# Enhancing robustness and sensitivity of metrics for WP applications.

Matthew Sottile

July 13, 2006

## Abstract

Quantitative comparison of experimental data and simulation output requires robust measures and metric algorithms. Often differences that we ask a metric to reveal are subtle features, thus we also seek metrics that are sensitive. Developing methods that do not sacrifice one for the other is a difficult process. This document introduces the issues that arise in this development process, the research that we have identified as necessary based on our experience constructing and applying metrics, and a proposed set of research topics we would like to pursue to move in this direction.

## 1 Introduction

Understanding of data and simulation output requires a disciplined method for characterizing features within data sets and defining a method for comparative analysis of simulation or data ensembles. Domain experts typically obtain qualitative comparisons by looking for features in data such as apparent differences in value statistics, spatial distributions of intensity, and small-scale features such as vortex and shock structures. Automation of this process is often desired. To automate the computation of metrics and measures, one must posess tools that are robust to inherent noise and fuzziness present in data while maintaining sensitivity to the subtle features of interest.

## 2 Robustness and sensitivity

Robustness of a metric boils down to the behavior of the metric under subtle data or parameter changes. Consider a set of images of a static object, where the only differences in the images are due to natural fluctuations in the imaging device or process. A user of the metric would often desire the metric to not vary wildly and unpredictably due to these noise-like fluctuations in the image. They would seek a metric that is *robust* to such noise.

Sensitivity on the other hand requires that a metric is able to identify subtle changes in images. If one has a large set of images, all of which differ in subtle ways, then a metric is desired that captures these nuances of the data. For example, one may be measuring characteristics of a hydrodynamical jet structure. The shape of this jet may vary in a subtle, but physically important way. It is desirable that a metric be *sensitive* to these subtle changes.

### 2.1 Issues that degrade robustness

Robustness is a subjective term. From a purely context-free image processing perspective, most metrics are reasonably stable and robust. Specialized analysis of data carries with it stricter requirements on the processing algorithms from which a measure or metric is constructed. Frequently an analyst will use a metric to probe data for a specific feature of interest, and to tease out differences between data sets relative to this specific feature from a richly structured feature set. Physical effects within data sets can introduce a certain degree of fuzziness or ambigu-

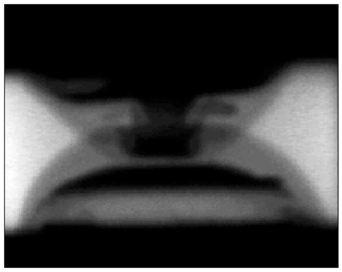

Figure 1: Experimentally obtained data on which metric computations are to be performed.


Figure 2: High gradient magnitude regions of the image are shown in white.

ity into the structure of interesting features. Diffuse boundaries, mixing of materials, and time integration due to imaging exposure times are all potential sources of fuzziness that arise in practice.

Heuristic decisions in analysis algorithms that act as "tie breakers" in these fuzzy regions lead to problems with respect to metric behavior. This becomes most apparent when analyzing time series data, where metrics can experience sharp discontinuities in otherwise smoothly changing images due simply to a heuristic decision. This jump and jitter in the metric often appears to disagree with the intuition of the expert, and determining whether it is the intuition or metric that is misleading is an open and important question.

A concrete example of these issues arises in shape extraction for image data. Segmentation algorithms that perform this separation of shape from background are forced to make a single decision from this set of possible candidates. There may exist many curves that can be drawn within the image that delinate background from shape, each of which is valid from the perspective of an expert.

# 3 Areas of research

In this section, three important research directions are discussed that are motivated by experiences with our metrics work in the context of the Wilde jets and shocktube problems.

## 3.1 Robust shape definition

In Figure 2 we see the high gradient regions of an experimentally obtained jet image shown in Figure 1. The white region corresponds to the region of the image containing edge-like features, making it very clear that no single curve captures the shapes precisely. All metrics that require derivation of a shape before metric computation force algorithms to make this decision as to where the curve should be placed.

We would like to investigate methods for using all information about the candidate regions of the shape boundary. We could then compute a metric that attempts to capture the values it would take on for a range of reasonable boundaries that fall within the edge-like region. Thus instead of producing a single scalar that varies depending on the heuristic used to define the boundary curve, we hope to provide a more stable interval. This is more a robust method, as it is less likely to vary due to metric parameters since less metric parameters and heuristics are required. Furthermore, yielding an interval and possible distribution function on the metric value provides not only a metric, but a measure of error and uncertainty on the information that the metric provides.

## 3.2 Using what we know about the past and future

We are also interested in the construction of metrics that are stable with respect to the time evolution of the images they are applied to. Often a time series of images contains smoothly varying features, and one would desire a metric that varies smoothly in some (likely nonlinear) manner with respect to the image evolution. Unstable metrics that do not take into account time evolution of features and the characteristics of neighboring time snapshots will have a tendency to contain sharp discontinuities, jitter, or general “noise” that obscures the relationship of the images in terms of their distance in time. We are interested in exploring the use of the time axis in data sets to stabilize metrics when applied to data where temporal information is available. Furthermore, we would like to understand the root causes of instabilities that are observed in order to determine if they are actually yielding some unexpected but important insight into the data.

## 3.3 Beyond $\mathcal{L}^2$

Often it is the case that when comparing images or data sets, one derives a set of measures that take on the form of a one dimensional sequence such as a histogram or ordered sequence of points along a boundary curve. The comparison required to compute a metric from these measures often turns into the problem of determining a measure of similarity between two sequences. Trivial metrics such as $\mathcal{L}^2$ or RMS differences are easily fooled by phase shifts, single local dilations, circular shifts, and other features that are often not considered large scale differences at all. We are exploring the use of novel methods for sequence similarity metrics that are most appropriate for problems such as ours where three important characteristics exist: scientific data is real valued, conservation laws may be necessary to obey, and we would like the similarity metric to consider *all* data points. These requirements make it apparent that naive application of traditional sequence similarity metrics from speech processing and bioinformatics is unwise and leads to unpredictable and unexpected outcomes.

# 4 Motivating contexts

Clearly metrics and measures are required for quantitative comparisons of data. Often a metric user can use a suite of metrics to understand precisely why two data sets are different. The metric helps guide their expert eye and is used as a tool. Many other instances exist where the metric is not used directly by the user, but within a larger tool that performs a higher level task that requires the existence of a metric for computing distance functions. When a metric becomes embedded within a larger tool, it becomes more important that the metric and its constituent algorithms be robust and stable enough such that they can run predictably and reliably with minimal hand tuning by the user.

## 4.1 Classification

Many methods exist for automatically determining the manner by which a data ensemble can be partitioned into groups that contain data sets exhibiting common features. The heart of these algorithms is a similarity metric (distance function). It is often interesting to have an arsenal of metrics available in order to explore how partitioning changes under different metrics in order to understand something about the commonalities and differences between the data sets in question. Classification and general metric-based data mining is necessary to extract as much information as possible from existing experimental data given the prohibitive environment for conducting new experiments. A great deal has yet to be learned from the large archives of data that the lab posesses. A similar problem exists for large scale simulation output.

## 4.2 Optimization

Optimization procedures seek a set of parameters for a function or simulation that produce output best matching some target objective. Each parameter set that is provided to the function as input yields an output data set, and a comparison must be made with the objective data set in order to evaluate the quality of the parameter set being tested. Metrics that elucidate different features of interest allow users to tune the optimizer to seek context specific optimal solutions, versus generic metrics that lack the ability to focus on features of interest.

# 5 Proposed research tasks

These examples are only two of many that exist where robust metrics and measures are necessary in order to build stable tools that users can have a high degree of confidence in. We hope that a continuing participation in projects with the end-users of these methods with support for general research into issues with algorithms will allow us to move easier in this desirable direction.

The concrete set of activities that we propose to perform are as follows.

1. **Shape extraction:** We would like to address the problem of unstable shape extraction algorithms by investigating methods that use all possible information about the set of potential shapes, yielding a metric that characterizes this family. Extension of existing segmentation and boundary parameterization algorithms is our first step.

2. **Whole image methods:** The DDMA team is beginning to consider methods that do not require shape extraction, removing a significant source of instability and robustness degradation that impact current methods. We would like these methods to compliment shape-based methods that currently have significant advantages with respect to computational efficiency. At least three new methods are in development by the DDMA team.

3. **Robust 1-d metrics:** The phase where metrics are computed from measures often boils down to deriving a scalar distance from two one dimensional sequences. Existing methods for sequence comparison are unstable and difficult to tune for specific problems. We are investigating warping techniques that address the shortcomings of current methods.

4. **Computable higher-dimensional metrics:** A large number of metrics that are computed from higher dimensional measures have been developed. Unfortunately, the algorithms tend to be computationally intractable for the volumes of data that we have considered. We are currently investigating algorithmic improvements in two and three dimensional warping algorithms, and multi-scale geometric measures for two dimensional data.