Efficient Control of PTZ Cameras in Automated Video Surveillance Systems

Musab S. Al-Hadrusi Nabil J. Sarhan

Media Research Lab, Dept. of Electrical and Computer Engineering
Wayne State University Detroit, Michigan 48202
Email: {hadrusi, nabil}@wayne.edu

***Abstract*—This paper deals with the camera control problem in automated video surveillance. We develop a solution that seeks to optimize the overall subject recognition probability by controlling the pan, tilt, and zoom of various deployed Pan/Tilt/Zoom (PTZ) cameras. Since the number of subjects is usually much larger than the number of video cameras, the problem to be addressed is how to assign subjects to these cameras. This control of cameras is based on the direction of the subject's movement and its location, distances from the cameras, occlusion, overall recognition probability so far, and the expected time to leave the site, as well as the movements of cameras and their capabilities and limitations. The developed solution works with realistic 3D environments and not just 2D scenes. We analyze the effectiveness of the proposed solution through extensive simulation.**

***Keywords*-Automated Video Surveillance; Camera Network; Camera Scheduling; Subject Recognition.**

I. Introduction

Recent advances in networking, video sensors, and networked video cameras have enabled the development of large scale surveillance systems. In contrast with traditional video surveillance, automated video surveillance (AVS) serves as an efficient approach for the realtime detection of suspicious activities and events by employing computer vision algorithms.

High-end video surveillance systems may employ a number of Pan/Tilt/Zoom (PTZ) cameras, which offer a high degree of flexibility. Since the number of subjects is usually much larger than the number of video cameras, the problem to be addressed is how to assign subjects to these cameras. Numerous studies have dealt with the control problem of PTZ cameras [1], [2], [3] (and references within). Unfortunately, these studies had many limiting assumptions and considered small subsets of the factors. The overwhelming majority of these studies analyzed metrics, such as the captured resolution, and the number of times a subject is viewed, but in AVS systems, the main objective should be maximizing the overall threat detection/recognition probability.

This paper addresses the camera control problem in AVS. We develop a solution that seeks to optimize the overall subject recognition probability by controlling the pan, tilt, and zoom of various deployed PTZ cameras, based on the characteristics of the subjects in the surveillance area. This paper focuses primarily on face recognition. This control of cameras is based on the direction of the subject's movement and its location, distances from the cameras, occlusion, overall recognition probability so far, and the expected time to leave the site, as well as the movements of cameras and their capabilities and limitations. Camera movement is broadly defined here as the time required to reach the assigned pan, tilt, zoom, and focus. Our overall solution includes three new alternative schemes for scheduling the cameras: *Brute-Force Grouping* (BFG), *Grid-Based Grouping* (GBG), and *Elevator-Based Planning* (EBP).

This work was supported in part by U.S. NSF grant CNS-0834537.

The rest of this paper is organized as follows. Section II discusses the background information and related work. Section III presents the proposed solution. Section IV discusses the performance evaluation methodology. Finally, Section V presents the main results.

II. Background and Related Work

Many face recognition algorithms [4] (and references within) are proposed to recognize faces in controlled environments. Many solutions have been investigated to deal with specific problems in face recognition, such as illumination, pose, resolution, occlusion, and expression. Most of these solutions require one or more of the following: (i) pre-enhancement using image processing algorithms, (ii) training on different images for the same subject, (iii) having more than one entry on the dataset for the same subject, and (iv) extensive execution time. Deploying these algorithms in video surveillance systems is still facing performance problems.

Recent research has opened wider doors for PTZ cameras by utilizing their potential power and flexibility [5]. PTZ cameras have appealing characteristics, such as (i) large *fields-of-regard* (FoR), (ii) enhanced views, and (iii) multiresolution views. FoR can be defined as the set of all possible combinations of *field-of-view* (FoV) that can be reached using different camera's settings, where FoV is the extent of the observable view as seen by the camera using one specific settings (pan, tilt, and zoom). PTZ cameras are currently controlled either manually by using special joysticks and keyboards or automatically by proceeding in pre-programmed tours [6].

In video surveillance systems, a limited number of cameras is deployed. A problem appears when the scene has more subjects than the available cameras. Subject grouping is one approach used to tackle the scalability problem. In this method, the system groups more than one subject in the same frame and assigns them to one PTZ camera. The work in [1] used a master-slave system with a wide-angle camera employed to collect scene information and pass it to the main server. The server scans the data and runs a *Lattice-Based Grouping* (LBG) algorithm to generate a set of frames that represent the maximum "Objective Satisfaction". That paper assumed that all PTZ cameras have exactly the same FoR (as they are all located in the same place and can cover the scene

978-0-7695-4875-3/12 $26.00 © 2012 IEEE
DOI 10.1109/ISM.2012.72

equally likely). This assumption oversimplifies the problem and substantially reduces the search domain.

Paper [2] is another work that considered subject grouping. The work modeled the surveillance system as a bipartite graph. The subject camera assignment was obtained by applying a maximum matching algorithm. Maximum subject coverage was sought and face recognition was not a target.

III. Proposed Solution

A. System Overview

The proposed AVS system employs wide-angle cameras, PTZ cameras, and a processing architecture. Wide-angle cameras are used to observe the site and send information to a processing system, which analyzes the scenes and then controls the PTZ cameras. The PTZ cameras capture higher quality frames of the subjects.

The processing architecture performs the following tasks.

- It analyzes the images of the wide-angle cameras to determine the locations and types of subjects and their speeds and directions of movement. This task utilizes the widely researched pedestrian detection algorithms [7] for determining the various attributes of subjects.
- It predicts the locations of the subjects when the cameras will start recording.
- It runs a camera planning and assignment algorithm and controls the PTZ cameras. This task may include generating sets of possible frames according to the states of the PTZ camera and their capabilities and limitations.
- It runs the computer vision algorithms, such as face recognition.

The system operates into two alternating periods: a *pre-recording period* of T_p seconds and a *recording period* of T_R seconds, as in [1]. In the first period, the processing system reads the state and performs some prediction and algorithmic calculations. The PTZ camera frame assignment is also done in this period. In the second period, the processing system starts receiving high quality frames from the PTZ cameras and considers them for further processing. During the recording period, the cameras track the subjects.

The system employs a *Watch List*, which includes an image database of subjects that are deemed dangerous to the surveillance site.

B. Overall Solution Overview

The proposed solution targets the camera scheduling and assignment problem in 3D environments. Its main objective is to optimize the overall subject recognition probability by controlling the pan, tilt, and zoom of various deployed PTZ cameras, based on the characteristics of the subjects in the surveillance area. This control of cameras is based on the direction of the subject's movement and its location, distance from the cameras, occlusion, overall recognition probability so far, and the expected time to leave the site, as well as the movements of cameras and their capabilities and limitations.

As shown in Figure 1, the solution consists of four phases.

- *Frame Generation*– In this phase, the processing system utilizes the information provided by the fixed wide-angle cameras to detect all subjects captured by these cameras and determine their attributes, including, location, speed, and direction of movement. The attributes can be determined using one of the widely researched pedestrian detection algorithms [7]. Subsequently, the processing system predicts the locations of the subjects at the future time when the PTZ cameras enter the recording and tracking phase. The processing system then uses the prediction results to generate the set of all possible frames that can be captured by each PTZ camera through the examination of the different combinations of the camera's settings. This last step utilizes the capabilities and limitations of the PTZ cameras. Note that each PTZ camera can capture a different frame at each camera setting. The frame here is basically the *projection* of the view as seen by the camera at a specific FoV. By examining different settings, a set of frames encompassing the entire FoR will be generated. Combining all possible frames from all cameras produces the entire frame domain that will need to be analyzed in later phases. In this stage, a projection triangulation is used to estimate the 2D frame characteristics from the 3D site. Rotation and translation matrices are used to map the 3D coordinates to each camera coordinates [8].
- *Frame Filtration*– After generating the sets of all possible frames, the processing system filters these frames by eliminating the frames that do not capture any subject in the surveillance site and by selecting only the best camera frame among the sets of frames that capture exactly the same set of subjects. The selected frame is the one that achieves the maximum aggregate recognition probability of all subjects.
- *PTZ Camera Scheduling*– In this phase, the processing system carries out camera scheduling and frame assignment. The proposed schemes for this task are detailed later in this section.
- *Recording and Tracking*– In the this phase, the PTZ cameras apply the settings of camera planning and scheduling to capture the target subjects and start recording and tracking these subjects.

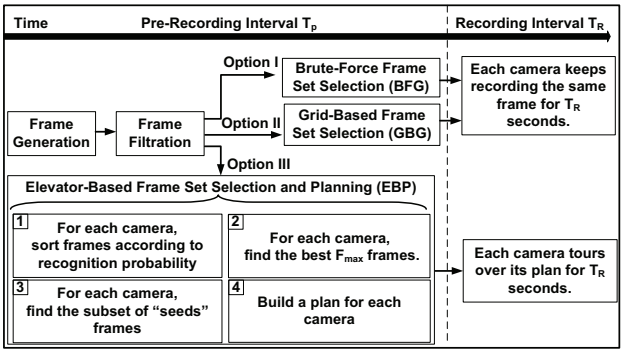


Fig. 1. Illustration of the Proposed Solution

C. Formulation of the Objective Function

Since the main objective of the proposed solution is to optimize the overall recognition probability, we need to formulate this probability as an objective function. In this paper, we focus primarily on face recognition.

We form a utility function $\psi(j)$ to measure the total recognition probability of all subjects included in camera frame j:

$$\psi(j) = \sum_{i=1}^{N_s} \gamma(i,j), \tag{1}$$

where N_s is number of subjects in frame j, and $\gamma(i,j)$ is the recognition probability of subject i captured in camera frame j. This equation sums the recognition probability value contributed by each covered subject i in frame j. The recognition probability $\gamma(i,j)$ of subject i captured in camera frame j can be formulated as follows:

$$\begin{aligned}\gamma(i,j) =& W(i) \times R_{pose}(\theta_{tij}) \times R_{pose}(\theta_{pij}) \\ & \times R_{zoom}(z_{ij}) \times u_{cov}(i,j) \times u_{occ}(i,j),\end{aligned} \tag{2}$$

where $W(i)$ is a weighting factor, $u_{cov}(i,j)$ and $u_{occ}(i,j)$ are unit functions indicating the subject coverage and occlusion status, respectively.

Based on the empirical data from [9], we derive the following model for the relationship between the recognition probability R and the zoom z:

$$R_{zoom}(z_{ij}) = a_1 \times e^{(b_1 \times z_{ij})}, \tag{3}$$

where z_{ij} is the zoom adopted by camera j to capture subject i with a resolution of 60 inter-ocular pixels, and a_1 and b_1 are constants.

Based on the empirical data in [10], which were verified by [3], and by limiting the poses to ± 25°pan or tilt boundaries, we derive the following model for the impact of pose on recognition probability:

$$R_{pose}(\theta_{ij}) = \begin{cases} e^{(-\frac{\theta_{ij}}{b_2})^2} & |\theta_{ij}| \leq 25 \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where θ_{ij} is the angle between subject i (face center) and the optical axis of camera j, b_2 is constant. The same function is used for pan and tilt angles.

D. PTZ Camera Scheduling

Camera scheduling is the core of our proposed solution. We present three schemes for camera scheduling: *Brute Force Grouping* (BFG), *Grid-Based Grouping* (GBG), and *Elevator-Based Planning* (EBP). BFG scans the filtered frame list from all C PTZ cameras and chooses a set of C frames, including one frame from each camera. This set must achieve the highest aggregate recognition probability of all subjects. This probability is measured by considering the recognition probability of each subject only once at its maximum occurrence in the set regardless of its number of occurrences. GBG is an adaptation of the algorithm in [1] (previously discussed in Section II) with several enhancements. The work in [1] assumed that all cameras capture the same set of frames in the same fashion. In particular, a frame can be captured by any PTZ camera located at any position in the scene, and it will have the same "satisfaction value". It also assumed a frame domain derived from 2D scenes and not realistic 3D environments. We enhance the algorithm in the following aspects: (i) dealing with cameras having different FoRs, (ii) generating frames according to cameras capabilities and not site area dimensions, (iii) grouping based on 3D scene by considering each camera's FoV and the locations of subjects, and (iv) assessing the overlap among frames in 3D scenes. EBP refines the filtered frames and uses them to build detailed plans for each camera. The scheme captures the current settings of each camera, and its moving, zooming, and focusing speeds. In contrast with BFG and GBG, EBP allows each camera to view a different set of subjects at different times during one recording period rather than repeatedly tracking the same subjects.

Let us now discuss the proposed EBP scheme in more detail. As shown in Figure 1, this scheme proceeds into two main stages: (i) generating plan seeds for each PTZ camera and (ii) building the plan for each PTZ camera from these seeds. The generation of the plan seeds encompasses Steps 1 to 4 in Figure 1. A simplified algorithm for seed generation is introduced. The algorithm sorts frames in each camera list according to their overall recognition probabilities ψ and picks only the best F_{max} frames for each camera to reduce the implementation complexity of subsequent steps. F_{max} can be selected such that the maximum number of frames for each camera is considered within an implementation overhead of T_p seconds for camera scheduling. For each camera, out of the F_{max} frames, the algorithm selects a subset of $seeds$ frames (where $seeds < F_{max}$) such that this subset achieves the maximum aggregate recognition probability ($MARP$) of all subjects. The aggregate recognition probability of all subjects is measured as follows by considering the recognition probability of each subject only once at its maximum occurrence in the subset regardless of its number of occurrences:

$$MARP = \sum_{i}^{S_N} \sum_{j}^{Seeds} (\bar{\gamma}(i,j) \times u(i)_{max}), \tag{5}$$

where $\bar{\gamma}(i,j)$ is the unweighted version of the $\gamma(i,j)$ defined in Equation 2, $seeds$ is the number of frames in the subset, S_N is the number of subjects in the subset, and $u(i)_{max}$ for subject i is 1 only when it has the highest recognition probability ($\bar{\gamma}$) and 0 otherwise.

IV. Performance Evaluation Methodology

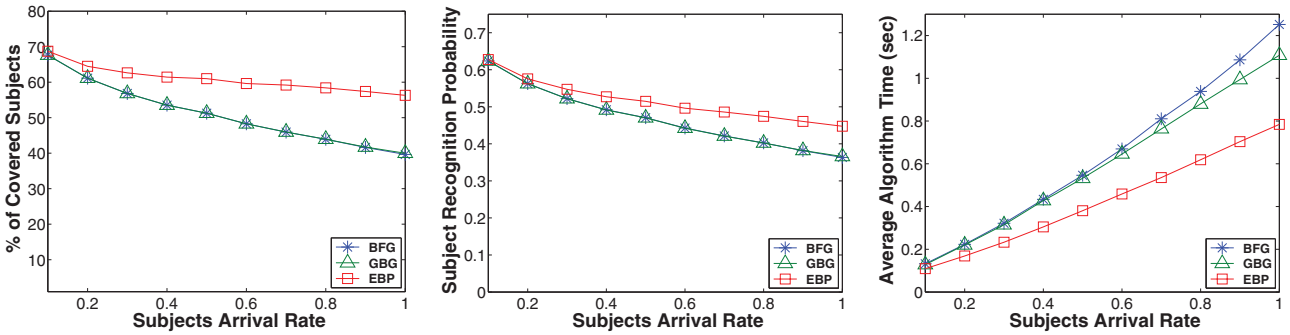
We developed in C++ a simulator, called *AutoSurvSim*, for an AVS system supporting the proposed solution for controlling the PTZ cameras. Without loss of generality, we assume a surveillance site with a rectangular area. Subjects arrive randomly from any of the four directions and enter the site with a random direction and a random speed. Poisson distribution is used to model subject arrivals, while a truncated Gaussian distribution is used for the speed, and a truncated uniform distribution is used for the direction [1]. The PTZ cameras are located at the perimeter of the site with different (x,y,z) locations. Moreover, a detailed characterization for each camera is implemented, including (Pan-Tilt-Zoom) speeds with maximum and minimum limits and a step size, refocusing speed, and sensor dimensions. The results are collected by running the simulator on a system with 32-bit 2.4 GHz dual-core CPU and 2.5 GB RAM. The simulations are run for the time required to process 50000 subjects.

The performance metrics are the average subject recognition probability, the percentage of subjects covered/captured at least one, and the algorithms computation time.

Table I summarizes the parameters used in the evaluation.

V. Result Presentation and Analysis

In this section, we compare the effectiveness of the three proposed schemes under different scenarios. Only the main results are shown.



(a) Percentage of Covered Subjects (b) Average Subject Recognition Prob. (c) Time Complexity

Fig. 2. Effect of Subject Arrival Rate

TABLE I
SUMMARY OF WORKLOAD AND SYSTEM CHARACTERISTICS

Parameter	**Model/Value(s)**
Site Area	Variable, Default = $80 \times 60\ m^2$
Request Arrival	Poisson Process
Request Arrival Rate	Variable [0.1-1.0], Default = 0.5 Req/second
Request Speed	Truncated Normal Distribution [0.5-2.5]
Request Speed rate	0.5 m/sec
Request Direction	Truncated Uniform Distribution [-40°40°] degree with $\perp$ to entrance side
PTZ Cameras	Variable [2-8], Default = 4
PTZ Pan	Variable [0-180] degrees
PTZ Tilt	Variable [0-90] degrees
PTZ Zoom	Variable [1-50] levels
Sensor Resolution	$[640 \times 480]$ pixels
Inter-Ocular Distance	60 pixels
Focus Time	0.5 second
T_p, T_R	Variable [1-10] second, Default = 5
R_{thresh}	0.9
F_{max}	15
Eq 3 constants a_1,b_1	0.693 , -0.0154
Eq 4 constant b_2	48.31

The effectiveness of various scheduling schemes in dealing with scalable workloads is shown in Figure 2. The figure compares the three schemes under different subject arrival rates. As expected, the three metrics become worse as the arrival rate is increased because more subjects leave the surveillance site without being captured, less time is spent on the captured subjects, and the search space becomes larger, respectively. These results demonstrate that EBP handles higher arrival rates much better than BFG and GBG, and that the gaps in performance and time complexity between EBP and the other two schemes become wider as the arrival rates increases. Note that EBP captures more subjects during a recording period, whereas GBG and BFG need to search the surveillance site one more time to capture new subjects, causing them to degrade much faster than EBP.

VI. CONCLUSIONS

We have analyzed the effectiveness of the proposed solution through extensive simulation. The main results can be summarized as follows.

- The proposed EBP scheduling scheme achieves the best recognition probability.
- EBP achieves the best scalability when increasing the subject arrival rate and recognizes the biggest number of subjects.
- EBP has the lowest time complexity.

REFERENCES

[1] Y. Xu and D. Song, "Systems and algorithms for autonomous and scalable crowd surveillance using robotic ptz cameras assisted by a wide-angle camera," *Auton. Robots*, vol. 29, no. 1, pp. 53–66, 2010.

[2] H. El-Alfy, D. Jacobs, L. Davis, and L. Davis, "Assigning cameras to subjects in video surveillance systems." in *Proc. of International Conference on Robotics and Automation (ICRA)*, 2009, pp. 837–843.

[3] S. L. N. Krahnstoever, T. Yu and K. Patwardhan, "Collaborative control of active cameras in large-scale surveillance," in *Proc. of NetworksPrinciples and Applications in Multi-Camera*, H. Aghajan and A. Cavallaro, Eds. Elsevier, 2009, pp. 165–188.

[4] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, pp. 399–458, December 2003.

[5] R. Cucchiara, "Multimedia surveillance systems," in *Proc. of the third ACM international workshop on Video surveillance and sensor networks*, ser. VSSN '05, 2005, pp. 3–10.

[6] F. Nilsson, *Intelligent Network Video: Understanding Modern Video Surveillance Systems*. CRC Press, 2009.

[7] B. S. P. Dollar, C. Wojek and P. Perona, "Pedestrian detection: A benchmark," in *Proc. of Computer Vision and Pattern Recognition Conference (CVPR)*, June 2009.

[8] J. P. D. A. Forsyth, *Computer Vision: A Modern Approach*, 1st ed. Prentice Hall, 2002.

[9] Y. Yao, B. R. Abidi, N. D. Kalka, N. A. Schmid, and M. A. Abidi, "Improving long range and high magnification face recognition: Database acquisition, evaluation, and enhancement," *Computer Vision and Image Understanding*, vol. 111, pp. 111–125, August 2008.

[10] R. Gross, S. Baker, I. Matthews, and T. Kanade, "Face recognition across pose and illumination," in *in handbook of face recognition*. Springer-Verlag, 2004.