \xi_2)\hat{\psi}_1(\xi_1)\hat{\phi}_1(\xi_2) \approx \hat{\psi}_1(\xi_1)\hat{\psi}_2(\tfrac{\xi_2}{\xi_1}) \tag{1.6}$$

relating this construction of the generator function ψ with the definition of classical Shearlets as in 1.3. Here ψ_1 and ϕ_1 are defined so that $\psi^{\text{sep}} = \psi_1 \otimes \phi_1$, with ψ_1 and ϕ_1 being respectively a 1D wavelet and a 1D scaling function. To ensure $\mathcal{F}$ to be a frame it is necessary to have some technical condition on the smoothness of ϕ_1 and on the vanishing momenta of ψ_1, see [KL12].

Similarly to what we have introduced in the two-dimensional case, to move to the three dimensions scenario we consider the case in which we adopt a *pyramidal-like partition* of the frequency domain.

The Shearlet frame $\mathcal{F}$ is now defined in terms of four different subfamilies $\mathcal{F}_\ell$ with $\ell = 0, \dots, 3$ as it follows. The first family $\mathcal{F}_0 = \{\varphi_m\}$ takes care of the low frequencies cube

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

and it is given by

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

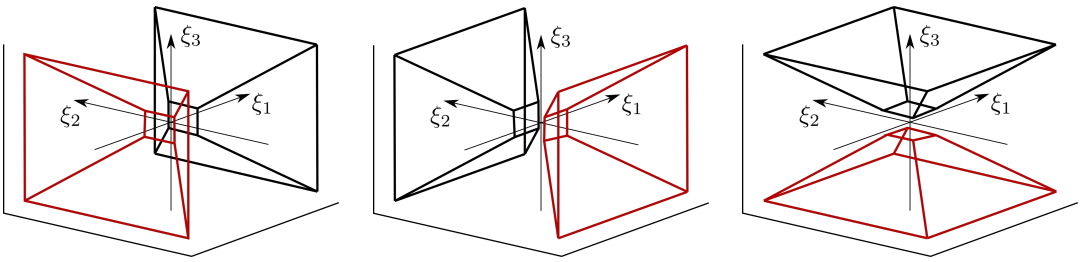
where $m = (m_1, m_2, m_3) \in \mathbb{Z}^3$ labels the translations, $c > 0$ is a step size, and

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

where ϕ_1 is a $1D$-scaling function.

The other three families $\mathcal{F}_\ell$ 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 Figure 1.2. Thus, the three pyramids are

$$\mathcal{P}_1 = \{(\xi_1, \xi_2, \xi_3) \in \widehat{\mathbb{R}}^3 \mid |\xi_1| > 1, |\frac{\xi_2}{\xi_1}| \leq 1, |\frac{\xi_3}{\xi_1}| \leq 1\},$$
$$\mathcal{P}_2 = \{(\xi_1, \xi_2, \xi_3) \in \widehat{\mathbb{R}}^3 \mid |\xi_2| > 1, |\frac{\xi_1}{\xi_2}| \leq 1, |\frac{\xi_3}{\xi_2}| \leq 1\},$$
$$\mathcal{P}_3 = \{(\xi_1, \xi_2, \xi_3) \in \widehat{\mathbb{R}}^3 \mid |\xi_3| > 1, |\frac{\xi_1}{\xi_3}| \leq 1, |\frac{\xi_2}{\xi_3}| \leq 1\},$$



(a) (b) (c)

Figure 1.2: The 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, the low frequency cube $\mathcal{P}_0$ is not shown here.

Fixed $\ell = 1, 2, 3$, each $\mathcal{F}_\ell = \{\psi_{\ell,j,k,m}\}$ is defined in terms of parabolic dilations

$$A_{1,j} = \begin{pmatrix} 2^j & 0 & 0 \\ 0 & 2^{j/2} & 0 \\ 0 & 0 & 2^{j/2} \end{pmatrix}, \quad A_{2,j} = \begin{pmatrix} 2^{j/2} & 0 & 0 \\ 0 & 2^j & 0 \\ 0 & 0 & 2^{j/2} \end{pmatrix}, \quad A_{3,j} = \begin{pmatrix} 2^{j/2} & 0 & 0 \\ 0 & 2^{j/2} & 0 \\ 0 & 0 & 2^j \end{pmatrix},$$

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

$$S_{1,k} = \begin{pmatrix} 1 & k_1 & k_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad S_{2,k} = \begin{pmatrix} 1 & 0 & 0 \\ k_1 & 1 & k_2 \\ 0 & 0 & 1 \end{pmatrix}, \quad S_{3,k} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ k_1 & k_2 & 1 \end{pmatrix},$$

where the index $k = (k_1, k_2) \in \mathbf{K}_j$ controls the shearing and runs over

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

Explicitly, in this case the functions $\psi_{\ell,j,k,m}$ are given by

$$\psi_{\ell,j,k,m}(x,y,z) = 2^j \psi_\ell \left(S_{\ell,k} A_{\ell,j} \begin{pmatrix} x - c_1 m_1 \\ y - c_2 m_2 \\ t - c_3 m_3 \end{pmatrix} \right), \tag{1.7}$$

where $c_\ell = c$ and $c_\alpha = c_\beta = \widehat{c}$ if $\alpha, \beta \neq \ell$, $\widehat{c}$ is another step size as in the two-dimensional case and, as for the family $\mathcal{F}_0$, $m = (m_1, m_2, m_3) \in \mathbb{Z}^3$ labels the translations.

Following [KLR16], the generating vector ψ_1 is of the form

$$\widehat{\psi}_1(\xi_1,\xi_2,\xi_3) = \widehat{\psi}_1(\xi_1) \left(P(\frac{\xi_1}{2},\xi_2)\widehat{\phi}_1(\xi_2) \right) \left(P(\frac{\xi_1}{2},\xi_3)\widehat{\phi}_1(\xi_3) \right), \tag{1.8}$$

where, P is a suitable polynomial 2D Fan filter similarly to the 2D case, ψ_1 is the 1D wavelet function associated with the scaling function ϕ_1 defining the family $\mathcal{F}_0$. Similar equations hold for $\ell = 2, 3$ by interchanging the role of ξ_1, ξ_2 and ξ_3.

The 3D Shearlet Transform of a signal $f \in L^2$ is given by

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

where $j \in \mathbb{N},\ k \in \mathbf{K}_j,\ m \in \mathbb{Z}^3$. In the experiments we use the digital implementation of the 3D Shearlet Transform described in [KLR16], 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.*

$$\phi_1(x) = \sqrt{2} \sum_{n \in \mathbb{Z}} h(n) \phi_1(2x - n) \tag{1.9}$$

$$\psi_1(x) = \sqrt{2} \sum_{n \in \mathbb{Z}} g(n) \phi_1(2x - n). \tag{1.10}$$

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

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

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 (1.8) and the low pass filter h associated with the scaling function ϕ_1 by (1.9). A further degree of freedom is the possibility to fix for each scale j a different number of shearings $k \in \mathbb{Z}^2$.

Our algorithm is based on the following nice property of the Shearlet coefficients. As shown in [GL11, GL12, KLL12, KP15, DST10] if the signal 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 $k \in \mathbf{K}_j$. 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 $SH[f](\ell, j, k, m)$ 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 given by

$$(\cos\alpha \cos\beta, \cos\alpha \sin\beta, \sin\alpha) \qquad \alpha, \beta \in [-\frac{\pi}{2}, \frac{\pi}{2}]. \tag{1.11}$$

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

$$\tan\alpha = \frac{2^{-j/2} k_2}{\sqrt{1 + 2^{-j} k_1^2}} \quad \tan\beta = 2^{-j/2} k_1 \quad \alpha, \beta \in [-\frac{\pi}{4}, \frac{\pi}{4}]. \tag{1.12}$$

For $\ell = 2, 3$ the angles α and β are calculated in the same way, plus an additional rotation of 90 degrees around an axis, which depends on ℓ.

In real applications, the ability to detect singularities strongly depends on the choice of the generating function ψ_1 and, hence, on the mirror filter pair (h, g). In this thesis, we adopt the mirror filter introduced in [MZ92] in the context of wavelets, which behaves like a bank of first derivative of Gaussians. In this way, every single Shearlet $\psi_{j,k,m}$ will extract some information about how our signal is varying along the direction associated with the shearing parameter k and with a level of granularity ($\sim$scale) j, allowing us to highlight the discontinuities characterizing our signal.

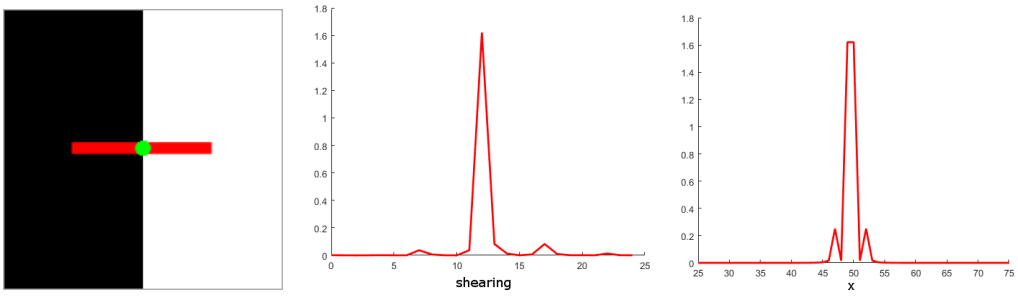
1.4 Geometrical Interpretation of Shearlet Coefficients

The sensitivity of the Shearlet coefficients of a signal to the spatial singularities contained in is has been deeply explored in the past for the two-dimensional case [DODV15, ELL08, YLEK09c]. The Shearlet Transform has shown itself to be a good tool in characterizing locally the behavior of two-dimensional singularities in images, and we want to move on to understand its capabilities in the three-dimensional case. Thus, in this section we explore the possibility to use the information carried by the Shearlet Transform so to understand the nature of spatio-temporal primitives in the three-dimensional domain.

For this aim, we consider synthetic spatio-temporal signals, representing a black object embedded in a white (background) space. The synthetic entity we are considering can be seen as a three-dimensional cube, while the white part of the signal can be seen as "empty" space around it. An example of a single slice of the object is represented in Figure 1.3(a), where the black part on the left represents the content of the three-dimensional cube, while the white portion on the right is the space in which the object has been embedded. We base our work on the ShearLab3D framework, and thus we can treat separately the Shearlet coefficients which belong to the three different pyramidal partitions $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$.

We begin our analysis on a specific point on a side of the cube parallel to the yz plane (the green dot in Figure 1.3 and following ones), which we know belongs to one of the surface characterizing the synthetic object. We then compute the Shearlet coefficients while we move our focus along the normal direction outside the cube (which is represented by the red line in Figure 1.3). The behavior of the coefficients is shown in Figure 1.3, where on the first column we show a sample of a xz-section of the synthetic cube at a given depth z. In the second column we plot the value of the Shearlet coefficients in the first

pyramid $\mathcal{P}_1$ (corresponding to the partition aligned along the x-axis) at the point selected on the surface, while we vary the shearing parameter. Note that the number coefficients in this plot is related to the number of shearings corresponding to the partition $\mathcal{P}_1$, which is associated with a grid of 5x5 shearings. The coefficients of the other two pyramids contain negligible values in the order of $\sim 10^{-16}$, thus we do not show the same plots in the case of $\mathcal{P}_2$ and $\mathcal{P}_3$. In the third column, we fix the shearing k^* corresponding to the peak value in Figure 1.3(b) and see how the coefficients evolve by moving along the normal direction corresponding to the red line in Figure 1.3(a), which also corresponds to the direction associated with the shearing k^*. The coefficients decay as we move away from the discontinuity, giving us an empirical evidence of the appropriateness of the analysis of the 3D Shearlet coefficients to localize interest points in correspondence of three-dimensional surfaces.



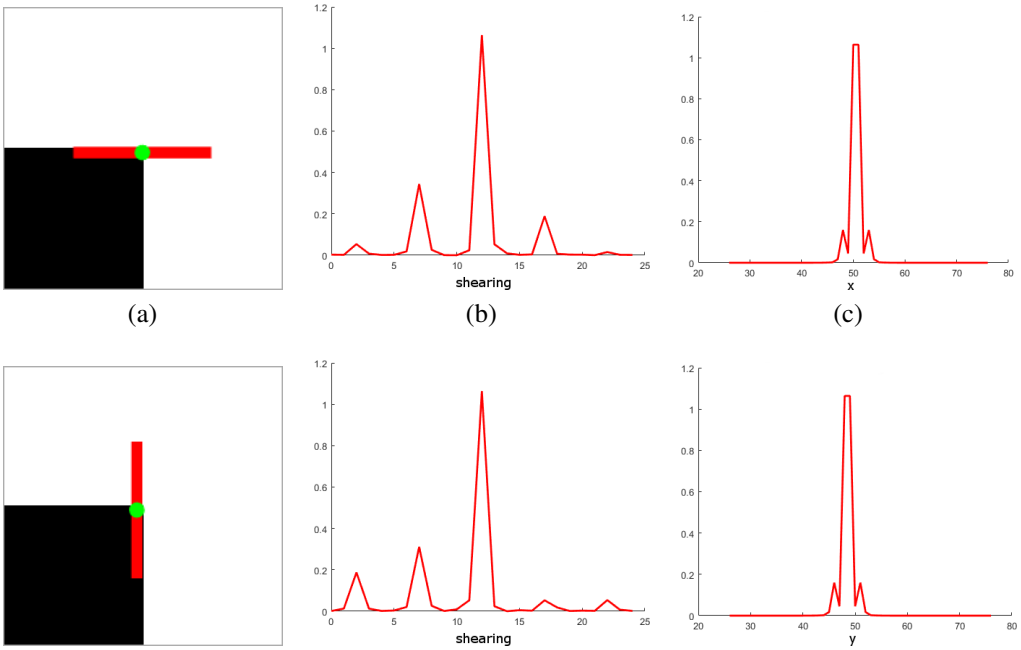
(a) (b) (c)

Figure 1.3: Coefficients analysis on a 3D surface (see text). (a) A section of a surface parallel to the yz 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 $\mathcal{P}_1$. (c) The coefficients decay for neighboring points along the surface normal (the red line), which corresponds to the shearing parameters k=(0,0) and the index 12 in the coefficients vector unrolled in (b).

We now consider a slightly different synthetic object. Figure 1.4 shows a similar analysis on a 3D edge produced by two surfaces, one which is parallel to plane xz and the other parallel to plane yz. In this case we identify two significant peaks in two different pyramids (the main peaks in (b) and (e), corresponding to pyramids $\mathcal{P}_1$ and $\mathcal{P}_3$).

Within the two pyramids $\mathcal{P}_1$ (in Figure 1.4 (a-c)) and $\mathcal{P}_3$ (in Figure 1.4 (d-f)) we see a similar behavior to the case of the 3D surface in Figure 1.3 (b). 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 coefficients associated with a 2D grid of directions) as a 1D function, thus they appear to be distant on the one-dimensional visualization.

The plots we show 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 procedure applicable to all points of an image sequence.



(a) (b) (c)

(d) (e) (f)

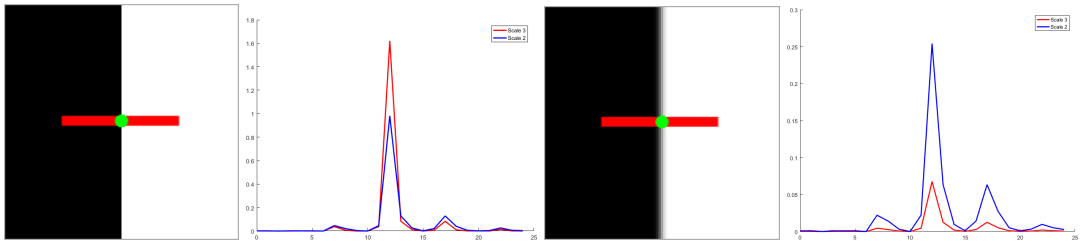
Figure 1.4: Coefficients analysis on a 3D edge (see text). (a) and (d) show a section of the edge parallel to xy plane, where we highlight the two normal vectors. (b) and (e) show the coefficients varying at different shearings in the two meaningful pyramids. (c) and (f) show the decay of coefficients for neighboring points along the corresponding normal directions.

In the previous examples we have considered a single scale, while displaying the behavior of the Shearlet coefficients while the shearing and the translation parameters vary. As explained in the first chapter Shearlets provide a multi-scale representation and, for 2D signals, this property allows to efficiently implement many algorithms[2] in image processing as deblurring [FB12, HHZ14], denoising [ELC09, NL12b, ELN13, EL12], blob detection [DPNODV17, DNOD17] and signal reconstruction [GL13], to name a few. From a theoretical point of view the same framework holds for 3D signals, however the computational cost of the 3D Shearlet transform forces to have only a few scales available. For this reason a complete analysis of the behavior of the Shearlet coefficients across scales is out of the scope of this thesis. Here, we only focus on what happens when we consider more than one scale and when the signal is also characterized by blurring.

[2]and a publicly available framework to carry on experiments, which is also the one we consider within this thesis, is available at `http://www.shearlab.org/`

What we suppose is that the coefficients at the finer scales (the ones corresponding to higher-frequency details) should be affected by the fact that the blurring effect spreads the finer details in the neighborhood of each point.

In Figure 1.5 we have an example of this behavior. In Figure 1.5(a-b) we show both the second and the third scale coefficients (in different colors, check the caption for details), trying to understand what happens to them while we focus our attention on the spatio-temporal point marked by the green dot and we vary the shearing parameter. We can see that the coefficients for the third scale are higher than the ones for the second one. In the case we introduce some blurring in our synthetic signal, as in Figure 1.5(c-d), the third scale coefficients decrease considerably, falling almost to zero, while those associated with the second scale (the *coarser* one, capturing lower frequency singularities) maintain higher values.



(a) (b) (c) (d)

Figure 1.5: A comparison of the coefficients for the two different scales, in the case a signal has or not blurring artifacts: in **blue** the ones corresponding to the second scale, and in **red** the ones corresponding to the third scale (corresponding to higher frequency details).

These simple experiments give us insights about the fact that the coefficients of the 3D Shearlet Transform of a three-dimensional signal are meaningful to characterize the local behavior of a spatio-temporal point, belonging it to surface-like or to more complex singularities. We limited our analysis to a small number of scales for computational reasons, for the calculation of the shearlet coefficients is demanding both in memory and in time, with the former limiting the most. All this is also related with the number of shearings considered, thus we will fix it for the experiments that we will carry on in the next sections.

Here we have considered only synthetic examples, so we need to move on and develop the basis about how to carry on the same kind of analysis on real data.

Chapter 2

Spatio-Temporal Structures

The idea of modeling spatial structures has been addressed in the image processing community since its early days. In this chapter, we quickly review the main contributions made in the field, starting from the two-dimensional case and describing which spatial elements have been analyzed and universally recognized in the past. Usually, these spatial elements have been the "building blocks" for further processing, for this reason the approaches developed have tried to extract and represent them more and more precisely over the years. We then proceed introducing one of the contributions in our work, the description and the modeling of primitives in the 2D+T-dimensional case, that are relevant in the analysis of video sequences. Within this scenario, we will show how to exploit the information brought by the Shearlet Transform to analyze such a kind of signals.

2.1 Singularities in 2D Signals

While the scientific community started to explore the image processing field, the first steps have been taken in the direction of exploring and understanding which types of spatial singularities could be meaningful to detect, analyze and describe in the two-dimensional scenario. Here, *meaningful* refers to the fact that these elements characterize the scene being analyzed, thus allowing for a better understanding of its content.

A large amount of work was devoted to try to understand how physical quantities could influence the way three-dimensional shapes in the real world are mapped on an hypothetical image plane [BT78, SA93]. After putting a lot of effort into wondering how to

model these physical and optical transformations, researchers moved on to a more algorithmic approach, with the objective to describe mathematically the appearance of a set of specific elements within an image so that, for each given task, to separate meaningful parts characterizing it from portions with no interesting behavior. The work leading this direction is the one carried on by D. Marr, which notes have been gathered in [MV82]. He explored several aspects of the problem of representing, recognizing and modeling elements mapped from the 3D world to images, setting the ground for the research in the years to come.

Seminal papers in this field tried to categorize and identify really simple elements within the image frame. The first methods developed have been related to the detection of region boundaries (namely "edges") [Can86, PM90, MZ92, Lin98a], corners (or, more generally, interest points) [FG87, HS88, TK91, WB95, LSC95], ridges [EGM$^+$94, Lin98a] and subregions of pixels with similar characteristics [Lin98b, Low04] (which might refer to groups of adjacent pixels with a similar behavior in terms of light *intensity*). Amongst these methods, a few of them are more of interest to us, in particular the approaches grounded on wavelets [MZ92, LSC95] or those which base their computation on scale-space theory [PM90, Lin98b]. Examples of the elements that we listed above can be seen in Figure 2.1. The wide range of works made in this direction showed the feasibility of a taxonomy of the spatial singularities which can be found in a two-dimensional image signal, thus setting the ground for further developments.

(a) image (b) edges (c) corners (d) regions

Figure 2.1: Examples of different kinds of spatial singularities.

The majority of the approaches developed in the past relied on the definition and the creation of ad hoc solutions for the analysis of each given type of singularity. On the other hand, more recent approaches have tried to *learn* meaningful features from data, trying to rely on less as possible a priori information about their shape. In this scenario, we report here only two main kinds of approaches: the ones based on dictionary learning, and those relying on neural architectures.

Dictionary learning, in the image processing scenario, refers generally to a set of techniques with the aim to find a sparse representation of a signal and adapting to the data being analyzed by learning which features are the most representatives. This kind of approach has been considered in the past to carry on object categorization and image classification tasks [WCM05, YYGH09], to infer discriminative features for different types of objects [MBP+08], for image denoising tasks [EA06], to learn a hierarchy of spatial primitives [JMOB10], and many others.

Methods based on deep learning architectures [LBBH98, KSH12, ZF14] have developed a set of models able to learn and infer the spatial primitives which are descriptive the most of what is represented in a given dataset of images. This has been possible by exploiting a precise type of networks, namely the *convolutional neural networks*. While a nontrivial subset of all these learnt primitives can be often associated with existing, well-known interesting elements (*i.e.* edges, corners or intersections, and so on), the features to which the hidden units in the deeper layers of these networks become more sensitive to represent more complex behaviors characterizing the data these architectures have been trained on. An example of low-level features learnt from such a network is represented in Figure 2.2, where it is possible to see edge-like objects and regions of constant color, together to elements with a more complex behavior.

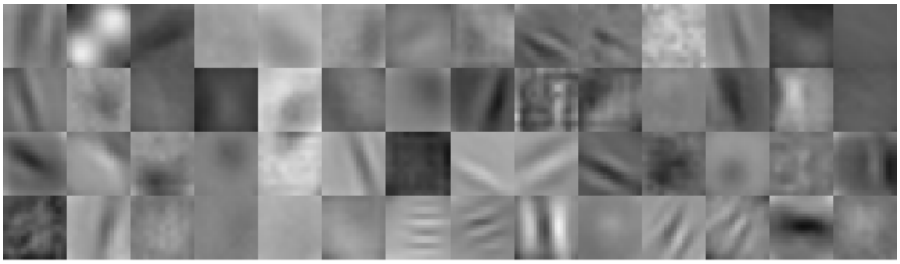


Figure 2.2: Examples of features learned in the first layer of a CNN architecture (from an implementation of AlexNet [KSH12]).

While this has shown the capabilities of such a kind of architectures, which are able to adapt themselves to the various nuances which can arise in each set of data, these models lack interpretability. We would like to be able to visualize the set of features which we are learning from our data, while at the same time being confident about the kind of objects we are looking at.

For their sensitivity in capturing anisotropic features, we consider the Shearlet framework to carry on such a task. We want to exploit their ability to detect spatial singularities in the spatio-temporal domain, while processing video sequences. In the next section, we try to understand the kind of information the Shearlet coefficients are carrying with them, as a ground base to build our further approaches onto.

2.2 Video Sequences as 3D Signals

Since we are considering the analysis of video sequences, we have to think about how to extend the previous ideas also to the 2D+T-dimensional case. In this thesis we consider a whole sequence as a three-dimensional object, the content of which changes along three different dimensions (two spatial and a temporal one). We don't perform tracking of matching tasks between different images, so that to correlate temporally elements which belong to subsequent frames. Previous papers in this field have followed a similar approach, exploiting the ability to calculate a foreground mask and to *cut off* the target from the whole scene [BGS$^+$05][Dav01] or actually considering a sequence in the same way we do [Lap05], as an actual three-dimensional object.

A video sequence can be regarded as 3D signal by stacking each 2D frame along the third direction (the t/z-axis). A region of interest moving in time generates a 3D volume. For example, in the sequence in Figure 2.3, taken from a popular action recognition dataset [SLC04], the boxing man generates the green volume depicted in Figure 2.3(d). This shape is obtained by selecting, for every frame in the sequence, only the pixels corresponding with the body of the boxer. This produces a binary mask representing pixels within the subject (value 1) or outside it (value 0). If we stack all the binary masks we obtain the three-dimensional shape as in Figure 2.3(d), which can be analyzed to understand what the actor in the scene is doing (e.g. which action it is executing).

This is just an example of visualization of how a single object evolves within the sequence. In this thesis, while developing our spatio-temporal analysis framework, we consider video sequences as real three-dimensional signals, without an a priori knowledge about the concept of background/foreground for them.

Regarding a video sequence as a 3D signals, it is natural to identify spatio-temporal primitives as the boundaries of the volumes generated by the regions of interest in the

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

Figure 2.3: (a-c) Sample frames of the *boxing* sequence and (d) corresponding shape generated from the movement.

video. Surfaces, edges and vertices of the 3D shape correspond to different behaviors in space and time.

The volume in Figure 2.3(d) is clearly representative of the motion happened in the scene. The building block which is missing is a formal definition of the different elements which characterize the surface of such volume, which is the topic of the next Section. While doing this, we have to keep into account of the fact that we are not working exactly with three-dimensional data, since one of the dimensions that we are considering (the temporal one) has a strong, different meaning w.r.t. the other two.

2.3 Primitives in 3D Signals

One of the first questions we have to ask ourselves was how to extend the idea of singularity (or primitive) to the spatio-temporal domain, since this is the scenario that we are considering. We have been able to address this problem by means of an in-depth understanding of the information that the Shearlet Transform is extracting through the corresponding decomposition of a video signal.

Therefore, if we analyze the behavior of the signal in space-time, we may observe that several different types of primitives arise (see also Figure 2.4):

- **Spatio-temporal surfaces**, caused by 2D edges with a smooth velocity profile, thus spanning surfaces in space-time.

- **Spatio-temporal edges** either caused by 2D corners moving smoothly or by 2D edges undergoing an abrupt velocity change. These two primitives could be discriminated by detecting the orientation of the 3D edge, see Figure 2.4 (b) and Figure 2.4 (c).
- **Spatio-temporal corners or vertices** caused by 2D corners undergoing a strong 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 cluster these primitives in still and moving entities (corresponding to different orientations in the 2D+T space). Also, the third component (corresponding to the time dimension) has a different intrinsic scale, and very precise constraints, since spatial features do not disappear all of a sudden and the time dimension can only proceed forward.

In the reminder of the work 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, 3D corners or spatio-temporal corners in space-time.

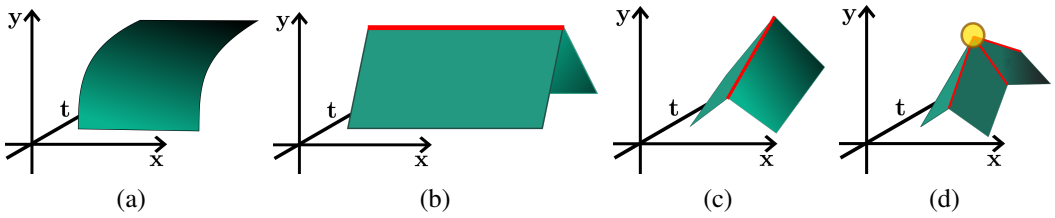


Figure 2.4: Spatio-temporal primitives which can take place in the space-time domain, by considering how the image in the background of each one of these moves over time: (a) a 2D edge moving smoothly spawns a spatio-temporal surface (b) a 2D edge undergoing a velocity change thus producing a 3D edge, (c) a 2D corner moving smoothly also producing a 3D edge, (d) a 2D corner undergoing a velocity change providing a 3D vertex.

We assume that the region of interest is a rigid planar body $\mathcal{C}$ moving in the time interval $[0, T]$ (in Figure 2.5 (a) we can see a visualization of the object at time t, with a few

quantities depicted on it). We further assume that the boundary of $\mathcal{C}$ can be parametrized at the initial time $t = 0$ by the simple closed curve

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

where L is the length of the boundary, s is the oriented arc-length and the curve is oriented so that the interior of the body is on the left side, see Figure 2.5. We denote by $\mathbf{i}$ and $\mathbf{j}$ the canonical unit vectors of the x-axis and y-axis, respectively. Since we assumed that 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 (see Figure 2.5 (b)) whose boundary is the surface parametrized by

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

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 $\sigma(s,t)$

$$N(s,t) = \frac{\partial\sigma}{\partial s}(s,t) \wedge \frac{\partial\sigma}{\partial t}(s,t) = \det \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial x}{\partial s}(s,t) & \frac{\partial y}{\partial s}(s,t) & 0 \\ \frac{\partial x}{\partial t}(s,t) & \frac{\partial y}{\partial t}(s,t) & 1 \end{bmatrix} \tag{2.1}$$

$$= n(s,t) + \tau(s,t) \wedge v(s,t) \tag{2.2}$$

where we have

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

Here, $\tau(s,t)$ and $n(s,t)$ 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, and 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, for it is obtained by clockwise rotating the tangent vector $\tau(s,t)$ by $\pi/2$, see Figure 2.5.

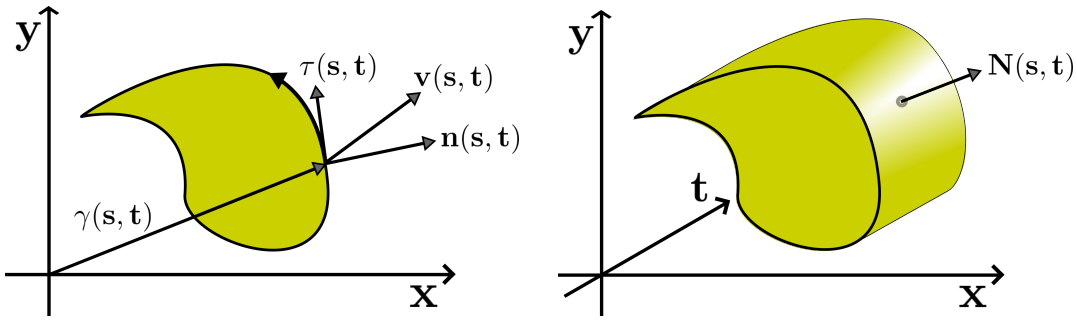


Figure 2.5: (left) A body at time t with the main relevant geometrical and dynamical quantities, (right) the spatio-temporal surfaces spanned by the movement of the body over time.

We now consider four prototypical setups (or "behaviors"), thanks to which we try to relate the evolution over time of the rigid body $\mathcal{C}$ with the Shearlet coefficients calculated for a given point $\gamma(s,t)$:

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 each point there is a tangent plane whose normal vector is given by $N(s,t)$, (see Figure 2.4 (a)); 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 corresponding to the shearing parameter associated with the direction corresponding to 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

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

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

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

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 non-zero component only along the t-axis and lives on the 3D edge (see Figure 2.4 (b)). In this case the Shearlet coefficients would include two maximum values, associated with the shearing parameters corresponding to the normals of 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

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

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 Figure 2.4 (c)). Again, the Shearlet coefficients would include two maximum values associated with the two surfaces described above.

4. The boundary has a 2D corner at point $(x(s_0), y(s_0))$ and there is a (significant) change of velocity at time $t = t_0$ either in the direction or in the speed. At the spatio-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

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

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 Figure 2.4 (d)) corresponding to four maxima in the coefficients.

This toy model may be adapted to real data, as we will see in the next sections. Here we provide a few examples of different behaviors in real video sequences. In Figure 2.6 (top) we may observe the evolution of the tip of a foot changing direction at the end of a step; this object, which actually looks like a 2D corner, behaves in a way so to produce a spatio-temporal corner, or 3D 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.

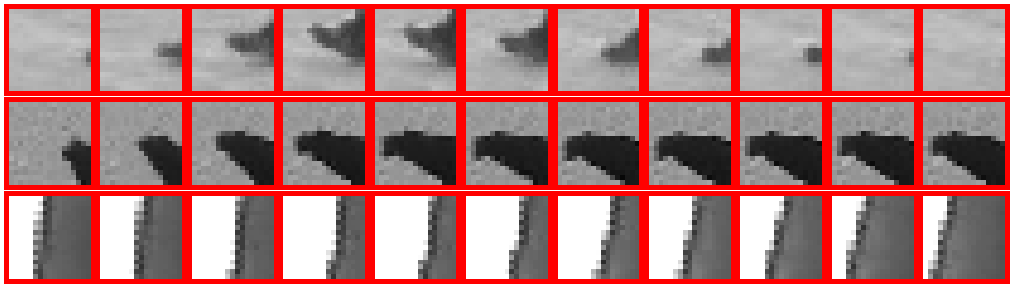


Figure 2.6: Space-time features in real data. Top: the tip of a foot changing direction at the end of a step produces a spatio-temporal corner; middle: the tip of a fist in the extension phase of a punching action produces a spatio-temporal edge; bottom: the side of an arm translating as a person is walking leads to a spatio-temporal surface.

Chapter 3

A Video Analysis Framework based on Shearlets

In this chapter we present one of the main contributions of this work: the development of a set of video analysis functionalities solely based on Shearlets. We devise different video processing methods to exploit as most as possible the information carried by the Shearlet Transform, showing a set of experimental results for every technique that has been developed.

3.1 Introduction

An important aim of this thesis is to provide evidence of the richness of information provided by the Shearlet decomposition of a signal. Indeed thanks to this, starting from a single, shared computational baseline (the decomposition itself), we are able to analyze a video signal at different levels, extracting appearance as well as dynamic information.

We start by addressing feature detection, and then consider feature representation. Here we include both appearance and motion. More specifically, we propose these methods:

- the selection of the **points** in a sequence which are **interesting** in the space-time, as a preprocessing step for sparser analysis of a video sequence,

- the development of a **shearlet-based representation**, so that to understand to which kind of spatio-temporal primitive each point belongs, following the theoretical analysis of Section 2.4,
- the **estimation of the apparent motion**, by analyzing the information related to the temporal changes which occur within the signal.

The three approaches have been developed incrementally, by an increased understanding of the capabilities of the Shearlet Transform after each experiment. We first focused on the detection of spatio-temporal singularities [MODV17a], inspired by some approaches developed in the past in the field [YLEK09a, DODV15]. We observe how the directional sensitivity of the Shearlet decomposition is an efficient tool to capture local spatio-temporal meaningful behaviors, and we exploited this property for the sake of finding elements with a behaving interestingly both in space and in time.

The results we obtain bring to the question whether this sensitivity could be exploited further also to describe every point in a video sequence with respect to its spatio-temporal behavior. This leads us to the development of a shearlet-based pointwise representation [MODV17b], grounded on the theory reported in Section 2.4.

We finally observe that while analyzing the feature behavior on the temporal axis, we are in fact considering its dynamics. Thus, we develop an algorithm for the extraction of the information related with the motion which is happening in the scene. While doing so, we notice that the amount of motion information which we can estimate is limited, due to some properties of the Shearlet transform that we better highlight in the corresponding section.

The following sections dive deeper and describe the methods that we have developed, by also showing their capabilities by means of a set of examples, for which we drew a set of sequences from several datasets, some of them which have been widely adopted in the past in the video processing community to carry on action or gesture recognition tasks. The reason for we chose such sequences is related to the fact that each one of them is helpful in highlighting the properties of the different methods that we introduce in the next sections. For more details, refer to the Appendix at the end of this thesis.

3.2 Detection of Spatio-Temporal Interesting Points

In this section, we introduce the shearlet-based dense method that we have developed to detect a sparse set of points which can be considered *interesting*, by first summarizing the approach which inspired us.

3.2.1 Spatio-Temporal Corners

Following the seminal work carried out by Laptev and his co-workers [Lap05] we are interested in structures which behave meaningfully both in the spatial and in the temporal dimension.

We start by considering a single given frame, where it has been shown by Shi and Tomasi [ST94] that corners posses good properties in terms of detection over scales, stability, and more. These points are also considered *interesting* in the spatial domain since they usually have been considered to be easier to track and describe, for in the past a lot of effort has been put into understanding which features would have been the best to be followed over time.

To extend the analysis to the temporal dimension, in their novel definition, and extending the work made by Harris and Stephens [HS88], a new family of points called *spatio-temporal corners* is described, representing spatial corners which direction of movement changes abruptly over time. Within their work and in following articles [LMSR08, MLS09] it has been shown how these points are meaningful to sparsify the analysis of a given sequence while still understanding efficiently which kind of movement or action was performed in it.

This approach is appropriate to our research, because of the strong directional sensitivity of the Shearlet decomposition. Following these considerations, we propose a method able the select points which vary considerably along three directions, the two spatial and the temporal ones, and the information carried by the Shearlet decomposition suits perfectly with our needs.

3.2.2 Shearlet-based Detection Method

As we have shown in Section 1.2, 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 [HL16]. 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 [MZ92] for wavelets and in [YLEK09a, DODV15] for shearlets, by taking inspiration from and revisiting the algorithm developed in the latter one. More precisely, we consider the details in [MZ92] during the construction of our Shearlet system, adapting the work in [DODV15] to the video analysis case, following a standard procedure [MODV17a].

More in details, we first define an *interest measure* IM representing a response function calculated for each point $m = (x, y, t)$ in our signal. We want this measure to rise in case the point m is placed on a spatio-temporal point with a rich behavior, thus we combine the contributions related to the different shearings, belonging to the three pyramidal partitions $\mathcal{P}_\ell$. 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 3.1 and summarized in Algorithm 1:

a) We compute $IM_j[f]$ for $j = 1, 2$ - for we want to 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, because it is related to really low-frequency information about the signal and it is not helpful in characterizing locally the behavior of a spatio-temporal point.

b) Then, we define an overall interest measure by multiplying the values calculated for $IM_j[f]$, for $j = 1, 2$

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

c) We perform 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 $IM[f](m) > \tau$.

Input: The coefficients $SH[f](\ell, j, k, m)$ for a video sequence f,
a window size w, a threshold τ
Output: The interest points selected in the signal
for $j \leftarrow$ *1* ***to*** *2* **do**
  for $m = (x, y, t)$ ***in*** f **do**
    $IM_j[f](m) = \prod_{\ell=1}^{3} \sum_{k=(k_1,k_2)} |\, SH[f](\ell, j, k, m)\,|$
  end
end
$IM[f](m) = IM_2[f](m) \cdot IM_1[f](m)$ Fig. 3.1(b)
$maxima =$ `NonMaximaSuppression` $(IM[f], w)$ Fig. 3.1(c)
return `Threshold` *(maxima, τ)* Fig. 3.1(d)

Algorithm 1: Spatio-Temporal Interest Point Detection.

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

Figure 3.1: A summary of the detection pipeline we developed shown on an example. (a) a frame I_t from the original video (from ChaLearn dataset [EBG+14]); (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.

Since we only have three scales, the analysis across scales in [MZ92] is not meaningful. We observe experimentally that the points of interest produce high values in both the scales $j = 1, 2$ and this observation is the root of the above definition.

Notice that in Figure 3.1 (b) the IM measure is shown for a fixed time step 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 3.1 (c)) as they are not maxima with respect to the temporal direction (*i.e.* they will be marked as candidates in some neighboring time instant).

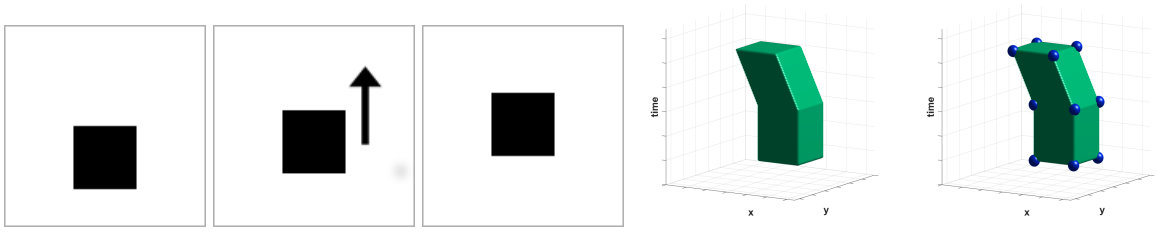
3.2.3 Experimental Assessment

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

Detecting features in action sequences

We start by considering the use of a synthetic sequence, built on purpose to spawn a shape which contains a precise and well-localized set of 3D vertices (or spatio-temporal corners). The sequence represents 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 3.2 (a-c) shows a selection of meaningful frames in the synthetic sequence, while Figure 3.2 (d) shows the 3D shape we obtain by stacking the video frames one on top of the other, finally Figure 3.2 (e) shows the spatio-temporal corners detected by our approach. Our method precisely detects the 12 spatio-temporal elements which represent the 3D corners belonging to this scene.

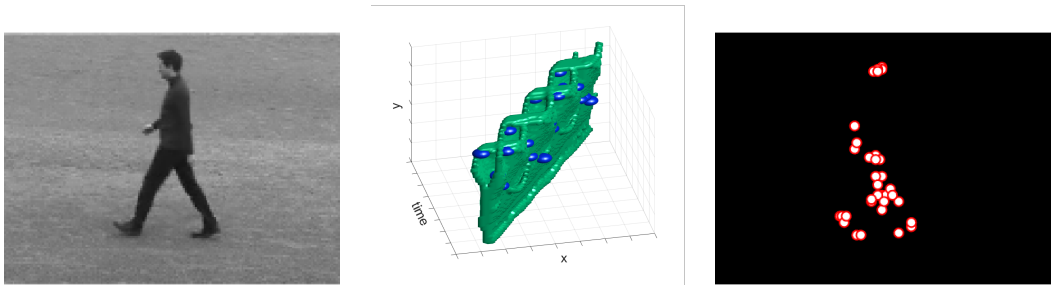
We also show examples of the extracted features in human action sequences. Figure 3.3 shows the results on a *walking* sequence from the KTH dataset, in two different visualization modalities: a 3D shape of the person silhouette evolving in time, where the detected features are marked as blue ellipsoids; a map where the positions of detected points across the whole sequence are merged, centering them with respect to the centroid of the silhouette of the subject. It can be noticed how all meaningful points (in particular all the points corresponding to a change in direction of the foot) have been nicely detected. Similarly, Figure 3.4 shows an example of a different human action, a *handwaving* one, where most spatio-temporal interest points are detected on the tip of the hands, on the



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

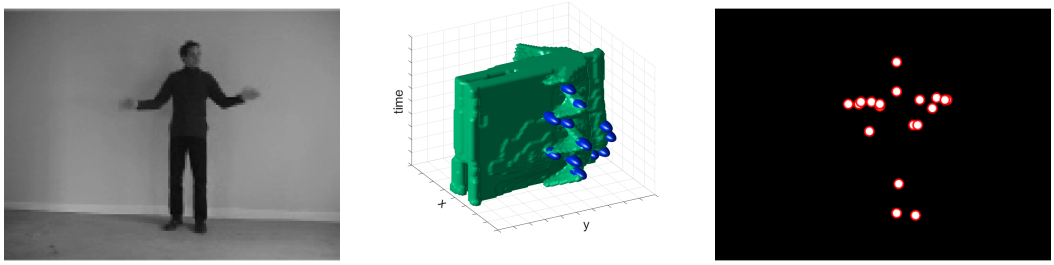
Figure 3.2: (a-c) Sample frames of the video sequence used to generate the shape taken into account in this section, (d) the shape resulting from the behavior of the black square within the sequence and (e) the spatio-temporal points detected by our method.

elbows and on the armpits.



(a) (b) (c)

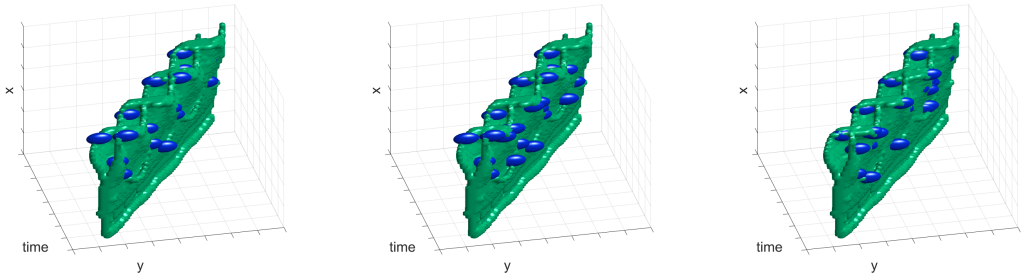
Figure 3.3: A *walking* action (a) observed as (b) feature detection on a 2D+T surface (where we flipped the surface upside down to better show the points detected) and (c) summarized on a reference time instant (detected features are translated w.r.t the body centroid).



(a) (b) (c)

Figure 3.4: 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.

For the sake of evaluating our method, we compare its results with the STIPs detector developed by Laptev and colleagues [Lap05]. In Figure 3.5(a) and (b) we slightly varied the parameters of our detection method, while in Figure 3.5(c) we report the results of the STIPs detector. It is possible to see how the set of points overlap in several points, in correspondence of the tip of the feet changing velocity and on the fist of the subject, swinging back and forth during the strides.

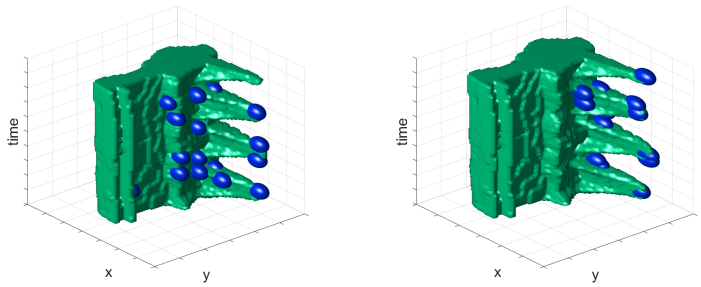


(a) (b) (c)

Figure 3.5: Spatio-temporal interest points detected on a *walking* action by (a,b) our method with two different settings of the parameters, and (c) Laptev's STIPs detector.

Our methods achieves similar results with respect to the STIPs detector. Both approaches can be tuned at the parameters level, leading to different results, and our approach is able to behave similarly to the STIPs detector given that we tune in advance the parameters of our procedure.

Another comparison is the one in Figure 3.6, where we consider a *boxing* sequence. In both visualization it is possible to see how the movement of a fist might cause the detection of two spatio-temporal interest points, for it stops its movement, so that to start again a few seconds later. Our method in Figure 3.6(a) misses the first spatio-temporal point, on the fist on the top of the visualization. This is due to the fact that we ignore a few frames at the beginning of every sequence, so to avoid the boundary conditions which characterize the Shearlet coefficients when near to the boundaries of a signal. While the STIPs detector fires only for the points belonging to the arms and to the fists, our method also detects the lower corner of the jacket worn by the subject, which oscillates during the execution of the movement.



(a) (b)

Figure 3.6: Spatio-temporal interest points detected on a *boxing* action by (a) our method and (b) Laptev's STIPs detector.

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 [Lap05]. Their presence can be used as a guideline on the importance of a given frame in a video summarization process. We evaluate the number of space-time interest points detected in each frame and select the most meaningful frames as the ones containing a large number of those points. While doing so we also apply a non maxima suppression on a spatio-temporal neighborhood of size ω to avoid the selection of frames too close in time.

Figure 3.7 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 the ChaLearn dataset) ones.

For the sake of the experiment, we select three frames in the sequences with the highest number of points detected. Figure 3.8 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 3.9 shows the three most meaningful frames of the *che 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 the starting position. Similarly to the previous case, the three frames identified highlight very peculiar elements of the acquired action.

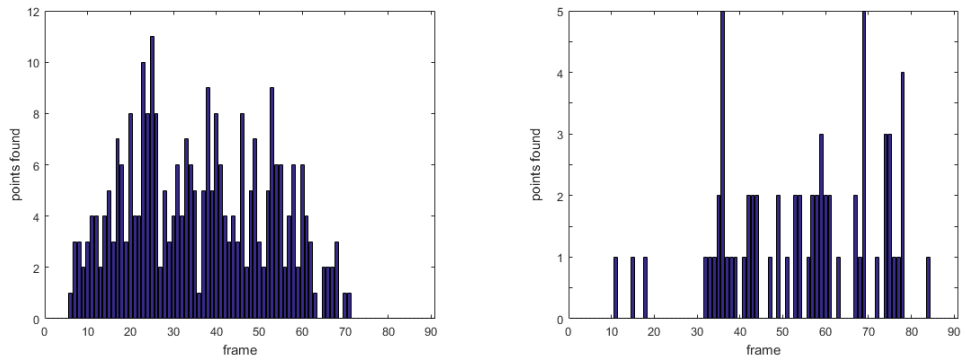


Figure 3.7: Distribution of interest points found over time (a) in the *walking* and (b) in the *che vuoi* sequences.

Figure 3.8: Salient frames selected for the *walking* sequence (KTH) .

Figure 3.9: 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 [RYS02]. Figures 3.10 and 3.11 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.

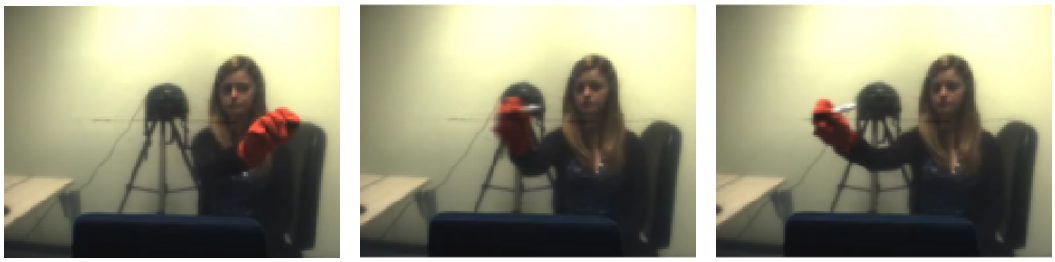


Figure 3.10: Frames corresponding to a change in action primitive on the *drawing line* sequence.

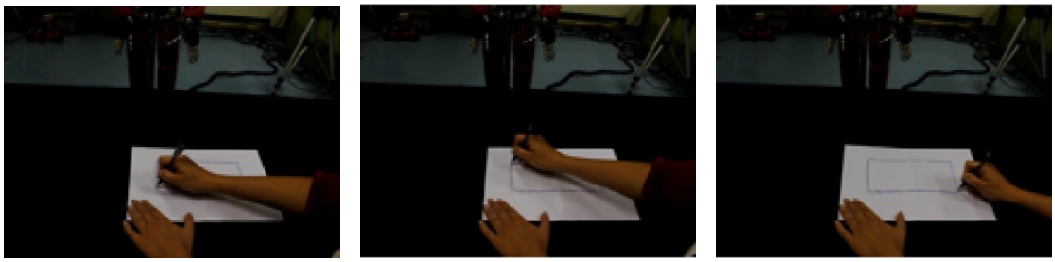


Figure 3.11: Frames corresponding to a change in action primitive on the *drawing rectangle* sequence.

3.3 Local Spatio-Temporal Representation

In the previous sections we have seen how Shearlet coefficients can be descriptive of the local spatio-temporal behavior of every point in a 2D+T signal. We can exploit this property of the Shearlet Transform to develop a novel method to represent a space-time point by means of the corresponding Shearlet coefficients, which entangle how the signal varies in its neighborhood. We carry out this procedure in a *dense* way, for the Shearlet Transform of a signal provides us with the coefficients associated for every point in the signal and for every value of the scale parameter j and for all the shearings in the sets $\mathbf{K}_j$ we are considering.

Our objective is to explore the possibility to better understand the nature of the spatio-temporal singularities which may arise in the 2D+T domain, as we described them in Section 2.4. Within this section we detail all the steps needed to build our representation and we show some brief results on the descriptive power of the approach we built.

3.3.1 Representation Analysis

The experiments we have discussed in Section 3.2 both on synthetic and real data showed us how sensitive the Shearlet coefficients are in correspondence of local changes in the signal, thus helping us to highlights singularities within it. We exploit this sensitivity to develop a local spatio-temporal description for every point.

In this section we describe a method which allows us to aggregate the local spatio-temporal information provided by the Shearlet Transform in order to enhance different types of discontinuities within a $2D + T$ signal.

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

$$\mathbf{K}_{\hat{j}} = \left\{ k = (k_1, k_2) \mid k_1, k_2 = -\lceil 2^{\hat{j}/2} \rceil, \ldots, \lceil 2^{\hat{j}/2} \rceil \right\}, \tag{3.1}$$

where $M = 2\lceil 2^{\hat{j}/2} \rceil + 1$ is the cardinality of $\mathbf{K}_{\hat{j}}$, where we suppressed the dependence on $\hat{j}$ from $\mathbf{K}_{\hat{j}}$ and M. The procedure we carry out in the discrete case is depicted in Figure 3.12 and consists of three 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, finally we reduce the dimensionality of our representation by aggregating meaningfully the information carried by the Shearlet coefficients. This representation should be meaningful of a specific space-time primitive. The following steps detail what is described in Algorithm 2.

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

(a) We reorganize the information provided by $SH[f](\ell, \hat{j}, k, \hat{m})$ in three $M \times M$ matrices, namely C_1, C_2 and C_3, each one associated with a pyramid $\ell = 1, 2, 3$, where each entry is related to a specific shearing. The association is given by the following formula, $C_\ell(r, c) = SH[f](\ell, \hat{j}, k_{rc}, \hat{m})$ with $\ell = 1, 2, 3$, where $r, c = 1, \ldots, M$ and k_{rc} is the corresponding shearing in $\mathbf{K}_{\hat{j}}$ defined in 3.1. As usual in this kind on analysis, we discard the information related to the Shearlet coefficients in the low frequency pyramid $\ell = 0$ since they are related to the smoothness of the signal. Figure 3.12 (a) 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 (Figure 3.12 (b)). For a given scale j and a fixed set of shearings $\mathbf{K}_{\hat{j}}$, the three matrices C_1, C_2, C_3 are tiled in a bigger matrix $\mathbf{C}$, which is then shifted both horizontally and vertically, so that 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 $k \in \mathbf{K}_{\hat{j}}$. This property is needed to obtain a rotation invariant representation in the next steps of this pipeline, since the values in $\mathbf{C}$ are redistributed similarly when considering two similar spatio-temporal primitives, 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 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 $\mathbf{C}$. We will see how different kinds of spatio-temporal elements can be associated with different kinds of local variations in $\mathbf{C}$.

2 - Compute a 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 in Figure 3.12 (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 3.12 (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 close 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 $\mathbf{C}$ 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 $\mathit{coeff}_{\bar{s}_i}$ 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 with each shearing $k_{\bar{s}_i}$. Then, we set

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

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.

3 - Derive a final reduced representation The representation $\mathbf{D}(\hat{m})$ entangles the relationships between the direction of maximum variation k_{max} for a given point $\hat{m}$ and the directions corresponding to the other shearings $k \neq k_{max}$, organized in squared rings of increasing side, see Figure 3.12 (c) 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 3.12 (e)). For example, the first entry of the vector $\mathbf{F}(\hat{m})$ is simply the Shearlet coefficient corresponding to k_{max} (the yellow pixel in Figure 3.12 (c)), 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 3.12 (c)), 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.

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.

Input: The coefficients $SH[f](\ell, j, k, m)$ for a video sequence f,
the selected scale $\hat{j}$ of the coefficients
Output: The Shearlet based representation $\mathbf{D}(\hat{m})$ for every point in the signal
for $\hat{m} = (x, y, t)$ ***in*** f **do**
  for $\ell \leftarrow$ *1* ***to*** *3* **do**
    C_ℓ = `GatherCoefficients` $(SH[f](\ell, \hat{j}, k, \hat{m}))$ Fig. 3.12(a)
  end
  $\mathbf{C}$ = `TileMatrices` (C_1,C_2,C_3)
  $\hat{\mathbf{C}}$ = `ShiftMatrix` ($\mathbf{C}$) Fig. 3.12(b)
  $\mathbf{D}(\hat{m})$ = `UnrollMatrix` ($\hat{\mathbf{C}}$) Fig. 3.12(c-d)
end

Algorithm 2: Calculation of the representation $\mathbf{D}(\hat{m})$.

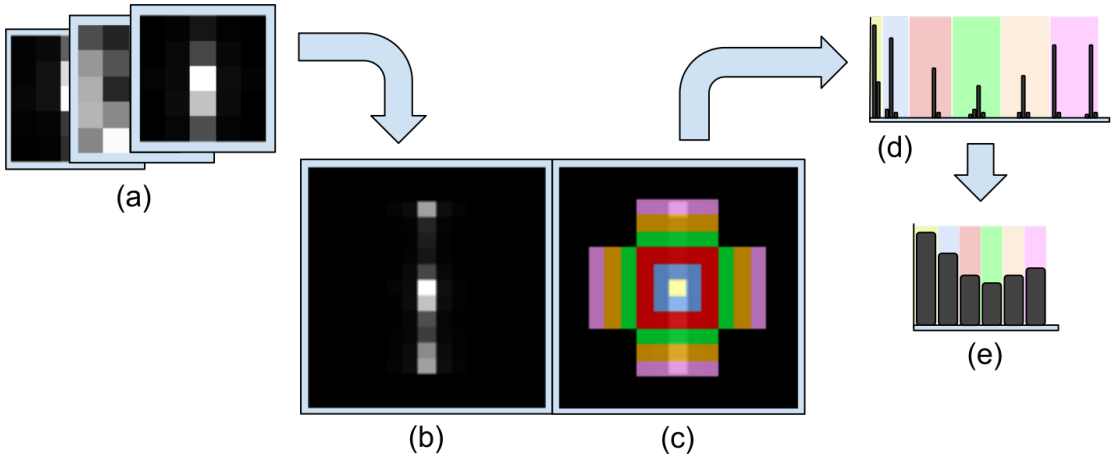
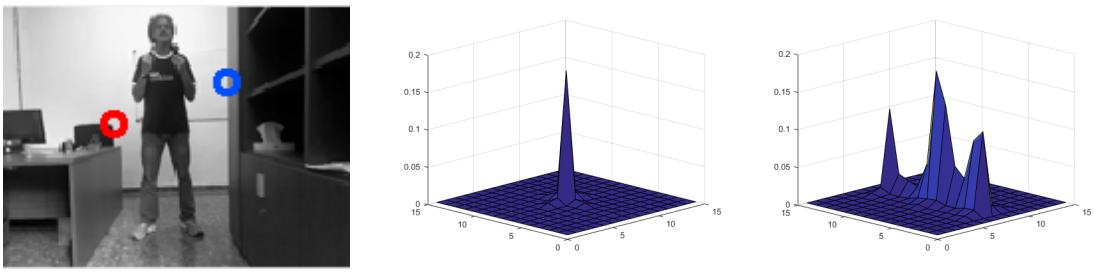


Figure 3.12: The main steps of the $2D + T$ signal representation procedure: (a) we compute matrices $C_1(r, c)$, $C_2(r, c)$ and $C_3(r, c)$, (b) we create the object $\mathbf{C}$, (c-d) we map subsets of elements (*i.e.* Shearlet coefficients) of $\mathbf{C}$ to different parts of a vector, (d) we obtain the representation $\mathbf{D}(\hat{m})$ for our point, finally (e) we create the compact descriptor $\mathbf{F}(\hat{m})$.

3.3.2 Geometrical Representation

At this point, the object $\mathbf{D}(\hat{m})$ entangles the relations between the direction of maximum variation k_{max} for a given point $\hat{m}$ and the directions corresponding to the other shearings $k \neq k_{max}$. Geometrically, it contains all the information provided by the Shearlet coefficients and related to how the signal changes around the selected spatio-temporal point.



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

Figure 3.13: 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.

Coherence with respect to Manually Detected Point Sets

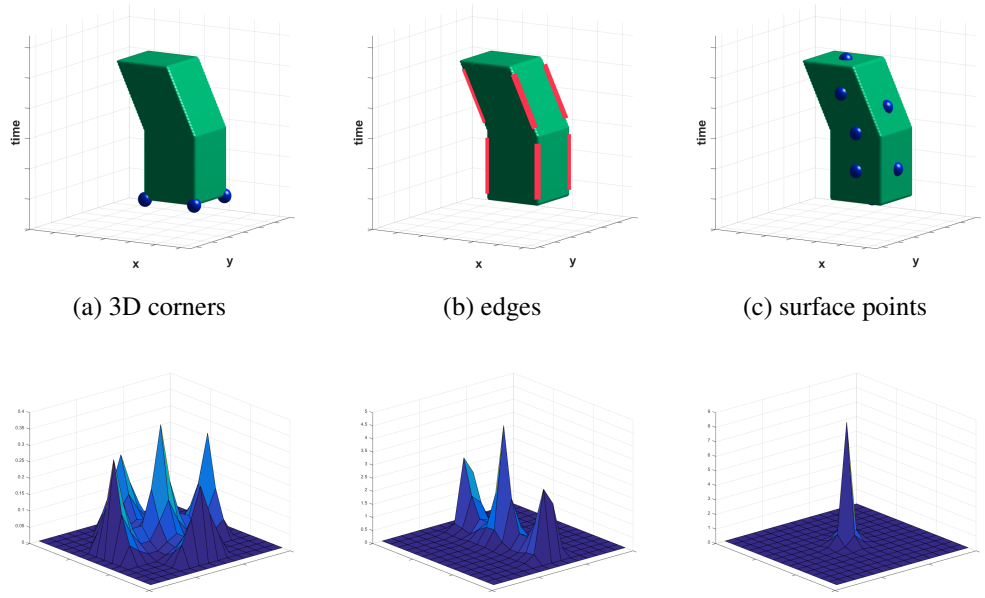
Figure 3.13 shows a possible way to visualize the values contained in the matrix **C** for two different points, the idea is to view the object **C** as a height-map so that to have an insight about the directions in which we found the highest variations (the visualization in Figure 3.13 (c) is the one corresponding to the object **C** shown in Figure 3.12 (b)).

The very simple synthetic sequence represented in Figure 3.2 contains three spatio-temporal features, which can be easily identified on the 3D shape: 3D corners, 3D 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 3.14 (a-c), while in Figure 3.14 (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 nature of different spatio-temporal features.

We now consider a real video from the KTH dataset [SLC04]. In the video sequence a subject is executing a *boxing* action, repeatedly moving his arms back and forth. In Section 2.3 (Figure 2.3) we introduced the idea of stacking the subject's silhouette as the action takes place, and this will come in handy now.

As in the case of synthetic data, we select points which are associated with a different spatio-temporal behavior and, for each of them, we compute our shearlet-based descriptor. The results can be appreciated in Figure 3.15, this time we sampled four points located on the red line in Figure 3.15 (b) to create the corresponding representation in Figure

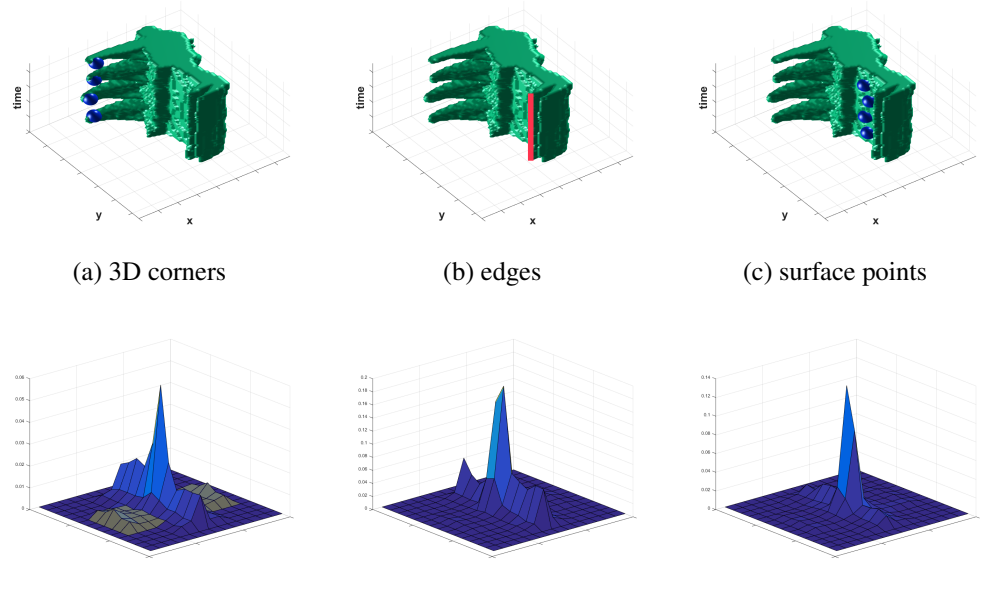


(a) 3D corners (b) edges (c) surface points

(d) (e) (f)

Figure 3.14: Examples of points on the 3D shape considered (a-c) and corresponding average shearlet-based representation (d-f).

3.15 (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)

Figure 3.15: 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 as in the synthetic case.

This can be seen in Figure 3.15 (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 3.13 (c), where the changes along the temporal dimension contribute for values lower than 10^{-3}).

Coherence with respect to Automatically Detected Feature Sets

As a further evidence we analyze the average $\mathbf{C}$ over sets of key points automatically detected by well know algorithms in image processing and computer vision. We consider two spatial features, edges [Can86][1] and corners [HS88][1] and a space-time feature, STIP [Lap05][2].

Edges. Figure 3.16 shows the average object $\mathbf{C}$ for all edge points obtained by the Canny detector applied to a 2D frame extracted from a video sequence. It is worth noting that, since the algorithm described in [Can86] (when applied to a single frame of a whole sequence) also detects elements which like corner points and moving edges (over time, but at a frame level, this behavior cannot be detected), the 3D visualization also includes small lateral peaks.

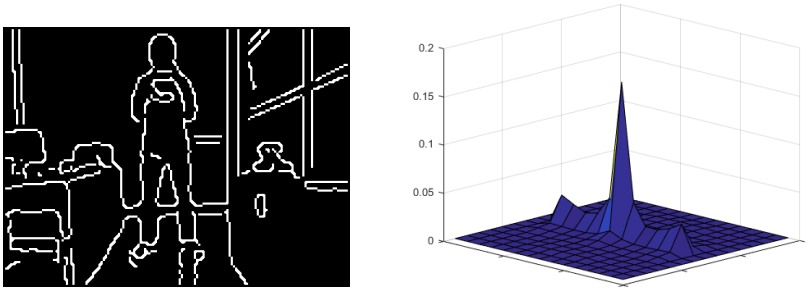
Corners. Figure 3.17 shows the behavior of corner points, automatically detected by the classical Harris algorithm. In this case we report the visualization for the subset of still and moving corners, which are more distinctive as expected, since our representation takes into account the space-time information which the Harris corner detector does not.

STIP. Figure 3.18 shows the average descriptor for the points selected by Laptev STIP detector on a different image frame. It is well known that the STIP detector identifies very few points, which are meaningful both in space and time. The

[1] we considered MATLAB's implementation of these algorithms

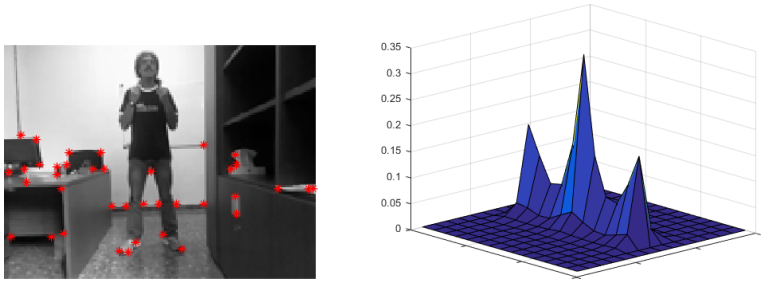
[2] implementation available at `https://www.di.ens.fr/~laptev/actions/`

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 strong peaks both in space and time directions.

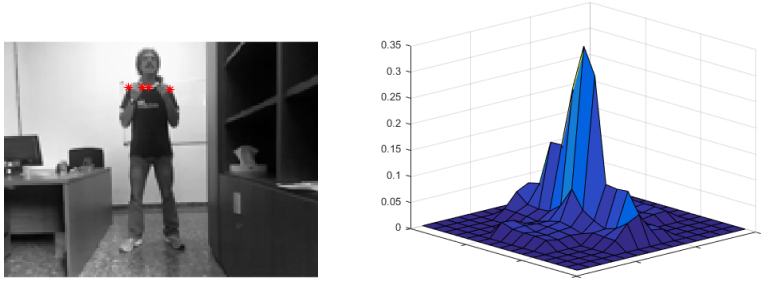


(a) edges mask (b) average descriptor

Figure 3.16: (a) Frame points automatically extracted by Canny edge detector; (b) a 3D visualization of C averaged on all the edge points.



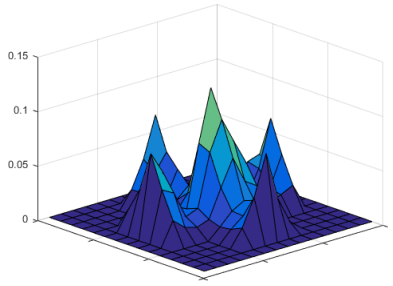
(a) still corners (b) average descriptor



(c) moving corners (d) average descriptor

Figure 3.17: 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.

(a) STIPs



(b) average descriptor

Figure 3.18: Laptev STIPs and a 3D visualization of $\mathbf{C}$ averaged on all the edge points.

3.3.3 Identifying Groups of Coherent Spatio-Temporal Primitives

In the previous section we saw how points with a different spatio-temporal behavior are actually characterized by shearlet coefficients which behave differently. Here we discuss how we can group sets of points by similarity, with the goal of identifying automatically possibly new types of spatio-temporal primitives.

To do so, we follow this method:

(a) we select a frame t in the sequence.

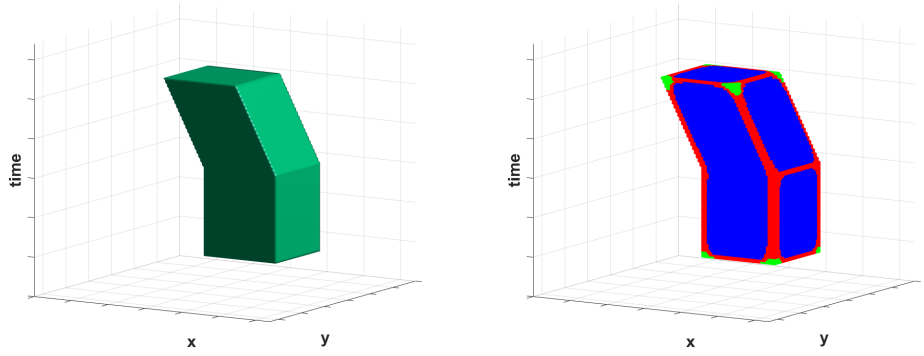
(b) we calculate the chosen representation ($\mathbf{D}(\hat{m})$ or $\mathbf{F}(\hat{m})$) for all the points at time t.

(c) we group the representations in p clusters via k-means.

We consider the resulting p cluster centroids as an unsupervised estimate of the space-time primitives which can be found within the selected video frame. The reason for we chose an unsupervised method (in particular, k-means) is that we do not have an a priori exact knowledge about the spatio-temporal primitives which may arise in the space-time domain. About these, we introduced several definitions in Section 2.3, and we consider the use of our shearlet-based representation to discover more about this, by means of a set of experiments.

As a first thing, we can again consider one of the synthetic shape we introduced before. We try to classify each point of its surface by calculating the distance between its representation $\mathbf{D}(m)$ and the three average representations in Figure 3.14, then each point is colored on the basis of the representation it is most similar to. The results are shown in

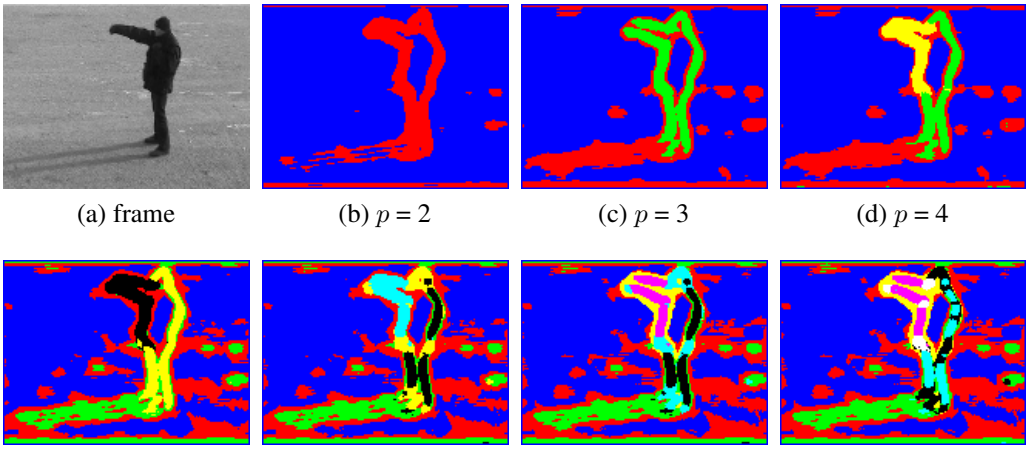
Figure 3.19, where it is possible to see the clear separation between the three kinds of spatio-temporal primitives.



(a) shape (b) classification

Figure 3.19: Example of classification of the surface points of our shape: surface points (blue), edges (red) and 3D corners (green).

Figure 3.20 shows the results obtained for different choices of p while clustering the points belonging to a single frame of a *boxing* sequence, Note that the color code used is not associated with the nature of each spatio-temporal primitive, while instead it represents the frequency of appearance of every given kind of primitive (the ordered color code is the following, from the most present to the rarest: blue, red, green, yellow, black, cyan, purple, white).



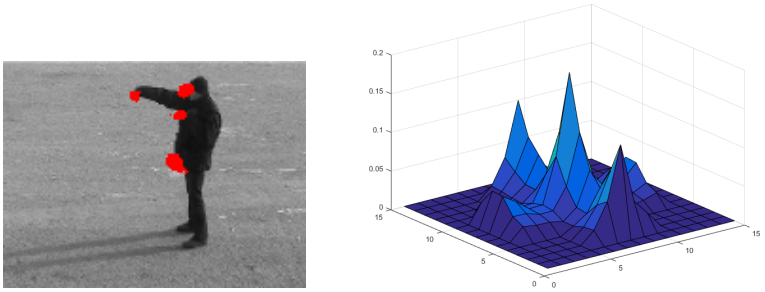
(a) frame (b) $p = 2$ (c) $p = 3$ (d) $p = 4$

(e) $p = 5$ (f) $p = 6$ (g) $p = 7$ (h) $p = 8$

Figure 3.20: Clusters of space-time primitives for different choices of p, the color code is unrelated with respect to the type of the primitive while it only depends on the cardinality of each cluster being shown.

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 — as in Figure 3.20 (a). 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 with respect to 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 subject is punching). These points are also highlighted in Figure 3.21, where we also show the average $\mathbf{C}$, and their similarity with the STIP points shown in Figure 3.18 is apparent.



(a) cluster # 8 from p=8 (b) average $\mathbf{C}$

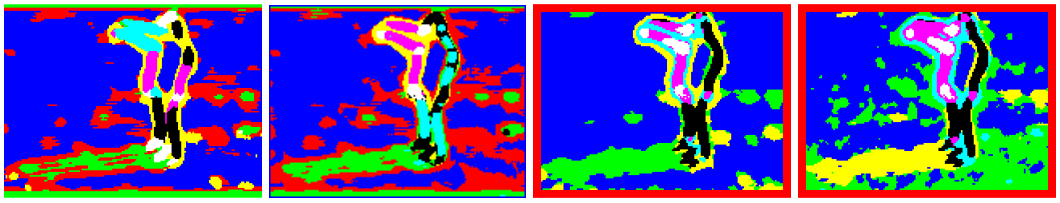
Figure 3.21: The points in the cluster corresponding to the most severe space-time variations, and the corresponding average object $\mathbf{C}$ (see text).

This result highlights many nice properties of our representation: 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.

The results shown here above are due to the analysis of the single representation $\mathbf{D}(\hat{m})$. We want to compare the results of the clustering process considering as input data both the original Shearlet coefficients and the compact representation $\mathbf{F}(\hat{m})$ we introduced in Section 3.2.2.

While carrying on this comparison, it can be noticed how in Figure 3.22 (a), where we show the result of the clustering process for the raw Shearlet coefficients, 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.

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 with respect to 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 3.22 (c)). In this last case

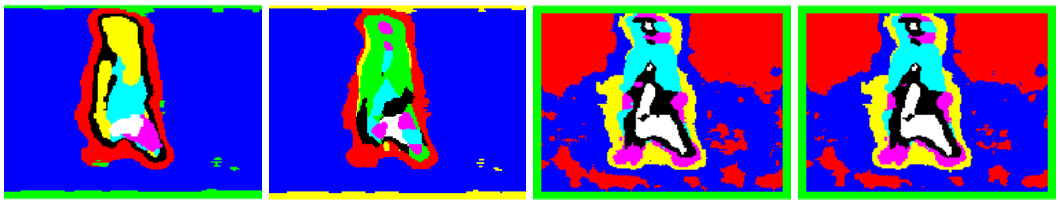


(a) $SH[f]$ (b) $\mathbf{D}(\hat{m})$ (c) $\mathbf{F}_2(\hat{m})$ (d) $\mathbf{F}_{2,3}(\hat{m})$

Figure 3.22: Results of a *k-means* clustering executed with k=8 clusters, on a *boxing* action: (a) using the Shearlet coefficients as they are provided by the 3D Shearlet Transform, (b-c) exploiting our representation and considering a single scale, (d) using our representation and considering the two finest scales available.

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 3.22(d)).

Similar considerations can be done by considering a different action, the *walking* one, from the same dataset. Figure 3.23 considers a specific instance of that execution. In the frame selected the subject is executing a stride, moving from right to left within the image frame. While carrying on a clustering process based on the original Shearlet coefficients (Figure 3.23(a)) and on the representation $\mathbf{D}(\hat{m})$ (Figure 3.23(b)) separates the subject from the background, it fails in characterizing the different parts of the shape, for it has a behavior which spreads more over time, with respect to the *boxing* action. The reduced representations (Figure 3.23(c-d)), instead, better capture the local nature of the primitives contained in the video sequence, by also highlighting elements with a more peculiar behavior, like the hands or the feet of the subject.



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

Figure 3.23: Results of a *k-means* clustering executed with k=8 clusters, on a *walking* action: (a) using the Shearlet coefficients as they are provided by the 3D Shearlet Transform, (b-c) exploiting our representation and considering a single scale, (d) using our representation and considering the two finest scales available.

3.3.4 Building a Dictionary to Encode Video Sequences

After evaluating the capability of our method in describing the local behavior of each point of a video sequence, we now assess its behavior over time to understand whether it can be used to describe the evolution of a movement within a sequence over time. The objective is to evaluate to which spatio-temporal primitive each point can be associated, and to consider how this association evolves over time and whether it may help us in understanding the kind of action carried on in the scene. The reason to build a dictionary of such primitives is also to understand whether exist commons spatio-temporal structures which can be found within different video sequences, and maybe in heterogeneous scenarios.

We start by introducing the concept of a *dictionary* of spatio-temporal primitives, so that to evaluate their presence within each frame of the video sequence being analyzed. The method is reminiscent of classical bag-of-words [CDF+04], for we look for representative elements within our signal so that to gather their distribution over frames and extract some meaning out of it. Our approach has also been described in [MGV+17], where we also carry on some additional experiments that can be found within the Conclusion and Further Work section at the end of this thesis.

Learning a dictionary of space-time primitives

The dictionary creation procedure is detailed in Algorithm 3, and can be described as follows:

(a) We consider a set of meaningful frames in a (set of) sequence(s) $\mathbf{V}$ (Figure 3.24(a)). The frames are chosen automatically through the spatio-temporal detection procedure we introduce previously. We select the N_f frames with the highest number of interest points, for a chosen scale $\hat{j}$, and we assume that these are the most representative of an action event.

(b) We represent each point $\hat{m}$ of every selected frame by means of $\mathbf{D}(\hat{m})$, for a fixed scale $\hat{j}$. On each frame, we apply k-means and obtain a set of K cluster centroids, which we use as space-time primitives or atoms (Figure 3.24(b)). The number K is chosen empirically, with the objective to obtain a set of centroids which is

meaningful and which captures the different interesting spatio-temporal behaviors within the sequences that we are considering.

(c) We re-apply K-means on all the previously obtained atoms. We end up with a dictionary $\mathcal{D}$ of N_a space-time primitives (Figure 3.24(c)). Here, N_a is chosen so that the dictionary is rich enough and at the same time to avoid too much redundancy in its components.

Input: A set of sequences $\mathbf{V}$, a number of frames to select N_f,
the number K of cluster to build per selected frame,
a number N_a of clusters in the final dictionary, the selected scale $\hat{j}$
Output: The dictionary $\mathcal{D}$ of spatio-temporal primitives
$\mathbf{C}$ = empty set of clusters
for v ***in*** $\mathbf{V}$ **do**
  $salient$ = `FramesWithMostInterestPoints` (v,N_f, $\hat{j}$) Fig. 3.24(a)
  for t ***in*** $salient$ **do**
    $clusters$ = `KMeansForPointsAtTime` (v,t,K, $\hat{j}$) Fig. 3.24(b)
    $\mathbf{C} = \mathbf{C} \cup clusters$
  end
end
$\mathcal{D}$ = `KMeans` ($\mathbf{C}$, N_a) Fig. 3.24(c)

Algorithm 3: The creation of a dictionary of spatio-temporal primitives.

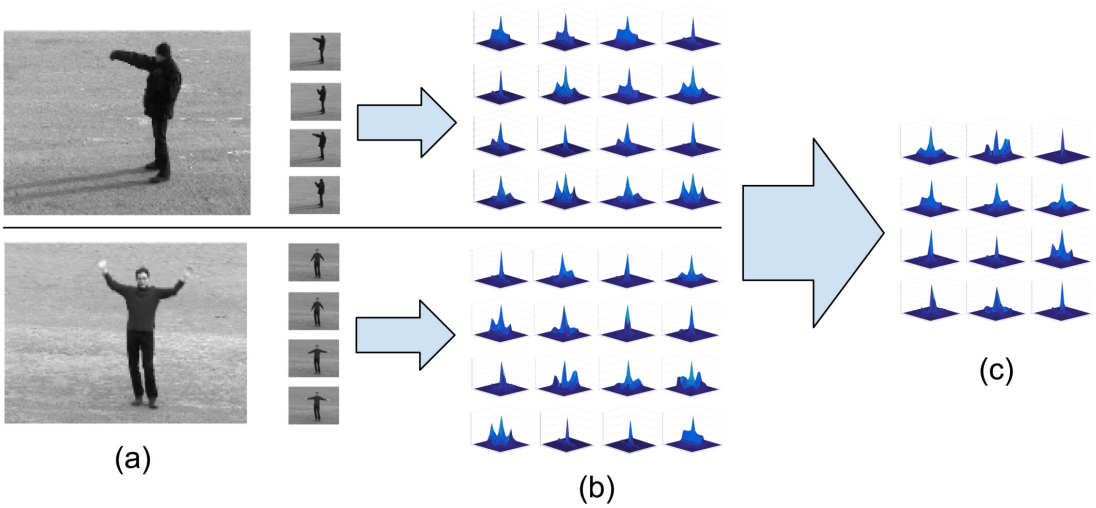


Figure 3.24: Learning the dictionary. (a) Automatic selection of meaningful frames from the training set; (b) Atoms learnt by each sequence; (c) Dictionary summarization on the whole training set.

Our assumption is that the elements resulting in the dictionary $\mathcal{D}$ are representative of the spatio-temporal primitives which characterize the set of sequences on which the creation

process is based. To evaluate such a descriptive capability, we have to consider it to understand how the primitives within a sequence evolve over time.

Encoding a video sequence with respect to a dictionary.

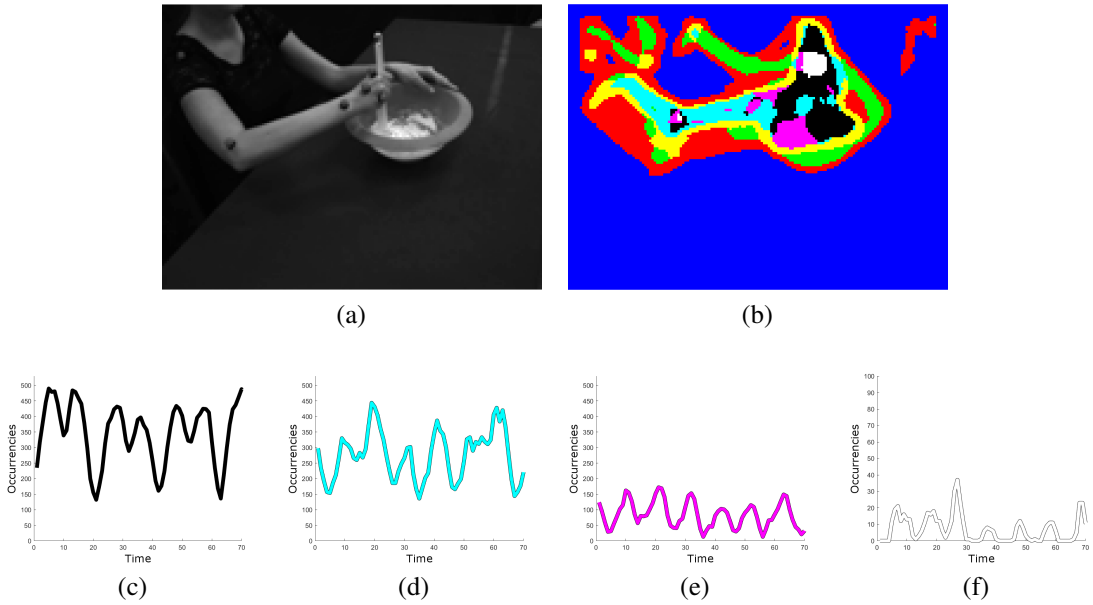
We now consider a sequence v representing a given action. The aim of this method is to build a description of what happens in a scene by combining the information that we are able to extract thanks to all the techniques that we have developed. To do so:

(a) For each image frame $I_t \in v$ we follow a bag-of-words approach and quantize points of I_t w.r.t the dictionary atoms in $\mathcal{D}$, obtaining F_i^t frequency values (how many points in frame I_t can be associated with the i-th atom).

(b) We filter out still primitives that are not useful to our purpose. To do so, we exploit the information about the motion happening in a scene carried by the Shearlet coefficients (this technique will be introduced in detail in Section 3.4). Finally, we compute temporal sequences of frequency values across time, obtaining N_a time series or profiles $\{P_j\}_{j=1}^{N_a}$, which summarize the content of the video sequence.

In Figure 3.25 we show an example of the result of this procedure. Figure 3.25 (a) represents a sample frame at time t of a *mixing* sequence from a dataset of in-house recorded cooking actions, and Figure 3.25 (b) shows a color-coded example resulting after associating every point at time t with one of the atoms in the reference dictionary $\mathcal{D}$. Several elements are associated with points belonging to the background, or to structures with a less interesting behavior. However, it is possible to display the occurrence of only a subset of the atoms in $\mathcal{D}$, and Figures 3.25 (c-f) show the profiles associated with the 4 less frequent atoms, showing how their periodic behavior might highlight some characteristics of the movement being executed.

The reason for we show only the less occurring atoms in Figure 3.25 is that the most frequent ones are usually related with the most present primitives in the sequence, which also are the one which characterize the less the movement occurring in the scene. This statement also holds for the sequence displayed above, since the most frequent points are the ones characterized by a non interesting behavior (like background, still elements, or far away from the strongest edges in each frame).



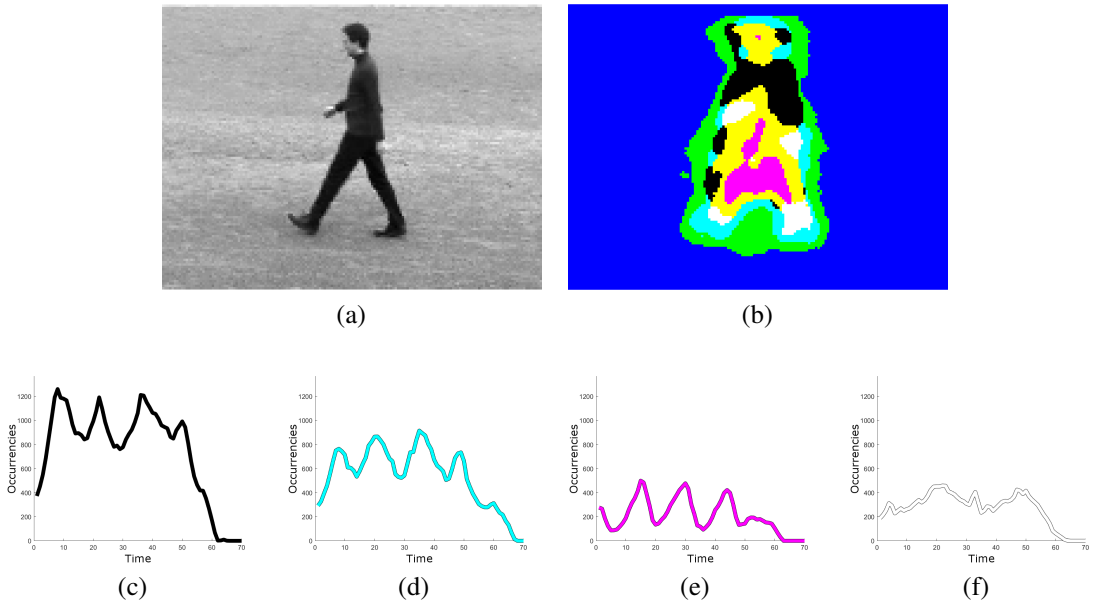
(a) (b) (c) (d) (e) (f)

Figure 3.25: Encoding of a *mixing* action: (a) A sample frame; (b) The quantization with respect to the dictionary $\mathcal{D}$ atoms; (c-f) Examples of temporal profiles (see text for details).

As additional examples, we also consider two sequences from the KTH dataset, a *walking* and an *handwaving* one, and we follow the same approach. Figures 3.26 and 3.27 show the results for this second experiment. It is possible to see the profile in Figure 3.26 (f) is the less representative of the periodicity of the strides within the walk, similarly to the profile in Figure 3.27 (d) for the *handwaving* case. In this last situation, it is also possible to see how the primitives represented by profiles P_7 and P_8 follow an antagonist behavior, with one raising when the other is decreasing.

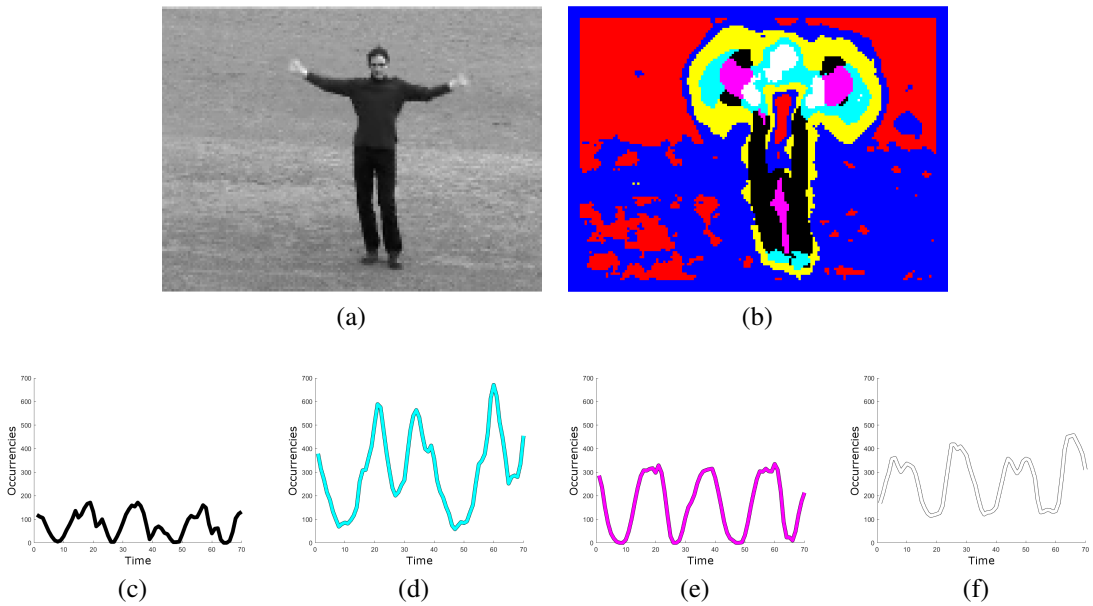
In the two previous examples, we considered two different dictionaries. In the first case, we only considered sequences from the cooking actions dataset for the creation of $\mathcal{D}$, while for the other two videos we have used samples from the KTH dataset for the construction of the dictionary.

These experiments showed us that our representation and the use of dictionaries of spatio-temporal built on top of it could be meaningful to represent the dynamic events which are taking place within video sequences. Each spatio-temporal primitive is associated with a specific spatio-temporal behavior of different moving parts belonging to the subject executing the actions. We further explore this in [MGV+17], where we consider the use of dictionaries of shearlet-based primitives to compare actions coming both from the same dataset the dictionaries are created from, and from sequences belonging to



(a) (b) (c) (d) (e) (f)

Figure 3.26: Encoding of a *walking* action: (a) A sample frame; (b) The quantization with respect to the dictionary $\mathcal{D}$ atoms; (c-f) Examples of temporal profiles (see text for details).



(a) (b) (c) (d) (e) (f)

Figure 3.27: Encoding of an *handwaving* action: (a) A sample frame; (b) The quantization with respect to the dictionary $\mathcal{D}$ atoms; (c-f) Examples of temporal profiles (see text for details).

different datasets and scenarios. In those preliminary experiments, we notice that a set of spatio-temporal primitives created on top of a given dataset can be used to describe meaningfully also sequences belonging to a different setting.

3.4 Motion Estimation

To estimate the motion taking place within an images sequence means to understand how a pixel's position evolves over time. This procedure is the first step towards more complex motion analysis tasks, like tracking[CZLK98, SES12], segmentation[SM98, MP02], structure from motion[LH81, KVD91] or 3D motion estimation[AKW$^+$17]. We want to explore how much information about the motion happening within the scene we are able to extract only starting from the information provided by the Shearlet Transform.

We introduced the geometrical meaning of the Shearlet coefficients in Chapter 2, and we ground the development of our motion estimation algorithm on that. In the following we explain what are the elements of our signal for which we can estimate the motion, showing also some brief results on existing datasets.

The Shearlet coefficients associated with every spatio-temporal point give us information about how the point varies with respect to its neighborhood, but we focus only one of the three dimensions, the temporal one, since it contains the information related to *what is changing* within the sequence, which is strictly related to dynamic information, and may help in motion estimation.

3.4.1 Normal Flow

In this section we summarize the main background concepts, necessary to understand the reminder of the section. The interested reader is referred to text books such as [TV98]. We focus on the problem of estimating the motion field from image sequences. To do so, we start by introducing one of the key assumptions in this scenario, which is that the appearance of the image patches within the frame do not change over time. This assumption is called the **image brightness constancy equation**, and is defined as:

$$\frac{d\,I}{d\,t} = \frac{d\,I(x,y,t)}{d\,t} = 0 \tag{3.2}$$

where the image I is regarded as a function of three coordinates, two spatial and a temporal one. By applying the chain rule of differentiation, we rewrite

$$\frac{d\,I(x,y,t)}{d\,t} = \frac{\partial\,I}{\partial\,x}\frac{d\,x}{d\,t} + \frac{\partial\,I}{\partial\,y}\frac{d\,y}{d\,t} + \frac{\partial\,I}{\partial\,t} = 0 \tag{3.3}$$

The partial spatial derivatives $I_x = \frac{\partial I}{\partial x}$ and $I_y = \frac{\partial I}{\partial y}$ are nothing else than the components of the gradient ∇I of the image. The components $u = \frac{d\,x}{d\,t}$ and $v = \frac{d\,y}{d\,t}$ are called the **motion field**, which we denote as $\mathbf{u}$, they are the *unknown variables* which we want to calculate and the aim is to estimate them as most precisely as possible. We can now write:

$$\begin{aligned} I_x u + I_y v + I_t &= 0 \\ \nabla I^T \mathbf{u} &= -I_t \end{aligned} \tag{3.4}$$

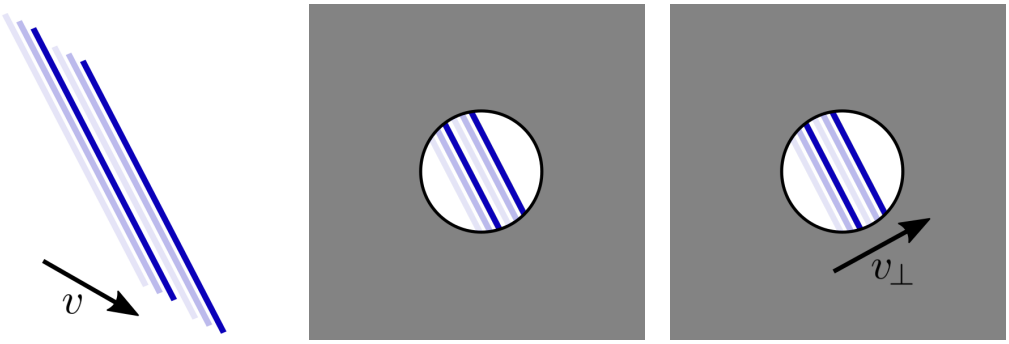
By isolating the known and unknown components of the Eq. (3.4):

$$\frac{\nabla I^T \mathbf{u}}{\|\nabla I\|} = -\frac{I_t}{\|\nabla I\|} = v_\perp. \tag{3.5}$$

Thus the only components of the motion field which can be estimated are the ones along the direction of the spatial image gradient, which we denote $v_\perp$, since we are trying to calculate two unknowns with the help of a single equation [TV98].

3.4.2 Shearlet-based Normal Flow Estimation

The algorithm we developed is based on the formalisms that we introduced in Chapter 1. For the sake of this explanation, we consider a rigid body (shown in Figure 3.29(a)) which describes a 3D volume while moving over time. Recalling the results shown in Section 1.2, we can relate to any shearing vector $k^* = (k_1, k_2)$ the direction (without orientation) of the normal vector $\mathbf{N}$ to the surface, for any point belonging to it. If we consider the body in Figure 3.29, we can parametrize unit vector corresponding to $\mathbf{N}$ (which we recall is the normal vector to the surface at spatial-temporal point $\sigma(s,t)$) by



(a) (b) (c)

Figure 3.28: The aperture problem: (a) two bars and their real velocity v, (b-c) the apparent velocity $v_\perp$ once the information about the structure of the two objects is limited.

the *latitude* angle α and *longitude* angle β as

$$\mathbf{N} = \pm\|\mathbf{N}\|(\cos\alpha\cos\beta, \cos\alpha\sin\beta, \sin\alpha) \tag{3.6}$$

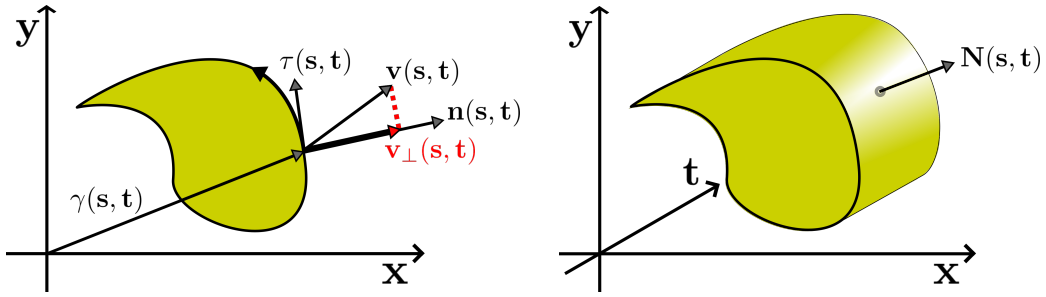


Figure 3.29: (left) A body at time t with the main relevant geometrical and dynamical quantities, the quantity highlighted in red is the projected component of the velocity on the normal vector $\mathbf{n}(s,t)$), (right) the surface spanned over time by the movement of the body and the normal vector on the surface $\mathbf{N}(s,t)$.

Eq. (2.2) shows that the xy-component of $\mathbf{N}$ is the normal vector $\mathbf{n}$ to the boundary of the body, so that it depends only on the geometry of the object, whereas the t-component of $\mathbf{N}$ is

$$\mathbf{N}(s,t)\cdot\mathbf{k} = -\mathbf{v}(s,t)\cdot\mathbf{n}(s,t) = -v_\perp(s,t), \tag{3.7}$$

we recall the definition of $\mathbf{N}$ from Eq. (2.2)

$$\mathbf{N}(s,t) = \mathbf{n}(s,t) + \tau(s,t)\wedge v(s,t) \tag{3.8}$$

i.e. $v_\perp(s,t)$ is the normal component of the velocity, up to a sign. Hence it depends both on the geometry and the dynamics, and the sensitivity of the Shearlet Transform to orientation can be used to extract both geometric and dynamic informations. More precisely, denoted by α the latitude angle of $\mathbf{N}$ as in (3.6), then

$$\sin\alpha = -\frac{v_\perp(s,t)}{\sqrt{1+v_\perp(s,t)^2}} \tag{3.9}$$

from which we derive

$$v_\perp(s,t) = -\frac{\sin\alpha}{\sqrt{1-\sin^2\alpha}} \tag{3.10}$$

so that it is possible to reconstruct the normal velocity from the direction of $\mathbf{N}$, whereas the longitude angle β of $\mathbf{N}$ gives the direction of the normal vector $\mathbf{n}(s,t)$.

From an algorithmic point of view, we fix the scale j on the basis of the sequence chosen, with respect to the characteristics of the motion taking place, be it represented by large scale movements or small ones. Then, for every spatio-temporal point $\hat{m}$ in the video sequence, we choose the shearing $k^* = (k_1, k_2)$ associated with the maximum absolute value of the Shearlet coefficient vector $\{SH[f](\ell, j, k, \hat{m}) \mid \ell = 1,2,3, k \in \mathbf{K}_j\}$. According to Eq. (1.12), k^* defines two angles, the latitude $\alpha(k^*) \in [0,\pi]$ and the longitude $\beta(k^*) \in [-\pi,\pi]$, associated with the normal vector $\mathbf{N}$ to the surface at the point $\hat{m}$. Hence we get

$$\mathbf{n}(s,t) = (\cos\beta(k^*), \sin\beta(k^*)) \qquad v_\perp(s,t) = -\tan\alpha(k^*). \tag{3.11}$$

We note that Eq. (3.8) implies that $\mathbf{N}$ is oriented "inside" the volume generated by the evolution of the body and this allows us to choose the correct sign in Eq. (3.6) and, hence, in Eq. (3.11). We want to stress the fact that, with this formulation, we are only able to estimate the velocity component $v_\perp(s,t)$ projected on the normal vector $\mathbf{n}(s,t)$. This limitation is related to the kind of information the Shearlet coefficients are carrying, and it is similar to the situation in which we rely solely the assumption leading to the image brightness constancy.

Input: The coefficients $SH[f](\ell, j, k, m)$ for a video sequence f,
the selected scale $\hat{j}$ of the coefficients
Output: The estimate of the direction and of the magnitude of the motion for every point in the signal
for $\hat{m} = (x, y, t)$ ***in*** f **do**

$$k^*, \ell^* = \text{argmax}_{k,\ell}\ SH[f](\ell, \hat{j}, k, \hat{m})$$
$$\hat{k}_1, \hat{k}_2 = \texttt{NormalizeShearings}\ (k^*)$$
$$num = 2^{-\frac{j}{2}} * \hat{k}_2$$
$$den = \sqrt{1 + 2^{-\frac{j}{2}} * \hat{k}_1^2}$$
$$\alpha = \arctan\left(\frac{num}{den}\right)$$
$$\beta = \arctan\left(2^{-\frac{j}{2}} * k_1\right)$$
$$\hat{\alpha}, \hat{\beta} = \texttt{AdjustByPyramid}(\alpha, \beta, \ell^*)$$
$$direction(\hat{m}) = (\cos(\hat{\beta}), \sin(\hat{\beta})) \qquad \text{Fig. 3.30(b)}$$
$$magnitude(\hat{m}) = -\tan(\hat{\alpha}) \qquad \text{Fig. 3.31(c)}$$

end

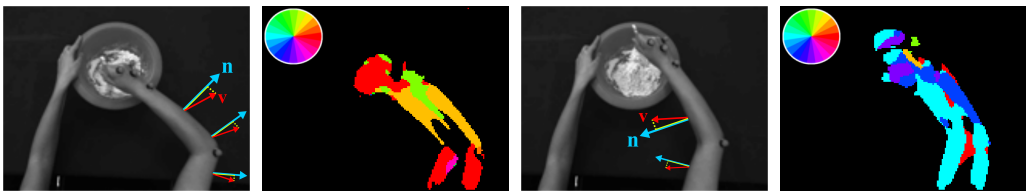
Algorithm 4: Estimation of the motion.

The above formulas hold if the maximum belongs to the pyramid with $\ell = 1$, otherwise one needs to perform a suitable rotation. Equations (3.11) show that it is possible to recover the orientation of the boundary and the normal component of the velocity directly from the Shearlet coefficients without any further processing of the representation.

3.4.3 Experimental Assessment

To provide an intuition of the applicability of our estimates to real world scenarios, Figures 3.30(b,d), 3.31(b), 3.32(b,d) and 3.33(b) report color-coded maps of the estimated normal flow $v_\perp$. In these visualizations, the estimated direction of motion for every point is represented with a different color taken from the color wheel on the upper left of each picture, where every color is associated with a specific direction of motion. For Figures 3.30, 3.32 and 3.33 we have considered the contributions coming from scale $j = 2$, while for the sequence in Figure 3.31 and Figure 3.34 we have used the coefficients calculated at scale $j = 3$. The reason for this choice is due to the intrinsic nature of the movement represented in the sequences we considered, since in the motorway scenario the movements are executed at a different pace with respect to the other two cases, thus the choice of a finer scale, more sensitive to high frequency space and time changes.

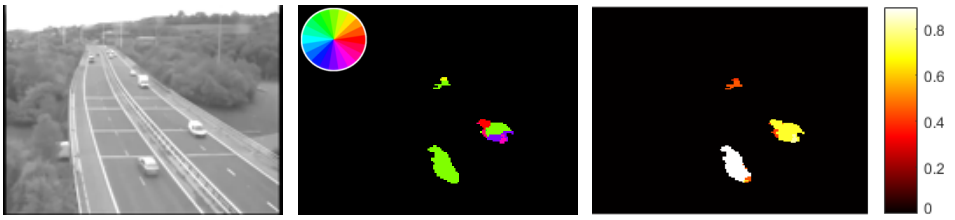
Since we want to understand the capability of our method to extract relevant motion information, we carry on a few experiments to evaluate how precisely our approach can estimate the *direction* and the *intensity* of the apparent motion in the scene.



(a) (b) (c) (d)

Figure 3.30: Two frames (a) and (c) from the *mixing* sequence and overlaid examples of the corresponding motion taking place; (b) and (d) show the motion estimated along the normal for each point $v_\perp$, with direction and orientation color coded with respect to the color wheel on the top left of each map.

In Figure 3.30 the recorded subject is mixing flour in a bowl, executing a circular movement with her right arm, while holding a spoon. Here it is possible to see one of the characteristics of our approach. For the way we analyze the Shearlet coefficients, we are able to extract some information about the motion occurring only in presence of spatio-temporal singularities. The inner uniform part of the arm, which lacks texture is depicted as black in Figures 3.30(b), meaning it has not been possible to extract any motion information about it.

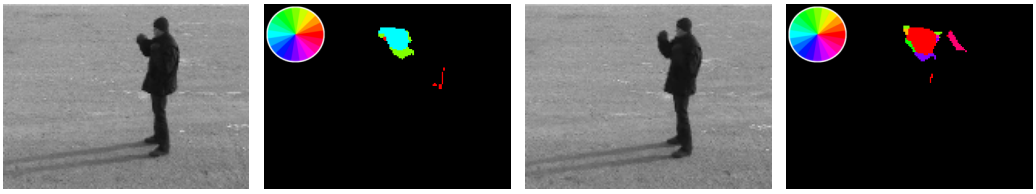


(a) (b) (c)

Figure 3.31: (a) A frame from the *motorway* sequence; (b) our approach segments the moving parts in the image; (c) it also gives an estimate about the speed of the different objects (a brighter object is associated with a higher speed).

A very nice result is the one shown in Figure 3.31 (c). In this motorway sequence, several cars are moving far away from the camera, each one crossing the scene at a different speed. Here our approach has not been only able to detect correctly which parts of the scene are moving (in this case, the cars), but also we could roughly estimate the speed of the elements. So objects which are actually both quick and near to the camera are

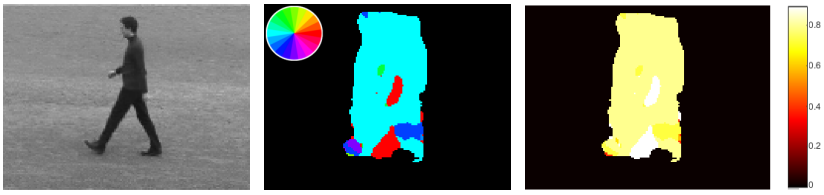
shown in bright colors, meaning an higher speed (please refer to the color map in the figure). Further object, appearing slow and moving toward the camera, are shown in darker colors, meaning a lower velocity.



(a) (b) (c) (d)

Figure 3.32: Two frames from the *boxing* sequence and the corresponding estimated motion; the two frames look similar, but in the first one the subject is extending his arm while in the other one he his contracting it, our method correctly labels the two set of points differently.

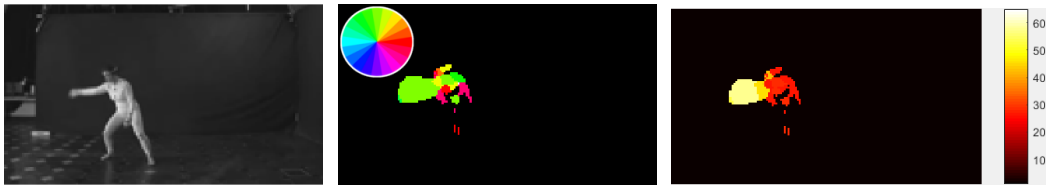
In Figure 3.32 we show how our method correctly labels two similar groups of points as "moving leftwise" or "moving rightwise", respectively the cyan and green ones in Figure 3.32 (b) and the red and purple ones in Figure 3.32 (d). By focusing on the frame in Figure 3.32(a-b), it is also possible to notice the sensitivity of our method, which is also able to capture and show the motion happening to the rear side of the jacket, which is oscillating while the subject is punching in the air.



(a) (b) (c)

Figure 3.33: (a) A frame from a *walking* sequence; (b) the estimated direction for the different parts of the subject; (c) the estimated speed for the subject walking.

Figure 3.33 represents a walking action from the KTH dataset, with a subject crossing the scene from right to left with a few strides. Here it is possible to see how different parts within the body, moving in different directions, are color coded accordingly. In Figure 3.33(b) the body is colored in cyan, since it is moving on toward the left side of the scene, the front foot is stomping on the ground, and it is moving downward, thus is it colored in violet. The knee of the leg on the back is slowly advancing and moving downward, so it is represented in blue.



(a) (b) (c)

Figure 3.34: (a) A frame from a *dancing* sequence; (b) the estimated direction for the different parts of the subject; (c) the estimated speed for the dancer.

Finally, Figure 3.34 shows a sample sequence in which a dancer is waving is body left and right, continuously. In this particular frame, the subject is raising her right arm above her head, while slowly starting to move her body on her left. As a consequence, the arms is drawn in green, representing a movement on the upper direction, while the body is painted in red, for it is starting a movement on the right part of the image. The magnitude of the movement is also estimated correctly, with the arm shown to be quicker than the body of the dancer.

We want to focus on the fact that the information that we are extracting from these video sequences can be rough, but the way we are analyzing these signals makes everything trickier. We have to remember that, while estimating the speed of one of the cars in the previous pictures, we are actually analyzing a three-dimensional volume, checking how things evolve *inside* it. However, the descriptive power of the Shearlet Transform and the framework that we have developed shows how it is possible to extract a very rich amount of information from this single decomposition.

Conclusion and Further Work

Within this thesis, we have explored the capabilities of the three-dimensional Shearlet Transform to extract meaningful information about a video signal. Our aim has been to understand the level at which it is possible to inspect a given sequence only through the information carried by its Shearlet decomposition.

What we have seen is really encouraging. We have successfully used the 3D Shearlet Transform to carry out different tasks, by developing novel algorithms, inspired by the previous works made in the field. The information extracted also helped obtain a better understanding about the analysis of spatio-temporal primitives. Our initial formalization of what happens in the 3D domain has been confirmed and validated by our experiments both on synthetic and on real world sequences.

The sensitivity to anisotropic singularities of the Shearlet Transform has been exploited in several ways.

At first, we have considered it as a mean to look for spatio-temporal meaningful points in a classical feature detection task. This led to a preliminary approach aimed at selecting a sparse set of points from a whole video sequence, so that to understand which elements could have been key into understanding what is represented in the scene.

Since the results we obtained showed us the potential of such an analysis, we moved on and developed a novel way to represent a spatio-temporal point with respect to its behavior in the space-time domain. This helped us to understand better the spatio-temporal which may arise in the space-time domain, spanned by the spatial elements which evolve over time within the scene. In this direction, we have seen that our shearlet-based representation was really powerful in describing which spatio-temporal elements appear in the sequences we analyzed. Also, this representation showed some stability,

giving us the ability to recognize patterns related to how these structures appear and disappear over time.

During the experiments related to the representation that we have developed, we also noticed that we were able to separate coherently elements which belong to structures which are moving from those which are not. This gave us an insight, which finally evolved into an algorithm for the estimation of the motion taking place into a scene. While afflicted by some limitations, this approach showed some really nice results while considering different sequences drawn from different scenarios. The capability of the Shearlet decomposition to describe the direction and the magnitude characterizing a movement exceeded our expectations, and opened the door for further developments.

Even if our achievements are satisfactory, further work could extend the results that we obtained within this thesis. In the next sections, we highlight some of the limitations of our method which we discovered carrying on the work for this thesis, and we propose which paths could be followed to further improve the results that we have obtained.

Limitations of Our Approach

Computing the Shearlet Transform of a video signal is an expensive operation, in terms of both memory space and time required. First, to calculate the Transform one needs to have available the whole sequence to be analyzed, since the calculation of the coefficients regarding a given point $m = (x, y, t)$ is carried out by both considering previous, current, and future information (*i.e.* frames in the sequence). This brings to the fact that the whole video sequence has to be considered within each of these calculations, and since each operation involves the calculation of a few forward and backward 3D Fourier Transform (see [KLR16] and the available code for details) it is trivial to see how the memory and space requirements of the whole computation explode easily.

For the same reason, another issue of this whole approach is that the computations can not output information in real-time, for to calculate the information at time t both previous and future frames are needed.

Moreover, computing all the Shearlet coefficients for a whole video is way too much memory consuming. Thus, we had to *slice* long sequences in subsequences, a few seconds

long. This is needed to contain the space required in memory to store all the results of the decomposition, this making the calculation feasible.

We put a lot of effort into improving the speed at which the computations are carried out (see Appendix for details), by understanding better the nature of the framework we have been using and by optimizing the source code for matrix-matrix operations (a "good practice" when developing code in the MATLAB environment). However, the calculations involved in the Shearlet decomposition of a 3D signal are time costly, keeping our method far away from real-time performance.

Studying the use of Spatio-Temporal Dictionaries

In [MGV+17] we considered the use of the dictionaries of spatio-temporal primitives that we introduced in Section 3.3.4 to gain more insights about them. In that work, we mainly used a dataset of cooking actions acquired by Vignolo and colleagues [Vig] and we carried on a set of experiments to evaluate preliminarily some properties of our shearlet-based representation.

First, we wanted to understand how a dictionary of primitives learnt on a given set of sequences could be meaningful in representing the spatio-temporal primitives (and their evolution) on another set of videos. To do so, we considered a dictionary $\mathcal{D}$ built from instances drawn from the KTH dataset [SLC04] to classify spatio-temporal points and to calculate their shearlet-based representation when belonging to sequences drawn from a different dataset. In particular, we used $\mathcal{D}$ to classify and describe the evolution of points in sequences from the Cooking Action dataset. Again, the results have been encouraging, with particular instances of actions which have been better represented with respect to other ones.

Another characteristics that we wanted to evaluate in [MGV+17] was the invariance of the spatio-temporal primitives learnt while changing point of view on the scene. The Cooking Action dataset is equipped with 3 different points of view for every instance of an action, thus we had the chance to understand a little bit more about the fact that while a given primitive looks in a particular way -while looking at the scene from a set point of view- it may look differently when changing the point of view of about 90 degrees. Within these experiments, we only considered a subset of the actions available, but we obtained promising preliminary results which opened the door for further investigation.

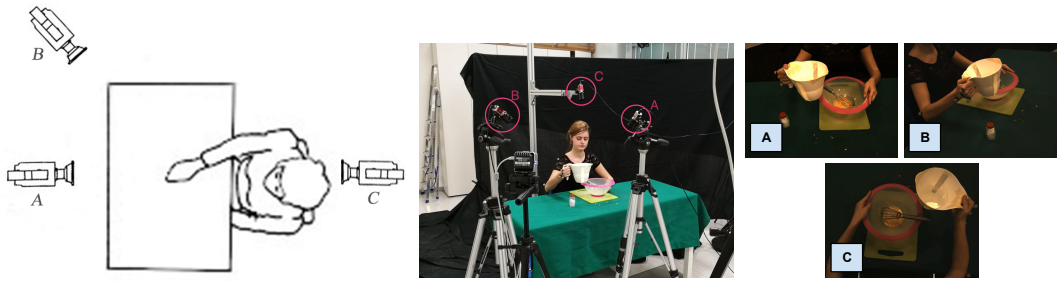


Figure 3.35: Acquisition setup for the Cooking Action dataset, where for each recordings three synchronized tracks are available, each one associated with a different camera displayed here.

Different Ways to Learn Meaningful Features

We considered a single method to build what we have called a *dictionary* of spatio-temporal primitives. While the results we have obtained are promising [MGV+17], the approach we developed can be further refined to keep into account alternative methods to learn a sets of meaningful features.

This can be done at two levels, when:

- learning a subset of **representative primitives**.
- analyzing the **evolution** over time of the **profiles** representing the number of points associated with the primitives belonging to the dictionary of reference.

In the first case, the dictionary learning has been carried out in a completely unsupervised way. The reason we followed such an approach is due to the fact that we did not have an a priori knowledge about the nature of the structures which may belong to such spatio-temporal signals. We formalized a subset of elements, as we did in Section 2.3, but we were unaware of the results that we might have achieved.

While discovering the primitives arising in the 2D+T domain, we started to notice some patterns. The strength of our shearlet-based representation is to coherently represent similarly elements which behave in a similar way in space-time. Thus, the results we obtained showed us how different learnt primitives could be associated with the same "kind" of spatio-temporal structure, but behaving in a slightly different way. As

an example, strong edges (*i.e.* sharp ones, characterizing an high frequency change in the signal) can have to the same shearlet-based representation of soft ($\sim$smooth) ones, the only difference lying on the absolute value of the coefficients composing the representation, or the coefficient scale at which the primitive turns out to be highlighted the most.

For the second entry of the previous list, as we experimentally showed in Section 3.3.4, repetitive actions show some interesting patterns in the evolution of the profiles of occurrence of each primitive within the dictionary. One step ahead in this direction could be to employ dictionary learning based techniques to learn sub-atoms within the evolution of these profiles, so that to be able to develop a "global" description of what is happening in the scene. For example, the sequences contained in the KTH dataset contain repetitive, periodic movements which reflected onto precise patterns in the evolution of the associated primitives distribution profiles. A more in-depth study of these behavior could lead to an approach able to segment and recognize the single instances of each movement, bringing our ability of analysis to the next level, for example the one of understanding what is happening in the scene.

An Improved Way to Estimate Motion

The method to estimate motion that we have developed within this thesis has a few limitations. First of all, it only produces the normal flow even for points that are not affected by the aperture problem, allowing us to estimate solely the velocity component along the normal to the spatio-temporal structure each given point is lying onto. Secondly, our method is very local, since at each time, to estimate the normal flow at a given point $\hat{m}$, we only consider the information provided by the shearlet transform (*i.e.* the coefficients) for the point $\hat{m}$. In this way, we can not overcome the problem of being unable to estimate any motion component beside the one along the normal at the point.

One solution could be to combine for every point the information given us by the representation $\mathbf{D}(\hat{m})$ and from the motion estimation algorithm. We know that there are a set of points for which it is possible to give more confidence to the motion estimated for them [LK$^+$81, TV98]. We find ourselves in a similar scenario, for there are spatio-temporal points where it is surely harder to give a detailed and precise information about the kind of motion they belong to. Other points, like *spatio-temporal corners*, have a

more characteristic and less ambiguous behavior, so it is possible to rely more on the information extracted by our approach. Thus, we believe that further effort could be put into developing a refined algorithm, able to improve the quality of the estimated motion for a subset of all the spatio-temporal points in the scene, more precisely the ones which are characterized by a richer behavior in the space-time domain.

Appendix

Computational Details

The information provided by the shearlet transform of a three-dimensional signal is very rich. The parameters chosen while executing the decomposition influence directly the amount of processing which is required. In this section, we focus on one of the aspects which required the most of the engineering, namely the computation of the pointwise local representation that we have introduced in Section 3.3.

Computing the Representation

The computation of the representation requires quite a bit of resources to be allocated for it. Within this section, we show how the computation is carried on, then we explain how these operations have been optimized.

The initial coefficients, calculated through the shearlet decomposition of the input signal f, are passed through different steps before reaching the final form $\mathbf{D}(\hat{m})$, or $\mathbf{F}(\hat{m})$, of the representation. To explicitly detail these steps, we need to introduce a little bit of notation, and we will strictly refer to the implementation in ShearLab3D and to the way we have considered its use.

Let M, N be, respectively, the number of rows and columns in each frame belonging to the sequence f that we are considering. Let T be the number of total frames in the sequence, S the total number of shearlet in our system (which is related to the number of shearings per scale, and to the number of scales considered) and S_j the number of shearlets associated with the scale of index j. Finally, $S_{\mathcal{P},j}$ is the dimensionality of the side of each matrix C_i represented in Figure 36(a) for a given scale j.

Given an input video sequence $f \in \mathbb{R}^{M\times N\times T}$, the corresponding shearlet decomposition gives as a result an object $SH \in \mathbb{R}^{M\times N\times T\times S}$. This means that, for each frame in the sequence, for every spatial point within that frame, we have S coefficients as a result of this decomposition. To create the grid of shearings as in Figure 36(a), we have to consider only a subset of these S coefficients, precisely the ones associated with the scale parameter $\hat{j}$ of interest. Once we have extracted these S_j coefficients, we gather them in three matrices C_1,C_2 and C_3 with respect to the pyramidal partition $\mathcal{P}_i$ they belong to. This first step induces one mapping, which we refer to as

$$\phi_1(SH[f](\cdot)) = (C_1, C_2, C_3) : \mathbb{R}^{M\times N\times T\times S} \to \mathbb{R}^{S_{\mathcal{P},j}\times S_{\mathcal{P},j}\times 3}$$

taking the raw coefficients as an input and producing the three foretold matrices.

The next step is represented by the first red arrow in Figure 36 between (a) and (b). This steps represents how the three matrices are transformed in the object $\mathbf{C}$. Again, this is a transformation defined as

$$\phi_2(C_1, C_2, C_3) = \mathbf{C} : \mathbb{R}^{S_{\mathcal{P},j}\times S_{\mathcal{P},j}\times 3} \to \mathbb{R}^{(3\cdot S_{\mathcal{P},j})\times(3\cdot S_{\mathcal{P},j})}$$

The factor 3 in the resulting dimensionality of the mapping ϕ_2 is related to an implementation choice, in particular to the way we *unroll* this object, which is the following. Starting from the object $\mathbf{C}$, knowing that it is representative of the spatio-temporal behavior of the selected point in space-time, we unroll it to the 1-dimensional vector $\mathbf{D}(\hat{m})$ (the second arrow in Figure 36) in this way: starting from the central element (which also corresponds to the shearing associated to the coefficient with the maximum value among those resulting for the selected point) and we unroll the matrix in a counter-clockwise manner. This unrolling procedure can be associated with a mapping

$$\phi_3(\mathbf{C}) = \mathbf{D}(\hat{m}) : \mathbb{R}^{(3\cdot S_{\mathcal{P},j})\times(3\cdot S_{\mathcal{P},j})} \to \mathbb{R}^{S_j}$$

All these steps turn out to be computationally heavy, for they involve several matrices tiling, shifting, and unrolling operations, which have to be carried on for all the $M \times N$ pixels in a given frame at time t. These computations can be sped up, by considering that

in case two different points $\hat{m}_1$ and $\hat{m}_2$ have their corresponding maximum coefficient associated with the same shearing parameter k_{max}, the mappings ϕ_1, ϕ_2 and ϕ_3 will permute the input shearlet coefficients from $SH[f](\cdot)$ to $\mathbf{D}(\hat{m})$ in the same exact way. Thus, it is possible to precompute the effect of the three subsequent mappings on the order of the input raw coefficients that leads to the final representation $\mathbf{D}(\hat{m})$ given k_{max}.

$$\phi(SH[f](\cdot)) = \mathbf{D}(\hat{m}) : \mathbb{R}^{M \times N \times T \times S} \rightarrow \mathbb{R}^{S_j}$$
$$\phi(SH[f](\cdot)) \equiv \phi_3(\phi_2(\phi_1(SH[f](\cdot)))).$$

This consideration led to a quicker implementation of the whole representation creation algorithm. The whole framework, both the algorithmic and experimental part, has been implemented in MATLAB[3], which is well-known for handling in a better way matrix operations w.r.t. iterative methods. Thus, the mapping $\phi(SH[f](\cdot))$ has been implemented in a way to exploit this characteristic of the programming environment we chose.

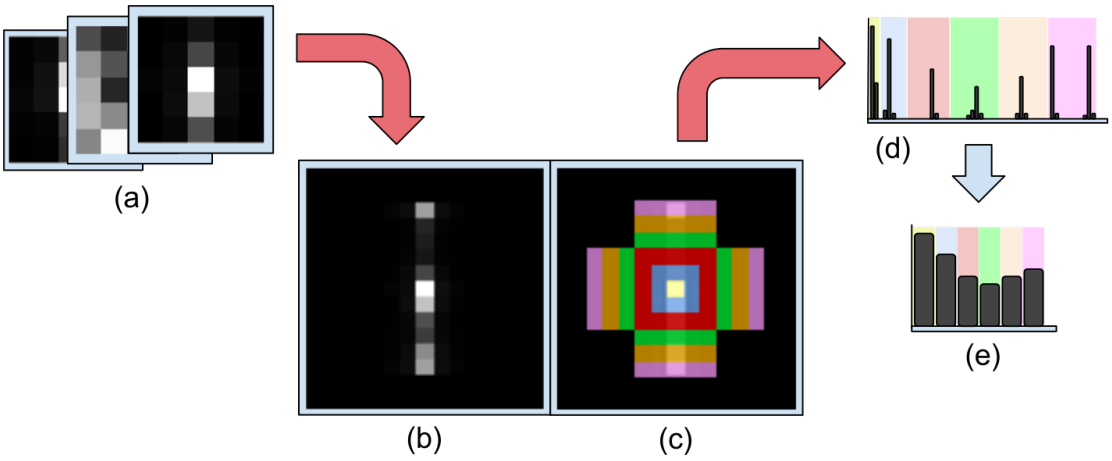


Figure 36: The representation creation pipeline, in red the two arrows the text in this section refers to, while describing the algorithm (see text for details).

[3] https://www.mathworks.com/products/matlab.html

Datasets

To show the potential of our approach, we have selected sequences from benchmark datasets. Some of them are widely known, others have been recorded with the specific purpose of carrying on a precise task, and we consider their use for our needs. Here we report the main characteristics of each dataset, with a few sample frames to give an hint of the dynamic which characterize each sequence. We want to specify that we selected only a few sequences from each dataset, which are acquired with heterogeneous sensors, in different environments, in various illumination conditions and that represent distinct kinds of motion. We followed this approach so that to highlight the different capabilities of our framework.

KTH dataset

The KTH dataset [SLC04] is characterized by videos belonging to six different full-body actions: *walking*, *boxing*, *handwaving*, *clapping*, *running*, *jogging*. In a sequence only a single action is reproduced and repeated for several times. We considered only a subset of all them, so to highlight the most important characteristics and capabilities of the methods that we have developed. The three sequences in Figure 37, 38 and 39 represent three different kinds of motion, which however share a common characteristic, that is they are related only to the upper or the lower part of the body. Thus, for example, spatio-temporal interest points arise only in a subpart of the corresponding 2D+T-dimensional structure, allowing for a better visualization.

Figure 37: A *boxing* action, where the subject extends his arms in front of him repeatedly.

Figure 38: Frames from an *handwaving* action.

Figure 39: A *walking* action, with the subject crossing the scene in a few strides.

ChaLearn dataset

The ChaLearn dataset [EBG+14] has been developed and enriched over the years for a human action recognition contest. The sequences contains several instances of different Italian gestures, each one of them which is modeled on the basis of what is commonly used within spoken Italian. Even if we considered only the RGB stream between all the ones available, the ChaLearn dataset also contains both the depth and the Microsoft Kinect skeleton data associated with the recordings. Again, we considered such a sequence since spatio-temporal interest points are really sparse, and exist only around the only parts of the subject which are moving, the arms and the fists.

Figure 40: Frames from a *che vuoi* gesture: the male depicted raises bot hands and shakes them, before lowering them again.

Cooking and Close-up Actions datasets

These two human-robot interaction dataset have been recorded by Vignolo and colleagues [VNR+17]. One of the datasets contains close-up registrations of cooking actions, while

the second one represents a set of prototypical movements carried on in front of the camera, with the subject performing periodic movements following a different patterns. These have been chosen for two different reasons. Figure 41 represent a sequence in which the movement can be clearly segmented in left-to-right and right-to-left executions, so it is a good candidate to try to extract meaningful frames from it. Figure 42 has been chosen because the video sequence it is associated with represent a periodic motion, where the subject is mixing flour in a bowl: this leads to periodic movements, thus it is meaningful to inspect how the points in a sequence are associated with different spatio-temporal primitives over time, looking for a periodicity pattern within the corresponding profiles (as shown in Figure 3.25 and following ones).

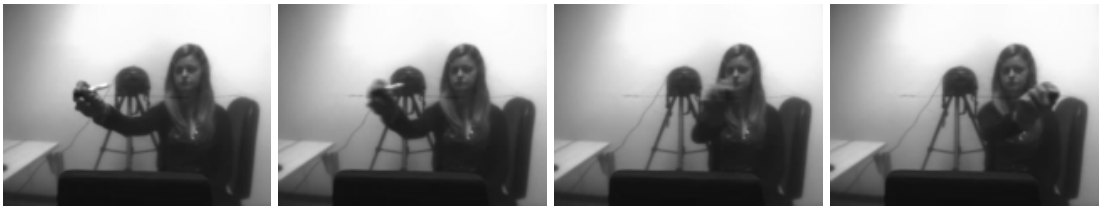


Figure 41: Frames from a *drawing line* action, with the subject drawing an horizontal line in front of her, keeping her arm raised.

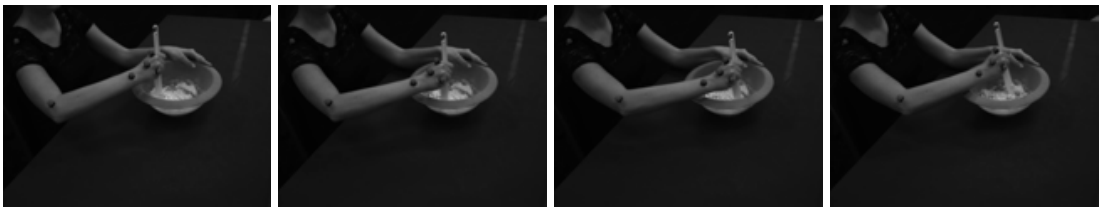


Figure 42: Frames from a *Mixing* action, where the subject is mixing flour in a bowl with a periodic circular motion.

DANCE dataset

This dataset has been recorded by Casa Paganini research lab[4] team for the objectives of the European funded DANCE project [CVP+16], which investigates how affective and relational qualities of human full-body movement can be expressed by the auditory channel. We consider a single sequence from it, one in which a dancer oscillates back and forth and develops a fluid, smooth movement over time with her body.

[4] http://www.casapaganini.org/

Figure 43: Frames from a sequence where a dancer is moving slowly, describing an arc with her right arm.

Roadway sequence

This sequence does not belong to any particular dataset, it has been downloaded from the Internet specifically to evaluate the sensitivity of our method to more than a single element moving at different speed within the same time frame. The sequence in Figure 44 is the one considered in Figure 3.31.

Figure 44: Frames from a roadway sequence, the vehicles depicted appear as moving at different speed.

Bibliography

[AKW+17] T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, and M. Magnor. Optical flow-based 3d human motion estimation from monocular video. In *German Conference on Pattern Recognition*, pages 347–360. Springer, 2017.

[AM96] J.-P. Antoine and R. Murenzi. Two-dimensional directional wavelets and the scale-angle representation. *Signal processing*, 52(3):259–281, 1996.

[BBC13] U. Bonde, V. Badrinarayanan, and R. Cipolla. Multi scale shape index for 3d object recognition. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 306–318. Springer, 2013.

[BBM09] T. Brox, C. Bregler, and J. Malik. Large displacement optical flow. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 41–48. IEEE, 2009.

[BBPW04] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. *Computer Vision-ECCV 2004*, pages 25–36, 2004.

[BGS+05] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395–1402. IEEE, 2005.

[BT78] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2, 1978.

[BWF+05] A. Bruhn, J. Weickert, C. Feddern, T. Kohlberger, and C. Schnorr. Variational optical flow computation in real time. *IEEE Transactions on Image Processing*, 14(5):608–615, 2005.

[Can86] J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1986.

[CD99] E. J. Candès and D. L. Donoho. Ridgelets: A key to higher-dimensional intermittency? *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 357(1760):2495–2509, 1999.

[CD04] E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise c2 singularities. *Communications on pure and applied mathematics*, 57(2):219–266, 2004.

[CDF+04] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[CDMNT06] E. Cordero, F. De Mari, K. Nowak, and A. Tabacco. Reproducing groups for the metaplectic representation. In *Pseudo-differential operators and related topics*, volume 164 of *Oper. Theory Adv. Appl.*, pages 227–244. Birkhäuser, Basel, 2006.

[CHMG12] B. Chakraborty, M. Holte, T. Moeslund, and J. Gonzàlez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 116(3):396 – 410, 2012. Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

[CHS13] 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.

[CVP+16] A. Camurri, G. Volpe, S. Piana, M. Mancini, R. Niewiadomski, N. Ferrari, and C. Canepa. The dancer in the eye: towards a multi-layered computational framework of qualities in movement. In *Proceedings of the 3rd International Symposium on Movement and Computing*, page 6. ACM, 2016.

[CZLK98] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference On*, pages 396–401. IEEE, 1998.

[Dav01] J. W. Davis. Hierarchical motion history images for recognizing human motion. In *Proceedings IEEE Workshop on Detection and Recognition of Events in Video*, pages 39–46, 2001.

[DCZD06] A. L. Da Cunha, J. Zhou, and M. N. Do. The nonsubsampled contourlet transform: theory, design, and applications. *IEEE transactions on image processing*, 15(10):3089–3101, 2006.

[DDMGL15] S. Dahlke, F. De Mari, P. Grohs, and D. Labate, editors. *Harmonic and applied analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, Cham, 2015. From groups to signals.

[DHST15] S. Dahlke, S. Häuser, G. Steidl, and G. Teschke. Shearlet coorbit theory. In *Harmonic and applied analysis*, Appl. Numer. Harmon. Anal., pages 83–147. Birkhäuser/Springer, Cham, 2015.

[DNL09] P. S. Dhillon, S. Nowozin, and C. H. Lampert. Combining appearance and motion for human action classification in videos. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 22–29. IEEE, 2009.

[DNOD17] M. Duval-Poo, N. Noceti, F. Odone, and E. De Vito. Scale invariant and noise robust interest points with shearlets. *IEEE Trans. Image Processing*, 26(6):2853–2867, 2017.

[DODV15] M. Duval-Poo, F. Odone, and E. De Vito. Edges and corners with shearlets. *IEEE Trans. Image Processing*, 24(11):3768–3780, 2015.

[DPNODV17] M. A. Duval-Poo, N. Noceti, F. Odone, and E. De Vito. Detection and description of scale invariant interest points with shearlets. In *Sampling Theory and Applications (SampTA), 2017 International Conference on*, pages 294–298. IEEE, 2017.

[DRCB05] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.

[DST10] S. Dahlke, G. Steidl, and G. Teschke. The continuous shearlet transform in arbitrary space dimensions. *Journal of Fourier Analysis and Applications*, 16(3):340–364, 2010.

[DV05a] M. Do and M. Vetterli. The contourlet transform: An efficient directional multiresolution image representation. *Trans. Img. Proc.*, pages 2091–2106, 2005.

[DV05b] M. N. Do and M. Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on image processing*, 14(12):2091–2106, 2005.

[EA06] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

[EBG+14] S. Escalera, X. Baro, J. Gonzalez, M. Bautista, M. Madadi, M. Reyes, V. Ponce-Lopez, H. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *ECCV Workshops*, 2014.

[EGM+94] D. Eberly, R. Gardner, B. Morse, S. Pizer, and C. Scharlach. Ridges for image analysis. *Journal of Mathematical Imaging and Vision*, 4(4):353–373, 1994.

[EL12] G. R. Easley and D. Labate. Image processing using shearlets. In *Shearlets*, Appl. Numer. Harmon. Anal., pages 283–325. Birkhäuser/Springer, New York, 2012.

[ELC09] G. Easley, D. Labate, and F. Colonna. Shearlet-based total variation diffusion for denoising. *IEEE Transactions on Image processing*, 18(2):260–268, 2009.

[ELL08] G. Easley, D. Labate, and W.-Q. Lim. Sparse directional image representations using the discrete shearlet transform. *Applied and Computational Harmonic Analysis*, 25(1):25–46, 2008.

[ELN13] G. Easley, D. Labate, and P. Negi. 3D data denoising using combined sparse dictionaries. *Math. Model. Nat. Phenom.*, 8(1):60–74, 2013.

[FÖ5] H. Führ. *Abstract harmonic analysis of continuous wavelet transforms*, volume 1863 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2005.

[Far00] G. Farneback. Fast and accurate motion estimation using orientation tensors and parametric motion models. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 135–139. IEEE, 2000.

[FB12] A. Firouzmanesh and P. Boulanger. Image de-blurring using shearlets. In *Computer and Robot Vision (CRV), 2012 Ninth Conference on*, pages 167–173. IEEE, 2012.

[FBK15] D. Fortun, P. Bouthemy, and C. Kervrann. Optical flow modeling and computation: a survey. *Computer Vision and Image Understanding*, 134:1–21, 2015.

[FG87] W. Förstner and E. Gülch. A fast operator for detection and precise location of distinct points, corners and centres of circular features. In *Proc. ISPRS intercommission conference on fast processing of photogrammetric data*, pages 281–305, 1987.

[GL11] 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.

[GL12] 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.

[GL13] K. Guo and D. Labate. Optimal recovery of 3D X-ray tomographic data via shearlet decomposition. *Adv. Comput. Math.*, 39(2):227–255, 2013.

[GLL+04] K. Guo, D. Labate, W.-Q. Lim, G. Weiss, and E. Wilson. Wavelets with composite dilations. *Electronic research announcements of the American Mathematical Society*, 10(9):78–87, 2004.

[GLL09] K. Guo, D. Labate, and W. Lim. Edge analysis and identification using the continuous shearlet transform. *Applied and Computational Harmonic Analysis*, 27(1):24–46, 2009.

[GPLC14] X. Gibert, V. M. Patel, D. Labate, and R. Chellappa. Discrete shearlet transform on gpu with applications in anomaly detection and denoising. *EURASIP Journal on Advances in Signal Processing*, 2014(1):64, 2014.

[GS08] L. Grafakos and C. Sansing. Gabor frames and directional time–frequency analysis. *Applied and Computational Harmonic Analysis*, 25(1):47–67, 2008.

[HHZ14] C. He, C.-H. Hu, and W. Zhang. Adaptive shearlet-regularized image deblurring via alternating direction method. In *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, pages 1–6. IEEE, 2014.

[HL16] 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.

[HS81] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, 1988.

[IMS+17] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[JMOB10] R. Jenatton, J. Mairal, G. Obozinski, and F. R. Bach. Proximal methods for sparse hierarchical dictionary learning. In *ICML*, number 2010, pages 487–494. Citeseer, 2010.

[KKL13] E. King, G. Kutyniok, and W. Lim. Image inpainting: theoretical analysis and comparison of algorithms. In *Wavelets and Sparsity XV*, volume 8858, page 885802, September 2013.

[KL10] G. Kutyniok and W.-Q. Lim. Image separation using wavelets and shearlets. In *International Conference on Curves and Surfaces*, pages 416–430. Springer, 2010.

[KL12] G. Kutyniok and D. Labate. *Shearlets*. Appl. Numer. Harmon. Anal. Birkhäuser/Springer, New York, 2012.

[KLL12] 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.

[KLR16] G. Kutyniok, W. Lim, and R. Reisenhofer. Shearlab 3d: Faithful digital shearlet transforms based on compactly supported shearlets. *ACM Trans. Math. Softw.*, pages 5:1–5:42, 2016.

[KP15] G. Kutyniok and P. Petersen. Classification of edges using compactly supported shearlets. *Applied and Computational Harmonic Analysis*, 2015.

[KSH05] Y. Ke, R. Sukthankar, and M. Hebert. 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, pages 166–173, Washington, DC, USA, 2005. IEEE Computer Society.

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[KVD91] J. J. Koenderink and A. J. Van Doorn. Affine structure from motion. *JOSA A*, 8(2):377–385, 1991.

[KvD92] J. J. Koenderink and A. J. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557 – 564, 1992.

[Lap05] I. Laptev. On space-time interest points. *Int. J. Computer Vision*, 64(2):107–123, 2005.

[LBBH98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LH81] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133, 1981.

[Lin96] T. Lindeberg. Scale-space: A framework for handling image structures at multiple scales. 1996.

[Lin98a] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):117–156, 1998.

[Lin98b] T. Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision*, 30(2):79–116, 1998.

[LK+81] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[LLKG05] D. Labate, W. Lim, G. Kutyniok, and W. Guido. Sparse multidimensional representation using shearlets. *SPIE Proc. 5914, SPIE, Bellingham*, 2005.

[LLKW05] D. Labate, W.-Q. Lim, G. Kutyniok, and G. Weiss. Sparse multidimensional representation using shearlets. *Optics & Photonics*, 2005:59140U–59140U, 2005.

[LMSR08] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[Low04] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[LSC95] J.-S. Lee, Y.-N. Sun, and C.-H. Chen. Multiscale corner detection by using wavelet transform. *IEEE Transactions on Image Processing*, 4(1):100–104, 1995.

[MBP+08] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[MGV+17] D. Malafronte, G. Goyal, A. Vignolo, F. Odone, and N. Noceti. Investigating the use of space-time primitives to understand human movements. In *International Conference on Image Analysis and Processing*, pages 40–50. Springer, 2017.

[MLS09] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2929–2936. IEEE, 2009.

[MMF+17] J. Ma, M. März, S. Funk, J. Schulz-Menger, G. Kutyniok, T. Schaeffter, and C. Kolbitsch. Shearlet-based compressed sensing for fast 3d cardiac mr imaging using iterative reweighting. *arXiv preprint arXiv:1705.00463*, 2017.

[MODV17a] D. Malafronte, F. Odone, and E. De Vito. Detecting spatio-temporally interest points using the shearlet transform. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 501–510. Springer, 2017.

[MODV17b] D. Malafronte, F. Odone, and E. De Vito. Local spatio-temporal representation using the 3d shearlet transform. In *Sampling Theory and Applications (SampTA), 2017 International Conference on*, pages 585–589. IEEE, 2017.

[MODVar] D. Malafronte, F. Odone, and E. De Vito. Space-time signal analysis and the 3d shearlet transform. *Journal of Mathematical Imaging and Vision*, to appear.

[MP02] E. Mémin and P. Pérez. Hierarchical estimation and segmentation of dense motion fields. *International Journal of Computer Vision*, 46(2):129–155, 2002.

[MV82] D. Marr and A. Vision. A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company*, 1(2), 1982.

[MZ92] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 710–732, 1992.

[NL12a] P. Negi and D. Labate. 3D discrete shearlet transform and video processing. *IEEE Trans. Image Processing*, 21(6):2944–2954, 2012.

[NL12b] P. S. Negi and D. Labate. 3-D discrete shearlet transform and video processing. *IEEE Trans. Image Process.*, 21(6):2944–2954, 2012.

[PKG15] S. Pejoski, V. Kafedziski, and D. Gleich. Compressed sensing mri using discrete nonseparable shearlet transform and fista. *IEEE Signal Processing Letters*, 22(10):1566–1570, 2015.

[PLPS16] A. Pein, S. Loock, G. Plonka, and T. Salditt. Using sparsity information for iterative phase retrieval in x-ray propagation imaging. *Optics express*, 24(8):8332–8343, 2016.

[PM90] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on pattern analysis and machine intelligence*, 12(7):629–639, 1990.

[RYS02] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226, 2002.

[SA93] P. Sinha and E. Adelson. Recovering reflectance and illumination in a world of painted polyhedra. In *Computer Vision, 1993. Proceedings., Fourth International Conference on*, pages 156–163. IEEE, 1993.

[SES12] T. Senst, V. Eiselein, and T. Sikora. Robust local optical flow for feature tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(9):1377–1387, 2012.

[SLC04] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, 2004.

[SM98] J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *Computer Vision, 1998. Sixth International Conference on*, pages 1154–1160. IEEE, 1998.

[ST94] J. Shi and C. Tomasi. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, Jun 1994.

[SZ14] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems 27*, pages 568–576. Curran Associates, Inc., 2014.

[TK91] C. Tomasi and T. Kanade. Detection and tracking of point features. 1991.

[TV98] E. Trucco and A. Verri. *Introductory techniques for 3-D computer vision*, volume 201. Prentice Hall Englewood Cliffs, 1998.

[Vig] A. Vignolo. Cooking actions dataset. to appear.

[VNR+17] A. Vignolo, N. Noceti, F. Rea, A. Sciutti, F. Odone, and G. Sandini. Detecting biological motion for human–robot interaction: A link between perception and action. *Frontiers in Robotics and AI*, 4:14, 2017.

[WB95] H. Wang and M. Brady. Real-time corner detection algorithm for motion estimation. *Image and Vision Computing*, 13(9):695 – 703, 1995.

[WC07] S. F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007.

[WCM05] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1800–1807. IEEE, 2005.

[WKSL11] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[WPZ+09] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers. An improved algorithm for tv-l 1 optical flow. In *Statistical and geometrical approaches to visual motion analysis*, pages 23–45. Springer, 2009.

[WRHS13] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large displacement optical flow with deep matching. In *IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia, December 2013.

[WS13] H. Wang and C. Schmid. Action recognition with improved trajectories. In *2013 IEEE International Conference on Computer Vision*, pages 3551–3558, Dec 2013.

[WTG08] G. Willems, T. Tuytelaars, and L. Gool. 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.

[YLEK08] S. Yi, D. Labate, G. R. Easley, and H. Krim. Edge detection and processing using shearlets. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 1148–1151. IEEE, 2008.

[YLEK09a] S. Yi, D. Labate, G. Easley, and H. Krim. A shearlet approach to edge analysis and detection. *IEEE Transactions on Image Processing*, 2009.

[YLEK09b] S. Yi, D. Labate, G. R. Easley, and H. Krim. A shearlet approach to edge analysis and detection. *IEEE Transactions on Image Processing*, 18(5):929–941, 2009.

[YLEK09c] S. Yi, D. Labate, G. R. Easley, and H. Krim. A shearlet approach to edge analysis and detection. *IEEE Transactions on Image Processing*, 18(5):929–941, 2009.

[YYGH09] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801. IEEE, 2009.

[ZF14] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.