Modeling User Context with Applications to Media Retrieval

Ankur Mani Hari Sundaram
Arts Media and Engineering Program
Arizona State University
e-mail: {ankur.mani, hari.sundaram}@asu.edu

Abstract

In this paper, we develop a theoretical understanding of multi-sensory knowledge and user context and their inter-relationships. This is used to develop a generic representation framework for multi-sensory knowledge and context. A representation framework for context can have a significant impact on media applications that dynamically adapt to user needs.

There are three key contributions of this work: (a) *theoretical analysis*, (b) *representation framework* and (c) *experimental validation*. Knowledge is understood to be a dynamic set of multi-sensory facts with three key properties – multi-sensory, emergent and dynamic. Context is the dynamic subset of knowledge that affects the communication between entities. We develop a graph based, multi-relational representation framework for knowledge, and model its temporal using a linear dynamical system. Our approach results in a stable and convergent system. We applied our representation framework to a image retrieval system with a large collection of photographs from everyday events. Our experimental validation with against two reference algorithms indicates that our context based approach provides significant gains in real-world usage scenarios.

Keywords

User context, media retrieval, multi-sensory knowledge representation

1. INTRODUCTION

In this paper we develop a representation framework for multi-sensory knowledge and user context, and apply this framework in a media retrieval application. The development of a generic representation for context is an important problem in multimedia. Intuitively, context is critical in explaining the "*why*" in communication – i.e. it represents the set of concepts that fully explain the concept that is contained in the message. Thus context play an important role in the exchange of messages in any communication or interaction. In face-to-face human communication, contextual cues play a key role [13,31], allowing us to act and participate in conversations in a rich manner. Similarly, a flexible representation framework that additionally addresses temporal evolution of context, can lead to powerful media applications that gracefully adapt to user interaction.

In this paper, we focus on developing a theoretical understanding of context, and its relationship to multi-sensory knowledge. Then we develop a graph based representation of knowledge, and show how we can model context as a dynamic subset of knowledge. We apply this framework on a photo browsing application with a large database of photographs from ordinary everyday events. We now summarize the unique and novel contributions of this paper are as follows.

- **Theoretical analysis:** We present a detailed understanding of multi-sensory knowledge and context in the communication between two entities. We define knowledge as a dynamic set of multi-sensory facts. There are three key aspects of knowledge – (a) multi-sensory, (b) emergent and (c) dynamic. We define context as "*the finite and dynamic set of multi-sensory and inter-related conditions that influences the exchange of messages between two entities in communication*." Context forms a dynamic subset of multi-sensory knowledge that is central to the semantics of the communication.
- **Representation framework:** We develop a graph based representation of multi-sensory knowledge. Each concept is a node in this graph, and a pair of nodes can have multiple relationships (feature based as well as semantic and each relationship represented as a weighted edge). Context is represented as the subset of nodes in attention that affect the communication. A key innovation is the development of a *linear dynamical system* to model the evolution of knowledge. We show that the system is stable and convergent.
- **Experimental validation:** We have applied our representational framework to a practical image retrieval application. The dataset comprised a large collection (~4000) of personal photographs of ordinary everyday events. This dataset is challenging as it is not organized into discrete categories, but is a better indicator of real-world results. We tested our algorithm of context aware retrieval against relevance feedback [27] and random browsing (baseline). Our user studies with graduate students indicate that our approach has significant gains.

It is important to note that the specific attributes of context *will always be application dependent,* driven by the real-world application needs. Our research instead focuses on a generic representation framework that allows for a systematic mechanism to model knowledge with *arbitrary*

attributes and their temporal evolution. This allows our contributions to context and knowledge be applied to other applications.

The rest of this paper is organized as follows. In the next section we begin by developing our notion of context – we provide real-world examples and then define context, multi-sensory knowledge and their inter-relationships. In section 3, we discuss related work. In section 4, we develop the key contribution of this paper in terms of a context model for media retrieval. In section 5, we present the architecture of our photo retrieval application. In section 6, we present our experiments on a large real-world data set and in section 7, we discuss the limitations of this work. Finally in section 8, we present our conclusions and future work.

2. MULTI-SENSORY CONTEXT

In this section we introduce the definition of multi-sensory and discuss some of its properties. We first provide two examples that demonstrate the context and the role of context. Then, we define context, multi-sensory knowledge and its relationship to context.

2.1 Examples of Context

We now discuss two user tasks and determine context in each task and then discuss the role that context plays in fulfilling the tasks.

Media retrieval: In the first task, the user wants to search photographs in a media database containing photographs, music, movies and others. The key problem here is to deliver to the user a set of media of elements most relevant to the user. The user provides query as either text or other by selecting media through an interface that allows for both modalities.

There are concepts and events that help in the query formulation. The user has some information about the set of events where such photographs were taken or the people or things visible in the photograph or how the photograph may look like. This set of information may provide a vague query, however most times it is difficult to provide query this way. Thus, the query is related to the current set of information in the user's short-term memory [4]. The user associates the photographs to other available media in a unique way. The strengths of these associations change with time and are related to the way the user associates things in his short-term memory. The multi-sensory and interrelated information set in the user's short-term memory influences the query provided by the user [4] and at the same time is influenced by the user's activity and the media the user consumes. *This set of multi-sensory and interrelated information forms the user query context*. The system can use estimates of the query context to provide a more comprehensive set of results as a response to the query. In this example attributes that do *not* form the query context include the temperature of the room, the sound of the air conditioner, or the last news article read.

Human Communication: In the second example, let us examine the case of two people Mary and Jane engaged in a face to face communication about a legal contract. In this case, Mary's spoken communication is highly dependent on several factors. It depends on the interaction immediately prior to Mary's spoken utterance, as well as multi-sensory cues such as Jane's gestures, clearly discernable affect states (e.g. anger). Mary *incorporates estimates of Jane's interpretation context* when composing her next spoken utterance – Mary's *context of construction*, is thus influenced by Jane's *context of interpretation*, as well as their past message history. Examples of attributes in this conversation that are *not* contextual include the color of the sky, the name of Mary's brother, or the name of Jane's graduate school – these attributes are part of the overall knowledge but are not relevant to the conversation

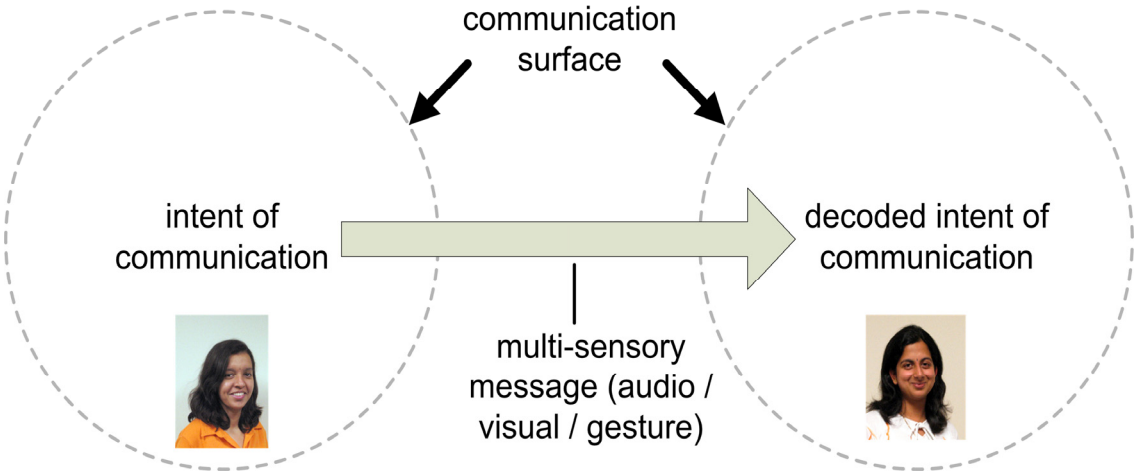

Figure 1: The general framework of multi-sensory communication between two entities. The intent of communication is dependent on the context of construction while the decoded intent depends upon the context of interpretation. We can draw a *communication surface* around each entity, through which all communication to the entity must pass.

about the legal contract.

Generalization: These examples can be generalized into the communication problem between any two entities [16]. When one person sends messages to the other person using a set of the available media, the messages are influenced by the person's current short-term memory, the task at hand, the messages the person has sent so far, the messages the person has received and the person's knowledge about the subject and the other person. This set of conditions is multi-sensory (represented in more than one medium) and inter-related and influence the messages originating from the communicator forms the context of the communicator. The interpretation of these messages by the other person (receiver) also depends upon a similar set of conditions for the receiver (ref. Figure 1). This set of conditions forms the context of the receiver. The more is the overlap between the contexts of the two people, the more effective is the communication. Note that the context of construction and the context of interpretation always represent a small part of the overall knowledge available to both entities. The communication is also influenced by the environment knowledge and the current state of the environment. Social knowledge such as language and the social codes form a part of the environment knowledge and influence communication [16].

2.2 What is multi-sensory Context?

We define context as "*the finite and dynamic set of multi-sensory and inter-related conditions that influences the exchange of messages between two entities in communication*." This is the set of conditions that can be estimated by the receiver based upon the messages from the originator and the set of conditions that influence the messages originating from the transmitter of messages. We observe in the above examples that the conditions were inter-related, dynamic, emergent (i.e. lead to the formation of new knowledge) and multi-sensory and influence the origin and interpretation of messages.

In the above examples the set of conditions was a subset of a large number of possible inter-related conditions. The superset of these conditions is knowledge about the communicating entities and the environment. We now discuss the definition of knowledge and its relationship to the context.

2.3 Knowledge and its properties

Knowledge is a dynamic set of multi-sensory facts. A fact is a statement that holds true for an entity (communicators or environment). Knowledge has three important properties; it is multi-sensory, emergent and dynamic.

- **Multi-sensory**: The environment we live in represents knowledge in multiple senses. The color, sounds produced, the texture, the shape, the structure and organization of different components, all these represent knowledge about the object. The functions of objects and their interactions also represent knowledge. Similarly, in our minds, we represent knowledge in multiple modalities. When we talk of the concept "bird," we associate it with the visual representations of a bird, the chirping sound of a bird, act of flying and others.
- **Emergent**: With increasing interaction with the environment, we begin to develop *new* associations and gather facts previously not known to us. For example, consider a child walking on the beach for the very first time. The child who has never been exposed to the sea (visuals / sound / wind and their inter-relationships), or the specific nature of the sand, quickly begins to learn the properties of this environment and use it to navigate the beach.
- **Dynamic**: Through interaction with the environment, we refine and modify our understanding of the *existing* relationships between the different observed concepts. For example Mary may associate sunshine with happiness, in a region where there is very little sunshine, but this relationship may weaken if she moves to a country with abundant sunshine.

In a media retrieval scenario, knowledge can be divided into three overlapping sets namely user knowledge, environment knowledge and the application knowledge. The user knowledge consists of the set of interrelated facts about the user such as the user interests and the set of associations the user has. The application knowledge consists of the application code and the media database. The application knowledge also influences the user knowledge as the user learns new facts about the media while exploring the database. The environment knowledge consists of the physical and social environment the user is in and that influences the way user knowledge has developed and will change in future.

2.4 Context and its relationship to knowledge

Context is the dynamic subset of knowledge that is in attention and influences the exchange of messages between the entities in communication (ref. Figure 2).

When two entities are communicating, each entity maintains an estimate of the knowledge and context about the other entity and the environment. The accuracy of the estimate influences the effectiveness of the communication. For example, if one person knows that the other person is deaf, he/she will communicate with the other person using sign language. The importance of communication for a reliable estimate of context is valid even in the absence of a human in communication. For example, in a location based adaptive system such as a mobile phone that displays active map information, it is critical to have a continuous stream of information about the location of mobile phone provided by a Global Positioning System especially when the user is highly mobile. For simplicity of notation we refer to the

estimate of knowledge as 'knowledge' and the estimate of context as 'context' in the rest of the paper.

Not all the knowledge about the entities is needed for a limited set of message exchanges. In fact, only a limited

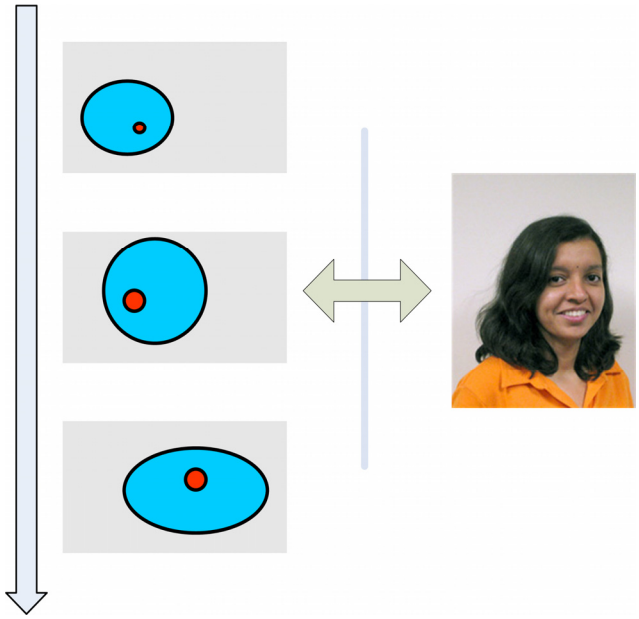

Figure 2: The figure shows relationship between knowledge, context with user interaction over a period of time. We see that the knowledge grows over time, and that context forms a dynamic, highly evolving subset of the knowledge.

subset of knowledge that covers all the possible contexts for the *specific communication* is sufficient. For example, when a child studying in first grade communicates with a physicist about birds, the physicist will not refer to his understanding of 'plasma physics' to communicate with the child. Additionally, the child need not be informed about the physicist's knowledge about 'plasma physics'.

In a media retrieval application the messages from the user are the selection of a particular media element and explicit textual query. These messages are influenced by the user's current attention and the way she associates different media elements. For the retrieval application the user knowledge is circumscribed by the media concepts, and the commonsensical relations connected to those concepts. An estimate of this knowledge by the application can make it more efficient. The environment knowledge can be limited to the semantic knowledge about the environment that influences the user interaction. This can be the linguistic, social, physical and commonsensical knowledge. The user always has some estimate of the knowledge and context of the application, such as how to construct the query, what to expect from the application. A user context-aware media retrieval application will also maintain an estimate of the user knowledge and query context. We next review related work.

3. RELATED WORK

Early work on context [10,11,12,28] treat context as pieces of information like location, identity, activity and time and has been successfully used for the purposes of configuration and adaptation is the areas of context aware ubiquitous computing. An attempt to integrate all such information was made in [9]. There the context was modeled for ubiquitous computing environments and hence was limited to the anticipated needs of the relevant ubiquitous computing applications. The context was defined as "any information that can be used to characterize the situation of an entity, where an entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and application themselves." This definition is very broad as pointed out in [34]. In [34], the author bounded the definition of context as the set of set of information that is relevant for the current communication. This definition is still broad and does not focus on two important properties of construct – (a) Context is a dynamic construct [14] and (b) Context is related to knowledge and cannot be discussed independent of it [26]. Our definition of context recognizes these important properties of context.

The generic model for context is not a well posed problem [17]. In [17] the authors pointed out the trade-off between the desire for abstraction (to make modular and tractable systems) and the desire for context-sensitivity. This also points towards an irony in making a generic context model (making a generic context model is actually motivated by this desire for abstraction) which is actually a drift away from the context sensitivity. *Thus the attributes of context can only be decided on a per application basis.* In this paper we focus on models of query context, with a model easily adapted to other applications. The needs of context in a media retrieval application are concerned with semantic inter-relationships between concepts, which can be arbitrary. Also the relationships that we wish to explore in multi-media are linguistic, statistical as well as common sense rules such as from the CyC database [2]. We address these issues in our work on context models.

The context model would thus consist of two structures:

- A knowledge base (personal or shared ontology) consisting of the concepts and the relationships between them
- A temporally evolving context representation that is in relation to this ontology.

Most recent attempts on knowledge representation and context modeling fall into one of the two categories namely the logical representation [7,21] and statistical representation [24,33] The logical representations are mostly driven by linguistic interpretations and the construction of meaning in language. On the other hand, the statistical knowledge representations are driven by pattern classification applications. For multimedia applications, we need a multi-modal knowledge

representation that integrates different approaches in single modalities. In [5], the authors proposed a multimedia knowledge framework involving arbitrary relationships between concepts. They suggested a learning framework in which the topology of the knowledge framework and the relationships between the concepts are statistically learned into a Bayesian network. This gives an ontology depicting purely arbitrary statistical relationships between the concepts but does not provide a meaningful interpretation and evolution of the knowledge framework. The other attempts to represent semantic knowledge include Wordnet [23], Cyc [16] and OpenMind Commonsense [20]. These knowledge representations focus on one or more relationships such as lexical relationships and commonsensical relationships.

Based upon such knowledge bases, several attempts have been made to represent and use context. This has been used in web-base learning environments, e-commerce applications, media recommendation applications, intelligent agents and others. However, the knowledge frameworks do not evolve dynamically with the user interactions in the system, nor do these models consider the personalized and shared nature of the multi-modal knowledge of each individual.

A significant survey on the content-based image retrieval has been given in [30]. Prior works on adaptive media retrieval have used relevance feedback [8,27] for a better query estimate and refinement of distance metrics in the low-level feature space. Recent work on retrieval has focused on context-aware similarities [35], however the work focuses on learning the similarity metric using a kernel trick and the model of context is not proposed. These works on media retrieval consider that image relationships exist in low-level feature space. However, most of the times people associate images semantically and not using low-level features. Media retrieval that uses semantics for retrieval [3,19] uses general and static knowledge bases like ConceptNet [20]. The problem with this approach is that it does not model the differences in the similarities between the media elements seen by different people in different contexts. We build a media retrieval system based upon the dynamic multimodal model of the user context as a subset of the personalized user knowledge. Our experiments show that media retrieval using the models of user context as in our approach produces better results as compared to relevance feedback as in [27].

Some recent media retrieval approaches [15,25] use a significant amount of prior knowledge for image retrieval. In [15], the authors use a pure metadata approach for image retrieval. The description of the images is in from the metadata of the images. Hierarchically, structured ontologies, obtained from the domain and from the media annotation define a set of properties of the images such as time, place and others. The query is also represented in a similar structure. Creation of such ontologies and in particular the annotation of all the images in the database is a significant problem. Often the image database is sparsely annotated. Hence, such approaches should be merged with content-based approaches (). The search and recommendations are then performed by a set of logical rules (recommendation rules, hierarchy rules and mapping rules). In [25] the authors present an interesting task based approach to define the context of the user query and the context of the images and their annotations. The authors also have a dynamic notion of task context that captures the user's task history. However, the approach is driven by annotation and relies on continuous annotation by the users and hence several tools are provided for the purpose of annotation in the application. Both the approaches [15,25] need significant amount of information from the user continuously. On the other hand, our approach makes limited demands from the user and grows the knowledge automatically by analyzing the user interactions. This is very crucial for keeping the users interested and at the same time provides opportunities for integration with incentive based strategies [29] for gathering information, for knowledge growth, from the user.

4. CONTEXT FOR MEDIA RETRIEVAL

In this section we introduce our model for representation and evolution of multi-sensory knowledge (text, images) and user query context.

4.1 Knowledge Representation

User knowledge in the model is represented as a graph as

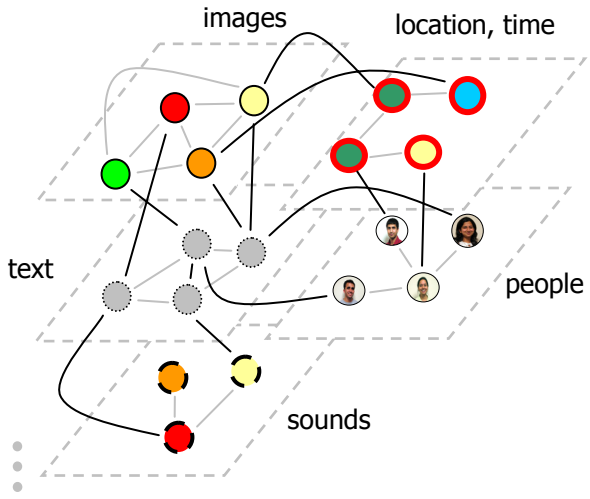

Figure 3: Multimodal User knowledge, nodes represent concepts and edges the relationships between concepts.

shown in Figure 3. The nodes in the graph are the instances of concepts in one modality and the weighted edges (weights represent the similarity between the end nodes along the edge) are the relationships between those instances. The knowledge can be divided into planes as shown in Figure 3, where each plane consists of concepts

in one modality and the relevant relations between them. The planes are connected through inter-planar edges between some nodes like an image in the images plane and its annotation in the text plane.

The relations between concepts in the text plane are linguistic and commonsensical, in the images and audio planes are low-level feature based, in the people plane are social and in the location and time planes are spatio-temporal. For now the knowledge is restricted to only text and image planes.

The description of the knowledge graph is as follows. Let C be the set of concepts in the knowledge and R be the set of relation types among the concepts in the knowledge. Each concept $c_i \in$ C and each edge $e_j \in$ E = C *x* C *x* R. Each edge e_j also has a weight $s_j(c_m,c_n)$ that denotes the similarity between the concepts c_m and c_n across the edge e_j due to the edge e_j. The weights of the edges in the text plane and the inter-planar edges are all set to 1 while the weights of the edges in the image plane are equal to the low-level feature based similarity between the images. The image plane is fully connected but the text plane is sparsely connected. The user knowledge is then defined as K = {C, E}.

Media knowledge is represented as a graph similar to the user knowledge and consists of the media elements and user specified annotations as the nodes and the feature based relationships (e.g. color histograms), commonsensical relationships and user specified annotation based relationships as edges. Environment knowledge at present is the commonsensical knowledge obtained from the ConceptNet [16] and it consists of text nodes and commonsensical relationships between those nodes.

4.2 User Query Context Representation

User query context is represented as the subset of the nodes and edges in the knowledge graph that are in attention. Although the user knowledge consists of a large set of concepts and relationships, at any time only a few of these concepts are in the user attention with different levels of attention. The concepts influence the user's interactions and the amount of influence depends upon their attention level. The association between the concepts is also different at different times and depends upon the level of attention on the relationships types connecting the concepts. An estimate of this set of concepts and relationships and their attention levels is represented as the user query context.

The attention on concepts and relationship types are represented as weights of the respective concepts and the biases on the types of relationships. The user attention at a concept c_i in the knowledge is modeled as the weight w_i of the concept. The sum of weights of all concepts in the user knowledge at any given time is constant. This is based upon the assumption that the amount of short-term memory in any individual and hence the total amount of attention given to all the concepts is constant over a small duration. The importance of a type of relationship $r_k \in$ R is modeled as the bias wr_k on the relationship type. The biases on the relationship types at any point are estimated as

$$wr_k = \alpha + 2(1-\alpha)\sum_i \sum_j w_i w_j \ , \qquad <1>$$

where w_i and w_j are the weights of the neighboring concepts c_i and c_j connected by the relationship r_k and α is a constant between 0 and 1. Thus wr_k lies between 0 and 1. Thus, a relationship type connecting concepts with high attention has a higher bias as compared to the relationship type connecting concepts with low attention levels.

The similarity between the neighboring nodes is the weighted sum (weights are the biases on the types of edges) of similarities between the nodes along all the edges

$$S(c_i, c_{i+1}) = \sum_{j=1}^{N} wr_j S(c_i, c_{i+1} \mid j) \ . \qquad <2>$$

$S(c_i,c_{i+1})$ is the similarity between the concepts c_i and c_{i+1}, N is the number of edges between the two concepts and $S(c_i,c_{i+1} \mid j)$ is the similarity of the j^{th} relationship (along the j^{th} edge) between any two nodes.

4.3 Context and Knowledge Initialization

Initial multimodal user knowledge is created using a set of multimodal (text, images and possibly other media such as audio) seed concepts provided by the user and the relation between the concepts either provided by the user or already present in the environment knowledge space. Each user provides some seed concepts describing herself such as her profession, interests, some images of herself and her friends, place and others. A set of concepts are extracted from the environment knowledge space (ConceptNet [1,20]) that are in the neighborhood of the textual seed concepts. The relationships between these concepts are also extracted from the environment knowledge space. All the seed concepts and extracted concepts and the relationships between them form the initial User Knowledge. Initially the whole user knowledge is set as the user context with equal weight of the nodes and the edges. This is a conservative estimate of the user context as the system does not have any information about the user's current context. Thus initially, each concept in the user query context has equal weight.

4.4 Context Evolution

We now discuss the evolution of the user context as the change in the weights of the nodes and biases of the relationships and discuss the convergence and stability of the model and intuitively discuss the appropriateness of the model. Earlier work on determining context of concepts in ontology has used activation spreading [7,8]. Thus the context in these cases is static as opposed to real situations where the context changes with time. We introduce a method for modeling the dynamics of activation spreading that helps us determine and track context through time both

in the presence and absence of any information about the user activity.

The change in weights of the nodes and edges follow the change in attention on the concepts and relationships in the short-term memory and are modeled as a linear dynamical system. The weight of a concept in the knowledge is affected by the weight of its neighbors. A concept strongly connected with other concepts with higher attention, gains attention while a concept strongly connected with concepts with lower attention, loses attention. The user activity related to a particular concept increases attention at the concept at the cost of the attention over all other concepts in the knowledge. Thus the rate of change of weight of a concept is

$$\frac{dw_i}{dt} = \sum_{j=1}^{N} C(S(c_i, c_j))(w_j - w_i) + A(i) - Bw_i \quad <3>$$

where w_i is the weight of the concept i at time t, $S(c_i,c_j)$ is the similarity between the neighboring concepts as given by equation <2>, c_i and c_j and $C(x)$ is a bounded monotonically increasing function. $C(S(c_i,c_j))$ represents the coupling between two neighboring concepts that causes the flow of weights between the two concepts. $A(i)$ is the activation function that is non-zero only for a certain set of concepts related to the current user activity and B is a constant that is related to $A(i)$ as

$$\sum_{i=1}^{N} A(i) = B . \quad <4>$$

$A(i)$ and B model the attention that is diverted from the existing set of concepts to the new set of concepts representing the current user activity. The first term of equation <3> models the spreading of the attention from the points of attention across the knowledge space. The context converges asymptotically on a suitable distribution of weights over all the nodes and edges in the knowledge under the presence of the messages from the user.

We now discuss the steady state distribution of weights of the concepts. The sum of weights of all the concepts is constant. It can be verified using the equation <3> that

$$\sum_{i=1}^{N} \frac{dw_i}{dt} = 0 . \quad <5>$$

The state-space equations in <3> can be represented in the vector form as

$$\begin{aligned} \frac{d\mathbf{w}}{dt} &= (\mathbf{C}\mathbf{w} - diag(\mathbf{C1})\mathbf{w} - B\mathbf{w})) + \mathbf{A} \\ &= (\mathbf{C} - (diag(\mathbf{C1}) + B\mathbf{I}))\mathbf{w} + \mathbf{A} \end{aligned} , \quad <6>$$

where **w** is the vector of the concept weights, **C** is a matrix whose elements $c_{i,j}$ represent the coupling between the neighboring concepts c_i and c_j given as $C(S(c_i,c_j))$, **1** represents a unit column vector and diag(**x**) represents a diagonal matrix whose diagonal elements are the elements of vector **x**. **A** is the activation vector whose elements are $A(i)$. When no information about the user activity is available, **A** is a zero vector and B is 0. Thus equation <6> reduces to

$$\frac{d\mathbf{w}}{dt} = (\mathbf{C} - diag(\mathbf{C1}))\mathbf{w} . \quad <7>$$

C is a sparse and symmetric matrix and the non-zero elements correspond to the coupling between neighboring concepts. At steady-state $d\mathbf{w}/dt = 0$. Thus the steady-state solution lies in the right null space of the matrix **C** - diag(**C1**). The given matrix is the weighted adjacency matrix of the knowledge graph where the weights are the coupling values between the adjacent concepts. It can be easily verified that the rank of such a matrix **C** - diag(**C1**) is $N_R - N_C$: number of rows – number of connected components. Hence the dimension of the solution space (i.e. the null space of the matrix **C** - diag(**C1**)) is N_C. The steady state solution thus is the uniform weight distribution across all the concepts in any connected component. In the current representation of knowledge, there is only one connected component in the User Knowledge network. Hence under the steady state, all concepts have the same weight. Further, since the sum of the weights of all the concepts is constant, this weight is unique and is 1/Number of concepts. This is intuitive since if we do not have any information about the user activity for a long period of time, then the estimate of the context is the just the overall knowledge that we have about the user.

4.4.1 Analysis of Stability and Convergence

We now analyze the stability and convergence of the user context in the general case when the information about the user activity is available and **A** and B is not 0. A Linear Time Invariant system with the state space equation as in eq. <6> is stable if all the eigenvalues of matrix **C** - diag(**C1**) – B**I** are less than or equal to 0. The system is not time invariant because the coupling matrix **C** changes with changes to context. However the change in matrix is small enough to be neglected and a time-invariant approximation is still justified. It can be easily verified that diag(**C1**) – **C** is positive semi-definite given the elements of **C** are all positive. Adding a diagonal matrix to this matrix whose diagonal elements are all positive or 0 gives a matrix which is also positive semi-definite. Therefore, diag(**C1**) + B**I** – **C** is a positive semi-definite matrix with all eigenvalues positive or 0. Therefore, all eigenvalues of **C** - diag(**C1**) – B**I** are less than or equal to 0 and the system is stable.

In the general case when **A** is not **0**, in the steady state, the weights of the concepts are the solution to the equation:

$$(\mathbf{C} - (diag(\mathbf{C1}) + B\mathbf{I}))\mathbf{w} = -\mathbf{A} . \quad <8>$$

To determine if the dynamical system converges and an unique solution exists, let us consider the rank of **C** - diag(**C1**) – B**I**.

$$(\mathbf{C} - diag(\mathbf{C1})) - B\mathbf{I}$$
$$= V\Sigma V^T - B\mathbf{I} \quad ,<9>$$
$$= V(\Sigma - B\mathbf{I})V^T$$

where Σ is the diagonal matrix whose diagonal elements are the eigenvalues of (**C** - diag(**C1**)) and V is the matrix whose columns are the eigenvectors of (**C** - diag(**C1**)). We have discussed earlier that the rank of (**C** - diag(**C1**)) is N_R – 1 as the knowledge graph has only one connected component. Thus, one of the eigenvalues is 0. From the discussion on stability, we also found that the eigenvalues of (**C** - diag(**C1**)) are less than or equal to 0. Therefore, all the eigenvalues of (**C** - diag(**C1**) – B**I**) are negative (B is a constant non-negative number) and the matrix has full rank. Hence, the estimate of the user context converges on a unique distribution of the weights over the concepts. The context estimate peaks at a small subset of the knowledge closer to the concepts related to the information about the user activity. The sharpness of the peaks can be controlled by the value of activation vector.

An important property of our context model is that with suitable activation vector, any weight distribution across the concepts in the knowledge can be achieved. This is true because the dynamical system is controllable. To verify controllability, we use the Popov-Belevitch-Hautus test [6] which suggests that the dynamical system in eq. <6> is controllable if there exists no left eigenvector of (**C** - diag(**C1**) – B**I**) that lies in the left null space of **I**. Since, there exists no null vector for **I**, therefore the system is controllable. In this section we discussed the evolution of the user context and analyzed the dynamics of the context. We now discuss the evolution of user knowledge.

4.5 User Knowledge Evolution

New facts about the user are exposed to the system through interaction and at the same time the user is exposed to new media elements. Thus the user knowledge is an increasing set of concepts.

The user knowledge estimate changes as new facts about the user are discovered. Each activity of the user (for example a query) provides information about the user. If the query concept is not present in the user knowledge, the user knowledge is expanded as follows:

- If the query concepts are textual, use them as seed concepts, extract a set of concepts in the environment knowledge that are in the neighborhood of the query concepts. If the query concepts are images and have annotations, the annotations of the images or the annotations of the images form the seed concepts. If the images do not have annotations, add the images to the User Knowledge, no textual concepts or relationships are added.
- Obtain the relations between the new concepts and the existing concepts in the user knowledge. For textual annotations use the ConceptNet neighborhood and add the concept and the relationships. For the image features, calculate the similarities with other images that are part of the user knowledge.
- Add the new concepts and relationships to the user knowledge.

The user knowledge estimate is also expanded as the user is exposed to new media elements. The new media elements that the user encounters and their annotations are added to the user knowledge. The new concepts are linked to the rest of the knowledge through the feature level relationships between the new media elements and the older ones in the user knowledge.

4.6 Similarity of non-neighboring concepts

The similarity of two non-neighboring concepts in the knowledge graph is computed as discussed in this section. Note that the similarity between the concepts depends upon the current context.

In the first case, when two concepts are in the current context, the similarity between them is computed as the maximum similarity between the two concepts computed over all the paths between the two concepts. Let $p(k,m)$ be a path that begins with concept c_k and ends with concept c_m. Let $l(p)$ be the length of the path. Then,

$$S(c_k, c_m) = \max_{p(k,m)} \left(S(c_{n-1}, c_m) \prod_{i=1}^{n-1} w_i S(c_{i-1}, c_i) \right), \quad <10>$$

where $S(c_{i-1},c_i)$ is the similarity between the two successive concepts along the path and w_i is the weight of the concept c_i along the path and where $n = l(\mathrm{p})+1$, and $c(0) = c_k$. This maximization can be done using a variant of the Dijkstra's shortest path algorithm. Thus, the path with the high attention concepts and strong similarity between the concepts is the maximum similarity path.

In the second case, when one or both concepts are not present in the current context we enlarge the context to encompass both concepts. The context is enlarged by the evolution process as given by equation <7> without any activation. During the evolution process all the concepts would lose their weights by a certain amount and the similarity calculated this way will be less than the similarity if there was a path between them in the context region. Note that this evolution process is only used to compute the similarity between the concepts and does not reflect as the change in the user context.

5. A PHOTO RETRIEVAL APPLICATION

We applied our context model in a photo retrieval system to improve the retrieval efficiency. Retrieval of photographs in a large personal collection is an important problem due to the easy availability of image capturing devices. We now discuss the architecture and different components of the photo retrieval application.

5.1 Architecture of the application

The system block diagram (ref. Figure 4) shows three main blocks namely the user interaction, the context-aware search engine and the context model as discussed earlier. The application is multi-threaded for efficiency – the three main blocks operate as three different threads.

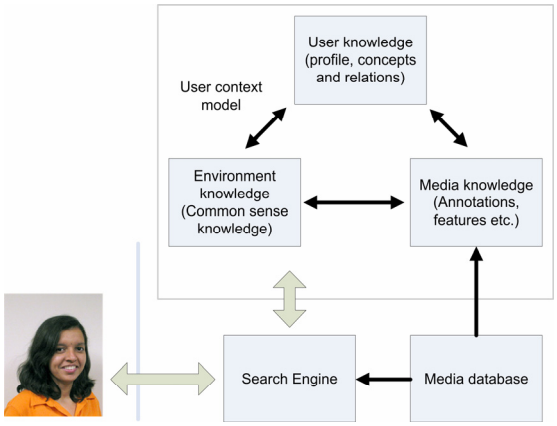

Figure 4: Block diagram of the photo retrieval application

The context model runs as a background thread while the User Interaction is the main thread. The search thread is launched when the user makes a query and ends when the results are obtained. The context model consists of the three different knowledge spaces namely the User Knowledge, the Environment Knowledge and the Media Knowledge. The User Knowledge is dynamic while the Media Knowledge and the Environment Knowledge are static. The user interactions provide information about the user and this information is used by the context model to update the User knowledge and the User context. The context evolution engine performs this task. The search engine uses the current user context and the media knowledge to provide the retrieval results. We now discuss the user interaction and the search engine.

5.2 User Interaction in the System

The user interacts with the system through an applet shown in Figure 5. In order to provide a text query, the query is typed in the text box in the upper left hand side of the screen. The displayed images can be selected by clicking the check box with the images. Finally the query is submitted by clicking the submit query button. The top nine results are displayed in the applet with the top most result also shown alongside. It has been studied that the human short term memory can efficiently process seven plus minus two units of information at a time [22], so we chose nine images to display as results.

The user time spent on the displayed pictures and the selection of pictures are the messages sent by the user to the system. These messages provide information about the current user activity and are used by the evolution engine to update the context.

5.3 Context-Aware Search in the Application

Given the query as a set of selected images, the context-aware search is performed in the media knowledge space to find the most relevant photographs. The search process first obtains the current context from the context model and modifies it using the user information obtained from the query. The modified context is then used to obtain the candidate concepts in the media knowledge space. The images close to the candidate concepts in the media knowledge space form the retrieval results. The complete search is as follows:

- Find the selected concepts from the selected images (the color histogram of the images and the annotations).
- Expand the user knowledge if the selected concepts are not present in the user knowledge.
- Evolve the user context with activation at the selected concepts. This evolution process is done by the search thread and does not reflect in the new user context model. The concepts (both images and text) in the user context that are also present in the media knowledge and have weight greater than α (optimized experimentally) of the highest weight form the set CC of candidate concepts.
- The score of an image in the media knowledge is now given as

$$S_c(I) = \max_k \{w_{cc} * S(I, CC_k)\}, \qquad <11>$$

where CC_k is the kth candidate concept and the $S(I, CC_k)$ is the similarity between the image and the candidate concept k in the media knowledge. The similarity is computed as in the user knowledge space as discussed in section 4.6. Nine images with highest scores are the result images. The displayed images form the new information viewed by the user. These images and their annotations are set as the current user information for the context evolution.

6. EXPERIMENTS AND EVALUATION

We now discuss the user experiments with the Photo retrieval application. We conducted a pilot user study with six graduate students. The experiments were conducted to demonstrate the effectiveness of the context-aware retrieval and hence the focus was not on using the optimal feature set.

The media repository was a set of 4000 images from the users' personal and shared photograph collection. Roughly 15% of these images were text annotated. Personal photographs makes the experiments more complex and better approximate real-world usage scenarios than the familiar Corel data set. Also importantly, the semantic associations are different for different users and we expect them to be captured by the User knowledge estimates.

The media retrieval application was implemented in Java and executed on Pentium 4 machine. The program structure is as discussed in section 4.

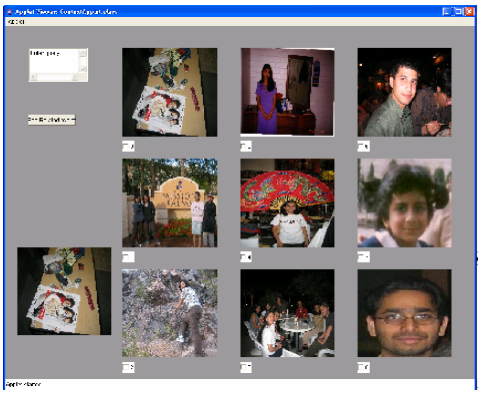

Figure 5: The user interface to the search application

6.1 Three Retrieval strategies

We implemented three retrieval strategies namely random retrieval, relevance feedback based retrieval and context-based retrieval.

- **Random:** In the random retrieval scenario, the images in the display set were selected at random. This strategy was selected as a baseline against which the other retrieval strategies were compared. The retrieval precision for this strategy should be close to the ratio of images in the database that are relevant to the query concept and is the minimum precision expected of any good retrieval strategy.
- **Relevance Feedback:** The relevance feedback based retrieval strategy was implemented as defined in the algorithm discussed in [27]. The color histograms of the images were the selected features for the experiments and the weights of the dimensions and similarities between the images were computed as discussed in [27]. However, for the consistency of the experiment and comparison with other approaches, we allowed only two levels of relevance score for each retrieved item. Thus a retrieved image is given a relevance score 1 if the user selects it as a relevant image. All other images in the database are given a relevance score 0.
- **Our Algorithm:** The context-based retrieval strategy used the context model and retrieved the images as discussed in section 4.

Thus, three experimental setups were made, each implementing one of the three retrieval strategies. The user interface and interaction instruction for the three setups were same for transparency to the user.

6.2 Description of Experiments

We now discuss the experimental set-up and the instructions given to the user. We required that each user provide a set of at least ten concepts. This set of concepts was the seed with which the user knowledge and context were initialized as discussed in section 4.3. Then each user was asked to follow the instructions given below:

- Select a query concept from among a set of choice concepts. These choice concepts were selected from the list of concepts in the media knowledge that had at least 100 relevant images. The first set of 9 images was selected randomly. We set the number of images to be nine due to the familiar property of short term memory [22].
- Once the images were presented, the users were asked to select among them, the images that were relevant, by selecting the associated checkboxes.
- The selected images were used as the query and the new images were presented. The images once shown are not repeated in the same set-up. This process is repeated four times. Thus the users see a total of 45 images.

Each user repeated the above experiment three times. Each time one of the three different search strategies namely, random retrieval, relevance feedback based retrieval and context-aware retrieval was used. The order of the experiments for each users was randomly selected and the order was not disclosed to the users.

6.3 Observations and Discussions

We now analyze the experimental results as both cumulative precision of the overall retrieved set and the change in the relevance score with increasing interaction and the personal priorities of different users.

6.3.1 Cumulative Precision

Table 1: Number of retrieved images for different queries and the % of relevant images in the database.

Query	% in database	Number of retrieved images in setup		
		Random	Relevance Feedback	Context
Home	10	6	12	14
Birthday	10	4	14	12
Graduation	3	1	9	14
Beach	5	2	15	16
Park	20	8	20	23
Office	5	1	8	10

We present the cumulative precision results as the number of relevant images that were retrieved in the complete experiment of five iterations and the mean relevance score of the retrieved images in the five iterations. The number of relevant images obtained in five iterations for three different search strategies are shown in Table 1.

We observe that the context-based retrieval gives the largest number of images relevant to the query. The number of relevant images retrieved using the random retrieval strategy depends solely on the percentage of relevant images in the database. It can be seen that the random retrieval strategy gives more number of relevant images for the query 'park' (20% images) as compared to the query 'beach' (5% images). The difference between the context-based retrieval and the relevance feedback method is not seen clearly as we do not consider the rank of the images.

A more revealing analysis of the results is possible by considering the rank of the retrieved images. The normalized relevance score for the retrieved set of nine images is computed as follows:

$$S = \frac{2}{N(N+1)} \sum_{j=1}^{N} (N+1-r_j), \qquad <12>$$

where N is the number of images and r_i represent the rank of the i^{th} image given as

$$r_i = \begin{cases} i & i \text{ is relevant to query} \\ 0 & \text{otherwise} \end{cases} \qquad <13>$$

Equation <12> helps us distinguish between two returned sets that have identical precision, as it takes the rank into account. The mean cumulative relevance score of the results for all the five iterations is shown in Table 2. The mean relevance score of the retrieved set for the context-based retrieval is significantly larger than the relevance feedback approach for all the queries. This demonstrates that context-based retrieval is more effective than the relevance feedback retrieval.

Table 2: Mean relevance score of the retrieved images for different queries

Query	% in database	Mean relevance score of retrieved image set		
		Random	Relevance Feedback	Context
Home	10	0.07	0.29	0.32
Birthday	10	0.07	0.24	0.38
Graduation	3	0.02	0.21	0.34
Beach	5	0.03	0.28	0.42
Park	20	0.16	0.44	0.51
Office	5	0.02	0.13	0.29

We now discuss the change in the relevance of the results with increasing user interaction.

6.3.2 Change in relevance with interaction

An important aspect of the context-based retrieval approach is that with increasing interaction more relevant images are retrieved. The relevance scores in different iterations for the three different search strategies are shown in Figure 6. The relevance score of the retrieved set is seen increasing with the increasing interaction in the context-based retrieval strategy. The relevance feedback based approach also shows and increasing trend but is not very consistent.

There are two reasons for the improved dynamic performance of the context-based approach against any other approach. Firstly, with increasing interaction, the estimate of the user context becomes more accurate and the

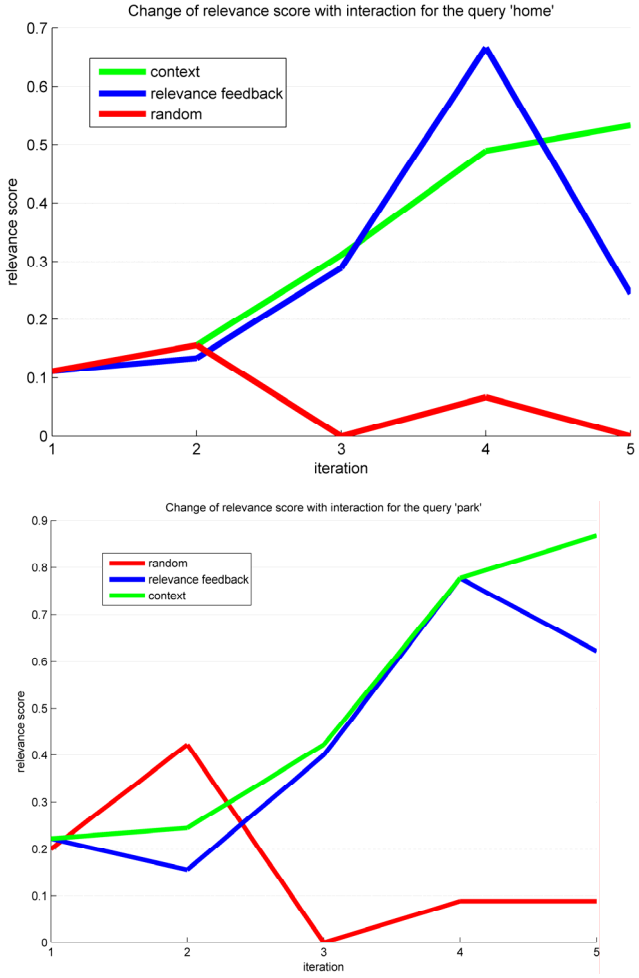

Figure 6: Plot of relevance score against user interaction for queries 'home' (top) and 'park' (bottom) queries.

retrieved set is therefore more relevant. Secondly, the context-based approach uses multi-modal relations (commonsensical relations using the annotations and feature based metrics) between concepts to estimate the similarities. This helps retrieving images that are more semantically similar to the query. We see an anomaly in bottom figure in Figure 6, where in the fourth iteration; the relevance score of the relevance feedback approach is higher than the context-based retrieval approach. This is an outlier and is ascribed to the human error in interaction.

7. LIMITATIONS

The use of the context model provided significant improvement in the photo retrieval over relevance feedback

base retrieval however, the model has the following limitations:

- The quality of the acquired user knowledge. For experimental purposes, we have used the concept net [20] to extract textual concepts and the commonsensical relationships between them while image concepts are derived from the media database itself. The textual concepts and their relationships in concept-net are not very dense and that limits the knowledge.
- The model needs continuous interaction with the user to have an accurate estimate of the user. In the absence of interaction, though the model maintains the knowledge about the user, until the past interactions, the estimate of the current context is very conservative.
- From the perspective of knowledge representation, while we provided a generic representation framework for knowledge in all modalities, the knowledge used in the application was limited to only text and image planes. In future work, we are planning to incorporate representations of human activities. Our current choice is guided by the needs of the image retrieval application.
- Knowledge is representable at multiple time-scales. We did not cover the multi-scale temporal nature of knowledge in our current framework, and plan to incorporate this into future work.
- Probabilistic: The presented framework does not incorporate a probabilistic knowledge representation framework. We are currently developing new statistical models to represent multi-sensory knowledge and user action context using partially observable decision markov processes [18,32].

We plan to address these limitation in future work.

8. CONCLUSIONS

In this paper, we presented a novel framework for modeling user context. We have developed a theoretical understanding of multi-sensory knowledge and user context and their inter-relationships. We presented our definition of context as "*the finite and dynamic set of multi-sensory and inter-related conditions that influences the exchange of messages between two entities in communication*." We stated properties of context and its relationship to knowledge and the application. We then presented a dynamic, emergent and multi-sensory knowledge representation framework in which context can be estimated and represented as a subset that is in focus. We modeled the dynamics of the context and knowledge as linear dynamical system. We provided proofs for the convergence, stability and controllability of the user context model under all circumstances. We tested the validity of our proposed user context model by applying it to a photo retrieval application. Our experiments demonstrate that the model of user context helps retrieve more relevant images for the user as compared to relevance feedback. The experiments also demonstrate the improvement in the relevance of the retrieved images with increasing user interaction.

We have identified that for a robust context model, we need a comprehensive framework for knowledge acquisition and representation. Hence, we plan to focus our future work mainly on improving the process of user knowledge acquisition, and multi-scale and multi-modal knowledge representation. We are also working on more exhaustive experiments with more users and a larger dataset and improved features to provide statistically more significant results.

9. REFERENCES

[1] *ConceptNet* http://web.media.mit.edu/~hugo/conceptnet.

[2] *OpenCyc* http://www.opencyc.org.

[3] P. APPAN and H. SUNDARAM (2004). *Networked multimedia event exploration*, Proc. ACM Multimedia 2004, also AME TR-2004-10, pp. 40-47, Oct. 2004., New York, New York.

[4] R. ATKINSON and R. SHIFFRIN (1968). *Human memory: A proposed system and its control processes*. The psychology of learning and motivation: Advances in research and theory(eds). New York, Academic Press.

[5] A. B. BENITEZ, J. R. SMITH and S.-F. CHANG (2000). *MediaNet: A Multimedia Information Network for Knowledge Representation*, Proceedings of the 2000 SPIE Conference on Internet Multimedia Management Systems (IS&T/SPIE-2000), Nov 6-8, 2000., Boston MA.

[6] C.-T. CHEN (1999). Linear Systems Theory and Design. Oxford University Press.

[7] W. CHRISTENSEN (2004). *Self-directedness, integration and higher cognition.* Language Sciences **26**(6): 661-692.

[8] I. J. COX, M. L. MILLER, T. P. MINKA, T. V. PAPATHOMAS and P. N. YIANILOS (2000). *The Bayesian Image Retrieval System, PicHunter: Theory, Implementation and Psychophysical Experiments.* IEEE Trans. Image Processing---Special Issue on Digital Libraries.

[9] A. K. DEY (2000). *Providing Architectural Support for Building Context-Aware Applications*. College of

Computing. Atlanta, Georgia Institute of Technology, PhD.

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

[11] A. K. DEY and G. D. ABOWD (1999). *Towards a Better Understanding of Context and Context-Awareness*, Proceedings of the 3rd International Symposium on Wearable Computers, pp. 21-28, October 20-21, 1999, San Francisco, CA.

[12] A. K. DEY, M. FUTAKAWA, D. SALBER and G. D. ABOWD (1999). *The Conference Assistant: Combining Context-Awareness with Wearable Computing*, Proceedings of the 3rd International Symposium on Wearable Computers, pp. 21-28, October 20-21, 1999., San Francisco, CA.

[13] P. DOURISH (2004). *What we talk about when we talk about context.* Personal and Ubiquitous Computing **8**(1): 19-30.

[14] S. GREENBERG (2001). *Context as a Dynamic Construct.* Human-Computer Interaction **16**: 257-268.

[15] E. HYVONEN, S. SAARELA and K. VILJANEN (2003). *Ontogator: Combining View- and Ontology-Based Search with Semantic Browsing*, Proc. of XML, October 30-31, 2003, Kuopio, Finland.

[16] R. JAKOBSON (1960). *Closing statement: linguistics and poetics*. Style in Language. T. SEBEOK. (eds). Cambridge MA, MIT**:** 350-377.

[17] H. LIEBERMAN and T. SELKER (2000). *Out of Context : Computer Systems that Adapt To, and Learn From, Context.* IBM Systems Journal **39**(3,4): 617-631.

[18] M. L. LITTMAN, A. R. CASSANDRA and L. P. KAELBLING (1995). *Learning Policies for Partially Observable Environments: Scaling Up*, Proceedings of the Twelfth International Conference on Machine Learning, July 1995, Lake Tahoe, CA.

[19] H. LIU and H. LIEBERMAN (2002). *Robust Photo Retrieval Using World Semantics*, LREC, Las Palmas, Canary Islands.

[20] H. LIU and P. SINGH (2004). *ConceptNet: a practical commonsense reasoning toolkit.* BT Technology Journal **22**(4): pp. 211-226.

[21] S. MAKARIOS and R. GUHA (2005). *A First Order Theory of Contexts*, Context 2005, Paris.

[22] G. A. MILLER (1956). *The magical number seven, plus or minus two: Some limits on our capacity for processing information.* Psychological Review **63**: 81-97.

[23] G. A. MILLER, R.BECKWITH and C.FELLBAUM (1993). *Introduction to WordNet : An on-Line Lexical Database.* International Journal of Lexicography **3**(4): 235-244.

[24] K. MURPHY and W. FREEMAN (2004). *Contextual Models for Object Detection using Boosted Random Fields*, NIPS'04,

[25] D. O'SULLIVAN, E. MCLOUGHLIN, M. BERTOLOTTO and D. WILSON (2005). *Context-Oriented Image Retrieval*, Context 2005, Paris.

[26] J.-C. POMEROL and P. BREZILLON (2001). *About some relationships between Knowledge and Context*, Context-01, Dundee, Scotland.

[27] Y. RUI and T. HUANG (1999). *A Novel Relevance Feedback Technique in Image Retrieval.*, Proc. ACM Multimedia 1999, Nov. 1999, Orlando, FL.

[28] B. N. SCHILIT and M. M. THEIMER (1994). *Disseminating active map information to mobile hosts.* IEEE Network **8**(5): 22-32.

[29] B. SHEVADE and H. SUNDARAM (2004). *Incentive Based Image Annotation*. Arts Media and Engineering Program, Arizona State University, AME-TR-2004-02, Jan. 2004.

[30] A. W. M. SMEULDERS, M. WORRING, S. SANTINI, A. GUPTA and R. JAIN (2000). *Content-based image retrieval at the end of the early years.* IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12): 1349-1380.

[31] L. A. SUCHMAN (1987). Plans and situated actions : the problem of human-machine communication. Cambridge University Press Cambridge Cambridgeshire ; New York.

[32] G. THEOCHAROUS, K. MURPHY and L. P. KAELBLING (2003). *Representing hierarchical POMDPs as DBNs for multi-scale robot localization*, Workshop on Reasoning about Uncertainty in Robotics, International Joint Conference on Artificial Intelligence, Acapulco, Mexico.

[33] S. VEERAMACHANENI, P. SARKAR and G. NAGY (2005). *Modeling Context as Statistical Dependence*, Context 2005, Paris.

[34] T. WINOGRAD (2001). *Architectures for Context.* http://hci.stanford.edu/~winograd/papers/context/context.pdf.

[35] G. WU, E. Y. CHANG and N. PANDA (2005). *Formulating Context-dependent similarity functions*, ACM Multimedia, 725-734, Singapore.