Omnidirectional 3D Visualization for the Analysis of Large-scale Textual Corpora: Tripitaka Koreana

Dr Sarah Kenderdine
ALiVE, City University
Hong Kong SAR,
China
skenderd@cityu.edu.hk

Prof Lew Lancaster
ECAI, Berkeley
University
San Francisco
buddhst@berkeley.edu

Howie Lan
ECAI, Berkeley
University
San Francisco
howielan@
socrates.berkeley.edu

Tobias Gremmler
City University
Hong Kong SAR,
China
gremmler@
syncon-d.com

Abstract— **This paper presents the research and development of a new omnispatial visualization framework for the collaborative interrogation of the world's largest textual canon, using the worlds' first panoramic stereoscopic visualization environment - the Advanced Visualization and Interaction Environment (AVIE). The work is being undertaken at a new research facility, The Applied Laboratory for Interactive Visualization and Embodiment (ALiVE), City University of Hong Kong. The dataset used is the Chinese Buddhist Canon, Koryo version (Tripitaka Koreana) in classical Chinese, the largest single corpus with 52 million glyphs carved on 83,000 printing blocks in 13th century Korea. The digitized version of this Canon (a project led by Berkeley University) contains metadata that links to geospatial positions, contextual images of locations referenced in the text, and to the original rubbings of the wooden blocks. Each character has been abstracted to a 'blue dot' to enable rapid search and pattern visualization. Omnispatial refers to the ability to distribute this data in 360-degrees around the user where the virtually presented visual space is in three dimensions (3D). The project's omnidirectional interactive techniques for corpora representation and interrogation offer a unique framework for enhanced cognition and perception in the analysis of this dataset.**

Keywords-digital humanities; immersive visualization; visual analytics; computational linguistics

I. INTRODUCTION

Research into new modalities of visualizing data is essential for a world producing and consuming digital data (which is predominantly textual data) at unprecedented scales [22, 37]. Computational linguistics is providing many of the analytics tools required for the mining of digital texts (e.g. [43, 44]). The first international workshop for intelligent interface to text visualization recently took place in Hong Kong, 2010 [32]. In the last five years, the visual analytics field has grown exponentially and its core challenges for the upcoming five years are clearly articulated [20, 40, 45]. It has been recognized that existing techniques for interaction design in visual analytics rely upon visual metaphors developed more than a decade ago [24] such as dynamic graphs, charts, maps, and plots. Moreover, interactive, immersive and collaborative techniques to explore large-scale datasets lack adequate experimental development essential to the construction of knowledge in analytic discourse [40]. Recent visualization research is constrained to 2D desktop screens and the ensuing interactions of "clicking", "dragging" and "rotating" [32, 43]. Furthermore, the number of pixels available to the user is a critical limiting factor in the human cognition of data visualizations [22], resulting in the recent development of gigapixel displays (e.g. Powerwall, StarCave, see [7]).

The project described in this paper, *Blue Dots AVIE*, exploits the opportunities offered by immersive 3D techniques to enhance collaborative cognitive exploration and interrogation of high dimensional datasets within the domain of the digital humanities [18]. The research takes core challenges of visualizing this large-scale humanities data inside a unique 360-degree 3D interactive virtual environment - AVIE [1] to provide powerful modalities for an omnispatial exploration responding to the need for embodied interaction, knowledge-based interfaces, collaboration, cognition and perception [40]. The research is taking place at a new facility, Applied Laboratory for Interactive Visualization and Embodiment (ALiVE) located at the Hong Kong Science Park [4].

A. Visual analytics

This research project responds to core challenges and potentials identified in Visual Analytics [24, 48]. Visual analytics is a rapidly expanding field applied to business intelligence, market analysis, strategic controlling, security and risk management, health care and biotechnology, automotive industry, environmental and climate research, as well as other disciplines of natural, social, and economic sciences (see [9, 11, 24, 48]). Websites such as *Visual Complexity* [70], and *Flowing Data* [12], and mainstream projects such as *Many Eyes* [36], *GapMinder* [13] and, *Wordle* [50] attest to the increasing interest in information visualization by multiple disciplines.

B. Corpora visualization

Most previous work in text visualization focused on one of two areas, visualizing repetitions, and visualizing collocations. The former shows how frequently, and where, particular words are repeated, and the latter describes the characteristics of the linguistic "neighborhood" in which these words occur. Word clouds are a popular visualization technique whereby words are shown in font sizes corresponding to their frequencies in the document. It can also show changes in frequencies of words through time [17] and in different organizations [6], and emotions in different geographical locations [15]. The significance of a word also lies in the locations at which it occurs. Tools such as *TextArc* [39], *Blue Dots* [28 - 31] and *Arc Diagrams* [48] visualize these "word clusters" but are constrained by the small window size of a desktop monitor. In a concordance or "keyword-in-context" search, the user supplies one or more query words, and the search engine returns a list of sentence segments in which those words occur. IBM's *Many Eyes* displays the context with suffix trees, thereby visualizing the most frequent n-grams following a particular word [49]. In the digital humanities, words and text strings is the typical mode of representation of mass corpora. However new modes of lexical visualization are emerging such as *Visnomad* [46], a dynamic visualization tool for comparing one text with another, and the *Visualization of the Bible* (Figure 2) by Chris Harrison where each of the 63,779 cross references found in the Bible are depicted by a single arc whose color corresponds to the distance between the two chapters [16].

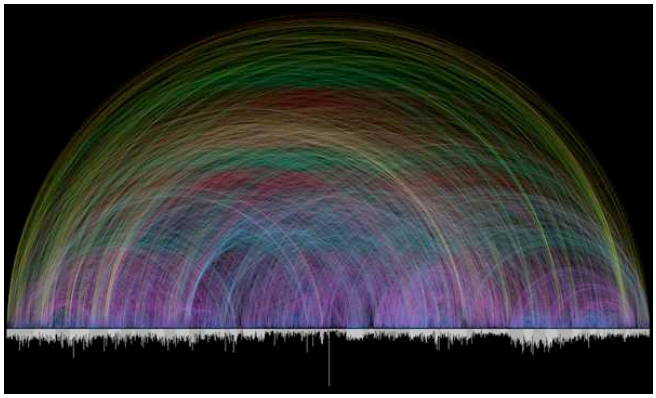


Figure 1. Visualizing the Bible, Chris Harrison (<http://www.chrisharrison.net/projects/bibleviz/index.html>).

C. Gesture based computing

The future for Visual Analytics is closely related to HCI and its development is related to gesture based computing for data retrieval [21]. Microsoft's *Project Natal* [42] and Pranav Mistry (MIT) *Six Sense* [41] are examples of increasing use of intuitive devices that promote kinesthetic embodied relationships with data.

D. High definition immersive visualization

Visualization systems for large-scale data sets are increasingly focused on effectively representing their many levels of complexity. This includes gigapixel tiled displays such as *HIPerSpace* at Calit2 [19] (Figure 1) and, next generation immersive virtual reality systems such as *StarCAVE* (UC San Diego) [7] and *Allosphere* (UC Santa Barabara) [2]. In the humanities, Cultural Analytics uses computer-based techniques for quantitative analysis and interactive visualization employed in sciences, to analyze massive multi-modal cultural data sets on gigapixels screens [35].

Figure 2. Using a unique 287 megapixel HIPerSpace at Calit2 (San Diego) for Manga research.

E. Advanced Visualization and Interaction Environment

The Advanced Visualization and Interaction Environment (AVIE [1]) is the UNSW iCinema Research Centre's landmark 360-degree stereoscopic interactive visualization environment spaces. [38] An updated active-stereo projection system together with camera tracking is installed at ALiVE and forms part of the core infrastructure for the Lab. The base configuration is a cylindrical projection screen 4 meters high and 10 meters in diameter, a 12-channel stereoscopic projection system and a 14.2 surround sound audio system (Figure 3). AVIE's immersive mixed reality capability articulates an embodied interactive relationship between the viewers and the projected information spaces. This system is used for the corpora visualization being undertaken at ALiVE.

In 2010, the social network visualizations of the Gaoseng Zhuan Corpus (biographies) [8] were integrated into the AVIE system as a prototype [26]. On the web, the social networks are displayed as spring loaded nodal points (JavaScript applet; Figure 4). However for immersive architectures such as AVIE, there is no intuitive way to visualizing such social network as a 2D relationship graph in the virtual 3D space provided by the architecture itself. Therefore this project proposed an effective mapping that projects the 2D relationship graph in to 3D virtual space and provide an immersive visualization and interaction system

for intuitive visualizing and exploring of social network using the immersive architecture AVIE. The basic concept is to map the center of the graph (the user) to the center of the virtual 3D world and project the graph horizontally on the ground of the virtual space. This mapping provides an immersive visualization such that the more related nodes will be closer to the center of the virtual world (where the user is standing and operating) Figure 5.

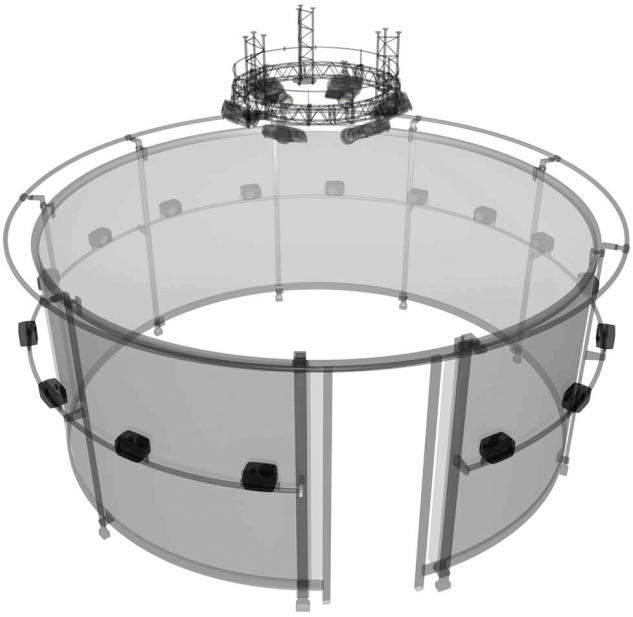


Figure 3. Advanced Visualization and Interaction Environment. Image © ALiVE, CityU

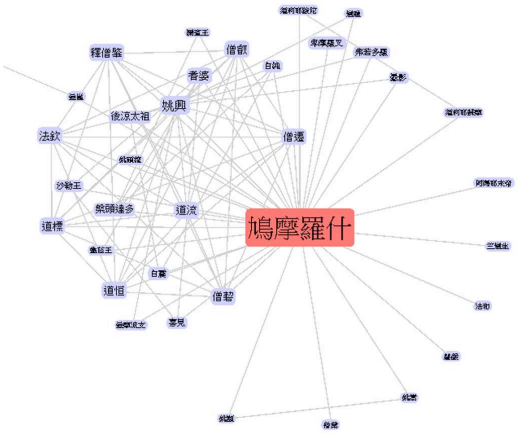


Figure 4. Social Networks of Eminent Buddhists JavaScript applet © Digital Archives, Dharma Drum Buddhist College.

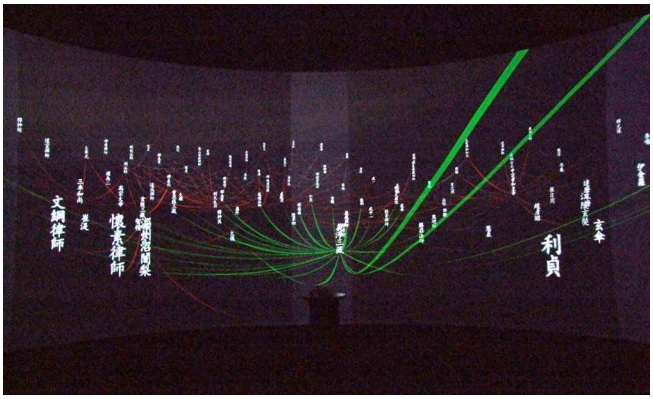


Figure 5. Social Networks of Eminent Buddhists, nodal networks distributed in 360-degrees AVIE © ALiVE, CityU.

II. THE BLUE DOTS

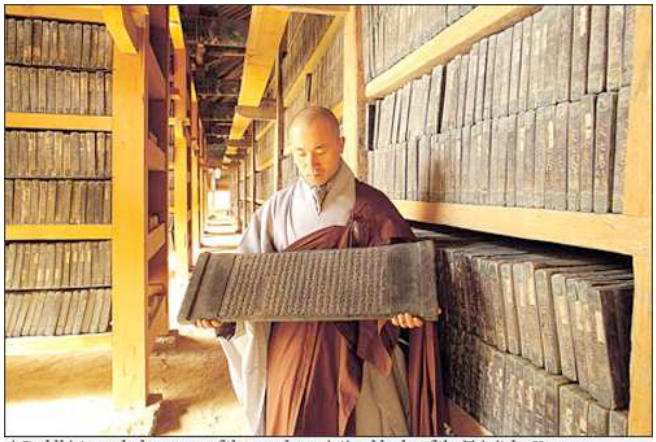
Blue Dots AVIE builds upon the *Blue Dots* visualization metaphor developed for the interrogation of the Buddhist Canon in which each character is represented as a blue dot [28 - 31]. This version of the Buddhist Cannon is inscribed as UNESCO World Heritage enshrined in Haeinsa, Korea. The 166,000 pages of rubbings from the wooden printing blocks constitute the oldest complete set of the corpus in print format (Figures 6 & 7). Divided into 1,514 individual texts the version has a complexity that is challenging since the texts represent translations from Indic languages into Chinese over a 1000-year period (2nd-11th centuries). This is the world's largest single corpus containing over 50 million glyphs and it was digitized and encoded by Prof Lew Lancaster and his team in a project that started in the 70s.

A. *Summary of content*

- 1.504 texts
- 160.465 pages
- 52.000.000 glyphs
- 1 text includes 107 pages (34674 glyphs)
- 1 page includes 324 glyphs arranged in 23 rows and 14 columns

B. *Contextual information*

- 1.504 colophons with titles, translators, dates, places, and other information.
- 202 people names (translators, authors, compilers)
- 98 monastery names



A Buddhist monk shows one of the wooden printing blocks of the Tripitaka Koreana, housed in Haein Temple, South Gyeongsang Province. The blocks are regarded as one of the nation's greatest treasures and a valuable religious heritage of the world. / Korea Times File

Figure 6. Tripitaka Koreana © Korean Times (http://www.koreatimes.co.kr/www/news/art/2010/03/293_61805.html).

Figure 7. Plate from the Tripitaka Koreana (http://www.dprk-tour.com/sub03/03.php).

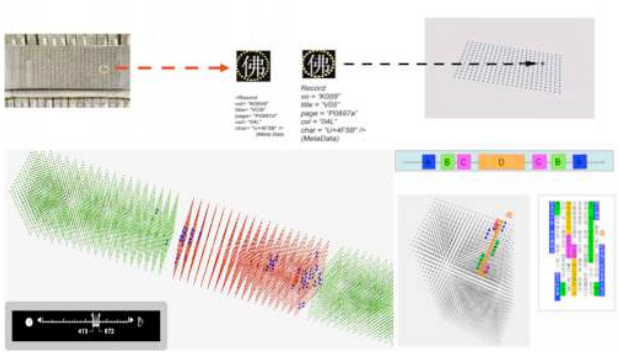


Figure 8. *Blue Dots*: abstraction of characters to dots and pattern arrays © ECAI, Berkeley.

The *Blue Dots* [5] project undertaken at Berkeley as part of the Electronic Cultural Atlas Initiative (ECAI) [10] which abstracted each glyph from the Canon into a blue dot, and gave metadata to each of these *Blue Dots* allowing vast searches to take place in minutes which would have taken scholars years (Figure 8). In the search function, each blue dot also references an original plate photograph for verification. The shape of these wooden plates gives the blue dot array its form.

As a searchable database, it exists on the Internet (5). Results are displayed in a dimensional array where users can view and navigate within the image. The image uses both the abstracted form of a "dot" as well as color to inform the user of the information being retrieved. Each blue dot represents one glyph of the dataset. Alternate colors indicate position of search results. Searching on a word results in a pattern where the distribution of that word is made evident as red dots within the blue dot matrix.

The use of color, form, and dimension for a fast understanding of the information is essential for large data sets where thousands of occurrences of a target word/phrase may be seen. Analysis across this vast text retrieves visual representations of word strings, clustering of terms, automatic analysis of ring construction, viewing results by time, creator, and place. The *Blue Dots* method of visualization is a breakthrough for corpora visualization and lies at the basis of the visualization strategies of abstraction undertaken in this project.

III. BLUE DOTS AVIE

Blue Dots AVIE will extend this methodology by developing tools for searching text in an embodied way. In AVIE, multiple users with gesture-based interfaces will be able to collaboratively search the text to find syntactic similarity and aforementioned linguistic phenomenon across the different corpora. The application of an omnispatial distribution of these texts solves problems of data occlusion, and enhances network analysis techniques to reveal patterns, hierarchies and interconnectedness (Figure 9 & 10). Using a hybrid approach to data representation audification strategies will be incorporated to augment interaction coherence and interpretation.

This project's omnispatial data browser prioritizes 'users in the loop' in an egocentric [22] and allocentric [27] model. To facilitate the flow of human reasoning, it provides a framework for 'enhanced human higher cognition' [3, 14] that extends the perceptual models common to visual analytics.

A. Project Objectives:

- Develop an omnispatial gesture based data browser for situated collaborative interaction with multilingual corpora.
- Develop visualization and audification strategies to represent linguistic phenomena that result from intra-corpora searching.
- Augment and supplement corpora metadata to enable visualization of contextual materials and heterogeneous materials.
- Develop network analysis algorithms to facilitate various types of pattern discovery and query.
- Evaluate the cognitive impacts and scholarly merits

of the browser.

Figure 9. Prof Lew Lancaster interrogates the preliminary prototype of Blue Dots AVIE © ALiVE, CityU. Image: Howie Lan

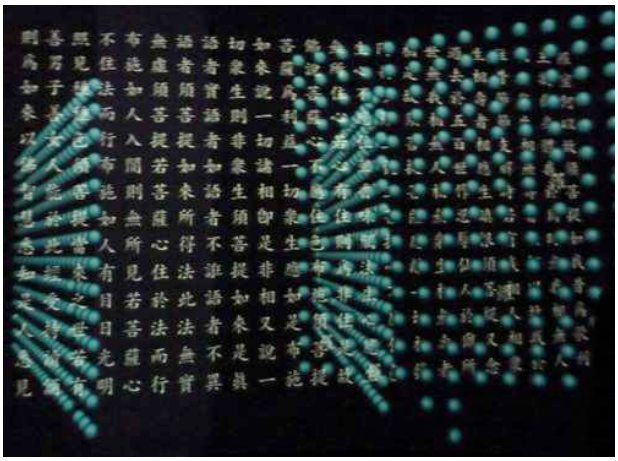


Figure 10. Close up of blue dots & corresponding texts, preliminary prototype of Blue Dots AVIE © ALiVE, CityU. Image: Howie Lan

The current search functionality ranges from visualizing word distribution and frequency, to other structural patterns such as the chiastic structure and ring compositions. In the *Blue Dots AVIE* version, the text is also visualized as a matrix of simplified graphic elements representing each of the words. This enables users to identify new linguistic patterns and relationships within the matrix, as well as access the words themselves and related contextual materials. The search queries will be applied across classical Chinese and eventually English, accessed collaboratively by researchers, extracted and saved for later re-analysis.

The data provides an excellent resource for the study of dissemination of documents over geographic and temporal spheres. It includes additional metadata such as present day images of the monasteries where the translation took place, which is will be included in the data array. The project will design new omnidirectional metaphors for interrogation and the graphical representation of complex relationships between these textual datasets to solve the significant challenges of visualizing both abstract forms and close-up readings of this rich data (Figures 11-13). In this way, it we hope to set benchmarks in visual analytics, scholarly analysis in the digital humanities and, the interpretation of classical text.

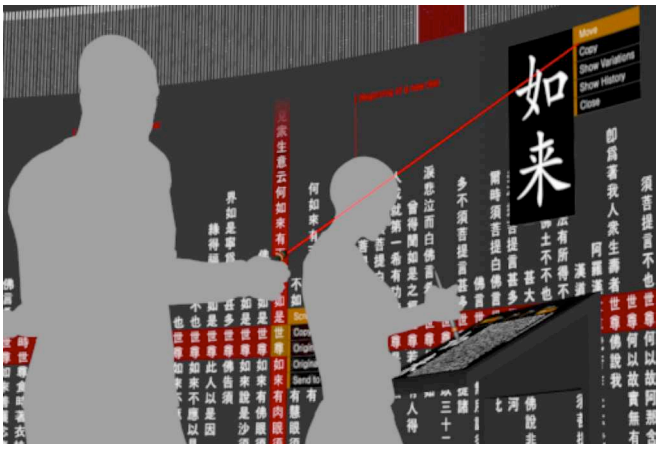


Figure 11. Graphic representation of Blue Dots Visualization in AVIE for close-up readings. Image Tobias Gremmler

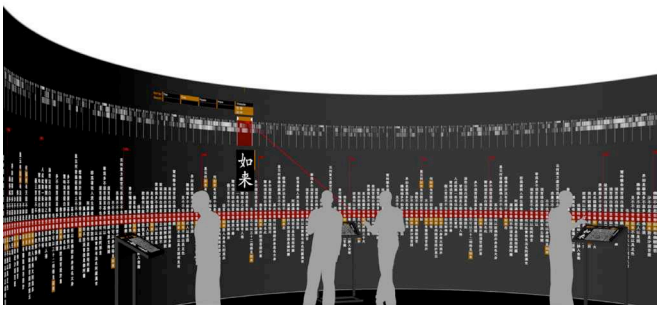


Figure 12. Graphic representation of *Blue Dots* Visualization in AVIE for close-up readings. Image Tobias Gremmler © ALiVE, CityU.

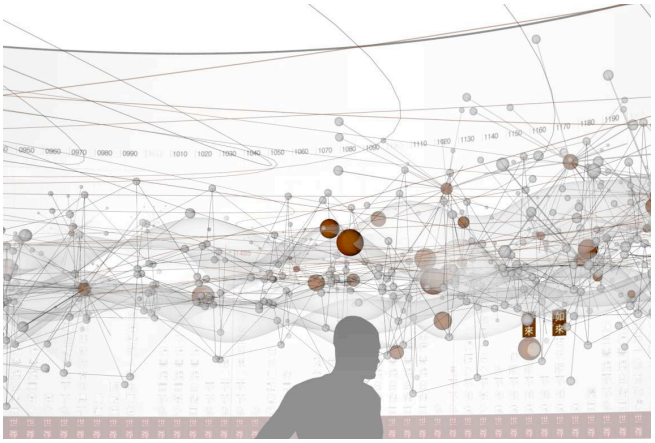


Figure 13. Graphic representation of Blue Dots Visualization in AVIE for abstract analysis. Image Tobias Gremmler

IV. Conclusion

This project is currently in its prototyping phase pending further funding which will see the methodology expanded to cross corpora searching and visualization inside AVIE. The researchers are currently experimenting with a variety of interface and interaction styles. With the complex integration of the full canon, scholars of the Tripitaka Koreana will be invited to use the system and will form part of a critical evaluation phase.

Acknowledgements

This paper contains information appearing in Kenderdine, Au & Shaw 2011 [25] and Kenderdine & Hart 2011 [26]. The principle author would like to acknowledge the contribution of colleagues at ALiVE. Also the contributions of members of the Department of Chinese, Translation and Linguistics, CityU, in relation to textual analytics, Prof Jonathan Webster and Dr John Lee and the Halliday Centre for Intelligent Applications of Language Studies.

Blue Dots AVIE: *Tripitaka Koreana*. Partners: ALiVE, City University of Hong Kong, UC Berkeley. Researchers: Dr Sarah Kenderdine, Prof Lew Lancaster, Howie Lan, Prof Jeffrey Shaw, Tobias Gremmler, Prof Jonathan Webster and Dr John Lee.

References

[1] Advanced Visualzation and Interaction Environment (AVIE). <http://www.icinema.unsw.edu.au/projects/infra_avie.html>

[2] Allosphere. <http://www.allosphere.ucsb.edu/>

[3] Anderson, M. 2003. Embodied Cognition: A field guide. *Artificial Intelligence,* 149(1), 91-130.

[4] Applied Laboratory for Interactive Visualzation and Embodiment (ALiVE). <http://www.cityu.edu.hk/alive>

[5] Blue Dots. <http://ecai.org/textpatternanalysis/>.

[6] Collins, C., *et al*. 2009. DocuBurst: Visualizing Document Content using Language Structure. *Proc. Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis) 28*(3), 1039-1046.

[7] DeFanti, T. A., *et al*. 2009. The StarCAVE, a third-generation CAVE & virtual reality OptIPortal. *Future Generation Computer Systems, 25*(2), 169-178.

[8] Digital Archives Projects, Dharma Drum Buddhist College. <http://www.ddbc.edu.tw/en/digital_archives/projects.html>

[9] Ebert, D. & Ertl, T. (Eds.). 2008. *IEEE Symposium on Visual Analytics Science & Technology: VAST '08*, IEEE. <http://conferences.computer.org/vast/vast2008/>

[10] ECAI. Electronic Cultural Atlas Initiative. <http://www.ecai.org>

[11] Ertl, T. (Ed.). 2009. *IEEE Transactions on Visualization & Computer Graphics*. IEEE. <http://www2.computer.org/portal/web/tvcg>

[12] Flowing Data. <http://flowingdata.com/>

[13] GapMinder. <http://ww.gapminder.org>

[14] Green, T. M., *et al*. 2009. Building and Applying a Human Cognition Model for Visual Analytics. *Information Visualization, 8*(1), 1-13.

[15] Harris, J. & Kamvar, S. 2009. *We feel fine*. New York, NY: Scribner.

[16] Harrison, C. & Romhild, C. 2008. The Visualization of the Bible. <http://www.chrisharrison.net/projects/bibleviz/index.hmtl>

[17] Havre, S., *et al*. 2000. ThemeRiver: Visualizing Theme Changes over Time. *Proc. IEEE Symposium on Information Visualization*, 115-123.

[18] Heer, J. & Agrawala, M. 2008. Design considerations for collaborative visual analytics. *Information Visualization, 7*, 49-62.

[19] HIPerSpace CALIT2. 2010. Research Projects: HIPerSpace. <http://vis.ucsd.edu/mediawiki/index.php/Research_Projects:_HIPerSpace>

[20] Johnson, C. R., et al. (Eds.) 2006. NIH/NSF Visualization Research Challenges Report. IEEE.

[21] Johnson, L., *et al*. 2010. *The 2010 Horizon Report*. Austin, Texas: The New Media Consortium. <http://wp.nmc.org/horizon2010/>

[22] Kasik, D. J., *et al*. 2009. Data transformations & representations for computation & visualization. *Information Visualization 8*(4), 275–285.

[23] Keim, D. A., *et al*. 2006. Challenges in Visual Data Analysis. *Proc. Information Visualization (IV 2006)*, 9-16. London: IEEE.

[24] Keim, D. A., *et al*. 2008. Visual Analytics: Definition, Process, & Challenges. *Information Visualization: Human-Centered Issues and Perspectives,* 154-175. Berlin, Heidelberg: Springer-Verlag.

[25] Kenderdine, S. & Au, K, Shaw, J. 2011, 'Cultural Data Sculpting: Omni-spatial Visualization for Cultrual Datasets, *Proceedings of IV10 - 15th International Conference on Information Visualisation*, IEEE, London, July 2011 (in press)..

[26] Kenderdine, S. & Hart, T. 2011, Cultural Data Sculpting: Omnispatial Visualization for Large Scale Heterogeneous Datasets, In D. Bearman & J. Trant (Eds.) *Museums and the Web, Selected papers from Museums and the Web 2011*. Philadelphia: Archives & Museum Informatics.

[27] Klatzky, R. L. 1998. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. *Spatial cognition - An interdisciplinary approach to representation and processing of spatial knowledge*, 1-17. Berlin: Springer-Verlag.

[28] Lancaster, L. 2010. Pattern Recognition & Analysis in the Chinese Buddhist Canon: A Study of "Original Enlightenment". *Pacific World*.

[29] Lancaster, L. 2007. The First Koryo Printed Version of the Buddhist Canon: Its Place in Contemporary Research. *Nanzen-ji Collection of Buddhist Scriptures and the History of the Tripitake Tradition in East Asia.* Seoul: Tripitaka Koreana Institute.

[30] Lancaster, L. 2008a. Buddhism & the New Technology: An Overview. *Buddhism in the Digital Age: Electronic Cultural Atlas Initiative*. Ho Chi Minh: Vietnam Buddhist U.

[31] Lancaster, L. 2008b. Catalogues in the Electronic Era: CBETA and *The Korean Buddhist Canon: A Descriptive Catalogue*. Taipei: CBETA (electronic publication).

[32] Lee, H., *et al*. 2010. Integrating Interactivity into Visualising Sentiment Analysis of Blogs. *Proc. 1st Int. Workshop on Intelligent Visual Interfaces for Text Analysis*, IUI'10.

[33] Lee, J. 2007. A Computational Model of Text Reuse in Ancient Literary Texts. *Proc. 45th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[34] Liu, S., et al. (Eds.). 2010. Proc. 1st Int. Workshop on Intelligent Visual Interfaces for Text Analysis, IUI'10.

[35] Manovich, L. 2009. How to Follow Global Digital Cultures, or Cultural Analytics for Beginners. *Deep Search: They Politics of Search beyond Google.* Studienverlag (German version) and Transaction Publishers (English version).

[36] Many Eyes. <http://www.manyeyes.alphaworks.ibm.com>

[37] McCandless, D. 2010. The beauty of data visualization [Video file]. <http://www.ted.com/talks/lang/eng/david_mccandless_the_beauty_of_data_visualization.html>

[38] McGinity, M., *et al.* 2007. *AVIE*: A Versatile Multi-User Stereo 360-Degree Interactive VR Theatre. *The 34th Int. Conference on Computer Graphics & Interactive Techniques, SIGGRAPH 2007*, 5-9 August 2007.

[39] Paley, B. 2002. TextArc. <http://www.textarc.org>

[40] Pike, W. A., *et al.* 2009. The science of interaction. *Information Visualization, 8*(4), 263–274.

[41] Pranav Mistry Six Sense. <www.pranavmistry.com/projects/sixthsense/>

[42] Project Natal. <http://www.microsoft.com/uk/wave/hardware-projectnatal.aspx>

[43] Speer, R., *et al*. 2010. Visualizing Common Sense Connections with Luminoso. *Proc. 1st Int. Workshop on Intelligent Visual Interfaces for Text Analysis* (IUI'10), 9-12.

[44] Thai, V. & Handschuh, S. 2010. Visual Abstraction and Ordering in Faceted Browsing of Text Collections. *Proc. 1st Int. Workshop on Intelligent Visual Interfaces for Text Analysis* (IUI'10), 41-44.

[45] Thomas, J. & Kielman, J. 2009. Challenges for visual analytics. *Information Visualization, 8*(4), 309-314.

[46] Visnomad. <www.visnomad.org>

[47] Visual Complexity. <www.visualcomplexity.com/>

[48] Wattenberg, M. & Viégas, F. B. 2008. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics 14*(6).

[49] Wattenberg, M. 2002. Arc Diagrams: Visualizing Structure in Strings. *Proc. IEEE Symposium on Information Visualization*, pp.110-116. Boston, MA.

[50] Wordle. <www.wordle.net>