DIVAS: MULTIMEDIA INFORMATION RETRIEVAL FROM MULTIMODAL DESIGN AND CONSTRUCTION KNOWLEDGE CORPORA

Renate Fruchter[1] and Zhen Yin[2]

ABSTRACT

Managing and reusing knowledge can lead to greater competitive advantage, and improved designs. However, design knowledge reuse often fails in the AEC (Architect, Engineering, and Construction) industry. We performed an ethnographic study to investigate the current practice of the shop drawing detailing and review process. Our observations show that most of the design and detailing knowledge among three key stakeholders: structural engineers, detailers, and architects, is communicated through discourse, sketch, and gesture. This multimodal design knowledge is not reusable because of two reasons: (1) the design knowledge is evolving through these informal and multimodal communication channels, which is hard to capture. (2) Even if the multimedia recording devices and technologies are used to capture these communications, the archived multimedia data is highly unstructured, and consequently to search and retrieve relevant content.

This study focuses on two research questions: (1) how to capture the stakeholders' knowledge and experience with high fidelity and least overhead? (2) how can stakeholders search, retrieve, and understand relevant knowledge from a large, unstructured, rich, and multimodal knowledge repository? We present DiVAS (**Di**gital **V**ideo **A**udio **S**ketch) a multimedia information capture and retrieval system to addresses these knowledge reuse issues. DiVAS is able to capture the multimodal design knowledge, improve the communications among stakeholders through effective contextual cross-media content retrieval, and significantly reduce the number of RFIs (Request for Information).

KEY WORDS

Information Retrieval, Multimedia, Multimodal, Knowledge Capture, Knowledge Reuse.

INTRODUCTION

As reflected in the Journal of Computing in Civil Engineering, modeling, as a persistent technical topic, is necessary for the civil engineer to fully understand new problems and situations [Abudayyeh et al. 2006]. Less research has been dedicated to model the multimodal design and construction knowledge in order to enable the knowledge reuse. The design knowledge reuse failure can result in substantial mistakes for both structure design and shop drawing design [Luth 2003], which eventually jeopardizes the construction process

[1] Director of Project Based Learning Laboratory (PBL Lab), Dept. of Civil and Environmental Engineering, Bldg. 550, CA 94305, USA, Phone 650/725-1549, FAX 650/723-4806, fruchter@stanford.edu

[2] PhD Candidate, Dept. of Civil and Environmental Engineering, Bldg. 550, Stanford, CA 94305, USA, Phone 650/725-1549, FAX 650/723-4806, zhenyin@stanford.edu

and the building. Recent research has been investigating the multimodal communication [Kondratova 2003 and Kondratova 2005] and multimodal engineering interfaces [Kamat et al. 2005, and Reinhardt et al. 2005]. However, these researches are either focusing on the construction process, or customized for the specific on-site construction tasks [Garrett et al. 2004]. It is necessary to design and implement a multimedia information retrieval system, which is able to capture the informal multimodal design and construction knowledge. This system should also provide an effective multimedia information retrieval mechanism to enable the knowledge reuse. We introduce the DiVAS prototype, which takes advantage of devices that enable multimedia and multimodal direct manipulation and capture of content. It enables the capture of content created during analog activities expressed through gesture, verbal discourse, and sketching into digital video, audio, and sketches. It provides an integrated digital video-audio-sketch environment for knowledge reuse.

RELATED WORK

In the domain of computing in civil engineering, recent research starts to focus on the multimedia technologies from retrieval perspective. For example, [Brilakis et al. 2005] proposed an image retrieval system for the on-site materials. Related research in information technology domain also proposes media specific analysis solutions, e.g., Video Traces for video content annotation [Stevens et al. 2002], Tracker for video content processing based on object segmentation [Zhong and Chang, 1997], Fast-Talk for audio search [Kim et al. 2003], text vector analysis and latent semantic indexing for information retrieval from text repositories [Landauer and Dumais, 1995], video object segmentation of video footage [Farin et al. 2003]. Nevertheless, there is little research in cross-media capture and retrieval approaches. Therefore, the knowledge communicated through multimodal channels is only partially captured or captured out of context. [Kondratova 2003 and Kondratova 2005] Kondratova's research provides users the voice and multimodal access to AEC (Architectural, Engineering, and Construction) project information. This research focuses on multimodal data collection and access for the construction field work. The interaction multimodalities captures by this approach are speech, stylus, and keyboard input. The design knowledge evolved through gestures and sketches with pencil and paper is not captures, though these two modalities frequently take place during engineers' informal communicative events.

The media specific analysis solution provides the recognition technology for single communication channels, such as the speech recognition technology for the discourse channel. The multimodal system actually can support more robust recognition by combing individual error-prone recognition technologies [Oviatt 1999]. Our observations show that during communicative events there is a continuum between gestures, discourse, and sketching as ideas are explored and shared. The link between gesture-discourse-sketch provides a rich context to express and exchange knowledge. This link becomes critical in the process of knowledge retrieval and reuse to support the user's assessment of the relevance of the retrieved content with respect to the task at hand.

PROBLEM IDENTIFICATION

Managing and reusing knowledge can lead to greater competitive advantage and improved designs. However reuse often fails, since knowledge is not captured, it is captured out of

context rendering it not reusable, or there are no formal mechanisms for finding and retrieving reusable knowledge. One of the worst case scenarios of knowledge capture and reuse failure in the AEC (Architect, Engineering, and Construction) industry was the Hyatt Regency Hotel collapse. During the shop drawing detailing and review process, the project manager of the structure design team was asked to confirm the details of a connection. He was not able to retrieve the relevant design knowledge to make the right decision. The connection was confirmed to be built, even though it was never designed. This connection failed and resulted in an accident, in which more than a hundred people died.

We performed an ethnographic study investigating the current practice of the shop drawing detailing and review process. Our observations show that most of the design and detailing knowledge among the three key stakeholders: structural engineers, detailers, and architects, is communicated through informal channels, such as gesture, discourse, and sketching channels. However, the multimodal design knowledge evolving through these communication channels is typically not captured. The current digital content management is limited to digital archives of formal documents (CAD, Word, Excel, etc.), and disconnected digital image repositories and video footage. These ignore the highly contextual and interlinked modes of communication in which people generate concepts, and reuse knowledge through gesture language, verbal discourse, and sketching. The analog design activities should be captured and converted into digital format without interrupting designers' common working behavior. Such a technology should be provided to capture the knowledge and the context in which this knowledge was originally created. Furthermore, an information retrieval mechanism should be provided to the user to support relevant knowledge reuse.

The hypothesis of this research is that if designers are able to understand the context in which the informal, multimodal design knowledge was originally created through the interaction with this rich content, i.e., interlinked gestures, discourse, and sketches, this knowledge can be assessed and reused in a meaningful way. The information capture and retrieval mechanism to address the knowledge reuse failure problems include: 1) Creative informal communicative events where people generate knowledge through gestures, dialogue, and sketches are captured and archived in a multimedia corpus; 2) Effective retrieval and reuse of knowledge from this multimedia corpus; 3) A multimedia interactive information retrieval interface that presents to the users the query results and the context of the query results

This study presents a framework and methodology to address the knowledge capture, retrieval, and reuse mechanism through three steps: non-intrusive knowledge capture, data analysis, and knowledge retrieval from large enterprise archives of rich, multimedia, unstructured content. We developed DiVAS, a multimedia information retrieval system.

Highly structured representations of design knowledge can be used for reasoning. However, these approaches usually require manual pre or post processing, structuring and indexing of design knowledge. In order to perform an integrated analysis and extract relevant content from digital video and audio footage, it is critical to convert the unstructured, informal content capturing gestures in digital video, discourse in audio, and sketches in digital sketches, into symbolic representations by converting video images of people into gesture vocabulary, audio into text, and sketches into sketch objects, respectively.

DIVAS

The DiVAS system presents an integrated multimedia environment and analysis methodology of indexed digital video-audio-sketch content in support of knowledge capture in context, and content understanding and knowledge reuse. The DiVAS system and approach takes advantage of:

- Innovative algorithms to index and capture audio and sketch developed by Dr. Fruchter and her team in previous research that led to a prototype called RECALL [Yen et al. 1999]. The content retrieval for knowledge reuse results in the sketch drawn up to the point where the corresponding discourse starts that is relevant to the knowledge reuse objective.
- Advanced techniques for object segmentation and automatic extraction of semantics out of digital video to develop a well defined, finite gesture vocabulary that describes a specific professional gesture language to be applied to the video analysis. This video analysis results in a video-gesture vocabulary.
- Advanced techniques for voice-to-text conversion (e.g., Dragon, MS Speech Recognition) and text search. As other studies have shown, text is most promising source for information retrieval. The information search applied to the audio/text portion of the indexed digital video-audio-sketch footage results in relevant discourse-text-samples linked to the corresponding video-gestures.

The research builds on previous ethnographic studies [Fruchter and Demian, 2002] of designers at work finding and understanding relevant knowledge from past experiences. We used a scenario-based approach to study the nature of the continuum between professional gesture vocabulary used in conjunction with verbal discourse, and sketching during of design concept generation and development activities. We developed an integrated analysis methodology of indexed digital video-audio-sketch content in support of knowledge capture in context, and content understanding and knowledge reuse from single designer sessions.

Most importantly, DiVAS integrates two technologies we developed, I-Gesture and I-Dialogue. I-Gesture [Fruchter and Biswas, 2005] enables the semantic video processing of captured footage during communicative events. I-Dialogue enables effective information retrieval from speech transcripts using Notion Disambiguation. Moreover, DiVAS constructs a cross-media relevance model, which correlates the contextual content from gesture, discourse, and sketch channels. This model significantly improves the retrieval performance. The following sections discuss the DiVAS system architecture, its modules, performance evaluation, and its contributions.

SYSTEM ARCHITECTURE

Based on the ethnographic study, this research formalizes three key activities in the knowledge life cycle as shown in the table 1. These three key activities provide the blueprint for the DiVAS system architecture as shown in figure 1. These key activities are supported by three major functional modules of DiVAS system. The capture activity is supported by the integration a knowledge capture technology called RECALL [Yen et al. 1999]. RECALL supports capture of the sketch, speech, and gestures into digital format, which represents the typical communication channel set used by architects, engineers, and detailers. The DiVAS

approach supports the retrieval activity through an integrated retrieval analysis of gesture vocabulary, verbal discourse, and sketch captured in digital video, audio, and sketch. It enables users to understand the explored content through an interactive multimedia information retrieval process. Each module is discussed in the following sections.

Table 1: Key Activities in Knowledge Reuse Life Cycle

Key Activities in Knowledge Reuse Life Cycle	Human Activities in Designers' Working Environment Supported by DiVAS
Capturing the knowledge creation process in a non-intrusive working setting	When a structural engineer finds a possible mistake on the shop drawings, she/he uses the engineering expertise to confirm this mistake. For example, she/he will check the capacity of a connection through calculations. The knowledge used in this process, such as back of the envelope calculation, a sketch that shows the load path, etc., will be captured together with the verbal explanation.
Retrieving or recalling reusable items from a repository of unstructured informal knowledge	A detailer wants to understand the structure engineer's markup. Assuming the markup generation process is captured, i.e., behind the engineer's rationale. The detailer searches the data archive and finds the material related to this markup, such as the load path sketches and the explanation.
Understanding these items.	By reviewing the load path sketches and back of the envelope calculation, the detailer understands that the connection she/he detailed doesn't provide enough capacity.

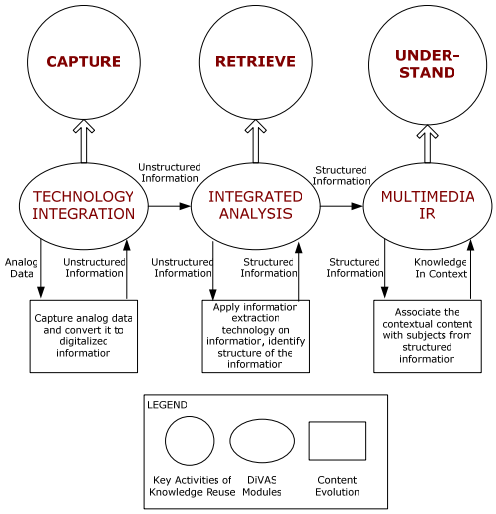

Figure 1: DiVAS System Architecture

CAPTURE MODULE

The capture environment of DiVAS considers the includes: (1) The designers' working environment is included in order to illustrate the usage scenario of this module. (2) I-Gesture prototype supports the video capture and processing modules. (3) The audio capture and processing modules are supported by V2TS (Voice to Text and Sketch) developed in the PBL Lab at Stanford. The sketch processing and synchronization are supported by RECALL. RECALL is a drawing application written in Java that captures and indexes each individual action on the drawing surface, such as sketching on paper. The drawing application synchronizes with audio/video capture and encoding through client-server architecture. Once the session is complete, the drawing and video information is automatically indexed and published on a Web server that allows for distributed and synchronized playback of the drawing session and audio/video from anywhere at anytime. The V2TS module processes the audio data stream captured by RECALL during the communicative event. The V2TS speech recognition module recognizes words or phrases from an audio file created in a RECALL session and stores the recognized occurrences and corresponding times to be used later during replay. Voice to text translation is performed by using standard speech recognition engine and speech recognition SDK. For this research, the Dragon NaturallySpeaking engine is used. The I-Gesture module enables the semantic video processing of captured footage during communicative events. Observations suggest that gesture movements performed by users during communicative events encode a large amount of information. Therefore by identifying the gestures, the context, and the times when they are performed can provide a valuable index available to the user to search for a particular issue. I-Gesture system has two modules: gesture definition module and video processing module. The gesture definition module uses advanced video segmentation and classification algorithms to extract and characterize video objects in order to develop the gesture database. The video processing module processes the video created during a communicative event session by comparing it with a gesture database, identifies the gestures performed in it and marks it up, i.e., store the occurrences and their timestamps to be used later during search, retrieval, and replay.

INTEGRATED ANALYSIS MODULE

The objective of the integrated analysis of gesture language, verbal discourse, and sketch captured in digital video, audio, and digital sketch respectively is to build up the index, both locally for each media and across media. The cross media index reflects whether content from gesture, discourse, and sketch channels are relevant to a specific subject. As shown in figure 2, the integrated analysis module is used to process video, audio, and sketch data. Before the cross media index is built, the integrated analysis module needs to construct the index for each media. The green path in figure 2 refers to the sketch processing. The index construction of the sketch uses RECALL algorithm. The blue path refers to the gesture processing. The first step is to identify the gesture vocabulary from designers' common practice. The next module formalizes the gesture lexicon based on this vocabulary. These two modules are designed and implemented by following the gesture reuse framework of I-Gesture prototype. The extracted gestures are used for later information retrieval in DiVAS. The red path refers to the discourse processing, which is defined and developed in the

framework of I-Dialogue. I-Dialogue is a DiVAS module that processes the speech transcript. The objective of I-Dialogue is to add structure to the unstructured speech transcripts. As shown in the red path, I-Dialogue uses vector analysis and LSI (Latent Semantic Analysis) to add clustering information to the unstructured speech transcripts. Consequently, the unstructured speech archive becomes a semi-structured speech archive. Then, I-Dialogue uses notion disambiguation to label clusters. The document inside the clusters is assigned the same labels. Both document labels and categorization information are used to improve information retrieval.

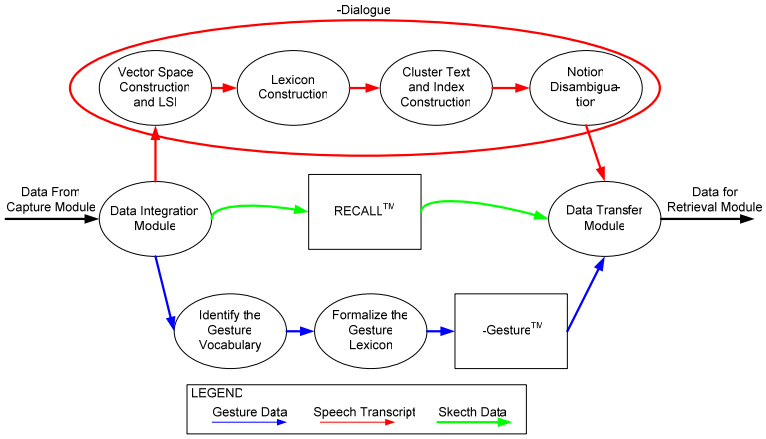

Figure 2: Integrated Analysis Module

RETRIEVAL MODULE

The gesture, discourse, and sketch data are stored in the DiVAS data archive after being processed from the capture and analysis module. Designers are able to start their query with a traditional text search engine. Key words can be input with logical expression, e.g. “roof + steel frame”. The retrieval module processes a query by comparing the query with all the speech transcript documents. The most similar documents are returned and ranked by similarity. In parallel DiVAS performs a search of the contextual content from gesture and sketch channel.

CROSS MEDIA RELEVANCE MODAL AND RANKING

Once the digital video-audio-sketch content is captured in real time by RECALL then the two macro level digital media, i.e., audio and video, are processed by the I-Dialogue and I-Gesture modules and all three channels are indexed and synchronized automatically. The DiVAS system provides an innovative cross-media search, retrieval and replay facility to capitalize on multimedia content stored in large, multimedia, unstructured corporate repositories. The user can search for a keyword (a spoken phrase or gesture). The system searches through the entire repository and displays all the relevant hits as gesture-text-sketch episodes. On selecting a session, DiVAS replays the selected session from the point where the keyword was spoken or performed. It is important to note that video and audio/text

provide a macro index and the sketch provides a micro index to large, unstructured repositories of rich multimedia content. The background processing and synchronization is performed by an applet that uses multithreading to manage the different streams. We develop a synchronization algorithm that allows us to use as many parallel streams as possible. Therefore there is the possibility of adding more streams or modes of input and output for a richer experience.

SYSTEM EVALUATION

To date few research efforts focused on modeling and knowledge capture and reuse of the multimodal communicative events in design and construction. This research builds on empirical studies in this specific design and construction domain. We tested and evaluated the DiVAS system in an academic test bed [Fruchter 1999], which simulates the industry working and computing environment. Within this test bed, students use DiVAS as part of their work environment in support of their building design project activities. Design knowledge has been archived and processed by DiVAS. For illustration, the notion labels for a document cluster after the integrated gesture-discourse-sketch analysis are: “architectural constraint”, “height limitation”, “concrete solution”, and “tension ring”. According to the expert definition, the topics for this cluster are: “Height limitation is one of the major architectural constraints”, and “Tension ring and X-bracing connect concrete structure and steel structure”. Similar observations are found for all the other clusters. Therefore, notion disambiguation is able to identify the clean form of notions. Although notions labels don’t fully represent the topics of the cluster, they emphasize a major portion of concepts mentioned in the topics.

In order to validate the information retrieval improvement of the cross media relevance ranking, a series of experiments are performed. Besides the output of V2TS, the audio part of RECALL sessions was manually transcribed in order to provide the comparison baseline for the information retrieval. Lucene (a full-featured text search engine provided by Apache Software Foundation) is used to construct the index over the archive and perform the query analysis. The archive is modified into six different forms as following:

- “clean” speech transcripts only – manually transcribed speech sessions, which are used as the comparison base
- “dirty” speech transcripts only – automatically transcribed speech sessions with transcription errors
- “dirty” speech transcripts and notion labels
- “dirty” speech transcripts and gesture labels
- “dirty” speech transcripts, notion labels, and gesture labels (integrated after I-Dialogue is applied)
- “dirty” speech transcripts, notion labels, and gesture labels (integrated before I-Dialogue is applied)

For the archive, there are eight notions predefined, e.g., “tension ring”. Each of them is used as the query terms. The ranking sequence corresponding to each queried notions for all six archives are compared. The recall and precision ratio is calculated and recall vs. precision curve is used for evaluation. Figure 3 shows the retrieval improvement for the query “tension ring”. Similar results were found for all the other queries. As a summary, the precision and recall ratio of using cross media relevance ranking are higher than using single media only. And cross media relevance ranking is able to overcome the inaccuracy due to single media data errors. These results show that DiVAS facilitates the multimedia retrieval activities for the multimodal design and construction knowledge corpora. We plan to deploy DiVAS system in industry pilot projects and perform usability analysis to further evaluate and improve the system.

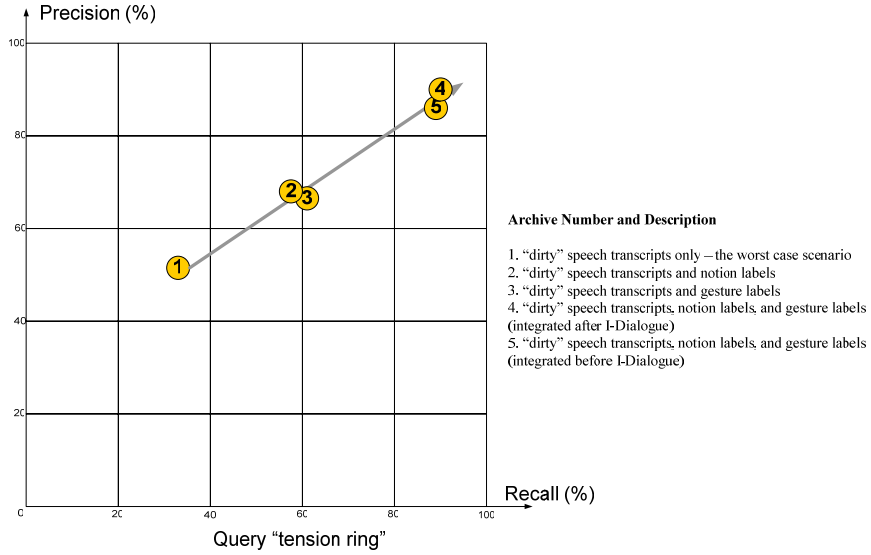

Figure 3: Recall vs. Precision Curve

CONCLUSION

The DiVAS system presents an innovative, integrated cross-media methodology of indexed digital video-audio-sketch rich content in support of knowledge capture in context, content understanding and knowledge reuse. DiVAS provides the AEC industry a ubiquitous knowledge capture and reuse environment, with least overhead to stakeholders' working practice. It enables stakeholders to quickly navigate through unstructured multimodal design knowledge archives and effectively reuse the design knowledge.

ACKNOWLEDGMENTS

This project was sponsored by PBL Lab, Media X at Stanford University, and KDDI, Japan.

REFERENCES

Osama Abudayyeh, Amber Dibert-DeYoung, William Rasdorf, and Hani Melhem. (2006). "Research Publication Trends and Topics in *Computing in Civil Engineering*". Journal of Computing in Civil Engineering. Volume 20, Issue 1, pp. 2-12.

Ioannis Brilakis, Lucio Soibelman, and Yoshihisa Shinagawa. (2005). "Material-Based Construction Site Image Retrieval". Journal of Computing in Civil Engineering. Volume 19, Issue 4, pp. 341-355.

Dirk Farin, Peter H. N. de With, and Wolfgang Effelsberg. (2003). "Recognition of User-Defined Video Object Models using Weighted Graph Homomorphisms". Proceedings of. SPIE Electronic Imaging 2003, Image and Video Communications and Processing Conference (Santa Clara, CA, 2003).

Dirk Farin, Thomas Haenselmann, Stephan Kopf, Gerald Kühne, and Wolfgang Effelsberg. (2003). "Segmentation and Classification of Moving Video Objects". Borko Furht, Oge Marques (eds.) Handbook of Video Databases (CRC Press 2003).

R. Fruchter. (1999). "Architecture/Engineering/Construction Teamwork: A collaborative Design and Learning Space". Journal of Computing in Civil Engineering, Volume 13, Issue.4, pp. 261-270.

R. Fruchter and Biswas. (2005). "Using Gestures to Convey Internal Mental Models and Index Multimedia Content". Proceedings of Social Inteligent Design workshop. (Stanford, CA, 2005)

R. Fruchter and P. Demian. (2002). "CoMem - Designing an Interaction Experience for Reuse of Rich Contextual Information from a Corporate Memory". Proceedings of AIEDAM International Special Issue on Human Computer Interaction in Engineering Context (2002) 16,127-147.

James H. Garrett, Jr. , Ian Flood, Ian F. C. Smith, and Lucio Soibelman. (2004). "Information Technology in Civil Engineering—Future Trends". Journal of Computing in Civil Engineering. Volume 18, Issue 3, pp. 185-186.

Vineet R. Kamat and Julio C. Martinez. (2005). "Dynamic 3D Visualization of Articulated Construction Equipment". Journal of Computing in Civil Engineering. Volume 19, Issue 4, pp. 356-368.

Jinmook Kim, Douglas Oard, and Dagobert Soergel. (2003). "Searching Large Collections of Recorded Speech: A Preliminary Study". Proceedings of ASIS&T 2003 Annual Meeting (October 19-22, 2003, Long Beach, CA)

I. L. Kondratova. (2005). "Speech-Enabled Handheld Computing for Fieldwork". Proceedings of the 2005 ASCE International Conference on Computing in Civil Engineering Lucio Soibelman, Feniosky Peña-Mora - Editors, July 12–15, 2005, Cancun, Mexico.

I. L. Kondratova. (2003). "Voice and Multimodal Access to AEC Project Information". Proceedings of 10th ISPE International Conference On Concurrent Engineering.

T. K. Landauer, and S.T. Dumais. (1995). "A Solution to Plato's Problem - The Latent Semantic Analysis Theory of Acquisition". Induction and Representation of Knowledge, in: Psychological Review (1995)104, 211-240.

Gregory P. Luth. (2000). "Chronology and Context of the Hyatt Regency Collapse". Journal of Performance of Constructed Facilities. Volume 14, Issue 2, pp. 51-61.

Sharon Oviatt. (1999). "Ten myths of multimodal interaction". Journa of Communication of the ACM. Volume 42, Issue 11, pp. 74-81.

Jan Reinhardt, James H. Garrett, Jr., and Burcu Akinci. (2005). "Framework for Providing Customized Data Representations for Effective and Efficient Interaction with Mobile Computing Solutions on Construction Sites". Journal of Computing in Civil Engineering. Volume 19, Issue 2, pp. 109-118.

Reed Stevens, Gina Cherry, and Janice Fournier. (2002). "Video Traces – Rich Media Annotations for Teaching and Learning". Proceedings of 2002 Computer Supported Collaborative Learning Conference (Boulder, Colorado, 2002).

S. Yen, R. Fruchter, and L. Leifer. (1999). "Capture and Analysis of Concept Generation and Development in Informal Media". Proceedings of ICED 12th International Conference on Engineering Design (Munich, Germany, 1999).

D. Zhong and S.F. Chang. (1997). "Video Object Model and Segmentation for Content-Based Video Indexing". Proc. of IEEE International Conference on Circuits and Systems (Hong Kong 1997).