ICICS-PCM 2003
15-18 December 2003
Singapore

Adaptive Monitoring for Video Surveillance

Jun Wang [1], Wei-Qi Yan[2], Mohan S. Kankanhalli[2], Ramesh Jain[3], Marcel J.T.Reinders[1]

[1]*Faculty of Information Technology and Systems, Delft University of Technology*
{j.wang, m.j.t.reinders}@ewi.tudelft.nl

[2]*School of Computing, National University of Singapore*
{yanwq, mohan}@comp.nus.edu.sg

[3]*School of ECE and College of Computing, Georgia Institute of Technology*
jain@ece.gatech.edu

Abstract

Adaptability is one of the key issues in the important area of surveillance systems. Based on attention and sensor samples, the experiential sampling technique provides a general framework for analyzing video data. In this paper, we present a scheme for adaptive monitoring of surveillance objects by utilizing the feedback of the experiential sampling based video surveillance results to change the video camera parameters. Our framework first detects the moving objects in the surveillance video. We then analyze the output of this step to determine the state of the video camera settings. The relevant parameters of the video camera are continuously adjusted based on a proportional feedback control system. The fixed camera is thus adaptively tuned so as to obtain a good quality surveillance video output. We centrally frame the target object by doing panning and zooming operations. Moreover, we utilize the experiential sampling approach to capture the moving objects by utilizing the context which is a function of the current state of the environment as well as past experiences. This adds tremendous flexibility to the surveillance process which can be applied in a wide variety of monitoring situations.

1. Introduction

There is a growing interest in the use of video surveillance technique aided by the decreasing cost of sensors and an increasing set of vulnerabilities. Video surveillance can potentially provide a cost-effective means of eliminating or mitigating potential security breaches. Video surveillance has applications in offices, stores, public spaces and homes. In a previous study, we have employed the experiential sampling technique to detect as well as track moving objects and human faces in surveillance videos [12]. Our scheme handles video surveillance by using attention sampling and attention saturation. The main advantage of this scheme is that we indeed are able to capture objects of interest in the form of a linear dynamical system. In this paper, we will utilize the experiential sampling technique in a feedback control system for automatic adjustment of the parameters of a fixed video camera. We will essentially control the setting of panning and zooming parameters of a video camera.

Our motivation for studying the automatic adjustment of camera parameters stems from the fact that a camera for video surveillance is usually not manipulated after its installation. Regular calibration of such cameras for object tracking is desirable but not feasible if done manually. As a result surveillance videos often are of low quality due to changing ambient conditions especially for outdoor settings. High quality videos in focus with appropriate positions of moving objects require the fixed surveillance cameras to capture and track those moving objects at its optimal framing position with correct zoom and pan settings. Thus the automatic setting and control of such video cameras is a difficult but extremely useful task. Manual control of cameras via internet, wireless or remote radio channels is a feasible but tedious operation. We aim to have precise control of surveillance cameras via these channels automatically in order to obtain a superior quality surveillance video.

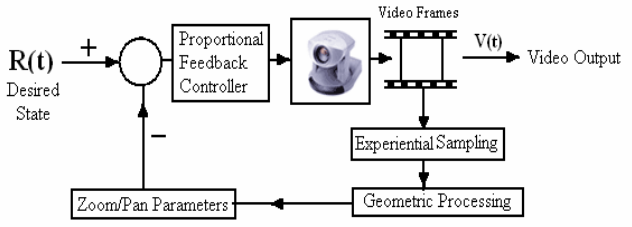


Figure.1 Feedback Control System for Surveillance

Figure 1 is the proposed framework to be discussed in this paper. Several techniques for the detection and tracking of moving objects have been investigated in the literature (see Section 2). We use the experiential

0-7803-8185-8/03/$17.00 © 2003 IEEE

sampling technique because it effectively takes advantage of the context information of surveillance videos. For example, we can predict the future motion of moving objects by using past knowledge that has been learned. Thus, our computation is performed in an experiential environment which is captured by both the sensor and attention samples. The number of attention samples used is therefore a function of the environment as well as past experiences. Consequently, experiential sampling creates a dynamical solution for real-time object detection and tracking.

The remainder of this paper will be structured as follows: Section 2 introduces the related work; our approach is described in Section 3; Section 4 demonstrates our experimental results, the conclusion and future work are presented at Section 5.

2. Related Work

Research in video surveillance focuses on day-to-day applications, i.e. uncontrolled indoor and outdoor scenarios. And it is moving away from the mere data collection with manual observation to intelligent analysis of events and human actions at a semantic level without the intervention of human beings. Regazzoni has organized several workshops related to video-based surveillance systems [1,2,3] while Chang et al. have some recent results on feature extraction and fusion for video surveillance [4][5].

Jain et al. have done pioneering work on video surveillance and achieved concepts of the multiple perspective interactive video [9,10]. Their approach is characterized as the early precursor of experiential computing [6]. Digital experience involves immersion in a rich set of data and information in a way that allows us to observe a subset of the data and information directly. Digital experience is a major but natural step in the evolution of technology.

In [7], the experiential computing paradigm has been elaborated. In an experiential computing environment, users directly apply all their sensors to observe data and information of interest related to an event. Furthermore, each user can explore the data according to his particular interest in the context of that event. Experiential environments free people from tediously managing enormous volumes of heterogeneous data. By dealing with spatial-temporal live data streams, experiential computing can address many new real world problems. The experiential sampling technique provides a formal basis for analysis in experiential environments.

In reference [11], we proposed a sampling based dynamic attention model for sensing the environment with voluminous spatial-temporal data. In this framework, sensor samples are used to gather information about the current environment and attention samples are used to represent the current state of attention. In this case, the task-oriented samples are inferred from the context and maintained by a dynamic sampling based system. Moreover, the past experiences and the current environment can be used to adaptively correct and tune the attention. This basically provides a powerful, yet compact, representation for experiential computing. The experimental results provided in [9] demonstrate the efficiency of this scheme for real-time applications.

In [12], we presented a new methodology for real-time video surveillance based on experiential sampling. We used the experiential sampling framework to dynamically model the evolving attention resulting in an efficient monitoring. We exploited the environment and past experience information in order to detect and track moving objects in surveillance videos. Moreover, we take multiple surveillance cameras into account and utilize the experiential sampling technique to decide which surveillance video stream needs to be displayed on the main monitor. This can tremendously assist the manual operator to deal with multiple video streams.

In this paper, we will employ experiential sampling for video surveillance in a feedback control system. First, the moving object in the video is captured by a fixed camera. Then the detected object (face in our experimental set-up) is analyzed and the corrective feedback required to be provided to the camera is deduced. After computing the feedback, we accordingly adjust the video camera parameters in order to capture the best possible surveillance video. This process works iteratively in a dynamical feedback control system. It must be noted that this work is different from the work on calibrating cameras for the pan and tilt operations [7] which focus more on the camera model than the feedback control aspects.

3. Adaptive Monitoring

In order to obtain robust and accurate results, video surveillance system should have *adaptability* in order to respond to the changes in the monitored environment. There can be potentially many variations in the monitored environment. These variations mainly come from two aspects: geometric aspects (location, size etc.) of the observed objects and the visibility quality (lighting, contrast) of the monitored environment.

Currently available video cameras can be manipulated to track the observed object of interest by performing panning and zooming operations. It is also possible to adjust to the visibility of the environment by automatically controlling camera parameters such as brightness and contrast. Therefore, it would be extremely useful to

systematically manipulate these camera parameters so as to adapt to the changing surveillance environment.

Our proposed adaptive monitoring method considers both geometric aspect (automatic panning and zooming) in order to follow the movements (position, size) of the monitored object and visibility aspect. Since the adjustment of visibility parameters is similar to our previous work about automatic adjustment video quality for home video in [13]. We do not experiment the automatic adjustment of the visibility parameters as opposed to only include it in our framework.

Our adaptive monitoring method is based on the experiential sampling (ES) [11] technique. Therefore, we first provide a brief introduction of the ES technique.

3.1 Experiential sampling (ES)

In this section, we only provide the outline of the experiential sampling algorithm. The interested readers may consult [11] for details. The ES algorithm can be briefly described as follows:

Algorithm: Experiential Sampling (ES):

1. Initialization: $t=0$
2. $\{SS(t)\}$ ← uniform sampling
3. $Asat(t)$ ← sum of $\{SS(t)\}$
4. $N_s(t)$ ← $Asat(t)$
5. *if* $N_s(t)$ =< 0 $t=t+1$; *goto* step 2
6. $\{AS(t)\}$← importance sampling from $\{SS(t)\}$
7. *for* each AS, perform analysis
8. $t=t+1$; *goto* step 2.

where $\{AS(t)\}$ and $\{SS(t)\}$ are sets of attention samples and sensor samples at time t respectively; $N_s(t)$ is the number of the attention samples decided by the current attention saturation. Current attention saturation is denoted as $ASat(t)$ which measures the task oriented attention at time t.

It can be understood that in the experiential sampling framework [11], the monitored environment is sensed by sensor samples. The number (possibly zero) of the attention samples aroused depends on the context (captured in step 4). The exact location of the attention samples also depends on the sensed environment (step 6). Therefore, it is clear that in the basic ES algorithm, $Env(t)$ the environment at time t is represented by the attention samples and the attention saturation. The relation is shown as follows:

$$Env(t) = \{ASat(t), AS(t)\} \quad (1)$$

3.2 Feedback Control with embedded ES

As we discussed earlier, in a surveillance system, there are two types of variations in the environment: geometric aspects of objects and the visibility. In order to adaptively monitor the environment, we embed the basic experiential sampling algorithm into a feedback control system. The state variable of this feedback control system are precisely the two factors of the environment which impact the surveillance task. Thus, the state vector of the feedback control system is represented by:

$$X(t) = \{Geom(t), Vis(t)\} \quad (2)$$

where $Geom(t)$ captures the geometric aspects of the surveillance object of interest and $Vis(t)$ represents visibility aspects of the object in the ambient environment. The state variables of this system (geometric aspects of object and the visibility of the environment) have to be estimated based on the previous time-instant attention samples and sensor samples. After obtaining those two factors, we can then adjust the camera parameters by providing the appropriate feedback According to:

$$Geom(t)=\{zoom(t), pan(t)\} \leftarrow \{AS(t\text{-}1)\} \quad (3)$$

$$Vis(t)=\{brig(t),contr(t)\} \leftarrow \{SS(t\text{-}1)\} \quad (4)$$

where $Geom(t)$, the geometric parameters at time slice t, includes the zoom parameters $zoom(t)$ and pan parameter $pan(t)$. $Vis(t)$, the visibility parameters to be tuned, includes $brig(t)$ the brightness and $contr(t)$ contrast parameters.

Having characterized the surveillance environment by these two state variables, the overall feedback control algorithm with experiential sampling embedded can be described as follows:

Input: surveillance video stream from a camera;

Output: surveillance object, camera feedback;

Procedure:

Feedback Control Algorithm:

1. $t=0$;Initialization
2. ***X(t)← {Geom(t), Vis(t)}***
3. $\{SS(t)\}$ ← uniform sampling
4. $Asat(t)$ ← *sum* of $\{SS(t)\}$
5. ***Vis(t+1) ← {SS(t)} [visibility feedback]***
6. $N_s(t)$ ← $Asat(t)$
7. *if* $N_s(t)$ > 0 *go to* step7
8. $t=t+1$; goto step 2
9. $\{AS(t)\}$← important sampling from $\{SS(t)\}$
10. *For* each *AS*, perform analysis.
11. **Geom(t+1) ← {AS(t)} [geometric feedback]**
12. $t=t+1$; *goto* step 2.

After we obtain the geometric parameters of the object at time *t*, we can decide the panning and zooming parameters of the camera for the next time instant *t*+1, and thus can use them to control the video display. We call this operation the *geometric processing*. The procedure is quite similar to the adjustment of the visibility parameters. We will describe the geometric aspect in detail now.

3.3 Geometric State Variables Feedback

Geometric processing aims to obtain the optimal framing for the surveillance object of interest. We try to frame a moving object (human face in our implementation) based on the attention samples. We then attempt to obtain the optimal zooming factor and translation vector to provide the feedback to the surveillance camera. We use the *proportional feedback control strategy* [8] for doing the corrective feedback. At first, we fix the desired state of the surveillance system. If the objects of interest are not framed properly (which will be indicated by the attention samples), we will reach our target state by appropriate modification of the pan and zoom parameters of the camera. For instance, if the tracked object is at the bottom of the video frame, we control the pan parameter of the camera and move the frame upwards. Similarly, if the tracked object is at the top of the video frame, we pan the frame downwards. Simultaneously, if the objects are not at the desired zoom level, we will control the zoom parameter to obtain the video with the appropriate size of the objects. The system settles down to a steady state after a period of transition when the feedback control system makes the appropriate corrective actions.

For our system, we have fixed the desired state of the system as follows: (a) the centroid of the object of interest should be located at the center of the video frame and (b) The ratio of the sides of the bounding box of the object of interest to the sides of the video frame should be 0.618. From the human visual system point of view, the object of interest at this framing state will exhibit the best observability. Such a video frame can then be suitably displayed on a monitor to be seen by a human operator. Based on the attention samples in the previous time instance, we can compute two parameters as follows:

$$O_c = \{AS(t\text{-}1)\}_c \quad (5)$$

$$AR = \text{Area}_i/\text{Area}_{\{AS(t\text{-}1)\}} \quad (6)$$

where the object centroid O_c is approximated by the centroid of all attention samples (denoted by $\{AS(t)\}_c$) while AR represents the area ratio of the total image area (Area_i) to the area of attention samples ($\text{Area}_{\{AS(t)\}}$).

We define $pan(t)$ as the camera pan position at time t. Therefore the new $pan(t)$ in step 11 of the feedback control algorithm can be formalized as

$$pan(t) = pan(t\text{-}1) + \alpha(O_c - \text{pan}(t\text{-}1)) \quad (7)$$

where α is the factor to control the panning speed. We have used α=0.3 in our experiments.

We define $zoom(t)$ as the object size at time *t*. The new $zoom(t)$ in step 11 of the feedback control algorithm can be obtained as follows:

Input: AR

Output: $zoom(t)$

Procedure:

if $AR > \lambda_{high}$ $zoom(t) = zoom(t\text{-}1)\text{-}1$ (zoom in)

if $AR < \lambda_{low}$ $zoom(t) = zoom(t+1)+1$ (zoom out)

where λ_{high} and λ_{low} are the two predefined zooming factors to adjust the ratio between the object and image. In our experiments, the value of λ_{high} used is 0.638 and the value of λ_{low} used is 0.598.

Notice that the error between the desired state of the system and the actual observed state of the system is used as the corrective stimulus to steer the system towards the desired state. This is basically due to the proportional feedback control strategy used.

4. Results

In this section, we show some resulting video clips for our experiments on adaptive video surveillance. The actual videos can be downloaded from our webpage [14]. We have used the AIPTEK HyperVcam video camera in our experiments. As shown in Figure 2, the camera automatically does a panning correction in order to frame the face appropriately. Figure 3 shows the predominant camera zooming based on the context. The zooming and panning parameters are adjusted by proportional feedback of the error of the centroid and size of the yellow bounding rectangle of the object of interest. Figure 4 shows the continuous adjustment of the zooming and panning parameters in order to the frame the object of interest based on the desired state of the feedback control system.

5. Conclusion and future work

In this paper, we have presented an adaptive monitoring system for video surveillance based on experiential sampling. The basis idea is to use a proportional feedback control strategy in order to appropriately adjust the zoom and pan parameters of the video camera so as to obtain the desired framing of the object of interest. The desired state of the system is up to the video surveillance system designer and it can vary

according to the desired application. Our experiments merely demonstrate one such possibility. Many improvements to this basic system are possible. In the future, we plan to utilize an ensemble of video cameras to do proper monitoring of a large spatial zone. The control strategy needs to be much more sophisticated in that case. We will investigate derivative and integral control strategies for such a system. Moreover, the video domain could be extended to include the adaptive adjustment of a symbiotic system of grouped multiple cameras and microphones for effective surveillance.

6. References

[1] P. Remagnio, G. Jones, N. Paragios and C. Regazzoni. Video-based Surveillance Systems, Computer Vision and Distributed Processing, Kluwer Academic Publishers, 2002, USA.

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

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

[4] G. Wu, Y. Wu, L. Jiao, Y.-F. Wang and E. Chang, Multi-camera Spatio-temporal Fusion and Biased Sequence-data Learning for Security Surveillance, Proc. of ACM Multimedia, 2003, Berkeley, USA, November, 2003.

[5] Y. Wu, L. Jiao, G. Wu, E. Chang and Y.-F. Wang, Invariant Feature Extraction and Biased Statistical Inference for Video Surveillance, Proc. of IEEE International Conference on Advanced Video and Signal Based Surveillance, Miami, July 2003.

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

[7] J. Davis and X. Chen, Calibrating Pan-Tilt Cameras in Wide-area Surveillance Networks, Proc. of IEEE International Conference on Computer Vision, 2003.

[8] B. C. Kuo, Automatic Control Systems, Wiley, 2003.

[9] P. H. Kelly, A. Katkere, D. Y. Kuramura, S. Moezzi, S. Chatterjee, R. Jain, An Architecture for Multiple Perspective Interactive Video, Proc. of ACM Multimedia 1995, San Francisco, California, USA, 1995: 201-212.

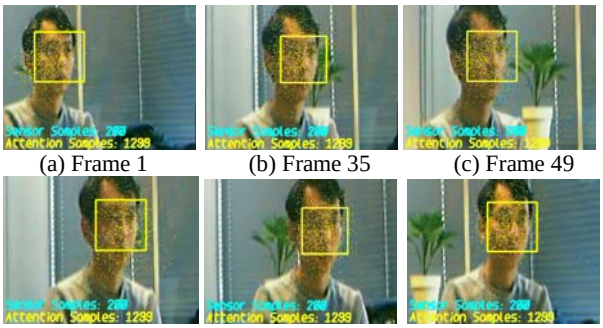
[10] S. Santini and R. Jain, A Multiple Perspective Interactive Video Architecture for VSAM, Proc. of the 1998 Image Understanding Workshop, Monterey, November 1998.

[11] J. Wang and M.S. Kankanhalli, Experience Based Sampling Technique for Multimedia Analysis, Proc. of ACM Multimedia Conference 2003, Berkeley, USA, November 2003.

[12] J. Wang, M.S. Kankanhalli, W.Q. Yan and R. Jain. Experiential Sampling for Video Surveillance, Proc. of First ACM International Workshop on Video Surveillance in ACM Multimedia 2003, Berkeley, November 2003.

[13]W-Q. Yan, M.S. Kankanhalli, Detection and Removal of Lighting and Shaking Artifacts in Home Videos, Proc. of ACM Multimedia 2002, Juan Les Pan, France, 2002: 107-116.

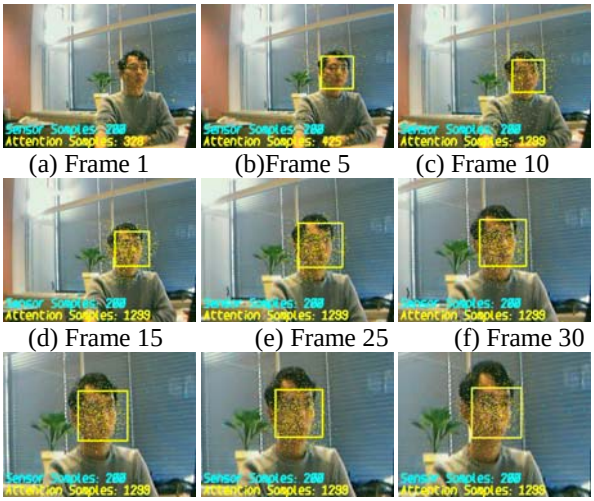
[14]http://diva.comp.nus.edu.sg/yanwq/PCM/index.htm



(a) Frame 1 (b) Frame 35 (c) Frame 49

(d) Frame 67 (e) Frame 80 (f) Frame 49

Figure. 2. Automatic Panning. (a)-(c) left pan (d)-(f) right pan



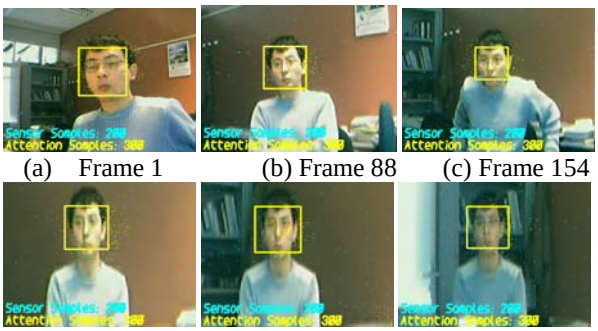
(a) Frame 1 (b)Frame 5 (c) Frame 10

(d) Frame 15 (e) Frame 25 (f) Frame 30

(g) Frame 35 (h) Frame 45 (i)Frame 49

(a) SS arouses AS according to motion and skin color (b) A face is detected and marked by a rectangle. (c)-(g) zooming and panning until ((g)-(i)) the face is properly framed.

Figure. 3. Automatic Zooming



(a) Frame 1 (b) Frame 88 (c) Frame 154

(d) Frame 222 (e) Frame 237 (f) Frame 265

(a)-(b) zoom out when the object moves.(b)-(c) pan right (d)-(e) pan left (e)-(f) pan left and zoom out

Figure. 4. Automatic Panning and Zooming.