Experiential Sampling for Video Surveillance

Jun Wang
Department of Mediamatics
Faculty of EEMCS
Delft University of Technology
j.wang@ewi.tudelft.nl

Mohan S Kankanhalli, Weiqi Yan
Department of Computer Science
School of Computing
National University of Singapore
{mohan,yanwq}@comp.nus.edu.sg

Ramesh Jain
School of ECE and College of Computing
Georgia Institute of Technology
jain@ece.gatech.edu

ABSTRACT

Due to the decreasing costs and increasing miniaturization of video cameras, the use of digital video based surveillance as a tool for real-time monitoring is rapidly increasing. In this paper, we present a new methodology for real-time video surveillance based on *Experiential Sampling*. We use this framework to dynamically model the evolving attention in order to perform efficient monitoring. We exploit the context and past experience information in order to detect and track moving objects in surveillance videos. Moreover, we take the situation of multiple surveillance cameras into account and utilize the experiential sampling technique to decide which surveillance video stream to be displayed on the main monitor. This can tremendously help in reducing the manual operator fatigue for multiple monitor situation. We have implemented the developed algorithms and experimental results have been presented to illustrate the utility of the proposed technique.

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.

General Terms

Algorithms, Performance, Design, Experimentation.

Keywords

Video Surveillance, Real-time, Sampling, Dynamical Systems, Experiential Computing.

1. INTRODUCTION

Video surveillance is being increasingly used for traditional and non-traditional security applications such as real-time temperature scanning during the SARS outbreak in East Asia, monitoring of shopping malls and ATMs as well as industrial supervisory use. The decreasing costs coupled with rapid miniaturization of the video camera have enabled its widespread use on highways, airports, railway stations and on-board vehicles. The recent trend of coupling video cameras to cell-phones will only accelerate this trend. Therefore, research in video surveillance is moving into the mainstream with the focus on day-to-day applications and uncontrolled outdoor scenarios. And it is moving away from mere data collection with manual observation to intelligent analysis of events and actions at a semantic level without the intervention of humans [4,20,21].

Without appropriate management, the huge volume of filed video data will be impossible to be analyzed for events of interest. Video surveillance chiefly deals with spatio-temporal data which have the following attributes:

- They possess tremendous volumes;
- The data is dynamic with temporal variations with a resultant history;
- Some data can be live with real-time processing and filtering requirements;
- It does not exist in isolation – it exists in its ambient context with other data.

If a video surveillance technique does not fully consider the above attributes, it can lead to tremendous computational inefficiency as well inflexibility in the processing. The inefficiency stems from the inability to filter out the relevant aspects of the data and thus considerable resources are expended on superfluous computations on redundant data. If the ambient context is ignored, the technique cannot react to the changing environment. Thus, the surveillance 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. The human visual system is particularly adept at understanding the surrounding environment at appropriate accuracy quite efficiently. This is due to many factors: 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* [10,11] play an important role for the human visual perception to understand the visual scene accurately and quickly. We would like to incorporate some of these strategies into the techniques for video surveillance. The main idea in this paper is to enable surveillance techniques with an ability to "focus" precisely on the data of interest while being

IWVS'03, November 7, 2003, Berkeley, California, USA.

simultaneously sensitive to the current context as well as the assimilated past experiences.

In order to achieve this, we utilize a novel technique called *experiential sampling*, i.e., sampling the data in the experiential environment. The main theoretical technique has been introduced in another paper [26]. In this paper, we apply this experiential sampling technique to the problem of video surveillance and demonstrate its utility for this problem. The basic idea is to sense the contextual information in the experiential environment in order to build a sampling based dynamic attention model to maintain the focus towards the interest of the current surveillance task. Only the relevant samples are considered for the surveillance analysis. These samples succinctly capture the most important data. Moreover, the past samples influence future sampling via feedback. This mechanism ensures that the analysis task benefits from past experience.

The experiential sampling technique can be utilized for many aspects of video surveillance such as object (face or vehicle) detection, object recognition and object tracking. For example, intruders can enter only from the boundary of the scene. Therefore, when there are no intruders in the scene, the analysis task for surveillance 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 experiential sampling technique. We present results for two applications – face detection in surveillance videos and for multiple cameras video surveillance.

Our contributions in this paper can be summarized as follows. We introduce the concept of experiential sampling which provides a new framework to tackle the problem of the adaptation and efficiency for video surveillance applications. This method can adaptively provide attended samples to help avoid exhaustive analysis which we have demonstrated by a face monitoring application. We have also shown how exactly the experiential sampling technique can be utilized for multiple camera video surveillance. We present experimental results to demonstrate the efficacy of our technique.

2. RELATED WORK

The main purpose of video surveillance is to allow for secure real-time monitoring. The primary research issue of video surveillance is the automated detection and tracking objects, events and patterns. Finally, alerting responses to signify the alarm conditions need to be presented to a decision maker. Thus a change detection model is a fundamental part of a video surveillance system [4,20,21].

In order to achieve these purposes, Wu et al. [27] provide invariant feature extraction and biased statistical inference schemes for video surveillance. There are two algorithms presented in Lienhart et al. [13], one is based on the heuristics and statistics of measurable features and the detection of commercial blocks within TV broadcasting, another is to perform detection and recognition of known commercials with a high accuracy. It relies on a database of known commercial spots. Both algorithms have been combined into a self-learning system.

Marchesotti et al. [14] provide an algorithm to track moving objects in a scene for dynamic occlusions between moving objects, especially tracking pedestrians moving in indoor environments. The tracking method is based on joint application of Kalman filtering and correlation-based shape matching techniques. A technique based on contour analysis and shape description is used to count people in the scene, shape matching algorithm is based on the maximization of a correlation function varying with the shape pose parameters.

A semi-automatic video surveillance approach has been presented in [15] to overcome the drawbacks of traditional video surveillance systems and an alarm generator has been implemented in order to prefilter the data. A symbolic representation of events has been created to generate alarms and to store them in a database for off-line consulting.

Event detection is considered in Stringa et al. [23]. The purpose of event detection is to send an alarm signal to the human operators, utilizing the fact that a pixel belongs to a moving object if it is different from both of its background and the previous acquired images. A sequence of interesting video shots is retrieved by using a feature vector to detect an event.

Isaac and Medioni [9] have proposed a graph representation of the moving regions extracted from a video acquired by a moving airborne platform. It allows dynamic inference of moving objects in order to perform robust tracking. It also provides a confidence measure characterizing the reliability of each extracted trajectory.

In [16], the authors propose an adaptive algorithm to robustly detect the illumination changes in outdoor video-surveillance images. This is achieved by using a background updating module together with a segmentation algorithm, especially for sudden changes.

In [17], an image mosaic is constructed for the panoramic multilayer background image allowing one to use common change detection algorithms to search for change detection. Three different change detection techniques using the multilayer background: mean criterion, minimum difference criterion and minimum distance criterion are proposed for change detection for a mobile camera. [19] also provides similar results, with the difference being that during the on-line phase, the acquired images are compared with a portion of the panoramic background. The minimum difference criterion is shown to give the best performance among the three candidates.

Granelli et al. [7] describe an algorithm to detect and remove the artifacts in surveillance videos. The corrupted video frames especially JPEG images are detection and recovered. The presented algorithm enhances the performance of the video surveillance system without affecting the real-time performance.

In video surveillance, not only is the image mosaic used to construct the background of scene, but image stabilization is also done. In [18], a motion composition and image registration methods has been provided. The displacements of image sequence are detected by using a new algorithm instead of global optical-flow estimation. Two evaluation methods have been suggested to measure the motion composition: Interframe Transformation Fidelity (ITF) and Global Transformation Fidelity (GTF).

Stringa et al. [24] present an interesting application of multimedia surveillance system. This system is used to alert the surveillance operator when an abandoned object is detected in the waiting room of an unattended railway station.

The Sampling Importance Resampling (SIR) method which can be used for modeling evolution of distributions was proposed by Rubin [22]. The dynamics aspects were developed by Gordon *et al* [6]. 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 [3]. A special case of this framework has been used for the purpose of visual tracking in [8]. Though we also utilize the sampling method, we use it to maintain the attention. Unlike these works, we use the sampling technique to maintain the dynamically evolving attention. In [26], we also utilize the sampling method to maintain the attention. Thus, unlike [8], the number of samples dynamically changes for the purpose of adaptively representing the temporal 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 [10,11,12]. One of the key technical challenges in experiential computing is information assimilation, i.e., how to process real time data from multiple sensors. Our research in this paper aims to provide a sampling based dynamical framework to solve this problem in the context of multiple camera video surveillance.

3. EXPERIENTIAL SAMPLING

In this section, we introduce the proposed method of monitoring the activity in the video surveillance environment based on the experiential sampling technique that we have initially proposed in [26].

The key idea of *experiential sampling* technique (ES) is to look in the direction that is semantically most rewarding and infer from context as well as experience when data streams do not provide reliable features. In video surveillance applications, the direction can be inferred from the whole environment of the surveillance system.

A surveillance system can include numerous sensors ranging from video, sound as well as other sources such as X-ray or infrared sensor data. Therefore, the surveillance experiential environment includes all type of source data and their relations (for instance, the positional relations among sensors).

Consider the scenario when we are dealing with video sensors data in a surveillance system. The ES approach can be stated as follows: first, sensor samples are scattered uniformly among those sensors and then they are fused to get the current state of the experiential environment in the surveillance system. Based on the information obtained by the sensor samples, the attention samples are aroused to pick up the most rewarding sensors or features of sensors (attention area) for further analysis.

In this section, we begin with the descriptions of surveillance environment sensing (Section 3.1) and saliency activity monitoring (Section 3.2 and 3.3). We then give our proposed *ES* based video surveillance algorithm (Section 3.4). Finally, a multi-camera surveillance application that utilizes our proposed method is described in Section 3.5.

3.1 Surveillance Environment Sampling via Sensor Sampling

In this section, we first use sensor samples and attention samples to represent the surveillance environment. We then give the method to utilize sensor samples to sense such environment.

3.1.1 Surveillance Environment

In a surveillance system, there are numerous sensor devices. We define the kth sensor devices as SD_k ,where k ranges from 1 to p and p is the total number of sensor devices in the surveillance system.

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

$$e_t = \{S_1(t), A_1(t), ..., S_p(t), A_p(t)\} \quad (1)$$

The environment e_t comprises of *sensor samples* $S_k(t)$ and the *attention samples* $A_k(t)$ for each sensor device. The sensor samples are basically uniform random samples at any time t which constantly sense the environment from one particular sensor. The attention samples are the dynamically changing samples which essentially represent the data of interest at time t in one particular sensor. 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 surveillance task at hand can work only with these samples instead of the entire sensor 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. In the video sensor, they can be sensed by multiple cues from the visual environment which can subsequently be fused. This will be introduced in the next section. The obtain of the attention samples will be discussed later in Section 3.2

3.1.2 Sensor Samples for Surveillance Environment

In this paper, we only analyze the video data in the surveillance system. However, other type data in the system can be used to infer the environments and help analyze the video data as well [26].

When we deal with video data, 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 currently processed visual data or information from their accompanying data. Basically, sensors can sense these cues in order to infer the state of the environment.

For simplicity, we will drop k, the index of each device, when we discuss a single video stream.

In our framework, $S(t)$ is a set of $N_S(t)$ sensor samples at time t which estimates the state of the multimedia surveillance 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 parameters and N_S denotes the number of sensor samples at any point in time. $S(t)$ is then defined as:

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

where $s(t)$ depends on the type of multimedia data. For spatial data, $s(t)=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$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 value of that particular cue at that sample coordinate. Now it can be easily seen that

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

where α_j is the importance of the j^{th} cue. So we basically employ the linear combination as the sensor fusion strategy. 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.

3.1.3 Motion Cue

It is well-known that motion is one of the most important activities need to be monitored in the surveillance system. In the simplest form, we can define it as:

$$w\_sp_{MT}(x,y) = |i_t(x,y) - i_{t-1}(x,y)| \quad (4)$$

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

Therefore, 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 $w\_sp^i_{MT}$ is obtained by using equation (4).

The algorithm of calculating the weight of each sensor samples lists below:

Algorithm *(calculating sensor samples)*

START: 1. FOR sensor samples i = 0 to N_S

Calculate $w\_sp^i_{MT}$ *using equation (4)*

IF $w\_sp^i_{MT} > T_1$ *THEN*

$\pi_i^S(t) = w\_sp^i_{MT} / f$

End.

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

3.2 Salient Activity via Attention Sampling

In this section, we use attention samples [26] to model the salient activity in the video surveillance environments.

3.2.1 Salient Activity

All past work on extraction of visual attention uses saliency map to denote the visual attention in an image. The saliency map is built by either linear combination of features or by training. 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. Secondly, the temporal variation of attention is not modeled.

Contrarily, we use attention samples to model the salient activity (attention). The reason is following. The salient activity 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 1. 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.

In our framework, the number of attention samples dynamically evolves so the number will be increased when more attention is required and vice-versa. This will be introduced in Section 3.3. Moreover feedback from the final analysis task is used to tune the attention model with time.

Thus, our attention sampling based model systematically integrates the top-down and bottom-up approaches to infer attention from the environment as well as experience.

3.2.2 Attention Samples for Representing Salient Activity

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

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

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.

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

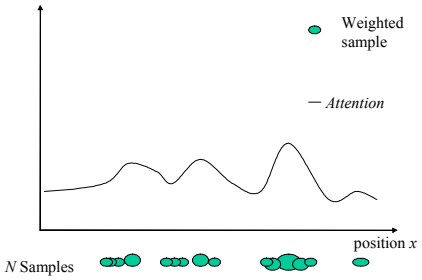


Figure 1. The multi-modal attention can be represented by *N* samples $a(t)=[s^1(t),\ldots,s^N(t)]$ and their weights $\Pi_{(t)}=[\pi^1(t),\ldots,\pi^N(t)]$.

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 [1] as shown in Figure 2. 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.

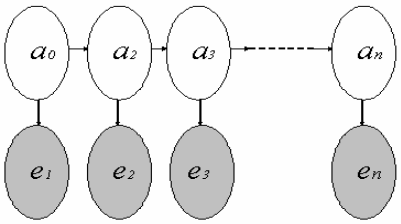


Figure 2. State-space model for attention.

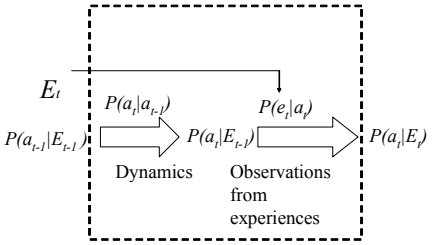
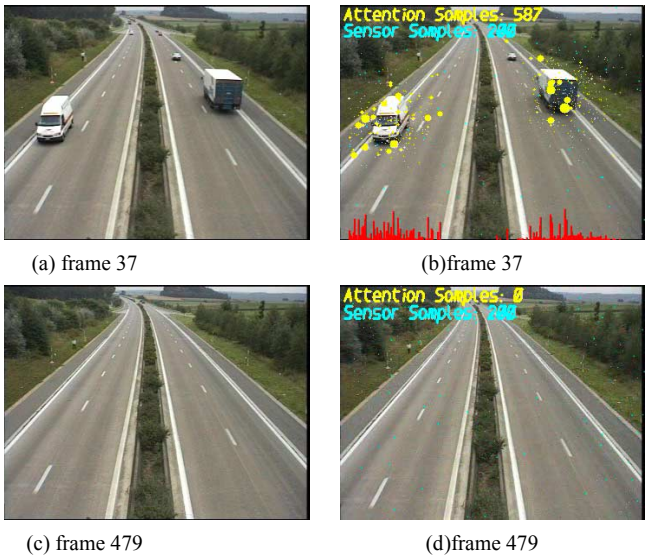


Figure 3. Iteration of calculating visual attention state density $P(a_t|E_t)$ in state-space model. By knowing previous state density $P(a_{t-1}|E_{t-1})$ and current experiences e_t, $P(a_t|E_t)$ can be approximated by a sampling method in the form of samples.

The dynamical system with its input and output are described in Figure 3. The explanation of the probability distribution is listed in Figure 4.

$P(a(t)\mid E(t))$:	The *a posteriori* probability of attention at time t given the experiential environment until time t
$P(e(t)\mid a(t))$:	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

Figure 4. The probability distribution



(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.

3.3 Attention Saturation

The temporal attribute of the spatio-temporal data requires the ability of having a varying amount of attention at different times. In [26], we use the average weight of all the sensor samples to represent the overall temporal fusion of the current state. In this paper, we refine this concept and introduce the notion of *attention saturation* to measure the attention in a given time slice. For instance, the attention saturation of motion in Figure 5(*a*) is higher than that in Figure 5(*d*). The attention saturation in this case can be calculated as the sum of attention in the spatial direction. Its value ranges from 0 (lowest, no attention) to 1 (highest, full attention). We define the attention saturation as $ASat(t)$:

$$ASat(t) = f_N\Big(\int_{Spatial} P(a(t)\mid E(t))\Big) \qquad (6)$$

where f_N is the mapping function which is used to normalize the value into range [0,1]. Usually f_N can be defined as a linear function.

$$f_N(x) = \xi \cdot x + b \quad (7)$$

where ξ is a scaling factor and b is the threshold. Those two parameters should be configured so that the output is uniformly scattered in the range [0,1].

The current attention is essentially captured by the sensor samples. Therefore, the discrete form is given below:

$$ASat(t) = f_N\left(\frac{1}{n}\sum_{q=[t-n,t]}\left(\frac{1}{N_S}\sum_{i=1}^{N_S}\pi_i^S(q) + \sum_j \beta_j w\_tp_j(q)\right)\right) \quad (8)$$

where β_j is the importance of the j^{th} temporal cue. 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.

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 $N_A(t)$ varies with time. For instance, in the traffic monitoring application shown in Figure 2, Figure 2 (a) has more motion activity and hence needs more attention samples to represent this motion attention. As shown in Figure 2 (b), 567 attention samples (marked as yellow dots) are required to represent this motion attention using our method. In contrast, Figure 2 (c) has less motion and needs fewer attention samples. As shown in Figure 2 (d), no attention samples are needed. Unfortunately, all previous image based attention models lack the ability to model this adaptive behavior.

Thus, we can utilize attention saturation to measure the attention at a given time instance (the temporal attention) as shown in the experiment in Fig 10.

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

$$N_A(t) = N_{Max}\, ASat(t) \quad (9)$$

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

3.4 Video surveillance by the experience based sampling technique

We are now ready to fully describe our video surveillance approach based on the background developed so far.

ES based video surveillance algorithm

t=0:

Experiential environment sensing

1. Initialization:

   FOR each SD_k(video sensor device)

   1.1 N_{sk}(number of sensor samples)← initialization.

   1.2 N_{Maxk} (maximum number of the attention samples)← initialization

   1.3 $N_{Ak}(t)$(number of sensor samples at time 0)← 0

   ENDFOR /* each video sensor device SD_k*/

   FOR each SD_k

2. Sense the surveillance environment:

   2.1 $S_k(t)$ (Set of sensor samples)← uniformly sampling .

   2.2. $\Pi^S(t)$ (Weights of sensor samples)← Employing Equation (3) (currently we use the motion cue).

   2.3 If $N_{Ak} > 0$, perform step 2.2 for $\Pi^A(t)$ (Weight of attention samples)

3. Compute the attention saturation:

   3.1. $ASat_k(t)$ ← employing equation (8).

   3.2. $N_{Ak}(t)$ ← using equation (9).

   ENDFOR /* each video sensor device SD_k*/

4. If $N_{Ak} = 0$, set $t = t+1$ and go to step 2. Otherwise go to step 5.

Building the attention model and attention driven analysis for the video sensor data from SD_k.

5. Perform re-sampling by using the method in [26] to obtain the attention samples:

6. Attention driven analysis: Perform the actual video analysis task on the $N_{Ak}(t)$ attention samples. For example, if the task is face detection, do it on the regions of the attention samples.

7. Treat sensor samples N_{Sk} and attention samples $N_{Ak}(t)$ as a whole and normalize their weights $\pi^S(t)$ and $\pi^A(t)$ to make $\sum \pi^A(t) + \sum \pi^S(t) = 1$

8. Create a new set of attention samples for time $t+1$ by propagating the attention samples $N_{Sk}(t)$ by dynamical evolution as expressed in Section 3.2.2.

9. $t = t+1$; go to step 2.

As a general analysis framework for video surveillance, our proposed approach can be used for a variety of video surveillance tasks, especially real-time applications like traffic monitoring and pedestrian surveillance. As a test case, we have first applied this framework for the face detection problem in videos whose results are presented in Section 4.1. And then we have developed the multiple camera surveillance scenario. We will first describe it before presenting the experimental results.

3.5 Multiple Camera Surveillance

In the real-world usage of video surveillance, multiple cameras are utilized in a wide spectrum of applications for the purpose of monitoring. As a result, multiple monitors are used to display the various output video streams. Manual monitoring of multiple screens is indeed tedious and operators are prone to fatigue which can have disastrous consequences. Thus, it would be extremely useful to reduce the data from multiple screens into one main screen for monitoring. This is reasonable since usually only one screen is of active interest which engages the attention of the operator. Even if there is no manual observer, finding the most

relevant camera data stream is useful for automated analysis such as object detection. Thus in this section, we will develop an algorithm to analyze the importance of the video frames from the multiple video cameras by using the experiential sampling technique.

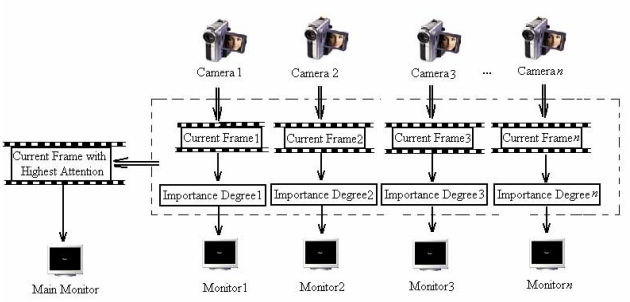


Figure 6. Our set-up for multiple camera surveillance

In this multiple camera surveillance scenario, there are p numbers of camera C_k. The experiential environment can be redefined as:

$$e_t = \{C_1(t), C_2(t), ..., C_p(t)\} \text{ and } C_k(t) = \{S_k(t), A_k(t)\} \quad (10)$$

where $A_k(t)$ and $S_k(t)$ are the attention samples and sensor samples of the camera $C_k(t)$ at time slice t. By this definition, we can use the algorithm introduced in the previous section to acquire the attention saturation of the motion activity. We assume that the motion attention saturation precisely reflects the importance of current video frame. At any point in time, we designate the output of one camera as the relevant camera whose output is displayed on the main monitor. The actual camera picked dynamically changes based on the data. Our strategy is illustrated in Figure 6. In figure 6, several cameras are fixed and their current frames are extracted for analysis. The attention saturation of these frames are computed and compared. Only the frame with the maximum attention saturation will be extracted and displayed on the main monitor.

Our algorithm for multiple camera video surveillance can be described as follows:

Algorithm: (Multiple Camera Video Surveillance)

START:

1. Collect all the video frames at this time instance;
2. Compute their attention saturation;
3. Find the most important frame;
4. Display the most important frame on the main monitor;
5. Continue till the end.

END

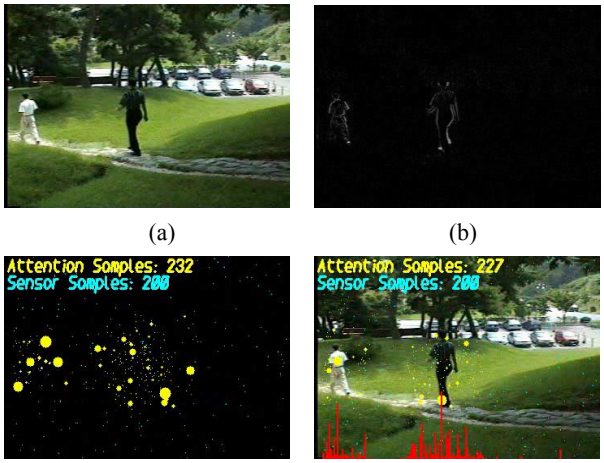
The advantage of our proposal is that we extract the most important frame at any given instance and render it on the main monitor which would greatly reduce the manual monitoring efforts. Note that this can be easily generalized for the n cameras and m monitors situation where $m < n$.

4. EXPERIMENTAL RESULTS

In this section, we present results from two experiments. The reader can see the all the results with actual videos at our website [28].

4.1 Single-camera scenario

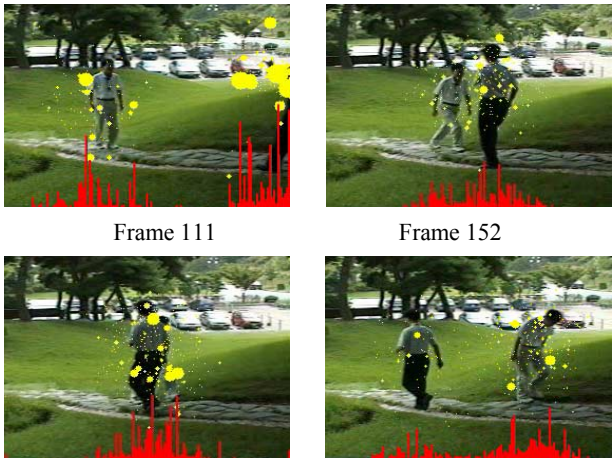
We test our method for the video of several pedestrian (Figure 7 and 8) and traffic monitoring sequences (Figure 9) in the one-camera scenario. There are 200 sensor samples randomly scattered spatially to sense the motion experience. Based on the sensor output, attention samples are created. Their numbers and spatial distribution are all determined by the motion experience. Figure 8 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 obtain the motion activity spatially.



(a) (b)

(c) (d)

Figure 7. 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 motion activity 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 8. Mpeg 7 Test sequence 2. Red bar shows the spatial motion activity 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 motion activity spatially. Both motion attention regions emerge and split during and after the crossing of the subjects

The weight of each attention sample is drawn using red bars along with the x direction to visualize the spatial motion activity in *x* direction. From Figures 7, 8 & 9, we can see that our approach can model multi-modal motion activity spatially quite well for the purpose of monitoring.

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 9. Traffic monitoring sequence. This figure illustrates the both spatial and temporal traffic activity inferred from motion experience. Blue points are sensor samples while yellow points are attention samples. Red bar shows the spatial traffic actvity 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.

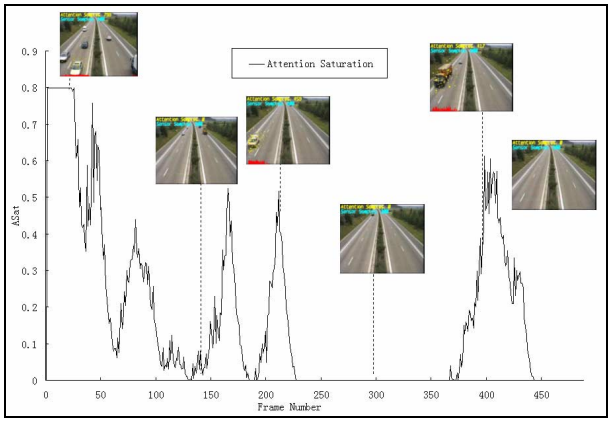
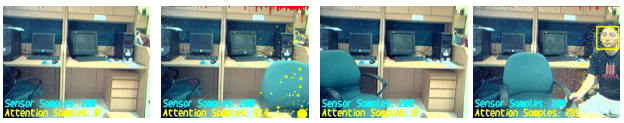


Figure 10. Motion attention saturation in the MPEG 7 clip.

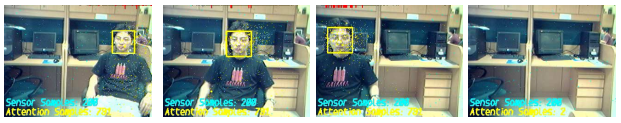
The evolution of temporal monitoring (represented by the number of attention samples) is shown in Figure 8 and 9. In Figure 9, the motion activity roughly reflects the traffic status at each time step. Therefore, our method here can be used for monitoring the traffic status also. It shows that the attention is only aroused when the cars come. At other times, when Attention saturation *ASat* is zero, there are no attention samples. During this time, the only processing and analysis done is the sensor sampling. Figure 10 shows that the attention saturation, calculated from the equation (8), evolves according to the motion activity in a traffic monitoring sequence.

It should be understood that all the results have been obtained by only processing a few samples in the visual data. There is no need to process the entire data. It fulfills our aims of providing analysis have the ability to select the data to be processed. In return, the processing is fast and easy to be real time.

Object detection is one of the important applications of video surveillance. We use our sampling technique to tackle the face detection problem. Sensor samples are employed to obtain the current visual environment from the skin color and motion cues. The face attention obtained by the skin color and motion cues is maintained by the attention samples.

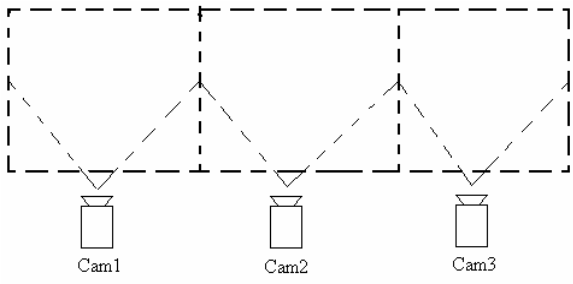


(a) Frame 76 (b) Frame 81 (c) Frame 106 (d) Frame 120

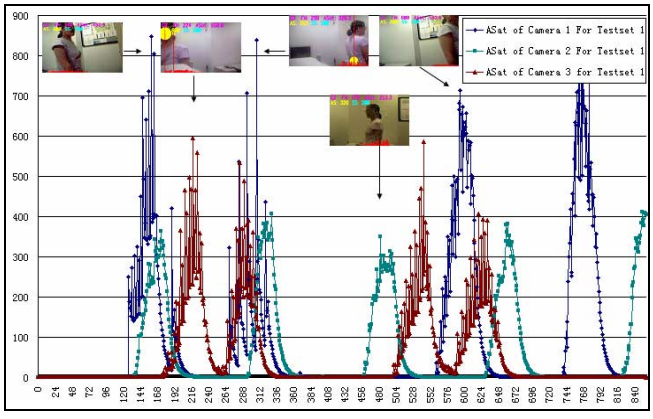


(e) Frame 147 (f) Frame 152 (g) Frame 158 (h)Frame 268

Figure 11. 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.



(a) Diagram **of Camera Position**

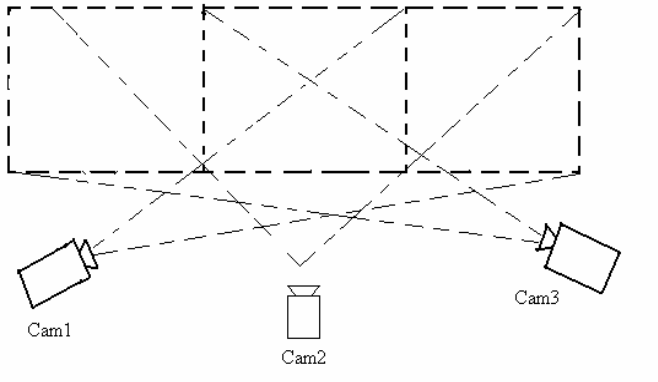


(b) Attention Saturation

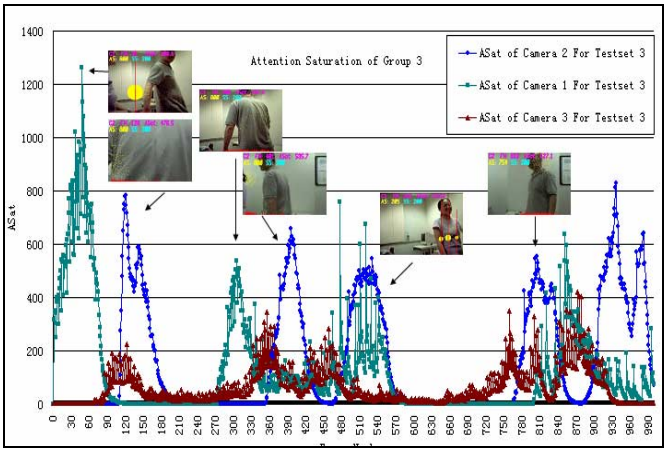
Figure 12. The attention saturation of surveillance videos (Group 1).

As shown in Figure 11, 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 11(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 11(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 11(c) as the chair stops, there is no motion and so the attention samples vanish. In Figure 11(d)-(h) attention samples come on with the face until the face vanishes.



(a) Diagram **of Camera Position**



(b) Attention Saturation

Figure 13. The attention saturation of surveillance videos (Group 2).

4.2 Multiple Camera based Video Surveillance

Figure 12 and figure 13 represent two kinds of experiments. Individual attention saturation of each frame in these three videos is calculated. We extract the important frames, namely the ones with the highest attention saturation to compose a new video, this new video will be displayed on the main monitor for the operators. Figure 12(a) is a diagram for the positions of three video cameras. Each camera covers only a limited field of view, and these viewing regions have no overlap. Figure 12(b) is the resultant diagram for the attention saturations of the three video streams. We can very clearly see the peaks occur at different times in the different cameras. A peak corresponds to the fact that someone is detected to be walking past that camera. Thus the main monitor can display only the output of the camera which is active at any instance of time (instead of constantly watching three monitors). Figure 13 shows another group of results. The three cameras have some overlaps in the field of view as shown in Figure 13(a). In this case, we can see the person walking on many monitors simultaneously. In this case, it becomes very difficult for a manual operator to decide which monitor to watch. However, by using attention saturation for experiential sampling, the peaks can be easily determined as shown in Figure 13(b). This information can be utilized to appropriately display the information on the main monitor.

5. CONCLUSIONS

In this paper, we describe a novel experiential sampling based framework for video surveillance. 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 (or the display monitor) can select its data of interest while immediately discarding the irrelevant data. The results establish the efficacy of the sampling based technique. We utilize this theory to detect human faces in surveillance videos. We also present an algorithm for multiple camera surveillance to display the most relevant camera output to be displayed on the main monitor. This can tremendously aid the manual monitoring of multiple cameras and can be used for alerting risky situations. We will continue to investigate the use of the experiential sampling technique for other video surveillance applications. Since other data (like audio, infrared or x-ray sensors) can be easily incorporated into our framework, we will further investigate multi-modal video surveillance using the experiential sampling framework.

6. ACKNOWLEDGEMENT

Weiqi Yan's work has been supported by a Fellowship from the Singapore Millennium Foundation.

7. REFERENCES

[1] 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.

[2] Duda, R.O. Hart, P.E. and Stork, D.G. Pattern Classification. Wiley Interscience, 2000.

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

[4] Foresti, G. L. Mahoen, P. and Regazzoni, C. Multimedia Video-Based Surveillance System, Requirements, Issues and Solutions. Kluwer Academic Publishers, 2002, USA.

[5] 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

[6] Gordon, N.J. Salmond, D.J. and Smith, A.F.M. Novel approach to nonlinear/Non-Gaussian bayesian state estimation, in IEEE Proceedings, 140(2):107-113, 1993.

[7] Granelli, F. Oberti, F. and Regazzoni, C.S. Adaptive post-processing error concealment based on feedback from a

video-surveillance system, in Proceedings of ICIP 2000, (Vancouver, Canada, September 2000).

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

[9] Isaac, C. Medioni, G. Detecting and tracking moving objects for video surveillance, in Proceedings of ICPR (Fort Collins CO. Jun. 23-25, 1999).

[10] Jain, R. Experiential Computing, Communications of the ACM, 46(7): 48-55, July 2003.

[11] Jain, R. Out-of the-Box Data Engineering. Keynote Address at International Conference on Data Engineering (ICDE 2003), March 2003.

[12] Jain, R. Semantics in Multimedia Systems. Keynote Address at International Conference on Multi-Media Modeling (MMM 2003), Taipei, January 2003.

[13] Lienhart, R. Christoph, K. and Wolfgang, E. On the detection and recognition of television commercials, in Proceedings of IEEE conference on multimedia computing and systems (Ottawa Canada, June 1997), pp.509 - 516.

[14] Marchesotti, L. Marcenaro, L. Regazzoni C. S. Tracking and counting multiple interacting pedestrian in indoor scenes, in Proceedings of third IEEE international workshop on performance evaluation of tracking and surveillance, (Copenhagen Denmark, June 1, 2002).

[15] Marchesotti, L. Marcenaro, L. Regazzoni, C.S. A video surveillance architecture for alarm generation and video sequences retrieval, in Proceedings of ICIP2002 (Rochester NewYork USA, September 2002).

[16] Marcenaro, L. Gianluca, G. Regazzoni, C. Adaptive change detection approach for object detection in outdoor scenes under variable speed illumination changes, in Proceedings of Eusipco 2000 (Tampere, Finland).

[17] Marcenaro, L. Oberti, F. and Regazzoni, C. Change detection methods for automatic scene analysis by using mobile surveillance cameras, in Proceedings of ICIP 2000, (Vancouver Canada, September 2000).

[18] Marcenaro, L. Regazzoni, C.S. Image-stabilization for video-surveillance applications, In Proceedings of ICIP2001 Thessaloniki, Greece, 2001.

[19] Oberti, F. Marcenaro, L. Regazzoni, C. S. Real-time change detection methods for video-surveillance systems with mobile camera, in Proceedings of XI EUSIPCO2002, Toulouse France, September 2002.

[20] Regazzoni, C. Fabri, G. and Vernazzza, G. Advanced Video-based Surveillance System. Kluwer Academic Publishers, 2002, USA.

[21] Remagnio, P. Jones, G. Paragios, N. and Regazzoni, C. Video-based Surveillance Systems. Computer vision and Distributed processing. Kluwer Academic Publishers, 2002.

[22] 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.

[23] Stringa, E. Regazzoni, C.S. Content-based retrieval and real time detection from video sequences acquired by surveillance systems, in Proceedings of ICIP98 (Chicago Illinois, Oct. 1998), Vol. III, pp.138-142.

[24] Stringa, E. Sacchi, C. Regazzoni, C.S. A multimedia system for surveillance of unattended railway stations, in Proceedings of Eusipco1998 (Rhodes Greece, 1998), pp. 1709-1712.

[25] Viola, P. and Jones, M. J. Robust real-time object detection. tech. rep. CRL 2001/01, Compaq Cambridge Research Laboratory, Cambridge, MA, 2001.

[26] Wang, J. Kankanhalli, M.S. Experience-based Sampling Technique for Multimedia Analysis, in Proceedings of ACM Multimedia Conference 2003.

[27] Wu, Y. Jiao, L. Wu, G. Chang, E. and Wang, Y.F. Invariant feature extraction and biased statistical inference for video surveillance, in Proceedings of IEEE international conference on advanced video and signal based surveillance (Miami USA, July 2003).

[28] http://www.comp.nus.edu.sg/~mohan/ebs/VideoSurveil.htm