Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
SWUTC/06/167452-1		

4. Title and Subtitle	5. Report Date
MICROSCOPIC DATASETS: A NOVEL APPROACH APPLIED TO VISUALIZATION OF SPATIOTEMPORAL FLOW REGIONS	May 2006
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Paul Nelson and Matthew J. Fields	Report 167452-1

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Texas Transportation Institute Texas A&M University System College Station, Texas 77843-3135	11. Contract or Grant No. 10727

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Southwest Region University Transportation Center Texas Transportation Institute Texas A&M University System College Station, Texas 77843-3135	14. Sponsoring Agency Code

15. Supplementary Notes

Supported by general revenues from the State of Texas.

16. Abstract

A parsimonious approach to microscopic traffic flow datasets is suggested. This approach is based on provision of minimal data, along with a set of supporting tools termed as feature extraction operators that are intended to provide researchers with the flexibility to extract those particular features of the data that they desire to study. As an illustration of the possibilities of this approach, an attempt is made to validate the hypothesis that traffic flow decomposes into spatiotemporal regions representing one of the four classes of congested flow, shock wave, acceleration wave or free flow. This approach employs the classical conceptual framework of the field of pattern recognition, as applied to microscopic datasets. The specific microscopic dataset employed is that labeled "I-405 Northbound at Mulholland Drive, Los Angeles" in the 1985 study conducted by JHK Associates for the FHWA. The classical approach of plotting and manually analyzing vehicle trajectories is initially employed, to establish some approximation to ground truth. Then it is demonstrated that speed alone is inadequate to support the desired classification. Finally, a 4-means cluster analysis in velocity-acceleration feature space is employed to demonstrate that a spatiotemporal plot of the resulting cluster numbers provides a decomposition more-or-less as expected.

17. Key Words	18. Distribution Statement
Microscopic Data Analysis, Traffic Patterns, Classes Of Traffic Flow, Pattern Recognition, *K*-Means Clustering	No restrictions. This document is available to the public through NTIS: National Technical Information Service 5285 Port Royal Road Springfield, Virginia 22161

19. Security Classif.(of this report)	20. Security Classif.(of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	33	

MICROSCOPIC DATASETS: A NOVEL APPROACH APPLIED TO VISUALIZATION OF SPATIOTEMPORAL FLOW REGIONS

by

Paul Nelson*
Department of Computer Science
Texas A&M University
College Station, TX 77843-3112
Phone: +1-979-845-4132
Fax: +1-979-862-3684
Email: pnelson@cs.tamu.edu

Kinetic Traffic Models
P. O. Box 13535
College Station, TX 77841-3535
Phone: +1-979-229-7421

Matthew J. Fields
Department of Computer Science
Texas A&M University
College Station, TX 77843-3112
Phone: +1-979-458-3611
Fax: +1-979-847-8578
Email: mjfields@neo.tamu.edu

Report 167452-1
SWUTC Project 167452

Southwest Region University Transportation Center
Texas Transportation Institute
Texas A&M University System
College Station, Texas 78743-3135

May 2006

* Corresponding author. Also affiliated with the Departments of Civil Engineering, Nuclear Engineering and Mathematics at Texas A&M University.

ABSTRACT

MICROSCOPIC DATASETS: A NOVEL APPROACH APPLIED TO VISUALIZATION OF SPATIOTEMPORAL FLOW REGIONS

A parsimonious approach to microscopic traffic flow datasets is suggested. This approach is based on provision of minimal data, along with a set of supporting tools termed as feature extraction operators that are intended to provide researchers with the flexibility to extract those particular features of the data that they desire to study. As an illustration of the possibilities of this approach, an attempt is made to validate the hypothesis that traffic flow decomposes into spatiotemporal regions representing one of the four classes of congested flow, shock wave, acceleration wave or free flow. This approach employs the classical conceptual framework of the field of pattern recognition, as applied to microscopic datasets. The specific microscopic dataset employed is that labeled "I-405 Northbound at Mulholland Drive, Los Angeles" in the 1985 study conducted by JHK Associates for the FHWA. The classical approach of plotting and manually analyzing vehicle trajectories is initially employed, to establish some approximation to ground truth. Then it is demonstrated that speed alone is inadequate to support the desired classification. Finally, a 4-means cluster analysis in velocity-acceleration feature space is employed to demonstrate that a spatiotemporal plot of the resulting clusters provides a decomposition more-or-less as expected.

EXECUTIVE SUMMARY

MICROSCOPIC DATASETS: A NOVEL APPROACH APPLIED TO VISUALIZATION OF SPATIOTEMPORAL FLOW REGIONS

The challenge considered is identification, from microscopic data, of spatiotemporal regions representing the four kinematic-wave macroscopic classes of traffic flow, namely *congested flow* (or *enqueued flow*)*, free flow, shock waves* and *acceleration waves* (AKA *queue discharge*). A two-step approach is proposed for investigation. In the first (microscopic) step individual vehicles at a given time (i.e., individual "patterns") are sorted into one of these four classes, depending upon their speed and acceleration at the time; the second (macroscopic) step then consists of an attempted visual identification of the relevant spatiotemporal regions from a (preferably color) space-time plot of the classifications of the individual patterns. A detailed description is provided (Section II) of a microscopic dataset, with associated incident, that was employed to test this approach. Kinematic-wave simulations of the flow following the incident are presented (Section III), in the form of time-space-speed plots. In Section IV the data following the incident are represented as both a time-space-speed plot and a vehicle trajectory plot, and the results are compared and contrasted to the preceding kinematic-wave simulation, and discussed in the light of expectations (e.g., difficulty distinguishing between acceleration waves and subsequent free flow) emerging from this simulation. Highlights are that it is difficult to distinguish between shock and acceleration waves from a spatiotemporal plot of speed alone, while the substantial fine structure (start-stop waves) present in the vehicle trajectories make it difficult to visually ascertain the boundaries of the larger scale congested-flow region. These deficiencies motivate a discussion (Section V) of the general elements of pattern recognition, including a detailed discussion of the implementation of each of these elements as subsequently (Sections VI and VII) employed in the present work, along with some consideration of how pattern recognition might in the future be applied to traffic flow. A canonical matrix-oriented conceptual organization of microscopic traffic-flow datasets is described (Section VI), and arguments are presented that suggest this organization is likely to more efficient than other alternatives. This argument hinges on the fact that most such datasets likely to be available in the near future will come from near-ground observations, and that for such observations the total number of frames of data and the total number of vehicles are likely to be considerably larger respectively than the typical number of vehicles per frame or of frames per vehicle. A conceptual library of feature-extraction operators, which are intended to facilitate extraction of pattern attributes from datasets organized in the canonical matrix format, is described. A MatLab implementation of this library is employed (Section VII) to extract the speeds and accelerations that are the basis for the proposed microscopic classification. A highly supervised k-means (k=4) algorithm is employed (also Section VII) for the microscopic classification step. Associated silhouette diagrams suggest the corresponding congested-flow and free-flow clusters are very well separated, but the shock-wave and acceleration-wave clusters not as well so. For the best values found for the parameters in the distance function employed for the k-means algorithm the two-step approach nonetheless was at least as effective in identifying the kinematic-wave spatiotemporal regions as the classical approach via speed plots. It is conjectured that oscillatory flow, particularly within acceleration waves and possibly due to start-stop waves, are a major cause for the lack of superiority to the classical approach. Modifications of the current approach, including a k-means algorithm with $k = 6$, are suggested as possible improved variants of the general two-step approach.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

List of Figures

List of Tables

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

ACKNOWLEDGMENTS

The authors recognize that support for this research was provided by a grant from the U.S. Department of Transportation, University Transportation Centers Program to the Southwest Region University Transportation Center which is funded 50% with general revenue funds from the State of Texas. The work of Paul Nelson was partially supported by Kinetic Traffic Models. The authors thank Dr. Henry Lieu, of the Turner-Fairbank Highway Research Center of the U.S. FHWA, for making available the collection of 1985 JHK datasets. This work was presented at the 85th Annual Meeting of the Transportation Research Board (Washington, D.C., January 22-26, 2006). An abbreviated version of this report appeared in the proceedings of that meeting (published in CD format), under the slightly different title "Microscopic Datasets: A Novel View Applied to Visualization of Spatiotemporal Flow Regions."

I. INTRODUCTION

At any given time, subject to precise definitions of the boundaries between "high" and "low" and between "negligible" and "significant," any individual vehicle within a traffic flow stream falls into exactly one of the following four classes:

- Negligible acceleration at low speed;
- negligible acceleration at high speed;
- significant negative acceleration; or
- significant positive acceleration

These behaviors for individual vehicles are the microscopic counterparts of the classical macroscopic flow classes generally termed respectively as "congested flow" (or "enqueued flow"), "free flow," "shock wave" and "acceleration wave" (or "queue discharge"). Drivers are familiar with each of these four macroscopic classes, and they are well reproduced, as spatiotemporal regions, by such mathematical models of traffic flow as the kinematic-wave model (KWM; cf. 1,2).† It is widely assumed, often tacitly, that any traffic flow pattern decomposes into spatiotemporal subregions within each of which the flow predominantly falls into exactly one of these classes. *The objectives of the work described here were to describe a novel scheme to represent and utilize microscopic datasets, and as an illustration of the utility of this view to employ the methods of pattern recognition* (3) *to facilitate the visual recognition, within microscopic traffic flow data, of spatiotemporal regions containing these various macroscopic classes.*

The structure of this report is as follows. The microscopic traffic flow dataset that we employ prototypically is described in Section II. In Section III we exemplify the spatiotemporal regions that the kinematic-wave model predicts for one interesting scenario arising from the dataset of the preceding section. Section IV is intended to illustrate the difficulties that arise from attempts to identify instances of the above-mentioned four macroscopic classes of traffic flow from the more-or-less classical approaches of visual analysis of vehicle trajectory plots and of space-time-speed plots. We subsequently take these difficulties as indicating the need for some type of preprocessing prior to attempting to visually identify, from microscopic data, spatiotemporal regions corresponding to the four classical macroscopic classes. We further seek tools for such preprocessing from the field of pattern recognition, and accordingly summarize the essence of this approach in Section V. One essential element of pattern recognition is "feature extraction," by which one intends the extraction – typically from a large number of both measurements per data point (pattern) and of data points – of values of the commonly much smaller number of attributes that the proposed classification algorithm is intended to be based on. In Section VI we describe a conceptual view of microscopic traffic data sets, and of an associated library of operators intended to facilitate the extraction of designated feature attributes from microscopic data sets. (While the feature-extraction operators are described conceptually, we also have developed MatLab implementations of these, which are employed for the work of Section VII.) In Section VII we first discuss application of the pattern recognition technique known as *k-means cluster analysis* to the automated assignment of individual vehicles, at any given time, to one of the four above-mentioned microscopic classes, based upon the choice of vehicular speed and acceleration as the designated features. Then we present and discuss the results of employing the *k*-means microscopic classification algorithm as a preprocessor for visual identification of spatiotemporal regions corresponding to the four classical macroscopic classes. Section VIII contains our conclusions, and suggestions for subsequent further work.

We thus consider here two distinct, hypothetically related, pattern recognition problems. The macroscopic pattern recognition problem of ultimate interest is to determine if an arbitrary traffic-flow dataset decomposes into spatiotemporal regions, each of which corresponds to one of the four classical macroscopic classes, and to identify the corresponding regions. The methodology applied to this macroscopic problem is visualization from space-time plots of some set of microscopic feature attributes; identification of patterns via visual recognition is an instance of what is sometimes known as Exploratory Data Analysis (4). Thus we are *not* applying the formal methodology of pattern recognition, as outlined in Section IV below, to the macroscopic pattern recognition problem; but see Section V below for further discussion of this possibility.

† Underlined Arabic numerals designate items in the list of references, which is located at the end of the body of this report.

We discuss in Section IV below the difficulties that arise from attempting to identify the macroscopic regions of interest from plots of the "raw" microscopic data. The approach we explore toward resolution of this issue is to apply the formal methods of pattern recognition, specifically a supervised version of *k*-means clustering, to inclusion of both vehicle speed and vehicle acceleration as an aid in classification. We then attempt to visually identify spatiotemporal regions corresponding to the classical macroscopic classes via space-time plots of the microscopic class of individual vehicles. The underlying hypothesis is that individual vehicles in spatiotemporal regions corresponding to one of the macroscopic classes will predominantly belong to the corresponding microscopic class, and that the human eye will detect the inevitable individual outliers and boundary regions for what they are.

Although the present work is phrased in terms of the problem of recognizing the macroscopic flow classes described above, we also intend it as an example of how techniques of pattern recognition can generally be used for automated classification. Alternate classification schemes for traffic flow include the well-known level of service (LOS) scheme (e.g., HCM 2000, p. 2-3), the scheme of B. S. Kerner and co-workers (5,6) and a scheme suggested by Schönhof and Helbing (7).

The validity or invalidity of one such classification scheme is *a priori* independent of the validity or invalidity of the others. The LOS classes seem to be sufficiently well-defined, for a particular type of facility, so that one could, subject to availability of appropriate data, develop a corresponding pattern recognition algorithm (particularly a classifier), in the sense of the elements of pattern recognition as outlined in Section V below. We have no doubt that the classifications schemes of Kerner *et al.* and of Schönhof and Helbing, as cited above, have significant merit, but it does not appear they are yet sufficiently well-developed to permit application of the formal methods of pattern recognition. More specifically, it is not clear how one could view the collections of classes suggested by these workers as comprising an algorithmically defined partition of some low-dimensional feature space.

II. A MICROSCOPIC DATA SET

Traffic flow data are commonly available, especially in a traffic operations setting, as macroscopic loop-detector data, in which traffic flow variables (e.g., flow, mean speed, occupancy) within distinct lanes at essentially point longitudinal locations separated by a kilometer or so are aggregated over some period of time, typically on the order of a few minutes. In order to best understand these data, and respond appropriately in an operational setting, it would be quite useful to have systems that would reinterpret these data as spatiotemporal regions within each of which exactly one of the classical classes of traffic flow prevails. At the present state of development it appears challenging to attain this objective in an automated algorithmic fashion, not least because of the spatially sparse and temporally aggregated nature of loop-detector data. For these reasons, and to investigate the ground truth of the classical four-class partitioning of traffic flow patterns, we employ microscopic data, which reflects the trajectories of individual vehicles over some period of time and section of freeway.

Microscopic data typically are obtained photographically from elevation, which makes them relatively expensive, both to obtain and to process. (Although significant strides are being made in addressing both of these issues, cf. (8,9) so that it is unclear how long it will be infeasible to use such data for operational purposes. The dataset used illustratively here is one of those described in a 1985 report submitted by JHK and Associates (10), to the US DOT/FHWA. The alternate candidate datasets considered were those currently available from the NGSIM program (11); however, the latter data were taken at a rate of 15 frames per second, as compared to the one frame per second of the JHK data set. Because of the resulting size of the NGSIM dataset, for ease of the presently intended proof of principle we therefore chose the JHK dataset.

The particular JHK dataset employed here was that labeled as "I-405 Northbound at Mulholland Drive, Los Angeles." The metadata description of this dataset in Ref. 10 is:

"This site is an upgrade, 1341 feet (408.5 m) long. The section is approximately one-half mile from the top of a 2.5-mile (4.04 km) long grade. The level of service was in the C to E range. The site consists of a tangent section, with five 11-foot (3.3 m) lanes and a narrow left shoulder.

Several traffic disturbances occurred during the filming. Approximately 22 minutes into the film there was a vehicle stalled in the right lane just downstream of the section. The vehicle was moved off onto the right shoulder 2 minutes later. Approximately 38 minutes into the film a truck traveling very slowly passed through the section and stopped on the shoulder approximately 1,000 feet (304.6 m) downstream of the section. There appeared to be some congestion developing in the left lanes as a partial result of these events, possibly caused by lane changing to avoid these problems."

See Ref. 10 for a diagram of the subject section. The JHK datasets were obtained by digitizing vehicle positions from time-lapsed aerial photographs taken from a fixed-wing aircraft. The sites were filmed at one frame per second, and data were reduced to digital form for one hour, at each of the 14 sites. The data files from each of these sites is a text (and therefore sequential) file, with each line formatted into nine fields: frame number, vehicle id, vehicle type code, vehicle length (feet), speed (mph), distance from beginning of section to front of vehicle (feet), distance from right edgeline to middle front of vehicle (feet), color code and lane number (right lane = lane 1).

Our generic parsimonious view of microscopic data sets is presented below (Section V). As this view is substantially motivated by a perceived need to provide flexibility in processing microscopic datasets, in order to enable application of the methods of pattern recognition, it seems appropriate first to illustrate the need for this flexibility, and to describe the elements of pattern recognition. These are the respective subjects of Sections IV and V, but first we explore somewhat further the implications of the foundational theory underlying the kinematic-wave classification scheme described in the Introduction.

III. REPLICATION VIA THE KINEMATIC-WAVE MODEL

As previously suggested, the objective here is to explore, in the context of the dataset described in the preceding section and especially the incident at 22 minutes into the observations, an approach to automated facilitation of visual recognition of spatiotemporal regions representing the four classes of kinematic-wave flows described in the Introduction. As an adjunct to this effort it will prove helpful to have in hand some expectation as to the spatiotemporal regions the KWM predicts for the target scenario. The objective of this section is to provide such predictions. First we provide a qualitative description, and then a quantitative description based upon computational solution of the KWM under a particular traffic stream model (TSM; also known as "flow-density relation," or *fundamental diagram*).

Qualitative Description

Before 22 minutes (1320 seconds) free flow prevails, as in the leftmost region of the sketch (line diagram) in Figure 1. At 22 minutes a shock wave forms (the free-flow/congested-flow boundary in Figure 1) and propagates upstream, leaving congested flow downstream. When the stalled vehicle is moved to the shoulder an acceleration wave forms, centered at the site of the incident. The upstream edge of this acceleration wave propagates faster upstream than the preceding shock wave, and eventually merges with it. In Figure 1 this merger is shown to occur within the study section, which is conceivable; however, it also could occur upstream of the entrant boundary, depending on the precise time the stalled vehicle was cleared, and the speeds of propagation of the various waves. This merger forms a shock that eventually propagates downstream. In Figure 1 this forward shock is caricatured as a straight line (constant speed), but for a classical concave fundamental diagram there will be curvature corresponding to increasing wave speed as this shock progresses downstream. Of course all of this tacitly assumes that conditions sufficiently far upstream and downstream of the study section remain in the original free-flow state sufficiently long so as not to introduce any additional effects within the study section for some four minutes or so following the time at which the stalling occurs.

Quantitative Description

A more quantitative, and somewhat alternative, portrayal of a kinematic-wave scenario compatible with the metadata of the preceding section is shown in Figure 2. These results were obtained by computational solution of a kinematic-wave model, by means of the computational approximation of Godunov; cf. [12] for a more detailed description of the computational methodology, with references to related prior work. For the present results a triangular [13, 14] TSM was employed, with flow given by

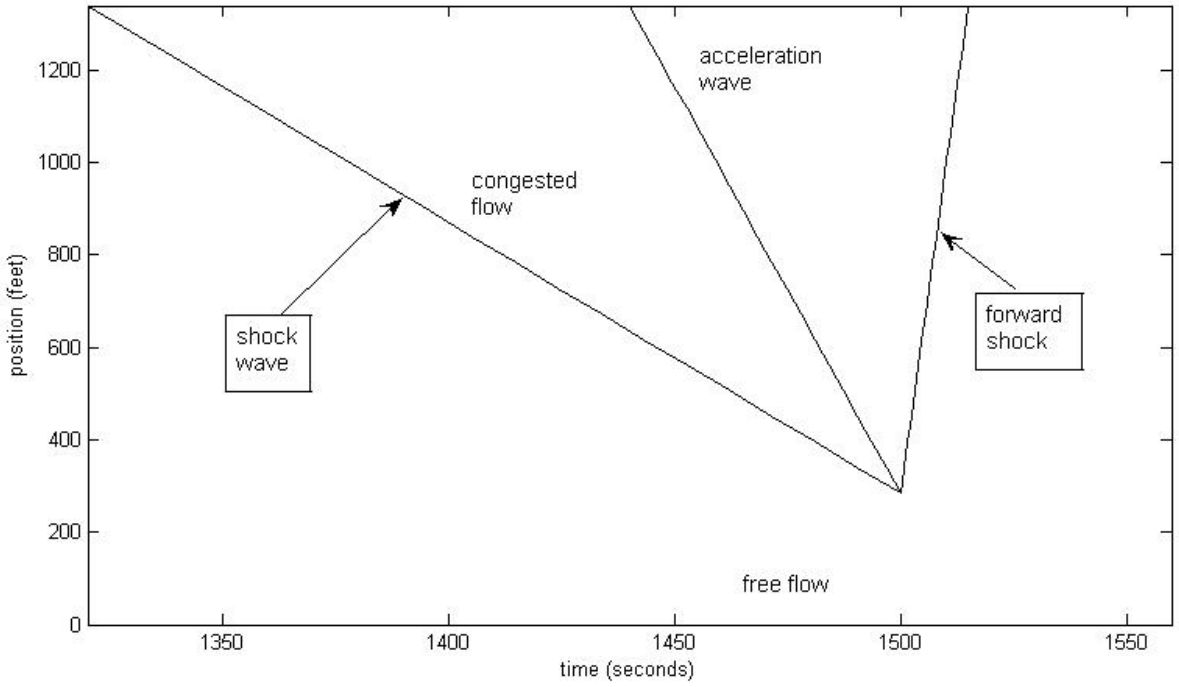

Figure 1 - Time (horizontal axis) and space (vertical axis) plot of the flow, as anticipated from the KWM. See text for detailed explanation.

$$q(c) = \begin{cases} v_{ff} c, \ c \le c_{cap}, \\ v_{ff} c_{cap} (c_{jam} - c) / (c_{jam} - c_{cap}), \ c \ge c_{cap}, \end{cases}$$

where

$$v_{ff} = 90 \text{ ft per sec (~61.4 mph)}$$

is the free-flow speed, and

$$c_{cap} = .0333... \text{ vehs per foot (over five lanes) and}$$

$$c_{jam} = .1875 \text{ vehs per ft (five lanes, or ~ 198 vpmpl),}$$

are respectively the density at which capacity flow occurs and the jam density. Initial conditions were taken as .03 vehicles per ft, the entry section was taken at -2500 feet (i.e., 2500 feet upstream of the observed section), and boundary conditions taken there were flow corresponding to the initial density (i.e. 2.7 vehs per sec, over the entire five-lane roadway), over the entire time of the simulation. This was justified because in this simulation the downstream edge of the congested-flow region (i.e., the boundary between it and the upstream acceleration wave) overtook the preceding shock wave propagating upstream at approximately -2000 feet, thus there was no spillback of the congestion from the incident to the assumed upstream boundary. The computational downstream boundary was set at 1500 feet (i.e., 159 feet further downstream of the downstream boundary of the observed section). The stalled vehicle was assumed to be located at this computational boundary, and to remain in the roadway there for four minutes. (This gives a better match to the data than does the two minutes described in the metadata quoted in the preceding section.) Downstream supply at this downstream computational boundary was taken as the capacity flow of 3 vehs per second, except that it was reduced to 4/5 of this value between times 1320 and 1560 seconds,

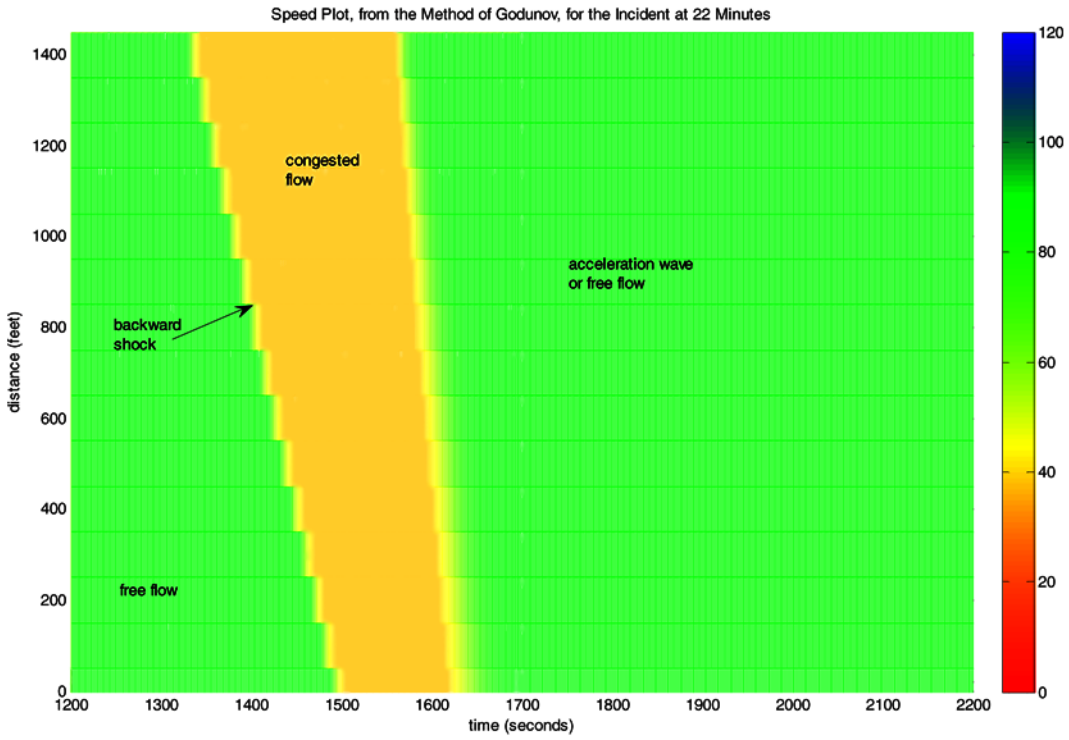

Figure 2 – Time-space-speed plot, from the method of Godunov and a triangular fundamental diagram, for the incident at 22 minutes. Speeds (feet per second) are displayed on a color axis, as defined by the colorbar.

corresponding to the assumption that the stalled vehicle blocked one lane, until it was cleared four minutes after the initial stalling. However, in order to facilitate comparisons with observations, only the portion of the predicted kinematic-wave speeds covering roughly the observed section of the roadway (more precisely from x=0 to x=1500 feet) is plotted in Figure 2.

A region of congested flow appears prior to the incident, as expected. The (backward) shock wave initiated by the incident also clearly is displayed, as the boundary between the free-flow and congested regions. In actuality we do not expect to see this boundary quite so cleanly. It will have a finite extent, corresponding to typical distances over which vehicles decelerate, and will be irregular, corresponding to variations in driving behavior and circumstances (e.g., the necessity to respond to possible forced merging). In one of the few studies of the structure of shock waves in traffic, Muñoz and Daganzo (15) describe "transition zones" approximately 1 km in width; that is somewhat wider than a typical deceleration distance, and also than the present work is able to verify (see Section VII).

The congested-flow region appears in the KWM simulation of Figure 2, more-or-less as expected. However, what significantly differs between Figures 1 and 2 is the absence in Figure 2 of any semblance of a boundary between the acceleration wave following the congested flow and the expected subsequently restored free flow. That is because the acceleration wave is, for the present case, entirely at the point of capacity flow in the underlying triangular TSM, and therefore the corresponding speeds are equal to the free-flow speed, exactly as in the free-flow region. Similar simulations with other TSMs yield a boundary between the acceleration wave and the trailing free-flow region that is discernible in time-space-speed plots, but the triangular results at least serve a warning that this distinction may be observationally difficult to make.

The fact that speeds are identical in the free-flow and acceleration-wave regions does not mean that other dependent variables have the same values. Figure 3 is a (three-dimensional black-and-white) time-space-density plot of the same scenario as Figure 2, except that now conditions are displayed along the entire length of roadway

between the entrant and exiting boundaries assumed for computational purposes. The boundary between the acceleration wave and the trailing free flow region now clearly appears as the lower ridge that propagates downstream (forward shock) very rapidly (in fact, at free-flow speed). This seems unlikely to provide a suitable basis for observationally distinguishing these two regions, because it is difficult to measure densities to the required level of accuracy (error < 10%). However, it does suggest that attributes other than speed might be useful in helping to expose this boundary. In Section VII below we explore the use of vehicular accelerations, along with speeds.

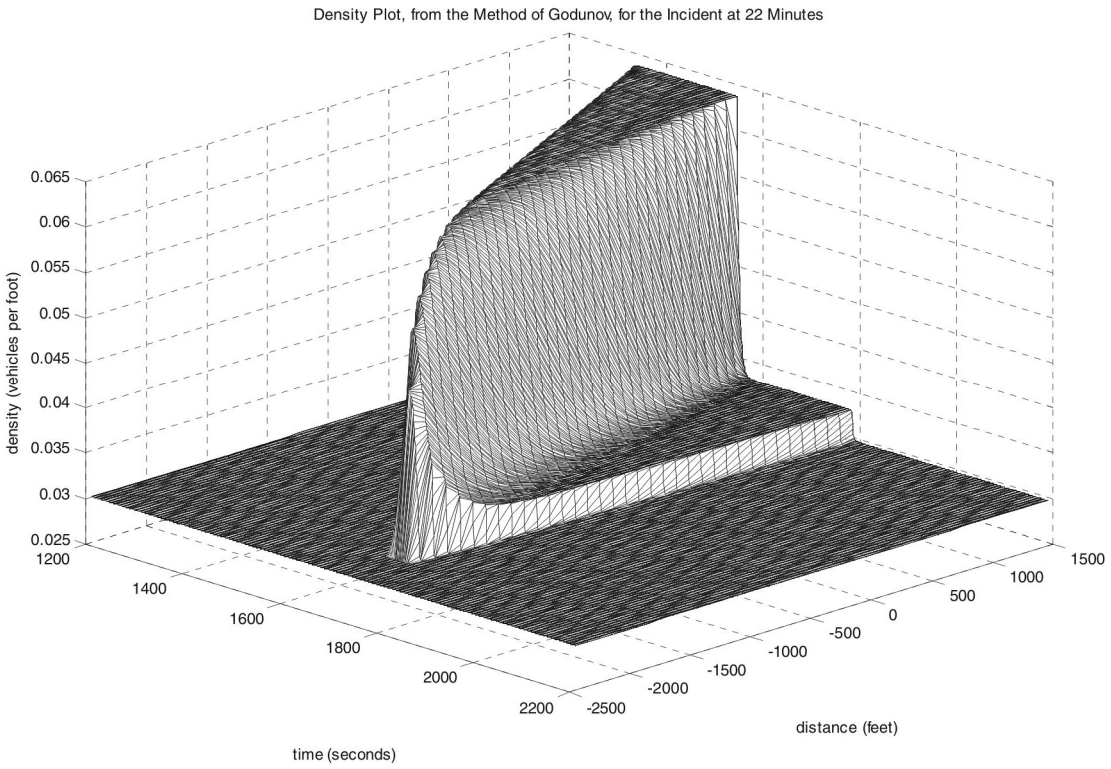

Figure 3 - Time-space-density plot, from the method of Godunov and a triangular TSM, for the incident at 22 minutes.

IV. CLASSICAL APPROACHES

In this section we explore the utility of time-space-speed plots, and of vehicle trajectory plots, to provide a kinematic-wave classification of the traffic flow patterns occurring in the vicinity of the incident at 22 minutes that is recorded in the metadata quoted in the preceding section.

Time-space-speed Plots

Figure 4 shows a (lane-aggregated) time-space-speed plot for the section and times surrounding the incident at 1320 seconds. Time (in seconds) is plotted horizontally and longitudinal position (in feet) is plotted along the vertical axis. Speed (of individual vehicles) is represented on the same red (stopped)-yellow(45 fps)-green(90 fps)-blue(120 fps) color scale employed for the KWM simulation of Figure 2, to facilitate visual comparisons between the two

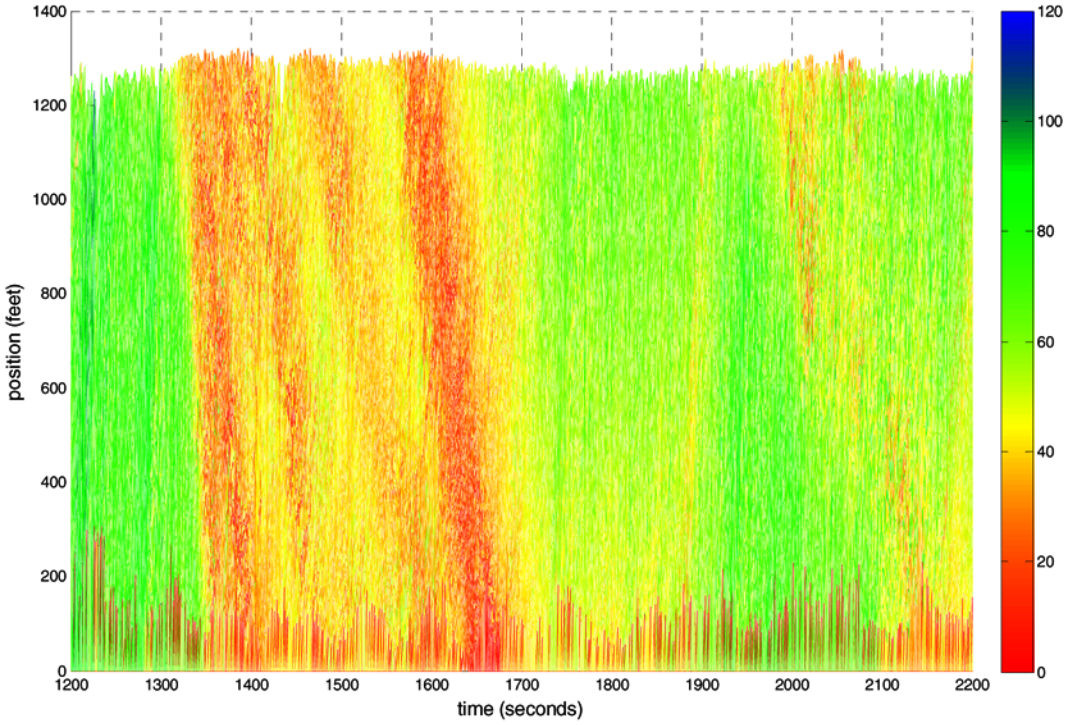

Figure 4 - Time-space-speed plot of the flow pattern, aggregated over all lanes, surrounding the incident at ~ 22 minutes. See text for further discussion.

figures. The initial free flow is more-or-less clearly visible, as the green(ish) region at the far left; however, there are some embedded yellow regions that presumably reflect variations due to either preferences of individual drivers or capabilities of individual vehicles, along with a couple of "streaks" of blue that presumably represent individual vehicles driving well above the nominal free-flow speed of 90 fps. The shock formed upstream of the stalled vehicle presumably appears as the narrow (width ≈ 50 ft) yellow region just to the right of that, which enters the observed section at its downstream boundary at approximately 1320 seconds. The reddish region immediately subsequent to the shock presumably is the congested flow downstream of the shock (and upstream of the stalled vehicle). There clearly is some oscillatory substructure within this region of congested flow, which presumably represents stop-and-go oscillations, perhaps from lane-changing.

The strongest (and last) of the stop waves (red bands) associated with the region of congested flow enters at about 1600 seconds, exits slightly before 1700 seconds, and is followed by a yellow-green region presumably corresponding to the region of acceleration wave relaxing to free flow. However, it is – as anticipated from the simulation of the preceding section – difficult to distinguish between the acceleration wave and the subsequent recovery to free flow. On the one hand it is not at all clear that the downstream (later) boundary of the congested-flow region is propagating faster than upstream (earlier) boundary, so that it is difficult to predict from the data when the former will overtake the latter, thereby generating the forward shock that is the boundary between the acceleration wave and the region of recovered free flow (cf. Figure 3). (In this regard the observed region of congested flow appears to have the distinguishing characteristic of one of the "wide jams" identified by Kerner (6).) On the other hand there is a narrow yellow region entering the upstream boundary at approximately 1880 seconds, and propagating at approximately free-flow speed, that is a candidate to be this forward shock. This putative forward shock is followed (temporally) by a region that is predominantly green (free-flow speed), and even with an occasional blue streaker, although punctuated by a diffuse yellow region entering from downstream at approximately 2000 seconds, and propagating upstream; no potential cause for this (weak shock?) is mentioned in the metadata.

These identifications of spatiotemporal regions observed in the time-space-speed plots provide a plausible mapping of the observed regions to kinematic-wave classes of flow. However, these identifications also hinge closely upon the specific KWM simulation of the preceding section, so that it is difficult to contend that they are either truly objective or entirely empirical. For example, a similar KWM simulation with the Dick-Greenberg TSM (not shown) suggests that the entire region of Figure 4 lying to the right of the region of congested flow comprises part of the acceleration wave, and it is difficult to argue that is inconsistent with Figure 4. Similarly, Figure 4 provides no solid basis for separating that portion of the shock wave consisting of vehicles going slowly but still decelerating from that portion consisting of vehicles going slowly under negligible deceleration. For these reasons we explore below (Section VII) one approach to the consideration of acceleration as a characterizing "feature," on more-or-less the same footing as speed has been employed in this subsection. However, first we conclude this section with consideration of yet another classical tool for analyzing traffic patterns, namely trajectory plots of individual vehicles, and then (Section V) introduce the basic elements of pattern recognition.

Trajectory Plots

Figure 5 shows a plot of vehicle trajectories, for vehicles in Lane 3 of the five-lane study section during any time between 1200 and 2000 seconds. One can attempt to identify visually the four classes of flow, with free flow and congested flow corresponding to regions in which the trajectories have negligible curvature and respectively large and small slopes, and shock waves and acceleration waves corresponding respectively to regions in which the trajectories collectively have significant negative and positive curvature. One of the deficiencies of this process is that "negligible" and "significant" are matters of judgment and depend on the scales used for plotting. With due recognition of these concerns, here is what we see in these data, based on the criteria describe.

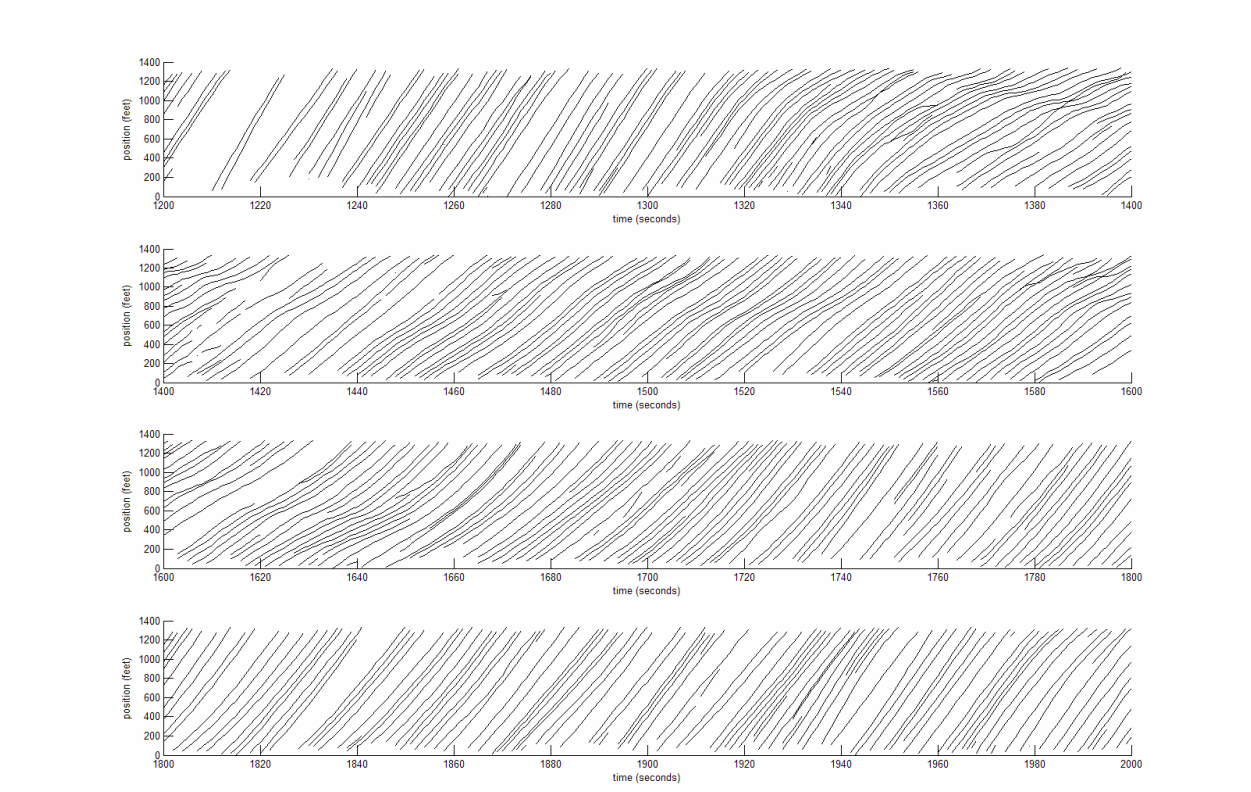

Figure 5 - Vehicle trajectories, in Lane 3, for times between 1200 and 2000 seconds into the measurements.

Prior to about 1300 seconds there is free flow (speed ≈ 90 fps). Beginning between 1300 and 1320 seconds a shock wave enters the section from upstream. This initial "stop wave" appears, as measured by the region of discernible deceleration for individual vehicles, to be approximately 200 feet wide. It seems to propagate through the observed region in about one minute, corresponding to a propagation speed of approximately 20 fps. It is followed by a region, extending approximately from 1360 seconds, that arguably corresponds to congested flow (speeds ≈ 50 fps), albeit with discernible oscillatory substructure. The last of the "stop waves" (small-scale shock waves) embedded in this congested flow exits at approximately 1670 seconds, and is followed by an acceleration wave that seems to be gradually relaxing toward a region of somewhat restricted free flow (speeds ≈ 70 fps). However, there is no evidence of a clear boundary between the acceleration wave and the recovered free flow.

Space does not permit insertion here of the trajectories for additional lanes, but they display roughly the same structure; however, there is one noticeable trend laterally across the lanes. As one moves from Lane 1 to Lane 5 (i.e., further away from the lane in which the stalled vehicle occurs at approximately 1320 seconds), the oscillatory substructure embedded in the congested flow beginning at approximately 1360 seconds becomes successively *more* pronounced.

Although such visual analyses of vehicle trajectories are the currently generally accepted method for establishing ground truth, the deficiencies of this approach that were previously described lead us to seek alternative and more algorithmic classification procedures. Given that our objective is to identify macroscopic patterns, the preceding discussion raises another challenge: Can one find an algorithm that will filter out small-scale substructure, such as the oscillations in the congested flow mentioned in the preceding paragraph. (Alternately, it would be interesting to explore possible connections between these oscillations and the various sorts of fine structure within congested flow that has been described by Kerner and co-workers, beginning with Kerner and Rehborn (16) and extending through Kerner (6). However, it is not clear that this previously reported fine structure has yet been sufficiently well characterized to support such an exploration.)

V. THE ELEMENTS OF PATTERN RECOGNITION

The discussion within this section of the generalities of pattern recognition owes much to the excellent treatment of Young and Calvert (17). Figure 6 illustrates our view, as influenced by Young and Calvert (17), of the essential components of any pattern recognition algorithm. This view best fits the classical statistical (decision-theoretic) approach to pattern recognition. In neural pattern recognition, which employs "the neural computing paradigm that has emerged with neural networks" (18), these elements tend to merge.

In the following subsections we discuss these components individually, first via generalities of that element, including a detailed description of the implementation of that component in the present work. This then is followed by a brief discussion of general classes of approaches to that element that have been adopted within the broad field of pattern recognition, typically beginning with the most naïve (manual) approach and progressing to the most sophisticated (automated) approach. We further attempt to flavor this discussion with some consideration of how pattern recognition might in the future be applied to traffic flow, with a higher degree of automation. A major motivation for this is to encourage others to seek similar algorithmic descriptions of their respective processes for macroscopic classification of traffic flow, toward obtaining some ability to make meaningful comparisons of proposed ontologies and taxonomies of traffic flow.

Measurement Space

Application of pattern recognition begins with selection of the particular *measurement space*, which is to say the attributes to be measured. If the measurements are to be generated *ab initio*, then one must make explicit decisions regarding these attributes. These decisions tend to be based on practical considerations, such as ease of measurement and perceived utility of measurements of particular attributes. Otherwise the selection of measurement space tends to be implicit within the choice of dataset (instance of measurements) to be employed. Again, practical considerations and nonautomated procedures tend to dominate the choice of dataset. ("The *selection of measurements* is based on our prior knowledge or experience on the particular pattern recognition problem" (17).) The opening two paragraphs of Section II illustrate the type of application of prior knowledge and experience that is typically involved in selecting a measurement space; see also the last sentence of the following paragraph.

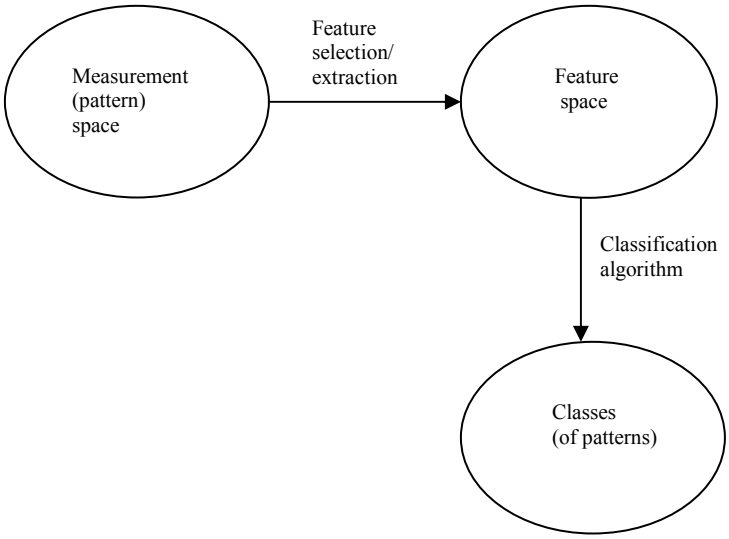

Figure 6 - The elements of a pattern recognition algorithm.

Our choice of dataset means implies our measurement space consists of the nine attributes ("fields") listed in Section II. Others involved in similar considerations have selected a different type of dataset (loop-detector), with a different measurement space; for example, attributes in a typical loop-detector measurement space might consist of flow, mean speed, and occupancy, at several different locations, and aggregated over some period of time. The use of such measurement spaces has led to considerable discussion in the literature (e.g., 19,20), primarily focused about existence of TSMs (e.g., bivariate density-flow relationships). Our choice of a dataset based on microscopic photographic measurements is partially motivated by the idea that a careful consideration of such data might further illuminate these controversies, and thereby aid in their resolution.

Feature Space

Feature space also consists of a set of attributes, albeit typically a smaller set of attributes than measurement space. Feature space comprises those attributes that will be employed to assign patterns to classes. In our case a pattern consists of a given vehicle at a given time, and the attributes of feature space are the speed and acceleration of that vehicle at that time.

This example illustrates two general points: First, feature attributes need not be measurement attributes, although they must be (at least approximately) computable from measurement attributes. For example, the two attributes of feature space in our application are speed and acceleration, but acceleration itself is not a measurement attribute. However, it is (approximately) computable from measured attributes, for example as either the first (e.g., forward) difference of two temporally consecutive speeds for the vehicle, or the second (e.g., centered) difference of three temporally consecutive positions for the vehicle.

The second general point is that not all measured attributes will contribute to values of some feature attribute. In our instance only the frame number, vehicle id, speed and distance from beginning of section to front of vehicle are the measured attributes employed to recognize the (four classical) classes for macroscopic traffic flow patterns. However, all of these plus the measurement attributes frame number, vehicle id, and lane number (if differentiation by lane is intended) may in principal be employed for the visual identification of the four classical macroscopic traffic flow classes.

Feature Selection

The choice of the attributes to be included in feature space is a process known as *feature selection*. This is an important step, as suggested by Fu *et al.* (21): "One important problem in pattern recognition is the selection of effective features from a given set of feature measurements. It is known that the performance of a pattern recognition system is closely related to the feature measurements taken by the classifier." Feature selection is often done heuristically, as is the case for our choice of speed and algorithm as the features for our attempt at sorting microscopic flow patterns into the four classical classes.

More algorithmic methods have been suggested; e.g., the "Relief method" of Huang *et al.* (22). In recent years much of the work on feature selection has been subsumed within the literature on neural networks. In fact much of the attraction of neural networks, at least for pattern recognition, lies precisely in their ability to subsume simultaneously, and more-or-less automatically, the functionality of feature selection, feature extraction and classifier (classification algorithm). On the one hand there is need for automated feature selection, simply because heuristic choices based on experience fail at some level of complexity. For an instance reasonably close at hand, candidate features for the problem of classifying flow patterns into the four classical macroscopic classes of free flow, congested flow, acceleration waves or shock waves include at least speed, density, flow, occupancy, and any of their first or second spatial or temporal differences; further, the temporal differences conceivably could be determined either at a fixed location or following a fixed vehicle. This is a total of 28 possible features, or 28*27= 756 possible two-dimensional feature spaces; a heuristic selection of feature attributes from among this number of different possibilities is a somewhat daunting task.

On the other hand automated feature selection has sometimes the disadvantage of loss of intuitive interpretation as to the underlying nature or cause of the different features, and perhaps the corresponding classes. However, often suitable interpretations can be identified, once the classification is completed.

Feature Extraction

We consider *feature extraction* as the process of extracting, from the values of the measurement attributes contained in the target dataset, the values of the feature attributes that have been identified for feature space in the preceding feature selection step. Feature extraction tends to be a process that is tailored to the particular measurement space and feature space, because the relationship between these different sets of attributes is highly dependent upon their individual nature.

If one limits measurement space to those attributes commonly measured for microscopic traffic-flow datasets, and feature space to those microscopic attributes commonly considered to be somehow closely related to macroscopic attributes (e.g., as vehicle spacing is to density), then there appears to be some opening for development of a general approach and associated conceptual tools. This idea is pursued further in Section VI below, where we describe a set of tools that were developed, in the course of the present work, for extraction of features of possible macroscopic interest from typical microscopic datasets.

Classification Algorithms

A *classification algorithm* decomposes the feature space into nonoverlapping sets whose union is the feature space. A collection of such sets is said to *partition* the feature space. A *classifier* is an implementation of a classification algorithm.

For our microscopic pattern recognition problem the partition of interest is intuitively described in the opening paragraph of the Introduction. In the first subsection of Section VII below we employ a highly supervised version of *k*-means clustering to obtain a partition that respects the spirit of this intuitive description, although it differs somewhat in detail. Our fundamental macroscopic pattern recognition problem is to determine ground truth regarding existence of some corresponding spatiotemporal macroscopic partition. That question is explored, based upon spatiotemporal visualization with preprocessing via the microscopic *k*-means clustering, in the third subsection of Section VII.

It would be of interest to develop more formal and highly automated pattern recognition techniques for direct application to the problem of identifying spatiotemporal regions corresponding to different macroscopic classes of flow (e.g., the four classical macroscopic classes). This appears to be a challenging problem, for at least two reasons. First is the necessity to select the feature space from the relatively large number of possibilities that initially present; see the preceding subsection on "Feature Selection." Second, the boundaries between adjacent spatiotemporal regions corresponding to different macroscopic flow classes will be both inherently somewhat imprecise and subject to complex forms of motion. In many ways this makes our macroscopic problem even more complex than the problem of recognition from spatiotemporal data of moving rigid objects, which is itself a subject of substantial current research interest (23).

VI. MICROSCOPIC DATASETS AND FEATURE EXTRACTION OPERATORS

Microscopic Datasets

We conceive a microscopic traffic-flow dataset to be organized as follows:

Metadata = <number_of_patterns, time_between_frames, units_for_time,
number_of_frames, length_of_section, units_for_lengths, number_of_lanes,
graphical_description_of_measured_section, textual_description_of_measured_section,
textual_description_of_external_circumstances_during_measurements>,

where a "pattern" refers to a unique <time, vehicle> pair, and

$$\text{Data}=A(5,1:\text{number\_of\_patterns}).$$

Here the columns of the traffic-flow-data matrix A contain measurements, as follows:

$A(1,i)$ = frame_index _for_ith _pattern <$1:$number_of_frames>,
$A(2,i)$= vehicle_id_for_$i$th_pattern <as_assigned_in_source_dataset>,
$A(3,i)$= length_of_vehicle_for_$i$th_pattern <continuous ≥ 0>,
$A(4,i)$= longitudinal_location_of_vehicle_for_$i$th_pattern <continuous $\geq 0, \leq$ length_of_section>,
$A(5,i)$= lane_location_for_$i$th_pattern <1:number_of_lanes>.

For example, values of the metadata parameters for the measurements (dataset) of Section II are as follows:

number_of_patterns=181142,
time_between_frames=1,
units_for_time='seconds',
number_of_frames=3600,
length_of_section=1341 feet (408.5 m),
units_for_lengths='feet',
number_of_lanes=5,
graphical_description_of_measured_section=as in Fig. 29 of Ref. (10),
textual_description_of_measured_section=
as in first paragraph of quotation in Section II,
textual_description_of_external_circumstances_during_measurements=
as in second paragraph of quotation in Section II.

We term the column index of A corresponding to a particular pattern as the *pattern_index* of that pattern. A principal use we make of the matrix data structure is the ability it provides to refer to a particular pattern by its pattern_index. For purposes of implementing the feature-extraction operators described below and carrying out the experiments described in the following section we reorganized the dataset described in Section II above as a MatLab matrix having the structure of a traffic-flow-data matrix. This permits more efficient access to individual patterns (e.g., those previously selected as members of a training sample selected randomly for initial identification of the means) than the alternative of sequentially searching through the original ASCI file for those patterns corresponding to unique <vehicle_id, frame_index> pairs. Of course this reorganization itself is not without computational cost, and this factor played a role in our choice of one of the JHK datasets, as opposed to other alternatives (see Section II).

Further efficiencies, especially in implementation of the feature-extraction operators, can be achieved by appropriate detailed organization of the traffic-flow-data matrix (A). We begin this discussion by noting that among the numerical data parameters that somehow reflect the size of the data, the meta-attribute number_of_patterns clearly will have the largest value. Thus one wishes to minimize the number of searches over all patterns. The parameter number_of_patterns can be expressed in either of the ways

$$\begin{aligned}\text{number\_of\_patterns} &= \\ &\sum_{\text{frame\_indices}} \text{no. of vehicles in frame} = \text{number\_of\_frames}*\overline{\text{vehicles per frame}} = \\ &\sum_{\text{vehicle\_indices}} \text{no. of frames vehicle in section} = \\ &(\text{number of vehicles})*\overline{\text{frames per vehicle}},\end{aligned} \tag{1}$$

where the overbars indicate mean values. For a quarter-mile (.40 km) four-lane section and data acquisition over one hour at a rate of one frame per second, typical values of the four parameters on the right-hand sides of Eq. (1) are

$$\begin{aligned}&\text{number\_of\_frames}=3600,\\ &\overline{\text{vehicles per frame}} = 40,\\ &\text{number of vehicles}=4000,\\ &\overline{\text{frames per vehicle}} = 20.\end{aligned}$$

These suggest the advisability of organizing the data so as to minimize the necessity of searches over all frames, or over all vehicles, that might be required. Although the present work employs only velocity and acceleration as features, and their extraction involves only searches for the same vehicle_id locally in time (i.e., at adjacent values of the frame_index), other feature-extraction operations (e.g., spacing between a vehicle and its leader) would involve a search for other vehicle_indices, but still would be local in time. For this reason we recommend, at least for purposes of optimizing extraction of microscopic analogs of macroscopic attributes, organizing the traffic-flow-data matrix (A) so that data for patterns contiguous in frame_index are contiguous in storage (pattern_index). Note that NGSIM (11) currently employs the alternate strategy of organizing the data so that patterns corresponding to the same vehicle are contiguous in storage. Some of the functionality of our operator approach is provided by fields for preceding and following vehicles at each frame; however, we believe the operator approach adopted here is more flexible, in light of the large number of possible attributes. (See the "Feature Selection" subsection of Section V.)

Feature-extraction Operators

Table I shows the feature-extraction operators that we have implemented in MatLab, as described by their names, required input and expected output. A pattern is identified either by its pattern_index, or by the unique corresponding pair < vehicle_id, frame_index>, as convenient. Rows of the traffic-flow-data matrix A are referred to by their associated attribute name, rather than row index. The four "Level 1" operators are the basic operators. Note that if patterns having contiguous values of frame_index are maintained in contiguous columns, as recommended above, then implementation of these requires only a search through a (typically not large) set of contiguous columns (pattern indices) in the traffic-flow-data matrix.

Table I. Description of operators, with name, level, required input and normal output

Operator Name	Level	Input	Output
NEXT	1	*j* = valid value for a pattern_index	pattern_index(vehicle_id(*j*), time_index(*j*)+1)
PREVIOUS	1	*j* = valid value for a pattern_index	pattern_index(vehicle_id(*j*), time_index(*j*)-1)
LEADER	1	*j* = valid value for a pattern_index	The vehicle_id value for the vehicle immediately downstream of vehicle_id(*j*) in frame frame_index(*j*)
FOLLOWER	1	*j* = valid value for a pattern_index	The vehicle_id value for the vehicle immediately upstream of vehicle_id(*j*) in frame frame_index(*j*)
VELOCITY	2	*j* = valid value for a pattern_index	The three-point centered approximation to the speed of vehicle_id(*j*) at the time frame_index(*j*).
ACCELERATION	2	*j* = valid value for a pattern_index	The three-point centered approximation to the acceleration of vehicle_id(*j*) at the time frame_index(*j*).

.

Level 2 operators are those constructed from Level 1 operators. For example the three-point centered-difference approximation to the speed of vehicle_id(*j*) at time frame_index(*j*) is

$$\text{VELOCITY}(j)=\frac{\text{location}(\text{NEXT}(j))-2*\text{location}(j)+\text{location}(\text{PREVIOUS}(j))}{\text{time\_between\_frames}},$$

where "location" is an abbreviation for the attribute longitudinal_location_of_vehicle. We employed this three-point centered-difference approximation to speeds in the pattern recognition experiments of the following section, for the following reason. We must use at least three points (patterns) to approximate the accelerations that also are an attribute of our selected feature space. The three-point centered-difference approximation is the natural such approximation. But if this is to be used for accelerations, then it is more natural to use the counterpart three-point approximation to the speeds, as opposed (say) to the two-point forward-difference approximation that is included in the source dataset we employ.

The feature-extraction operators that were employed directly for the present work were VELOCITY and ACCELERATION. When directed toward a traffic-flow-data matrix, with input pattern_index *j*, these return respectively the three-point centered-difference approximations to the speed and acceleration of the corresponding vehicle at the corresponding time. These particular operators are "Level 2" feature extraction operators, which utilize more basic (Level 1) operators such as NEXT, PREVIOUS, LEADER and FOLLOWER. The functionality of these operators is essentially suggested by their names; for example, NEXT(*j*) returns the pattern-index corresponding to the vehicle associated with pattern_index *j*, but for the frame following the frame associated with that value of pattern_index. Unfortunately space limitations preclude a more detailed discussion of feature extraction operators. (The "level" terminology is adapted from the taxonomy employed for the Basic Linear Algebra Subprograms (24).

We now turn to the application of the feature-extraction operators within the illustrative example of employing microscopic pattern recognition to facilitate visual recognition of spatiotemporal regions comprising instances of the classical macroscopic flow classes.

VII. VISUALIZATION VIA FOUR-MEANS CLUSTERING IN SPEED-ACCELERATION FEATURE SPACE

In the first subsection ("Four-means Clustering") of this section we describe a highly supervised *k*-means (*k*=4) algorithm, and its application to sorting patterns in speed/acceleration feature space into clusters that are candidates to be the four kinematic-wave classes of traffic flow. Silhouette values, and associated silhouette plots, are a commonly used tool for evaluating the quality of clusters produced by a *k*-means algorithm. In the second subsection ("Silhouette Plots") we present and discuss the silhouette plots corresponding to the four clusters produced by our application. The discussion especially focuses upon illumination provided by the silhouette plots in regard to the three significant difficulties already identified in Sections III and IV in regard to identifying kinematic-wave spatiotemporal regions, namely:

1. Difficulty in separating (relatively narrow) shock waves from both the preceding free flow and the subsequent congested flow;
2. difficulty in separating the acceleration wave succeeding the congested region from the subsequent recovered free flow; and
3. difficulty in filtering out the small-scale oscillatory structure within presumed regions of congested flow.

The final subsection treats the use of the four clusters to promote visual identification of spatiotemporal regions corresponding to the four kinematic-wave classes. Again the focus of the discussion is upon the three previously anticipated difficulties listed just above.

The Training Phase

The dataset described in Section II was reorganized as a matrix, as described in the preceding section. A MatLab® program (*km.m*) was written to perform the following functions on this traffic-flow-data matrix:

1. Extract a user-specified number of randomly selected patterns (i.e. vehicle-time combinations) from the dataset.
2. Employ the feature extraction modules described in Section V to extract values of velocity and acceleration corresponding to each of these randomly selected patterns.
3. Use the MatLab program "kmeans" to perform a *k*-means clustering analysis (25) on these random patterns in velocity-acceleration (feature) space, with a user-specified value of *k* (= number of clusters) and user-specified starting values for the cluster centroids. The distance function employed for this analysis was the weighted Euclidean norm

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\left(\frac{v_i - v_j}{v_{\text{norm}}}\right)^2 + \left(\frac{a_i - a_j}{a_{\text{norm}}}\right)^2} \tag{2}$$

where v_i and a_i are respectively the speed and acceleration associated with the feature vector $\mathbf{x}_i$, and the normalization constants v_{norm} and a_{norm} are user-specified.

As a training phase (i.e., to identify cluster centroids typical of the four individual kinematic-wave classes of flow), the program *km* was employed, with *k*=4 clusters, in a *training phase,* as follows. A number of collections of randomly selected datasets, each consisting of about 1000 points (patterns), were obtained per Step 1 above of the program *km*. The velocities and accelerations associated with each of these patterns were then extracted per Step 2. (Typically a few of the patterns randomly selected for each training phase we unusable, because of duplications in the random selection and a variety of data errors.) Each of the resulting cleansed velocity/acceleration datasets was then subject to a *k*-means analysis, per step 3 above, with starting values for the cluster centroids selected to represent intuitive expectations of those values for respectively congested flow, shock waves, acceleration waves and free flow, and with values of the normalization constants selected to approximately equalize the contributions from velocity and acceleration at values of speed and acceleration near the maximum expected to normally be encountered in the subject dataset.

For the prototypical instance that the remaining discussions, in this and the following subsection, are based on, the randomly selected collection of data consisted of 998 patterns (1000 minus two duplications in the randomly

® Trademark of The MathWorks, Inc.

selected training set), the normalization constants were $v_{\text{norm}} = 90$ ft/sec and $a_{\text{norm}} = 45$ ft/sec^2, and the starting values for the cluster centroids were

$$(v_{cf}, a_{cf}) = (10,0), (v_{sw}, a_{sw}) = (50,-15), (v_{aw}, a_{aw}) = (50,10) \text{ and } (v_{ff}, a_{ff}) = (90,0).$$

(Here and below the units for speeds and accelerations are respectively ft/sec and ft/sec^2. See the latter part of the following subsection for discussion of the choice of normalization constants.) The values returned for the cluster centroids were

$$(v_{cf}, a_{cf}) = (8.0, -0.3), (v_{sw}, a_{sw}) = (51.6,-4.9), (v_{aw}, a_{aw}) = (45.4,4.6) \text{ and } (v_{ff}, a_{ff}) = (76.4,0.95).$$

The numbers of points in the respective clusters were

$$\left(n_{cf}, n_{sw}, n_{aw}, n_{ff}\right) = \left(141, 250, 275, 332\right).$$

The numbers of points in the free-flow cluster were smaller than expected, and those for the other clusters were larger than expected; these tendencies were seen consistently, and were even further aggravated with many other values of the normalization constants and starting centroid values.

It is appropriate to indicate briefly how the indicated values of the normalization constants were determined. The value $v_{\text{norm}} = 90$ ft/sec was selected in order to give a maximum distance contribution from speeds on the order of unity. With this choice fixed, a considerable amount of experimentation was carried out with various values of anorm. Values on the order of the indicated 45 ft/sec^2 seemed to give the best results, as judged via the silhouette values to be discussed in the following subsection. We believe that is because values of a_{norm} in this range effectively assign patterns having small or large speeds to respectively the congested-flow or free-flow clusters based primarily upon the associated value of speed, whereas in the intermediate range of speeds they effectively discriminate between the shock-wave and acceleration-wave clusters based largely upon the associated value of acceleration.

Silhouette Plots

So-called "silhouette values" provide one useful means of judging the quality of a particular clustering of patterns (26). The silhouette value of a given pattern, relative to a specified clustering, is roughly a measure of how much less the mean distance of that pattern to other patterns in its cluster is than the corresponding mean distance to other patterns in its second-nearest cluster, by which is intended the cluster to which it has the next smallest mean distance. Here the relevant notion of distance between patterns is the distance function (2) in speed/acceleration attribute space, and for the example discussed the normalization constants were taken as described in the preceding subsection. Silhouette values can be negative, which means that the associated pattern is in fact, on average, closer to the patterns in its "second-nearest" cluster than to the other patterns in the cluster to which it is assigned. In such cases the term "second nearest" is technically correct, but if the clustering algorithm is well suited to the classification problem at hand then such instances will be rare.

For the particular instance of the training phase described in the preceding section the mean silhouette value, over all patterns in the randomly selected training dataset, was 0.56. This means that the average pattern was, in feature space and under the distance function (2), on average about 56% further from the patterns in its second-nearest cluster than to the patterns in its assigned cluster. This is perhaps the most meaningful single figure of merit for a clustering of patterns. Values for the normalization constants in the vicinity of those indicated in the preceding paragraph consistently gave the largest value for this mean silhouette value, from an extensive search over a reasonable range of values for these constants.

A more detailed representation of the quality of this clustering, in terms of silhouette values, is provided by the "silhouette diagram" of Figure 7. Such a diagram is constructed as follows. Each pattern assigned to a particular cluster corresponds to a horizontal line in the silhouette corresponding to that cluster. The length of that line is proportional to the associated silhouette value, with negative silhouette values corresponding to lines extending to the left from the point designating silhouette value zero. The lines are plotted in order of descending silhouette value, from top to bottom. Thus the height of the silhouette corresponding to a particular cluster is proportional to the number of patterns comprising that cluster. This is consistent with the numbers in each of the clusters, as delineated in the preceding subsection; i.e., approximately equal numbers of patterns in each of the shock-wave and acceleration-wave clusters, with about half as many in the congested-flow cluster and 25% more in the free-flow cluster.

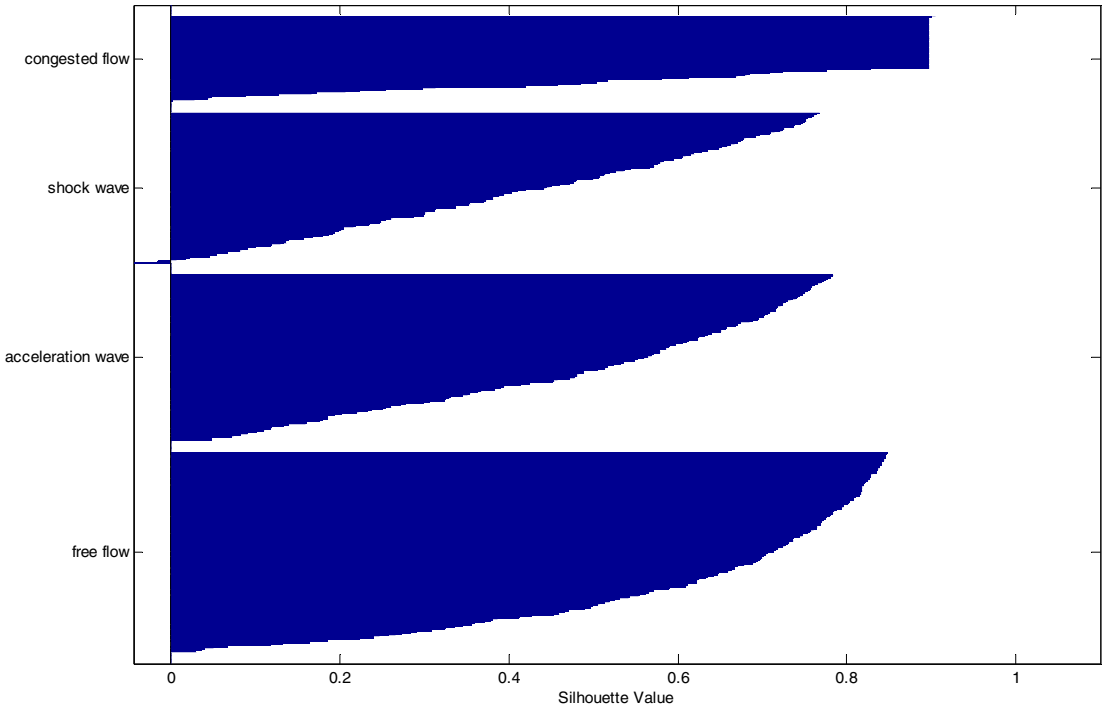

Figure 7 - Silhouette diagram for the typical 4-clustering described in the text.

Some discussion of the silhouettes for each of the individual clusters is also worthwhile, especially in regard to the anticipated difficulties listed in the opening paragraph of this section. It is convenient to begin with the free-flow cluster. The maximum silhouette value within this cluster is approximately 0.85, the median value is (graphically) about 0.75, the 25% percentile value is around 0.55, and the minimum value is around 0.05, so there are no negative silhouette values. Thus the free-flow cluster seems well separated, notwithstanding the anticipated difficulty in distinguishing between acceleration waves and the recovered free flow.

The corresponding values for the acceleration-wave cluster are ~ 0.79 for maximum silhouette value, ~0.57 for the median value, ~0.35 for the 25th percentile, and again no negative values, with the smallest value again around 0.05. Thus the acceleration-wave cluster also is well separated, although perhaps slightly less so than the free-flow cluster.

The silhouette plots for the free-flow and acceleration-wave clusters thus give little indication of the second anticipated difficulty of distinguishing between the acceleration wave and the subsequent recovered free flow. On the other hand the shock-wave cluster is arguably the least well separated of the four clusters, with values of ~0.75 for maximum silhouette value, ~0.5 for the median value, ~0.25 for the 25th percentile value, and ~-0.08 as a smallest value, with indication of approximately three patterns that are likely misclassified (negative silhouette values). This is consistent with the first of the anticipated difficulties listed above. It is also consistent with the hypothesis that the idiosyncrasies of individual drivers, or even the variations in behavior of individual drivers across different times, would tend to show most within spatiotemporal regions of high deceleration.

The congested-flow cluster has a somewhat qualitatively different nature, owing to the large number (~ 75%) of patterns having silhouette value ~0.9 (more precisely, 0.8975). These patterns correspond to stopped or nearly stopped vehicles. Although these patterns clearly belong to the congested-flow cluster, there also are several patterns assigned to the congested-flow cluster that have relatively small silhouette values; other instances (random

selections of the training set) even display occasional negative silhouette values. This is supportive of the hypothesized difficulty in separating congested flow from a preceding shock wave.

We turn now to the use of these 4-cluster training results to visually identify corresponding macroscopic (spatiotemporal) regions.

Application as a Preprocessor to Visualization

Figure 8 shows a spatiotemporal color plot of the cluster assignments, each pattern to the centroid determined by the training-set instance of the preceding subsection, for the patterns occurring in the earlier spatiotemporal plot of speeds (Figure 4). Now congested flow, shock waves, acceleration waves and free flow correspond nominally to respectively red, yellow, blue and green regions, as shown on the color bar at right. The cluster assignments for the individual patterns were as determined by the nearest centroid, under the distance function (2), with centroids determined by the specific instance of the training phase that was the basis of the discussion of the two preceding subsections. Again the initial free flow is more-or-less clearly visible, as the predominantly green region at the left. This region is punctuated by occasional small yellow, and now blue, bands, which presumably represent isolated instances of acceleration and deceleration. The shock wave emanating from the incident at approximately 2 minutes presumably is the narrow yellow band entering the observed section at approximately 1300 seconds. This band is approximately 300 feet (100 meters) wide, as seen from corresponding spatially expanded plots (not shown). It propagates upstream at about 20 feet per second (25 km per second), which is more-or-less consistent with the 20 km per hour (upstream) wave speeds in congested flow that have been described in the literature (27).

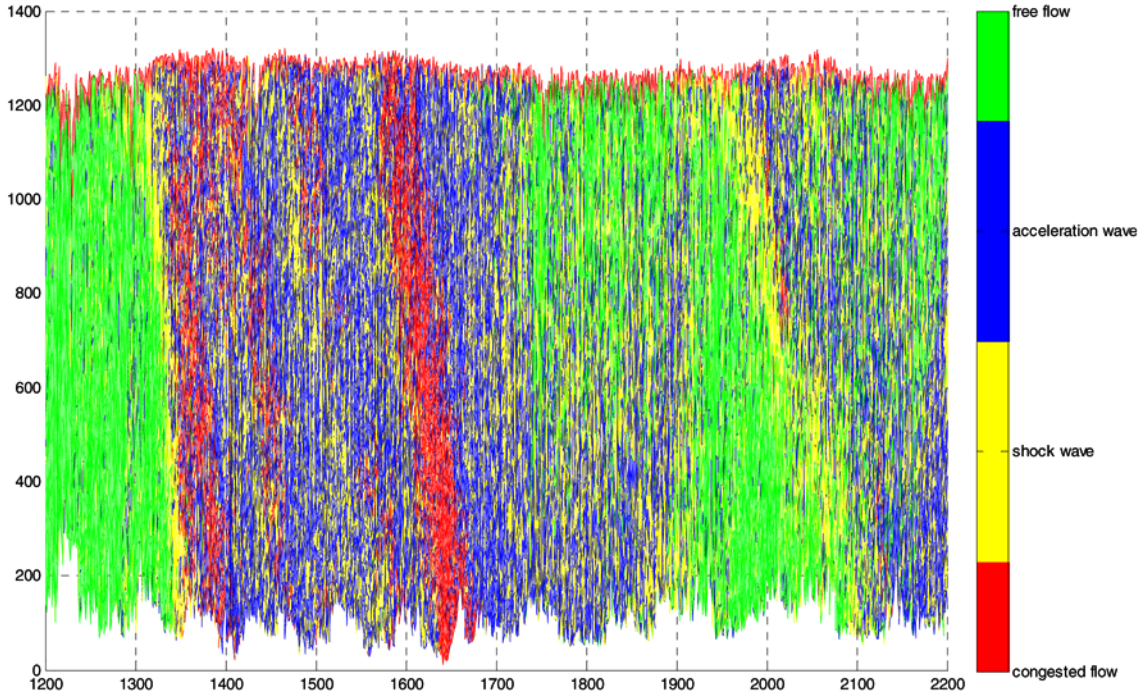

Figure 8 - Spatiotemporal cluster plot of the flow pattern, aggregated over all lanes, surrounding the incident at ~ 22 minutes. See text for further discussion.

The approximately 300 foot width of the putative shock wave is somewhat narrower than the 1 km "transition zone" widths described by Muñoz and Daganzo (15). This should not be taken too seriously, in view of the anticipated difficulty in clearly distinguishing between a shock wave and both the preceding free flow and the succeeding congested flow. The perhaps more surprising point, in view of the anticipated difficulty in

distinguishing shock waves from the preceding free flow and the succeeding congested flow, is the fact that nonetheless the initial shock wave is clearly visible. Although this remains to be verified, the shock presumably could be made wider, and therefore even more visible, by increasing the sensitivity to acceleration (i.e., decreasing a_{norm}) in the distance function (2). However, this seems to be counterproductive, in terms of the next issue to be discussed, as we proceed to discuss the time sequence seen in Figure 8.

To the right of this shock there is a region of more-or-less alternating yellow and blue bands (respectively shock and acceleration waves), bounded on either side by red bands (congested flow). This region presumably corresponds to the region of oscillatory congested flow that was observed in the preceding analysis of the trajectory plots. The appearance of the alternating yellow and (predominantly) blue bands (stop-and-go waves?) appears to signal failure in our objective (cf. last paragraph of Section IV) of filtering out such small-scale substructure. Somewhat surprisingly, decreasing the weight given to acceleration in the distance function employed in the *k*-means algorithm (specifically, increasing a_{norm} to 90 ft/sec^2) considerably resolves this issue (results not shown here), and results in somewhat higher quality clustering (mean silhouette value = 0.64, but with several negative silhouette values within the shock-wave cluster and one among the free-flow cluster), without noticeably increasing the anticipated difficulties in identifying the shock or distinguishing the acceleration wave from the following free flow. We have not followed up on that observation here, because it essentially amounts to a four-regime clustering, with vehicle speed as the only attribute. Perhaps a better approach, while still including acceleration as a significant attribute, would be to increase the number of clusters, with the expectation of distinguishing clearly between stop-start waves (deceleration/ acceleration regions at slow speeds), shock waves and accelerations waves (regions of respectively deceleration or acceleration, but now at predominantly intermediate speeds).

Following this oscillatory congested flow there is a band of predominantly blue (acceleration wave) that trends toward green (free flow) as time increases. This presumably is some mixture of the anticipated acceleration wave and recovery free flow. It is, just as anticipated, difficult to distinguish these two classes of flow within this region, although the expected time sequence (first acceleration wave, then free flow) is clearly discernible. Increasing the sensitivity of the underlying distance function to acceleration (specifically, decreasing a_{norm} to 22.5 ft/sec^2) seems to help very minimally with this, or with the difficulty of identifying the shock-wave region, but neither does it make noticeably worse the pronounced stop-start oscillations; it does somewhat decrease the overall quality of the clustering, as measured by a mean silhouette value of 0.52, and by significantly more patterns in both the shock-wave and acceleration-wave clusters that have associated negative silhouette values.

The trend toward (recovered?) free flow in Figure 8 ultimately is interrupted by yet another shock wave that enters from upstream at approximately 1950 seconds. This shock is just discernible in the trajectory plot (Figure 5), and also arose in our earlier discussion of the spatiotemporal speed plot of Figure 4, but the metadata are silent on any cause for it.

VIII. CONCLUSIONS

We have presented a view (the traffic-flow-data matrix of Section VI) of microscopic traffic flow datasets that rests on provision of minimal data, along with a library of feature extraction operators that are intended to provide researchers with the flexibility to extract those features of the data that they wish to study. Arguments are given to suggest feature extraction operators can most efficiently access traffic-flow-data matrices if those matrices are organized so that data from adjacent frames are adjacent in storage.

This viewpoint and accompanying set of tools was applied to attempts to visualize spatiotemporal regions corresponding to the four kinematic-wave classes of traffic flow (congestion, shock waves, acceleration waves and free flow) based upon spatiotemporal plots of speed alone. This was in some degree successful, particularly in identifying the shock wave generated by an incident; however, it was inadequate to provide empirical evidence concerning the nature of the observed oscillatory structure following this shock, and in empirically distinguishing the subsequent acceleration wave (queue discharge) from the following (recovered) free flow.

We therefore applied out tools to the task of visually identifying spatiotemporal regions corresponding to the classical kinematic-wave flow classes, through a 4-means clustering analysis of individual patterns (vehicle-frame) in speed-acceleration feature space. This approach identifies the oscillatory structure mentioned in the

preceding paragraph as small-scale shock and acceleration waves, presumably corresponding to what are often termed as stop-and-go waves. It also somewhat better separates the subsequent acceleration wave and recovered free flow, although there is room for more improvement in this regard.

The spirit of the approximations embodied in the kinematic-wave model of traffic flow seems to require identification of the nature of traffic flow over spatiotemporal regions larger than those characterizing the presumed stop-start waves. A promising approach toward this, within the spirit of the techniques explored here, would seem to be to increase the number of clusters, perhaps to six to provide for stop-start waves (regions of respectively deceleration and acceleration at slow speeds) as clusters distinct from the four kinematic-wave classes. One could then hope to achieve a kinematic-wave classification by "lumping" these two new classes with congested flow, which now would correspond to negligible acceleration at slow speeds.

REFERENCES

1. Lighthill, M. J. and G. B. Whitham. 1955. On kinematic waves II. A theory of long crowded roads. Procs. Roy. Soc. London A229 317-345.
2. Richards, P. I. 1956. Shock waves on the highway. Operations Research 4 42-51.
3. Webb, A.R. Statistical Pattern Recognition. John Wiley & Sons, Ltd, West Sussex, England, 2002.
4. *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, accessed July 2, 2005. See especially Chapter 1.
5. Kerner, B.S. and S. L. Klenov. 2003. Microscopic theory of spatial-temporal congested traffic patterns at highway bottlenecks. Physical Review E 68 036130-1 through 036130-20.
6. Kerner, B. S. *The Physics of Traffic.* Springer, Berlin, 2004.
7. Schönhof, M. and D. Helbing. Empirical features of congested traffic states and their implications for traffic modeling. Preprint, available at URL http://www.trafficforum.org/journalArticles/tf04052101.pdf, accessed May 14, 2006; see also Treiber, M., A. Hennecke and D. Helbing. 2000. Congested traffic states in empirical observations and microscopic simulations. Physical Review E 62 1805-1824; and Treiber, M., A. Hennecke and D. Helbing. 1999. Phase diagram of traffic states in the presence of inhomogenieties. Physical Review Letters 82 4360-4363.
8. Coifman, B., *et al.* Surface transportation surveillance from unmanned aerial vehicles. Procs. of the 83rd Annual Meeting of the Transportation Research Board, 2004; also Coifman, B., *et al.*. 2006. Roadway traffic monitoring from an unmanned aerial vehicle. IEE Journal on Intelligent Transport Systems 153 11-20.
9. Kim, Z. *et al.* A machine vision system for generating vehicle trajectories over extended freeway segments. Procs. 12th ITS World Congress, San Francisco, Nov. 6-10 2005.
10. U.S. DOT, "Freeway Data Collection for Studying Vehicle Interactions," Report No. FHWA/RD-85/108, JHK & Associates, May 1985.
11. "NGSIM, Home of the Next Generation SIMulation Community -> Data Sets". URL http://ngsim.fhwa.dot.gov/, accessed May 15, 2006.
12. Nelson, P. and N. Kumar. Point constriction, interface and boundary conditions for the kinematic-wave model. Procs. of the 83rd Annual Meeting of the Transportation Research Board, 2004. (To appear, in revised form, in Transportation Research Record.)
13. Newell, G.F. 1993. A simplified theory of kinematic waves in highway traffic, Part II: Queuing at freeway bottlenecks. Transportation Research B 27 289-303.
14. Muñoz, J. C. and C. Daganzo. 2002. The bottleneck mechanism of a freeway diverge. Transportation Research A 36 483-505.
15. Muñoz, J. C. and C. Daganzo. 2003. Structure of the transition zone behind freeway queues. Transportation Science 37 312-329.
16. B. S. Kerner and H. Rehborn. 1996. Experimental properties of complexity in traffic flow. Physical Review E 53 R4275-R4279.
17. Young, T.Y. and T. W. Calvert. *Classification, Estimation, and Pattern Recognition*. American Elsevier Publishing Company, Inc, New York, 1974.
18. FOLDOC, "Free On-Line Dictionary of Computing," http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?pattern+recognition, accessed May 15, 2006.
19. Drake, J. S. *et al.* 1967. A statistical analysis of speed density hypotheses. Highway Research Record 154 53-87.

20. Cassidy, M. J. 1998. Bivariate relations in nearly stationary highway traffic. Transportation Research B 32 49-59.
21. Fu, K.S., Min, P.J., and Li, T.J. Feature Selection in Pattern Recognition. *IEEE* Trans. System Science and Cybernetics SSC-6 1 (Jan. 1970), 33-39.
22. Huang, Y., McCullagh, P.J., Black, N.D. Feature selection via supervised model construction. In *Proc. Fourth IEEE International Conf. on Data Mining (ICDM'04)* (Brighton, England, Nov. 1-4, 2004).
23. Moscheni, F., S. Bhattacharjee and M. Kunt. 1998. Spatiotemporal segmentation based on region merging. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 897-915.
24. Lawson, C. L. *et al.* 1979. Basic linear algebra subprograms for FORTRAN usage. ACM Trans. Math. Soft. 308-323.
25. Weisstein, E. W. K-means clustering algorithm. *MathWorld*--A Wolfram Web Resource. http://mathworld.wolfram.com/K-MeansClusteringAlgorithm.html, accessed July 27, 2005.
26. Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Computational and Applied Mathematics 20 53-65.
27. Daganzo, C. F. 2002. A behavioral theory of multi-lane traffic flow. Part I: Long homogeneous freeway sections. Transportation Research B 36 131-158.