Chapter 2

Methodology for the study of the geometry of structures in turbulence

The starting point of this methodology is a three-dimensional scalar field obtained from a turbulence database. Three properties were sought in its development: multi-scale capability, non-local character, and geometry-based analysis. It consists of three main steps: extraction, characterization, and classification of structures. They are explained in each of the sections of this chapter.

2.1 Extraction of structures

The main requirement imposed on the extraction process is to enable eduction of structures associated with different ranges of scales. Although scale decomposition is commonly defined in Fourier space, the nature of Fourier basis functions, that are localized in wavenumber but not in physical space, makes top-hat window filtering in Fourier space inappropriate for the purpose of educing structures that are extended but compact in physical space. Thus, a transformation with basis functions that are localized both in Fourier space, where the ranges of scales are defined, and in physical space, where the structures are to be educed, is required. For this purpose, the curvelet transform (Candès & Donoho, 2003*a*,*b*) in its three-dimensional discretized version (Ying et al., 2005; Candès et al., 2005) is used. Owing to the multi-dimensional character of their definition, curvelets, unlike wavelets, are naturally suited for detecting, organizing, or providing a compact representation of intermediate multi-dimensional structures.

2.1.1 The curvelet transform

Curvelets, the basis functions of the curvelet transform, are localized in scale (frequency/Fourier space), position (physical space) and orientation (unlike wavelets). The frequency space is smoothly windowed in radial and angular spherical coordinates, providing the decomposition in different scales and orientations, respectively. For a given scale[1], j, the radial window smoothly extracts the frequency contents near the dyadic corona $[2^{j-1}, 2^{j+1}]$. A low-pass radial window is introduced for the coarsest scale, j_0. The unit sphere representing all directions in $\mathbb{R}^3$ is partitioned, for each scale $j > j_0$, into $O\left(2^{j/2} \cdot 2^{j/2}\right) = O\left(2^j\right)$ smooth angular windows, each with a disk-like support of radius $O\left(2^{-j/2}\right)$. In a discrete three-dimensional data field, of uniform grid of size n^3, the last scale, j_e, which extracts the highest frequency content, is given by $j_e = \log_2(n/2)$.

Denoting by $f(n_1, n_2, n_3)$ the scalar field, where $0 \leq n_i < n$, being n the number of grid points in each direction, the discrete version of the curvelet transform (see Ying et al., 2005) provides a set of coefficients $c^D(j, l, k)$ defined as

$$c^D(j, \ell, k) \equiv \sum_{n_1, n_2, n_3} f(n_1, n_2, n_3)\, \overline{\varphi^D_{j,\ell,k}(n_1, n_2, n_3)} \tag{2.1}$$

where $j, \ell \in \mathbb{Z}$, $k = (k_1, k_2, k_3)$ (j represents the scale, ℓ the orientation, and k the spatial location); $\varphi^D_{j,\ell,k}(n_1, n_2, n_3)$ are the curvelets, defined in Fourier space by

$$\hat{\varphi}^D_{j,\ell,k}(\omega) \equiv \tilde{U}_{j,\ell}(\omega) \exp\left(\frac{-2\pi i \displaystyle\sum_{i=1}^{3} \frac{k_i \omega_i}{L_{i,j,\ell}}}{\sqrt{\displaystyle\prod_{i=1}^{3} L_{i,j,\ell}}} \right) \tag{2.2}$$

for $\{0 \leq k_i < L_{i,j,\ell}, i = 1, 2, 3\}$—where ω is the wavenumber; $\tilde{U}_{j,\ell}(\omega)$ is the frequency window $\tilde{U}_{j,\ell}(\omega) = \tilde{W}_j(\omega)\, \tilde{V}_{j,\ell}(\omega)$, being $\tilde{W}_j(\omega)$ and $\tilde{V}_{j,\ell}(\omega)$ the radial and angular windows; and $\{L_{i,j,\ell}, i =$

[1] The term scale, when referred to the index j in curvelet space, denotes in fact the range of scales in physical space that results from the radial window filter in Fourier space.

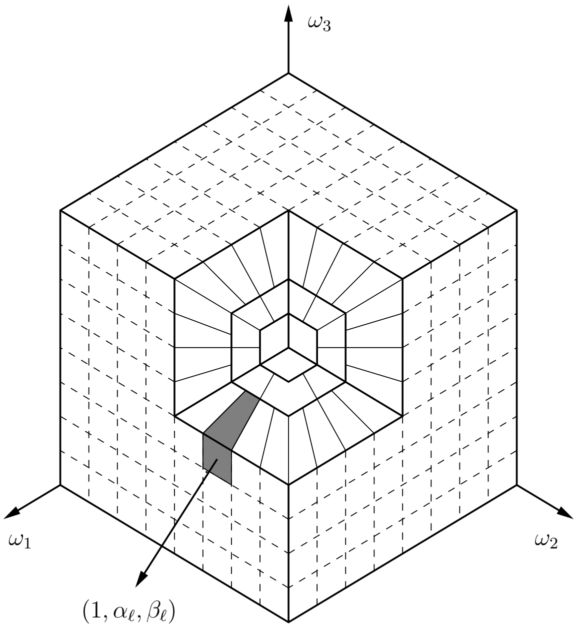


Figure 2.1: Frequency window $\tilde{U}_{j,\ell}$ (darkened region) defined in the three-dimensional discrete curvelet transform, extracting the frequency content near the wedge with center slope $(1, \alpha_\ell, \beta_\ell)$ (figure based on Ying et al. (2005))

$1, 2, 3\}$ are three positive integers such that: $i)\left\{ \nexists\, \omega, \omega' \,\middle|\, \tilde{U}_{j,\ell}(\omega) \geq \tilde{U}_{j,\ell}(\omega') \,\text{and}\, \omega_i \,\text{is multiple of}\, L_i, i = 1, 2, 3 \right\}$; $ii)\Pi_{i=1}^{3} L_{i,j,\ell}$ is minimal. A Cartesian coronae is used, so that:

$$\tilde{W}_{j_0}(\omega) = \Phi_{j_0}(\omega); \qquad \tilde{W}_j(\omega) = \sqrt{\Phi_{j+1}^2(\omega) - \Phi_j^2(\omega)}, \qquad j > j_0, \tag{2.3}$$

where $\Phi_j(\omega_1, \omega_2, \omega_3) = \phi(2^{-j}\omega_1)\,\phi(2^{-j}\omega_2)\,\phi(2^{-j}\omega_3)$, and ϕ is a smooth function such that $0 \leq \phi \leq 1$: it equals unity on $[-1, 1]$ and zero outside $[-2, 2]$. The angular window for the ℓth direction is defined (for example, in the $\omega_1 > 0$ face of the unit cube) as

$$\tilde{V}_{j,\ell}(\omega) = \tilde{V}\left(2^{j/2}\frac{\omega_2 - \alpha_\ell\omega_1}{\omega_1}\right)\,\tilde{V}\left(2^{j/2}\frac{\omega_3 - \beta_\ell\omega_1}{\omega_1}\right) \tag{2.4}$$

where $(1, \alpha_\ell, \beta_\ell)$ is the direction of the center line of the wedge (see Figure 2.1) defining the center slope for the ℓth wedge. Wherever three smooth angular windows $\tilde{V}_{j,\ell}$, $\tilde{V}_{j,\ell'}$, and $\tilde{V}_{j,\ell''}$ overlap, they are redefined as $\left(\tilde{V}_{j,\ell}, \tilde{V}_{j,\ell'}, \tilde{V}_{j,\ell''}\right)/\sqrt{\tilde{V}_{j,\ell}^2 + \tilde{V}_{j,\ell'}^2 + \tilde{V}_{j,\ell''}^2}$.

$\tilde{W}_j(\omega)$ and $\tilde{V}(\omega)$ satisfy:

$$\sum_{j \geq j_0} \tilde{W}_j^2(\omega) = 1, \qquad \sum_{\ell=-\infty}^{\infty} \tilde{V}^2(t - 2\ell) = 1. \tag{2.5}$$

Curvelets form a tight-frame in $L^2\left(\mathbb{R}^3\right)$. Any function $f \in L^2\left(\mathbb{R}^3\right)$ can be expanded as $f = \sum_{j,\ell,k} \langle \varphi_{j,\ell,k}, f \rangle \varphi_{j,\ell,k}$, where $\varphi_{j,\ell,k}$ is the curvelet at scale j, orientation ℓ, and position $k = (k_1, k_2, k_3)$. Parseval's identity holds: $\sum_{j,\ell,k} \mid \langle f, \varphi_{j,\ell,k} \rangle \mid^2 = \|f\|^2_{L^2(\mathbb{R}^3)}$. The effective longitudinal and cross-sectional dimensions (length and width), of curvelet basis functions in physical space follow the relation width≈length^2 (parabolic scaling). As a consequence of this parabolic scaling, curvelets are an optimal (sparse) basis for representing surface-like singularities of codimension one. These are three of the most remarkable properties of the curvelet transform.

We apply the curvelet transform to a scalar field, obtained from a turbulence database at an instant in time, but again emphasize its broader applicability to other fields. The curvelet transform allows a multi-scale decomposition by filtering in curvelet space the different scales of interest $j = j_0, ..., j_e$, individually or in groups. In addition, for anisotropic fields with privileged direction(s) (e.g., shear flows), a multi-orientation decomposition may be useful for studying structures according to their directionality (by using the angular window filtering in frequency space of the curvelet transform (index ℓ in curvelet space)). Throughout this thesis, only the multi-scale decomposition is used, which could be also attained by other multi-resolution techniques sharing the same choice of sub-band radial filtering decomposition in Fourier space. Nevertheless, those capabilities that set curvelets apart from other multi-resolution techniques, e.g., multi-orientation decomposition and compact representation of surface-like singularities, justify its early implementation within the frame of this methodology, enhancing its potential applications and possibilities of expansion.

For each scale $j = j_0, ..., j_e$, a new scalar field is obtained after filtering all other scales ($j\prime \neq j$) in curvelet space and inverse transforming to physical space. Thus, a set of $j_e - j_0 + 1$ filtered scalar fields results from the original field. The volume-based probability density functions (pdfs) of the filtered fields are, in general, different from each other and from the original field; their comparison can be useful in determining how the original scalar field is distributed among the different scales.

After this multi-scale analysis, a second step is applied in the extraction process, by which the structures of interest associated with each relevant range of scales are educed. Currently those structures of interest are defined as the individual disconnected surfaces obtained by iso-contouring each filtered scalar field at particular contour values (for example, the mean value of that filtered scalar field plus a multiple of its standard deviation). See Appendix B for a physical interpretation of the educed structures following this multi-scale decomposition plus iso-contouring procedure.

2.1.2 Periodic reconnection

In the case of scalar fields with periodic boundaries, an additional step is included in the extraction process, to reconnect those structures intersecting boundaries with their periodic continuation on the opposite boundaries. Figure 2.2 shows an example of the application of such periodic reconnection algorithm to a set of 3D structures obtained from a periodic scalar field.

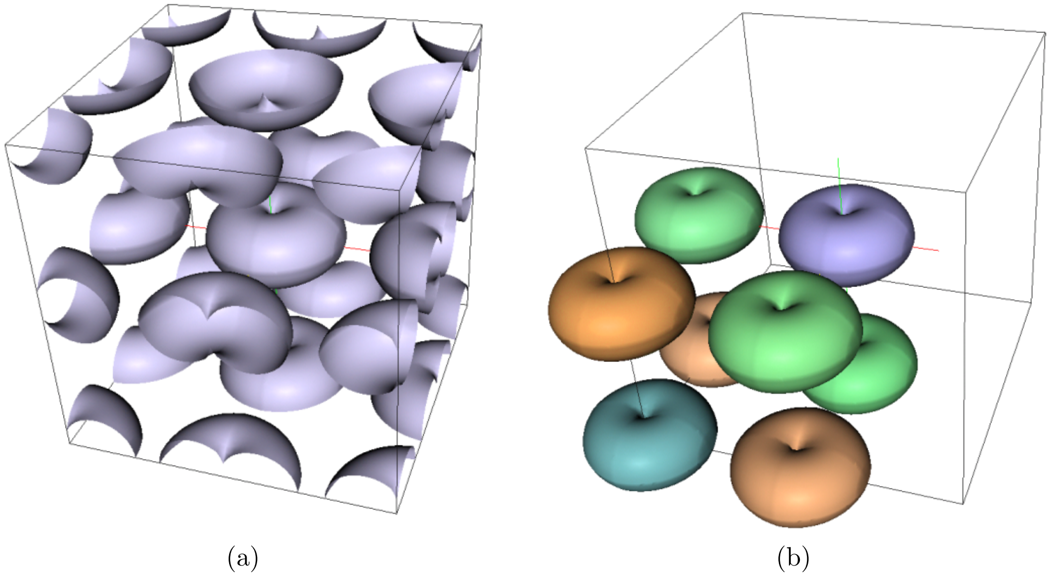


Figure 2.2: Example of application of the periodic reconnection algorithm to a set of boundary-intersecting structures obtained from a periodic three-dimensional scalar field. (a) Before periodic reconnection. (b) After periodic reconnection, where the color of each structure indicates the number of pieces involved in the reconnected structure for this particular scenario: blue = 1 (non-intersecting), green = 2, orange = 4, cyan = 8

Structures spanning across multiple repetitions of the periodic domain are accounted for in the algorithm, as shown schematically in Figure 2.3 for a 2D example. The only case that cannot be completely reconnected is the structure with infinite extent.

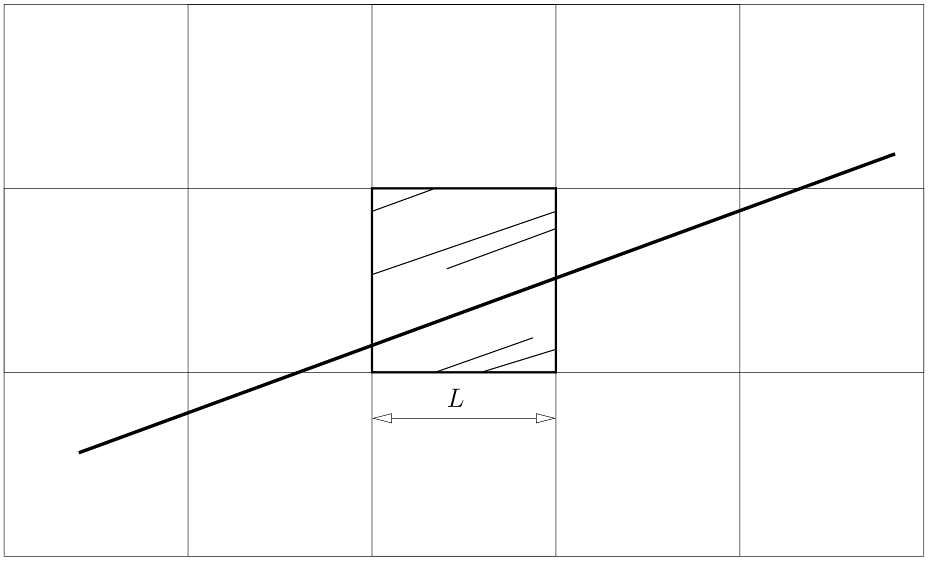


Figure 2.3: 2D example of reconnection of a periodic structure spanning across multiple extensions of the periodic domain. The original fragments of the structure are represented in the original domain (central square). The resulting structure, after reconnection, is represented by the thick line.

2.2 Characterization of structures

We seek a geometrical characterization able to distinguish structures based on their shape. A two-step method is used: first, a suitable set of differential-geometry properties is obtained locally (at all points of the surface), and then area-based probability density functions of those local properties are calculated, making the transition from local to non-local (in the surface sense) possible.

2.2.1 Shape index and curvedness

Shape index, Υ, and curvedness, Λ, (see Koenderink & van Doorn, 1992) are the differential-geometry properties chosen to represent locally the geometry of the surface. They are related to the *principal curvatures* $\{\kappa_1, \kappa_2\}$ of a surface at a given point by:

$$\Upsilon \equiv -\frac{2}{\pi} \arctan\left(\frac{\kappa_1 + \kappa_2}{\kappa_1 - \kappa_2}\right), \qquad \Lambda \equiv \sqrt{\frac{\kappa_1^2 + \kappa_2^2}{2}}. \tag{2.6}$$

Υ is scale-independent, whereas Λ is scale-dependent, having the dimensions of a reciprocal length. The scaling is such that, for example, Λ at every point on a sphere equals the absolute value of its reciprocal radius, $1/R$, whereas the cylinder of radius R presents $\Lambda = 1/(\sqrt{2}R)$ for all points. The principal curvatures, $\{\kappa_1, \kappa_2\}$, are obtained as the maximum and minimum values of the normal curvature, κ_n, in all possible directions of the tangent plane defined at the point P of the surface of study. The normal curvature, κ_n, at a point P in a given direction $\boldsymbol{a}$ of the tangent plane, defined as the division of the second and first fundamental forms of differential geometry applied in that direction, $\boldsymbol{a}$, can also be interpreted as the inverse of the radius of curvature, R, of the curve obtained as the intersection of the surface and the plane defined by the direction $\boldsymbol{a}$ and the normal $\boldsymbol{N}$ to the surface at the point P. Thus higher values of the curvedness correspond to smaller radius of curvature (and, therefore, more locally curved surface at P). All regular patches of a regular surface M map on the domain $(\Upsilon, \Lambda) \in [-1, +1] \times \mathbb{R}^+$, except for the planar patch, which has an indeterminate shape index and nil curvedness (since $\kappa_1 = \kappa_2 = 0$).

The mapping $(\kappa_1, \kappa_2) \rightarrow (\Upsilon, \Lambda)$ represents (see Figure 2.4) a transformation from Cartesian coordinates (κ_1, κ_2) to non-standard polar coordinates (Υ, Λ). For any point in the (κ_1, κ_2) plane, Υ contains the information on the direction (measured as the angle, ϕ, with respect to the axis $\kappa_1 - \kappa_2$, rescaled into the range $[-1, +1]$ by $\Upsilon = -2\phi/\pi$), whereas Λ contains the information on the distance, ϱ, to the origin (rescaled as $\Lambda = \varrho/\sqrt{2}$). The convention chosen when ordering the principal curvatures ($\kappa_1 \geq \kappa_2$) implies that only the region $\kappa_1 - \kappa_2 \geq 0$ of the (κ_1, κ_2) plane is accessible (see Figure 2.4). Therefore, the polar angle ϕ can only have values in the range $[-\pi/2, +\pi/2]$, and, consequently, $\Upsilon = -\frac{\phi}{\pi/2} \in [-1, +1]$ covers all the possible cases, thus making the mapping $(\kappa_1, \kappa_2) \rightarrow (\Upsilon, \Lambda)$ injective (excluding the point $(\kappa_1, \kappa_2) = (0, 0)$ from its domain) by eliminating the multi-valuedness of the arctan function used in the definition of the shape index. The absolute value of the shape index $S \equiv |\Upsilon|$ represents the local shape of the surface at the point P, with $0 \leq S \leq 1$. Its sign indicates the direction of the normal, distinguishing, for example, convex from concave elliptical points. Figure 2.5 shows the range of values of Υ and sketches of the local shapes associated with representative values, with the names of corresponding points. Figure 2.6 shows

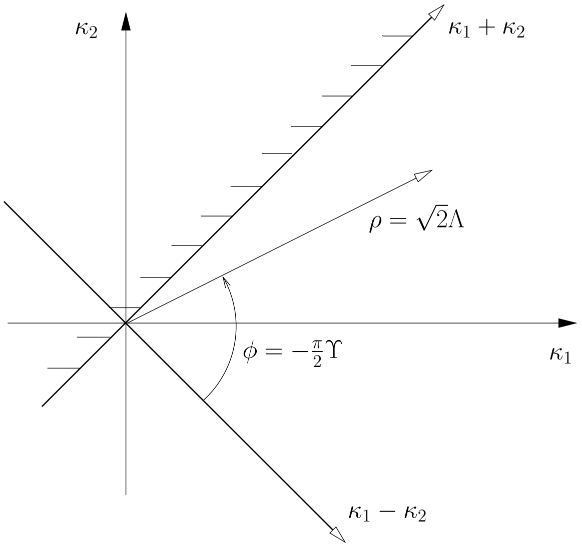


Figure 2.4: Transformation from (κ_1, κ_2) to (Υ, Λ)

the mapping of both Υ and Λ in the plane of principal curvatures, also with representative local shapes. A deeper mathematical background of these differential-geometry concepts is presented in Appendix C.

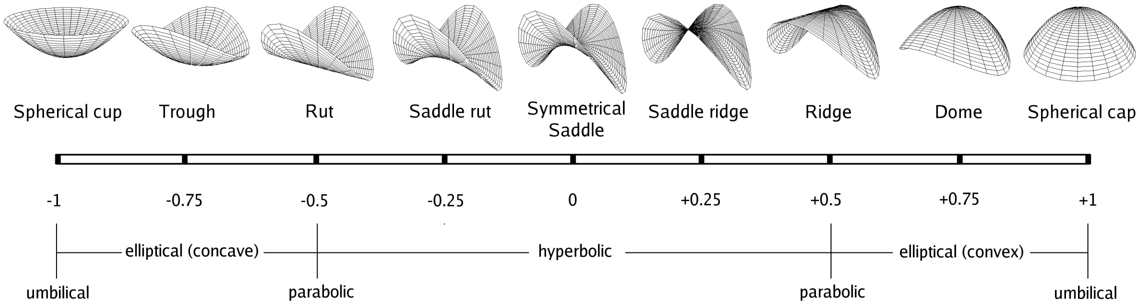


Figure 2.5: Range of shape index, Υ, with its most representative associated local shapes (figure based on Koenderink & van Doorn (1992))

2.2.2 Joint probability density function (jpdf)

From the pointwise Υ and Λ, a two-dimensional area-based joint probability density function in the space of (S, Λ) can be obtained (see Appendix D). Since Λ is scale dependent, in order to compare the shape of surfaces of different sizes, a non-dimensionalization is required for each surface. Selection of the appropriate length scale for this purpose is critical; several can be obtained from global

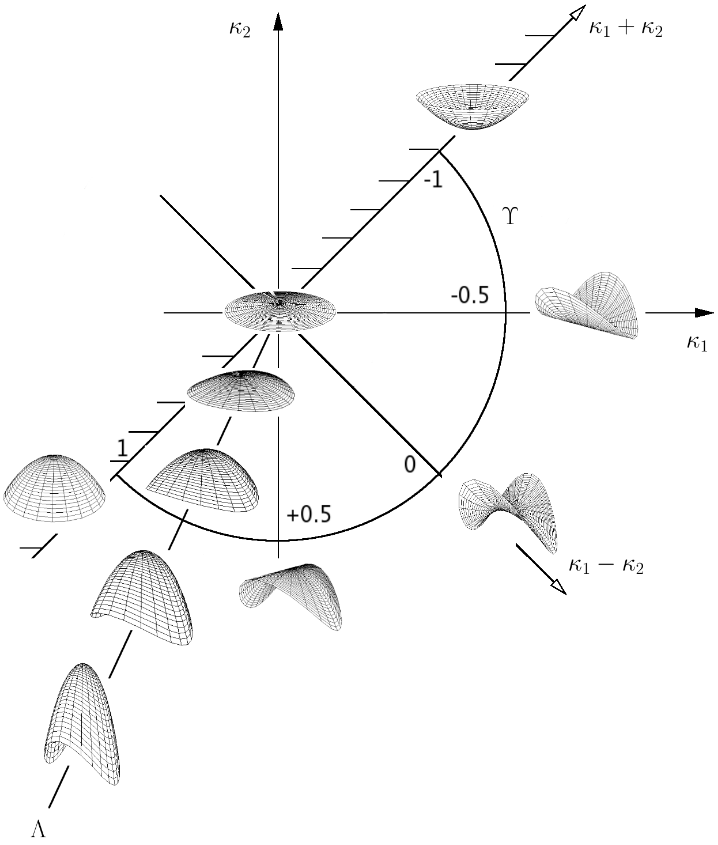


Figure 2.6: Representative local shape in the combined (κ_1,κ_2) plane

geometrical invariants of the surface, such as the square root of its area (A), the cubic root of its volume (V), etc. Presently we define

$$C \equiv \mu\Lambda, \qquad \mu \equiv 3\,\frac{V}{A}. \tag{2.7}$$

For the sphere, $C = 1$. The definition of a volume implies that the structure under consideration is a closed surface. Thus, only closed surfaces educed from the scalar field are studied. For periodic domains only those structures with infinite extent will not be closed. All others, following periodic reconnection, will be closed. For non-periodic domains of limited extent, those structures intersecting boundaries will not be closed, but they could still be considered in the analysis by closing them either with the boundaries that they intersect or with their mirrored extension across those boundaries, for example.

Another dimensionless global parameter useful in the characterization of the geometry of the educed closed structures is

$$\lambda \equiv \sqrt[3]{36\pi}\,\frac{V^{2/3}}{A}. \tag{2.8}$$

It represents the stretching of the structure; the lower its value, the more stretched the structure is. For the sphere $\lambda = 1$.

The area-based jpdf $\mathcal{P}(S,C)$, $\int\int \mathcal{P}(S,C)\,\mathrm{d}S\mathrm{d}C = 1$, contains non-local information on the geometry of the surface. $\mathcal{P}(S,C)$ can be geometrically interpreted as a representation of how the local shape, S, is distributed across the different (relative) length scales present on the surface, given by C, in terms of area coverage. For closed surfaces, their geometry and topology are related by the Gauss–Bonnet theorem, which imposes an integral constraint on the area-based joint probability density function of S and C (see Appendix E for details).

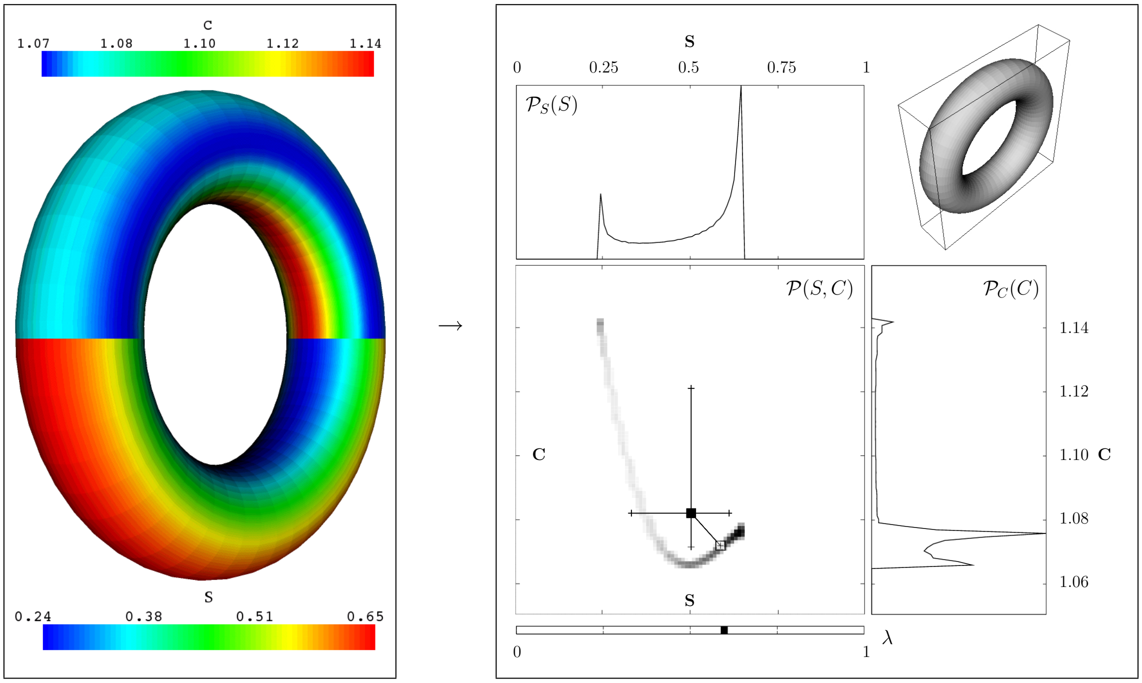
2.2.3 Signature of a structure

We consider $\mathcal{P}(S,C)$ plus its associated one-dimensional marginal pdfs,

$$\mathcal{P}_{\mathcal{S}}(S) = \int \mathcal{P}(S,C)\,\mathrm{d}C, \qquad \mathcal{P}_{\mathcal{C}}(C) = \int \mathcal{P}(S,C)\,\mathrm{d}S, \tag{2.9}$$

to be the signature of the structure. This is complemented with its area A and λ, representing the stretching of the structure. We find it useful to display $\mathcal{P}(S,C)$ mapped onto the (S,C)-plane with greyscale rendering of $\mathcal{P}$; white $\equiv 0$, black $\equiv \max(\mathcal{P})$. Additionally, we plot $\mathcal{P}_S(S)$ and $\mathcal{P}_C(C)$ on the S (top) and C (right) axes respectively; see Figure 2.7 for an example. This geometrical characterization is based on properties of the structure that are invariant with respect to translations and rotations of the reference system, and therefore, are suited for comparing structures based on their geometry, the basis of the next step of this methodology: the classification of structures.

Several methods have been proposed in the computer graphics literature for estimating curvatures of a discretized surface (such as the ones that represent our structures in the computational domain).



(a) (b)

Figure 2.7: Example of a three-dimensional surface (a) with S and C mapped onto it (bottom and upper halves, respectively) and its corresponding signature (b), for which a projection of the three-dimensional physical structure is shown at the top-right corner, its area-based joint two-dimensional probability density function (in terms of S and C) is presented in the bottom-left area, while the marginal probability density functions of both S and C, are drawn at its top and right sides, respectively. The value of the stretching parameter, λ, is represented below the jpdf by a black bar (in a scale from 0 to 1). Mean and feature centers, as well as upper and lower distances for each variable of the jpdf, are superimposed to the jpdf, as the filled and hollow squares, respectively (refer to §2.3)

A subset of them, applied to the case in which the discretized surface is a triangular mesh, was here implemented and tested (Chen & Schmitt, 1992; Dong & Wang, 2005; Taubin, 1995; Meyer et al., 2003). Finally, a modification of the algorithm proposed by Dong & Wang (2005) (based on Chen & Schmitt (1992)) is used. The only modification is the way in which the normal vector to each face of the discretized surface is computed, following the method proposed by Chen & Wu (2004).

2.3 Classification of structures

A process of classification assigns different elements of a given set to groups based on the similarities of their signatures. In our system, the elements to be classified are the educed structures, and the signatures are given by $\mathcal{P}(S,C), \mathcal{P}_S(S), \mathcal{P}_C(C)$, and λ, obtained in the characterization step for each structure. Among the different approaches to the problem of classification, we seek those involving as little a priori knowledge as possible of the relationships governing the different groups and of the number of groups present in the set of elements under evaluation. This leads to the utilization of learning-based clustering techniques. The idea behind this approach is to be able to detect other types of geometries apart from the known tube-like and sheet-like structures in turbulence databases, should they exist, by not imposing strong assumptions on the groups.

2.3.1 Clustering algorithm

The clustering algorithm used in this classification step combines several techniques found in the data mining, pattern recognition, and artificial intelligence literature (see, for example, Berkhin, 2002, for a survey of such clustering techniques). It is a locally scaled spectral partitional clustering algorithm that automatically determines the number of clusters. Its main steps are summarized below using the notation proposed by Ng et al. (2001) in their NJW algorithm, that conforms the core of our technique. Additions, particularizations, and modifications to the NJW algorithm are also described below. In what follows we denote a set of N structures at a particular scale by the N elements $E = \{e_1, \ldots, e_N\}$. For each member of this set we construct a set of parameters $\{p[k], k = 1, \ldots, N_P\}$ which will serve as the contracted computational signature of the structure. These will be a finite set of moments of $\mathcal{P}(S,C), \mathcal{P}_S(S), \mathcal{P}_C(C)$, to be defined subsequently, together with λ. The $p[k]$ will also define a *feature space of parameters* in which the elements e_i are mapped. Typically $N = O(10^2 - 10^5)$, depending on the scale, and it will be seen that $N_p = 7$.

1. Start from a set of N elements $E = \{e_1, \ldots, e_N\}$ and their corresponding contracted signatures $\{p_{e_i}[k], k = 1, \ldots, N_p\}$.

2. Construct the *distance matrix*, $d_{ij} = d(e_i, e_j)$, $e_i, e_j \in E$. The element d_{ij} of the distance matrix measures dissimilarity between the two elements e_i and e_j of E, based on their signatures.

Presently we define the distance

$$d_{ij} = F(\{p_{e_i}[k] - p_{e_j}[k], k = 1, \ldots, N_p\}) \tag{2.10}$$

where $p_{e_l}[k]$ is the kth parameter associated with element e_l. The weighting function F defines a distance in that space of parameters. For example, a functional dependence of F of the form $F(\boldsymbol{x}) = \left(\sum_i x_i^2\right)^{1/2}$ defines a Euclidean distance in the feature space of parameters.

3. Construct a *locally scaled affinity matrix* $\hat{\boldsymbol{A}} \in \mathbb{R}^{N \times N}$ defined by

$$\hat{A}_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma_i \sigma_j}\right) \tag{2.11}$$

where σ_l is a *local scaling parameter* introduced by Zelnik-Manor & Perona (2005) and defined as the distance of the element e_i to its rth closest neighbor, denoted by $e_{r,i}$, $\sigma_i = d(e_i, e_{r,i})$ (a value of $r = 7$ is used, following Zelnik-Manor & Perona, 2005). The purpose of introducing a local scaling parameter is to take into consideration the multiple scales that can occur in the clustering process, which is important, for example, when tight clusters are embedded within more sparse background clusters. Note that the elements of the diagonal of $\hat{\boldsymbol{A}}$ are null.

4. Normalize $\hat{\boldsymbol{A}}$ with a diagonal matrix $\boldsymbol{D}$ such that $D_{ii} = \sum_{j=1}^{N} \hat{A}_{ij}$, obtaining the *normalized locally scaled affinity matrix* $\boldsymbol{L} = \boldsymbol{D}^{-1/2} \boldsymbol{A} \boldsymbol{D}^{-1/2}$

5. For N_C varying between the minimum and maximum number of clusters considered, do the following loop:

   (i) Find the N_C largest eigenvectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_C}\}$ of $\boldsymbol{L}$ and form the matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_C}] \in \mathbb{R}^{N \times N_C}$. This step constitutes the spectral part of the algorithm. It is intended to map the elements e_i onto a different eigenspace where clusters can be better identified. It can be considered a pre-clustering step that, combined with the local scaling explained in one of the previous steps, allows clustering of elements with more complicated relationships among them (and to other clusters) than traditional clustering techniques that do not use these features. For example, concentric clusters can be easily educed by means of spectral clustering.

(ii) Re-normalize the rows of $\boldsymbol{X}$ so that they have unitary length, obtaining the matrix $\boldsymbol{Y} \in \mathbb{R}^{N \times N_C}$ as $Y_{ij} = X_{ij} / \left(\sum_j X_{ij}^2 \right)^{1/2}$.

(iii) Treat each row of $\boldsymbol{Y}$ as a point in $\mathbb{R}^{N_C}$ and cluster them into N_C clusters via K-means algorithm.

(iv) Assign the original element e_i to cluster k if row i of $\boldsymbol{Y}$ was assigned to cluster k in the previous step.

(v) Obtain *optimality score* for this number of clusters N_C (see §2.3.3).

6. After the previous step has been done for all the possible numbers of clusters under evaluation, determine the optimum number of clusters based on the minimization of the optimality score for each one of the possible numbers of clusters (as will be described in §2.3.3).

The K-means clustering algorithm mentioned above is one of the simplest partitional clustering techniques available. It first initializes the cluster centers (for a given number of them). Then it assigns each element to the cluster with the closest centroid to that element. After all elements have been assigned, it recalculates the position of the cluster centers. The last two steps are repeated until the cluster centers no longer move. Different implementations of the K-means clustering algorithm differ mainly in the initialization of the cluster centers: we choose the initial position of the first cluster center randomly among all the elements; initial positions of subsequent cluster centers are obtained as the farthest elements to the previously assigned cluster centers. Several initializations following that procedure are performed to avoid local minima.

2.3.2 Feature and visualization spaces; definition of the $p[k], k = 1, ..., N_p$

The selection of the $p[k]$ used to define the feature space plays a decisive role in the classification step. Each structure will be represented by a point in that feature space and its distance to the other points will define the similarity to their corresponding structures. The number of parameters (dimensions of feature space) should be sufficiently large to distinguish satisfactorily relevant groups of structures, but at the same time, it should be kept as small as possible to avoid the so-called 'curse of dimensionality' (see Bellman, 1961) that affects unsupervised learning algorithms, like the clustering method used in this methodology, compromising its success by making the points too disperse in such high-dimensional space. The set of (seven) parameters chosen here for each element e_i of E is

$$\{p[k], k = 1, \ldots, 7\} \equiv \{\hat{S}, \hat{C}, \lambda, d_u^S, d_l^S, d_u^C, d_l^C\} \tag{2.12}$$

where $\hat{S}, \hat{C}$ denotes the *feature center* of $\mathcal{P}(S, C)$ and d_u^S, d_l^S, d_u^C, d_l^C are the *upper* and *lower distances* of the jpdf in each variable. The feature center ($\hat{S}$,$\hat{C}$) takes into account the asymmetry of the jpdf, correcting the mean center ($\bar{S}, \bar{C}$) so that the feature center lies closer to the region of higher density of the jpdf. The upper and lower distances, d_u and d_l, can be regarded as the r.m.s. of the part of the jpdf above and below, respectively, its mean value. A graphical example can be seen in Figure 2.7, where the mean and feature centers have been superimposed to their corresponding jpdf. Definitions of feature center, upper and lower distances, together with a representative one-dimensional example can be found in Appendix F.

Based on the idea of the feature space of parameters used for educing clusters of similar structures, we define a *visualization space*, intended to provide a graphical representation of the distribution of individual structures in a three-dimensional space, providing qualitative and quantitative information. In general, the higher-dimensional character of the feature space prevents its use as visualization space, but the utilization of *glyphs*, scaling, and coloring allows more than just three dimensions to be represented in the visualization space.

We define the three axes of the visualization space by $\hat{S}$, $\hat{C}$, and λ. Owing to the choice of non-dimensionalization of the curvedness and the normalization factors (see §2.2), as well as the intrinsic meaning of the shape index, curvedness, and stretching parameter, it is possible to identify regions in the visualization space with a particular geometrical meaning for those structures whose representation lies in them. For example, blob-like structures occupy the region near the point

$(1,1,1)$ (which corresponds to spheres); tube-like structures are localized near the $(1/2,1,\lambda)^2$ axis (λ being an indication of how stretched the tube is) and the transition to sheet-like structures occurs as the curvedness and λ decrease. The plane $\hat{C}=0$ is the limiting case of planar structure; furthermore, any structure composed of (predominantly) planar regions, thus featureless in the curvature sense, will have a (nearly) nil $\hat{C}$, independently of its relative aspect ratios, that will nevertheless affect its λ value. See Appendix G for an analysis of these limiting values. Throughout this thesis, the visualization space is presented by a set of two-dimensional projections (see Figure 2.8 for an example).

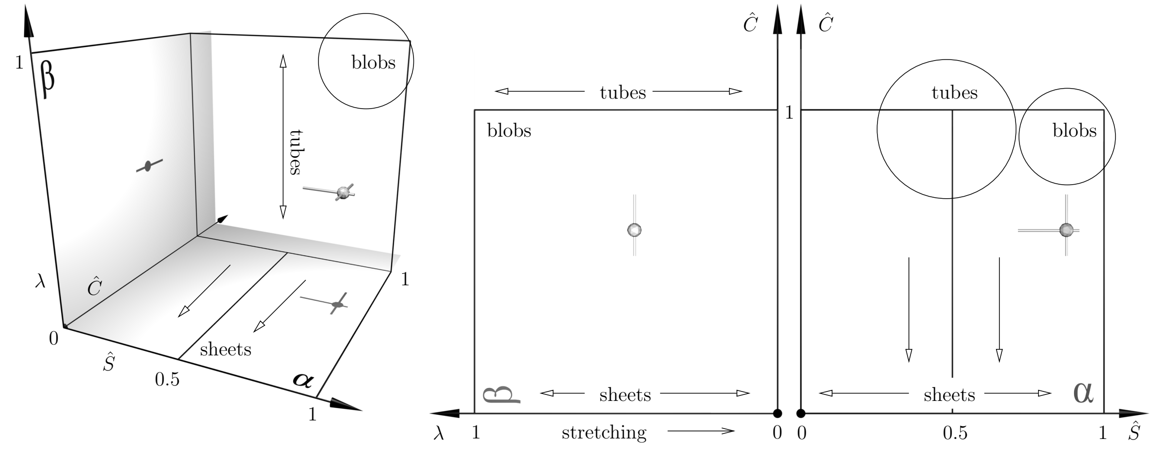


Figure 2.8: Projections of the *visualization space* with the predominantly blob-, tube- and sheet-like regions sketched: three-dimensional perspective projection (left), two-dimensional orthogonal projections (right) of the planes β (formed by the axes $\hat{C}$ and λ) and α (formed by $\hat{S}$ and $\hat{C}$). For example, a glyph consisting of a sphere and four bars along the $\pm\hat{S}$, $\pm\hat{C}$ axes can represent nine parameters of the characterization of the corresponding structure: $\hat{S},\hat{C},\lambda$ given by the center of the sphere, upper and lower distances of S and C given by each bar, the surface area A of the associated structure, given by the size of the glyph, and the group to which the structure belongs, given by the color of the glyph

2.3.3 Optimality score: silhouette coefficient

The determination of the optimum number of clusters is based on the minimization of an *optimality score*. Different approaches have been considered. Among them, probabilistic criteria that consider the relative increment of complexity of a model (set of clusters) when another parameter (cluster)

[2] Note that a value of the shape index equal to 1/2 corresponds to locally cylindrical shapes, that are predominant in tube-like structures. The dimensionless curvedness of a straight elongated circular cylinder of radius R reduces to $C \simeq 3V/A\sqrt{2}R \approx 3/2\sqrt{2} \approx 1.06$.

is added, such as the *Bayesian Information Criterion (BIC)* BIC (Schwarz, 1978), were found to provide unsatisfactory results. This is mainly due to the use of spectral techniques, since they map the elements to be clustered onto a different eigenspace whose dimensions change with the number of clusters considered, complicating the task of comparing the goodness-of-fit for different number of clusters by such probabilistic methods. Instead, the *silhouette coefficient* (Rousseeuw, 1987), SC, is used. It is a confidence indicator of the membership of an element to the cluster it was assigned to. It is defined, for each element e_i, as $SC_i = (b_i - a_i)/\max(a_i, b_i)$, where a_i is the average distance between element e_i and other elements in its cluster, and b_i is the average distance to the items in the closest cluster. It varies from -1 (lowest membership) to $+1$ (highest membership). Being a dimensionless quantity, the mean and variance throughout all the clustered elements can be used as indicators of the optimality of the clustering, and compared among results for different numbers of clusters to determine the optimum number of them. High values of the mean silhouette indicate a high degree of membership of the elements being clustered to the clusters they were assigned, and low values of its variance indicate that the majority of elements have a similar value of the silhouette coefficient. The combination of both indications reflects a successful clustering.

Once the cluster centers have been obtained, it is also possible to retrieve the closest elements to those cluster centers among the elements being classified. These closest elements to the cluster centers can be considered as representative elements of each cluster.