Chapter 3

Mining Temporally-Varying Phenomena in Scientific Datasets

R. Machiraju[*], *S. Parthasarathy*[*], *J. Wilkins*[×],
D. Thompson[†], *B. Gatlin*[†], *D. Richie*[‡], *T. Choy*[×], *M. Jiang*[*],
S. Mehta[*], *M. Coatney*[*], *S. Barr*[×], *K. Hazzard*[×]
[*]Department of Computer and Information Sciences, The Ohio State University
[×]Department of Physics, The Ohio State University
[†]Department of Aerospace Engineering, Mississippi State University
[‡]HiPTi Corporation

Abstract:
Simulation is enhancing and, in many instances, replacing experimentation as a means to gain insight into complex physical phenomena. Recent advances in computer hardware and numerical methods have made it possible to simulate physical phenomena at very fine temporal and spatial resolutions. Unfortunately, given the enormous sizes of the datasets involved, analyzing datasets produced by these simulations is extremely challenging. In order to more fully exploit simulation, the analysis of these large datasets must advance beyond current techniques that are based on interactive visualization.

We outline our vision for one such approach and describe progress on a unified framework that promises to provide a novel method to explore large simulation data sets. We illustrate its application to two disparate science drivers – temporally varying solid and fluid systems. In both applications, there are hidden hierarchies of *features* as well as many abstract multidimensional feature *characterizations* (e.g. shapes). Through this framework, we offer a systematic approach to detect, characterize, and track meta-stable features as well as formulate hypotheses about their evolution – an important step in extracting vital information from such complex systems.

Keywords: Feature mining, Spatio-temporal patterns, Shape-based
Mining, Physics-based mining, Categorization, Computational simulations

3.1 Introduction

The physical and engineering sciences increasingly study large, complex ensembles seeking to understand the underlying phenomena. These studies require analysis of the data generated by either experiments or computational simulations. In this chapter, we focus on the latter and provide motivation using applications from two disparate fields – numerical simulations of fluid flow and molecular dynamics. Computational fluid dynamics (CFD) seeks to understand flow patterns to enhance, for instance, drug delivery schemes for pulmonary treatments for asthma. Similarly, molecular dynamics (MD) seeks to understand the evolution of material defects that affect the properties or performance of industrial materials. In these data, patterns of interest arise and evolve over time as a result of the unsteady nature of the phenomenon under consideration.

Scientific discoveries are often best understood visually – from Galle seeing Neptune in 1846 to Binnig and Rohrer seeing atoms on a surface in the twentieth century. Both discoveries were not surprises in the sense that previous analysis had convinced most of their reality. However, each discovery stimulated future work more dramatically than any analysis might have done.

Unfortunately, the size of simulation datasets significantly challenges our abilities to explore and comprehend effectively the generated data. Analysis via interactive visualization sessions is tantamount to searching for the proverbial "needle in a haystack." Currently, a well-trained individual may need several days or even weeks to analyze the data generated by an MD simulation and create a list of viable defect structures. Similarly, in the extremely large datasets generated by simulations of complex fluid flows, locating and tracking relevant features are daunting tasks. In both cases, phenomena occur on multiple length and time scales. Some features persist sufficiently to have gross macroscopic effects. Other short-lived transients are precursor events central to the unsteady (in the temporal domain) behavior of the system. An additional complication is that currently available hardware does not have the prowess yet to provide even near real-time visualizations.

Therefore, we believe it is crucial that some degree of automation be incorporated into the exploration process for large datasets. One such successful approach is described in [Machiraju *et al.*2001] and is based on a representational scheme that facilitates ranked access to macroscopic features in the dataset. However, other than identifying, denoising, and ranking the features, no attempt is made to extract information about the features or track and catalog them.

An alternative approach would be to apply traditional data mining algorithms to these scientific datasets. However, it is our contention that existing data mining techniques, applied in isolation, are simply too general. Embedding domain expertise (i.e., via understanding the science) in the data mining process is critical to its success especially for the datasets characteristic of large-scale simulations. Moreover the application of existing data mining techniques may not be the most efficient of solutions, particularly for analyzing complex simulation data.

Thus, there is a paucity of general approaches that facilitate meaningful analysis of large and complex data describing physical phenomena. Traditional There is a need to explore a larger space of solutions that are based on the underlying physics and are enabled by computer science techniques from visualization, data analysis, and data mining. By incorporating application-specific physics into the mining effort, we can develop characterizations of physically-relevant features. In this chapter, we describe one such approach that we call feature mining.

The remainder of this chapter is organized as follows. In Section 3.2 we describe the two important applications that motivate this work. In Section 3.3 we describe the system and identify the key components. Section 3.4 documents the preliminary results we have for the two application domains, while Section 3.5 describes previous work conducted by other groups. Finally we provide conclusions and future directions in Section 3.6.

3.2 Example Applications

We now provide some background information on the two science drivers we have chosen – respiratory flow in multi-generational bronchial trees and defect evolution in materials. While our two science drivers would seem significantly different, we contend that a common framework can facilitate effective exploration for both problems.

3.2.1 Computational Simulation of Biomedical Fluid Flows

Respiration – specifically, airflow through the network of lung airways – produces surprisingly complex flow fields. Even though the flow is laminar through much of the bronchial tree, secondary currents can be dominant, particularly transverse vortex pairs that form due to axial curvature of the tubes and wall shear. These vortices migrate downstream and interact with new ones generated by repeated branching. These secondary flows are critical to the efficient filtering of inhaled air: aerosols, entrained in their trajectories, impact mucus-lined walls from which they can be expelled from the lungs through the action of cilia or coughing.

The analysis of these secondary flows is complex, both because of their not-yet-understood persistence and their branching into multiple vortices. Much of the computational modeling of flow through small airway bifurcations is that due to Gatlin, Hammersley, *et. al* [Hammersley *et al.*1993, Gatlin *et al.*1995, Gatlin *et al.*1997b, Gatlin *et al.*1997a]. When dealing with datasets generated by simulations of complex temporally varying fluid flows, it is challenging to locate and track relevant features. Existing techniques for vortex detection are typically based on local, flow-field parameters such as the velocity gradient tensor. The generation of new features and destruction of existing features present challenges for feature correspondence algorithms.

Potential Impact: New data mining techniques relevant to computed respiration flow data not only can enhance the understanding of known flow characteristics, but also may discover previously undetected features, just as visualization techniques revealed the long unknown secondary structures in the flow. Of particular interest are the longevity of vortex pairs generated by bifurcation and the mechanisms of interaction between vortices. Additionally, the detection of regions of flow separation are important for understanding the impaction of entrained particles and the interruption of laminar flow. Improved understanding of these flows has two important applications: (1) the health hazard posed by the inhalation of carcinogenic, disease-bearing, or lung-damaging aerosols and (2) the clinical delivery of both local and systemic aerosolized drugs through the lungs. While the depth of penetration into the tubular network of the lungs depends on the nature and concentration of particles, the aerodynamics of respiration plays a critical role.

3.2.2 Molecular Dynamics Simulations of Defect Evolution

The key complexity of real materials used in commercial applications is not that they are defected in the trivial sense of being imperfect or impure, but rather that their material properties depend critically on their nonideality. As an example, the enhanced diffusion of dopants in the presence of extended {311} defects in silicon is a limiting factor in the fabrication of shallow junction devices [Cowern *et al.*1999]. The growth of such extended defects involves the diffusion, capture and dissociation of silicon point defects [Arai, Takeda, & Kohyama1997,Kim *et al.*1999,Kim *et al.*2000]. This example can be repeated with variations in every material essential to current high technology.

Molecular dynamics simulations can track the nucleation and growth of defects but realistic time scales exceed computing technology. Emerging acceleration techniques [Montalenti, Sørensen, & Voter2001,Sørensen & Voter2000,Voter1997,Voter1998] can achieve realistic simulation times. Wavelet techniques [Richie, Kim, & Wilkins2001] can dramatically reduce the molecular dynamics data and detect persistent defect structures. The challenge is to identify and classify these structures and track their evolution and interactions.

Potential Impact: New data mining techniques can not only uncover fundamental defect nucleation and growth processes but also provide essential parameters for modeling macroscopic properties of materials. This need is well recognized in the the semiconductor industry in its “silicon roadmap” that identifies the short- and long-range problems necessary to continually pack more transistors on a chip. In structural materials used, for example, in turbine engines, there is a growing need to connect the microscopic and macroscopic scales. Indeed the phrase “multiscale methods” recognizes the wide spread importance of connecting complex microscopic processes to the design and optimization of materials properties.

3.3 Generalized Framework

Essentially, there is a need to deduce the presence of features and derive their shape characteristics from a large data repository that describes some time-varying, evolutionary phenomenon. In this context, shape refers to the salient characteristics of a feature including kinematical and dynamical characteristics along with a geometrical description. This abstract notion of shape allows us to apply more general data mining algorithms to the extracted features and their characteristics. It is our claim that a "shape-based" data-mining paradigm will prove fruitful in the analysis of complex unsteady phenomena. Kamath also makes similar remarks about the utility of feature-based approaches [Kamath2001].

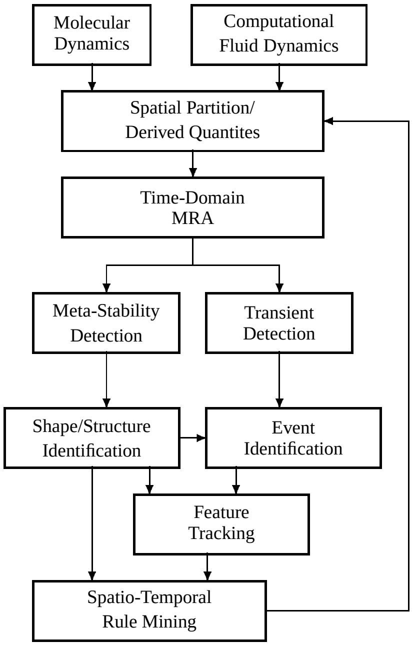
Molecular Dynamics
Computational Fluid Dynamics
Spatial Partition/ Derived Quantites
Time-Domain MRA
Meta-Stability Detection
Transient Detection
Shape/Structure Identification
Event Identification
Feature Tracking
Spatio-Temporal Rule Mining

Figure 3.1: Generalized Feature Mining Framework

Figure 3.1 illustrates our generalized framework applied to processing of physically-based simulation data. We contend that a common framework can compactly store and analyze data of evolutionary phenomena. We assume that certain locally computable quantities can detect precursor events. Our approach is novel in its flexibility and applicability across disciplines. The shape-based analysis converts the task of data management and analysis into one of choosing robust shape descriptors and being able to index features from a catalog. The descriptors will be derived from the application.

In addition to feature detection algorithms, aggregation or segmentation, tracking and characterization algorithms must be utilized in conjunction with traditional data mining algorithms to facilitate cataloging detected structures and expediting searches to gain scientific insights. Our framework synergistically brings to bear these techniques to address the problems associated with analyzing large datasets generated by simulations of physical phenomena. We now describe the key elements of our framework.

Spatial Partitioning to Exploit Locality

Fine spatial resolutions are often used to resolve features in computational simulations. Tracking features over the entire spatial domain is not viable and meaningful. Hence, through a process of partitioning, smaller sub-domains are considered for shape and feature evolution. This process is tantamount to dividing the bale of hay into smaller bales to search for the proverbial needle. Thus regular and irregular sized sub-domains derived from either just the domain or function values can be considered.

Multiscale Event and Feature Detection

A feature at a given temporal scale can be stable, meta-stable, or transient. The birth, evolution and death of a feature is often triggered by precursor events. It is therefore crucial to identify when such events occur. We have chosen a "trigger-based" multi-resolution analysis (MRA) using wavelets for event detection. Thus, a single derived quantity or a trigger is monitored for events at multiple time scales. For instance this quantity is *swirl* in a CFD simulation. In an MD simulation this quantity is *potential* or *energy* or *dislocations* of atoms. Multiscale trigger monitoring is needed given the range of feature lifetimes. Wavelet techniques are effective here and are already successfully working in molecular dynamics simulations [Richie, Kim, & Wilkins2001, Richie *et al.*2002].

Feature Mining

A systematic approach to feature mining [Thompson *et al.*2002], i.e., the process of detecting, characterizing and tracking relevant features, is being developed. Our intent is to exploit the physics of the problem at hand to develop highly discriminating, application-dependent feature detection algorithms and then use data mining algorithms to classify, cluster, and categorize the identified features. Our work parallels that of Yip and Zhao [Yip & Zhao1996] in some ways. It should be noted that our work relies on more physics-based understanding of features and exploits the underlying physics to a greater extent.

The most basic aspect of feature mining is **feature detection**. The output of any detection algorithm is a collection of many regions-of-interest (ROIs). The underlying physics is exploited to locate features using local operators or sensors to detect and non-local or global operators to verify. Verification is need in some cases to confirm that a given ROI indeed represents a feature. We consider defects at quenched states and finite temperatures for MD simulations and shocks and vortices for CFD.

A second component of feature mining is **shape-based feature characterization and categorization** in which the "shape" of a feature is described by characteristics, such as shape and structure, and kinematical and dynamical properties in an abstract *multidimensional shape space*. The descriptors for a MD simulation can include the number of atoms involved, their orientations, the connectivity between atoms, the trajectory, and history of its evolution. In a CFD simulation, vortices, the type of feature of importance for respiratory flows, can be characterized by their strength and sense of rotation as well as obvious geometrical parameters such as position, shape, and extent. These features can be categorized by notions of similarity. Shape categories enable synergistic understanding of events and features in the MD and CFD domains. To compute the similarity between shapes or structures we rely on spatial geometric hashing [Wolfson & Rigotsos1997] and clustering algorithms [Jain & Dubes1988]. To categorize the structures we rely on classification algorithms [Quinlan1996] using the generalized shape descriptors as input to the classifiers. For CFD data we employ a generalized shape descriptor for swirling regions and propose hierarchical shape matching algorithms.

A third component of feature mining is **corresponding and tracking of features over time**. The generation of new features and destruction of existing features pose major challenges to effective, feature-tracking algorithms. The essential problem is to determine how the position of a particular feature changes during a given time interval. In our datasets, this is non-trivial since fissures and fusions of features are extremely common. Furthermore, the structural descriptors of the same feature may change over time. Tracking and correspondence complete the construction of the multi-dimensional shape space for a given application. Relevant related work in feature tracking was reported in [Samtaney *et al.*1994, Silver & Wang1997]. Shapes were not considered therein and the method is, in general, expensive. Similarly in [Reinders, Post, & Spoelder1999, Reinders, Jacobson, & Post2000] the skeleton or an approximate medial axis was computed for vortices. However, this representation is very *crisp* and does not allow tangible matching and tracking. In [Thampy2003], a predictive algorithm was developed that utilized the evolution of selected kinematical and dynamical properties to enhance confidence in the correspondence algorithm.

Mining for Spatial and Spatio-Temporal Patterns

Over any time interval in a simulation, we need techniques that can identify important spatial patterns efficiently. Some patterns can be complex and not necessarily sequential. The aim is to derive predictive rules: combinations of features resulting in certain events, (e.g., fusion or fissure). To derive such rules requires identifying frequently occurring spatial patterns. Clustering, association [Agrawal *et al.*1996], and sequential pattern analysis [Parthasarathy *et al.*1999], and spatio-temporal analysis [Vlachos, Kollios, & Gunopulos2002] will be used to determine the important patterns. Our eventual goal is to correlate information from a shape categorization together with transition detection mechanisms to help discover novel axioms relating to the evolution of shapes over time. An example of such an axiom could be “a type-A feature evolves into a type-B feature through some particular mechanism.” Such rules can be found using event-based sequential and association pattern analysis. Equally important is to identify those axioms that dominate the particular simulation type. These data mining algorithms will operate on the shape space constructed in an earlier step and produce explanations of feature behavior and evolution.

Generality

The large-data exploration methodology we describe is appropriate for any data that can be transformed to a multiscale representation and consists of features that can be extracted through local operators and aggregated in spatial, scale, and temporal dimensions. Thus, one can consider domains in addition to CFD and MD.

3.4 Current Efforts

In the previous sections we described our vision for the generalized framework and described our motivating applications. In this section we describe current and ongoing work toward realizing our vision. The methods described below will partially construct the shape space for a given application. Work on tracking and correspondence is ongoing and is not described. We also do not describe our approach to conducting data mining tasks in the shape space as it is a work in progress.

3.4.1 Feature Mining

In this section, we focus on one component of feature mining and describe two distinct feature detection paradigms [Thompson *et al.*2002]. The common thread is that both are bottom-up feature constructions with underlying physically-based criteria. The two perform essentially the same steps, but in different orders. As will become evident, it is unlikely that non-physics-based techniques would provide the fidelity needed to locate complex flow field (CFD) or defect (MD) structures.

In general, a feature is a pattern occurring in a dataset that is of interest and that manifests correlation relationships among various components of the data. For instance, a shock in a supersonic fluid flow would be considered a significant feature: when such a shock occurs, the pressure increases abruptly in the direction of the flow, and the fluid velocity decreases in a prescribed manner. A significant feature also has spatial and temporal scale coherence.

For many applications, generic data mining techniques such as clustering, association, and sequencing can reveal statistical correlations between various components of the data. Returning to the shock example, we could use statistical mining to ferret out associations, but it might be difficult to attach precise spatial associations for the rules discovered. A fluid dynamicist, however, would like to locate features with a rather high degree of certainty. Such qualitative assertions alone will not suffice. This is where our approach to feature mining comes in: we take advantage of the fact that, for simulations of physical phenomena, the field variables satisfy certain physical laws. We can exploit these kinematic and dynamic considerations to locate features of interest.

The fidelity improvements garnered by tailoring these highly discriminating feature detection algorithms to the particular application far outweigh any loss of generality. The state of the art in feature detection and mining in simulation data is similar to what existed for image processing when edge detection methods were the main techniques. Much more is now understood, and mining for image data is often done in terms of the features, namely edges. This suggests that a blend of data- and feature-mining methods might have the potential to reduce the burdensome chore of finding features in large datasets.

Point classification techniques

The first feature detection paradigm, which we call point classification, requires several operations in sequence:

- Detection by application of a local operator at each point in the domain
- Binary classification of each point based on some criteria
- Aggregation of contiguous regions of similarly classified points
- Denoising to eliminate aggregates that are of insufficient extent, strength, etc.
- Ranking based on feature saliency
- Tracking identified features

This approach identifies individual points as belonging to a feature and then aggregates them to identify regions that are features. The points are obtained from a tour of the discrete domain and can be in many cases the grid points of a physical grid (CFD) or a lattice (MD). The operator used in the detection step and the criteria used in the classification step embody physically based point-wise characteristics of the feature of interest. In this context, classification accords membership of a discrete point in the dataset to a feature.

Aggregate classification techniques

We can best incorporate the global information needed to define a vortex into our second feature detection paradigm, the aggregate classification approach. Aggregate classification follows a somewhat different sequence of operations:

- Detection by application of a local operator at each point in the domain
- Aggregation of contiguous regions of probable candidate points
- Binary classification (or verification) of each aggregate based on some criteria
- Denoising to eliminate aggregates that are of insufficient extent, strength, etc.
- Ranking based on feature saliency
- Tracking identified features

This approach identifies individual points as being probable candidate points in a feature and then aggregates them. The classification algorithm is applied to the aggregate using physically based regional criteria to determine whether the candidate points constitute a feature. Thus, the operator deployed towards point classification can be efficient but less accurate. False positives generated at the earlier stages can be eliminated later in the verification stage. Classification in this context confirms that an aggregated subset of a domain indeed forms a relevant feature.

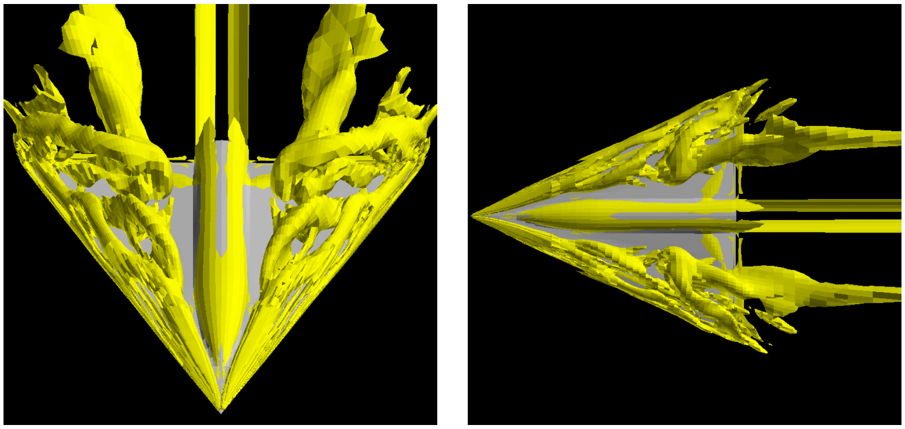

Figure 3.2: The results of our point classification algorithm applied to a delta wing dataset. The front and top views respectively are shown. The yellow regions indicate regions of swirling flow. There exist several regions which are falsely classified.

3.4.2 Fluid Dynamics

We now present two examples of feature detection algorithms as applied to CFD datasets. Although algorithms have been developed for other features, we focus on those for vortices because of the critical role they play in the bronchial airflow. Additionally, the vortex provides a direct way to contrast the two different feature detection paradigms.

CFD Example 1 : Vortex Detection using Point Classification The first technique we consider uses the eigenvalues of the local velocity gradient tensor. In regions of swirling flow, the eigenvalues of the velocity gradient tensor are complex. Berdahl and Thompson [Berdahl & Thompson1993] defined a swirl parameter that estimates the tendency for the fluid to swirl about a given point. The swirl has a nonzero value in regions containing vortices and attains a local maximum in the core region. In this point classification algorithm, the detection step consists of computing the eigenvalues of the velocity gradient tensor at each field point. The classification step consists of checking for complex eigenvalues and assigning a swirl value if they exist. The aggregation step then agglomerates contiguous grid points where the swirl parameter exceeds a threshold value into vortical regions. This method's primary shortcoming is that it–and all eigenvalue-based vortex detection techniques–can generate false positives. An example of this method is shown in Figure 3.2. Its local nature makes it unable to discriminate between locally curved streamlines and closed streamlines characteristic of a vortex. Other features, such as shocks, are more amenable to the point classification framework.

CFD Example 2 : Vortex Detection using Aggregate Classification We recently developed an aggregate classification-type vortex detection technique. We based its detection step on an idea derived from a lemma in combinatorial topology. Specifically, velocity vectors around core regions exhibit certain flow patterns unique to vortices,

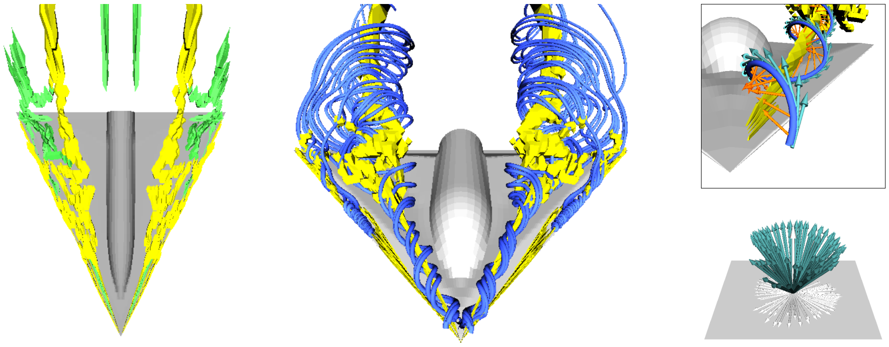

Figure 3.3: The results of our aggregate classification technique applied to the delta wing dataset. (left) All candidate core regions are shown. The verified cores are shown in yellow while the spurious ones are shown in green. (middle) Streamline tracing around verified cores. (right) The top image shows the verification algorithm at work through seeding and tracing, while the bottom image shows illustrates the use of projections and angles to verify vortices.

and it is precisely these flow patterns that we search for in the computational grid. Not surprisingly, our approach is related to critical point theory. However, critical points alone are not sufficient to detect a vortex. For each grid point, our algorithm examines its immediate neighbors to see whether the neighboring velocity vectors point in three or more direction ranges. The novelty of this method is its relative insensitivity to core direction. Therefore, very approximate core directions may be used in the detection step.

Our technique segments candidate core regions by aggregating points identified from the detection phase. We then classify (or verify) these candidate core regions based on the existence of swirling streamlines surrounding them. (For features that lack a formal definition, such as the vortex, we must choose the verification criteria so that it concurs with the intuitive understanding of the feature. In this case, verifying whether a candidate core region is a vortex core region requires checking for any swirling streamlines surrounding it.) Checking for swirling flow in three dimensions is a nontrivial problem since vortices can bend and twist. The technique we developed essentially checks to see if the local tangent to the streamline, when projected onto the plane normal to the local core tangent, spans 2π. The aggregate nature of this classification step is apparent. Checking for swirling streamlines is a global (or aggregate) approach to feature classification (or verification) because swirling is measured with respect to the core region, not just individual points within the core region. Figure 3.3 describes all steps of this paradigm.

3.4.3 Molecular Dynamics Simulations of Defect Evolution

The challenge of detecting features during an ongoing MD simulation was met with the application of multiresolution analysis (MRA) techniques. Wavelet analysis is exploited in the time domain to analyze dynamics. For each atom, its sequence of positions are projected on a wavelet basis, with the expansion coefficients generated incrementally using components supplied by the STORMRT Scientific Wavelet Package [Richie, Kim, & Wilkins2001]. These components treat streaming data more efficiently than more conventional "fast" wavelet transform (fwt) techniques.

The same feature mining procedures that worked well for CFD data work for MD too. In any persistent structure, "defect" atoms must be distinguished from "bulk" atoms. While this task might seem more challenging at finite temperature due to the thermal noise, a *single* rule works for all structures: thermal and quenched. For a bulk atom, precisely four atoms have bonds (with the bulk atom) less than 2.6 Å and the angles between any two bonds lie within 90-130 degrees. Any other atom is a defect. Similar definitions can be formulated for other systems. Atoms near the surface of a periodically repeated cell don't "see" the other atoms though. This problem is solved by padding the cell with a layer of periodic material.

Here, we illustrate how point classification procedures can be employed toward the defects in the quenched (cooled state) and finite thermal temperatures respectively. Each atom site is visited and the atom is tested for membership in a defect ensemble. The classification operator for this application is as follows. We define two conditions C_1 (bond angle as above) and C_2 (number of bonds as above) to accurately classify bulk atoms. The *conjunction* of the above two conditions as well as the *disjunction* are evaluated for all atom sites. The atom sites which satisfy the conjunction are the ones which definitely belong to the bulk. Those that satisfy the disjunction will with some likelihood belong to the bulk. The remaining atom sites are definitely part of the defect. Such atoms are referred to as *defect atoms*. The defect atoms are then spatially clustered to aggregate these into possible defect structures. We empirically verified that this method works well even on noisy data. Figure 3.4a shows a persistent structure at 1000 K. The black atoms are those identified as defect atoms. Figure 3.4b shows the same structure quenched with a first-principles approach; the quenching removes thermal noise at a heavy computational cost. The same atoms are marked in both structures which demonstrates this method works on non-quenched structures.

In a large scale simulation the challenge is to isolate separate defect clusters. A line is drawn connecting all defect atoms that lie within 4 Å of each other. Thus, a cluster is comprised of connected defect atoms, a computationally fast process. Figure 3.5 shows two defects embedded in a 512 atom lattice.

The deployment of aggregate classification techniques for MD data is far from clear. This is still an active area of research. The large number of resulting identified structures (from the detection phase) must be sorted into a smaller set of distinct types. Quenching solves this problem but is computationally expensive. In addition, some structures are stable only at high temperatures. Through quenching, these structures are lost. Identifying time averaged structures is a great challenge. Occasionally with noisy data too many or too few atoms are marked. Figure 3.6 demonstrates this problem. These two structures have different numbers of defect atoms marked, yet when

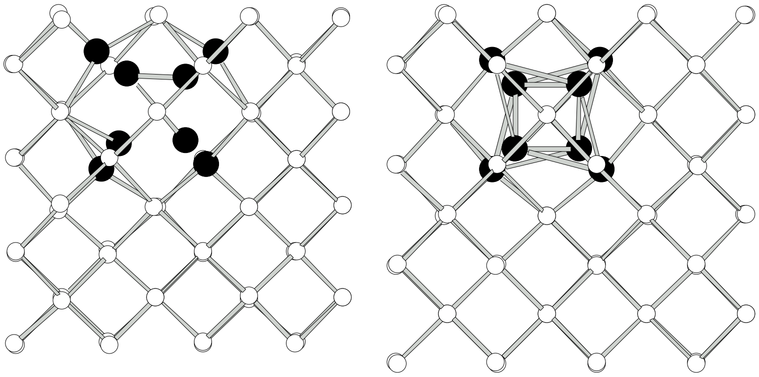

Figure 3.4: Black atoms are defect atoms. Top is a structure identified at 1000K. Bottom is the same structure quenched using first principles. Even though the atoms in the top structure are displaced due to thermal noise, the same atoms are marked as defect in both structures.

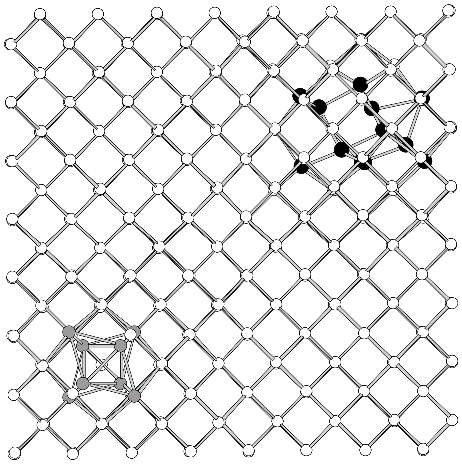

Figure 3.5: Two separated defects: black atoms are one cluster, grey atoms are different cluster.

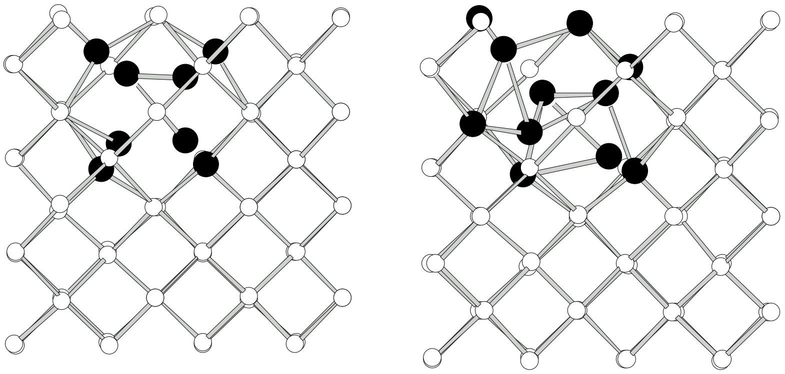

Figure 3.6: These two structures have a different number of defect atoms marked. When quenched however they are the same structure.

quenched are the same. Additionally, we are still exploring robust and viable shape descriptors and matching algorithms for MD data.

3.5 Related Work

The framework described here is related to the work being conducted by Marusic and his associates [Marusic *et al.*2001]. Event time-series tracking is employed to detect turbulent bursts which are then analyzed and tracked. However, they do not consider the detection and cataloging of features at multiple temporal scales.

Knowledge discovery and data mining (KDD) refers to the overall process of discovering new patterns or building models from a given dataset. Fundamental KDD research in the last decade has primarily focused on: i) new techniques to preprocess, mine and evaluate the data, ii) efficient algorithms that implement these techniques, and iii) applications of above techniques on business applications.

More recently researchers have started tackling the problem of mining scientific datasets. In particular approaches for mining astronomical [Burl *et al.*1998], physical (fluid flow) [Han, Karypis, & Kumar2001], biological [Wang *et al.*1997, Li & Parthasarathy2001, Parthasarathy & Coatney2002] and chemical [Dehaspe, Toivonen, & King1998] datasets have been recently proposed by various researchers. Few of the above methods actually account for the structural or spatial properties of the data. A straight-forward application of well-known data-mining techniques does not always yield the most efficient algorithms.

Computational fluid datasets have received more scrutiny from visualization and data-mining researchers than other computational domains. Significant progress has been made in the area of identifying regions of swirling flow. Algorithms described in [Berdahl & Thompson1993,Banks & Singer1995,Jeong & Hussain1995,Portela1997, Sujudi & Haimes1995,Jiang, Machiraju, & Thompson2002a,Jiang, Machiraju, & Thompson2002b] demonstrate the ability to identify regions of swirling flow in complex three-dimensional flow fields.

Consideration of time-varying data introduces additional complexity through the need for tracking of features. According to [Samtaney *et al.*1994], five distinct evolutionary events can occur to features in scientific simulations: continuation, creation, dissipation, bifurcation, and amalgamation. Each of these processes must be accounted for in the tracking algorithm. The work in [Silver & Wang1997] is applicable for general three-dimensional tracking of features. Other solutions to this problem exploit hierarchical data structures [Carr, Snoeyink, & Axen2000, Shen, Chiang, & Ma1999].

3.6 Summary

The steady increase in computing power available for science and engineering problems challenges our ability to learn new science from the massive data. We have proposed and are developing a generalized framework that facilitates the analysis of large-scale simulation data for time-varying, evolutionary phenomena. The key component of our approach is an abstract shape-based description of the relevant features. This abstract notion of shape allows us to apply more general data mining algorithms to the extracted features and their characteristics.

Our flexible approach is motivated by two disparate applications – respiratory flow and material defect simulation. Both drivers raise **central** issues that the components of the framework will necessarily address:

- Multiscale event detection
- Feature mining
- Shape-based feature characterization and categorization
- Correspondence and tracking of features over time
- Mining for spatial and spatio-temporal patterns

It should be noted that both science drivers have commonalities that are exploited by the techniques listed above.

Preliminary results have been very encouraging. However, more remains to be done to realize the complete unified framework. A systematic approach to feature mining was conceived to locate both local and global features. Currently, tracking features in a time-varying dataset is being investigated. Similarly, we are conceiving a comprehensive framework that will allow one to derive appropriate associations between the occurrence of transitionary events and the change in feature demographics. This framework will also include environmental parameters such as the underlying geometry. Also, of interest is the creation of tools which will control both the feature- and data-mining exercises. It is our belief that our proposed framework is likely to garner new insights from massive simulation datasets and allow for a better understanding of the underlying physical phenomena.

Bibliography

[Agrawal *et al.*1996] Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, A. I. 1996. Fast discovery of association rules. In *et al.*, U. F., ed., *Advances in Knowledge Discovery and Data Mining*. MIT Press.

[Arai, Takeda, & Kohyama1997] Arai, N.; Takeda, S.; and Kohyama, M. 1997. Self-interstitial clustering in crystalline silicon. *Phys. Rev. Lett.* 78:4265.

[Banks & Singer1995] Banks, D. C., and Singer, B. A. 1995. A Predictor-Corrector Technique for Visualizing Unsteady Flow. *IEEE Transactions on Visualization and Computer Graphics* 1(2):151–163.

[Berdahl & Thompson1993] Berdahl, C. H., and Thompson, D. S. 1993. Eduction of Swirling Structure using the Velocity Gradient Tensor. *AIAA J.* 31(1):97–103.

[Burl *et al.*1998] Burl, M.; Asker, L.; Smyth, P.; Fayyad, U.; Perona, P.; Aubele, J.; and Crumpler, L. 1998. Learning to recognize volcanos on venus. In *Machine Learning*, 165–195.

[Carr, Snoeyink, & Axen2000] Carr, H.; Snoeyink, J.; and Axen, U. 2000. Computing contour trees in all dimensions. In *Proc. 11th ACM/SIAM Symp. on Discrete Algorithms*.

[Cowern *et al.*1999] Cowern, N. E. B.; Mannino, G.; Stolk, P. A.; Roozeboom, F.; Huizing, H. G. A.; van Berkum, J. G. M.; Cristiano, F.; Claverie, A.; and Jaraiz, M. 1999. Energetics of self-interstitial clusters in si. *Phys. Rev. Lett.* 82:4460.

[Dehaspe, Toivonen, & King1998] Dehaspe, L.; Toivonen, H.; and King, R. 1998. Finding frequent substructures in chemical compounds. In *International Conference on Knowledge Discoverya and Data Mining*.

[Gatlin *et al.*1995] Gatlin, B.; Cuicchi, C. E.; Hammersley, J. R.; Olsen, D. E.; Reddy, R. N.; and Burnside, G. G. 1995. Computational simulation of steady and oscillating flow in branching tubes. In *The 1995 ASME/JSME Fluids Engineering and Laser Anemometry Conference and Exhibition: Bio-Medical Fluids Engineering*, volume FED-212, 1–8. American Society of Mechanical Engineers. Hilton Head, SC.

[Gatlin *et al.*1997a] Gatlin, B.; Cuicchi, C. E.; Hammersley, J. R.; Olsen, D. E.; Reddy, R. N.; and Burnside, G. G. 1997a. Computation of converging and diverging flow through an asymmetric tubular bifurcation. In *The 1997 ASME Fluids*

Engineering Division Summer Meeting, volume FEDSM97. American Society of Mechanical Engineers. Vancouver,BC.

[Gatlin *et al.*1997b] Gatlin, B.; Cuicchi, C. E.; Hammersley, J. R.; Olsen, D. E.; Reddy, R. N.; and Burnside, G. G. 1997b. Particle paths and deposition patterns for laminar flow through a branching tube. In *The 1997 ASME Fluids Engineering Division Summer Meeting*, volume FEDSM97. American Society of Mechanical Engineers. Vancouver,BC.

[Hammersley *et al.*1993] Hammersley, J. R.; Olson, D. E.; Reddy, R. N.; Arabshahi, A.; and Gatlin, B. 1993. Computational modeling of airflows in the smaller airways. *American Review of Respiratory Diseases* 145:A32.

[Han, Karypis, & Kumar2001] Han, E.; Karypis, G.; and Kumar, V. 2001. Data mining for turbulent flows. In *Data mining for scientific and engineering applications*, 239–256.

[Jain & Dubes1988] Jain, A. K., and Dubes, R. C. 1988. Algorithms for clustering data, prentice-hall, englewood cliffs. *NJ* 88:1988.

[Jeong & Hussain1995] Jeong, J., and Hussain, F. 1995. On the identification of a vortex. *J. Fluid Mech.* 285:69–94.

[Jiang, Machiraju, & Thompson2002a] Jiang, M.; Machiraju, R.; and Thompson, D. 2002a. A Novel Approach to Vortex Core Region Detection. In *Joint Eurographics-IEEE TCVG Symposium on Visualization*, 217–225.

[Jiang, Machiraju, & Thompson2002b] Jiang, M.; Machiraju, R.; and Thompson, D. 2002b. Geometric Verification of Swirling Features in Flow Fields. In *Proc. IEEE Visualization '02*, 307–314.

[Kamath2001] Kamath, C. 2001. On Mining Scientific Datasets. In *et al.*, R. L. G., ed., *Data Mining for Scientific and Engineering Applications*, 1–21. Kluwer Academic Publishers.

[Kim *et al.*1999] Kim, J.; Kirchhoff, F.; Aulbur, W.; Wilkins, J.; and Khan, F. 1999. Thermally activated reorientation of di-interstitial defects in silicon. *Phys. Rev. Lett.* 83:1990.

[Kim *et al.*2000] Kim, J.; Kirchhoff, F.; Wilkins, J.; and Khan, F. 2000. Stability of si-interstitial defects: From point to extended defects. *Phys. Rev. Lett.* 84:503.

[Li & Parthasarathy2001] Li, H., and Parthasarathy, S. 2001. Automatically deriving multi-level protein structures through data mining. In *HiPC Workshop on Bioinformatics and Computational Biology*.

[Machiraju *et al.*2001] Machiraju, R.; Fowler, J.; Thompson, D.; Soni, B.; and Schroeder, W. 2001. EVITA - Efficient Visualization and Interrogation of Terascale Datasets. In *et al.*, R. L. G., ed., *Data Mining for Scientific and Engineering Applications*, 257–279. Kluwer Academic Publishers.

[Marusic *et al.*2001] Marusic, I.; Chandler, G. . V.; Interrante, V.; Subbareddy, P. K.; and Moss, A. 2001. Real Time Feature Extraction For the Analysis of Turbulent Flows. In *et al.*, R. L. G., ed., *Data Mining for Scientific and Engineering Applications*, 223–238. Kluwer Academic Publishers.

[Montalenti, Sørensen, & Voter2001] Montalenti, F.; Sørensen, M.; and Voter, A. 2001. Closing the gap between experiment and theory: Crystal growth by temperature accelerated dynamics. *Phys. Rev. Lett.* 87:126101.

[Parthasarathy & Coatney2002] Parthasarathy, S., and Coatney, M. 2002. Efficient discovery of common substructures in macromolecules. In *IEEE International Conference on Data Mining*.

[Parthasarathy *et al.*1999] Parthasarathy, S.; Zaki, M.; Ogihara, M.; and Dwarkadas, S. 1999. Incremental and interactive sequence mining. ACM Confereince on Information and Knowledge Management (CIKM).

[Portela1997] Portela, L. M. 1997. *On the identification and classification of vortices*. Ph.D. Dissertation, Stanford University.

[Quinlan1996] Quinlan, J. R. 1996. Induction of decision trees. *Machine Learning* 5(1):71–100.

[Reinders, Jacobson, & Post2000] Reinders, F.; Jacobson, M. E. D.; and Post, F. H. 2000. Skeleton Graph Generation for Feature Shape Description. In *Joint Eurographics-IEEE TCVG Symposium on Visualization*, 73–82.

[Reinders, Post, & Spoelder1999] Reinders, F.; Post, F. H.; and Spoelder, H. J. W. 1999. Attribute-Based Feature Tracking. In *Joint Eurographics-IEEE TCVG Symposium on Visualization*, 63–72.

[Richie *et al.*2002] Richie, D.; Kim, J.; Hazzard, R.; Hazzard, K.; Barr, S.; and Wilkins, J. 2002. Large-scale molecular dynamics simulations of interstitial defect diffusion in silcon. volume 731, W9.10. Material Research Society.

[Richie, Kim, & Wilkins2001] Richie, D.; Kim, J.; and Wilkins, J. 2001. Applications of real-time multiresolution analysis for molecular dynamics simulations of infrequent events. volume 677, AA5.1. Material Research Society.

[Samtaney *et al.*1994] Samtaney, R.; Silver, D.; Zabusky, N.; and Cao, J. 1994. Visualizing Features and Tracking Their Evolution. *IEEE Computer* 27(7):20–27.

[Shen, Chiang, & Ma1999] Shen, H.-W.; Chiang, L.; and Ma, K.-L. 1999. Time-Varying Volume Rendering Using a Time-Space Partitioning Tree. In *Proceedings of Visualization '99*, 371–378.

[Silver & Wang1997] Silver, D., and Wang, X. 1997. Tracking and Visualizing Turbulent 3D Features. *IEEE Transactions on Visualization and Computer Graphics* 3(2).

[Sørensen & Voter2000] Sørensen, M., and Voter, A. 2000. Temperature-accelerated dynamics for simulation of infrequent events. *J. of Chem. Phys.* 112:9599.

[Sujudi & Haimes1995] Sujudi, D., and Haimes, R. 1995. Identification of Swirling Flow in 3D Vector Fields. In *AIAA 12th Computational Fluid Dynamics Conference, Paper 95-1715*.

[Thampy2003] Thampy, S. 2003. Feature Tracking in Two-Dimensional, Time-Varying Data Sets. Master's thesis, Mississippi State University.

[Thompson *et al.*2002] Thompson, D.; Machiraju, R.; Jiang, M.; Nair, J.; Craciun, G.; and Venkata, S. 2002. Physics-Based Feature Mining for Large Data Exploration. *IEEE Computing in Science and Engineering* 4(4):22–30.

[Vlachos, Kollios, & Gunopulos2002] Vlachos, M.; Kollios, G.; and Gunopulos, D. 2002. Discovering similar multidimensional trajectories. In *M. Vlachos, G. Kollios, and D. Gunopulos. Discovering similar multidimensional trajectories. In ICDE, San Jose, CA, 2002*.

[Voter1997] Voter, A. 1997. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* 78:3908.

[Voter1998] Voter, A. 1998. Parallel replica method for dynamics of infrequent events. *Phys. Rev. B* 57:13985.

[Wang *et al.*1997] Wang, X.; Wang, J.; Shasha, D.; Shapiro, B.; Dikshitulu, S.; Rigoutsos, I.; and Zhang, K. 1997. Automated discovery of active motifs in three dimensional molecules. In *Knowledge Discovery and Data Mining*, 89–95.

[Wolfson & Rigotsos1997] Wolfson, H., and Rigotsos, I. 1997. Geometric hashing: an overview. *IEEE Computational Science & Engineering* 10–21.

[Yip & Zhao1996] Yip, K., and Zhao, F. 1996. Spatial Aggregation: Theory and Applications. *J. of Artificial Intelligence Research* 5:1–26.