Experience based Sampling Technique for Multimedia Analysis

Jun Wang
Department of Computer Science, School of Computing, National University of Singapore
Singapore 117543
Tel: (65) 6874-4925
wangj@comp.nus.edu.sg

Mohan S Kankanhalli
Department of Computer Science, School of Computing, National University of Singapore
Singapore 117543
Tel: (65) 6874-6738
mohan@comp.nus.edu.sg

ABSTRACT

Voluminous spatio-temporal data make multimedia analysis tasks extremely inefficient and lack of adaptability. We present a novel *experience based sampling* technique which has the ability to focus on the analysis's task by making use of the contextual information and past experiences. Based on this, a sampling based dynamic attention model is built by sensing the experiential environment. Sensor samples are used to gather information about the current environment and attention samples are used to represent the current state of attention. In our framework, the task-attended samples are inferred from experiences and maintained by a sampling based dynamical system. The multimedia analysis task can then focus on the attention samples only. Moreover, past experiences and the current environment can be used to adaptively correct and tune the attention. As a prototypical multimedia analysis task, we tackle the face detection problem in videos. Face detection is only performed on the attended samples to achieve robust real time processing. Experimental results have been presented to demonstrate the efficacy of our technique. This experience based sampling based analysis method appears to be a promising technique for general multimedia analysis problems. The generality stems from the power of the sampling method which makes no assumptions about the form of distribution of attention which is usually multimodal in nature.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis-*Color, Motion, Sensor fusion, Time-varying imagery*. I.6.5. [**Simulation and Modeling**]: Model Development - *Modeling methodologies.*

General Terms

Algorithms, Performance, Design, Experimentation.

ACM Multimedia '03, Dec 1-2, 2003, Berkeley, California.

Keywords

Sampling, Dynamical Systems, Experiential Computing, Visual Attention.

1. INTRODUCTION

Multimedia processing often deals with spatio-temporal data which have the following attributes:

- They possess a tremendous amount of redundancy
- The data is dynamic with temporal variations with the resultant history
- Some data can be live with real-time processing requirements
- 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.

However, many current multimedia analysis approaches do not fully consider the above attributes which leads to two main drawbacks – *inefficiency* 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. If the ambient experiential context is ignored, the approaches cannot adapt to the changing environment. Thus, the processing cannot adapt itself to 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 [6]. Especially the human visual system is quite successful in understanding the surrounding environment at appropriate accuracy quite efficiently. This is due to many factors [16]: the excellence of the physical visual 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* [8] play an important role for the human visual perception to understand the visual scene accurately and quickly. The vision for experiential computing was introduced in [7], which envisages that multimedia analysis should also have the ability to process and assimilate sensor data like humans. Therefore, we would like to articulate the following goal for multimedia analysis:

"*In an experiential environment, analysis is based on sensing the data from the environment. Based on the observations and experiences, collate the relevant data and information of interest related to the task of the analysis. Thus, the analysis process interacts naturally with the data based on its interests in light of the past experiences of that analysis.*"

It is apparent that many current multimedia analysis approaches ignore their environments and do not have the ability to interact with the environments. They, instead of processing relevant data, perform pervasive non-focused computations. In this paper, we argue that like human perception, multimedia analysis should be placed in the context of its experiential environment. It should have the following characteristics: 1. The ability to "focus" (have attention), i.e., to selectively process the data that it observes or gathers based on the context. 2. Experiential exploration of the data. Past analysis should help improve the future data assimilation. In return, these two attributes would help the analysis to deal with the redundancy and diversity of the spatial-temporal data which is particularly important for real time applications.

In order to achieve this, we introduce a novel technique called *experience based sampling*, i.e., sampling multimedia data according to experiences. As shown in Figure 1, by sensing the contextual information in the experiential environment, a sampling based dynamic visual attention model is built to maintain the focus towards the interest of the current analysis task. Only the relevant samples survive for performing of the final task. These samples precisely capture the most important data. What is interesting is the past samples influence future sampling via feedback. This mechanism ensures that the analysis task benefits from past experience.

In this paper, as an illustrative example of an analysis task, the face detection problem in videos is tackled. Experiences (domain knowledge, speech, visual cues such as skin color and motion) are utilized to infer the attended samples. These samples are adaptively maintained by a sampling based dynamical system. Face detection is performed only in the attended samples to achieve the robust real time performance. Most importantly, previous face detection measurements serve as experiences and are used to adaptively adjust the skin color model to fit variety of changing visual environments. The experience based sampling technique can be utilized for many other applications such as object detection, object recognition, object tracking (face recognition or traffic sign recognition), context aware video streaming and surveillance. For example, in a surveillance application, intruders can enter only from the boundary of the scene. Therefore, when there are no intruders in the scene, the attention of visual analysis should focus on the boundary. If there is an intruder, the focus of attention should evolve to follow the person. These experiences can be easily modeled by using our proposed method.

Our contributions in this paper are as follows. We introduce the concept of experience based sampling which provides a new visual analysis framework to solve the problem of the adaptation and efficiency. This method can adaptively provide attended samples to help avoid exhaustive analysis. Thus, it can substantially improve the efficiency as well as help suppress noise. Though we present the experience based sampling technique for spatio-temporal (visual) analysis of multimedia data, it should be understood that the sampling technique can be applied to any data type.

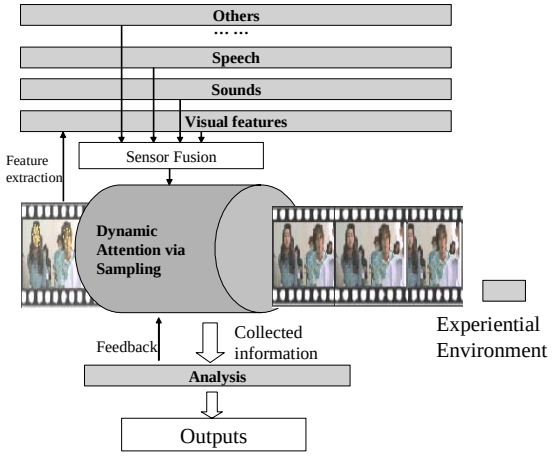


Figure 1. Experience based Sampling for multimedia processing.

2. RELATED WORK

Since human perception is greatly aided by the ability to probe the environment through various sensors along with the use of the situated context, it has inspired context aware computing in the Human Computer Interaction research community [15]. The basic idea there is to help the computer respond more intuitively to the human user based on the context. A comprehensive review of context aware computing can be found in [15, 16].

The ability to "focus" the "consciousness" in human visual perception has inspired research in non-uniform representation of visual data. Visual attention in human brains allows a small part of incoming visual information to reach the short-term memory and visual awareness, consequently providing the ability to investigate more closely. The computational modeling of visual attention has been investigated for potential usages in planning and motor control [14], video summarization [11] and object recognition [12]. The computational model of visual attention maintains a two-dimensional topographic saliency map by employing a bottom-up reasoning methodology [10]. Reference [13] attempts to model the influence of high-level task demands on the focal visual attention in humans. There is also the *foveation* technique [19, 21] for maintaining a high resolution area of interest in an image. A uniform-resolution image can be foveated to transform into a spatially varying resolution image by either a log-polar [19] or a wavelet approach [20]. However, in humans, attention varies with the nature of task. In addition, visual attention is adaptive. This means it will vary depending on the visual environment and has a self-corrective mechanism utilizing experiences. Thus, attention will vary over time. Unfortunately, the above saliency map based visual attention models and foveation approaches are image based that do not provide a mechanism to evolve and adapt attention dynamically. Contrastingly, our sampling framework naturally expresses the dynamics of attention of a system. What is particularly appealing is that the attention states as well as the state-transitions are captured as a closed-loop feedback system.

The Sampling Importance Resampling (SIR) method which can be used for modeling evolution of distributions was proposed by

Rubin [26]. The dynamics aspects were developed by Gordon *et al* [27]. In a SIR filter, a set of particles, which move according to the state model, multiply or die depending on their "fitness" as determined by the likelihood function. A general importance-sampling framework that elegantly unifies many of these methods has been developed in [25]. A special case of this framework has been used for the purpose of visual tracking in [18]. Though we also utilize the sampling method, we use it to maintain the visual attention. To the best of our knowledge, this is the first use of the sampling technique to maintain the dynamically evolving attention. Thus, unlike [18], the number of samples dynamically changes for the purpose of adaptively representing the temporal visual attention. This is in tune with the growing realization that computing systems will increasingly need to move from processing information and communication to the next step: dealing with insight and experience [6,7,8]. One of the key technical challenges in experiential computing is information assimilation, i.e., how to process real time disparate data from multiple sensors. Our research in this paper aims to provide a sampling based dynamical framework to solve this problem in the multimedia analysis domain.

3. EXPERIENCE BASED SAMPLING

In this section, we introduce our experience based 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). As stated earlier, we base this discussion on video which is a prototypical multimedia data type.

3.1 Definitions

Our definition of experience is based on [8].

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.*

There are some relationships among these experiences. The environment can be characterized by features extracted from visual scene and other accompanying 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 states encapsulate the experiences till the current state. The relationships are shown in Figure 2. These relationships can help us to define the experiential environments when we do 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 experiences.

In our framework, we allow the analysis to guide the attention on to regions or data of interest from the entire spatio-temporal data. Differing from other methods [9,10,12,13], we propose a sampling based attention model to obtain attention along both spatial and temporal axes from the experiential environments.

3.2 Sensor Sampling

Studies on human visual system show that the role of experiences 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 [7]. 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.

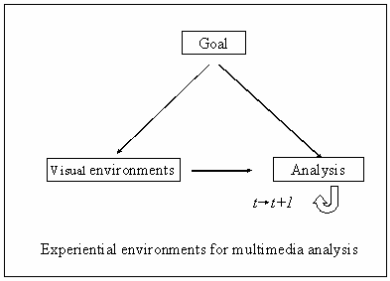


Figure 2. The relationships of experiences.

Experience Based Sampling: *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. Spatially, higher attended samples will be given more weight and temporally, attention is controlled by the total number of attention samples.*

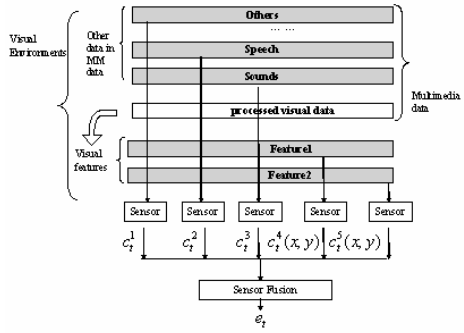


Figure 3. The framework of sensing environments.

In the sampling framework, we represent the environment e_t at time t as

$$e_t = \{S(t), A(t)\} \quad (1)$$

The environment e_t comprises of *sensor samples S(t)* and the *attention samples A(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. The attention samples are actually derived dynamically and adaptively at each time instance from the sensor samples in our framework through sensor fusion 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. Figure 3 shows our

framework for doing sensor fusion within the experiential environments. 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 visual attention. They can be sensed by multiple cues from the visual environment which can subsequently be fused.

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 current processing visual data or information from their accompanying data (speech, sounds, text etc.). Basically, sensors can sense these cues in order to infer the state of the environment.

In our framework, $S(t)$ is a set of $N_S(t)$ sensor samples at time t which estimates the state of the multimedia environment. These sensor samples are randomly and uniformly generated. 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. $S(t)$ is then defined as:

$$S(t) = \{s(t); \Pi^S(t)\} \tag{2}$$

where $s(t)$ depends on the type of multimedia data. For spatial data, $s(t) = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{N_S}, y_{N_S})\}$ 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 spatial cues $C(t)$ available from the multimedia data (like color, motion, texture etc.). Thus, the set of cues is given by $C(t)=\{c\_sp_1(t), c\_sp_2(t), \ldots, c\_sp_q(t)\}$ where each individual spatial 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

$$\pi_i^S(t) = \sum_{j=1}^{q} \alpha_j \cdot w\_sp_j^i \tag{3}$$

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 [22, 23], 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. For instance, the motion cue is a spatial cue since it varies according to its spatial position. It can be simply defined as

$$w\_mot(x, y) = |i_t(x, y) - i_{t-1}(x, y)| \tag{4}$$

Here the weight is the absolute difference of corresponding pixel intensity values of two neighboring frames.

3.3 Attention Sampling

All past work on extraction of visual attention use saliency map to denote the visual attention in an image [9,10,12,13]. The saliency map is built by either linear combination of features or by training [28]. 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 [10]. Secondly, the temporal variation of attention is not modeled.

Contrarily, our sampling based dynamic visual attention model systematically integrates the top-down and bottom-up approaches to infer attention from the environment as well as experience. Thus, the number of attention samples dynamically evolves so the number will be increased when more attention is required and vice-versa. Moreover feedback from the final analysis task is used to tune the attention model with time.

The visual attention in a scene can be represented by a multi-modal 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. Therefore, we adopt a sample-based method to represent the visual attention. For example, in the one dimensional case, the visual attention is maintained by N samples $a(t)=[s^1(t), \ldots, s^N(t)]$ and their weights $\Pi(t) = [\pi^1(t), \ldots, \pi^N(t)]$ as shown in Figure 4. It provides a flexible representation 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 within a dynamical system which can model the temporal continuity of visual attention.

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

$$A(t) = \{a(t); \Pi^A(t)\} \tag{5}$$

where $a(t)$ again depends on the type of multimedia data. For spatial data, $a(t) = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{N_A(t)}, y_{N_A(t)})\}$, 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, two key issues have not been addressed yet. One is how to determine the number of attention samples $N_A(t)$ which varies with time (measuring the attention at a given time instance called temporal attention). Secondly, it is not apparent how to determine the actual attention samples (for the spatial attention). Note that for sensor samples, the number of samples was fixed a priori at N_S and these samples are generated uniformly and randomly at every time instant. But the number of attention samples varies with time. For instance, in the traffic monitoring application shown in Figure 5, Figure 5 (a) has more motion activity and hence needs more attention samples to represent this motion attention. As shown in Figure 5 (b), 567 attention samples (marked as yellow points) are required to represent this motion attention using our method. In contrast, Figure 5 (c) has less motion and needs fewer attention samples. As shown in Figure 5 (d), no attention samples are needed. We determine the number of attention samples $N_A(t)$ based on the current state and the past experiences. The current state is essentially captured by the sensor samples. Therefore, the influence of the current state on the number of attention samples is by using the means of fusion of the sensor sample data:

$$\mathrm{T}(t) = \frac{1}{N_S}\sum_{i=1}^{N_S} \pi_i^S(t) \tag{6}$$

Thus, the overall temporal fusion of the current state is captured by the average weight of all the sensor samples.

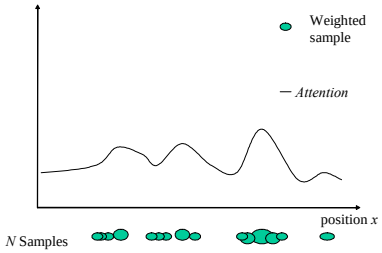


Figure 4. The multi-modal attention can be represented by *N* samples $a(t)=[s^1(t),\ldots,s^N(t)]$ and their weights $\Pi(t)=[\pi^1(t),\ldots,\pi^N(t)]$.

(a) frame 37 (b)frame 37 (c)frame 479 (d)frame 479

Figure 5. Temporal motion attention.(a) more motion activity (b) 567 attention samples are employed to represent this motion attention.(c) need less attention at this time(d) No attention samples are needed at this time. The number of attention samples is calculated by using equation (8).

Intuitively, we can use the squashing function [1] to normalize this relationship.

$$\hat{\mathrm{T}}(t) = \frac{1 - \exp(-\beta \cdot \mathrm{T}(t))}{1 + \exp(-\beta \cdot \mathrm{T}(t))} \tag{7}$$

where β is a scaling factor. As shown in Figure 6, by employing equation (6), the squashing function can map a very large input domain to the interval [0,1]. $\hat{\mathrm{T}}(t)$ can be adopted to measure the attention at a given time instance (the temporal attention) shown in Figure 13.

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

$$N_A(t) = N_{Max}\frac{1}{n}\sum_{i=[t-n,t]}\hat{\mathrm{T}}(i) \tag{8}$$

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

3.3.1 Dynamical Evolution of Attention

Attention is inferred from the observed experiences coming from the experiential environments. 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 is $E_t=\{e_1,\ldots,e_t\}$, a_t is the "attention" in the scene and $a(t)$ is the sampled representation of a_t. Attention has temporal continuity which can be modeled by a first-order Markov process state-space model [29] as shown in Figure 7. 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,

$$P(a_t \mid a_{t-1},\ldots,a_0) = P(a_t \mid a_{t-1}) \tag{9}$$

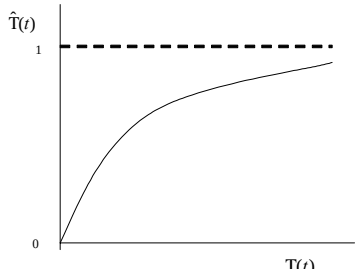


Figure 6. Relationship between the temporal experiences and the probability of the temporal attention.

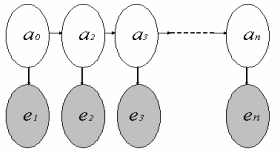


Figure 7. State-space model for attention.

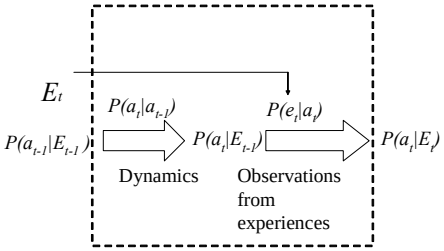


Figure 8. Iteration of calculating visual attention state density $P(a_t|E_t)$ in state-space model. By knowing previous state density $P(a_{t-1}|E_{t-1})$ and current experiences e_t, $P(a_t|E_t)$ can be approximated by a sampling method in the form of samples.

Based on the above state space model, the *a posteriori* density $P(a_t|E_t)$ can iteratively be obtained by knowing the observations (likelihood) $P(e_t|a_t)$, the temporal continuity (dynamics) $P(a_t|a_{t-1})$ and the previous state density $P(a_{t-1}|E_{t-1})$ as shown in Figure 8 and shown here:

$$\begin{aligned} P(a_t \mid E_t) &= P(a_t \mid e_t, E_{t-1}) = \frac{P(e_t \mid a_t, E_{t-1})P(a_t \mid E_{t-1})}{P(E_t \mid E_{t-1})} \\ &= \frac{P(e_t \mid a_t)P(a_t \mid E_{t-1})}{P(E_t \mid E_{t-1})} \qquad (P(e_t \mid a_t, E_{t-1}) = P(e_t \mid a_t)) \\ &= kP(e_t \mid a_t)P(a_t \mid E_{t-1}) \qquad \left(k = \frac{1}{P(E_t \mid E_{t-1})}\right) \end{aligned} \tag{10}$$

Since we are interested in the attention a_t, k becomes a normalization factor which does not depend on the attention. Now,

$$P(a_t \mid E_{t-1}) = \int_{a_{t-1}} P(a_t \mid a_{t-1}, E_{t-1}) P(a_{t-1} \mid E_{t-1})$$
$$= \int_{a_{t-1}} \int_{a_{t-2}} P(a_t \mid a_{t-1}) P(a_{t-1} \mid E_{t-1}) \quad (11)$$
$$= \int_{a_{t-1}} P(a_t \mid a_{t-1}) \int_{a_{t-2}} P(a_{t-1} \mid E_{t-1})$$
$$= \int_{a_{t-1}} P(a_t \mid a_{t-1}) P(a_{t-1} \mid E_{t-1})$$

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 term $P(a_t|a_{t-1})$ can be obtained by:

$$a_t = \Phi a_{t-1} + w \quad (12)$$

The term Φ is basically the deterministic part which is the state transition matrix and w is the stochastic component modeling noise. This formulation is quite similar to the Kalman filter. The problem of parameter estimation was explored in [24].

3.4 Experience Based Sampling Technique

We are now ready to fully describe the technique based on the background developed so far.

Algorithm (experience based sampling technique)

t=0:

Environment sensing

1. Choose the number of sensor samples N_s for sensing the environment e_t (in our experiments, we set N_S = 100 or 200) and set $N_A(0) = 0$ initially.
2. Perform sensor sampling of the environment by creating N_s random samples uniformly.
3. Compute the weight w_sp_i of each multimedia cue sensed from the environment.
4. Update the weight of each sensor sample, defined as $\Pi^S(t) = \{\pi_1^S(t), \pi_2^S(t), \ldots, \pi_{N_S}^S(t)\}$, by performing sensor fusion using equation (3). If $N_A > 0$, the attention samples' weights also need to be updated by utilizing sensor fusion.
5. Compute the influence of the current environment $T(t)$ using equation (6).
6. Obtain temporal environment $\hat{T}(t)$ by using equation (7).
7. Calculate the number of attention samples $N_A(t)$ needed using equation (8).
8. If $N_A = 0$, no attention samples are needed, so set $t = t+1$ and go to step 2. Otherwise go to step 9.

Building the attention model using sampling

9. Create or Resampling $N_A(t)$ number of attention samples $a(t) = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{N_A(t)}, y_{N_A(t)})\}$ and their weights $\Pi^A(t) = \{\pi_1^A(t), \pi_2^A(t), \ldots, \pi_{N_A(t)}^A(t)\}$ using the following resampling algorithm:
   - normalize $\pi^A(t)$ and $\pi^S(t)$ so that $\sum_{n=0}^{N_A(t)+N_S} \pi^S(t) + \pi^A(t) = 1$
   - Interpreting each weight as a probability, use $c_t^i = c_t^{i-1} + \pi^i(t)$ in order to obtain the cumulative probability distribution $c_t = [c_t^1, \ldots, c_t^N]$ where $N = N_A(t) + N_S$
   - for i =0 to i = 1, step-size $1/N_A(t)$ find by binary subdivision, the smallest j for which $c_{t-1}^j >= i$. Use this sample as the new sample.

   This algorithm [29] treats the weights as contiguous intervals of (0,1). These intervals are randomly ordered and it is sampled such that the weight of chosen samples in every interval is the same.
10. Perform the actual multimedia task analysis on the $N_A(t)$ attention samples. For example, if the task is face detection, do it on the regions of the attention samples. If the task is successful (e.g. a face is detected), increase the weights of the relevant attention samples and update the cue extraction model (such as skin color model which will be explained the in the next section). This is the feedback of the analysis.
11. Treat sensor samples N_S and attention samples $N_A(t)$ as a whole and normalize their weights $\pi^S(t)$ and $\pi^A(t)$ to make $\sum \pi^A(t) + \sum \pi^S(t) = 1$
12. Create a new set of attention samples for time $t+1$ by propagating the attention samples $N_S(t)$ by dynamical evolution. As described by equation (12), the temporal continuity $p(a_{t+1} \mid a_t)$ is modeled by a linear stochastic differential equation $a(t+1) = \Phi a(t) + w$ where Φ is the linear state transition matrix and w is a vector of standard normal random variables as stated before.
13. $t = t+1$; go to step 2.

4. APPLICATIONS

As a general analysis framework, our proposed experience based sampling technique can be used for a variety of multimedia analysis tasks, especially real-time applications like traffic monitoring and surveillance. As a test example, we have applied this framework for the face detection problem in videos.

In face detection problem [2,3,4,5], current robust detection methods all rely on exhaustively performing visual measurements on the entire image with different scale factors i.e. every position of the image is probed at different scales by employing either a Gaussian model [2], or Neural Networks [3] or boosted classifiers [4,5]. 193737 probe computations are needed for a single 320x240 sized image (using 20x20 size scan window and a scale factor of 1.2). However, in most of the cases, human faces only occupy a small part of a given frame. Obviously, most of these probes are conducted where faces do not possibly exist. The computations in such low probability areas are wasteful and can even lead to false detects. It would be ideal if the expensive face detection computations are carried out only where faces are very likely to occur. Our experience based sampling framework precisely facilitates this.

We utilize experiences (domain knowledge and accompanying audio (speech) and visual cues (skin color and motion)) to infer the attention samples. These attention samples are adaptively

maintained by the sampling based visual attention framework proposed in the previous section. We use the *adaboost* face detector [4] as the multimedia analysis task. Face detection is only performed on the attention samples to achieve robust real time processing. Most importantly, past face detection results serve as experiences to adaptively correct the attention samples and the skin color model in the attention inference stage. This adapts the face detector to cope with a variety of changing visual environments

4.1 Sampling for face attention

In the face detection application, the attended regions are those where the probability of finding a face is high. The face attention information can be inferred from the contextual information in the experiential environment. In this application, we would like to use the cues of visual features (motion and skin color) and accompanying audio data to sense the experiential environment. The methods of obtaining the cues are now described:

Skin color cue: Since skin color is clustered well in the color space [30], we use the 1-D histogram of hues (color) channel from HSV color system to represent skin color [30]. This histogram can not be fixed. We define it as H_t at time t. It could dynamically change according to the current environment. In our method, the face regions obtained by face detection in time *t-1* serves as the feedback to calculate the skin color histogram H_t for time t. Based on this feedback, the skin color model is dynamically updated.

We define

$$c\_sp_{Skin}(t) = \left\{(x^1_{Skin}, y^1_{Skin}, w\_sp^1_{Skin}), \ldots, (x^{N_S}_{Skin}, y^{N_S}_{Skin}, w\_sp^{N_S}_{Skin})\right\}$$

as the skin color cues. As show in Figure 9, the stored skin color histogram H_t in time t is employed as a lookup table to calculate the weight $w\_sp^i_{Skin}$. This lookup procedure (looking for bin's value) is defined as

$$w\_sp^{\ i}_{Skin}\ (x, y) = lookup\ \ (x, y, H_t) \qquad (13)$$

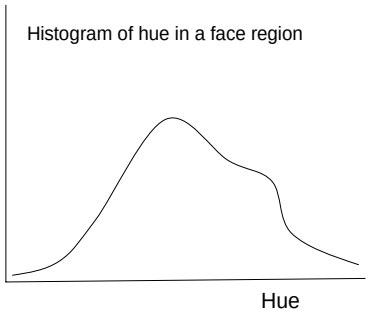


Figure 9. 1 D histogram of hue channel in a face region.

Motion cue: The motion cue is defined as

$$c\_sp_{MT}(t) = \left\{(x^1_{MT}, y^1_{MT}, w\_sp^1_{MT}), \ldots, (x^{N_S}_{MT}, y^{N_S}_{MT}, w\_sp^{N_S}_{MT})\right\}$$

The weight $w\_sp^i_{MT}$ is obtained by equation (4).

Speech cue: Speech implies the possible existence of the face in the visual data. The face detector should work on more samples during the time slice when speech is on. In our method, we dynamically increase the number of attention samples when speech is on.

By using the motion cue and skin color cue, the step 3 in our experience based sampling technique is adapted as follows:

Step 3 and 4:

START:

- For sensor samples $i = 0$ to N_S :

  Calculate $w\_sp^i_{MT}$ and $w\_sp^i_{Skin}$ using equations (4) and (13) respectively.

  IF $w\_sp^i_{MT} > T_1$ and $w\_sp^i_{Skin} > T_2$ THEN

  $$\pi^S_i(t) = \alpha_{MT} \cdot w\_sp^i_{MT} + \alpha_{Skin} \cdot w\_sp^i_{Skin}$$

- IF number of attention samples $N_S > 0$, the weights of the attention samples also need to be updated by using the above method.

RETURN.

T_1 and T_2 are the thresholds for removing noise, α_{Skin} and α_{MT} denote the importance of the cues.

The Step 10 is adapted as follows:

Step 10:

START:

- IF $N_A = 0$ THEN RETURN
- For attention samples $i = 0$ to N_A

  Perform face detection on the attention samples i.

  IF face *is* detected THEN

  $$\pi^A_i(t) = \lambda$$

- Merge all face samples to get the final face detection output.
- Using face region obtained by face detection to update the skin color histogram H_t (which will be used as the updated skin color model for the next time step)

RETURN

λ is a higher weight which will be given to the relevant samples. Since this feedback from the face detection is more reliable than motion and skin cues, λ here is set to be higher than the weight in step 3 and 4. By using this, the relevant samples proved by the face detection can still survive even when no motion or skin color attention.

In the step 10 shown above, there are two feedbacks occurring when face is detected. Relevant attention samples which are confirmed is proved by the face detector are given higher weight λ. Meanwhile the 1D skin color histogram model H_t is recursively updated.

By using the speech cue, the step 7 for the calculation of the number of the attention samples (attention at a given time instance) can be modified as:

Step 7

START:

- Detect whether the speech is on or off in the accompanying audio
- IF speech on THEN $N_A(t) = N_{Max}$

  ELSE Obtain $N_A(t)$ using the equation (8)

RETURN

5. EXPERIMENTS

In this section, we present results from two experiments. The reader can see the all the results with actual videos at our website [31].

5.1 Experience based Sampling Results

The experiments of our proposed experience based sampling technique are given in this section.

5.1.1 Experimental Setup

We use motion as the cue to show the effectiveness of our sampling based technique to maintain the motion attention in both spatial and temporal directions.

Sensors samples in the equation (2) again can be defined as $s(t) = \{(x_1, y_1), (x_2, y_2), \cdots, (x_{N_S}, y_{N_S})\}$. Their associated weight $\Pi^S(t) = \{\pi_1^S(t), \pi_2^S(t), \ldots, \pi_{N_S}^S(t)\}$ can be obtained by calculating the spatial cue of motion. We define again the spatial cue of motion as $c\_sp_{MT}(t) = \{(x_{MT}^1, y_{MT}^1, w\_sp_{MT}^1), \ldots, (x_{MT}^{N_S}, y_{MT}^{N_S}, w\_sp_{MT}^{N_S})\}$. The weight of the motion cue $w\_sp_{MT}^i$ is obtained by using equation (4). For motion attention sampling, the step 3 and 4 of our experience based sampling technique are changed as follows:

Steps 3 and 4:

START:

- For sensor samples $i = 0$ to N_S

  Calculate $w\_sp_{MT}^i$ using equation (3)

  IF $w\_sp_{MT}^i > T_1$ THEN

  $$\pi_i^S(t) = w\_sp_{MT}^i / f$$

- IF number of attention samples N_S >0, attention samples' weights also need to be updated using the above method.

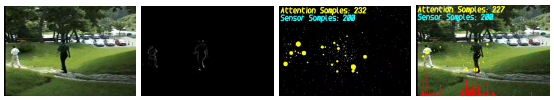
RETURN

(T_1 is the threshold for removing noise and f is a constant. They are set to 4 and 2 respectively in our experiments).

Since in these experiments, we want to show that our sampling method captures the motion attention, step 10 for final analysis is discarded to depict pure attention.

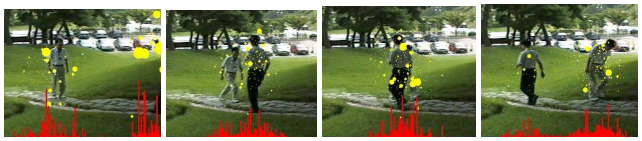
5.1.2 Experimental Results

We test our method for the video of several pedestrian (Figures 10 and 11) and traffic monitoring sequences (Figure 12). There are 200 sensor samples randomly scattered spatially to sense the motion experience. Based on the sensor output, attention samples are created. Their numbers and spatial distribution are all determined by the motion experience. Figure 11 shows that, unlike the saliency map based attention model (indicated by Figure 10(b)), only 227 attention samples and 200 sensor samples are sufficient to maintain the motion attention.



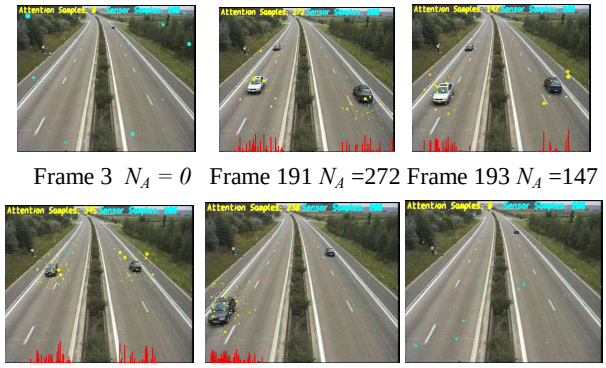
(a) (b) (c) (d)

Figure 10. Mpeg 7 Test sequence 1. (a) original frame.(b) saliency map for motion. (c) 232 attention samples (yellow points) for motion. (d) motion attention by attention samples with original frame. Red bar shows the spatial visual attention in *x* direction; yellow points show the 227 attention samples. Point size indicates the confidence of this sample. Blue points show the 200 sensor samples.



Frame 111 Frame 152 Frame 165 Frame 195

Figure 11. Mpeg 7 Test sequence 2. Red bar shows the spatial visual attention in *x* direction; Yellow points show the attention samples. Point size indicates the confidence of this sample. This figure illustrates the ability of maintaining multi-modal attention. Both visual attention emerge and split during and after the crossing of the subjects.



Frame 3 $N_A = 0$ Frame 191 N_A =272 Frame 193 N_A =147

Frame 203 N_A =345 Frame 212 N_A = 238 Frame 243 N_A =0

Figure 12. Traffic monitoring sequence. This figure illustrates the both spatial and temporal visual attention inferred from motion experience. Blue points are sensor samples while yellow points are attention samples. Red bar shows the spatial attention in *x* direction. It evolves according to the spatial experience. N_S number of sensor samples is set to 200. N_A number of attention samples changes each time according to the temporal experience.

The weight of each attention sample is drawn using red bars along with the x direction to visualize the spatial attention in *x* direction. From Figure 10, 11 &12, we can see that our sampling technique can model multi-modal motion attention quite well without maintaining the saliency map (which requires higher computation).

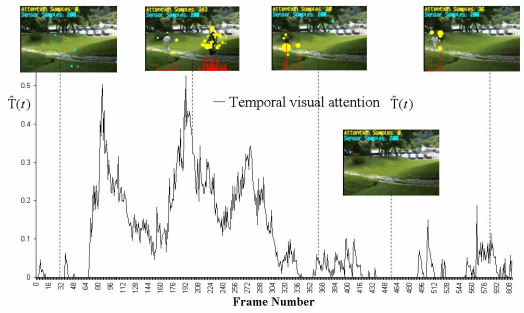


Figure 13. Temporal motion attention in the MPEG 7 clip.

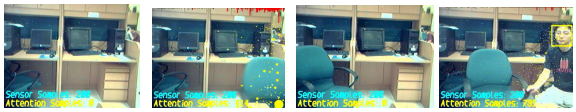
The evolution of temporal attention (represented by the number of attention samples) is shown in Figure 12 and 13. In Figure 12, the N_A roughly reflects the traffic status at each time step. Therefore, our method here can be used for monitoring the traffic also. It also shows that the temporal attention is only aroused when the cars come. At other times, when N_A is zero, there are no attention samples. During this time, the only processing and analysis done

is the sensor sampling. Figure 12 shows that the temporal motion attention, calculated from the equation (7), evolves according to the motion activity in a pedestrian sequence.

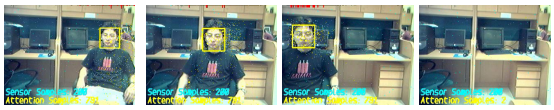
It should be understood that all the results are obtained by only processing a few samples in the visual data. There is no need to processing the entire data. It fulfills our aims of providing analysis have the ability to select the data to be processed.

5.2 Face Detection Results

We use our sampling technique to solve the face detection problem. Sensor samples are employed to obtain the current visual environment from the skin color, motion and speech cues. The face attention is maintained by the attention samples.



(a) Frame 76 (b) Frame 81 (c) Frame 106 (d) Frame 120

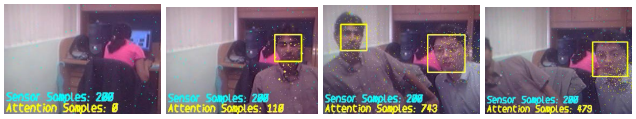


(e) Frame 147 (f) Frame 152 (g) Frame 158 (h)Frame 268

Figure 14.Face detection sequence 1. (a) static frame N_A=0 (b)A chair moves N_A=414 (c)the chair stopped. N_A=0 (d) a person comes. N_A=791 (e) a person. N_A=791 (f) one person. N_A=791 (g) one person. N_A=791 (h) static frame. N_A=2.

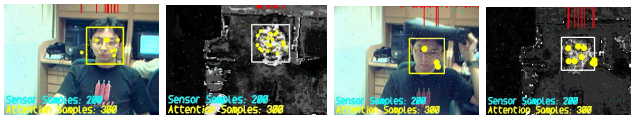
By our sampling method, the *adaboost* face detector is not applied on all the pixel and regions. The face detector is only executed on the attention samples which indicate the most probable face data.

As shown in Figure 14, N_S number of sensor samples is set to 200.The number and spatial distribution of attention samples can dynamically change according to the face attention. In Figure 14(a), there is no motion in the frame, so N_A, the number of attention samples is zero. No face detection is performed. In Figure 14(b), when a chair enters, it alerts the motion sensor and attention is aroused. N_A increases to 414. Face detection is performed on the 414 attention samples. But the face detector verifies that there is no face there. In Fig 14(c) as the chair stops, there is no motion and so the attention samples vanish. In Figure 14(d)-(h) attention samples come on with the face until the face vanishes. Note that depending on how much the attention is, the number of attention samples is different. For instance, N_A in Figure 15 (c) is 743 which is bigger than in Figure 15 (b) and (d) since Figure 15 (c) has two attention areas whereas Figure 15 (b) and (d) only have one. Figure 15 (c) also shows our sampling technique can maintain more than one attention region. Figure 16 (a) is a face under normal light. Figure (b) shows its skin color saliency map calculated by the equation (13). Figure 16 (c) and (d) are a shadowed face and its skin color saliency map. Figure 16 (b) and (d) shows that the feedback of face detector can update the skin color model H_t and make it more adaptive to the visual environment.(note that the skin color saliency map as shown in Figure 16 (b) and (d) is not necessary to be maintained in our method). Figure 17 shows that speech cue can help to create the attention samples when there is not any motion attention initially. As shown in Figure 17 (c), even speech is off, relevant attention samples still survive by being giving higher weights from the feedback of the previous face detection.



(a) N_A = 0 (b) N_A = 110 (c) N_A = 743 (d) N_A = 479

Figure 15. Face detection sequence 2.



(a) (b) (c) (d)

Figure 16. Skin color histogram H_t updated by feedback from the final analysis (face detection).

(a)Frame 2 (b)Frame 10 (c) Frame 42 (d) Frame 100

Figure 17. Speech experience (a) speech off N_A=0.(b) speech on. N_A becomes 1000. face detected.(c) speech off. N_A becomes 711(feedback from previous face detection). Face is detected (d) speech on. face detected. N_A becomes 1000.

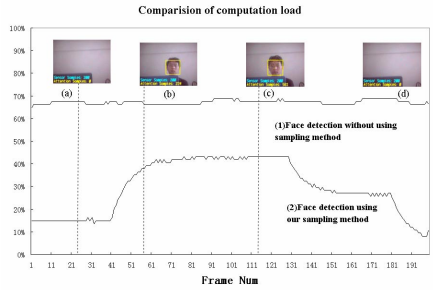


Figure 18. Comparison of computation speed.

We use a USB web camera to perform the real time face detection on a Pentium III 600MHz laptop. The graph of the computation load in this real time scenario is shown in Figure 18. In this experiment, curve 1 shows the computation load of the *adaboost* face detection while curve 2 indicates the computation load of our experience based sampling with *adaboost* face detector. This figure shows that by using our experience based sampling technique, computation complexity can be significantly reduced. In addition, the computation complexity also varies. When there is no face attention (see frame (a) and (d)), the only process is sensing by employing sensor samples. When the face comes (see frame (b) and (c)), the process includes attention samples and its load goes up.

6. CONCLUSIONS

In this paper, we describe a novel sampling based framework for multimedia analysis called experience based sampling. Based on this framework, we can utilize the context of the experiential environment for efficient and adaptive computations. Inferring

from this environment, the analysis procedure can select its data of interest while immediately discarding the irrelevant data. As an example, we utilize this framework for the face detection problem. The results establish the efficacy of the sampling based technique. In the future, other applications like adaptive streaming and surveillance and more sources of different modalities will be further investigated.

7. ACKNOWLEDGEMENT

We are grateful to Ramesh Jain of GeorgiaTech for his stimulating discussions on Experiential Computing during his visit to NUS in October 2002. This work has been inspired by these discussions.

8. REFERENCES

[1] Duda, R.O., Hart, P.E., and Stork, D.G. Pattern Classification. Wiley_Interscience, 2000.

[2] Sung, K.K., and Poggio, T. Example-Based Learning for View-Based Human Face Detection. Tech. Rep. 1532, M.I.T.: Artificial Intelligence Laboratory and Center for Biological and Computational Learning, 1994.

[3] Rowley, H., Baluja, S., and Kanade, T. Neural Network-based Face Detection. IEEE Trans. Pattern Analysis and Machine Intelligence, 1998.

[4] Viola, P., and Jones, M. J. Robust Real-time Object Detection. Tech. Rep. CRL 2001/01, Compaq Cambridge Research Laboratory, Cambridge, MA, 2001.

[5] Li, S.Z., Zhu, L., Zhang, Z.Q., Blake, A., Zhang, H.J., and Shum, H. Statistical Learning of Multi-View Face Detection. In Proceedings of The 7th European Conference on Computer Vision. Copenhagen, Denmark. May, 2002.

[6] Jain, R. Semantics in Multimedia Systems. Keynote at International Conference on Multi-Media Modeling, Taipei, January 8-10,2003.

[7] Jain, R. Experiential Computing.2003. . http://jain.faculty.gatech.edu/unpublished/experiential_computing.pdf

[8] Jain, R. Out-of the-Box Data Engineering. Key note at International Conference on Data Engineering, March, 2003.

[9] Chung, D., Hirata,R., Mundhenk, T. N., Ng, J., Peters, R. J., Pichon, E. , Tsui, A., Ventrice, T. , Walther, D., Williams, P., and Itti, L. A New Robotics Platform for Neuromorphic Vision: Beobots. In Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02), Tuebingen, Germany, 2002, in-press.

[10] Itti, L., and Koch, C. Computational Modeling of Visual Attention. Nature Reviews Neuroscience, Vol. 2, No. 3, pp. 194-203, Mar 2001.

[11] Ma,Y-F., Lu,L., Zhang, H-J., and Li,M-J. A User Attention Model for Video Summarization. ACM MM02, 2002.

[12] Walther, D., Itti, L., Riesenhuber, M., Poggio, T. and Koch, C. Attentional Selection for Object Recognition - a Gentle Way. In Proc. of 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02), Tuebingen, Germany, 2002.

[13] Navalpakkam, V., and Itti, L. A Goal Oriented Attention Guidance Model. In: Proc. 2nd Workshop on Biologically Motivated Computer Vision (BMCV'02), Tuebingen, Germany, 2002.

[14] Miller, E.K. The Prefrontal Cortex and Cognitive Control. Nature Rev. Neurosci. 1. 59-65(2000).

[15] Dey, A.K., and Abowd, G.D. Towards a Better Understanding of Context and Context-awareness. In H-W Gellerson, editor, Handheld and ubiqitous computing, number 1707 in Lecture Notes in Computer Science, pages 304-7. Springer, September 1999.

[16] Lieberman, H., and Selker, T. Out of Context: Computer Systems That Adapt to, and Learn From, Context. IBM Systems Journal 39, Nos. 3&4, 617-632 (2000, this issue).

[17] Want, R., Hopper, A., Falcao, V., and Gibbons, J. The Active Badge Location System. ACM Transactions on Information Systems 10(1) (1992) 91-102.

[18] Isard,M., and Blake, A. Condensation-conditional Density Propagation for Visual Tracking. International Journal on Computer Vision, 29(1):5-28,1998.

[19] Colombo, C., Rucci, M., and Dario, P. Integrating Selective Attention and Space-variant Sensing in Machine Vision. In Jorge L.C. Sanz, editor, Image Technology: Advances in Image Processing, Multimedia and Machine Vision,pages 109-128. Springer,1996.

[20] Chang, E-C., Mallat, S., and Yap, C. Wavelet Foveation. J. Applied and Computational Harmonic Analysis, volume 9, number 3, October 2000, pages 312-335.

[21] Schwartz,E.L., Greve,D.N., and Bonmassar,G. Space-variant active vision: definition, overview and examples. Neural Networks, Vol.8 No.7-8, pp.1297-1308, 1995.

[22] Wang, J., Achanta,R., and Kankanhalli, M.S. A Hierarchical Framework for Face Tracking Using State Vector Fusion for Compressed Video. In the 28th International Conference on Acoustics, Speech, and Signal Processing, 2003.

[23] Wang, J. Detecting and Tracking Human Faces in Compressed Domain for Content Based Video Indexing. Master Thesis, School of Computing, National University of Singapore, 2002.

[24] Ghahramani, Z., and Hinton,G. Parameter Estimation for Linear Dynamical Systems. Technical Report CRG-TR-96-2, Dept. Comp.Sci., Univ. Toronto, 1996. http://www.cs.toronto.edu/~hinton/absps/tr96-2.html

[25] Doucet, A., Godsill,S.J., and Andrieu, C. On sequential Monte Carlo sampling methods for Bayesian filtering. Statist. Comp., 10:197-208, 2000.

[26] Rubin, D.B. Using the SIR Algorithm to Simulate Posterior Distributions (with discussion), in Bayesian Statistics 3, eds. J.M. Bernard, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, New York: Oxford University Press, pp. 395-402.1998.

[27] Gordon, N.J., Salmond, D.J., and Smith, A.F.M. Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. IEEE Proceedings F, 140(2):107-113, 1993.

[28] Itti,L. and Koch, C. Feature Combination Strategies for Saliency-based Visual Attention Systems. J. Electronic Imaging,Vol. 10, No. 1, pp. 161-169, Jan 2001.

[29] Carpenter, J., Clifford,P., and Fearnhead, P. Building Robust Simulation-based Filters for Evolving Data Sets. Technical report, University of Oxford, Dept. of Statistics, 1999.

[30] Bradski, G. R. Computer Vision Face Tracking For Use in a Perceptual User Interface. Intel Technique Journal, 2nd quarter'98.

[31] http://www.comp.nus.edu.sg/~mohan/ebs/ExpSampling.htm