6

Visual-Interactive Analysis With Self-Organizing Maps - Advances and Research Challenges

Tobias Schreck
Technische Universitaet Darmstadt, Computer Science Department, Interactive Graphics Systems Group
Germany

Abstract
Based on the Self-Organizing Map (SOM) algorithm, development of effective solutions for visual analysis and retrieval in complex data is possible. Example application domains include retrieval in multimedia data bases, and analysis in financial, text, and general high-dimensional data sets. While early work defined basic concepts for data representation and visual mappings for SOM-based analysis, recent work contributed advanced visual representations of the output of the SOM algorithm, and explored innovative application concepts.
In this article, we review a selection of classic and more recent approaches to SOM-based visual analysis. We argue that important improvements have been achieved which allow effective visual representation and interaction with the output of the SOM algorithm. We identify promising directions for future research, which will support new application areas and provide additional advanced visualization approaches.

1. Introduction

Driven by technological advances in data acquisition, processing, storage and dissemination, increasingly large and complex data sets become available. Important fields in which huge amounts of data arise include the Multimedia, Science and Engineering, and Business domains. In Multimedia, digitization of existing and authoring of new content contribute toward formation of large repositories of digital audio-visual material. In Science and Engineering, vast amounts of primary research data arise in experiments, simulation and phenomena observation. And in the Business domain, a wealth of data arising from customer transactions, monitoring of productive processes etc. is captured by many companies. While *storage technology* allows to persistently save much of these data volumes, it is less obvious how *automatic analysis* and *interactive access* methods can support retrieval and discovery of interesting and useful information therein.
To effectively deal with growing amounts of complex data, besides application of purely automatic solutions, a promising approach is to apply visual-interactive data reduction and exploration techniques (Hinneburg et al. (1999)). The general idea here is to abstract large data sets to a condensed representation which captures important aspects in the data, abstracting from details and allowing better navigation and interpretation by the user. The *Self-Organizing Map* (SOM) (Kohonen (2001)) algorithm, to date applied on many different data types, has proven to be a very suited algorithm to this end. SOM has a strong disposition for visual cluster anal-

ysis, as it not only provides the data reduction, but also a spatialization of cluster prototypes forming a baseline for visualization and interaction with the data.
In this article, we survey applications of the SOM algorithm for visually supported retrieval and analysis tasks in a variety of data domains. Our survey is focused on two aspects in SOM-based data analysis. First, we aim to cover a range of different approaches to *visualization* of the output of the SOM algorithm, as visualization often forms the key interface between the algorithm output and the user. Second, we aim to cover a range of different *application* types, illustrating the principal applicability of the algorithm on any data for which a meaningful vector representation can be defined. The choice of examples presented is done to survey both classic, well-established visualization approaches and application types, as well as more experimental and innovative ones. The goal is to motivate the wide applicability of current SOM methodology, and identification of promising future research options leading to further advanced SOM visualizations and application concepts.
The remainder of this article is structured as follows. Section 2 discusses fundamental options for training of Self-Organizing Maps and for their visualization. In Sections 3 and 4, we survey a selection of SOM-based visualization approaches and application concepts, illustrating progress made over basic SOM visualization and application concepts. Section 5 gives an assessment of the surveyed techniques, identifying promising directions for future development of SOM visualization and application fields. Finally, Section 6 concludes.

2. Background

The Self-Organizing Map algorithm is one of the most popular visual cluster analysis algorithms. Its applicability on a broad range of data types has been empirically demonstrated by an abundance of research papers to date. Visualization of SOM output plays a key role in many successful applications of SOM. In Section 2.1, we recall the main issues in training SOMs from complex data, and in Section 2.2, basic options for visual analysis of SOM output are discussed. We also recall results from Information Visualization relevant to SOM visualization in Section 2.3.

2.1 Training of Self-Organizing Maps

The Self-Organizing Map algorithm is a scheme for training a neural network to represent a distribution of high-dimensional input data. A low-dimensional (usually, 2-dimensional) grid of reference vectors is learned from the input data by means of competitive, iterative adjustment of reference vectors to the input data space. Effectively, the SOM is a combined vector quantization and projection algorithm, as (a) many input records are represented by a fixed number of reference vectors, and (b) the reference vectors representing the input data are given a topological ordering by the SOM grid. The SOM yields (a) a clustering of the data and (b) approximately preserves the topology of the data points from the input space, and is therefore especially useful for data visualization and exploration purposes. We here do not recall the SOM algorithm details but instead refer to Kohonen's monograph (Kohonen (2001)).
Application of SOM on a data set requires two main prerequisites: A *vector representation* of the data to be analyzed (usually called Feature Vector, or descriptor), and the *definition of a set of training parameters*. If the data elements are already characterized by application-specific numeric attributes, then these can be used as a descriptor. In case of non-standard data, e.g., multimedia objects, a feature extraction step has to be performed. The latter usually involves heuristics in the definition of the features (Duda et al. (2001)), and can be optionally supported by methods of feature selection (Liu & Motoda (2007)). After the vector description has been

obtained, a number of SOM parameters needs to be set by the user. Main parameters include the resolution and topology of the prototype network, the learning rate and training kernel, and length of the training run. Note that both descriptor extraction and SOM parameterization are non-trivial tasks usually requiring experience of the user both in general data analysis as well as in the specific data domain to be addressed.

2.2 Basic Visual Analysis of SOM Output

The output of the SOM is a network of prototype vectors representing the input data set. Contrary to the output of other cluster algorithms such as e.g., k-means, the SOM output allows straightforward visualization. Usually, the SOM network structure is exploited to map the prototype vectors to the visual variable *position*, while other visual variables such as *shape* and *color* are used to encode properties of the prototype vectors and their relationships regarding the data vectors. Vesanto (1999) distinguishes three SOM-analytic use cases and supporting visualization approaches:

2.2.1 SOM Cluster Structure

Here the task is to assess the structure of the SOM output in terms of the relation between the prototype vectors, usually quantified by their relative vector distances. Distances between reference vectors can be visualized directly on the SOM network by gray- or color-coding (U-Matrix visualization) or by glyph-based approaches scaling e.g., the size of SOM prototype glyphs. Furthermore, the SOM prototype vectors can be compared for similarity by means of subsequent projection methods such as Multidimensional Scaling. Both techniques may support the user in discriminating between regions of the SOM.

2.2.2 Prototype Vector Analysis

Here, a detailed analysis of the properties of prototype vectors in terms of individual vector dimensions is aimed at. The spread of dimensions over the SOM network can be visualized by color-coding of single selected dimensions in form of so-called component planes. Multiple component planes can be visualized simultaneously by arranging them in a matrix layout. Besides the spread of component values, the contribution of selected components toward distances between cluster prototypes can be visualized by color-coding. An important analysis goal here is to understand the correlation between the cluster structure and individual components.

2.2.3 Cluster Structure and Data Distribution

Here, the distribution of data elements underlying the SOM analysis, or new data elements to be discussed in terms of an existing SOM structure, is considered. Single data elements can be positioned on the SOM by simple marks indicating the best matching unit, or response surfaces indicating the distance between the sample to all SOM prototype vectors by means of color-coding. Density histograms visualize the aggregated distribution of many data elements over the SOM by color-coding or glyph size as basic visualization options. Density itself can be estimated in the nearest neighbor sense, but also voting-based approaches forming smoothed density histograms are possible (Pampalk et al. (2002)).

2.2.4 Background Summary

These approaches comprise some of the most fundamental SOM visualization options, and are specifically useful if the vector components can be meaningfully interpreted by the user. In

case of more abstract vector components, which often is the case with derived Feature Vectors, other forms of visualization of represented data can be useful. *Thumbnail Maps* use a visual representation of selected data elements (if available) to visualize the spread to data samples over the SOM.

2.3 SOM and Information Visualization

Information Visualization is concerned with the visual representation and interactive exploration of information spaces that are not inherently spatial in nature. The goal is to foster insight into the data, and stimulate interactive visual data analysis (Ware (2004); Card et al. (1999); Spence (2006)). To date, SOM has been applied for data preprocessing in many Information Visualization systems. Important interaction and visualization concepts developed in Information Visualization are useful for application in SOM-based data analysis.
The ordered layout of data items is an important principle applied in many Information Visualization solutions, and often, Self-Organizing Maps are employed to provide ordering of elements of a data set. Besides direct visualization of the SOM grid as discussed above, coordinated multiple views have been developed which allow deeper exploration of data characteristics of the SOM output and represented data (Guo et al. (2006)). Specific interaction concepts have been developed in Information Visualization for user navigation in dense data spaces (Card et al. (1999)). Distortion techniques adapt the display such to visualize detail information at a focus point set by the user, and compressing the surrounding context information by data reduction or visual aggregation. Dynamic queries allow to interactively filter the data to instantly show data selections matching a user query. Visualization techniques developed for high-dimensional data such as Parallel Coordinates (Inselberg & Dimsdale (1990)), Glyph-based approaches (Ward (2002)) or Pixel-oriented techniques (Keim et al. (1995)) can be used for visualization of high-dimensional SOM prototype vectors.

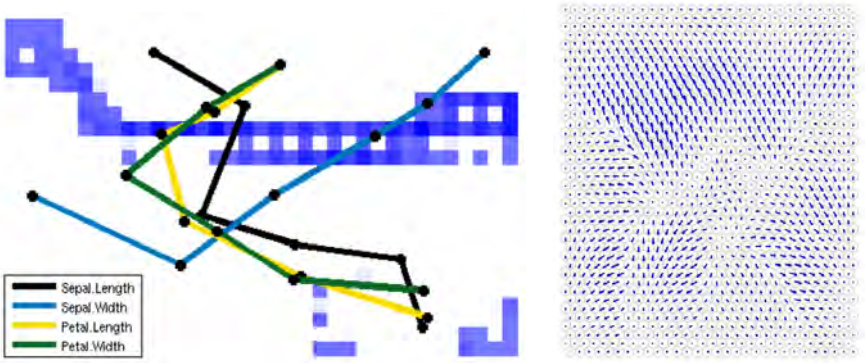
3. Approaches to Enhanced SOM Visualization

In this section, we recall several works we consider innovative on the side of visual SOM support, before we address innovative application domains in the next section. We note that while we made a distinction between visualization and application concept contribution, that distinction is not always clear cut, as often, novel visualization approaches foster new application concepts and vice versa.

3.1 Structural Enhancement for Component Planes

The consideration of characteristics of SOM reference vector components is among the most basic and important SOM analysis tasks (cf. also Section 2.2). Component planes and arrays thereof are a straightforward visualization for the spread of component values over the SOM lattice. However, the user needs to switch back and forth between multiple components to arrive at an understanding of multidimensional component spread, a task which incurs considerable cognitive load. Rauber in Neumayer et al. (2007) proposed to overcome this problem by jointly plotting characteristics of the spread of individual components. Specifically, by means of quantized component planes, path lines indicating the direction of increase or decrease of component values are calculated. Together with appropriate layout adjustments, multiple such path lines are overlaid over a baseline Umatrix visualization. The resulting visualization, called metro visualization, gives an aggregate view over several structural component spreads over the baseline SOM lattice. As it captures multiple component spread patterns, it reduces cognitive user load as no switching between component plane plots is required. Also, the

extracted path lines are an abstraction of the overall information contained in the component plane plots. Detail information is abstracted in favor of overall fundamental information. This in turn may help the user in correlating the individual components. Figure 1 (a) illustrates the concept.



(a) Structural component plane analysis (b) Vector field display

Fig. 1. (a) The so-called metro visualization overlays structural component spread paths over a baseline Umatrix visualization (Figure © 2007 IEEE, taken with permission from Neumayer et al. (2007)). (b) The vector field enhancement indicates the direction of attraction for each SOM cell on the lattice. The resulting visualization is useful for visual evaluation of overall cluster structure given by the SOM (Figure © 2005 IEEE, taken with permission from Polzlbauer et al. (2005)).

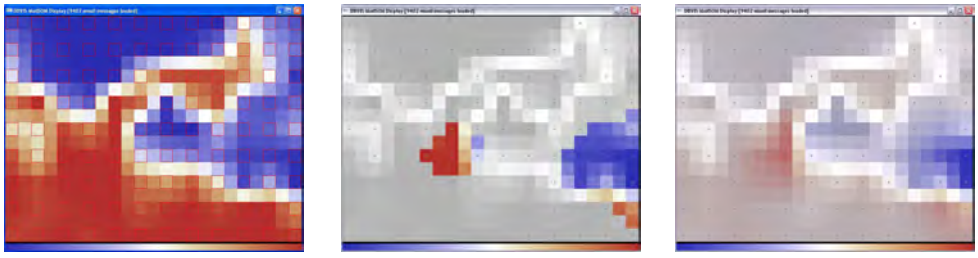
3.2 Vector Fields for Enhanced Umatrix Visualization

Another key analysis task in visual SOM analysis is determination of cluster structure by the user. As the SOM algorithm not directly returns a measure for the number or boundaries of data clusters, this is a key interpretation which the user has to perform. While automatic analysis may help during SOM post-processing (e.g, by applying a post processing cluster algorithm on the SOM prototypes), often the Umatrix visualization is interactively employed to this end. However, the Umatrix in its standard form considers only distances between adjacent SOM reference vectors. A more elaborate visualization to support interactive cluster analysis was proposed in Polzlbauer et al. (2005). For each prototype, the direction toward the most similar map area is evaluated by integration over a set of neighboring prototypes; that direction is then represented in form of a vector. If the direction vectors are plotted on the SOM grid, the resulting gradient vector field visualization gives visual hints to the user which may help in identification of the SOM cluster structure. The method comprises a user-settable parameter for the definition of the neighborhood size over which the similarity is evaluated. Figure 1 (b) illustrates a gradient vector field for a SOM lattice.

3.3 Bivariate Color Maps

Email nowadays is one of the most important means of communication, and many people maintain growing personal email archives constituting large information collections. Besides

the possibility to transport any kind of file, email can be considered a textual document. Consequently, information retrieval oriented techniques have been applied to the management of email archives. Based on term-frequency descriptors (Baeza-Yates & Ribeiro-Neto (1999)), SOM-based analysis of Email archives has been proposed in Nuernberger & Detyniecki (2006). The authors trained a SOM for an email archive, and presented to the user the SOM grid on which selected keywords describing the content of Email documents represented by each SOM cell were printed. Supported by methods to search for keywords, the proposed system supports explorative analysis of Email archives.
A problem with Email may be the occurrence of large amounts of unsolicited Email ("spam") which detriments productivity of the Email users. In Keim et al. (2005), a SOM-based Email visualization system was proposed that leverages a color mapping scheme to visualize the degree to which SOM clusters contain spam email. Specifically, a spam classifier software was used to calculate for each email document its probability of being spam (so-called spam score, treated as a continuous attribute for each email). A color-mapping scheme was proposed which allows to simultaneously visualize the probability of each SOM cell to contain spam email, as well as the occurrence of selected keywords in the represented mails. To this end, for each SOM cell the spam scores of all represented Emails were averaged. That value was then mapped to a bipolar colormap going from blue (low scores, good mail) over white to red (high scores, spam mail). Thereby, red areas on the map indicate the presence of spam email, blue indicate the presence of valid email,while white areas indicate the border between both classes of email. Optionally, it was proposed to enhance the display by overlaying also the averaged frequency of user-selected keywords over each SOM cell. This was done by weighting the saturation channel of the mapped color by the normalized weight of the selected index term. Figure 2 illustrates a SOM trained of an archive of 10000 emails. Note that the visualization effectively combines two SOM views in one: A spam histogram (overall SOM structure) and a component plane (magnitude of a selected vector component).



(a) Spam histogram (b) Thresholded saturation (c) Scaled saturation

Fig. 2. SOM-based analysis of a spam/non-spam classified E-Mail dataset. (a) spam-histogram (red indicates spam, blue indicates non-spam). (b)-(c) represent the component plane for frequency of a selected term overlaid over the spam histogram by thresholding (b) or linearly scaling the spam histogram via the saturation channel of the respective colors.

4. Innovative Application Concepts

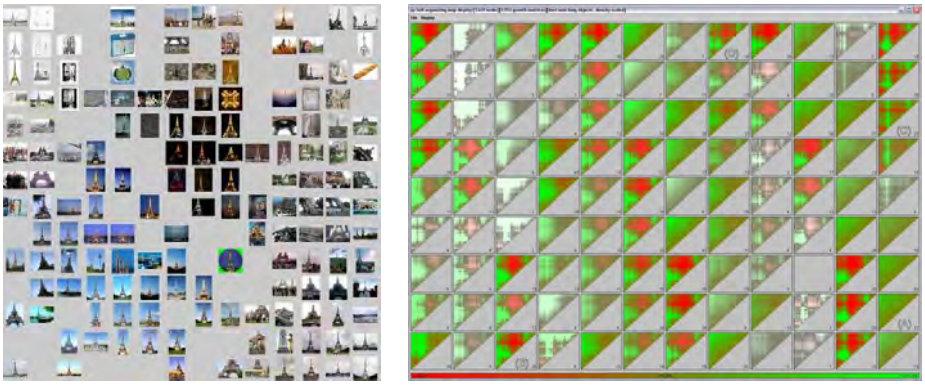
In this section, we review a number of recent SOM application concepts.

4.1 SOM-based Filtering for Image Retrieval

SOM to date has often been successfully applied in the Text (Nuernberger & Detyniecki (2006); Honkela et al. (1997)) and Multimedia Retrieval (Laaksonen et al. (2000); Pampalk et al. (2002); Bustos et al. (2004)) domains. Often, the goal is to provide an effective overview over the content of a document collection, to motivate the user to explore the collection (browsing), and to locate interesting documents which can in turn be used as query keys.
Image Sorter (Barthel (2008)) is an image search engine which shows particular deep integration of a SOM display of images with the core search engine. The overall goal is to provide efficient search in images on the Web by filtering the output of text-based Internet search engines. Specifically, Image Sorter uses Google image search to retrieve a seed set of images based on query by keyword. From the seed set, a SOM is trained using image color features. From the SOM-based overview, the user is then allowed to manually select a subset of the seed images (cf. Figure 3 for an illustration). These in turn are used in a subsequent retrieval stage. Specifically, a larger number of images is obtained using Google image search, yet this time the user selected image set is used to filter the larger answer set by content-based similarity with the selected image set. SOM training, image selection and retrieval are provided in real time, allowing effective image retrieval. Conceptually, the approach is regarded particularly appealing because it uses SOM to integrate an original content-based retrieval strategy with a keyword based web image search. The system is currently further developed within the Pixolution software suite (Pixolution. (2009)).

4.2 Small Multiple Views for Information Visualization

The sorting capabilities of the SOM algorithm make it applicable not only on general high-dimensional or multimedia data, but can also be used to sort sets of diagrammatic views or visualizations. In Information Visualization, often different views on the same data set are possible, and the user is confronted with selecting the most appropriate view. While one-by-one inspection of different views is a basic option, an overview of many views simultaneously can be helpful to quickly screen a large visualization space. As an example in this domain, in Schreck (2007) SOM was applied to sort many different views, effectively giving a SOM of many diagrams. The underlying diagrams are so-called Growth Matrices (Keim et al. (2006)), which encode in a color-coded triangular visualization all possible growth rates in a financial asset. The SOM was utilized to represent large sets of Growth Matrices in order to produce informative overviews. For training the SOM, the original growth matrix displays were down-sampled from matrices of edge lengths of several hundreds to matrices of size 20^2. The down-sampled matrices were used as Feature Vectors and made input to the SOM algorithm. A grid of 12×9 reference vectors was trained onto which the original data was mapped back. Figure 3 (b) illustrates a SOM obtained from the data described in Keim et al. (2006) by drawing the best matching Growth Matrix for each SOM node. Furthermore, a density histogram was overlaid over the display to show the frequency by which each Growth Matrix pattern is represented on the SOM. To this end, the saturation channel of each Growth Matrix thumbnail image was scaled proportional to the total number of objects matching the respective SOM cell.

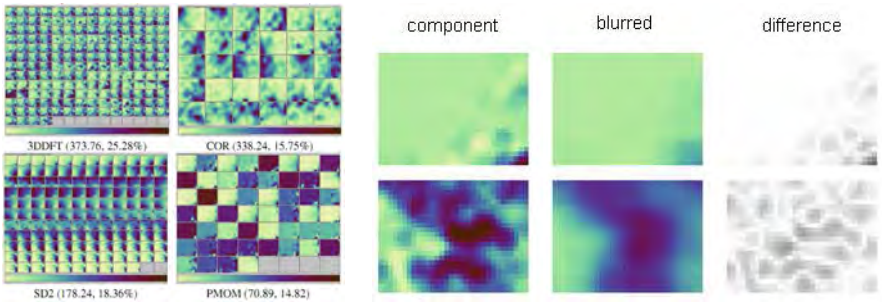


(a) Image Sorter system (b) Growth matrix display

Fig. 3. (a) The Image Sorter system uses color features to sort a set of images by similarity (Figure © 2008 IEEE, taken with permission from Barthel (2008)). The overview allows selection of query images for content based image retrieval. (b) Self-organizing map of 1.700 growth matrices. The SOM was trained from a down-sampled representation of the diagrams. Saturation scaling is performed on the thumbnail images to give additional information regarding the density distribution of the patterns.

4.3 Visual Feature Space Analysis

Complex data often needs to be described by Feature Vectors (FVs), before data analysis and similarity search tasks can take place. Often, many different FV representations of a given data set are possible, and the process of feature selection is expensive, often involving heuristics, human supervision, and benchmarking. In Schreck et al. (2006), Umatrix visualizations of a given data set described by different FV representations were compared. It was observed that Umatrices of different FV representations differed significantly regarding the distribution of distances between adjacent reference vectors (shown as gray- or color-codings in Umatrix displays). Regression experiments found a correlation between the uniformity of the distance distributions, and the discrimination capability of the underlying FV representations, as measured by ground truth benchmark information. The hypothesis was formulated that discriminative feature spaces can be expected to provide a rich mix of distances between adjacent reference vectors (forming high contrast images), while non-discriminating feature spaces tend to show more skewed Umatrix distance distributions. That hypothesis was further evaluated by experiments on synthetic data (Schreck, Fellner & Keim (2008)) as well as also considering distribution characteristics observed in component plane images (Schreck, Schneidewind & Keim (2008)). Figure 4 illustrates the latter analysis for several different feature representations of a benchmark data set. While more experimental and theoretical consideration is needed, the idea of applying unsupervised SOM analysis to support the feature selection process is considered promising.

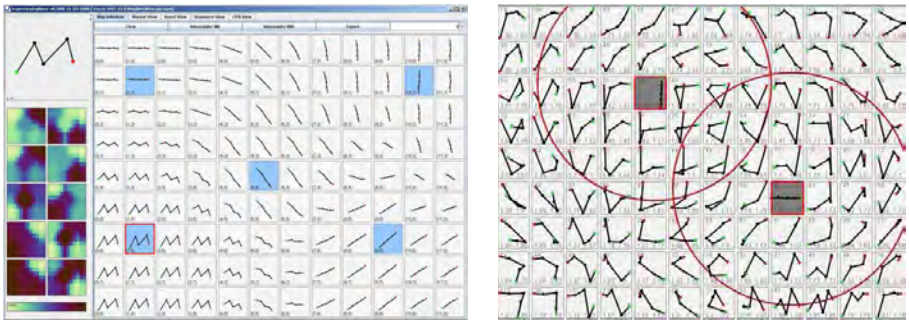


(a) Component plane arrays (b) Analysis of component plane images

Fig. 4. (a) Arrays of component planes for four different feature vector space representations of the same benchmark data set. (b) A *Difference-of-Gaussians* analysis yields a measure for the amount of information contained in a component plane image. An overall measure for each feature space is obtained by averaging over the Difference-of-Gaussian measures of each component plane image in that feature space. That score in turn can be used as a heuristic criterion for unsupervised feature selection, or to filter large sets of candidate feature spaces for analysis by an expert user.

4.4 Interactive Trajectory Analysis

Besides multimedia and general multidimensional data, SOM-based visual analysis of time-dependent data has also been successfully done in the past. A system considering SOM-clustering of one-dimensional time series data is described in Šimunić (2003). Two-dimensional time series in form of trajectories derived from scatter plot data have been analyzed with SOM in Schreck et al. (2007). Most SOM-based analysis systems operate by training the SOM in off line mode, and restrict to static visualization of the SOM analysis output. In Schreck et al. (2009), a concept for fully visual-interactive training of SOMs for trajectory pattern data is introduced. The basic idea is not only to visualize the SOM result as is evolves during training in real time, but also, to provide the user means to initialize and control the SOM algorithm interactively. To this end, an editor allows the user to visually initialize the grid of trajectories, and adjust important training parameters. At any time during the training, the user is able to pause the training, update specific parameters as is seen fit, and then continue or restart the training. Figure 5 (a) shows the interactive initialization of the prototype vector array, which is possible due to the specific vector representation of the trajectory data (Schreck et al. (2009)). Figure 5 (b) illustrates interactive control of the training process. In the example, the user manually adjusts two reference trajectories, and also updates the neighborhood kernel for the next training iterations to take place.



(a) Initialization (b) Interactive Control of Training

Fig. 5. (a) Interactive initialization of a grid of prototype trajectories prior to start of the SOM training. (b) Visual-interactive update of selected reference vectors and training parameters during online training.

5. Selected Research Challenges

In the previous sections, we have illustrated a number of examples which in our view represent innovative SOM-based visualization approaches and application concepts. For both aspects, we summarize the findings from the presented examples, and provide a selection of future work we consider promising to undertake.

5.1 SOM Visualization

Effective SOM-based data analysis relies on appropriate visualization of the output of the SOM algorithm. The visualization approaches presented in this article enhance standard SOM displays by (i) integrating structural properties of prototype components into the Umatrix display (Section 3.1); (ii) by using vector fields for improved visual cluster analysis (Section 3.2); and (iii) by mapping not one but two quantitative variables to color in a distribution view (Section 3.3). We observe that as the SOM algorithm provides rich information, the provision of displays adequately representing this information is important to allow effective interactive SOM analysis. The presented approaches aim to increase the density of the information visualized. Inspired by this review, interesting future work in SOM visualization may be identified along the following lines:

5.1.1 Visualization of High-Dimensional Data

The network of SOM prototype vectors and its associated sample distribution can be regarded as a high-dimensional visualization problem. Much potential is perceived in appropriately combining standard solutions from high-dimensional data visualization with the SOM data structure. E.g., appropriately designed glyph- or parallel coordinate plots, reflecting also the neighboring relationships, could help in packing more information into a single static SOM view.

5.1.2 Comparative SOM Visualization

Often, the data under concern can be described by many different feature representations, each reflecting different similarity notions. As it often is not clear which of these representations is the best suited, it is proposed to develop specific visualizations for comparative SOM

analysis. A connector-oriented approach as introduced in Holten (2006) could be suited as an appropriate tool to quickly compare the outcome of training several SOMs for the same data set described under different vector representations. Also on the interaction side, comparison of different SOMs of the same data set could be supported by e.g., means of animation. Then, how to transit between different SOMs is an interesting problem to solve.

5.1.3 Visualization of Sample Distribution Characteristics

A key SOM-based analysis task refers to analysis of the distribution of data samples over the SOM. In many cases, the number of data samples outperforms the number of SOM cells. For effective distribution visualization, appropriate visual representations of sets of samples need to be found, for overlay over the display. For many data types such as textual or multimedia documents, finding the right representations (e.g., labels or appropriately defined thumbnail previews) is a difficult problem, the solution of which could further contribute to advanced SOM visualization.

5.1.4 Visualization of SOM Training Process

Currently, in most cases only the final output of the SOM training, usually done in a black-box manner, is the basis for visualization. However, the online visualization of the training process is expected to help the user better understand emerging data structure yielded by the SOM. Also, visualization of the training could help in finding better parameterizations by the user, by means of interaction with the training algorithm itself. While one system that visualizes the SOM training on a specific data type (trajectories) has been presented in Section 4.4, the visualization of SOM training for general data types is an interesting problem for future work.

5.2 Application Concepts

On the application side, we have reviewed a selection of SOM-based application concepts that (i) showed the deep application integration of SOM analysis in a retrieval system (Section 4.1); (ii) SOM as a layout generator for diagram visualization (Section 4.2); (iii) visual feature space analysis (Section 4.3); and (iv) interactive SOM training (Section 4.4). These are only a few of many new SOM applications introduced which demonstrate that the potential for SOM applications is far from being exhausted. A number of interesting future research directions can be outlined in the application context as follows.

5.2.1 SOM quality assessment

The quality of a given SOM can be measured by many aspects, ranging from analytic measures of quantization error or topology preservation to user-oriented measures based on interpretability, subjective notions of data similarity, etc. The definition of a set of flexible quality assessment capabilities would benefit many application areas, in leading to better quality assessments by the user. Also, SOM quality assessment should be tightly coupled with interactive parameterization of the SOM training process, to allow the user to instantly compare and re-generate different SOM results, to obtain results best suiting the given application and data set.

5.2.2 SOM Postprocessing and Analysis Support

SOM analysis often includes considerable user effort during an extensive interpretation stage. Many post processing techniques can in principle support the interpretation stage. Examples include post processing for automatic detection of SOM clusters, or comparison of the SOM

output with those of other clustering algorithms. Integration of such analysis into the SOM application could increase effectiveness of SOM analysis by the user.

5.2.3 Applications in Information Retrieval

Browsing and searching in data are closely related. In multimedia retrieval, query-by-example is a popular query concept, in which the user supplies a multimedia object as a query key to search for. Obtaining appropriate query examples requires efficient browsing capabilities, allowing the user to quickly screen large data sets. While the SOM is well suited to provide the basic structure for browsing, issues relating to the presentation of sets of elements and the efficient selection of individual elements thereof is a non-trivial problem. We see a need to develop further methods to more tightly integrate SOM-based browsing facilities with query-by-example oriented retrieval.

6. Conclusions

We surveyed a set of SOM-based works illustrating, as we consider, innovative visualization approaches and application concepts. We argued that while the base SOM algorithm is already well-known and widely used, much innovation potential exists in further improving SOM visualization and defining novel application areas. Specifically, the Information Visualization field offers a rich background of visualization approaches which can be exploited for advanced SOM visualization. Deeper application integration and visual-interactive control of the SOM training process are regarded promising for future applications. The tight integration of automatic data analysis and visualization, as demonstrated by many SOM-based systems, offers much potential in handling increasing amounts of complex data, and is considered a promising basis for efficient information retrieval systems.

Acknowledgments

We thank Tatiana von Landesberger, Juergen Bernard, Sebastian Bremm, and Joern Kohlhammer for fruitful discussion of SOM-based visual analysis methods. Figures 1 (a+b) are courtesy of Andreas Rauber. Figure 3 (a) is courtesy of Kai Uwe Barthel.

7. References

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, Addison-Wesley.

Barthel, K. U. (2008). Improved image retrieval using automatic image sorting and semi-automatic generation of image semantics, *Image Analysis for Multimedia Interactive Services, International Workshop on* **0**: 227–230.

Bustos, B., Keim, D., Panse, C. & Schreck, T. (2004). 2D maps for visual analysis and retrieval in large multi-feature 3D model databases, *Proceedings of the IEEE Visualization Conference (VIS'2004)*, IEEE Press. Poster paper.

Card, S., Mackinlay, J. & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*, Morgan Kauffman.

Duda, R., Hart, P. & Stork, D. (2001). *Pattern Classification*, 2nd edn, Wiley-Interscience, New York.

Guo, D., Chen, J., MacEachren, A. M. & Liao, K. (2006). A visualization system for space-time and multivariate patterns (VIS-STAMP), *IEEE Transactions on Visualization and Computer Graphics* **12**(6): 1461–1474.

Hinneburg, A., Wawryniuk, M. & Keim, D. A. (1999). HD-eye: Visual mining of high-dimensional data, *IEEE Computer Graphics & Applications Journal* **19**(5): 22–31.

Holten, D. (2006). Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data, *IEEE Transactions on Visualization and Computer Graphics* **12**(5): 741–748.

Honkela, T., Kaski, S., Lagus, K. & Kohonen, T. (1997). WEBSOM—self-organizing maps of document collections, *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6*, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, pp. 310–315.

Inselberg, A. & Dimsdale, B. (1990). Parallel coordinates: a tool for visualizing multi-dimensional geometry, *VIS '90: Proceedings of the 1st conference on Visualization '90*, IEEE Computer Society Press, Los Alamitos, CA, USA, pp. 361–378.

Keim, D. A., Ankerst, M. & Kriegel, H.-P. (1995). Recursive pattern: A technique for visualizing very large amounts of data, *VIS '95: Proceedings of the 6th conference on Visualization '95*, IEEE Computer Society, Washington, DC, USA, p. 279.

Keim, D., Mansmann, F. & Schreck, T. (2005). Mailsom - visual exploration of electronic mail archives using Self-Organizing Maps, *Second Conference on Email and Anti-Spam (CEAS 2005), Stanford University, Palo Alto, CA, USA, July 21-22*. Short paper.

Keim, D., Nietzschmann, T., Schelwies, N., Schneidewind, J., Schreck, T. & Ziegler, H. (2006). A spectral visualization system for analyzing financial time series data, *Proceedings of the EuroVis 2006: Eurographics/IEEE-VGTC Symposium on Visualization, Lisbon, Portugal, May 8-10 , 2006*, IEEE Computer Society.

Kohonen, T. (2001). *Self-Organizing Maps*, 3rd edn, Springer, Berlin.

Laaksonen, J., Koskela, M., Laakso, S. & Oja, E. (2000). PicSOM—content-based image retrieval with self-organizing maps, *Pattern Recogn. Lett.* **21**(13-14): 1199–1207.

Liu, H. & Motoda, H. (eds) (2007). *Computational Methods of Feature Selection*, Data Mining and Knowledge Discovery, Chapman & Hall/CRC.

Neumayer, R., Mayer, R., Poelzlbauer, G. & Rauber, A. (2007). The metro visualisation of component planes for self-organising maps, *Proceedings of International Joint Conference on Neural Networks*, IEEE.

Nuernberger, A. & Detyniecki, M. (2006). Externally growing self-organizing maps and its application to e-mail database visualization and exploration, *Applied Soft Computing* **6**(4): 357–371.

Pampalk, E., Rauber, A. & Merkl, D. (2002). Using smoothed data histograms for cluster visualization in self-organizing maps, *Proc. Int. Conf. on Artifical Neural Networks (ICANN'02)*, Vol. 2415 of *Lecture Notes in Computer Science*, Springer.

Pixolution. (2009). Visual image search software. http://www.pixolution.de/.

Polzlbauer, G., Dittenbach, M. & Rauber, A. (2005). A visualization technique for self-organizing maps with vector fields to obtain the cluster structure at desired levels of detail, *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, Vol. 3, pp. 1558–1563 vol. 3.

Schreck, T. (2007). *Effective Retrieval and Visual Analysis in Multimedia Databases*, PhD thesis, University of Konstanz, Germany.

Schreck, T., Bernard, J., von Landesberger, T. & Kohlhammer, J. (2009). Visual cluster analysis of trajectory data with interactive kohonen maps, *Information Visualization* **8**(1): 14–29.

Schreck, T., Fellner, D. & Keim, D. (2008). Towards automatic feature vector optimization for multimedia applications, *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, ACM, New York, NY, USA, pp. 1197–1201.

Schreck, T., Keim, D. & Panse, C. (2006). Visual feature space analysis for unsupervised effectiveness estimation and feature engineering, *IEEE International Conference on Multimedia and Expo (ICME'2006). Toronto, Canada, July 9-12.*

Schreck, T., Schneidewind, J. & Keim, D. (2008). An image-based approach to visual feature space analysis, *Proc. Int. Conference in Central Europe on Computer Graphics, Visualization and Computer Vision.*

Schreck, T., Tekušova , T., Kohlhammer, J. & Fellner, D. (2007). Trajectory-based visual analysis of large financial time series data, *SIGKDD Explorations* **9**(2): 30–37.

Spence, R. (2006). *Information Visualization: Design for Interaction*, 2nd edn, Prentice Hall.

Vesanto, J. (1999). SOM-based data visualization methods, *Intelligent Data Analysis* **3**(2): 111–126.

Šimunić , K. (2003). Visualization of stock market charts, *Proc. Int. Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG).*

Ward, M. O. (2002). A taxonomy of glyph placement strategies for multidimensional data visualization, *Information Visualization* **1**(3/4): 194–210.

Ware, C. (2004). *Information Visualization*, Morgan Kaufmann.