CONTEXT-AWARE DYNAMIC PRESENTATION SYNTHESIS FOR EXPLORATORY MULTIMODAL ENVIRONMENTS

Harini Sridharan *Ankur Mani* *Hari Sundaram* *Jennifer Brungart* *David Birchfield*

Arts Media and Engineering Program

Arizona State University, AZ 85281

e-mail: {first.last}@asu.edu

ABSTRACT

In this paper, we develop a novel real-time, interactive, automatic multimodal exploratory environment that dynamically adapts the media presented, to user context. There are two key contributions of this paper – (a) development of multimodal user-context model and (b) modeling the dynamics of the presentation to maximize coherence. We develop a novel user-context model comprising interests, media history, interaction behavior and tasks, that evolves based on the specific interaction. We also develop novel metrics between media elements and the user context. The presentation environment dynamically adapts to the current user context. We develop an optimal media selection and display framework that maximizes coherence, while constrained by the user-context, user goals and the structure of the knowledge in the exploratory environment. The experimental results indicate that the system performs well. The results also show that user-context models significantly improve presentation coherence.

1 INTRODUCTION

In this paper we develop a framework for a context-aware, user-centric, automatic multimodal presentation system. The problem is important in immersive multimodal environments (eg. computer games) as well as learning environments for children.

There has been prior work on dynamic presentation schemes [3]. While the work presents an efficient spatial arrangement of media, it is limited to tailoring the presentation based upon user query. There is no attempt to model the user context based upon the short term memory. Related work on context [4] focuses on a very narrow scope of context like location, identity, activity and time and has been successfully used is the areas of context aware ubiquitous computing. However, there is no framework to model multi sensory user context.

In our approach, we limit our domain to concepts in geography. The environment is created with an expert and the environmental knowledge is structured. The multimodal environment is created as audio-visual collage. We develop a multimodal user-context model that evolves with user interaction with the environment. We also develop new metrics for media distances to user context. The dynamic presentation framework for the environment is dependent on the user context, goals and knowledge in the environment. We show an optimal media selection and presentation algorithm to maximize the presentation coherence. The audio-visual collage is dynamically created. The user results are very good and indicate that user context models are crucial in such adaptive presentation systems.

The organization of the rest of the paper is as follows. In the next section, we describe our environment. In section 3, we shall present our model of user context. In section 4, we shall discuss our framework for dynamic presentation synthesis and we shall conclude with section on experiments and conclusions.

2 EXPLORATORY ENVIRONMENT

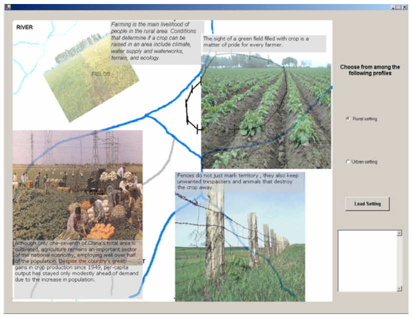
We built an exploratory environment to allow students to actively explore concepts in geography. Within geography, we focused on the abstract concept of *population density* and related key concepts (e.g. water).

2.1 Structure in Knowledge

The knowledge in the environment is structured by a domain expert. The expert encodes concept relations using a *Knowledge Flow Graph* (KFG). This is a directed acyclic graph with vertices representing concepts and the edges representing the progressive knowledge flow. Each of the key concepts related with population density has an '*associated set*' of concepts. Some concepts in this associated set are critical to understanding the corresponding key concept related to population density. The initial importance scores, associated set for goal concepts and the KFG are given by an expert. Note that there can be an intersection between the associated sets of the various key concepts.

2.2 Visualization

Our exploratory environments are map-based. As population density is tightly coupled to the setting, we chose three profiles – rural, sub-urban and urban, each widely different from the other in terms of environmental setting. Each setting is represented as a map and characterized by a set of locations. Each map was carefully constructed in consent with a graphic designer in a manner so as to typify the environmental setting of the profile. On 'mouse-over' a location, a subset of the media associated with the location and related to the user's context is presented in the form of collages.



Farming is the main livelihood of people in the rural area. Conditions that determine if a crop can be raised in an area include climate, water supply and waterworks, terrain, and ecology.

The sight of a green field filled with crop is a matter of pride for every farmer.

Although only one-seventh of China's total area is cultivated, agriculture remains an important sector of the national economy, employing well over half of the population. Despite the country's great gains in crop production since 1949, per-capita output has stayed only modestly ahead of demand due to the increase in population.

Fences do not just mark territory , they also keep unwanted trespassers and animals that destroy the crop away.

● Rural setting

○ Urban setting

Figure 1: A map-based visualization that presents media collages on interaction, based upon user's current context.

2.3 Auditory Mapping

In our environment, we build upon our prior work [2]. We create sonic collages for conveying information about geographical concepts and population density, and increase their effectiveness by ensuring that: (a) the selection of audio samples and their presentation is informed by the context (b) we use auditory information in addition to connotation to shape sonic environments and exaggerate important concepts and (c) the sonic environment is constantly evolving and transforming in real time as a user explores the virtual space.

0-7803-9332-5/05/$20.00 ©2005 IEEE

3 CONTEXT MODEL

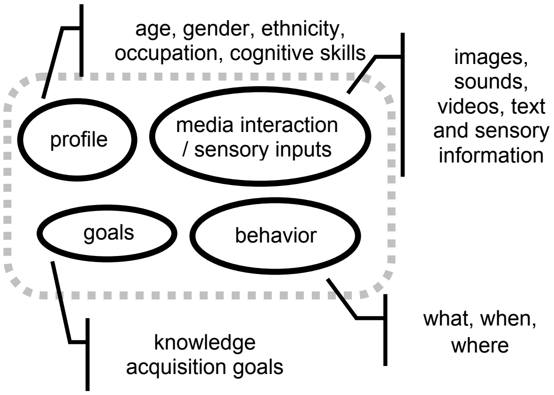


Figure 2: The user context model has four components – (a) profile (b) media history (c) behavior and (d) goals

The Merriam-Webster dictionary [1] defines context as "the interrelated conditions in which something exists or occurs." In our framework, we define *current user context to be the subset of the space of multimodal assertions that are true for the user at this time, which affect the user's interaction with the environment*. This builds upon earlier work on context, that only dealt with textual concepts [7].

3.1 User context

In our proposed framework, the user-context is a graph with concepts at each node, and where the nodes are connected with a specific relationship. The relationship between concepts can come from different sources – (a) linguistic (e.g. WordNet [5], (b) feature based relationships

The formal model is defined using a semantic-net – a graph $G = <V,E,W>$ where the nodes $v_i \in V$ represent the concepts, the edges $e_{ij} \in E$ represent the type of relationship (semantic, spatio-temporal, feature-level) between the nodes i and j and $w_{ij} \in$ W, specifies the strength of the relationship between the two nodes. As depicted in Figure 2, we develop a simple model where we define the user context to comprise four semantic nets (a) the initial user profile (stating the user's interests, background etc.), (b) the viewing history (it establishes the importance of knowledge the user has acquired while browsing the environment), (c) user behavior (locations visited and time spent on each media element) and (d) user's learning goals. The context is the union of such semantic-nets.

3.1.1 Concept Covers

The context model is seeded with the initial user profile (details like gender, ethnicity, age, profession, cultural interests) and the user goals. We introduce the idea of concept covers for media sets – these are concepts that effectively represent the media sets.

Text: The text in the environment is parsed, and after stop-word removal, a simple term frequency analysis is used to determine the dominant concepts. The covers are created using the generalization / specialization relations in WordNet. The dominant textual concepts are expanded on their synonym sets (synset) and also generalized using WordNet . The generalized synset (the *synset* for a word are a group of words which sufficiently characterize the semantics associated with the word) that covers these concepts is then the concept cover.

Images: The color histogram of the images in the user context is computed. We use the HSV color space and the histogram is calculated with 166 bins. Image histograms are clustered using the K-Means clustering algorithm . The cluster center defines the concept cover for the images that fall into that cluster. *Audio*: The audio clips are clustered similar to images. However, we use MFCC [6] as the features space. The distance is defined as the mean-squared cepstral distance [6]. The concept cover of the set of audio clips belonging to a cluster is defined by their cluster center. The distance between an image (audio clip) and its cluster center defines the clip's relationship to the concept cover.

3.2 Media distance to User Context

We now show how the distance between a media element and the user context is computed. The distance between a text and the (textual concepts in) the user profile is defined using the idea of WordNet implication distance [7]. Hence the textual distance d_t between two concepts α and β is given by

$$d_t = 1 - I(\alpha \mid \beta = T), \qquad <1>$$

where $I(\alpha \mid \beta = T)$ is the *implication* that the concept α is true given that another concept β is true. This distance is normalized by the knowledge priors for concepts α and β. The text concept distance is the average of the distances to all the text concepts in the current user context.

The distance between two images, d_i is defined as their low-level color histogram distance. We use the HSV color space with 166 bins. The distance between an image and (the images in) the current user context is then the average distance between this image and all the image cluster centers in the current user context.

We define the auditory distance d_a between two audio clips as the mean-squared cepstral distance between them [6]. The distance is calculated using the first two seconds of the corresponding clips. The audio sequence is divided into 200ms overlapping frames (100ms overlap) and the cepstrum of frames is computed. The distance between an audio clip and (the audio clips in) the current user context is the average distance between this clip and all the audio concept covers (cluster centers) of clips in the current user context. The final distance is then determined to be a weighted sum of the distances and is computed as follows:

$$d(m,U) = \omega_1 d_t + \omega_2 d_i + \omega_3 d_a, \qquad <2>$$

where m is the media element, U is the current user context and where ω_i represent the normalized weights. A media concept is considered close to the user if its distance to the user context was less than a threshold. This threshold was chosen by experimentation.

3.3 Context evolution

As the user interacts with the environment, the user context changes based upon the media consumed and the time spent. We use our prior work on context evolution [7], according to which context evolves over time in a way that is analogous to the human memory. For each collage visited by the user, the system creates a semantic net from the media; it also measures the time spent by the user on the collage. This results in certain new concepts being introduced in her user context (newly gained knowledge), certain concepts getting reinforced (due to associations with similar concepts in the user's context), and certain other concepts to decay (put behind over time).

4 DYNAMIC PRESENTATION SYNTHESIS

In this section we discuss the process of multimodal presentation synthesis. For any given media set and a presentation scheme (collages, in our case), there are various ways in which the collection can be organized and structured - spatially and temporally. The process of presentation synthesis requires the following to be determined – (a) the media elements to be displayed, and the (b) the presentation order and duration.

There are three key components that affect the synthesis – (a) the specific goals that the user is interested, (b) the current state of the user context and (c) the specific knowledge structure of the environment – i.e. the system of relationships amongst concepts.

4.1 Optimal Media selection

In this section, we discuss the criteria for selecting media elements for the next collage. In our multimodal environment, a concept is represented by a set of media elements and their interrelationships.

Concepts are represented using *media sets*. The semantics of a media element depends upon other media elements present along with it. A singular concept could be represented using various media elements. These elements have inter-relationships that represent a single concept and should occur together – spatially and temporally – to be able to convey the concept. Hence, the media elements are grouped together into co-occurring media sets such that each set represents a particular concept. This is done by an expert in the current system. Note that a concept may have multiple media sets associated with them.

A set of concepts in a collage can be represented by a large number of media sets. However, to represent concepts in a collage the media sets are chosen to maximize coherence – this is done by minimizing the distance with respect to the current user context as well as being relevant to the current spatial location (e.g market, mountains etc.) on the map.

A media set that represents concepts close to the user's context is relevant to him or her. The distance between a media element and the user context (ref. section 3.2) defines the media's "closeness" to the user context. We set a threshold on distance for each of the three media types and chose those media elements whose distance from the user context falls below this threshold. Each media element in the media repository is associated with a location. We keep track of the position of the mouse on the map and pick only those media elements that are associated with that location.

4.2 PRESENTATION ORDER

In this section we discuss how the media elements are ordered in time when presented as a collage. Let us assume that the user is interested in exploring the environment, but has a set of goal concepts in mind – we assume that they form a proper subset of the knowledge in the environment. It is clear that we need to ensure that while the user is exploring the environment, the presentation must be coherent, make progress towards the goal, and additionally reflect the knowledge structure in the environment (ref. section 2.1).

The knowledge structure in the environment is exploited in the following manner. We maximize the number of those *critical* concepts (ref. section 2.1) that are related to the user goal concepts. Within each critical concept, there are sub-concepts that have a logical order e.g. if concept A is needed to understand concept B, then it must be shown first, where A and B are sub-concepts of the critical concept. Secondly, the sequence of critical concepts shown to the user, also constrained by the knowledge flow graph.

With the current context state, the goals and the knowledge flow graph known, we pick the optimal media sequence with the additional constraint of minimizing the distance of the user context to the goal concepts. We then use the following procedure:

1. For the user goal concept, choose the critical concept, C_1 closest to the user's current context and determine the subtree of the KFG where C_1 is present.
2. Present to the user, media representing the concepts in the KFG in the order of knowledge progression. Note that this media is chosen from the set of media elements picked using the media selection procedure (ref. section 4.1).
3. It is possible that additional constraints (e.g. location, knowledge progression order) cause the media set associated with the currently picked concept C_1 to be *not* displayed. In this case the environment indicates to the user additional locations in the environment that need to be explored, that are logical antecedents of the current concept. This ensures that the currently selected media set may be displayed at a later stage.
4. If the user chooses to stay on in the same location, present to the user the non-critical concept closest to her context.
5. As the presentation progresses reevaluate the importance scores of the goal concepts dynamically. The new importance score is given by:

$$I_{new} = I_{old} * \frac{C_{seen}}{C_{present}} \qquad <3>$$

where I_{old} is the initial importance score of the goal concept, C_{seen} is the number of critical concepts seen and $C_{present}$ is the total number of critical concepts for that goal concept.

This process is repeated as long as the user interacts with the environment. Note that *as the user interacts with the system, the user context is being dynamically updated, thus critically affecting the media selection procedure*. The change in user context will affect all media / concept distances. In this work, we use related work [8,9] on joint media presentation duration models. Briefly, the media duration are related to the Kolmogorov complexity of the media. Experimental studies help establish the mapping between media complexity and presentation duration..

5 EXPERIMENTS AND EVALUATION

In this section we discuss our experiments with users and the experimental results. The environment was created using Visual J#. We used an annotated media repository picked by an expert for our presentation. The expert also provided us with the knowledge flows between the various possible concepts represented by the media set.

We conducted two experiments that analyzed the qualitative and quantitative aspects of our proposed system. The user goal was a set of concepts related to water, a key factor of population density. Water is a proper subset of the entire environmental knowledge space. This goal, became part of the user context.

For the first experiment, four different exploration environments were created: (1) the proposed dynamic presentation synthesis, (2) a system without incorporation of user-context, but where the environment was structured (3) the system without incorporation of the knowledge flow graph, but where user-context was incorporated and (4) a presentation system with a totally random presentation order.

We evaluated our models through a pilot user study with five users, all graduate students at ASU. They were asked interact with all four systems and compare them in terms of coherence (i.e. intelligibility) with respect to learning the goal concept and comprehension (i.e. proper knowledge progression) of concepts presented. The users were asked to rate these aspects of the system on a scale of 1-7, 1 representing strongly disagree and 7, strongly agree. The experiment was double blind – neither the authors or the users were aware of the order, or the type of the presentation shown. The results obtained are tabulated below:

Table 1: Average Rating of users for evaluating the quality of presentation, for four different presentation scenarios. Only the first and the third presentations incorporate the user context.

Environments	Coherence	Comprehension
Optimal System	**5.75 / 7**	**4.50 / 7**
No context	2.50 / 7	4.25 / 7
No KFG	5.50 / 7	4.50 / 7
Random	1.25 / 7	1.25 / 7

It is interesting to note that the users felt that the coherence of the systems without context (2nd and 4th rows of Table 1) to be very low. This demonstrates the importance of context (present in the optimal and the system with the KFG; 1st and 3rd rows of Table 1) in the presentation system. Also noticeable is that there is little difference in the comprehension values between the systems with and without KFG. We conjecture that this represents a "mismatch" between the knowledge due to the expert, present in the environment and the user understanding of the concepts of population-density. i.e. the users do not agree with the knowledge flow sequence fixated by the expert or the strength of the knowledge relationships.

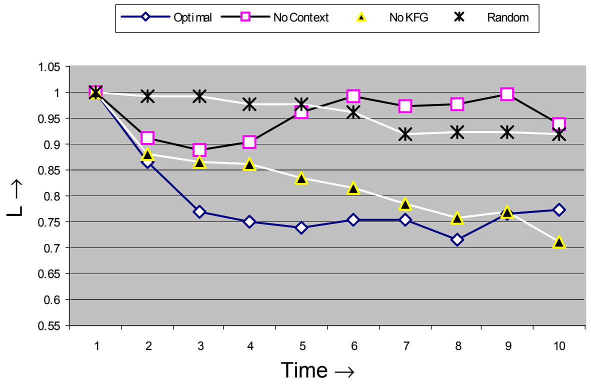


Figure 3: A plot of the norm of the distance to the target concepts against interaction time in minutes for the four presentations. The optimal presentation performs best.

In the second experiment (ref. Figure 3), the extent of learning (concepts imbibed) by the user with respect to the goal was measured for the same four user interactions as mentioned above. Each user interacted with the system for 10 minutes each. We computed the presence of the goals concepts by computing a simple weighted norm L on the concepts:

$$L(\vec{\boldsymbol{\omega}}) = \vec{\boldsymbol{\omega}}^{t}\mathbf{S}\vec{\boldsymbol{\omega}}, \qquad \text{<4>}$$

where $\boldsymbol{\omega}$ is defined as $[1\text{-}\omega_1,\ 1\text{-}\omega_2\ ,\ \ldots,\ 1\text{-}\omega_n]$ where ω_i is the weight of i^{th} goal concept in the user profile, and where **S** is a similarity matrix of the goal concepts.

The experiments indicate that the optimal system performs the best. It is very interesting to note that the *system with the user context but without the knowledge flow graph, also performs well.* However, note that the random case and the case with the knowledge flow graph, do not perform well. Clearly, this is an indicator that context is crucial in adaptive presentation systems, and that the presentation differences are clearly discernable.

6 CONCLUSIONS

In this paper, we presented a framework for a dynamic, multimodal, and context aware automatic presentation environment. We also presented models for: (a) multimodal user context, (b) optimal media selection to maximize coherence and (c) a presentation synthesis mechanism to enable users to explore concepts in geography. The experimental results are very good indicating that user context models can significantly improve the quality of the presentation. We plan to develop more sophisticated user context models, as well as allow user modification to the environmental knowledge by allowing new relations.

7 REFERENCES

[1] *Merriam Webster Dictionary* http://www.m-w.com.

[2] D. BIRCHFIELD (2003). *Generative Model for the Creation of Musical Emotion, Meaning, and Form*, ACM SIGMM 2003 Workshop on Experiential Telepresence, Berkeley, CA.

[3] M. G. CHRISTEL, A. G. HAUPTMANN, H. D. WACTLAR, et al. (2002). *Collages as dynamic summaries for news video*, Proceedings of the 10th ACM international conference on Multimedia, 561-569, Dec. 2002, Juan Les-Pins, France.

[4] A. K. DEY (2001). *Understanding and Using Context.* Personal and Ubiquitous Computing Journal **5**(1): 4-7.

[5] G. A. MILLER, R. BECKWITH, C. FELLBAUM, et al. (1993). *Introduction to WordNet: An On-line Lexical Database.* International Journal of Lexicography **3**(4): 235-244.

[6] L. R. RABINER and B. H. JUANG (1993). Fundamentals of speech recognition. Prentice Hall Englewood Cliffs, N.J.

[7] H. SRIDHARAN, H. SUNDARAM and T. RIKAKIS (2003). *Computational models for experiences in the arts and multimedia*, 1st ACM Workshop on Experiential Telepresence, in conjunction with ACM Multimedia 2003, Nov. 2003, Berkeley CA.

[8] H. SRIDHARAN, A. MANI and H. SUNDARAM (2005). *A Multimodal Complexity Comprehension-Time Framework For Automated Presentation Synthesis*, Proc. International Conference on Multimedia and Expo 2005, also AME-TR-2005-03, Jan. 2005, Amsterdam, The Netherlands.

[9] H. SUNDARAM and S.-F. CHANG (2001). *Condensing computable scenes using visual complexity and film syntax analysis*, Proc. IEEE International Conference on Multimedia and Expo, pp. 273-276, Aug. 2001, Tokyo, Japan.