Improving the Usability of Hierarchical Representations for Interactively Labeling Large Image Data Sets

Julia Moehrmann[1], Stefan Bernstein[2], Thomas Schlegel[3], Günter Werner[2], and Gunther Heidemann[1]

[1] Intelligent Systems Group, University of Stuttgart, Universitätsstr. 38, 70569 Stuttgart, Germany
{moehrmann,heidemann}@vis.uni-stuttgart.de
[2] University of Apllied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany
{sbernste,gwerner}@hs-mittweida.de
[3] Softwareentwicklung ubiquitärer Systeme, TU Dresden, Nöthnitzer Straße 46, 01187 Dresden, Germany
thomas.schlegel@tu-dresden.de

Abstract. Image recognition systems require large image data sets for the training process. The annotation of such data sets through users requires a lot of time and effort, and thereby presents the bottleneck in the development of recognition systems. In order to simplify the creation of image recognition systems it is necessary to develop interaction concepts for optimizing the usability of labeling systems. Semi-automatic approaches are capable of solving the labeling task by clustering the image data unsupervised and presenting this ordered set to a user for manual labeling. A labeling interface based on self-organizing maps (SOM) was developed and its usability was investigated in an extensive user study with 24 participants. The evaluation showed that SOM-based visualizations are suitable for speeding up the labeling process and simplifying the task for users. Based on the results of the user study, further concepts were developed to improve the usability.

Keywords: Self-organizing map, SOM, user study, image labeling, ground truth data.

1 Introduction

The importance of image recognition systems increases with the ubiquity of webcams and camera phones. However, the extensive development of image recognition systems for non-industrial purposes fails due to time and cost. One important factor is the labeling (or annotation) of the training images. Labeling the image data is necessary for building a ground truth data set on which the classifier for a specific recognition system can be trained. Since correctness in the ground truth data is crucial for the recognition rate, labeling has to be performed manually. Since image recognition systems need large amounts of images to cover the appearance space in all its details (variances in lighting, rotation, occlusion ...) it takes a lot of effort to assign correct labels to all images.

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2011, LNCS 6761, pp. 618–627, 2011.

Conquering the complexity of labeling large image data sets is possible by using a semi-automatic approach. Here, the image labeling is not performed manually on the individual images, but instead features are used for clustering the image data unsupervised. A resulting cluster represents similar images and can be labeled by the user in one go. The labeling itself is still performed manually, since an unsupervised clustering may result in heterogeneous clusters, i.e. clusters containing errors. However, the necessary interaction can be dramatically reduced.

We implemented a semi-automatic approach, using a self-organizing map (SOM) to cluster image data. Based on this system we developed concepts for investigating and labeling the clustered image data set. For this purpose, we provide a sophisticated user interface which allows for an easy and intuitive labeling of images and gives a good overview of the complete data set, and especially of the parts that still need to be labeled. We conducted a user study with 24 participants where the users were asked to annotate three different data sets. The goal of the user study was to find out whether the developed concepts simplify the interactive labeling task for the user. Based on the study results further concepts were developed to optimize the user experience.

2 Related Work

The creation of ground truth data sets has been neglected as a research topic for a long time. The idea of human computation has led to the development of a few tools for the creation of general purpose ground truth data sets, like the ESP Game [1], Peekaboom [2], and systems by Ho et al. [3] and Seneviratne et al. [4]. People participate in these games because they are entertaining, not because of the actual benefit. Russell et al. provide a web-based interface (LabelMe) for labeling images [5]. The incentive for contributing is the data itself. Similarly, Yao et al. [6] provide an annotation tool for the creation of a large ground truth data set. A semi-supervised labeling process is realized through a hierarchical segmentation of images. All of these approaches aim at the creation of general purpose data sets. They are, however, not applicable to the task of labeling specific data sets which include less variance.

In the area of content based image retrieval, Koskela et al. [7] developed a system for the automated semantic annotation of images based on the PicSOM system [8]. An existing ground truth data set is used to annotate other images in the data set. Heidemann et al. developed VALT (Visualization and Labeling Toolkit) [9] for labeling images clustered with a self-organizing map. Category labels can be assigned to nodes or individual images. However, the system was intended for AR scenarios.

A variety of systems and techniques for browsing large digital photo collections have been proposed. Most photo browsing applications use hierarchical representations. Among the best known applications are PhotoMesa [10], PhotoTOC [11], and CAT [12]. All three cluster images in order to provide a suitable interface for browsing. Both PhotoMesa and CAT use hierarchical structures and representative images for clusters in each hierarchy.

Despite these developments the problem of labeling large specialized data sets has not yet been solved. Therefore concepts are necessary to handle the complexity of large data sets without the possibilities of classical human computation approaches.

3 Data Preprocessing and Self-Organizing Map

For conducting our user study we extracted color histogram features as the input for clustering. However, any other type of features may be used. The feature vectors are processed by a standard SOM as proposed by Kohonen [13]. The SOM Toolbox for Matlab[14] is used for SOM calculation.

A SOM is a type of unsupervised neural network which projects high-dimensional input data to a low-dimensional (usually 2D) map (or grid). The input to a SOM are *n*-dimensional feature vectors. Each unit of the SOM grid is associated with one weight vector of the same size as the input vectors. The SOM is trained by iteratively calculating the best matching unit (BMU), i.e. the unit with the smallest euclidean distance to an input vector. The weight vectors of the BMU and its neighboring units are adapted towards the input vector. The shape and size of the neighborhood, as well as the learning rate, is controlled with a neighborhood function, usually a Gaussian.

An important result of this training process is that SOMs are topology preserving, i.e. data vectors that are close in the input space will be close in the resulting map. This aspect makes SOMs extremely valuable in data exploration tasks.

4 Interaction and Visualization Concepts

Figure 1a) displays the visualization of a trained SOM using a representative image for each node. A representative image is taken from the set of images which were assigned to this node. Additionally, we implemented a second visualization using the Unified Distance Matrix (U-Matrix) [15] (see Figure 1b)). Basically, the U-Matrix is a color-coded visualization of the mean distance to neighboring map units. The U-Matrix is widely used for visualizing clusters (dense areas) on trained SOMs. Therefore we wanted to investigate whether the color-coding is also suitable if image data is used. The arrangement of images is the same for both visualizations.

As mentioned, the quality of the ground truth data is crucial for training classifiers. Since the SOM clusters the data unsupervised, it is important that it is still possible for users to take a closer look at and label each image individually. Therefore, the developed user interface is zoomable and allows the user to view the SOM representation in both levels of detail (LOD), the map view and the detail view.

Figure 1a) shows the SOM where each unit is displayed by a representative image, c) shows the detail view of three selected units. In the detail view all images which were projected onto these map units are displayed. The utilization of more levels of detail were considered but not realized due to the increasing complexity and the inferior overview of the data set. In general more levels of hierarchy are preferable for image browsing tasks, however, if every single image needs to be processed by the user, a simple structure is desirable.

Although the SOM visualization itself has two LODs only, the interface provides a natural zooming behavior, since images move apart as the zoom factor is increased. However, the images are not scaled because larger images would require the user to draw wider selection areas.

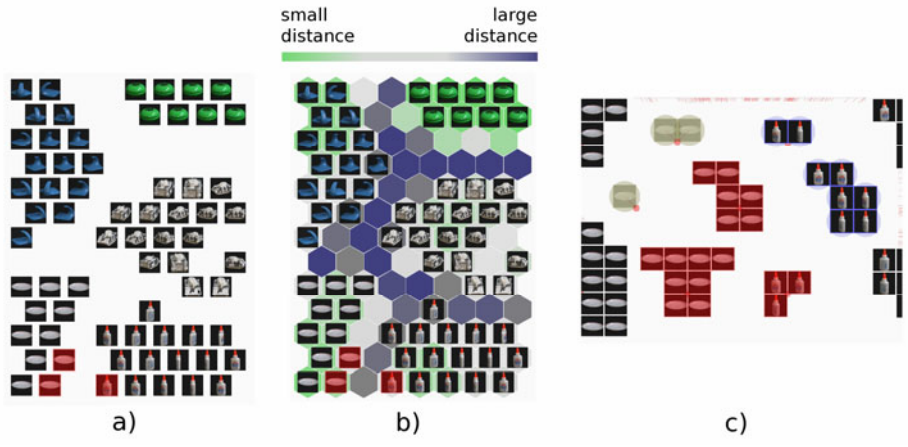

Fig. 1. SOM clustered visualization of COIL-100 subset [16]. a) shows the plain standard visualization of the map view, b) displays the color coded U-Matrix in the background on the hexagonal map. Selected map units are highlighted in red. c) shows the detail view of the area with the selected map units. Grayed out images are labeled, as well as images with blue boundaries. Subsidiary lines are displayed on upper boundary of c).

Inherent to a zoomable user interface for labeling large image data sets is the fact that only a subset of all images can be displayed at once. To simplify the navigation subsidiary lines were introduced to indicate the direction of remaining unlabeled images outside the current viewport (see Figure 1c)).

The user interface allows the direct selection of images, as well as the use of a rectangle and a free selection (lasso) tool. There is no difference in the selection mechanism of units or individual images in the detail view. If a node is selected and labeled, the label is assigned to all images attached to this node. However, in the detail view images may be selected and labeled individually, thus allowing the user to correct possible clustering errors.

Users can define labels in the interface. For each label a button with the name and a customized color is displayed in a separate panel. Labeled images are highlighted in the color of the assigned label and are drawn semitransparent as shown in Figure 1c) for plate images (gray). The images are displayed semitransparent because they are no longer selectable, thus simplifying the selection of the remaining images. In case a user wants to reassign or remove a label from an image, the visibility may be toggled and the images are again drawn opaque, though still with highlighted boundary, and can be selected. Toggling the visibility refers to all images assigned to the selected category label, as in Figure 1c) where all images assigned to *bottle* are opaque, i.e. selectable. This function was also intended to simplify the examination of all images that were assigned to a certain category. If only the images belonging to a certain category are displayed opaque, it is likely that incorrectly assigned images stand out visually and can be detected more easily in a final check. An option to hide all labeled images was included in the interface to further reduce visual clutter.

5 Evaluation

The concept of an unsupervised clustering with a SOM as the basis for a zoomable user interface is quite intuitive and promising. Nevertheless, unsupervised learning algorithms provide no guarantee that the resulting clusters agree with the users' expectations. A nice clustering in the map view might give users a false sense of security regarding the correctness of the unsupervised arrangement. Incorrectly clustered images may be missed by users and lead to inaccurate ground truth data sets. Additionally, it is important to investigate whether our proposed concepts are actually suitable for simplifying and accelerating the labeling task. For comparison, a third visualization was realized, which will be referred to as "unsorted representation".

The study comprises different labeling tasks. Table 1 shows a summary with the data sets used. In data sets 1 and 2 the number of images for individual objects varies from 1 to 72 and was chosen randomly. Data set 3 is a real data set with webcam images showing windows.

The sequence of the tasks was the same for all participants. The tests within each task, i.e. the different visualizations, were randomized to compensate for learning effects.

Table 1. Summary of labeling tasks and the data sets used. The same visualizations were tested in T2 and T3. The discussion of hypothesis 1-3 can be found in Section 5.2.

Task	Data set	#Images	#Labels	Visualizations	Hypotheses
T1	COIL-100	432	5	plain SOM	H3
T2	COIL-100	857	18	Unsorted, SOM, U-Matrix	H1, H2
T3	Webcam/ Windows	666	2	Unsorted, SOM, U-Matrix	H1, H2

5.1 Study Method and Procedure

The user study we conducted used a within-subjects design. In this study, we investigated the usability of the SOM based labeling interface, both with and without the U-Matrix visualization. Since no standard approach exists for labeling specific image data sets the visualization denoted as unsorted representation was used as baseline. The behavior of the user interface is exactly the same as for the SOM-based visualizations except that there is only one level of hierarchy. Instead all images are visible in one plane, exactly in a quadratic grid aligned by rows. Images were not randomized but instead displayed according to their filenames. The COIL set is special due to the labels being embedded in the filenames. Labeling of this data could be simplified using an explorer like interface. However, the COIL data is well suited for presenting the functionality of the system. In contrast, the data set in T3 did not include such a structure.

24 persons participated in the study (21 male, 3 female). 20 participants are computer scientists, four participants are students. The average age was 29 years (min. 16, max. 50). Ten participants were familiar with image recognition. All participants reported normal or corrected-to-normal vision. Two participants suffer from dyschromatopsia. Their results were within range of the other results and are therefore included in all calculations.

The study began with every participant answering personal statistical questions. The participants were then introduced to the topic of the study in a 15min tutorial. The tutorial included the SOM visualizations and the unsorted representation. Participants were asked to use the GUI in order to get familiar with the controls.

Completion times for individual tests and the assigned labels were recorded. After each test (each visualization), participants were asked to judge the simplicity (Q1) and speed (Q2) on a 7-point Likert scale (1 -- strongly disagree, 7 -- strongly agree). The last part of the evaluation consisted of general questions about the user interface.

Participants needed approximately 1.5 hours to complete the study. Labels were predefined for all tasks. In three cases participants interchanged labels for whole categories. These labels were manually corrected since they were caused by mere misunderstanding.

Table 2. *p*-values calculated using paired t-tests. Simplicity and speed were judged by participants on a 7-point Likert scale after each test. Average values are given for all parameters, standard deviation is given in brackets. Error rate calculated as η = #incorrectLabels / #images. Participants were asked for their favorite visualization after the completion of whole tasks.

	T2			T3		
	Unsorted	SOM	SOM w. U-Matrix	Unsorted	SOM	SOM w. U-Matrix
p	< 0.001			< 0.001		
duration (avg)	727s (237)	412s (146)	430s (131)	503s (147)	293s (144)	309s (121)
p	0.009			<0.001		
Q1 (avg)	4.31 (1.49)	5.17 (1.02)	5.57 (0.89)	3.35 (1.46)	4.95 (1.43)	5.05 (1.19)
p	<0.001			<0.001		
Q2 (avg)	2.7 (1.29)	5.35 (1.15)	5.74 (0.81)	2.8 (1.44)	5.25 (1.25)	5.05 (1.15)
η Error	0.0045	0.011	0.016	0.279	0.27	0.275
Favorite vis.	1	13	10	4	5	15

Experiment Setup. The tests were performed on two laptop computers. The first one has an Intel Core 2 Duo 2.8GHz with 4GB RAM, an NVIDIA GeForce 9600M GT, and a 15 inch display with 1440 x 900 pixels resolution. The second computer has an AMD Athlon II Dual Core M300 2.0 GHz with 4GB RAM, an ATI Mobility Radeon HD 4200, and a 15,6 inch display with 1366 x 768 pixels resolution. Custom laptops with standard display sizes were used to investigate the usability under usual conditions. The differences of the GPUs were regarded as negligible since only two-dimensional graphics are used.

Groups of 12 participants completed the study on each laptop. There was no statistically significant difference in the completion times or error rates of both groups. The following results, therefore, refer to all 24 participants.

5.2 Hypotheses

The evaluation was conducted to test a total of three hypotheses. The tests and the evaluation for each hypothesis are given individually in the following sections.

H1: Arranging images with a SOM simplifies the labeling task. This section investigates the unsorted representation and the plain SOM visualization. Considering the completion times in tasks 2 and 3 (see Table 2), the results are statistically significant. In both tasks the average time to complete the labeling using the SOM visualization was less than 60% of the time needed for the unsorted representation.

Additionally, the subjective preferences (Q1 and Q2) in both tasks clearly favor the SOM visualization. The large differences in the judgment of speed are most probably due to the fact that users did not have to look at every single image.

In T2 the SOM visualization exhibits a larger error rate than the unsorted representation. In the unsorted representation all images are displayed in a two-dimensional plane without any hierarchy. Every image is therefore immediately visible, which leads to only a few incorrectly labeled images. The SOM clustering contained a few nodes that were not correctly clustered. Several participants chose to trust the SOM clustering and neglected the detail view, while others did look at the detail view but missed some errors anyway. In T3, the SOM visualization exhibits a slightly lower error rate. This may be due to the SOM representation reaching a very good separation of closed and opened windows.

H2: Visualizing the U-Matrix helps identifying clusters. It is not possible to get statistically significant results for both SOM based visualizations due to their inherent similarity. Nevertheless, to give an overview of the results, Table 2 gives the average and standard deviation of the duration, as well as for the questions about simplicity and speed, and the relative error.

When asked for their favorite visualization at the end of T2 and T3 the U-Matrix version was mentioned 25 times (see Table 2). In the final interview all participants were asked whether they found the U-Matrix visualization helpful. 15 participants answered that they did not look at the colors or did not care about them, but found the hexagonal grid structure helpful for navigating inside the data set. The grid helped participants to estimate the current zoom level since hexagons scale as the zoom factor is increased. Three participants mentioned that they found the colors distracting but liked the hexagonal structure. Only four participants stated that they found the distance encoding at least partially helpful. Since images themselves are easy to grasp it is not surprising that the U-Matrix color is of assistance in specific cases only.

H3: Drawing labeled images semitransparent and making them non-selectable enhances visual clarity. In T1 participants labeled the data set twice: once labeled images were drawn semitransparent and were non-selectable, the other time labeled images were highlighted and drawn opaque. The two representations were described above in the context of toggling the visibility of labeled images. In this test the representations were used exclusively and could not be switched.

The clustering in this example included no errors and the separation of clusters was clearly visible. When asked about their impression 14 participants answered that they preferred the version with semitransparent and non-selectable images because it provided a better overview of remaining unlabeled images. It was mentioned 8 times that the selection of remaining images was simpler if labeled images were non-selectable. 6 participants felt there was no significant difference in using both versions.

These results imply that the semitransparent representation of labeled images provides a benefit. However, participants used the option to hide all labeled images in a majority of the following tests. We therefore conclude that the additional colors present visual clutter even if it is less strong.

5.3 Further Enhancements

Although participants clearly favored SOM based visualizations over the unsorted representation, the risk of a higher error rate cannot be underestimated. One major drawback is the representation of nodes through one random image. We therefore developed four alternative representations which allow users to detect clustering errors more easily. These representations are average images, eigenimages and subset images. Subset images were calculated from the first four eigenvectors for one representation. Another subset image is calculated using *k*-means clustering on each unit ($k = 4$) and combines images closest to the cluster centers. Examples are given in Figure 2.

Although all representations allow the detection of clustering errors, the average images were not appreciated by users. Representations based on eigenimages were experienced as least suitable. Although clustering errors of the SOM are well visible, they complicate the labeling of correctly clustered nodes, since the object cannot directly be identified (see Figure 2a-b)). The most intuitive and helpful representation proved to be the subset image based on *k*-means clustering. It displays the images in a very intuitive and familiar way and allows easy identification of incorrect clusters. By deploying these representations another disadvantage was eliminated. Users are now able to grasp their position inside the hierarchy, since it is obvious whether the map or detail view is displayed.

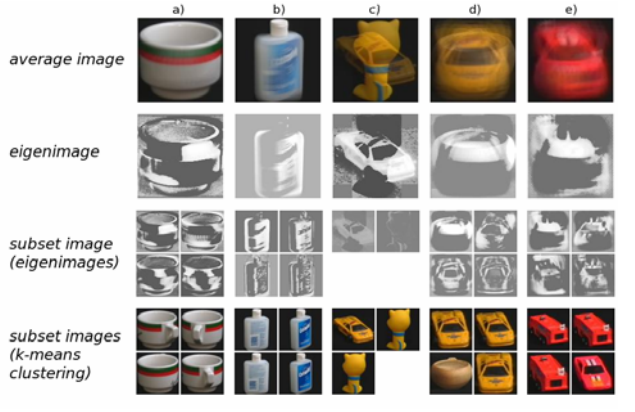

Fig. 2. Alternative representations for map units. Average images allow the detection of clustering errors, but cause a nauseating feeling, Eigenimages complicate the identification of the actual object. Subset images from k-means clustering seem intuitive and display four very different images belonging to this unit. Identical objects as in a) and b) can be grasped immediately by users.

Another improvement was the adaption of the U-Matrix color scheme. Initially a continuous red to blue gradient was used, which was impractical and too distracting. Instead, only very dense and very sparse areas are highlighted. Medium distances are displayed in neutral gray to reduce the amount of information transported via color.

6 Conclusion

We developed a graphical user interface concept for the interactive semi-automatic labeling of large image data sets. The interface uses a clustering calculated by a self-organized map (SOM) as the basis for a visualization, with or without the U-Matrix to transport additional information about the node distances.

The interface was developed to investigate concepts for optimizing the usability when labeling large image data sets. This was realized by drawing labeled images semitransparent and making them non-selectable. The zoomable interface provides an overview of the SOM and still lets the user explore all images belonging to one unit. To investigate whether the SOM based visualizations are applicable to real image labeling tasks a user study with 24 participants was conducted. Although the SOM visualizations yield a risk for a slightly higher error rate, the usability was preferred over labeling images sorted by filename. SOM visualizations achieved shorter completion times and were subjectively experienced as faster and easier to use. The drawback of the SOM visualizations is that they might yield a false sense of security regarding the unsupervised clustering. Therefore, alternative representations for node images were developed which reduce the risk of missing clustering errors. Additionally, the U-Matrix color coding was adapted to better fit user needs.

Future work will include research on whether existing visualizations used in SOM based data exploration, like hit histograms or component planes, can be of advantage in supporting the image labeling task. A major topic will also be the integration of user feedback for an iterative recalculation of the SOM.

We believe the developed concepts will become more important in the future as the need to annotate large image data sets increases. Nevertheless, it will be necessary to extend the proposed concepts in the future to further increase the usability of labeling systems, especially for very large data sets.

References

1. von Ahn, L., Dabbish, L.: Labeling Images with a Computer Game. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 319–326 (2004)
2. von Ahn, L., Riu, R., Blum, M.: Peekaboom: a Game for Locating Objects in Images. In: SIGCHI Conference on Human Factors in Computing Systems (CHI 2006), pp. 55–64 (2006)
3. Ho, C.-J., Chang, T.-H., Lee, J.-C., Hsu, J.Y.-J., Chen, K.T.: KissKissBan: a Competitive Human Computation Game for Image Annotation. In: ACM SIGKDD Workshop on Human Computation, pp. 11–14 (2009)
4. Seneviratne, L., Izquierdo, E.: An Interactive Framework for Image Annotation through Gaming. In: International Conference on Multimedia Information Retrieval, pp. 517–526 (2010)

5. Russel, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: A Database and Web-based tool for image annotation. Int. J. on Computer Vision 77(1), 157–173 (2008)
6. Yao, B., Yang, X., Zhu, S.-C.: Introduction to a Large-Scale general Purpose Ground Truth Database: Methodology, Annotation Tool and Benchmarks. In: International Conference on Energy Minimization Methods in Computer Vision and Pattern Recgnition, pp. 169–183 (2007)
7. Koskela, M., Laaksonen, J.: Semantic Annotation of Image Groups with Self-Organizing Maps. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 518–527. Springer, Heidelberg (2005)
8. Laaksonen, J., Koskela, M., Laakso, S., Oja, E.: PicSOM – Content-based Image Retrieval with Self-Organizing Maps. Pattern Recognition Letters
9. Heidemann, G., Saalbach, A., Ritter, H.: Semi-Automatic Acquisition and Labelling of Image Data using SOMs. In: European Symposium on Artificial Neural Networks, pp. 503–508 (2003)
10. Bederson, B.: PhotoMesa: a Zoomable Image Browser using Quantum Treemaps and Bubblemaps. In: ACM Symposium on User Interface Software and Technology, pp. 71–80 (2001)
11. Platt, J., Czerwinski, M., Field, B.: PhotoTOC: Automatic Clustering for Browsing Personal Photographs. In: Joint Conference of the International Conference on Information, Communications and Signal Processing, and the Pacific Rim Conference on Multimedia, vol. 1, pp. 6–10 (2003)
12. Gomi, A., Miyazaki, R., Itoh, T., Li, J.: CAT: a Hierarchical Image Browser using a Rectangle Packing Technique. In: International Conference on Information Visualisation, pp. 82–87 (2008)
13. Kohonen, T.: The Self-Organizing Map. Proceedings of the IEEE 78(9), 1464–1480 (1990)
14. Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J.: Self-Organizing Map in Matlab: the SOM toolbox. In: Matlab DSP Conference, pp. 35–40 (1999)
15. Ultsch, A., Sieman, H.P.: Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. In: Proceedings of International Neural Networks Conference, pp. 305–308. Kluwer Academic Press, Dordrecht (1990)
16. Nene, S., Nayar, S., Murase, H.: Columbia Object Image Library (COIL-100). Technical report, Columbia University (1996)