

Feature-based Automatic Identification of Interesting Data Segments in Group Movement Data

T. von Landesberger and S. Bremm and T. Schreck and D. W. Fellner



Fig. 1. Example of the proposed visual analysis process: (a) Display of original point movement data using trajectories. (b) Plot of features extracted from point movement data. (c) Result of feature analysis using temporal pattern search function in time view. (d) Automatically selected movement data suggested for inspection by the analyst.

Abstract— The study of movement data is an important task in a variety of domains such as transportation, biology, or finance. Often, the data objects are grouped (e.g., countries by continents, etc.). Then, the analysis focuses on three main aspects: (a) Behavior of an individual in the context of its group, (b) dynamics of a given group, and (c) comparison of the behavior of multiple groups. Analysis of group movement data can be effectively supported by data analysis and visualization. Feature-based approaches have shown useful to describe and analyze movement data. However, previous approaches were limited as they did not cover a broad range of situations and required manual feature monitoring. We extend the set of movement analysis features and add automatic analysis of the features for filtering for interesting parts in the movement data. Users can easily detect new interesting characteristics such as outliers, trends and task-dependent data patterns even in large sets of data points over long time horizons. We demonstrate usefulness of our system on a real-world data set from financial domain.

Index Terms—Spatio-temporal data, Visual Analytics, time-dependent data, movement data.

1 INTRODUCTION

Movement data arises in many important applications areas such as transportation, traffic analysis, biology, or finance. Movement data describes time-dependent 2D data that can be geo-referenced (e.g., animal movements) or abstract (e.g., stock market dynamics given in risk-return diagram space). Such data often includes relevant grouping information (animal herds, asset categories, etc.) Then, often it is required not only to examine the data elements individually, but to analyze them in context of the group they belong to, including the analysis of relationships between different groups. In financial applications, it may be interesting to analyze changes of individual stock prices w.r.t. the changes exhibited by all stocks of the respective country. Likewise, one may want to compare the dynamics of a selected country against those of other countries.

We identify three generic **analytical tasks**:

- T1 *Analyze dynamics of an individual data element in context of its corresponding group over time* such as relative position of an object in a group (center or boundary)).
- T2 *Analyze dynamics of a group of data elements*, its direction, speed, or shape and area spanned by the group, etc..

- T3 *Compare the movement of multiple groups* such as correlation of group movements or overlap of groups.

To date, movement data has been analyzed using Visual Analytics approaches including statistical analysis of the data distribution in combination with specialized visual mappings. However, these approaches mainly concentrate on analysis of individual object movements and often require manual monitoring of the data.

We present an encompassing approach for analysis of movement data with grouping information. In particular our contributions comprise:

- We significantly extend a set of available base features for identification of interesting data views. Additional new, meaningful features covering a wide variety of use cases and explicitly considering grouping structure are defined.
- We introduce automatic identification of interesting patterns in the monitored features. It enables faster analysis of the monitored data also in case of long time horizons and many features to be analyzed.

We present a real-world use case demonstrating the application potential of our approach.

2 RELATED WORK

Our work relates to visualization and analysis of movement data with grouping. As a main area of study, this data is considered in the geographic context [1, 2]. Movement data also prominently arises in the domain of time-dependent diagram data, and in particular, time-dependent scatter plot data. Often, the time-dimension of scatter plot data is shown by animation [5], usage of trajectories [13], or a empirical comparison of both approaches [14]. For visualization of groups of

- T. von Landesberger is with Technische Universität Darmstadt, Germany, e-mail: tatiana.von.landesberger@gris.tu-darmstadt.de.
- S. Bremm is with Technische Universität Darmstadt, Germany, e-mail: sebastian.bremm@gris.tu-darmstadt.de.
- T. Schreck is with Universität Konstanz, Germany, e-mail: tobias.schreck@uni-konstanz.de.
- D. W. Fellner is with Technische Universität Darmstadt and Fraunhofer IGD, Darmstadt, Germany, e-mail: d.fellner@igd.fraunhofer.de.

points, hull-formation methods such as bubble sets [4] or other compact shapes can be applied.

Wilkinson et al. [16] proposed a set of analytical point data measures for grouping and heuristically filtering of static scatter plots for expected interestingness to the analysis. The approach is applicable to static (non-time-dependent) and one-group data. For a comparative analysis of several moving objects, a number of time-dependent movement parameters (features) were computed, such as direction, speed and acceleration [11]. The parameters are visually represented in a matrix-oriented view allowing to compare moving objects and identifying correlated or concurrent movement patterns. Dodge et al. [7] applied time series analysis of movement parameter (feature) data to decompose trajectories into meaningful segments for pattern clustering and similarity computation.

Previously, we introduced the concept of feature-based visual analysis of movement data, explicitly also considering groups of data [15]. The approach was based on extraction of potentially interesting features from visual data abstractions such as hulls and point distributions. Interesting views on the time-dependent data were identified in a purely interactive way, i.e., the analyst inspecting line-graph visualizations of extracted features.

In this paper, we substantially extend our baseline previous work by an extended set of analytical features. Furthermore, we apply time series analysis [8] on the extracted features for automatic identification and recommendation of potentially interesting movement data segments.

3 APPROACH

We present an approach for visual analysis of time-dependent, grouped 2D movement data. It is based on combined visual analysis of the base data with derived and automatically analyzed movement features. It comprises four steps:

1. Movement data visualization in a scatter plot-based display (see Figure 1a). Individual objects are displayed as points, and groups are shown using enclosing hulls. The time dimension is reflected by animation and trajectories (see Section 3.1).
2. Movement feature computation (see Section 3.2) is performed on input data objects (or groups). Extracted features are shown by time line plots linked to the scatter-plot view (see Figure 1b).
3. The movement features are automatically analyzed using user-selected time series analysis functions (see Section 3.4).
4. Interesting data parts are highlighted in the line and scatter plot displays based on time-series analysis (see Figure 1c&d). This guides the user to the interesting data parts, and is particularly useful for analysis of large and complex movement data too large to be inspected only manually.

This workflow can be used iteratively by analysts. Figure 2 illustrates the overall approach. As compared to previous work [15], steps 3 and 4 are new and the set of movement features is extended.

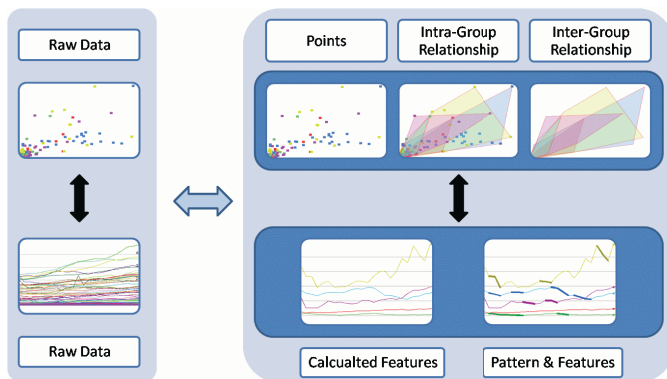


Fig. 2. Proposed analytical workflow.

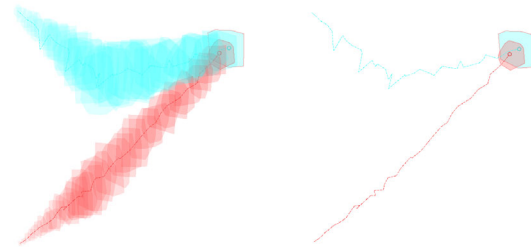


Fig. 3. Visualizing dynamics of point groups by traces of group hulls (left) or trajectories of the centers of mass (right).

3.1 Interactive Visualization of Grouped Movement Data

Visualization of movements of objects and groups follows approach presented in [15]. Position of objects in a group is displayed in a scatter plot, where each group is distinguished by color. Time is reflected by animation or trajectories. Groups of entities are shown using enclosing hulls of different shapes (minimum bounding boxes, circles, and convex or alpha hulls). Distribution of points within hulls can be indicated by density plots. For larger data sets, the groups can be reduced to their center of mass to avoid overplotting. Group dynamics are shown by tracing the path covered by the group hull movements using transparency or trajectories of the center of mass movements (see Figure 3).

3.2 Extended Features for Group Movements

We propose new useful movement features in addition to previously proposed ones [3, 16, 15]. We next present all features used in our system, marking the new ones by “(*)”. A large set of features is useful to cover a wide variety of possible use cases and data characteristics.

3.2.1 Extended Features for Object Movement Within a Group

Our set of features describing point dynamics in groups extends trajectory features for single entities described by Andrienko et al. [1, 3]:

1. *Movement length*: (the distance covered by an object). Includes last step length, total path and distance from start.
2. *Direction*: The direction is measured as an angle to the X axis. Includes direction of the current step and the direction from start.
3. (*) *Position in group*: The relative position of an object in a group measured by distances to group boundary and to center.
4. (*) *Co-movement with group*: Indicates whether the object moves in coherence with the group as approximated by the movement of the group center of mass. Includes length and direction of movement.
5. *Outlying*: indicates outliers in the groups.

3.2.2 Extended Features for Single Group Movement

Features for static point groups from [16] are extended by state and movement measures based on mid-point and PCA description of the group, and measures for assessing the group dynamics.

1. *Group size* (*): group size features include area and diameter (*) of the group hull. They show compactness of the entities.
2. *Point distribution*: measures the homogeneity of point positions in a group by skewness, sparsity, and strait.
3. *Group shape* (*): describes the shape of the group using convexity, skinniness, and stringiness measures from [16]. We also use PCA based measures (*) for the shape calculated as relative proportion of first and second eigenvalues of points. A large number indicates long narrow shapes, numbers close to one the opposite.
4. *Group alignment* (*): alignment of the points is measured. We extend this measure by the direction of the points based on PCA major axis (*) direction.
5. (*) *Movement of the group*: indicates direction and speed of group movement approximated by mid point dynamics.

3.2.3 Extended Features for Multiple Group Movements

When analyzing multiple groups, we either analyze each group separately and compare their measures, or consider the measures for the inter-group relations. When comparing more than two groups, measures for all pairs of groups can be computed and average, minimum and maximum are used.

1. *Overlap*: shows group overlap. Measured by absolute and relative overlapping area as the (relative) area of the group intersection to the group union.
2. (*) *Distance*: The distance between groups is measured as distance between their center of mass points.
3. (*) *Relative position*: measured by direction and distance between mid points of groups (applicable when comparing two groups).

3.3 Feature Normalization and Calculation Complexity

When analyzing the features their scale may be relevant. Where possible, normalization is done according to Wilkinson et al. [16], alternatively a simple linear min-max normalization of features is applied. The computational complexity of the proposed features ranges from $O(n)$ to $O(n^3)$. For large datasets, features can be pre-calculated offline to keep analysis interactive.

3.4 Automatic Feature Analysis

We use time series analysis to support the user in spotting interesting sections or time intervals in the data. A set of analysis functions was chosen so that various data patterns can be discovered. Specifically, they identify correlated movements of the features or data points, trends in the data, sharp changes in direction or speed, other extreme values, and user-specified search patterns in the data.

3.4.1 Search for Pre-defined Patterns

In the literature, a variety of meaningful characteristics in time series has been introduced. Our approach proposes to use a broad spectrum of these known techniques to identify interesting patterns in the data. Our time series analysis functions include:

- *Maximum and minimum*: identification of the extrema.
- *Moving averages*: smoothing of time series supports analysis of patterns also for noisy and high-frequency data.
- *Bollinger bands*: expected mean, expected highest and lowest values [6], often used in financial data analysis. Yields two time series for the upper and lower bands.
- *General trend*: We calculate linear, constant, polynomial, or exponential trends in the feature data using linear regression. E.g., increasing linear trend [12] in the total distance indicates constant speed of data dynamics. The type of the trend is user-specific. Results in a time series for the trend path.
- *Outlier detection*: identification of outlying feature values. Defined as values (time points) that are out of bounds compared to some calculated data trend (e.g., linear, polynomial, exponential) or Bollinger-Bands.
- *Local trend*: allows for spotting short-term increase, decrease or stable periods in the data. Defined as a consecutive set of time moments for which the consecutive difference of feature values are above, below, or within a user-given tolerance threshold.

Time points or intervals identified by these analysis functions are shown in the display by markers, or by highlighting the curve segments in the respective time lines. Derived time series like Bollinger bands or trends are shown in the view using the same color as the original time series (see Figure 4). illustrates.

3.4.2 Search for User-defined Patterns

We also provide the search for custom data- or task-dependent patterns (e.g., a sharp peak followed by a slow decrease).

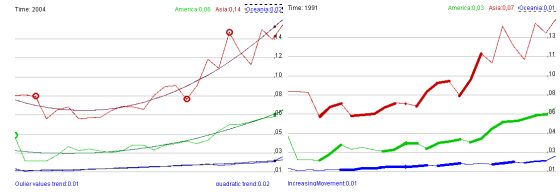


Fig. 4. Feature visualization with analytical functions. Left: Original data with quadratic trend and outlier identification. Right: Temporal trend search result highlighted using bold line segments.

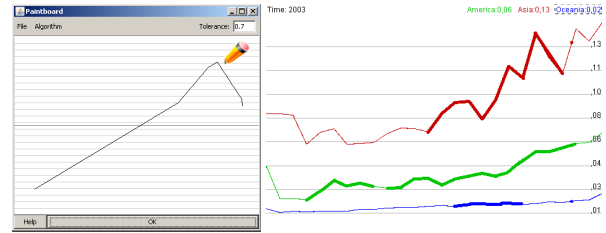


Fig. 5. Pattern definition and search. Left: Pattern drawing window. Right: Pattern search result marked using bold line segments.

Time series queries can be defined in two ways:

- (1) marking existing time series segments similar to [9], or
- (2) by interactively drawing a curve sketch (see Figure 5 left). Identification of user-defined time series patterns is done by two alternative similarity measures: The correlation measure or the Euclidean distance of SAX-descriptor representations [10] of the patterns, both in conjunction with a sliding window search procedure. The retrieved patterns are highlighted in bold or filtered in the monitoring window (see Figure 5 right).

4 APPLICATION

We apply our approach to the analysis of the development of 2D stock market indicators. To support financial investment decisions, the analysis of *risk-return indicators* of financial assets (stocks) is of interest. We here analyze asset indicator movements in risk-return diagram space, a widely used diagram in financial data analysis. A data set for 30 blue chip German assets listed in the German Stock Index DAX between June 2005 and August 2007 is used.

We first analyze *asset movements within a country*. Owing to the high volatility of the market movements, the visualization of asset trajectories in the risk-return plot does not reveal any particular insights into the data (see Figure 6 left). We therefore analyze step length and distance to mid-point features for identification of interesting data patterns (see Figure 6 right). The mid-point of the group represents an approximation of the so-called market portfolio, and is a relevant property to compare individual group members against. Although several exceptional movements stand out in the line-chart feature plot, the plot in general is still cluttered. Therefore, algorithmic analysis of the movements is performed, revealing several stocks with interesting behaviors (see Figure 7 right). We use linear trend and outlier detection methods. The stocks with large movements during short periods include *Schering* and *Altana*. Based on our consulting of various financial news archives, Schering's extraordinary movements can be explained by its potential acquisition by Merck AG at the beginning of March 2006. The volatilities of Altana can be explained by then speculations on large dividend yield to be paid after an outstanding result in the first quarter of 2006. The linear trend of the feature reveals that the distance to the mid point of most assets stays constant, however some assets move closer to the center (*Infineon*) or further away from the center (*ThyssenKrupp*). These movements could hardly be recognized from neither the movement visualization nor the standard line-chart feature monitoring, but automatic analysis of movement features provides effective analysis in this case.

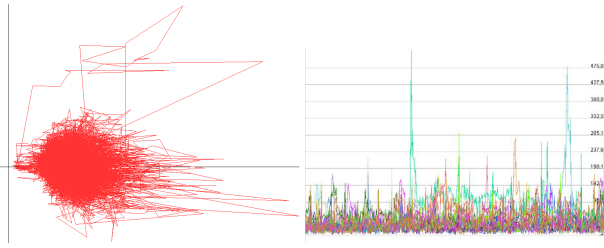


Fig. 6. Left: Movements in risk-return diagram space show too high volatility for meaningful analysis. Right: Distance from mid point feature.

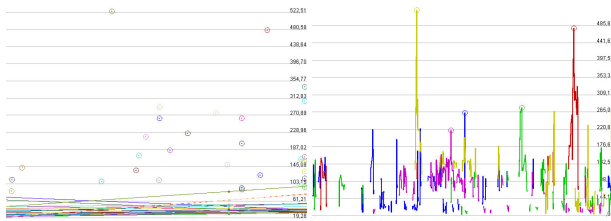


Fig. 7. Distance from mid point. Left: Maximum values and linear trend. Right: Maximum values and outliers linear trend.

Similar, direct visualization of *group development* over the whole time period is difficult to interpret due to cluttering (see Figure 8a). Therefore, we calculate and visualize group features (area, PCA relative eigenvalue and convexity) for these data (see Figure 8c). Owing to the very volatile dynamics of the features, it is difficult to recognize individual values. To reveal underlying movements and extreme values, we calculate and examine the moving averages and the maximum values (see Figure 8d). This exhibits several interesting time points. On 8.5.2007, we see a high density of assets in a small area and an outlier which causes an increase of the area while decreasing convexity (see Figure 8b). It is *Altana*, already spotted in the analysis of individual stocks. On 16.8.2007, the maximum area indicator stands out. Interestingly, the high area at this time point is connected with low levels of the PCA eigenvalues. This date shows strong movements on the whole market owing mainly to uncertainties during subprime mortgage crisis (see also Figures 8e and 8f).

5 CONCLUSIONS AND FUTURE WORK

We presented a novel visual analytics approach for analysis of group movement data. An encompassing set of features for describing data dynamics is introduced and time-series analysis of movement features helps users identify interesting data segments even in case of large data and cluttered displays of the raw data. The methodology is widely applicable in areas dealing with 2D time-dependent data (e.g., geospatial or financial data analysis). Future work will consider extending the approach to include prediction of movement data and deriving guidelines which features to use for specific application contexts.

REFERENCES

- [1] G. Andrienko, N. Andrienko, I. Kopanakis, A. Ligtenberg, and S. Wrobel. *Mobility, Data Mining and Privacy*, chapter Visual Analytics Methods for Movement Data, pages 375–410. Springer Berlin Heidelberg, 2008.
- [2] N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data - a systematic approach*. Springer, 2006.
- [3] N. Andrienko, G. Andrienko, N. Pelekis, and S. Spaccapietra. *Mobility, Data Mining and Privacy*, chapter Basic Concepts of Movement Data, pages 15–38. Springer Berlin Heidelberg, 2008.
- [4] C. Collins, G. Penn, and S. Carpendale. Bubble sets: Revealing set relations with isocontours over existing visualizations. In *Proceedings of IEEE Information Visualization*, 2009.
- [5] P. Craig and J. Kennedy. Coordinated graph and scatter-plot views for the visual exploration of microarray time-series data. In *Proceedings of IEEE Symposium on Information Visualization*, pages 173–180, 2003.

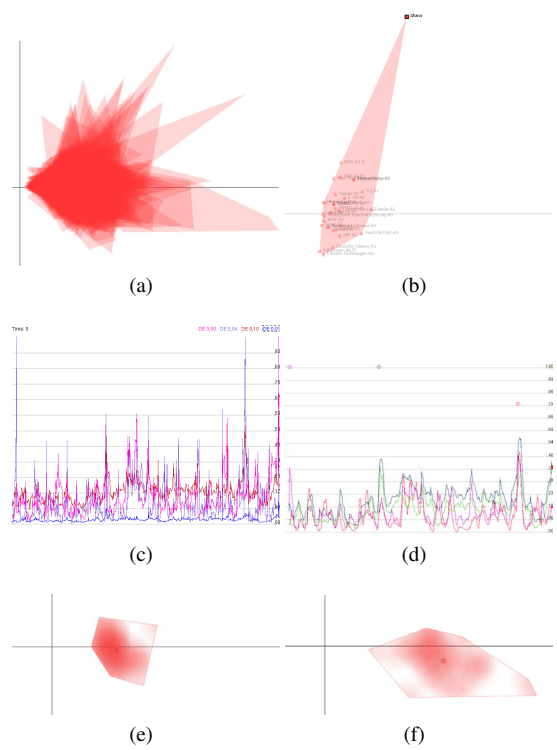


Fig. 8. (a) (Cluttered) hull trace of the full group development. (b) Situation on 8.5.2007. (c) Display of the extracted features (area, PCA relative eigenvalue and convexity). (d) Analysis of features showing maxima and moving averages. (e, f) Density plot of August 15 and 16 2007, showing rapid movements caused by subprime mortgage crisis.

- [6] E. D. Dobson. *Understanding Bollinger Bands*. Traders Press, 1994.
- [7] S. Dodge, R. Weibel, and E. Foroootan. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*, 33(6):419–434, 2009.
- [8] J. Hamilton. *Time series analysis*, volume 2. Cambridge Univ Press, 1994.
- [9] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, Mar. 2004.
- [10] E. Keogh. A decade of progress in indexing and mining large time series databases. In *Int. Conference on Very Large Data Bases (VLDB)*, 2006.
- [11] P. Laube, S. Imfeld, and R. Weibel. Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, 19(6):639–668, 2005.
- [12] D. Montgomery, E. Peck, and G. Vining. *Introduction to linear regression analysis*, volume 49. John Wiley & Sons, 2007.
- [13] C. North, U. Farooq, and D. Akhter. Datawear: Revealing trends of dynamic data in visualizations. In *Proceedings of IEEE Symposium on Information Visualization (InfoVis 2001)*, pages 8–11, 2001.
- [14] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. Effectiveness of animation in trend visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1325–1332, Nov.-Dec. 2008.
- [15] T. von Landesberger, S. Bremm, P. Rezaei, and T. Schreck. Visual analytics of time dependent 2D point clouds. In *Proc. Computer Graphics International*, pages 97–101, 2009.
- [16] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1363–1372, Nov.-Dec. 2006.