Experiential Sampling in Multimedia Systems

Mohan S. Kankanhalli, Jun Wang, and Ramesh Jain, *Fellow, IEEE*

***Abstract*—Multimedia systems must deal with multiple data streams. Each data stream usually contains significant volume of redundant noisy data. In many real-time applications, it is essential to focus the computing resources on a relevant subset of data streams at any given time instant and use it to build the model of the environment. We formulate this problem as an experiential sampling problem and propose an approach to utilize computing resources efficiently on the most informative subset of data streams. First, in this paper, we focus on theoretical background and develop a theoretical framework for a single data stream. We generalize the notion of static visual attention in a dynamical systems setting and propose a dynamical attention-orientated analysis method. This is achieved by a sampling representation that utilizes the current context and past experience for attention evolution. Hence, the multimedia analysis task at hand can select its data of interest while immediately discarding the irrelevant data to achieve efficiency and adaptability.**

***Index Terms*—Dynamical systems, experiential computing, experiential sampling, sampling, visual attention.**

I. Introduction

MULTIMEDIA information processing usually deals with spatio–temporal data which have the following attributes.

- It consists of a multiplicity of usually correlated data streams. Thus, it does not exist in isolation—it exists in its *context* with other data. For instance, visual data comes along with audio, music, text, etc.
- They possess a tremendous amount of redundancy.
- The data is dynamic with temporal variations with the resultant history.

However, many current approaches towards multimedia analysis do not fully consider the above attributes which lead to two main drawbacks—*lack of efficiency* and *lack of adaptability*. The inefficiency arises from the inability to filter out the relevant aspects of the data and thus considerable resources are expended on superfluous computations on redundant data. Hence speed-accuracy tradeoffs cannot properly be exploited. The lack of adaptability stems from the fact that the context of the data is often ignored. As a result, rigid computational procedures are employed for analyses that remain fixed when the environment itself is changing. Moreover, the context of multiple correlated data streams is not fully harnessed in order to perform the task at hand.

On the other hand, we have solid evidence that humans are superb at dealing with large volumes of disparate data using their sensors [3]. For instance, the human visual system is quite successful in understanding the surrounding environment at an appropriate accuracy quite efficiently. This is due to many factors [9]: the excellence of the physical visual sensing system, the richness of fusion information from perception, implicit understanding of every visual object, and the common understanding of how the world works. These attributes in the *experiential environments* [4] play an important role for the human visual perception to understand the visual scene accurately and quickly under fairly adverse conditions. The vision for experiential computing was introduced in [4], which envisages that multimedia analysis should also have the ability to process and assimilate sensor data like humans. Examples of such problems being currently tackled are speaker recognition [20], speech event detection [20], speaker change detection [20], monologue detection [21], and cross-modal information retrieval [22]. Many tasks, like remote monitoring, understanding semantics, and adaptive presentations, also fall under this paradigm. Therefore, we would like to articulate the following goal for such multimedia systems:

> "*In an experiential computing environment, the system should sense the data from the environment. Based on the observations and experiences, the system should collate the relevant data and information of interest related to the task. Thus, the system interacts naturally with all of the available data based on its interests in light of the past states in order to achieve its designed task.*"

Our ideas are articulated using some important concepts that Neisser [19] introduced in 1976 in his work on the notion of perceptual cycle to model how people perceive the world. He presented the idea that a perceiver builds a model of the world by acquiring specific signals and information to accomplish certain tasks in the natural environment. The perceiver continuously builds a schema that is based on the signals that he has received so far. This schema represents the world as the perceiver sees it at that instant. The perceiver then decides to get more information to refine the schema for accomplishing the task that he has in mind. This sets up the cycle as shown in Fig. 1. The perceiver gets signals from the environment, interprets them using the current schema, uses the results to modify the schema, uses the schema to decide to get more information, and continues the cycle until the task is done. To formulate the problem precisely, we will define the scenario more formally in Section II.

Due to space constraints, we split our presentation over two papers. In this paper, we only concentrate on the theoretical development of the technique on single data stream while we pro-

Manuscript received April 28, 2004; revised December 4, 2005. The associate editor coordinating the review of this paper and approving it for publication was Dr. Benoit Macq.

M. S. Kankanhalli is with the School of Computing, National University of Singapore, Singapore 117543 (e-mail: mohan@comp.nus.edu.sg).

J. Wang is with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft 2628 CD, The Netherlands (e-mail: jun.wang@tudelft.nl).

R. Jain is with the Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA 92697-3425 USA (e-mail: jain@ics.uci.edu).

Digital Object Identifier 10.1109/TMM.2006.879876

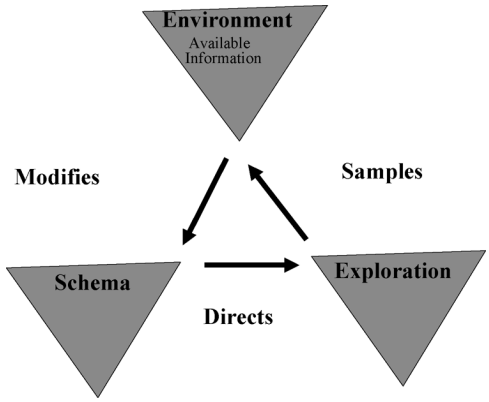

Fig. 1. Neisser's perceptual cycle. (Based on [19, Fig. 1]).

vide its generalization on multiple data streams, related work, and the experimental results in [26].

II. Experiential Sampling

A. Defining Experiential Sampling

Experience is defined as the accumulation of knowledge or skill that results from direct participation in events or activities [23]. Direct participation implies having access to the environment of the event in order to observe it using all potential sensory mechanisms available to the perceiver or the experiencer. In such an environment, the experiencer is driven by the goal of maximizing the efficacy of building the schema with minimal efforts to accomplish the most efficient mechanism to accumulate the knowledge. This task translates into selecting appropriate data streams at any given time, based on the current schema, for paying attention.

We define experiential sampling as the process of identifying the most relevant data stream among the available streams at a given instant to utilize for interpretation to refine the current model of the environment.

In this section, we introduce our experiential sampling technique. There are two major components in this technique. The first is how to sense and fuse experiences (contextual information) in the experiential environment. The second is how to build a dynamic attention model to select the data (or region) of interest.

1) Experience: Our definition of experience is based on [4].

Experience in Multimedia Analysis: is any information that needs to be specified to characterize the current state of the multimedia system. It includes the current environment, *a priori* knowledge of the system domain, current goals, and the past states.

Although experience and experiential environments are domain dependent and their components are not clear in general, we define three main components as follows.

Current contextual information: is the current existing information about the environment that needs to be specified to characterize the current state of the multimedia system with respect to the current goal.

Past experience: is the accumulated experience of the multimedia analysis task performed in the past.

Goal: is the purpose of the current analysis task. It is used to define what the related experiences are, and what analysis technique should be employed to accomplish the task.

There are some relationships among these components. The current contextual information can be characterized by features extracted from the visual scene and other accompanying multimedia data (audio, speech, text etc.). The current goal and prior knowledge provide a top-down approach to analysis. It also determines which features of the visual scene and other accompanying data type should be used to represent the environment. The past experiences encapsulate the experiences till the current state. These relationships can help us defining the experiential environment when we perform multimedia analysis. More importantly, when we consider the experiential environment, the analysis process systematically integrates the top-down and bottom-up approaches by employing the context and history.

2) Goal Oriented Attention From Experiential Environments: As mentioned in the introduction, we allow the system to sense the data from the experiential environment. Based on the observations and experiences, it collates the relevant data and information of interest related to the task of the analysis and discards irrelevant information. In this regard, a central problem is the allocation of the goal oriented attention within the experiential environments. Note that attention is intimately related to the goal—generic attention does not make sense. We base this discussion on video which is a prototypical multimedia data type.

In our framework, we allow the analysis task to guide the attention onto regions or data of interest from the entire spatio–temporal data. We first introduce a vector to represent the spatial position of the goal oriented attention in a given time t as

$$a_G(t) = [x, y]'$$

where $x = 1, \ldots, X$ and $y = 1, \ldots, Y$ are spatial coordinates and $t = 1, \ldots, T$ is temporal position. $a_G(t) = [x, y]'$ indicates the current a*ttended* position is $[x, y]'$ in a time slice t. $'$ denotes the transpose operator. Without loss of generality, the stream dimension $\{1, \ldots, n\}$ can be further added when multiple streams are considered, while the spatial coordinates x, y can be dropped when nonspatial streams are considered.

To infer the attention from the environment, we define the current contextual information with respect to the attention at the time t as:

$$e(t) = \{e(x, y, t) | x = 1, \ldots, X; y = 1, \ldots, Y\}$$

where again x, y are spatial coordinates and t is the temporal position. It includes any contextual information which could help in inferring the goal oriented attention (we will show later it is a combination of different feature cues). Therefore, it can also be considered as the measurement (e.g. motion, colors etc) of the attention with respect to the given spatial coordinates and time. For this, the values of the elements are required to be normalized to the range of [0,1). The sum total of accumulated contextual information for the attention is defined as $E(t) = \{e(1), \ldots, e(t)\}$.

In this paper, we attempt to infer the attention from the experiential environment. By employing probabilistic reasoning,

we define the *a posteriori* probability $P(A_G(t)|E(t))$ with $A_G(t) = \{a_G(1), \ldots, a_G(t)\}$ as the goal oriented attention up to time t. For real time applications, we need to estimate $P(a_G(t)|E(t))$ rather than $P(A_G(t)|E(t))$. Here we assume that the attention at each spatial position $\{x, y\}$ is only dependent on the context measurement around the position $\{x, y\}$. Then we have the following:

$$P\left(a(t) = [x, y]'|E(t)\right) = P\left(a(t) = [x, y]'|E(x, y, t)\right).$$

Note that this notion of attention is a generalization of visual attention [6] in the sense it can be applied to any multimedia stream which may be nonvisual. For example, this definition subsumes the notion of aural attention which is also related to the cocktail party effect in digital audio processing. And this generalized attention concept can be applied to non-visual, non-audio data as well. Also, it is a phenomenon which dynamically varies with time unlike the notion of static image attention dealt by the bulk of the visual attention literature. Moreover, attention is always goal-driven.

B. Goal Oriented Attention Driven Analysis

In this section, we formulate the goal oriented attention driven analysis by using the Bayesian framework.

1) Signal to Symbol Matching: The central problem of multimedia content analysis is the signal to symbol matching. Fundamentally, it involves mapping the relationships between the digitized spatial-temporal data and *semantic symbolic identity*. We define this mapping function as S_M. Many analysis approaches only unite the *local content intrinsic* features to perform content analysis. Here "local" and "intrinsic" refer to the fact that these features come from the information of the symbolic identity itself. By employing probabilistic reasoning, such analysis approaches, which we classify as *local feature centered approaches*, can be expressed as maximizing the *a posteriori* probability

$$SID = S_M(f_L) = \arg\max_H P(H|f_L) \tag{1}$$

where SID is the estimated true semantic symbolic identity, f_L denotes the local intrinsic features and H is the hypothesis of the symbolic identity. For instance, in face detection, the hypothesis is face region and non-face region. Note that in this section, since we only discuss the situation within a given time slice, we simply drop the entire notation related to time.

For instance, the local feature centered approach, which has been the dominant theme in computer vision for many years, exclusively uses object intrinsic features to represent the objects and to perform object detection/recognition tasks [1], [2], [17].

2) A Bayesian Framework for Integrating Attention: However, the symbolic identities physically exist in their environment and not in isolation. It is a well-known fact that focus of attention plays an important role in the human visual system to understand the visual scenes. It can selectively process the data that it observes or gathers based on the context. The illusions in Fig. 2 shows that the role of goal oriented attention in top-down visual system increases in importance and can become indispensable when the viewing conditions deteriorate

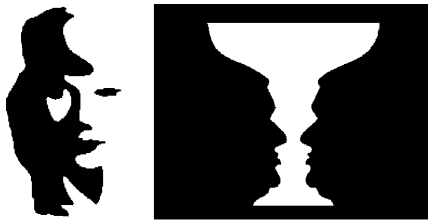

Fig. 2. Attention helps analysis. (a) Woman's face or a saxophone player. (b) Vase or head to head?.

or when ambiguity exists. In Fig. 2 (courtesy of http://members.lycos.co.uk/brisray/optill/othis.htm), if we look at the entire image (process all the data in the image), we maybe confused whether there is a saxophone player or a woman's face. However, if we just focus our attention on the dark region, we instantly identify that there is a saxophone player. Contrarily, if we focus our attention on the white region towards the right, it could convince us that it is a woman face. Similar ambiguity exists in the second illustration of Fig. 2 as well.

In some respects, in the visual scene, the object intrinsic features and their differences with respect to the global environment features make the object distinct from the environment. In the early vision of human brain, by making use of these features, goal driven focus of attention allows human visual perception to quickly become aware of objects of interest from large volumes of visual data in the visual environment [6]–[8]. Recently, Jepson *et al.* in [16] have stated that contextual information plays an important role to make reliable inferences in situations where the measurements produce ambiguous interpretations. Torralba [15] mainly interpreted the scene information as context and developed contextual priors for object detection.

Therefore, it is absolutely necessary to build in the attention phenomenon into the multimedia analysis process. Based on this, we extend the signal to symbol mapping function formulated in (1) by adding the attention A. Therefore, the multimedia analysis problem, as shown in (1) essentially becomes maximizing the symbolic identity's posterior probability $P(H|f_L, a)$. That is, the probability of identity H, given the current intrinsic feature F_L and the current attention a. According to this model, we will use a Bayesian reasoning framework to embed the attention and experiential environment E into the multimedia analysis tasks. Bayes' theorem can be used to factorize the probability $P(H|f_L, a)$

$$P(H|f_L, a) = \frac{P(f_L|H, a)}{P(f_L|a)} P(H|a). \tag{2}$$

The identity feature is directly affected mainly by the identity. There is very little influence coming from the attention. Here we assume that the local feature f_L is independent of the attention A. Therefore, (2) can be rewritten as

$$P(H|f_L, a) = \frac{P(f_L|H)}{P(f_L)} P(H|a). \tag{3}$$

Therefore, the probability of the hypothesis H given local feature f_L and the attention a is factorized into two components. The first component is the effect from the local feature f_L on

hypothesis H. The second component is the attention oriented priors on the hypothesis.

It can also be further factorized as follows:

$$P(H|f_L,a) = \frac{P(f_L|H)}{P(f_L)} \cdot \frac{P(a|H)P(H)}{P(a)} = \frac{P(f_L|H)P(H)}{P(f_L)} \cdot \frac{P(a|H)}{P(a)}. \quad (4)$$

In the end, we have the final equation

$$P(H|f_L,a) = P(H|f_L) \cdot \frac{P(a|H,E)}{P(a|E)} \quad (5)$$

where we treat attention in the experiential environment E. Therefore, we add the dependence of E in the probability of the attention. The numerator of the second component in (5) is the attention aroused by both the symbolic identity and its experiential environment. The denominator of the second component is the attention aroused by the experiential environment only. By this denominator, the attention aroused by the environment is inhibited. Therefore, we can see that these arousing and inhibiting attentions can contribute to the multimedia analysis task. We call this attention *goal-driven attention*. From Section II-A-1, our experiential environment E includes the goal. It means the goal about obtaining the symbolic identity SID has been considered in this framework. Therefore, we denote

$$P(a_G|E) = \frac{P(a|H,E)}{P(a|E)}. \quad (6)$$

We can now rewrite (1) as

$$\begin{aligned} SID &= S_M(f_L, a_G) \\ &= \arg\max_H P(H|f_L, a_G) \\ &= \arg\max_H P(H|f_L) \cdot \frac{P(a|H,E)}{P(a|E)} \\ &= \arg\max_H P(H|f_L) \cdot P(a_G|E). \end{aligned} \quad (7)$$

From the above equation, we can see that the final posterior probability has two components. The first component is the local posterior probability which can be inferred from the symbolic identity's local features. In general, local feature centered approaches exclusively concentrate on obtaining this probability. The second component is the impact coming from the goal-driven attention. This part serves as an amplification factor on the identity centered approach of the first component.

C. Sampling Based Dynamical Attention Driven Analysis

From above analysis, we can see that the attention helps the multimedia analysis task. Given that our task in this paper is identifying the most relevant data stream among the available streams at a given instant, based on the above discussion, we treat the information which makes the term $P(a|E)$ (for the sake of simplicity, we will drop the subscript G later on; however, $a(t)$ and $A(t)$ will *always* denote goal oriented attention) smaller as the *irrelevant information*. We discard it since we would not like to do the time-consuming processing [shown in (1)] on the irrelevant information which give a lower value for $P(a|E)$. Contrarily, we treat the information which gives higher value on $P(a|E)$ as the *relevant information* and perform detailed processing (to obtain $P(H|f_L)$) on it.

There are two steps involved in performing this attention driven analysis. Firstly, we use samples and their weights to dynamically maintain the attention with respect to the experiential environment. Secondly, we propose the use of a re-sampling approach to obtain relevant information captured in the samples, which is employed to perform the multimedia analysis task based on the attention. The (visual or otherwise) attention in a scene can be represented by a multimodal probability density function. Any assumptions about the form of this distribution would be limiting. However, not making any assumption about this distribution leads to intractability of computation.

All the past work on extraction of visual attention uses the saliency map representation to denote the visual attention in an image [5]–[8]. The saliency map is built by either linear combination of features or by training [13]. There are two weaknesses of these approaches. First, most of the methods perform bottom-up computation which does not take into account the past experiences of the system [6]. Secondly, the temporal variation of attention is not modeled.

On the other hand, based on the sequential importance sampling (*SIS*) algorithm [12], [14], [18], we use *attention samples* to represent the probability of attention $P(a|E)$. For example, in the one-dimensional case, the probability of attention $P(a|E)$ is maintained by N attention samples $AS(t) = [as^1(t), \ldots, as^N(t)]$, as well as their weights $\Pi(t) = [\pi^1(t), \ldots, \pi^N(t)]$. It provides a flexible representation of the probability with minimal assumptions. The number of samples employed can be adjusted to achieve a balance between the accuracy of the approximation and the computation load. Moreover, it is easy to incorporate this representation within a dynamical system which can model the temporal continuity of attention if we consider each sample as a particle and each particle having its own dynamics.

In this sampling representation, the location of samples and their associated weights are employed to represent the attention probability $P(a \ldots E)$. This means that for a particular region, the more samples fall into this region and the higher their weights are, the higher is the probability of attention in this region. Apparently, the probability distribution is not fully represented by the distribution of the samples. It also relies on the weights of the samples. However, since we use attention to get the relevant information, we would like the probability of the attention be fully represented by the distribution of the attention samples, not partially on their weights. That is the highly attended regions should have more samples and vice versa. A re-sampling method is therefore introduced to let only the distribution of samples reflect the distribution of attention. In addition, since the attention is inferred from experience (which will be discussed in Section II-C.3) and experience itself encapsulates the goal and environment, our sampling based dynamical attention model systematically integrates the top-down and bottom-up approaches.

$P(A(t)\mid E(t))$:	The *a posteriori* probability of attention given the contextual information up to now.
$P(a(t)\mid E(t))$:	The *a posteriori* probability of attention at time t given the contextual information up to now.
$P(e(t)\mid a(t))$:	The likelihood of the attention at time t with respect to the current contextual information.
$P(a(t)\mid a(t-1))$:	The dynamics of the evolution of attention.

Fig. 3. Probability distributions.

The entire probabilistic notation used in this section is shown in Fig. 3. In the remaining part of this section, we first provide the solution to the static case in Section II-C1. We then extend the solution to the dynamic case in Section II-C2. We treat attention as a Bayesian inference problem and develop an approach to obtain relevant information from the approximated dynamical attention probability. In Sections II-C3–II-C5, a sampling based approach is introduced to maintain the probability of the dynamic attention. Important concepts like *environment sampling*, *sensor sampling*, *attention sampling*, *attention saturation*, as well as *past experiences*, are described in Sections II-C3–, respectively.

1) Static Attention Driven Analysis: In our sampling technique, the second factor in (7), called *goal driven attention*, is represented by samples and their associated weights. Those samples which have higher weights can survive as the samples in the next time slice.

Therefore, samples represent the higher task driven attention data. That is they will contribute more in the final computation of (7). In contrast, other regions having less attention value have less impact on (7). Based on this, we perform the multimedia analysis task (indicated by $P(H|f_L)$) only on these samples and treat other data as the irrelevant data which is to be discarded from the analysis point of view.

The entire algorithm including the dynamics will be discussed in the next section. But, first let us consider the simple static case. Here we assume that we know $P(a|E)$ and we are able to simulate N *i.i.d.* (independently and identically distributed) random samples $\{a_1, a_2, a_3, \ldots, a_N\}$ according to $P(a|E)$. For instance, in a spatial case, they are a set of spatial coordinates. Their associated weights $\{w_1, w_2, w_3, \ldots, w_N\}$ can be obtained by $w_i = P(a|E)$. So the weight w_i is directly proportional to attention probability $P(a|E)$ such that the sum of the N weights is equal to the total attention at that time.

In this case, the distribution of the samples actually reflects the distribution of the attention probability. Differing from the classical perfect Monte Carlo sampling [12] which uses samples to approximate the distribution and consequently get its expectation, we use the sampling method to maintain our attention probability and consequently collect relevant information while discarding irrelevant information. By choosing a proper number of samples, the samples will only exist in the higher attended regions as shown in Fig. 4(b) since the high attention data is given by the distribution $P(a|E)$. These samples intuitively represent the relevant information to be processed. Note that selection of the number of random samples N depends on current overall attention (measured by the *attention saturation* which will be introduced later), as well as the tradeoff between the computation load and the representation accuracy.

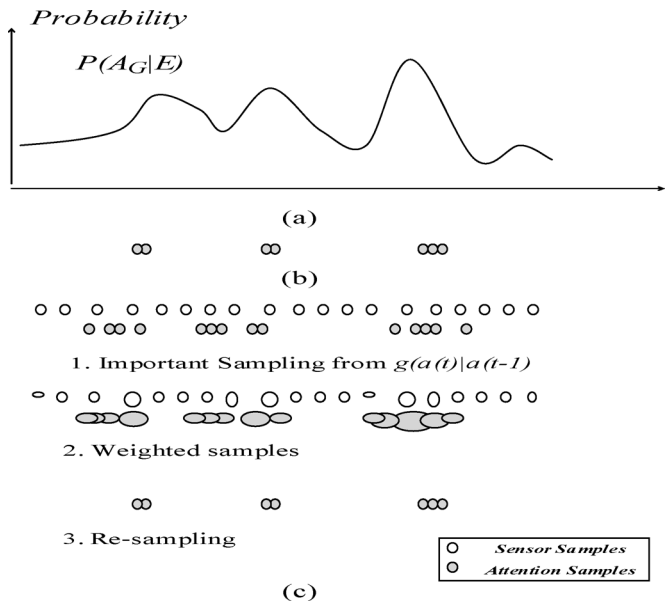

Fig. 4. Sampling based dynamical attention model. (a) Attention. (b) Samples as relevant information (static case). (c) Samples as relevant information (dynamic case).

It is clear that our method allows performing of the actual multimedia analysis task on the N samples. For example, if the task is face detection, it can be performed on the regions of the thresholded N samples. But when dealing with spatio–temporal multimedia data, the focus of attention dynamically varies not only along the spatial axes but also along the temporal axis. A dynamic attention model needs to be investigated in order to achieve effective and efficient spatio–temporal data analysis.

2) Dynamical Attention Driven Analysis: In this section, we aim to infer the current attention $a(t)$ from the contextual information $E(t)$ until the time t, i.e. calculate the *a posteriori* probability $P(a(t)|E(t))$.

a) Dynamical evolution of attention: The attention is inferred from the observed experiences coming from the environment. That is, we try to estimate the probability density of the attention (which is the *state variable* of the system) at time t using $P(a(t)|E(t))$. Note that $E(t)$ consists of all the observed experiences until time t which means $E(t) = \{e(1), \ldots, e(t)\}$, and $a(t)$ is the "attention" in the scene. Attention has temporal continuity which can practically be modeled by a first-order Markov process state-space model [29]. The value of $a(t)$ may not be observed though the experience $e(t)$, which influences the attention $a(t)$, is observable. In this model, the new state depends only on the immediately preceding state, independent of the earlier history. This still allows quite general dynamics, including stochastic difference equations of arbitrary order. Therefore

$$P\left(a(t)|a(t-1), \ldots, a(0)\right) = P\left(a(t)|a(t-1)\right). \quad (8)$$

Our target, the posterior probability $P(a(t)|E(t))$, can be factorized by using the Bayes' rule. The formalization is shown in (9) as follows:

$$\begin{aligned} P(a(t)|E(t)) &= P(a(t)|e(t), E(t-1)) \\ &= \frac{P(e(t)|a(t), E(t-1))\, P(a(t)|E(t-1))}{P(E(t)|E(t-1))} \\ &= \frac{P(e(t)|a(t))\, P(a(t)|E(t-1))}{P(E(t)|E(t-1))} \end{aligned} \tag{9}$$

where

$$\begin{aligned} (P(e(t)|a(t), E(t-1)) &= P(e(t)|a(t))) \\ &= kP(e(t)|a(t))\, P(a(t)|E(t-1)) \end{aligned}$$

where $(k = (1/P(E(t)|E(t-1))))$

Since we are interested in the attention $a(t)$, k becomes a normalization factor which does not depend on the attention.

The prior probability $P(a(t)|E(t-1))$ in (9) can be further formulated as follows (a detailed explanation can be found in [18])

$$\begin{aligned} &P(a(t)|E(t-1)) \\ &\quad = \int_{a(t-1)} P(a(t)|a(t-1), E(t-1)) \\ &\qquad \times P(a(t-1)|E(t-1))\, da(t-1) \\ &\quad = \int_{a(t-1)} P(a(t)|a(t-1)) \\ &\qquad \times P(a(t-1)|E(t-1))\, da(t-1). \end{aligned} \tag{10}$$

According to (10), $P(a(t)|E(t-1))$ is dependent on the probability $P(a(t)|a(t-1))$ and $P(a(t-1)|E(t-1))$.

From the above two equations, we know that the posterior density $P(a(t)|E(t))$ can be iteratively obtained by knowing the observation (likelihood) $P(e(t)|a(t))$, the temporal continuity (dynamics) $P(a(t)|a(t-1))$, and the previous state density $P(a(t-1)|E(t-1))$.

Initially, we assume that the $P(a(1)|E(1))$ is zero. During each iteration, the three probabilities for obtaining the posterior density $P(a(t)|E(t))$ are calculated as follows.

$P(a(t)|a(t-1))$: Since we have assumed a Markov state-space model, the dynamics of attention evolution is described by a stochastic differential equation where the deterministic part models the system knowledge and the stochastic part models the uncertainties. Thus, the dynamics $P(a(t)|a(t-1))$ can be obtained by

$$\begin{aligned} P(a(t)|a(t-1)) &= (2\pi)^{-k/2}|Q| \exp \\ &\times \left(-\frac{1}{2}[a(t) - \Phi a(t-1)]' Q^{-1} [a(t) - \Phi a(t-1)]\right) \end{aligned} \tag{11}$$

where Q is the covariance matrix of the random noise and the term Φ is basically the deterministic part which is the state transition matrix. This formulation is same as that of the Kalman filter. The problem of parameter estimation has been explored in [25].

$P(e(t)|a(t))$: As mentioned before, since we use the current contextual information $e(t)$ to infer the goal oriented attention $a(t)$, we select the contextual information regarding the attention. In another words, the contextual information can be considered as the measurements of the attention coming from the experiential environment. If we assume that the context measurement is independent on each other, we then can define the likelihood of attention in each position to follow the Gaussian distribution

$$\begin{aligned} &P(e(t)|a(t)) \\ &\quad = \prod_{x=1,y=y}^{x=X,y=Y} P(e(x,y,t)|a(t)) \\ &P(e(x,y,t) = 1|a(t) = [x_a, y_b]') \\ &\quad = L \exp\left\{-\frac{(x-x_a)(y-y_b)}{\delta^2}\right\} \end{aligned} \tag{12}$$

where δ^2 is the constant which is used to control the randomness level and L is the normalizing constant.

When the situation that the measurement $e(x,y,t)$ is not binary, the above equation can be modified as follows:

$$\begin{aligned} P(e(x,y,t) = 1|a(t) = [x_a, y_b]') = L \cdot e(x,y,t) \\ \cdot \exp\left\{-\frac{(x-x_a)(y-y_b)}{\delta^2}\right\}. \end{aligned}$$

$P(a(t-1)|E(t-1))$: This is the posterior probability of attention during time $t-1$.

b) Sequential simulation-based solution: Instead of using Kalman filters, the sequential simulation method (SIS) [12], [14], [18] can be invoked to generate a numerical solution for dynamically approximating the density $\pi(a(t)) = P(a(t)|E(t))$. The approach has an advantage in terms of the capacity for generalization.

Let $S(t-1) = \{s_1(t-1), s_2(t-1), \ldots, s_N(t-1)\}$ denote N random draws that are properly weighted by the set of weights $W(t-1) = \{w_1(t-1), w_2(t-1), \ldots, w_N(t-1)\}$ with respect to $\pi(a(t-1))$.

At time t, firstly, a set of samples $S(t)$ is drawn from a so-called *importance function* $g(a(t)|a(t-1))$ [12], [14], [18] (as shown in Fig. 4(c).1). The importance function is defined depending on the application. Secondly, their associated weights are obtained by

$$w_i(t) = w_i(t-1)\frac{P(e(t)|a(t))\, P(a(t)|a(t-1))}{g(a(t)|a(t-1))} \tag{13}$$

where $i = 1, \ldots, N$ and the definitions of $P(e(t)|a(t)$ and $P(a(t)|a(t-1))$ have been provided in the previous section. The discussion of $g(a(t)|a(t-1))$ will be introduced later. This weighting is shown in Fig. 4(c) step 2. Note that in the initial step, $w(t) = P(e(t)|a(t))$.

It has been shown that [18] the above obtained set of random draws and their weights $\{S(t), W(t)\}$ *is properly weighted* with respect to $\pi(a(t))$. It means that the following equation is true:

$$\lim_{n\to\infty} \frac{\sum_{j}^{n} h(s_j)w_j}{\sum_{j}^{n} w_j} = E_\pi\left(h(a)\right) \tag{14}$$

where h is any integrable function, E_π is the expectation, and the notation of time t has been dropped for the sake of simplicity of the expression.

The fundamental idea of the *SIS* algorithm is to use both a set of discrete samples obtained by the importance function $g(a(t)|a(t-1))$ and the weights obtained by (13) to approximate the *a posteriori* density. In other words, the distribution information is embedded both in the samples $S(t)$ and the weights $W(t)$. It is suitable for the applications which only require to get the expectation $E(h(a(t))$ like in tracking problems. However, in our application, our final aim is to obtain the relevant data on which the analysis task can be performed. We need the samples $S(t)$ (*i.e.* location of the samples) themselves to fully cover the entire information about the distribution of the attention $\pi(a(t))$. To this end, after the *SIS* algorithm, a re-sampling step is required to relocate the samples in the higher attended regions as shown in Fig. 4(c) step 3. The re-sampling step works as follows by using arithmetic coding [24]: the weights of all samples are normalized such that their sum is equal to 1 and thus they can be treated as contiguous intervals of [0,1). A random value is obtained by uniformly sampling from the range [0,1). A new sample will be re-created if the random value lies in the interval of the sample's weight. Re-do this over time until N_A (will be discussed in Section VI) new samples are obtained.

Next, we will discuss how to update the samples from the current experiential environment according to (13).

3) Environment Sampling: Since we obtain the attention value from the experiential environments, samples used in our approach have two tasks: sense the environment and maintain the attention. Therefore, we define samples $S(t)$ to include both sensor samples $SS(t)$ and attention samples $AS(t)$

$$S(t) = \{SS(t), AS(t)\}. \tag{15}$$

The samples $S(t)$ comprises of *sensor samples* $SS(t)$ and the *attention samples* $AS(t)$. The sensor samples are basically uniform random samples at any time t which constantly sense the environment. The attention samples are the dynamically changing samples which essentially represent the data of interest at time t.

Since both the types of samples have different uses, we define different importance functions $(g(a(t)|a(t-1)))$ for them. The sensor samples are used to constantly sense the environment. Therefore, we define a uniform importance function $g_S(a(t)) = uniform\ sampling$ for sensor samples. It allows the sensor samples to quickly notice any changes in the environments. Thus, sensor samples constantly scan the environment, looking out for sudden changes in the attention. For example, in the video face-detection scenario, the sensor samples can alert the fact that a new face has entered the scene which cannot be inferred merely by the dynamical evolution of the attention samples of the previous time instant. So sensor samples perform the task of current context estimation from the extracted clues $c_t^n, n = 1, \ldots, N$. The attention samples are the dynamically changing samples which essentially represent the data of interest at time t. The attention samples are therefore derived dynamically and adaptively at each time instance from the sensor samples in our framework through sensor fusion of the current environmental context and the assimilation of the past experience. Once we have the attention samples, the multimedia analysis task at hand can work only with these samples instead of the entire multimedia data. These focused attended samples are the most relevant data for that purpose. It should be understood that our data assimilation process is sampling based. Not all data need to be processed. Our aim now is to obtain these sensor samples to infer the attention. They can be sensed by multiple cues from the environment which can subsequently be fused to create $e(t)$.

The cues for obtaining experiences in the *visual environments* can be classified as temporal cues and spatial cues. They can be visual features extracted from the visual data or information from its accompanying data (speech, sound, text etc.). Basically, sensors can sense these cues in order to infer the state of the environment. Based on the above, the experiential sampling technique is summarized as follows:

The current environment is first sensed by uniform random sensor samples and based on experiences so far, compute the attention samples to discard the irrelevant data. Higher attended samples will be given more weight and temporally, attention is controlled by the total number of attention samples.

4) Sensor Sampling: Studies on human visual system show that the role of experience used in top-down visual perception increases in importance and can become indispensable when the viewing conditions deteriorate or when a fast response is desired. In addition, humans get information about the objects of interest from different sources of different modalities [4]. Therefore, when we analyze one particular data type (say spatio–temporal visual data) in multimedia, we cannot constrain our analysis to this data type only. Sensing other accompanying data like audio, speech, music, and text can help us find out where is the important data. Therefore, it is imperative to develop a sampling framework which can sense and fuse all environmental context data for the purpose of multimedia analysis.

In our framework, $SS(t)$ is a set of $N_S(t)$ sensor samples at time t which estimates the state of the multimedia environment. As mentioned above, these sensor samples are randomly and uniformly generated in order to sense the changes in the environments. Therefore, we define a uniform importance function $g_S(a(t)) = uniform\ sampling$ for them. It makes sensor samples to quickly spot any changes in the environments.

Since we do not change the number of the sensor samples with time, we will drop the time parameter and N_S denotes the number of sensor samples at any point in time. $SS(t)$ is then defined as

$$SS(t) = \left\{ss(t); \Pi^S(t)\right\} \tag{16}$$

where $ss(t)$ depends on the type of multimedia data. For spatial data, $ss(t) = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{N_S}, y_{N_S})\}$ at time t, this is the set of spatial coordinates of the sensor samples. These coordinates are generated randomly and uniformly at every time instance. $\Pi^S(t)$ is the associated weight or the importance of each sample which is represented as $\Pi^S(t) = \{\pi_1^S(t), \pi_2^S(t), \ldots, \pi_{N_S}^S(t)\}$. Now each $\pi_i^S(t)$ is obtained by performing sensor fusion of the q cues $C(t)$ available from the multimedia data (like color, motion, texture etc.). Thus, the set of cues is given by $C(t) = \{c\_sp_1(t), c\_sp_2(t), \ldots, c\_spq(t)\}$ where each individual cue $c\_sp_i(t)$ is given by $c\_sp_i(t) = \{(x_i^1, y_i^1, w\_sp_i^1), \ldots, (x_i^{N_S}, y_i^{N_S}, w\_sp_i^{N_S})\}$ Note that the coordinates x and y refer to the spatial coordinates of the sensor samples and w_sp_i refers to the weight of that particular cue at that sample coordinate. Now it can be easily seen that

$$\pi_i^S(t) = \sum_{j=1}^{q} \alpha_j \cdot w\_sp_j^i \tag{17}$$

where α_j is the importance of the j^{th} cue. So we basically employ the linear combination as the sensor fusion strategy. But this can be replaced by a more sophisticated sensor fusion strategy, which has been investigated in our previous research in [10], [11], if the application so requires. Also, note that if the cue is not spatial, then instead of the spatial coordinates, an appropriate reference (e.g. time) can be used for that cue. Usually, spatial cues are obtained from visual features. This can be denoted as

$$w\_sp_j = VF_j\left(I_t(x,y), \ldots, I_1(x,y), m_j\right) \tag{18}$$

where VF_j is the feature extraction function of the jth cue and m_j is its function parameters. $I_t(x,y)$ denotes the image intensity at time t.

For instance, in a video, the motion cue is a spatial cue since it varies according to its spatial position. It can be simply defined as

$$w\_mot(x,y) = \left|I_t(x,y) - I_{t-1}(x,y)\right|. \tag{19}$$

Here, the feature-extraction function is the absolute difference of corresponding pixel intensity values of two neighboring frames. However, there is no adjustable parameter in this function.

5) Attention Sampling: We know that the attention changes dynamically. In a manner different from that of the sensor samples, which use uniform random sampling as the importance function, we use another probability distribution as an importance function $g_A(a(t)|a(t-1))$ to create the attention samples

$$g_A\left(a(t)|a(t-1)\right) = P\left(a(t)|a(t-1)\right) \tag{20}$$

where $P(a(t)|a(t-1))$ is the dynamics of attention which can be obtained by (11). Consequently, the equation to compute the weights [in (13)] becomes

$$w_i(t) = w_i(t-1)P\left(e(t)|a(t)\right). \tag{21}$$

The notation for attention sampling is introduced as follows.

We represent the dynamically varying $N_A(t)$ number of attention samples $AS(t)$ using

$$AS(t) = \left\{as(t); \Pi^A(t)\right\} \tag{22}$$

where $as(t)$ again depends on the type of multimedia data. For spatial data, $as(t) = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{N_A(t)}, y_{N_A(t)})\}$ is the set of spatial coordinates of the attention samples. $\Pi^A(t)$ is the associated weight or the importance of each sample, which is represented as $\Pi^A(t) = \{\pi_1^A(t), \pi_2^A(t), \ldots, \pi_{N_A(t)}^A(t)\}$. Again, each of the $\pi_i^A(t)$ value is obtained by performing sensor fusion of the q cues $C(t)$ available from the multimedia data.

However, there still have one question: how to determine the number of attention samples $N_A(t)$ which varies with time? $N_A(t)$ intuitively models the *attention saturation* which is defined in the next section.

6) Attention Saturation: The temporal attribute of the spatio–temporal data requires the multimedia system to possess the ability of varying the amount of attention at different times. We introduce the concept of *attention saturation* to measure the attention in a given time slice. The attention saturation in this case can be calculated as the sum of attention in the spatial extent. Its value ranges from zero (lowest, no attention) to one (highest, full attention). We define the attention saturation as $ASat(\mathrm{t})$

$$ASat(t) = f_N\left(\int_{Spatial} P\left(a(t)|E(t)\right)\right) \tag{23}$$

where f_N is the mapping function which is used to normalize the value into range [0,1]. f_N is defined as follows:

$$f_N(x) = \frac{1 - \exp(-\lambda \cdot x)}{1 + \exp(-\lambda \cdot x)} \tag{24}$$

where λ is a scaling factor. The benefit of employing (24) is that it can map a very large input domain to the interval [0, 1]. We select λ so that the output scatters in the interval [0, 1] as much as possible.

The current attention is essentially captured by the sensor samples. The sensor samples are updated by each of the cues. Of course, some cues may only have temporal attributes and no spatial coordinate (e.g. audio volume). Such cues can be defined as $c\_tp_j(t) = \{w\_tp_j\}$, where w_tp_j is the weight of the jth cue. Therefore, the discrete form of (23) is given as follows:

$$ASat(t) = f_N\left(\frac{1}{n}\sum_{t'=[t-n,t]}\left(\frac{1}{N_S}\sum_{i=1}^{N_S}\pi_i^S(t') + \sum_{j=1}^{p}\beta_j w\_tp_j(t')\right)\right) \tag{25}$$

where β_j is the importance of the j^{th} temporal cue and p is the number of the temporal cues. Thus, the attention saturation of the current state is captured by the average weight of all the sensor samples and temporal cues. The value n is the

temporal neighborhood. The aim of averaging n number of recent temporal attention epochs is to suppress noise and to maintain temporal continuity. In our audio-visual face detection, we set $\beta_j = 0.8$ for the sound volume cue and $n = 3$ for the web-camera video stream.

Note that for sensor samples, the number of samples was fixed a priori at N_S in (25) and these samples are generated uniformly and randomly at every time instant. But the number of attention samples should vary with time. However, all previous image-based attention models [5]–[8] lack the ability to model this adaptive behavior.

We are now ready to determine the number of attention samples at time t using

$$N_A(t) = N_{\text{Max}} ASat(t) \tag{26}$$

where N_{Max} is the maximum number of samples the system can handle.

7) Past Experiences: We have introduced how the attention guides the analysis task. Contrastingly, in this section, we will discuss how the local analysis task guides the attention in the form of the past experiences. This is also an important concept in Neisser's Perceptual Cycle, *i.e.*, how the perceiver use the results of analysis to modify the current schema (current environment model).

Our attention model is employed to obtain attention from the experiential environment. The current environment model in our case is the attention model. As formulated in Section II-C.4, the attention model is parameterized by each cue's feature extraction function VF_j, its function parameter m_j and its importance α_j [see (17) and (18)]. The data to be dealt with is dynamic with temporal variations. Therefore, the attention model itself should change dynamically. It is nontrivial to accurately model the dynamical evolution of the attention model itself due to these variations. Thus, we want to simultaneously model the dynamically varying attention as well as the evolving attention model (from which the attention is derived). We add the time variable t to our formulation and define the parameters of the attention model for q feature cues at time t as $APara(t) = \{\alpha_1, \ldots, \alpha_q . m_1, \ldots, m_q\}$.

The local analysis task, though time-consuming, provides us the most reliable measurements about the multimedia data. Like human beings, the results of the analysis can be stored as the accumulated knowledge. This knowledge can be utilized as the past experience when a future data assimilation process starts. In our framework, we want those past experiences to help in *adjusting* (*adapting*) the attention model and let the analysis task guide that attention model evolution. Fig. 5 describes this process graphically.

Suppose we are doing multimedia analysis by mapping low level features to a *semantic symbolic identity*, named Tar (target) in the spatio–temporal data. The attention represented by the attention samples should be focused on regions which have concentrated relevant information about the identity Tar. Due to the reliability of the local analysis task, we can actually employ the local analysis task to judge the accuracy of the current attention samples. At time t, after performing the local analysis task on the attention samples $AS(t)$, we divide the attention samples $AS(t)$ into two sets: $AS_+(t)$ containing *reliable attention samples*, and $AS\_(t)$ containing *unreliable attention samples* by using the following equations:

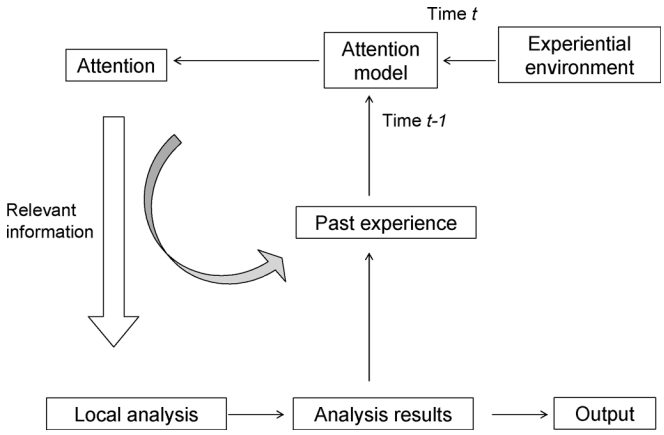

Fig. 5. Analysis guides attention model evolution by past experiences.

$$\begin{aligned} AS_+(t) &= \{AS(t) : S_M(f_L, AS(t)) = Tar\} \\ AS\_(t) &= \{AS(t) : S_M(f_L, AS(t)) \neq Tar\} \end{aligned} \tag{27}$$

where S_M is the feature to semantics mapping function defined in Section II-B.

By employing (27), we treat the attention samples which are finally proven to have the relevant information about the target Tar as the reliable attention samples and the others are not reliable. We call these classified attention samples $AS(t) = \{AS_+(t), AS\_(t)\}$ the past experience. Intuitively, the *past experience* can be used as labeled training samples to learn or update the attention model parameters $AP(t+1)$ for next time slice. This procedure is defined as follows:

$$APara(t+1) = L(AS_+(t), AS\_(t)) \tag{28}$$

where L denotes the inductive learning method to be used to obtain the parameters of the attention model.

III. Conclusions

In this paper, we describe a novel sampling based framework for multimedia analysis called experiential sampling. Based on this framework, we can utilize the context of the experiential environment for efficient and adaptive computations. Inferring from this environment, the multimedia system can select its data of interest while immediately discarding the irrelevant data. In the future, other applications like adaptive streaming and surveillance with more sources of different modalities will be further investigated.

References

[1] P. Viola and M. J. Jones, Robust Real-Time Object Detection Compaq Cambridge Res. Lab., Cambridge, MA, Tech. Rep. CRL 2001/01, 2001.

[2] S. Z. Li, L. Zhu, Z.-Q. Zhang, A. Blake, H.-J. Zhang, and H. Shum, "Statistical learning of multi-view face detection," in *Proc. 7th European Conf. Computer Vision*, Copenhagen, Denmark, 2002.

[3] B. J. Scholl, "Objects and attention: the state of the art," *Cognition*, vol. 80, pp. 1–46, 2001.
[4] R. Jain, "Experiential computing," *Commun. ACM*, vol. 46, no. 7, pp. 48–55, 2003.
[5] D. Chung, R. Hirata, T. N. Mundhenk, J. Ng, R. J. Peters, E. Pichon, A. Tsui, T. Ventrice, D. Walther, P. Williams, and L. Itti, "A new robotics platform for neuromorphic vision: beobots," in *Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02)*, Tuebingen, Germany, 2002.
[6] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.
[7] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition—a gentle way," in *Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02)*, Tuebingen, Germany, 2002, pp. 472–479.
[8] V. Navalpakkam and L. Itti, "A goal oriented attention guidance model," in *Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02)*, Tuebingen, Germany, 2002, pp. 453–461.
[9] H. Lieberman and T. Selker, "Out of context: computer systems that adapt to, and learn from, context," *IBM Syst. J.*, vol. 39, no. 3&4, pp. 617–632, 2000.
[10] J. Wang, R. Achanta, and M. S. Kankanhalli, "A hierarchical framework for face tracking using state vector fusion for compressed video," in *Proc. 28th Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP 2003)*, 2003.
[11] J. Wang, "Detecting and Tracking Human Faces in Compressed Domain for Content Based Video Indexing," M.S. thesis, School of Computing, National Univ. Singapore, , Singapore, 2002.
[12] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for bayesian filtering," *Statist. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.
[13] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. Electron. Imag.*, vol. 10, no. 1, pp. 161–169, 2001.
[14] J. Carpenter, P. Clifford, and P. Fearnhead, Building Robust Simulation-Based Filters for Evolving Data Sets Dept. Statistics, Univ. Oxford, Oxford, U.K., 1999 [Online]. Available: http://www.stats.ox.ac.uk/pub/clifford/Particle_Filters/jj.Abstract.html, Tech. Rep.
[15] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, 2003.
[16] A. Jepson, W. Richards, and D. Knill, "Modal structures and reliable inference," in *Perception as Bayesian Inference*, D. Knill and W. Richards, Eds. New York: Cambridge Univ. Press, 1996, pp. 63–92.
[17] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 696–710, 1997.
[18] J. Liu and R. Chen, "Sequential Monte Carlo for dynamic systems," *J. Amer. Statist. Assoc.*, vol. 93, pp. 1031–1041, 1998.
[19] U. Neisser, *Cognition and Reality*. San Francisco, CA: W. H. Freeman, 1976.
[20] C. Neti, B. Maison, A. Senior, G. Iyengar, P. Decuetos, S. Basu, and A. Verma, "Joint processing of audio and visual information for multimedia indexing and human-computer interaction," in *Proc. RIAO (Computer Assisted Information Retrieval)*, France, 2002.
[21] G. Iyengar, H. Nock, and C. Neti, "Audio-visual synchrony for detection of monologues in video archives," in *Proc. ICASSP*, 2003.
[22] D. Li, N. Dimitrova, M. Li, and I. Sethi, "Multimedia content processing through cross-modal association," in *Proc. ACM Int. Conf. Multimedia (ACM MM 2003)*, Berkeley, 2003.
[23] R. Jain, "Semantics in multimedia systems," in *Keynote talk at Int. Conf. Multi-Media Modelling*, Taipei, Taiwan, Ja. 8–10, 2003.
[24] D. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[25] Z. Ghahramani and G. Hinton, Parameter Estimation for Linear Dynamical Systems Dept. Comp. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. CRG-TR-96-2, 1996 [Online]. Available: http://www.cs.toronto.edu/~hinton/absps/tr96-2.html
[26] M. S. Kankanhalli, J. Wang, and R. Jain, "Experiential sampling on multiple data streams," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 947–955, Oct. 2006.

Mohan S. Kankanhalli received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kharagpur, and the M.S. and Ph.D. degrees in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY.

He is a Professor with the Department of Computer Science, School of Computing, National University of Singapore. He has worked at the Institute of Systems Science in Singapore and at the Department of Electrical Engineering, Indian Institute of Science, Bangalore. His current research interests are in multimedia systems (content processing, multimedia retrieval) and information security (media watermarking and authentication).

Dr. Kankanhalli is on the editorial board of several journals, including the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Multimedia Systems* journal, and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.

Jun Wang received the B.E. degree in electrical engineering from Southeast University, Nanjing, China, and the M.Sc. degree in computer science from the National University of Singapore. He is currently pursuing the Ph.D. degree with the Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science (EWI), Delft University of Technology, Delft, The Netherlands.

His current research topic is personalized multimedia systems and recommender systems.

Ramesh Jain (F'92) is the Bren Professor of Information and Computer Sciences in the Department of Computer Science, University of California, Irvine. He has been an active Researcher in multimedia information systems, image databases, machine vision, and intelligent systems. While he was at the University of Michigan, Ann Arbor, and the University of California, San Diego, he founded and directed artificial intelligence and visual computing labs. He has co-authored more than 250 research papers in well-respected journals and conference proceedings. Among his co-authored and co-edited books is *Machine Vision*, a textbook used at several universities. He enjoys working with companies, is involved in research, and enjoys writing. His current research is in experiential systems and their applications.

Dr. Jain was also the founding Editor-in-Chief of *IEEE Multimedia* magazine and the *Machine Vision and Applications* journal. He serves on the editorial boards of several magazines in multimedia, business, and image and vision processing. He is a Fellow of ACM, IAPR, AAAI, and SPIE.