Video Segmentation using Morse-Smale Complexes

Steve Gu
steve@cs.duke.edu
Project report for CPS234
Instructor: Pankaj K. Agarwal

Abstract

In this paper, we regard a space-time block of video data as a piecewise-linear 3-manifold, and we interpret video segmentation as the computation of the Morse-Smale complex for the block. In the generic case, this complex is a decomposition of space-time data into 3-dimensional cells shaped like crystals, and separated by quadrangular faces. The vertices of these are Morse critical points. In practice, video data is discrete, and we devise an algorithm that adapts Morse theory to this reality. The resulting cell decomposition provides an efficient representation of the space-time data, and separates topology from geometry. Critical points paired by the edges of the complex identify topological features and their importance. We use topological persistence over the Morse-Smale complex to build a video segmentation hierarchy through successive simplification. This hierarchy provides a new, promising handle for visual saliency, useful for video summarization, simplification, and recognition. In the report, we first give a brief introduction to the Morse theory in d dimensions and present several theoretical fundamental results derived from the theory. We then present an O(n log n), practically efficient algorithm to construct the 3D Morse-Smale complex, and show results on real video.

1. Introduction

The importance of digital video has recently exploded thanks to YouTube, cameras on nearly every cell phone, live satellite imagery, military surveillance drones, new motion-aware medical imaging devices, and much more. All this rich video content waits to be organized, retrieved, edited, and generally analyzed for recognition or other purposes.

Analysis requires in turn a representation that captures the visual *structure* of the data. A structured representation splits video into different takes and scenes, describes camera motion separately from the motion of objects in the scene, separates foreground from background, and isolates different objects from one another. Partial structure in video has been captured sometimes through the definition of points of interest in the spatio-temporal xyt volume of moving images. These points, however, provide a sparse representation of the data. Segmentation is a more complete description, and captures structure without sacrificing detail. However, a single segmentation incurs the cost of an irrevocable commitment to a specific level of detail.

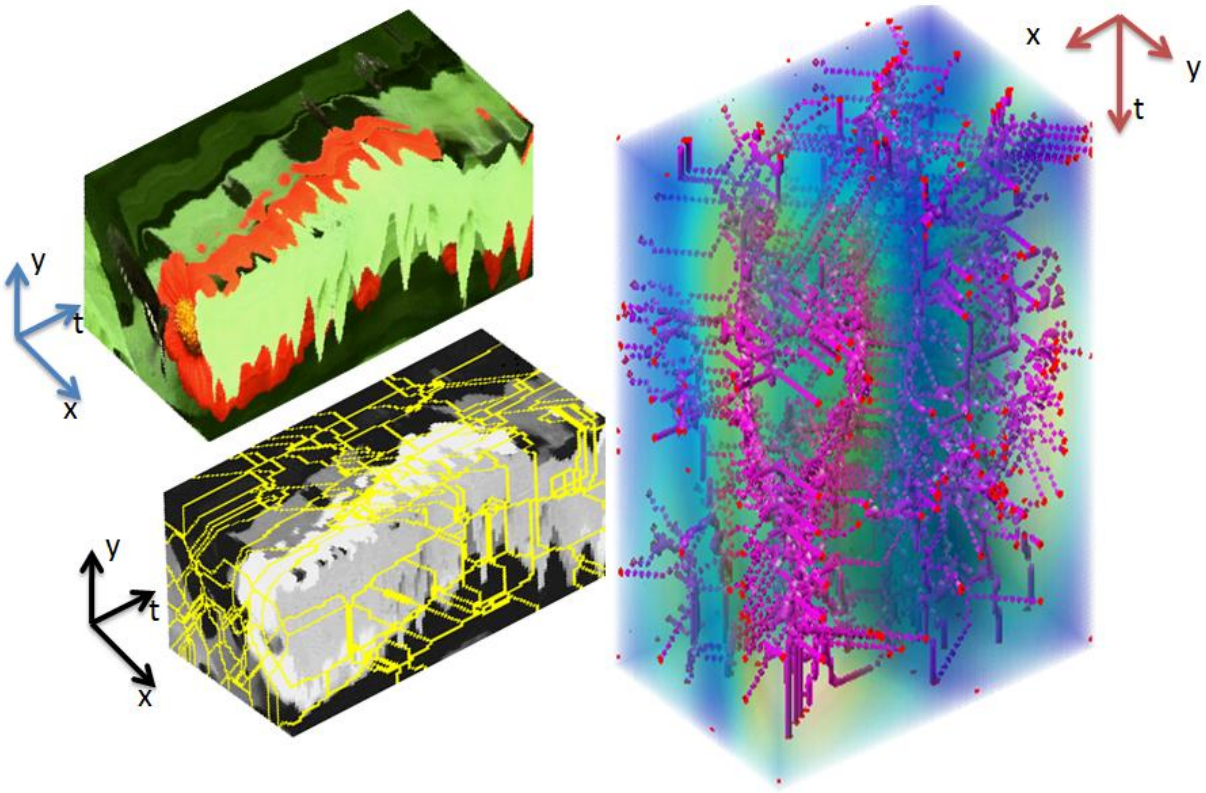

Figure 1. Morse-Smale cell decomposition of video streams. The top left image shows the video cube viewed in three dimensional space. The right image plots the density of the video sequences overlayed with Morse critical points, which densely cover the whole video volume and cluster naturally in the motion area. The bottom left image shows the original video cube overlayed with Morse-Smale cell boundaries. The volume enclosed by the boundaries are formed as Morse-Smale 3-cells.

Instead, a good structured representation should provide a handle on level of *detail*, so that full information can be retained in regions of interest, but only an approximate sketch is preserved of surrounding areas. This is useful not only for reasons of efficiency in storing, transmitting, and processing video, but also to control emphasis, to layer analysis into different levels of understanding, and to defer a decision of what is important to each of a variety of applications. In computer vision, detail has been often identified with scale, and so-called scale-space approaches have achieved simplification by blurring. However, blurring obliterates small detail together with the boundaries of large regions. A more discriminating treatment of scale would instead eliminate small detail while keeping large parts crisply delineated.

In this paper, we propose a representation of video that captures structure and affords flexible control of detail, without sacrificing crispness. Specifically, we represent structure through the so-called *Morse-Smale complex* of a scalar function $f(x, y, t)$ associated with the video data, and we obtain a *hierarchy* of increasingly simple complexes through the process of *topological simplification*.

A (three-dimensional) *complex* is a set of compact regions in (x, y, t) space called the *cells*. Cells are bounded by two-dimensional faces, bounded by one-dimensional curve segments, bounded by isolated points. In a Morse-Smale complex, in particular, the isolated points are maxima, minima, and saddles of the function f, and the curve segments and faces are loci that follow the gradient of f away from the

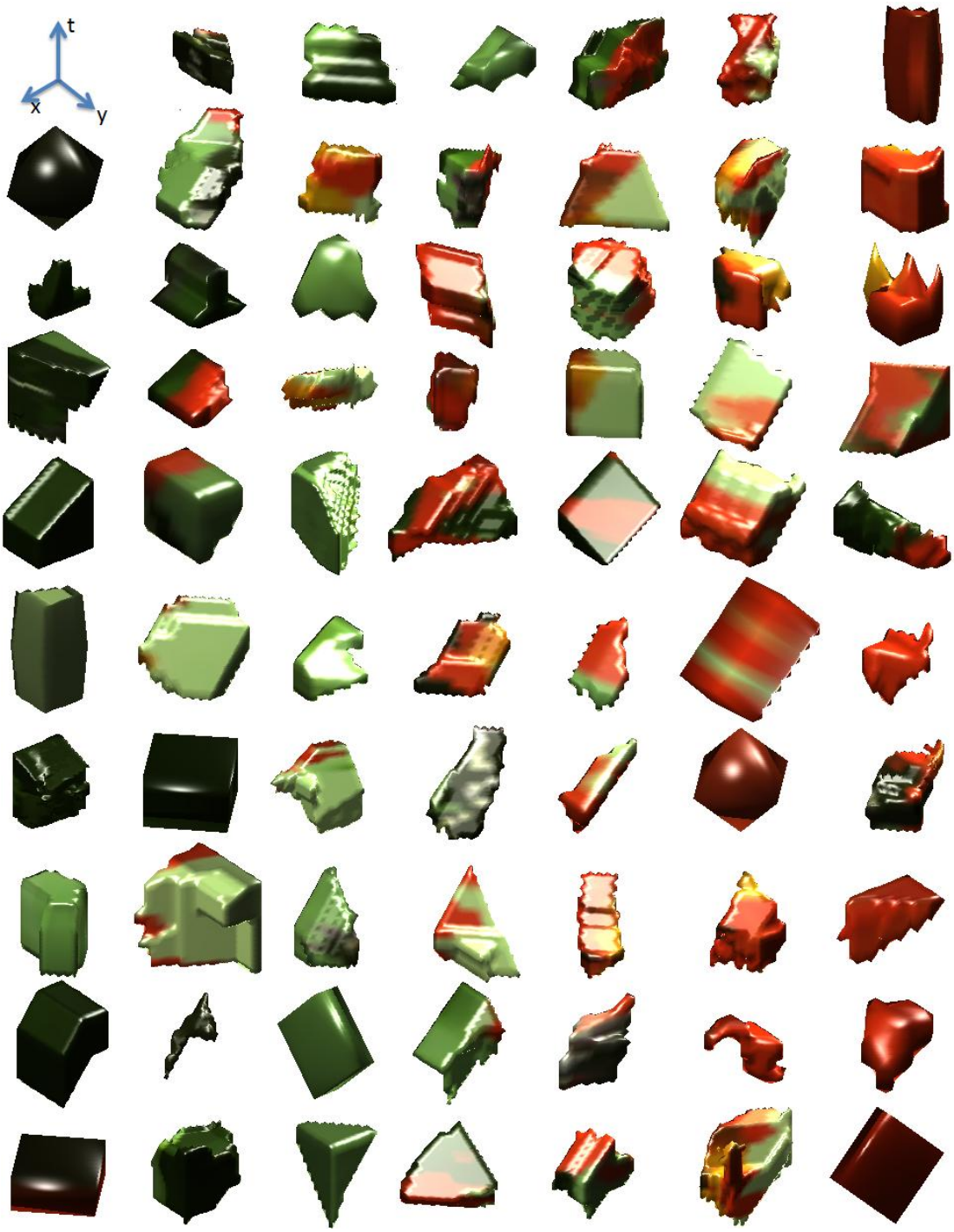

Figure 2. Sample Morse-Smale cells generated from video sequences. 69 among 947 Morse-Smale cells from a video sequence of size $54 \times 80 \times 102$ are randomly selected to give a flavor of what Morse-Smale cells look like. Although the abstract setting for each Morse-Smale cell is cubic alike, the geometric realization is usually complicated in 3D space due to different behaviors of gradient flow

saddle points. The intent is to choose f so that the cells of the complex track through time in image regions of approximately uniform appearance. While these regions do not correspond to "objects" *per se*, later analysis of their motion can group them into regions that share the same fate over time. Figure 1 shows one of the video sequences used in this paper and its Morse-Smale cell decomposition. Note that here we chose the function f conceptually as the density of the video volume which is defined in section 5. Figure 2 shows some sample Morse-Smale cells constructed from the density of the video sequence. Note that the geometric realization of each Morse-Smale cell can appear very different in shape and volume due to the rich flexibility in the gradient field of the underlying spatial temporal domain.

Topological simplification is based on the technical notion of *persistence*, which measures the difference in the value of f between extrema (maxima or minima) and neighboring saddle points. Regions of low persistence can be merged into neighboring regions, thereby simplifying the Morse-Smale complex, and a filtration from full detail to a trivial, single-cell complex can be obtained by merging regions in order of increasing persistence. The intent here is for persistence to measure the importance of a region, or its resilience to perturbation. A different filtration would rank regions by size, rather than persistence, thereby embodying the notion of detail more directly.

The seeds of several of these ideas have already appeared in applications other than vision and, in some form, in the field of computer vision itself. The next Section traces these intellectual relations. Section 3 introduces the celebrated Morse theory along with several fundamental results derived from the theory. Section 4 describes an algorithm that builds 3D Morse-Smale complexes efficiently and performs persistence-based simplification successively, leading to a segmentation hierarchy adapted to video streams. Results and conclusions are presented in Sections 5 and 6.

2. Related Work

Segmenting video stream into interesting "events" or spatial temporal interest points has gained much recent attention [16, 15, 23, 29, 6, 24]. These approaches extract representative points or regions that summarize the rich information contained in the video data. For example, in [16], the Harris corner detector is extended to the (x, y, t) spatio-temporal domain, and local maxima with high eigenvalues are extracted as interest points. In [15], a "visual event" is detected through the computation of space-time volumetric features. In [23], the SIFT operator [18] is extended to 3D in order to represent locally distinctive features. In [24], local regions that have structure similar to that of regions found in a sample query video are determined with a similarity measure based on the eigenvalues of a Gram matrix. All these approaches find isolated points or small, separate regions. Everything between them is lost.

Segmentation, on the other hand, outlines structure, but preserves full information. In our discussion, we focus on approaches that segment the (x, y, t) volume as a whole [12, 5, 11, 26, 1, 21, 25], rather than individual frames. For example, in [12], a mixture of 3D Gaussians is fit to the data through expectation maximization. In [5], segmentation is performed through hierarchical mean-shift analysis [3]. In [11], spectral graph methods are used for segmentation, and the Nyström method is used to accelerate the computation. All these approaches yield a single segmentation for a given data volume. For the reasons mentioned in the introduction, a hierarchy of segmentations is often preferable instead.

Our proposed way to achieved this is based on recent developments in computational topology, which in turn rest on the much older field of Morse theory (see [20] and [19] for good introductions). The recent developments aim at applying Morse theory, developed for smooth functions, to the piecewise-linear triangulations typically used to interpolate discrete data. Edelsbrunner *et al.* [9, 10, 8] define the Morse-

Smale complex for piecewise-linear two-dimensional and three dimensional manifolds by considering these as limits of suitable series of smooth functions. They also give an efficient algorithm to compute the Morse-Smale complex, restricted to edges of the input triangulation, and to build a hierarchical representation by repeated cancellation of pairs of critical points. In [2], the authors improve on the algorithm and describe a multiresolution representation of the scalar field. Recent applications using Morse-Smale complex to visualize scientific data can be found in [13, 14]. All these algorithms make the technical, simplifying assumption that the underlying function satisfies the Morse-Smale criterion. This requires generic function values, isolated critical points, and transversality of the cell decomposition. All these assumptions are routinely violated in video data. For instance, saturation of image values leads to maximal and minimal regions, rather than isolated points, and quantization of pixel values implies non-generic function values.

Paris and Durand [22] introduce Morse theory into computer vision, and use the notion of topological persistence for hierarchical segmentation using mean shift. However, they apply these concepts to a "density" function in a five-dimensional space that combines the geometric, x, y position of a pixel with its r, g, b color components. Also, rather than the Morse-Smale complex, they construct a cell complex that is very similar to the watershed transform [28], with which they also share the general philosophy for simplification. While it is known[4] that the Morse-Smale complex can be computed by the intersection of two watershed transforms (for f and $-f$), the computations become very different in three dimensions. While Paris and Durand do perform experiments on video segmentation, they segment each frame separately instead of treating the (x, y, t) volume as a single object, and thereby lose any notion of inter-frame consistency.

The notion of topological persistence is used directly in image segmentation in [17], where a split-then-merge paradigm is applied to a Delaunay triangulation of the edge map, which is simplified based on persistence. In contrast with this and all the existing work on image or video segmentation, we evaluate the Morse function over functions defined on $\mathbb{R}^3$. This is a non-trivial extension from $\mathbb{R}^2$, and requires special care for image data, which as mentioned earlier do not satisfy the Morse-Smale conditions assumed in the computational topology literature. We also compute the Morse-Smale complex explicitly, generate a hierarchy of simplification based on the notion of topological persistence, and analyze the implications for video analysis.

3. An Introduction to Morse Theory

We regard interest point detection and spatiotemporal segmentation as different ways of *simplification* of the video data. In some sense, we can think of those interest points as *topological* features of the underlying data and the shapes of each region in the segmentation as the *geometrical* features. We introduce the use of Morse theory to interpret the initial video segmentation as a Morse-Smale complex where each cell in the Morse-Smale complex groups gradient flows sharing the same two critical points. Using this result, we not only link the topological features with geometric features but also identify the connection among those topological features. Morse theory has been well studied in the context of smooth scalar functions. However, the video stream data is indeed a set of discrete samples over the spatial-temporal domain. We follow the description of Morse theory for piecewise linear 3-manifolds presented in [8] and apply it to the spatiotemporal segmentation. But first, let us briefly review the Morse theory in finite d dimensions in general and several key concepts exploited in this paper as well.

3.1. Morse Functions

A d-dimensional manifold $\mathbb{M}$ of class C^∞ is a generalization of the concept of smooth surface to a higher dimension d where for each $p \in \mathbb{M}$, there exists a smooth (C^∞) coordinate system: $(x_1, x_2, ..., x_d)$ isomorphic to the Euclidean space.

Definition 3.1. Let $f : \mathbb{M} \to \mathbb{R}$ be a smooth map. Given a local coordinate system $(x_1, x_2, ..., x_d)$, the *gradient* of f at $p \in \mathbb{M}$ is: $\nabla f(p) = \left[\frac{\partial f}{\partial x_1}(p), \frac{\partial f}{\partial x_2}(p), ..., \frac{\partial f}{\partial x_d}(p)\right]^T$. p is *critical* if and only if $\nabla f(p) = 0$. Otherwise it is *regular*. A real number c is a *critical value* of f if $f(p_0) = c$ for some critical point p_0.

Definition 3.2. The *Hessian* of f at a critical point p_0 is the $d \times d$ matrix:

$$H_f(p_0) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(p_0) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(p_0) \\ & \ddots & \\ \vdots & \frac{\partial^2 f}{\partial x_i \partial x_j}(p_0) & \vdots \\ & \ddots & \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(p_0) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(p_0) \end{pmatrix} \tag{1}$$

A critical point p is *non-degenerate* if $\det H_f(p_0) \neq 0$.

Note that whether a critical point is degenerate doesn't depend on the choice of coordinate system as for any two coordinate system $(y_1, y_2, ..., y_d)$ and $(x_1, x_2, ..., x_d)$, the corresponding Hessian $\mathcal{H}_f(p_0)$ and $H_f(p_0)$ are related as:

$$\mathcal{H}_f(p_0) = J(p_0)^T H_f(p_0) J(p_0) \tag{2}$$

where $J(p_0)$ is the *Jacobian* matrix of the coordinate transform from $(y_1, y_2, ..., y_d)$ to $(x_1, x_2, ..., x_d)$:

$$J(p_0) = \begin{pmatrix} \frac{\partial x_1}{\partial y_1}(p_0) & \cdots & \frac{\partial x_1}{\partial y_d}(p_0) \\ & \ddots & \\ \vdots & \frac{\partial x_i}{\partial y_j}(p_0) & \vdots \\ & \ddots & \\ \frac{\partial x_d}{\partial y_1}(p_0) & \cdots & \frac{\partial x_d}{\partial y_d}(p_0) \end{pmatrix} \tag{3}$$

Therefore, $\det \mathcal{H}_f(p_0) \neq 0$ if and only if $\det H_f(p_0) \neq 0$ since $\det J(p_0) \neq 0$.

Definition 3.3. (Morse Function) The smooth function f is called a *Morse function* if all critical points are non-degenerate.

We have the following Morse lemma[20] basically saying that all the critical points of f are isolated.

Lemma 3.4. *(Morse Lemma) Let p_0 be a non-degenerate critical point of $f : \mathbb{M} \to \mathbb{R}$. Then we can choose a local coordinate system $(x_1, x_2, ..., x_d)$ about p_0 such that: $f(x_1, x_2, \ldots, x_d) = f(p_0) \pm x_1^2 \pm x_2^2 \pm \ldots \pm x_d^2$*

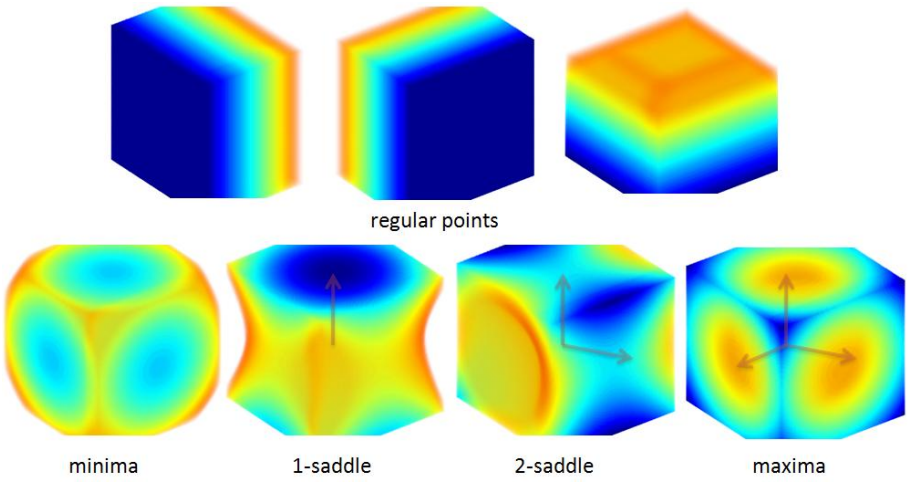

Figure 3. Local pictures of regular points and different types of critical points. The top row shows the typical neighborhood of regular points. The bottom row shows different types of critical points determined by the number of negative eigenvalues.

We can classify all the critical points based on the following definition:

Definition 3.5. Let $\lambda_1, \lambda_2, \ldots, \lambda_d$ be the eigenvalues of $H_f(p)$, the *index* $i_f(p_0)$ of f at the critical point p_0 is defined to be: $i_f(p_0) \triangleq \sum_{j=1}^{d} \text{sign}(-\lambda_j)$ where $\text{sign}(x) = 1$ if $x > 0$ and 0 otherwise.

Note that $i_f(p_0)$ is equal to the number of principal directions under which f decreases in the neighborhood of p_0. Therefore, the notion of index classifies the types of critical points of a Morse function in d dimensions into $d+1$ categories: *minima* have index 0, *maxima* have index d and *k-saddles* have index k. We also call *minima* and *maxima 0-saddles* and *d-saddles* for consistency. Figure 3 shows the local pictures of a regular point and four types of non-degenerate critical points in $\mathbb{R}^3$.

Before we proceed to state the fundamental results from Morse theory, one may wonder if such a Morse function exists and how many are there. The following theorem shows that the Morse function is indeed dense in the functional space.

Theorem 3.6. *(Existence of Morse Function) Let $\mathbb{M}$ be a closed d-manifold and let $g : \mathbb{M} \to \mathbb{R}$ be a smooth function defined on $\mathbb{M}$. Then there exists a Morse function $f : \mathbb{M} \to \mathbb{R}$ arbitrarily close to $g : \mathbb{M} \to \mathbb{R}$ in $\mathbb{C}^2$ sense. Since smooth functions are dense, so are the Morse functions.*

Given a Morse function $f : \mathbb{M} \to \mathbb{R}$, we consider a corresponding "gradient-like vector field", which plays an important role when studying how critical points of f are related with each other and how the manifold is decomposed to a set of handles, which we defined later. By a *vector field* on $\mathbb{M}$, we mean a correspondence which assigns to each point p of $\mathbb{M}$ a tangent vector $\mathbf{v}$ at p. More formally:

Definition 3.7. A vector field X on $U \subseteq \mathbb{M}$ with coordinate system $(x_1, x_2, ..., x_d)$ is described by

$$X = \xi_1 \frac{\partial}{\partial x_1} + \xi_2 \frac{\partial}{\partial x_2} + \ldots + \xi_d \frac{\partial}{\partial x_d} \tag{4}$$

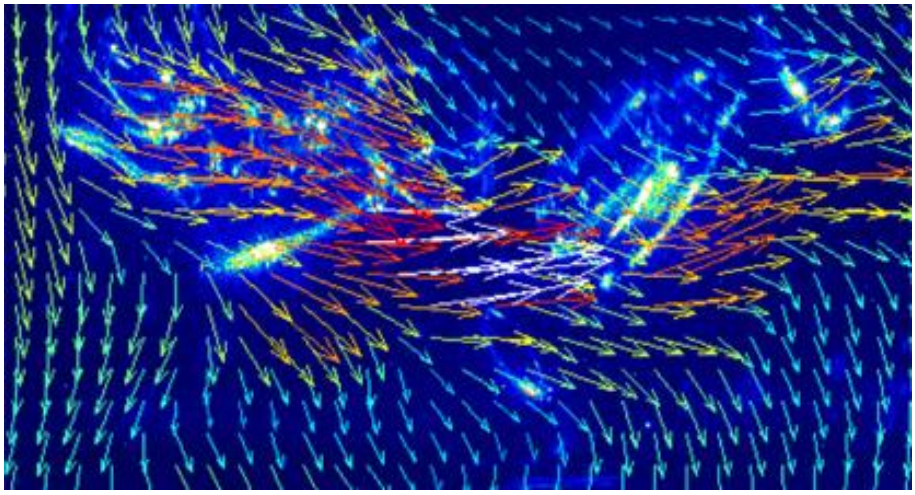

Figure 4. Gradient vector field associated with a 2D surface

In particular, let $\xi = \frac{\partial f}{\partial x_i}$, the *gradient vector field* of the function f, X_f is:

$$X_f = \frac{\partial f}{\partial x_1}\frac{\partial}{\partial x_1} + \frac{\partial f}{\partial x_2}\frac{\partial}{\partial x_2} + \ldots + \frac{\partial f}{\partial x_d}\frac{\partial}{\partial x_d} \tag{5}$$

Note that a vector field itself is sort of a differential operator. For example, we have: $X_f \cdot f = \left(\sum_{i=1}^{d} \frac{\partial f}{\partial x_i}\frac{\partial}{\partial x_i}\right) \cdot f = \sum_{i=1}^{m} \left(\frac{\partial f}{\partial x_i}\right)^2 \geq 0$ where $(X_f \cdot f)(p) > 0$ unless p is a critical point of f. Figure 4 shows the gradient vector field associated with a 2D surface. As is stated in lemma 3.4, f takes the standard form as $f = -x_1^2 - \ldots - x_\lambda^2 + x_{\lambda+1}^2 + \ldots + x_d^2$ around a λ-saddle point. Therefore, the gradient vector fields can be written as: $X_f = -2x_1\frac{\partial}{\partial x_1} - \ldots - 2x_\lambda\frac{\partial}{\partial x_\lambda} + 2x_{\lambda+1}\frac{\partial}{\partial x_{\lambda+1}} + \ldots + 2x_d\frac{\partial}{\partial x_d}$ around some critical point. The following definition generalizes the gradient vector field to the global manifold:

Definition 3.8. X is *gradient-like* vector field if $X \cdot f > 0$ away from the critical point and X is locally a gradient vector field taking the form:

$$X_f = -2x_1\frac{\partial}{\partial x_1} - \ldots - 2x_\lambda\frac{\partial}{\partial x_\lambda} + 2x_{\lambda+1}\frac{\partial}{\partial x_{\lambda+1}} + \ldots + 2x_d\frac{\partial}{\partial x_d} \tag{6}$$

Note that the gradient-like vector field X always points "upward", leading to the direction into which f is increasing. Following theorem justifies the existence of a gradient-like vector field.

Theorem 3.9. *Suppose that $f : \mathbb{M} \to \mathbb{R}$ is a Morse function on a compact manifold $\mathbb{M}$. Then there exists a gradient-like vector field X for f*

The gradient-like vectors fields admits a natural domain decomposition based on the behavior of gradient flows. Such a decomposition is also called the handle decomposition of manifolds. We need the notion of integral lines to understand the basic construction:

Definition 3.10. An *integral line* $\gamma : \mathbb{R} \to \mathbb{M}$ is a maximal path whose tangent agree with the gradient:

$$\forall s \in \mathbb{R}^3, \frac{d\gamma}{ds}(s) = \nabla f(\gamma(s)) \tag{7}$$

Both ends of an integral line are necessarily critical points of f and all the integral lines are pairwise disjoint while the union of all the integral lines cover $\mathbb{M}$, indicating that the integral lines attached with critical points can be used to decompose $\mathbb{M}$ into regions of similar flow patterns.

Definition 3.11. The *ascending* and *descending* manifolds of a critical point p_0 are defined to be the union of all the integral lines starting and ending at p_0 respectively:

$$A(p_0) = \{p_0\} \bigcup \left(\bigcup_{\lim_{s\to-\infty}\gamma(s)=p_0} \gamma \right) \tag{8}$$

$$D(p_0) = \{p_0\} \bigcup \left(\bigcup_{\lim_{s\to+\infty}\gamma(s)=p_0} \gamma \right) \tag{9}$$

Since integral lines are open at both ends, the ascending and descending manifolds obtained through concatenation of integral lines are still open. Similar to the classification of critical points, we can classify the ascending and descending manifolds based on their dimensions where $\dim A(p_0) = d - i_f(p_0)$ and $\dim D(p_0) = i_f(p_0)$. We call the ascending and descending manifolds of dimension k as ascending k-manifolds and descending k-manifolds denoted as $A^k(p_0)$ and $D^k(p_0)$. Obviously $\{p_0\} = D^k(p_0) \bigcap A^{d-k}(p_0)$. When the critical point is clear, we simply write ascending and descending k-manifolds as A^k and D^k. Note that A^k and D^k are isomorphic to each other and therefore are the same objects from the topological point of view.

3.2. Handle Decomposition of Manifolds

We now investigate how the Morse function defined on a manifold $\mathbb{M}$ determines the underlying topological structure. Let $\mathbb{M}$ be a close manifold and $f : \mathbb{M} \to \mathbb{R}$ be a Morse function on $\mathbb{M}$. Let $M_t = \{p \in \mathbb{M} | f(p) \leq t\}$ represent the level set up to t associated with $\mathbb{M}$. Naturally we obtain a *filtration* as: $\emptyset = M_0 \subseteq M_{t_1} \subseteq M_{t_2} \subseteq ... \subseteq M_{t_k} \subseteq M_1 = \mathbb{M}$ for $0 < t_1 \leq t_2 \leq \ldots \leq t_k < 1$. The goal is to study how M_t changes as the parameter t changes and in the meantime manifest the topological structures of $\mathbb{M}$. Of course the critical points play an important role here.

Theorem 3.12. *If f has no critical values in [a,b], then M_a and M_b are* diffeomorphic*: $M_a \cong M_b$*

Note that diffeomorphism is a bijection which is of class C^∞ in both directions (map and inverse map). The above theorem says that only when f reaches a critical point will the topological structure be changed. We order the critical vales of f in ascending order and label the corresponding critical points as: $p_1, p_2, \ldots, p_n$ with $c_i = f(p_i)$. Therefore, we have: $c_0 < c_1 < \ldots < c_n$ where c_0 is the minimum value and c_n is the maximum value. Let λ-handle be a direct product between a descending manifold of dimension λ and an ascending manifold of dimension $d-\lambda$: $D^\lambda \times A^{d-\lambda}$, we have the following theorem:

Theorem 3.13. *The set $M_{c_i+\epsilon}$ is diffeomorphic to the manifold obtained by attaching a λ-handle to $M_{c_i-\epsilon}$ for ϵ small enough such that there are no critical vales in $(c_i - \epsilon, c_i) \bigcup (c_i, c_i + \epsilon)$:*

$$M_{c_i+\epsilon} \cong M_{c_i-\epsilon} \bigcup_{\varphi} D^\lambda \times A^{d-\lambda} \tag{10}$$

where φ is a smooth attaching map gluing the boundary of the λ-handle to the boundary of $M_{c_i-\epsilon}$:

$$\varphi : \partial D^\lambda \times A^{d-\lambda} \to \partial M_{c_i-\epsilon} \tag{11}$$

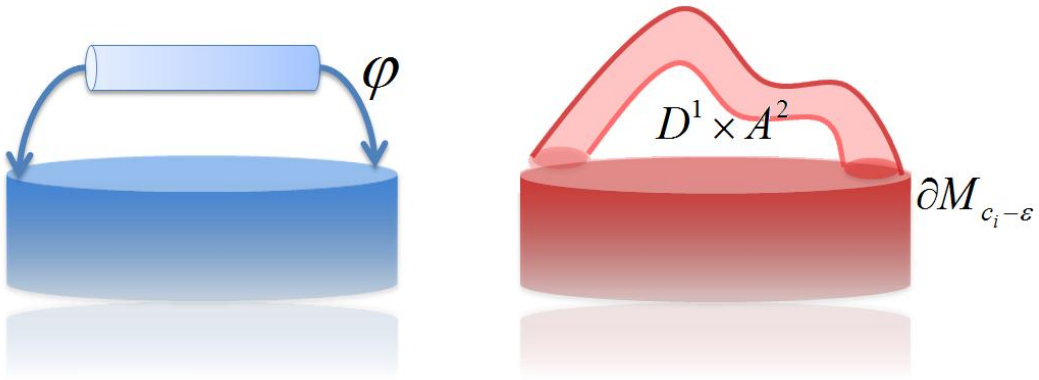

Figure 5. Attaching a 1-handle

Figure 5 illustrates the process of attaching a 1-handle to $M_{c_i-\epsilon}$ in 3-dimensional space where each $p \in \partial D^1 \times A^2$ is identified with the point $\varphi(p) \in \partial M_{c_i-\epsilon}$. Theorem 3.13 characterizes the topological change across each critical point. More vividly we can view the evolution of M_t as a successive attaching of handles of different types. We call such a collection of handles *handlebody*.

Definition 3.14. (Handlebody) A manifold obtained from A^d by attaching handles of various indices one after another:

$$A^d \bigcup D^{\lambda_1} \times A^{d-\lambda_1} \bigcup D^{\lambda_2} \times A^{d-\lambda_2} \bigcup \ldots \bigcup D^{\lambda_n} \times A^{d-\lambda_n} \tag{12}$$

is called a d-dimensional handlebody. More precisely, a handlebody is defined recursively as follows:
(i) A^d is a d-dimensional handlebody.
(ii) If $N = \mathcal{H}\left(A^d; \varphi_1, \varphi_2, \ldots, \varphi_{i-1}\right)$ is a d-dimensional handlebody, then the manifold: $N \bigcup_{\varphi_i} D^{\lambda_i} \times A^{d-\lambda_i}$ obtained from N by attaching a λ_i-handle $D^{\lambda_i} \times A^{d-\lambda_i}$ with an attaching map of class C^∞, $\varphi_i : \partial D^{\lambda_i} \times A^{d-\lambda_i} \to \partial N$, is a d-dimensional handlebody, denoted by $\mathcal{H}(A^d; \varphi_1, \varphi_2, \ldots, \varphi_{i-1}, \varphi_i)$.

With the notion of handlebody, we now present one of the fundamental results from Morse theory, which states that a manifold equipped with a Morse function admits a handlebody structure.

Theorem 3.15. *(Handle decomposition of a manifold). When a Morse function $f : \mathbb{M} \to \mathbb{R}$ is given on a closed manifold $\mathbb{M}$, a structure of a handlebody on $\mathbb{M}$ is determined by f. The handles of this handlebody correspond to the critical points of f, and the indices of the handles coincide with the indices of the corresponding critical points, that is, a critical point p_0 corresponds to a $i_f(p_0)$-handle: $D^{i_f(p_0)} \times A^{d-i_f(p_0)}$.*

3.3. Handle Sliding and Cancelation

The handlebody structure determined by $f : \mathbb{M} \to \mathbb{R}$ is unique up to a diffeomorphism. Such a decomposition indeed gives one much flexibility to rearrange the critical points without changing the diffeomorphism type of a handlebody. More precisely, let $\mathbb{M} = A^d \bigcup D^{\lambda_1} \times A^{d-\lambda_1} \bigcup D^{\lambda_2} \times A^{d-\lambda_2} \bigcup \ldots \bigcup D^{\lambda_n} \times A^{d-\lambda_n}$ be a handlebody decomposition, each handle $D^{\lambda_i} \times A^{d-\lambda_i}$ can be "slid" through a perturbation of the attaching map $\varphi_i : \partial D^{\lambda_i} \times A^{d-\lambda_i} \to \partial \mathcal{H}\left(A^d; \varphi_1, ..., \varphi_{i-1}\right)$ by an "isotopy".

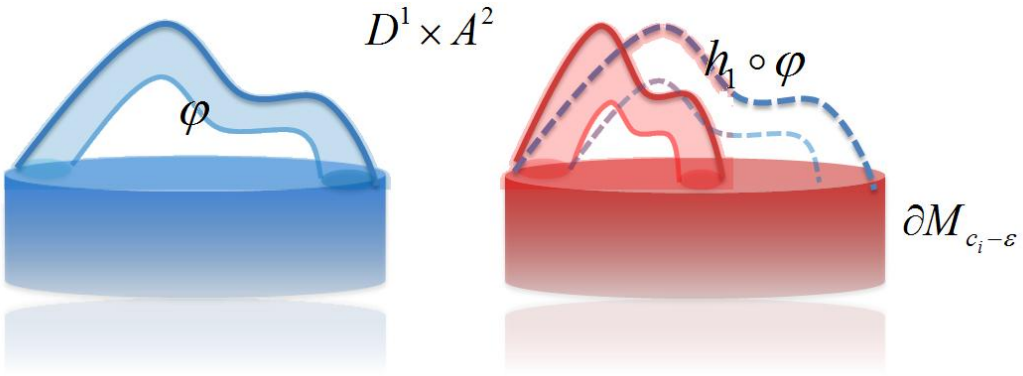

Figure 6. Handle sliding

Definition 3.16. (Isotopy). Let K be a k-dimensional manifold and $\{h_t\}_{t\in[0,1]}$ be a family of diffeomorphisms. $\{h_t\}_{t\in[0,1]}$ is called an *isotopy* if (i) $h_0 = id_K$, the identity map and (ii) The map $H : K \times [0,1] \to K \times [0,1]$, defined by $H(x,t) = (h_t(x), t)$, is a diffeomorphism. The map h_t depends on the parameter t smoothly.

Figure 6 illustrates the idea of handle sliding where the attaching position of a 1-handle is moved along the 2-manifold smoothly. The following theorem states that handles can be slid without changing the diffeomorphism type.

Theorem 3.17. *(Sliding handles). Fix one of the subscripts i of the critical points $(0 \leq i \leq n)$. Given an isotopy $\{h_t\}_{t\in[0,1]}$ of the boundary ∂N_{i-1} of a subhandlebody N_{i-1}, the attaching map φ_i of the handle $D^{\lambda_i} \times D^{d-\lambda_i}$ on N_{i-1} can be replaced by $h_1 \circ \varphi_i$. The replacement of the i-th attaching map doesn't change the diffeomorphism type of the subhandlebodies $N_j (0 \leq j \leq n)$, that is, $\forall 0 \leq j \leq n, \mathcal{H}\left(A^d; \varphi_1, \varphi_2, ..., \varphi_{i-1}, h_1 \circ \varphi_i, \varphi_{i+1}, \ldots, \varphi_j\right) \cong \mathcal{H}\left(A^d; \varphi_1, \varphi_2, ..., \varphi_{i-1}, \varphi_i, \varphi_{i+1}, \ldots, \varphi_j\right)$.*

Using theorem 3.17, one can show that all the critical points can be arranged in such a way that indices increase as the critical vales increase.

Theorem 3.18. *(Arrangements of critical points). Let $\mathbb{M}$ be a d-dimensional closed manifold and $f : \mathbb{M} \to \mathbb{R}$ a Morse function on it. Then f can be perturbed in such a way that for any critical points p_i and p_j, $f(p_i) < f(p_j)$ implies that $i_f(p_i) \leq i_f(p_j)$. In other words, any handlebody can be modified in such a way that the new one is constructed first from a disjoint union of 0-handles, and then a disjoint union of 1-handles are attached on them, and then a disjoint union of 2-handles are attached, and so forth, so that handles are attached in ascending order of indices.*

Figure 7 illustrates the process of arranging critical points such that critical points of higher indices get promoted to higher positions without changing the diffeomorphism types. In addition to the rearrangement of critical points, one can also cancel a pair of critical points without affecting the diffeomorphism type. We have the following theorem:

Theorem 3.19. *(Canceling handles) Suppose that a manifold N' is obtained from a d-dimensional manifold N with boundary by attaching a λ-handle, and suppose further that a manifold N'' is obtained from N' by attaching a $(\lambda+1)$-handle, that is: $N' = N \bigcup_{\varphi} D^{\lambda} \times A^{d-\lambda}$ and $N'' = N' \bigcup_{\psi} D^{\lambda+1} \times A^{d-\lambda-1}$.*

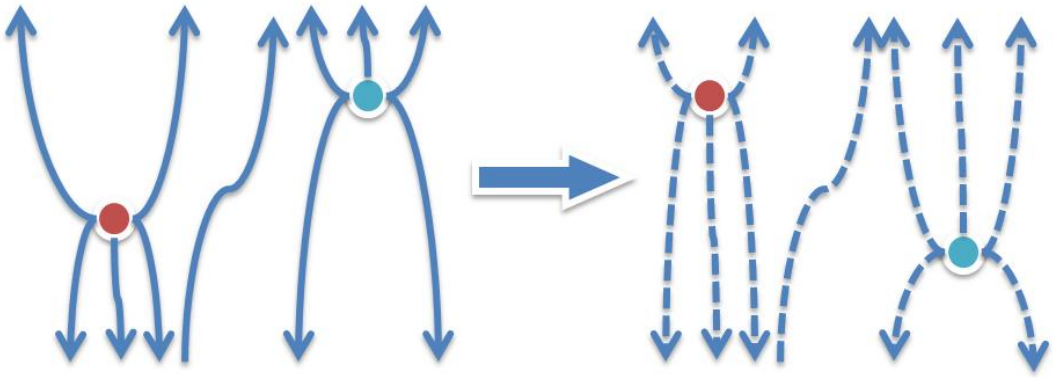

Figure 7. Arrangement of critical points

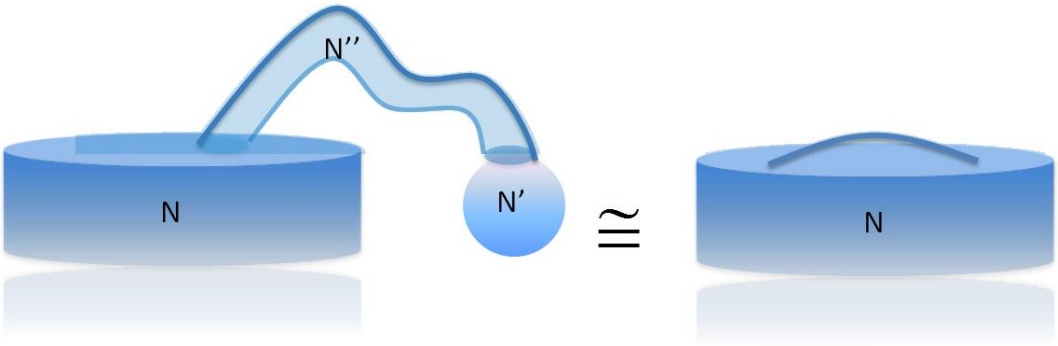

Figure 8. Canceling handles

If $\mathbf{0} \times \partial A^{d-\lambda}$ *of the* λ*-handle and* $\partial D^{\lambda+1} \times \mathbf{0}$ *of the* $(\lambda+1)$*-handle intersect transversally at a single point in the boundary* $\partial N'$*, then* N'' *is diffeomorphic to* N*:* $N'' \cong N$.

Figure 8 illustrates the process of canceling a 0-handle and a 1-handle where both handles fade out. Note that the arrangement of critical points can be used to perturb the function to be a desired Morse function as demonstrated in [7] while the handle canceling process naturally leads to the notion of topological persistence and the corresponding persistence based simplification described in section 4.2.

3.4. Morse-Smale Complex

A Morse function f is *Morse-Smale* if the descending and ascending manifolds intersect only transversally, meaning that the intersection (if not empty) between a descending i-manifold and an ascending $(d-j)$-manifold is again an open manifold of dimension $|i-j|$. The *Morse-Smale complex* is the collection of *cells* obtained by the intersection of ascending and descending manifolds over all the critical points of f. Let $\mathcal{C}(f)$ be the set of critical points of f, the Morse-Smale complex of f, $\mathcal{MS}(f)$ is then constructed as:

$$\mathcal{MS}(f) = \left\{ D(p) \bigcap A(q) | p, q \in \mathcal{C}(f) \right\} \tag{13}$$

Note that the Morse-Smale complex is also a cell system satisfying the following three properties:

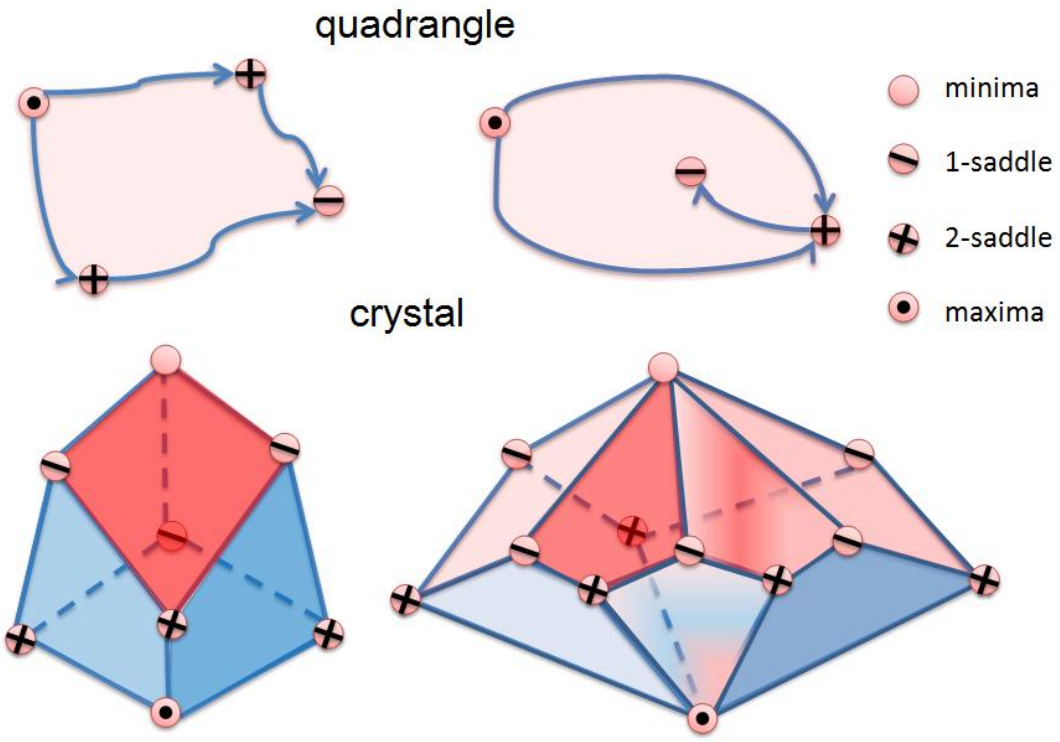

Figure 9. Quadrangles and crystals. The first column shows the typical setting of a quadrangle and a crystal and the second column shows some possible variations

(a) $\forall \gamma_1, \gamma_2 \in \mathcal{MS}(f), \gamma_1 \bigcap \gamma_2 = \emptyset$ whenever $\gamma_1 \neq \gamma_2$
(b) $\forall \gamma \in \mathcal{MS}(f)$, the *boundary* $\partial\gamma \subseteq \mathcal{MS}(f)$
(c) $\forall \gamma_1, \gamma_2 \in \mathcal{MS}(f), \partial\gamma_1 \bigcap \partial\gamma_2 \subseteq \mathcal{MS}(f)$

In particular, each cell of the Morse-Smale complex is a union of integral lines sharing the same critical points as origin p and destination q by definition. The dimension of each cell is then computed as $|i(p) - i(q)|$. We call the cells of dimension 0 to 3 *nodes*, *arcs*, *quadrangles* and *crystals* and sometimes we simply call the cell of dimension k as k-cell for generality. The quadrangle lemma [9] says that each 2-dimensional cell is indeed a quadrangle with nodes of type $i-1, i, i+1, i$ with the possibility that the boundary may be glued to itself. The prototypical case of crystal is a cube whose boundaries are the union of a set of quadrangles sharing common critical points. More interesting cases are possible. Figure 9 shows the typical setting of a quadrangle and a crystal and some possible variations.

Note that the boundary operator ∂ maps a k-cell to the union of $(k-1)$-cells, reducing the number of dimensions by one. For example, the boundary of crystals is composed of a set of quadrangles while the boundary of a quadrangles is composed of arcs, etc. In the case of video segmentation, we treat 3-cells in the Morse-Smale complex as basic units. However, in order to identify 3-cells we need to also identify 2-cells, 1-cells and 0-cells living on its boundary. Also note that the 0-cells must be the critical points of f. Based on the classification of the type of critical points, a minimum is the intersection of an ascending 3-manifold and a descending 0-manifold; a 1-saddle is the intersection of an ascending 2-manifold and a descending 1-manifold; a 2-saddle is the intersection of an ascending 1-manifold and a descending 2-manifold; and a minimum is the intersection of an ascending 0-manifold and a descending

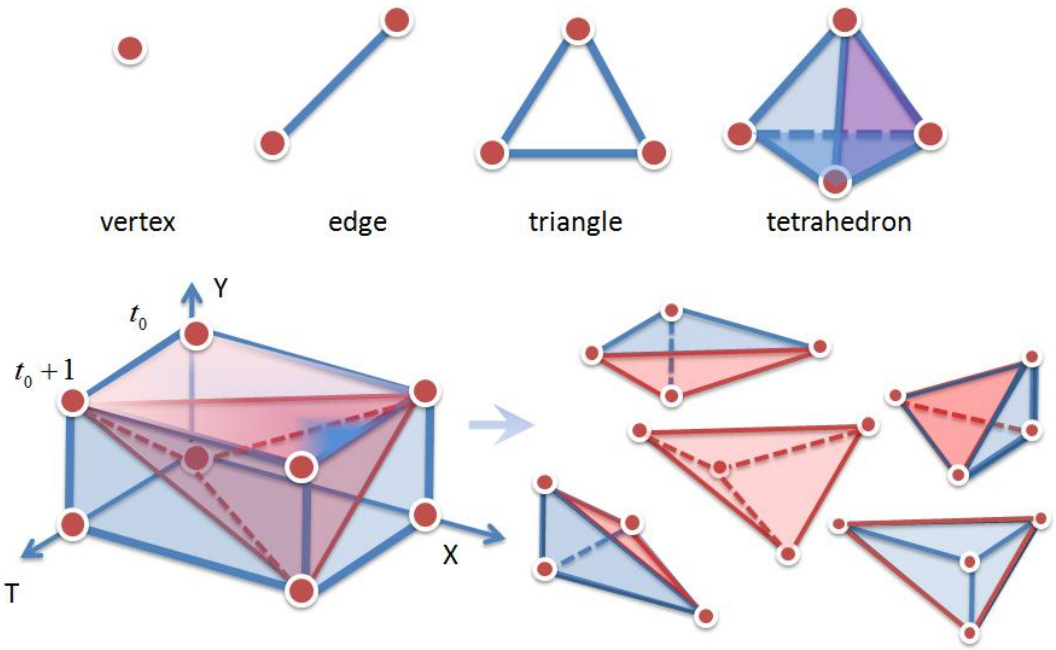

Figure 10. Triangulation of video data in the spatial-temporal domain

3-manifold. These observations lead one to design efficient algorithms to compute such a Morse-Smale complex.

3.5. Quasi Morse-Smale complex

Video data is always available as a set of discrete voxels confined by the spatial resolution in X and Y and the frame rate in T. We think of those points in the spatial-temporal domain as a set of discrete samples over a smooth manifold $\mathbb{M}$, represented by a triangulation K whose underlying space is homeomorphic to $\mathbb{M}$. K is indeed a simplicial complex consisting of simplices of dimension 0,1,2, and 3, which we refer to as *vertices*, *edges*, *triangles*, and *tetrahedra*. Although the function defined over K is not a Morse function any more, a perturbation technique can be performed [7] to ensure that all critical points are non-degenerate. A quasi Morse-Smale complex for a piece-wise linear function is a segmentation of K where each segment is made up of simplices of K. In practice, the video data is usually presented as a regular grid and hence always admits a triangulation which can be computed in linear time. Figure 10 illustrates one of the possible schemes to triangulate the video data in the spatial-temporal domain.

4. Algorithm for Quasi 3D Morse-Smale Complexes

In this section, we present an algorithm for finding quasi Morse-Smale complexes and show how to construct a hierarchy of complexes by merging 3-cells.

4.1. Build Morse-Smale Complex

Algorithm overview. The Morse-Smale complexes of f is a decomposition of K into cells consisting of integral lines that share common origin and destination. Those cells can be obtained by intersecting the

descending and ascending manifolds, where the ascending and descending manifolds meet transversally. Therefore, our approach consists of the following two steps. (1) Compute ascending manifolds, and meanwhile identify the minimum and maximum. (2) Compute descending manifolds with the guide of ascending manifolds constructed in the first step and identify the saddles. The overlay of ascending and descending manifolds form the cells of Morse-Smale complex.

For ease of discussion, we introduce the following notations. Denote K_i as simplicial complex of dimension i. Then, K_0 is the set of vertices. Denote A_i as the set of ascending i-manifolds and D_i as the set of descending i-manifolds. Let $Lk(p)$ be the set of the neighbors of point $p \in K_0$, and $ULk(p) \subseteq Lk(p)$ be the set of the neighbors p' of p where $f(p') > f(p)$. We use $Int(A(p))$ for the interior points of the ascending manifold $A(p)$ and $Int(A_i)$ for the interior points of the ascending i-manifold.

Ascending manifolds. The ascending manifolds are constructed in two steps. (1) For each minimum $p \in K_0$, we construct the ascending 3-manifolds $A(p)$ and classify each point $p \in K_0$ as one of the three types, the minimum or boundary or interior of an ascending 3-manifold. (2) We recursively compute the ascending i-manifold, where $i = 2, 1, 0$, by identifying the boundary between each pair of adjacent ascending $(i+1)$-manifolds.

The idea to construct the ascending 3-manifolds A_3 is so-called *region growing*, which is similar to watershed. First, sort the points in K_0 in the ascending order of their mapping value $f(p), p \in K_0$. Then, make a single sweep over K_0 in the ascending order of the mapping value and classify a vertex p according to the following criteria: (i) p is a minima if and only if $|ULk(p) \cap Int(A_3)| = 0$. Note that, this criteria captures the local minima p. In this case, we obtain a new ascending 3-manifold $A(p)$; (ii) p is in the interior of an ascending 3-manifold $A(p')$, i.e. $p \in Int(A(p'))$, if and only if $Int(A_3) \cap ULk(p) \subseteq Int(A(p'))$; (iii) Otherwise, p is on the boundary. The points on the boundary of the ascending 3-manifolds form the ascending 2-manifolds A_2. Denote $I(p)$ as the set of the ascending 3-manifolds that are incident on p, i.e. $I(p) = \{A(p') \in A_3 | \exists p'' \in Lk(p), p'' \in Int(A(p'))\}$. A point p is in the interior of an ascending 2-manifold if and only if $|I(p)| = 2$ and $\forall p' \in Lk(p) - Int(A_3) : I(p) \subseteq I(p')$. The boundary of the ascending 2-manifolds ∂A_2 form the ascending 1-manifolds A_1. And, we further extract the boundary (or endpoints) of A_1 as the maximum. Note that several ascending 1-manifolds may converge to points neighboring to each other due to the discrete representation of $\mathbb{M}$. In this case, we only choose the local maxima.

Descending manifolds. We compute the descending manifolds with the guidance of the ascending manifolds and ensure that they meet transversally. First, we determine whether a point on the ascending 1-manifolds A_1 is on the boundary of the descending 3-manifolds D_3 by performing region growing in the descending order on A_1. The boundary points classified by region growing are 2-saddles, i.e. the intersections of descending manifolds and ascending 1-manifolds. Next, we again perform region growing in the descending order on A_2 on the basis of the growing result in the first step. The new boundary points are regarded as 1-saddles, i.e. the intersection of descending manifolds and ascending 2-manifolds. In the end, region growing in the descending order in the interior of A_3 are invoked to obtain the overlay of the descending and ascending manifolds. Points classified in the same ascending and descending manifolds form the 3-cell of Morse-Smale complex.

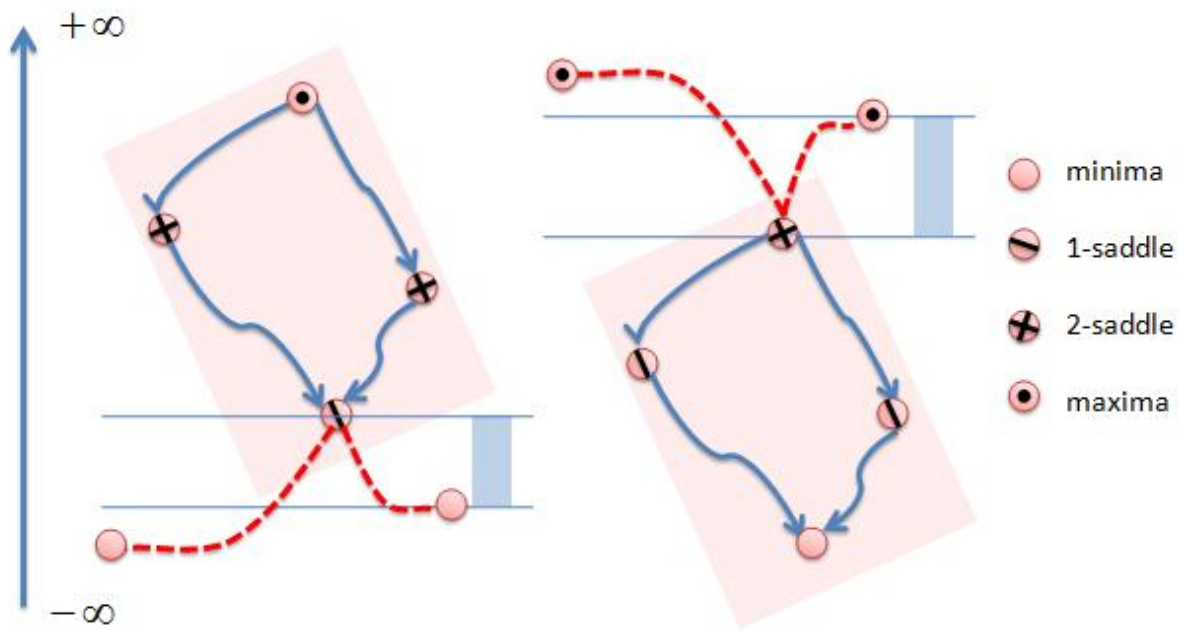

Figure 11. Topological persistence defined over neighboring 3-cells. The quadrangles are the two typical intersections of two neighboring 3-cells and the blue shaded distance denote the persistence between two neighboring 3-cells

4.2. Persistence-based Simplification

A hierarchy of segmentations can be created through successive merging of neighboring pairs of Morse-Smale cells based on the topological persistence. This is equivalent to canceling a pair of handles through the modification of the underlying gradient flow as stated in theorem 3.19.Intuitively the notion of persistence quantifies the stability of pair of topological features, meaning how much f is allowed to be perturbed so that the features will disappear. In a sense the persistence based simplification captures the amount of "topological noise" being removed. Let $\gamma_1, \gamma_2 \in \mathcal{MS}(f)$ be two 3-cells. We say γ_1 and γ_2 are neighboring cells if $\partial\gamma_1 \bigcap \partial\gamma_2$ is a 2-cell. In the three dimensional case, the 2-cell can be a quadrangle enclosed by a maxima, two 2-saddles and one 1-saddle or a quadrangle enclosed by one 2-saddle, two 1-saddles and a minima. In the first case, let s be the common 1-saddle point and m_1, m_2 be the minima of each 3-cell, the persistence is defined to be: $p(\gamma_1, \gamma_2) = \min\{|f(s) - f(m_1)|, |f(s) - f(m_2)|\}$. Similarly, let S be the common 2-saddle in the second case and M_1, M_2 be the maxima of each 3-cell, the persistence between γ_1 and γ_2 is: $p(\gamma_1, \gamma_2) = \min\{|f(S) - f(M_1)|, |f(S) - f(M_2)|\}$. One then will be able to construct the dual graph of the Morse-Smale complex such that each 3-cell becomes a node and neighboring 3-cells γ_i, γ_j share an edge whose weight is $p(\gamma_i, \gamma_j)$. The simplification process can be imagined as an edge contraction process such that cells with low persistency get merged together in the early stage, leaving important topological features unchanged.

4.3. Analysis

Both Morse-Smale complex construction and the simplification can be performed in $O(n \log n)$ time. The sorting step takes $O(n \log n)$ time and the sweep takes linear time assuming the number of neighbors per vertex is a constant. The simplification process can be implemented efficiently using the Union-Find

data structure. Therefore, the overall time complexity is $O(n \log n)$.

5. Results

In this section we report the results of video segmentation by successively merging the Morse-Smale cells using persistence-based simplification described in section 4.2. But first of all, let's briefly describe the ways to smooth the data and compute the density function.

5.1. Data smoothing and Density computation

We use the 3D version of the bilateral filter [27] to perform edge-preserving smoothing for the video data, that is:

$$\hat{I}_p = \frac{1}{W_p} \sum_{q \in \mathcal{N}(p)} G_{\sigma_s}(\|p-q\|) G_{\sigma_r}(I_p - I_q) I_q \tag{14}$$

where $p \in \mathbb{R}^3$, I_p is the video pixel at p, $\mathcal{N}(p)$ is the neighborhood of p and $G_{\sigma_s}, G_{\sigma_r}$ are Gauss function operated on spatial-temporal and range domain. Also, $W_p = \sum_{q \in \mathcal{N}(p)} G_{\sigma_s}(\|p-q\|) G_{\sigma_r}(I_p - I_q)$ is the weight for normalization. We compute the density as follows:

$$f(p) = \sum_{q \in \mathcal{N}(p)} G_{\sigma_r}(\hat{I}_p - \hat{I}_q) \tag{15}$$

Intuitively the density function constructed this way measures the similarity between each pixel and its neighbors based on color coherency. Therefore, the density value will appear high inside some uniform regions and drop down when approaching the boundary. However, the density function generated this way has lots of critical points. To reduce the number of critical points, we further smooth the density function as follows:

$$\widetilde{f}(p) = \sum_{q \in \mathcal{N}(p)} G_{\sigma_{x,y}}(\|p-q\|) G_{\sigma_t}(\|p-q\|) f(q) \tag{16}$$

Note that we distinguish the $\sigma_{x,y}$ and σ_t in order to emphasize the difference between the spatial and temporal domain. In the experiment, we always set σ_t twice as $\sigma_{x,y}$ in order to facilitate the formation of Morse-Smale cells which are elongated along the time axis. Figure 12 illustrates the process of computing the density function. We do not focus on the choice of density functions in this paper. Therefore, one has the full flexibility to chose a better density function in order to achieve better results or tailer to some specific applications. We refer this direction as our future work.

5.2. Video segmentation and applications

We build explicitly the Morse-Smale complexes for video sequences and apply persistence-based simplification. As an application, we see how the merging of Morse-Smale cells can form meaningful objects and hence can be extracted for later analysis. The initial complex can be treated as an over-segmentation and through successive simplification some meaningful objects emerge. In Figure 2 we show some shapes of the Morse-Smale cells generated for the video sequence. In Figure 13, we demonstrate the hierarchical segmentation of the same video sequence. In particular, by the simplified result one can trace the trajectories of the bud and leaves through time in the very shaky and noisy data.

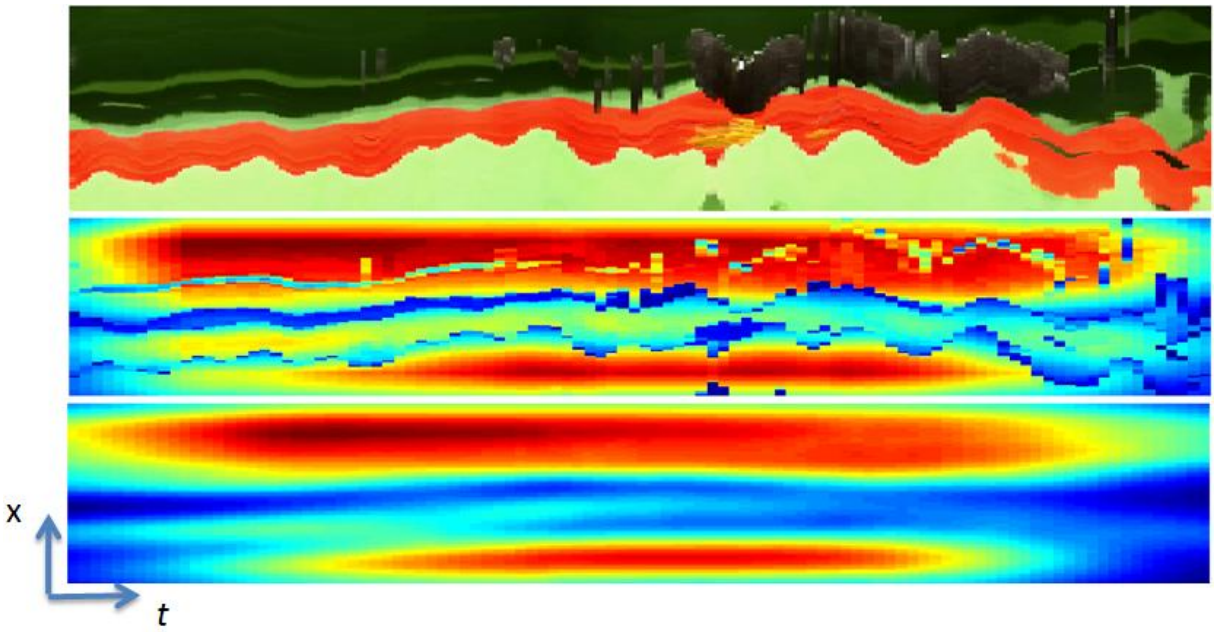

Figure 12. The top row shows the spatial-temporal graph for the 10^{th} row in the video. The next row shows the density function computed by 15. The third row is the smoothed version of the density function computed by 16.

However, the butterfly is grouped into the leaves too probably because part of its wings merge into the background. In Figure 14, where the video data comes from [24], the simplification is performed until only 11 cells are left. One of the cells exactly matches the trajectories of the walking person. Figure 15 shows the shape of the 11 cells through the top view. A trace of the foot prints can be clearly seen in the last cell.

6. Conclusions and Future Work

In this paper, we interpret video segmentation hierarchy as the computation of a Morse-Smale complex with persistence-based simplification. We review the Morse theory in d-dimensions and design efficient algorithms to compute the Morse-Smale complex for functions which are piecewise linear in the spatial temporal domain. Also, we demonstrate how to simplify the structure formed by the Morse-Smale cells in order to produce meaningful results. The experimental results in this paper are promising in terms of video segmentation. In the future work, we plan to investigate how the Morse-Smale complex can serve as a fundamental representation supporting video retrieval applications and recognition. Moreover, the existing approach to construct the Morse-Smale complex is problematic because a pixel (or voxel) can be classified to belong some boundary while in mathematics the boundaries of all the Morse-Smale cells have measure zero. Therefore, a better approach seems to build Morse-Smale complex implicitly in the interpolated domain rather than the simplicial complex domain. Many methods such as splines either quadratic or cubic, Gaussian mixtures, Beizer or NURBS surface representation can serve as potential interpolating functions. We leave this investigation as our future work as well.

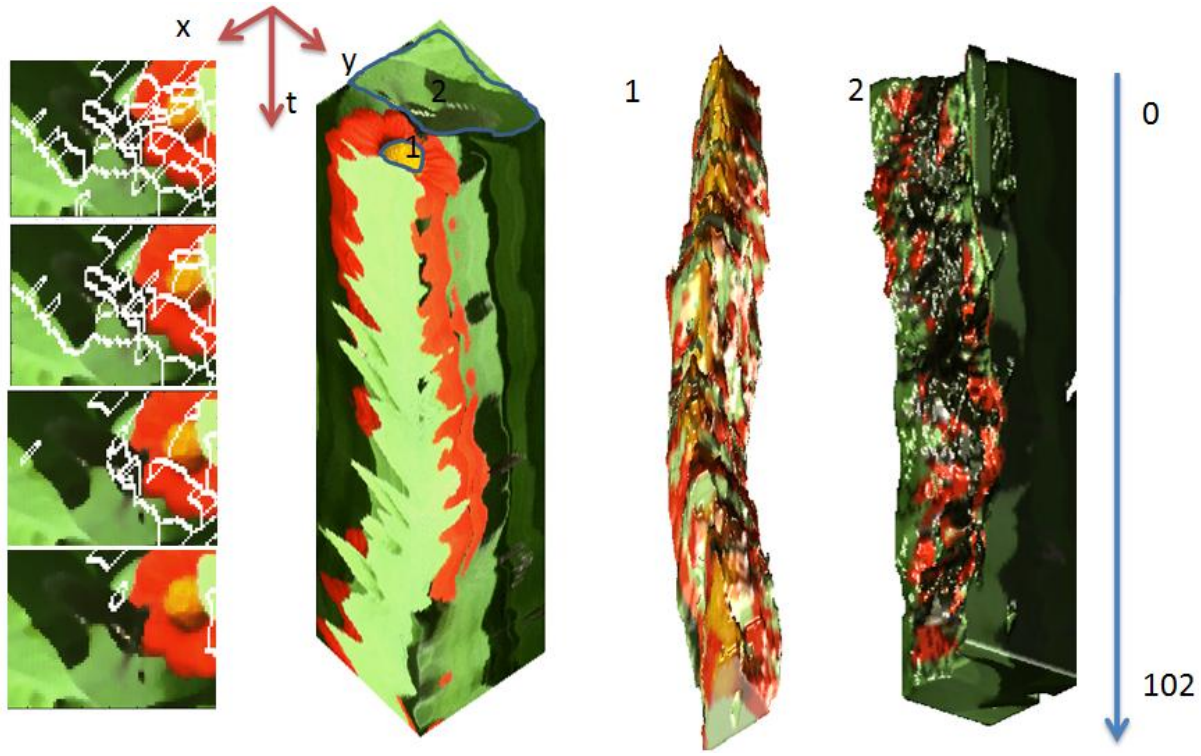

Figure 13. The left column shows the first frame after successive simplifications. The right two columns show two trajectories formed during the cell merging process.

References

[1] N. Apostoloff and A. W. Fitzgibbon. Learning spatiotemporal t-junctions for occlusion detection. In *CVPR*, pages 553–559, 2005. 4

[2] P.-T. Bremer, H. Edelsbrunner, B. Hamann, and V. Pascucci. A multi-resolution data structure for 2-dimensional morse functions. In *IEEE Visualization*, pages 139–146, 2003. 5

[3] D. Comaniciu and P. Meer. Mean shift analysis and applications. In *ICCV*, pages 1197–1203, 1999. 4

[4] E. Danovaro, L. De Floriani, and M. Vitali. Multi-resolution morse-smale complexes for terrain modeling. *Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on*, pages 337–342, Sept. 2007. 5

[5] D. Dementhon. Spatio-temporal segmentation of video by hierarchical mean shift analysis. In *Porc. of Statistical Methods in Video Processing Workshop*, 2002. 4

[6] D. DeMenthon and D. S. Doermann. Video retrieval using spatio-temporal descriptors. In *ACM Multimedia*, pages 508–517, 2003. 4

[7] H. Edelsbrunner. *Geometry and Topology for Mesh Generation*. Cambridge University Press, 2001. 12, 14

[8] H. Edelsbrunner, J. Harer, V. Natarajan, and V. Pascucci. Morse-smale complexes for piecewise linear 3-manifolds. In *Symposium on Computational Geometry*, pages 361–370, 2003. 4, 5

[9] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical morse complexes for piecewise linear 2-manifolds. In *Symposium on Computational Geometry*, pages 70–79, 2001. 4, 13

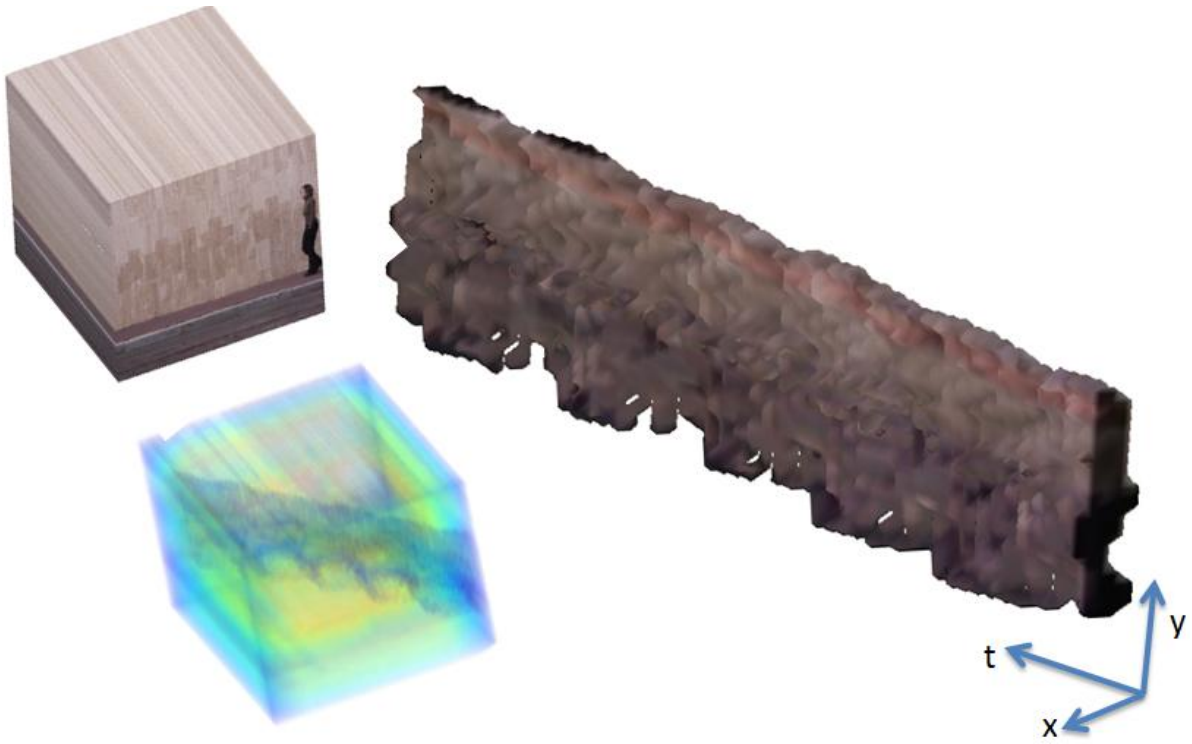

Figure 14. The left column shows the video cube and its density function. The right column shows the trajectory of the walking person extracted from the video sequences

[10] H. Edelsbrunner, D. Morozov, and V. Pascucci. Persistence-sensitive simplification functions on 2-manifolds. In *Symposium on Computational Geometry*, pages 127–134, 2006. 4

[11] C. Fowlkes, S. Belongie, and J. Malik. Efficient spatiotemporal grouping using the nyström method. In *CVPR*, pages 231–238, 2001. 4

[12] H. Greenspan, J. Goldberger, and A. Mayer. A probabilistic framework for spatio-temporal video representation and indexing. In *ECCV*, pages 461–475. Springer-Verlag, 2002. 4

[13] A. Gyulassy, V. Natarajan, V. Pascucci, P.-T. Bremer, and B. Hamann. A topological approach to simplification of three-dimensional scalar functions. *IEEE Trans. Vis. Comput. Graph.*, 12(4):474–484, 2006. 5

[14] A. Gyulassy, V. Natarajan, V. Pascucci, and B. Hamann. Efficient computation of morse-smale complexes for three-dimensional scalar functions. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1440–1447, 2007. 5

[15] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV*, pages 166–173, 2005. 4

[16] I. Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005. 4

[17] D. Letscher and J. Fritts. Image segmentation using topological persistence. In *CAIP*, pages 587–595, 2007. 5

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 4

[19] Y. Matsumoto. *An Introduction to Morse Theory*. American Mathematical Society, 2002. 4

[20] J. Milnor. *Morse Theory*. Princeton Univ.Press, 1963. 4, 6

[21] F. Moscheni, S. Bhattacharjee, and M. Kunt. Spatiotemporal segmentation based on region merging. *PAMI*, 20(9):897–915, 1998. 4

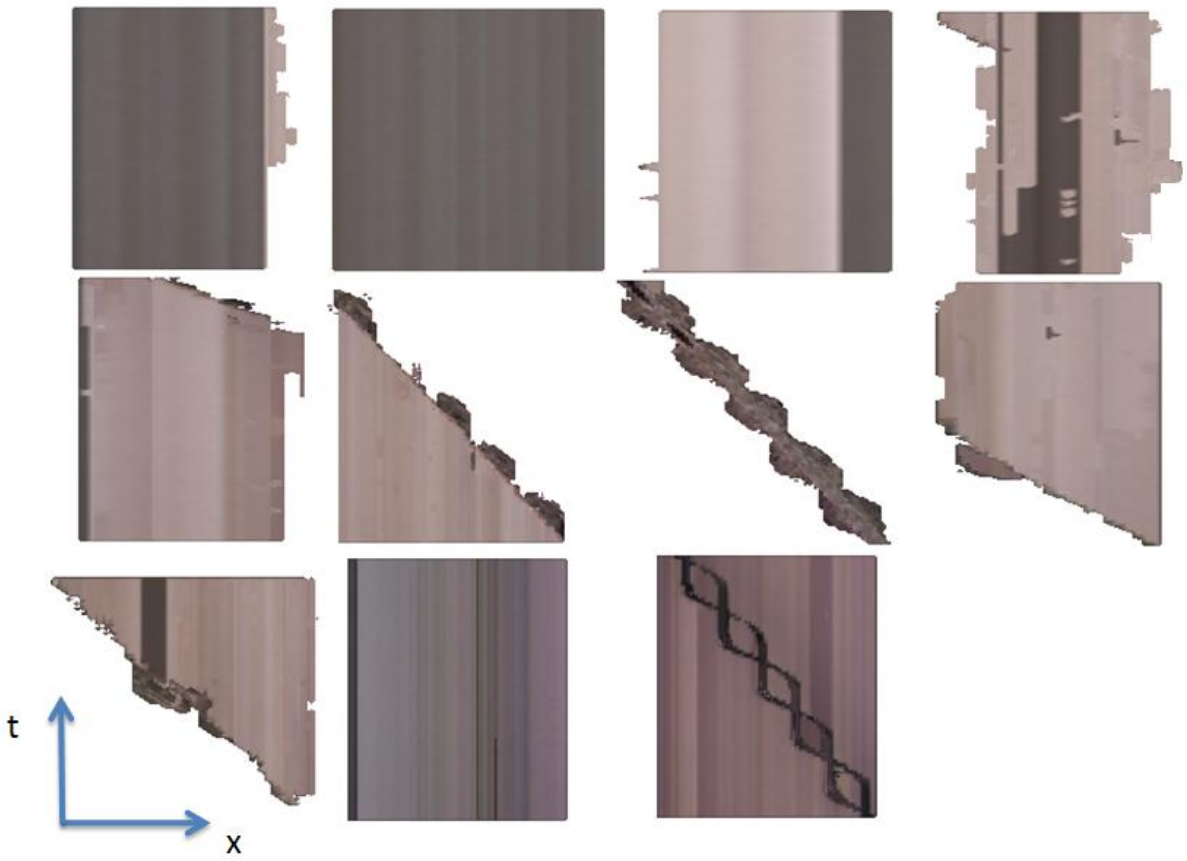

Figure 15. In the simplification, 11 cells are formed and kept. Each cell is shown from the top view toward the X-T domain

[22] S. Paris and F. Durand. A topological approach to hierarchical segmentation using mean shift. In *CVPR*, 2007. 5

[23] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ACM Multimedia*, pages 357–360, 2007. 4

[24] E. Shechtman and M. Irani. Space-time behavior based correlation-or-how to tell if two underlying motion fields are similar without computing them? In *PAMI*, volume 29, pages 2045–2056, November 2007. 4, 18

[25] J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *ICCV*, pages 1154–1160, 1998. 4

[26] E. Sifakis, I. Grinias, and G. Tziritas. Video segmentation using fast marching and region growing algorithms. *EURASIP Journal on Applied Signal Processing*, 4:2002, 2001. 4

[27] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, pages 839–846, 1998. 17

[28] L. Vincent and P. Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(6):583–598, 1991. 5

[29] A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *CVPR*, pages 984–989, 2005. 4