Computational Visualization Interpretation

Jeffrey Heer, Fei-Fei Li Stanford Computer Science Seed Grant Proposal – Spring 2010

Visual depictions of data such as charts, graphs, and statistical maps are a ubiquitous means of representing structured information. For at least the last 300 years, these visualizations have been a primary vehicle for communicating scientific and statistical results. While well-designed data graphics leverage human visual processing to more effectively convey data [27], in many cases they are the *only* publicly available representation of the underlying data, hindering reuse and analysis. This is unfortunate, as hundreds of publications (both printed and online) present data graphically, not to mention centuries of books and articles.

In this project, we propose to develop computational models of graphical perception to extract data from twodimensional information graphics. We seek to combine computer vision techniques with models of chart decoding developed by psychologists, statisticians, and computer scientists. Our goal is to enable the automatic extraction of data from visualization images. We will build a corpus of examples for training and testing models, investigate textual and image features for chart recognition, and develop routines for data extraction from recognized visualization types. As perfect recognition is unlikely, we will also investigate visualization and interaction techniques for understanding and correcting our computational models, enabling interactive data extraction and incorporating active learning techniques to allow human input in the training loop.

Our research could enable a variety of compelling applications. Imagine a climate scientist reviewing prior work on atmospheric carbon dioxide levels—with our techniques, she could extract data from a relevant chart embedded in a PDF file and compare that to her own results. Alternatively, a social entrepreneur reading The Economist might view growth predictions for developing regions, take a picture of a line chart with his mobile phone, annotate a region of interest, and e-mail extracted data to himself for further use. We might also apply our techniques at scale, extracting data from publication archives or scanned books, either automatically or with human supervision via crowdsourcing. Other applications include *retargeting*—extracting data and then re-visualizing it using a more effective representation—and *visualization evaluation*—our models may also indicate which visualizations will be easier for human viewers to decode, aiding design optimization.

In short, we seek to develop general techniques for computational extraction of data from information graphics and apply these techniques to benefit domains such as scientific research, public policy, education, and design. In the spirit of the call for proposals, we believe that socially-relevant domains such as health, energy, education, and the environment could all benefit from improved access to data contained in published data graphics. In addition, the project will spur collaboration among the department's Vision, Visualization, and HCI groups.

Motivation and Prior Work

Our research will draw on prior work in both perceptual psychology—particularly the study of *graphical perception*, or how people decode information in graphs—and computer vision techniques.

Psychological Models of Graphical Perception

A great deal of prior research has investigated how visual variables such as position, length, area, shape, and color impact the effectiveness of data visualizations. Inspired by Bertin's [1] systematic identification of visual variables, researchers in cartography [10,16,21], statistics [4,5,25], and computer science [17] have derived perceptually-motivated rankings of the effectiveness of the visual variables for encoding nominal, ordinal and quantitative data. These rankings were initially based on psychophysical models of human perception such as the Weber-Fechner law [19, 27] and Steven's Power Law [19, 24]. For example, the latter predicts that position encodings are more accurate than 2D area encodings and that area encodings are more accurate than 3D volumetric encoding techniques (e.g., [3, 14, 22, 23, 25]). These experiments typically measure how each visual encoding variable affects the accuracy and/or response time of value comparisons of the underlying data. For example, Cleveland and McGill [5] showed subjects several types of charts including divided bar charts, stacked bar charts and pie charts and asked them to compare the values encoded by two of the marks. They measured response accuracy

and found that position encoding along an aligned scale resulted in higher accuracy than length judgments, while angle judgments, as found in a pie chart, were more error-prone than position judgments.

Other researchers have formulated cognitive models of graph perception at varying levels of detail. Kosslyn's [13] model of chart perception identifies three levels of analysis: syntax, semantics, and pragmatics. Syntactic analysis requires identifying the distinct visual marks and categorizing them as labels, framing elements such as axes, or representations of the data. Semantic analysis involves associating the syntactic properties with the data they represent and turning perceptual inferences into statements about the data. Pragmatic analysis focuses on connotations beyond the semantic interpretation. Pinker's [20] model describes a process of decoding visual information into a propositional language that is matched against a graph schema representing one's familiarity with the chart type. Unfamiliar graphics require inferring this graph schema and are therefore harder to decode. Other models further divide chart decoding into simple perceptual operations and use these models to predict performance. Lohse [15] formulated a timing model for chart perception using cognitive operations such as move eyes, identify shape, retrieve memory, etc. Simkin and Hastie [22] posit that low-level graph perception can be described in terms of mental image operations such as anchoring, scanning, projection, and super-imposition. Gillan and Lewis [9] propose a Mixed Arithmetic-Perception (MA-P) model that combines perceptual inferences with mental arithmetic. The model predicts that the most effective visualizations minimize the need for mental math. However, these models are largely abstract; while some provide computational performance models, none specify concrete mechanisms for identifying syntactic visual features from input images.

Computational Perception of Data Graphics

A handful of computer graphics and computer vision researchers have attempted to automatically identify chart types and extract encoded data. For example, Zhou et al. [28] apply the Hough transform on an image edge map to model and extract data in bar charts. Huang et al. [11] apply off-the-shelf OCR software to identify text and image regions. They then strip all text from image regions, vectorize the image edge map, and apply a hand-constructed chart-specific procedure (i.e., for bar, line, and pie chart) to read off data values. More recent work has focused solely on the problem of accurately classifying a chart type [12, 26], similar in spirit to Pinker's concept of identifying a graph schema. For example, Vitaladevuni et al. [26] construct low-level image features such as SIFT and HOG (Histogram of Oriented Gradients) descriptors which they then classify using multi-class Support Vector Machines. They achieve ~80% accuracy across five classes: bar, line, pie, scatter, and surface plots. Misclassifications commonly occur when multiple encoding types are used, such as a regression line in a scatter plot or a Pareto curve plotted over a bar chart. As described in the next section, we believe we can achieve better results by creating models of graphical perception informed by psychological theories of chart decoding.

Research Plan

In this section, we present our preliminary research plan. While we intend to follow the specified path, we will of course evaluate progress and re-assess our strategy as the research progresses. We plan to start with simple data graphics first (e.g., bar charts, scatter plots, and line charts without overplotting), increasing both the sophistication of the charts (e.g., adding error bars, multiple encoding variables, etc) and the diversity of chart types (e.g., processing map displays) as the research progresses. Though comprehensive treatment of each of the steps below may be beyond the scope of this seed grant, we hope to jumpstart continuing research on this topic.

Corpus Generation. Our initial task will be to create a corpus of chart images and corresponding data for the purpose of training and testing perceptual models. We will create a diverse corpus of 2D data graphics, including standard charts and graphs, maps, and network diagrams. First, we will generate our own collection of chart types using the Stanford Protovis toolkit [2]. This step is important because it will provide a corpus for which we have perfect knowledge of both the data and geometry of the chart. We will also introduce distortions to the generated images (e.g., shearing, blurring, noise, contrast and brightness adjustment) to create more difficult test cases. Second, we will construct a corpus of real-world examples from the web, books, and publication archives. We will use crowdsourcing methods such as Amazon's Mechanical Turk to collect various examples. We plan to collect at least hundreds of images for each visualization type. Furthermore, all relevant examples will be stored and categorized in the Stanford ImageNet database [7], making the results accessible to other interested researchers.

Visualization Type Recognition. Existing approaches to chart type recognition use seemingly sensible categories such as bar chart, line chart, etc. However, we contend that this classification is too coarse-grained and at odds

with psychological research. A more deep-seated distinction is the recognition of framing elements: the use of metric spaces such as Cartesian and radial coordinates (c.f., Kosslyn [13] and Pinker [20]), or of container spaces that use Gestalt grouping features such as enclosure or connectivity to encode relationships among data. Once the suitable frame has been recognized, it provides a prior for how decoding should proceed. Our research will include an investigation of the image features indicative of these visualization types, with a focus towards those deemed semantically meaningful in prior cognitive models. We will especially draw upon our knowledge in scene and texture classification algorithms [8] for classifying encoding types. Oddly, existing techniques (e.g., [11,26]) explicitly ignore textual cues in the data. In addition, we intend to use existing OCR techniques to identify text in the display; we hypothesize that both label positions (e.g., regularly spaced along a Cartesian grid) and content (e.g., numbers vs. letters or dates) provide salient features for identifying both visual encoding and data types.

Data and Reference Element Recognition. Once the visualization type has been established (either automatically or via user input), type-specific strategies can be applied. A next step is to separate reference elements such as gridlines and labels from data representative marks. We will explore a combination of edge detection, image segmentation techniques, and shape matching approaches for this purpose. We will attempt to classify common shape types such as bars, wedges, and poly-lines which represent data in the display. In the case of maps, we plan to apply shape matching techniques to learn common outlines for geographic maps, such as state and country boundaries. We also will develop models to identify legends, e.g., to identify color or shape encodings.

Data Extraction. Once framing elements and data representative marks have been separated, we can attempt to interpret the data display. For example, in a Cartesian coordinate system, point positions, bar extents, and polyline vertices can be interpreted in terms of normalized [0,1] image coordinates. These extracted coordinates can then be mapped through a scale (either user-provided or automatically extrapolated from label text via OCR) to determine data values. We also plan to model extraction error; as a single pixel may cover multiple values in the data space, our extraction process should include this uncertainty. Similar strategies may be applied for other visual variables such as angle, area, color hue, and luminance. Our model will incorporate contextual and prior knowledge based on the visualization type we have identified. For example, the locations of relevant data points differ depending on whether we are analyzing a bar chart or a line chart. By incorporating such contextual information, we hope to improve significantly the recognition and extraction accuracy.

Interaction and Error Handling. We may find it difficult to achieve high accuracy in all cases. Accordingly, we will also investigate interaction techniques for aiding ambiguous classifications or faulty recognition. We will construct visualizations that depict the model's current assessment of recognized text and graphical marks. Users will be able to manually indicate the visualization type, encoding variables of interest, and also select erroneously recognized elements and provide corrective feedback, such as selecting an alternate interpretation from an N-best list. Based on this user input, we hope to use active learning methods to improve our models' recognition ability on difficult cases [6]. We expect our interface will also provide a useful debugging tool as we refine our models. If successful, we hope to also explore the effectiveness of using the interface to crowdsource human corrective feedback. The goal would be to achieve greater scalability of batch data extraction, as when processing publication archives.

Applications

Interactive Data Extraction and Visualization Retargeting. A first application of our techniques is an interactive tool for data extraction and visualization retargeting. A user may find a chart in an image file or PDF, or take a picture using their mobile phone. We will apply our model to extract data from the chart. Once the data has been extracted, we can further use the data to generate a space of alternative visualizations by applying automated visualization design algorithms [17, 18].

Extracting Data from Archival Publications. A long-term goal of this research is to "bring to life" data graphics contained in thousands of old books and articles. We hope to assess the feasibility of providing automated or semiautomated tools for extracting data sets from previous published charts. A useful starting point will be existing accessible publication archives such as the ACM Digital Library.

Visualization Evaluation Tools. In addition to extracting data from charts, we will explore the use of our techniques for evaluating visualization designs. For example, a line chart with insufficient contrast between gridlines and plotted data may result in inaccurate extraction. Similarly, overplotting will make it difficult to extract

values. Human viewers may experience similar difficulties, and so we plan to assess to what degree our models can be applied to rank-order multiple design alternatives. Note that this may result in a compelling use case even for models exhibiting insufficient accuracy for unsupervised extraction.

Budget and Work Plan

We seek seed funding for a 1-2 year period beginning Fall 2010. We ask for a budget of \$100k to support two graduate students (either PhD or MS): one specializing primarily in computer vision and another in visualization and HCI. We will hold weekly research meetings in which we will share and evaluate research progress as well as read and discuss relevant related work. In addition to resulting publications, our primary deliverable will be an interactive proof-of-concept system for extracting data for multiple visualization types.

References

- [1] J. Bertin. *Semiologie Graphique*.Gauthier-Villars, 1967. English translation by W.J. Berg as Semiology of Graphics, University of Wisconsin Press: Madison, WI, 1983.
- [2] M. Bostock and J. Heer. Protovis: A graphical toolkit for visualization. IEEE TVCG, 15(6):1121–1128, 2009.
- [3] C. A. Brewer. Color use guidelines for data representation. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*, pp. 55–60, 1999.
- [4] W. S. Cleveland, P. Diaconis, and R. McGill. Variables on scatterplots look more highly correlated when the scales are increased. *Science*, 216:1138–1141, Jun 1982.
- [5] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79:531–554, 1984.
- [6] B. Collins, J. Deng, L. Kai, and L. Fei-Fei. Towards scalable dataset construction: An active learning approach. *European Conference on Computer Vision (ECCV)*, 2008.
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR, 2009.
- [8] L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. IEEE Computer Vision & Pattern Recognition (CVPR), 2005.
- [9] D. J. Gillan and R. Lewis. A componential model of human interaction with graphs: 1. Linear regression modeling. *Human Factors*, (36):419–440, 1994.
- [10] P. P. Gilmartin. Influences of map context on circle perception. Annals of the Association of American Geographers, 71:253–258, 1981.
- [11] W. Huang, C.L. Tan, and W.K. Leow. Model-based chart image recognition. In *Graphics Recognition*, pp. 87–99, 2004.
- [12] W.Huang, S. Zong, and C.L. Tan. Chart image classification using multiple-instance learning. In *IEEE Workshop on Applications of Computer Vision*, 2007.
- [13] S. M. Kosslyn. Understanding charts and graphs. Applied Cognitive Psychology, 3:185–225, Jul-Sep 1989.
- [14] S. Lewandowsky and I. Spence. Discriminating strata in scatterplots. *Journal of the American Statistical Association*, 84:682–688, Sep 1989.
- [15] J. A. Lohse. Cognitive model for the perception and understanding of graphs. In ACM CHI, pp. 137–144, Apr/May1991.
- [16] A. M. MacEachren. How Maps Work: Representation, Visualization, and Design. Guilford, 1995.
- [17] J. D. Mackinlay. Automating the design of graphical presentations of relational information. *ACM TOG*, 5:110–141, 1986.
- [18] J. D. Mackinlay, P. Hanrahan, and C. Stolte. Show Me: Automatic presentation for visual analysis. *IEEE TVCG*, 13:1137–1144, Nov/Dec 2007.
- [19] S. Palmer. Vision Science: Photons to Phenomenology. MIT Press, 1999.
- [20] S. Pinker. A theory of graph comprehension. In Artificial intelligence and the future of testing, pp. 73–126. LEA, 1990.
- [21] B. G. Shortridge. Stimulus processing models from psychology: Can we use them in cartography? *The American Cartographer*, 9:155–167, 1982.
- [22] D. Simkin and R. Hastie. An information-processing analysis of graph perception. *Journal of the American Statistical Association*, 82:454–465, Jun 1987.
- [23] I. Spence and S. Lewandowsky. Displaying proportions and percentages. Applied Cognitive Psychology, 5:61–77, 1991.
- [24] S. S. Stevens. The psychophysics of sensory function. American Scientist, 48(2):226–253, 1960.
- [25] L. Tremmel. The visual separability of plotting symbols in scatterplots. J. Comp. & Graph. Statistics, 4:101–112, Jun 1995.
- [26] S. N. Vitaladevuni, B. Siddiquie, J. Golbeck, and L. S. Davis. Classifying computer generated charts. In CBMI Workshop, 2007.
- [27] C. Ware. Information Visualization: Perception for Design. Morgan-Kaufmann, 2ndedition, 2004.
- [28] Y. P. Zhou and C. L. Tan. Bar chart recognition using Hough-based syntactic segmentation. In *Theory and Application of Diagrams*, volume 1889, pp. 494–497, 2000.