Operator Attention Based Video Surveillance

Ulas Vural and Yusuf Sinan Akgul

GIT Vision Lab, http://vision.gyte.edu.tr/,
Department of Computer Engineering, Gebze Institute of Technology,
41400, Kocaeli, Turkey

Abstract

We introduce a surveillance video tractability adjustment system that employs a dynamic operator attention model. The tractability of the surveillance video is adjusted according to the instantaneous attention level of the operator (Fig.1). Our system has two novel major parts: (i) dynamic measurement of the operator attention levels and (ii) an online video tractability adjustment system that employs measured attention levels. We estimate the attention levels of the operators by measuring the blink rates and the saccade information of the eyes. Using well known theories from human psychology, we estimate the amount of video tractability that the operator can easily handle. We tested the proposed system on volunteer operators using both synthetic and real surveillance data. The results are very promising and we plan to extend the system with multiple operators and multiple video streams. Supplementary material is available at the project website.

1. Introduction

Digital surveillance systems have become an important part of our daily lives[19]. The lowering costs of surveillance systems makes them even more popular which would produce a very large amount of visual data to be processed. However, a fully automatic analysis of the surveillance videos is not possible with today's available technology [17]. As a result, manual processing of the resulting visual data becomes inescapable and the human labour emerges as a dominating part of the surveillance system costs [9]. In addition, human errors have to be factored in for such systems [34, 35]. Video summarization methods have been used as a partial solution for lowering the costs of human labour for surveillance systems. Linear summary systems [20] attempt to drop video frames with less activity and they cannot compress video without time-lapse effects which is error prone [18]. Many of these systems produce long summaries to keep clear of the visual errors. One popular and relatively new solution for compact summarization is non-linear video synopsis [2]. These methods can bring actions from different time intervals to the same frame and they can produce much shorter video summaries without losing any action.

One major problem with the plain linear and non-linear video summary methods is their static human attention models. Psychological studies show that multiple object tracking capabilities of human beings depend on the number of objects [3, 8], their speeds [4], and spatial configurations [28], individual operator differences [15, 35], and operator attention levels [31, 36]. There are a few static human attention models for selecting best key-frame sequences[23, 27]. These methods discard unattractive frames according to their human attention model. However, it is very difficult for these systems to adjust for different operators because the parameters for the static attention model are fixed during the system design. In addition, static attention models assume that humans have constant and continuous attention levels. Holding sustained attention at high levels requires too much energy so operator performance degrades after several minutes. Forcing to stay at high attention levels causes another unwanted situation called stress. An operator attention model should include operator's initial condition and workload [31]. Operator's attention can also be affected from other stimuli in the working environment [24].

This paper defines a new concept called *video tractability* which is the number of moving objects, their directions, speeds, and relative spatial positions. We argue that the tractability of the manually viewed surveillance videos should be adjusted according to the instantaneous attention level of the surveillance operator (Fig.1). Our system has two novel major parts: (i) dynamic measurement of the operator attention levels and (ii) an online video tractability adjustment system. We estimate the attention levels of the operators by collecting position and velocities of eye-gaze points, pupil diameter, blink rate, and saccade information from the operator. Using well known theories from human psychology, we estimate the amount of video tractability

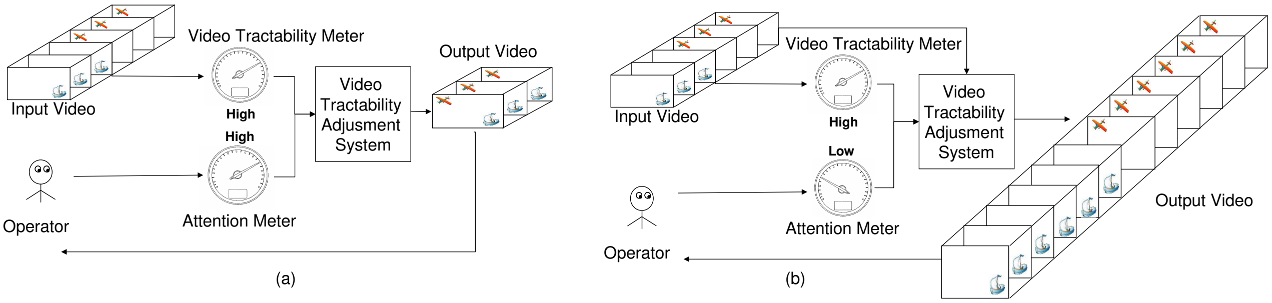

Figure 1. Two cases for video tractability adjustment: (a) An operator with high attention level can track denser videos (b) while an operator with low attention can track sparser videos accurately.

that the operator can handle. This information is used to synthesize a surveillance video stream that the operator can easily handle.

There are several advantages and novel contributions of the proposed system. First, our system puts the surveillance operators in an active feedback loop, which solves many problems associated with systems with static operator attention models. The human operator is the critical element in a surveillance system and including this element in the overall system loop is a big advantage. Our system is an example of *adaptive automation based systems* which are very successful for preventing operators from fatigue and stress [7]. These types of systems use human resources at maximum limits while preserving system's reliability[25]. Second, we introduce a more general concept of video tractability adjustment instead of video summarization. We argue that if the original surveillance video contains too many moving objects and actions, it should be re-synthesized so that the operator can monitor the video without getting overloaded in order not to miss any moving objects. The resulting video might get longer than the original video with tractability adjustment (Fig.1-b) . If the input video is too sparse in terms of moving objects and their spatial configurations, the re-synthesized video will be shorter (Fig.1-a). In contrast, the classical video summarization methods always try to produce shorter videos regardless of the operator workloads, which will cause human errors and stress. Third, we introduce a novel non-linear video tractability adjustment method that can work on the fly. To achieve high frame rates without sacrificing system performance, we use a meta-heuristic called iterated local search. Given the desired video tractability level, our system can bring objects from different video frames into the same output video frame or it can separate objects from the same frame into different frames without dropping any input surveillance video frames.

In a similar study, [33] tracks eye-gaze movements of operators and determines monitored or overlooked parts of a screen. It produces non-linear summaries of monitored/ overlooked parts of input video stream. The system does not extract any knowledge about operator's mental situation, workload, stress, and fatigue. [6] uses operator eye gaze positions to decide which camera view is most important among many cameras. There also exists a parallel variation of [33] that works on high resolution surveillance videos [32] but none of them attempt to adjust video tractability.

We performed several experiments on volunteers using real and synthetic videos. A synthetic video based multiple object tracking experiment is done on several different groups. These experiments show us how human attention metrics vary with time and workload. We also measure task success rates on different workloads.

The rest of the paper organized as follows: We describe psychological background and our action model in Section 2. We explain our video tractability adjustment framework in Section 3. In Section 4 we give both visual results of our tractability adjustment system and quantitative results of psychological experiments.

2. Human Psychology on Multiple Object Tracking

Many manual surveillance monitoring tasks are based on tracking multiple independent objects which forces us to understand the human psychology of object tracking. Luckily, there is a large amount of psychology literature on the subject[29, 12, 22].

Psychology is interested with multiple object tracking in terms of attention, tracking limits and strategies, vigilance, and workload. Humans do not perform perfectly on these tasks because their visual and the mental capabilities are limited. In many studies it is shown that a human can track only a small number of independently moving objects at the same time[26, 14] (Fig.2(a)). Some studies argue that divided attention is used for multiple object tracking and according to these studies humans can track 4 to 6 dependent objects [8, 4]. On the other hand, humans can only track

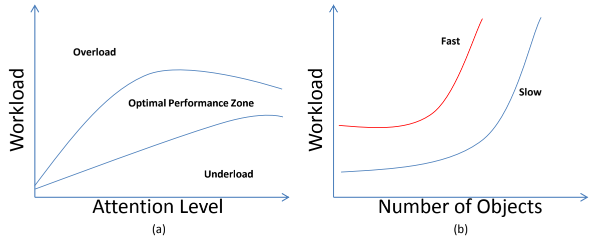

Figure 2. (a) Humans have an optimal performance zone where they are both reliable and effective. (b) Humans can track more objects if the objects are slow.

one object with sustained visual attention. Multiple moving objects with similar trajectories form a group called virtual object. Humans can track a larger number of objects if they are all in virtual objects [28] by looking at the centre of the virtual objects[11]. There are also constraints other than the number of objects. It is shown that a larger number of objects can be tracked if they are slow (Fig.2(b)). Eyes make smooth pursuit for tracking slow objects and accuracy is increased [4]. For high speed and spatially distributed objects, saccadic eye movements are required, which demand more mental resources [16]. Tracking is also hard for independently moving objects that collide and occlude each other [13].

Attention levels and periods are subjective and depend on expertise [3], emotional situation [15], and perception of workload [35]. An expert user generally can track a larger number of objects than a novice user. Experts can also keep their attention high for longer periods of time. If an operator is underloaded he or she gets bored [31]. On the other hand, attention levels of overloaded operators will drop in a short time [36].

Our operator attention level measurement method uses the above human psychology theories. We estimate the operator attention levels by monitoring the operator eye-gaze metrics. The output videos are re-synthesized such that the operators are always kept at optimal performance zone (Fig.2(a)).

3. Video Tractability Adjustment System

A surveillance video tractability adjustment system modifies the speeds, the directions, and the spatio-temporal configurations of moving targets according to the operator attention level and the system places the modified targets into the best tractable places. We define the video tractability adjustment problem as a kind of facility location problem which is finding the optimum locations of given objects in a definite search space[21]. We view each action a in the input video buffer V_{in} as a facility and extract the set of all candidate locations L for a given size of output video V_{out}. When the input video buffer V_{in} is ready, a set of actions is obtained by running a blob based object tracking and segmentation algorithm on the buffered video. These segmented actions are appended to a queue of waiting actions Q. According to the attention criteria, a subset of actions A is obtained randomly from the Q. Video tractability adjustment system then assigns a location label l from L to each action a in the action set A. A configuration C represents the labelling assignments of all actions in A. The video tractability adjustment problem is to find an optimal configuration C^* which has minimum total energy E_T. We solve an approximation of this problem using the iterated local search approach [30].

3.1. Problem Definition

Facility location is an NP-Hard problem. In 3D x-y-t volume of a given output video V_{out}, there are too many pixel locations where a modified action a can be placed. Therefore, limiting the search space is essential for a fast surveillance video tractability system. We prefer to work in 2D domain which represents the x-t trajectories of moving targets. The 2D trajectory images are used in surveillance video systems for efficiency [33]. Our tractability adjustment system first extracts motion trajectories of moving objects from the surveillance videos in $x - y - t$ space (3D space time volume). For each frame and each action in the input video V_{in}, bounding boxes of the actions and their centroids are computed. The y components of the centroids are then dropped to move to the x-t space (2D space-time). In the video reconstruction phase, the bounding boxes in the x-y domain is used for the given object and time. (See supplementary material-1 for dimension reduction.)

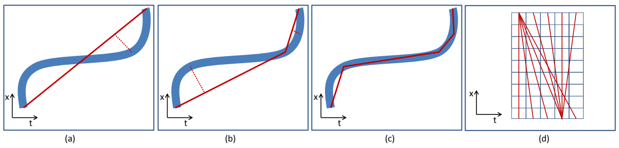

Figure 3. Line Segment Model. (a) First iteration of Douglas-Peucker line approximation algorithm. Dashed line represent the maximum distance between the trajectory curve and the fitted line.(b,c) Second and third iterations of the algorithm. (d) 2D space-time domain of output video V_{out} and possible locations.

We prefer to use a line segment based model for representing motion trajectories in the $x - t$ projection space. A line model ℓ represents $x = \ell.m\ t\ +\ \ell.b$ in 2D space-time domain where $\ell.m$ is the slope and $\ell.b$ is the phase of the line l. Line segment model is an effective model for manipulating action properties. Speed of an action a can be adjusted by changing the slope magnitude of its line model or its direction can be reversed by changing the sign of the slope. We obtain line segment models of actions from 2D space-time trajectories. These action trajectories generally form a higher order curve that cannot be approximated by a single line segment model. A set of line segments $a.segm$ are used for each action a. The iterative curve approximation algorithm of Douglas-Peucker is used[10] to fit line segments to these trajectories (Fig. 3). For each action a, the proposed method checks the signs of slopes for each line segment in $a.segm$ to determine the direction of target's movement. A direction change in an action trajectory is determined if there are two consecutive line segment parts with different signs of slopes. The line model divides action into parts so each action is forced to have a single direction.

The video tractability adjustment system also uses the line segment model for representing candidate locations L. Our method extracts candidate location segments from the 2D space-time as shown in (Fig. 3(d)), which includes only a subset of possible line segments for clarity. Each red coloured line segment is a location label.

The line models of actions and the candidate line labels are used in an energy formulation for finding optimum configuration C^*. Total energy E_T of tractability adjustment system is a linear composition of the video tractability energy E_{tract}, the video similarity energy E_{sim} and the energy of operator comfort E_{comf}. Total energy of a given configuration C, set of actions A, and operator attention level Att is

$$\begin{aligned} E_T(A,C,Att) &= \alpha_1 E_{tract}(C) \qquad (1)\\ &+ \alpha_2 E_{sim}(A,C) + \alpha_3 E_{comf}(C,Att). \end{aligned}$$

where $a, b \in A$ and $C[a]$ is a location label $l \in L$.

In facility location problem, a facility searches for a minimum cost location that maximizes its profit. For video tractability adjustment problem, action a should be placed on the best viewable location according to the operator attention while preserving its initial action characteristics. The similarity term E_{sim} in the total energy E_T tries to keep the configuration C as similar as possible to the action's original speed, direction, and chronological order. On the other hand, the tractability energy E_{track} forces actions in A to the new location labels for better tractability. The tractability term only deals with configuration properties of a given video and it is free from the operator's attention level. In facility location problem, facilities can also require outputs of other facilities, so compact establishment of related facilities is crucial for minimizing run-time costs. The tractability term also evaluates the interactions between the location labels in the configuration C. The total energy E_T includes E_{comf} which regularizes the tractability of each frame according to the attention level.

3.2. Energy Functions

The total energy E_T of the tractability adjustment system has three main parts: E_{tract}, E_{sim}, and E_{conf}. These terms are formulated as a weighted sum of several sub energy terms and in this section we describe them in detail.

We define a linear model for evaluating the tractability level. The model is both based on psychological theories and empirical results of our experiments. The tractability energy E_{tract} is defined in terms of speed, direction, collisions and object density. Tractability level E_{tract} of a given configuration C is

$$\begin{aligned} E_{tract}(C) &= \beta_1 E_{den}(C) + \beta_2 E_{vel}(C)\\ &+ \beta_3 E_{dir}(C) + \beta_4 E_{col}(C). \qquad (2) \end{aligned}$$

Number of moving targets in a video frame is an important tractability factor. Tracking accuracy of human operators decreases on crowded scenes, so a tractable video should consist of fewer number of objects per frame. The video density term E_{den} is estimated from the configuration C as an average number of moving targets per frame.

The tractability adjustment system evaluates the velocities of moving targets and forces actions to move slowly. The velocity term E_{vel} is the average speed of moving targets (Eq. 3).

$$E_{vel}(C) = \sum_i abs(C[i].m)/sizeof(C)\ ,\ \ \forall C[i] \in C \qquad (3)$$

where $C[i]$ is a label $l \in L$.

Psychological studies on multiple object tracking show that humans can track more objects if the objects move in similar directions. The term E_{dir} evaluates the direction differences of moving objects in a frame. We compute it as the number of frames which include at least two moving targets with different directions. The direction of a moving object can simply be determined by checking the sign of its line model's slope m.

Multiple object tracking capabilities of human operators decrease when objects collide. To prevent collisions, we apply a penalty to intersecting line models.

$$E_{col}(k,l) = M[k][l] \qquad (4)$$

where k and l are two distinct labels in L. We pre-compute the intersections of all line pairs in L for efficiency and hold them in a collision matrix M. (Please, see supplementary-3 for the effects of the energy terms.)

The proposed system measures the attention level of the operator by using two different eye metrics. Our psychological experiments show that operators success rate decreases when their blink rates or saccade frequencies increase. Attention level Att is increased if operator gaze metrics are regular and decreases when he or she is overloaded or tired (Eq.5).

$$Att = Att + \begin{cases} -1 & if\ (saccade \geq T_1\ or\ blink \geq T_2), \\ 1 & otherwise. \end{cases} \tag{5}$$

We form an array H with the size of the video. The elements of H correspond to the frames of V_{out} and their values represent the number of actions on those frames. Each location label l in configuration C increases the array elements corresponding to its time interval.

$$E_{comf}(C, Att) = \sum_{f=0} |(H[f] - Att)| \tag{6}$$

The video tractability adjustment system uses the similarity energy term E_{sim} to preserve the original action characteristics. The similarity energy term consists of three terms.

$$\begin{aligned} E_{sim}(A, C) \quad &= \quad \gamma_1 \sum_a E_{dirSim}(A[a], C[a]) \\ &+ \quad \gamma_2 \sum_a E_{velSim}(A[a], C[a]) \\ &+ \quad \gamma_3 \sum_a E_{chronoSim}(A[a], C[a]). \end{aligned} \tag{7}$$

Direction of an action can be manipulated to form a virtual group of objects but this manipulation can produce unusual results like a human walking backwards. A cost is applied for preserving original moving direction. A change on direction is detected by using Eq. 8.

$$E_{dir}(a, l) = \begin{cases} 1 & if\ a.segm[0].m\ l.m \leq 0, \\ 0 & otherwise, \end{cases} \tag{8}$$

where $a.segm[0].m$ is the slope of the first line segment of a and $l.m$ is the slope of candidate location label l.

The tractability system adjusts the speeds of actions for compact video outputs with less collisions or for accurate tracking. Accelerating the actions makes a more compact video possible for an operator with high attention level. A more tractable output can also be obtained by slowing down objects for an operator with low attention level. This term tries to limit to change in the velocity so a fast target is accelerated before a slow one or a slow object will be chosen for deceleration.

Chronological orders of actions change in the video tractability adjustment system but the system tries to preserve their order while synthesizing output videos. This is done by assigning a waiting time priority to the actions in the waiting queue Q

3.3. Optimization

Video tractability adjustment for a given attention level is a complex optimization problem. A minimum energy configuration C^* should be computed efficiently.

$$C^* = \underset{\boldsymbol{C}}{\operatorname{argmin}}\, E_T(A, C, Att) \tag{9}$$

We use iterated local search (ILS) for finding minimum energy configuration. ILS is a simple yet powerful metaheuristic and it is fast enough for video tractability adjustment. It has three important parts:

1-Initialization: The initial configuration can be a random location label l assignment to each action a in A. We use the output of the greedy method that only optimizes similarity energy E_{sim} as the initial configuration.

2-Local Search: In the local search phase, each action a is selected in order as an active action. Activation order of actions is determined randomly once in the beginning of each local search phase. Active action a in configuration C searches for a better location l while other actions preserve their locations.

3-Modification: In the modification step, a number of actions in configuration C is selected randomly. These selected elements of configuration C change their location labels l randomly. This step can increase the energy level of configuration C and can be considered as a diversification step.

In our video tractability adjustment system, local search and modification steps are iterated for several times. We randomly group actions in set of all actions A in such a way that all groups have Att number of actions.

The final synthesized output video is formed from the 2D projection space that produces C^*.

4. Experiments

We have a number of volunteers whose eye-gaze metrics are recorded by using the head-mounted eye-gaze tracker of Arrington Research [5]. Our first experiment is a psychological experiment for extracting the relationship between eye-gaze metrics and human attention level. A multiple object tracking based task is prepared. The task also measures the short-term memory capabilities of volunteers which is directly related with the attention level. Human-like moving targets have three different labels on their chests and they move on a plain background (Fig. 5(a)). Each object has different speeds and they can change their directions or they can stop for a while. They also produce alerts by hiding their labels for a while as shown in Fig. 5(b) . Operators should catch these random alerts on time. Operator should also remember the true label of the object and press its label key on the keyboard, which is called *catch.* If the operator saw the alert but does not remember its label, he or she

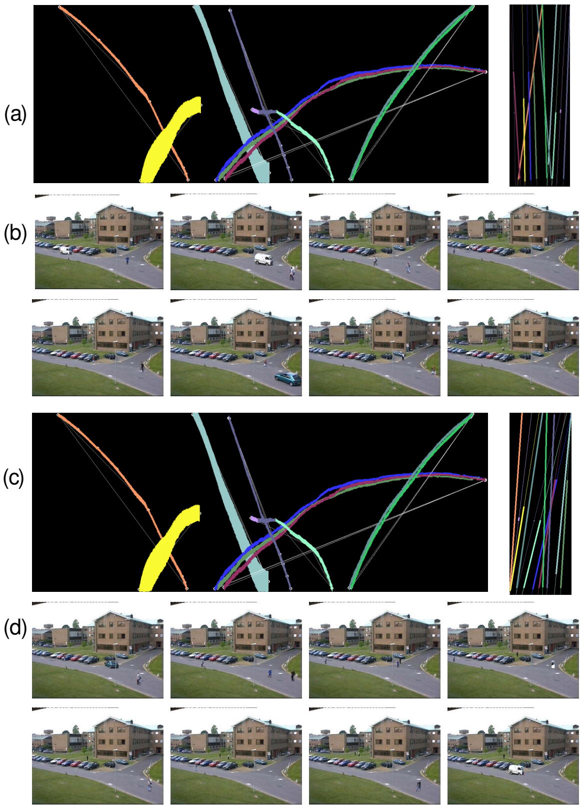

Figure 4. Visual results of the proposed method. (a, c) Trajectories of actions in the input video and their optimized trajectories for configuration C^* (b) Sample frames for the trajectories of (a). There are generally two moving objects in each frames. (d) Some sample frames from the tractability adjustment system's output for the trajectories shown in (c). This video output is optimized for three objects per frame.

presses a special key, which is called *monitored*. If the operator does not press a key, the alert ends in two seconds and we mark this alert as *missed*. The goal of the operator is to minimize the number of missed alerts.

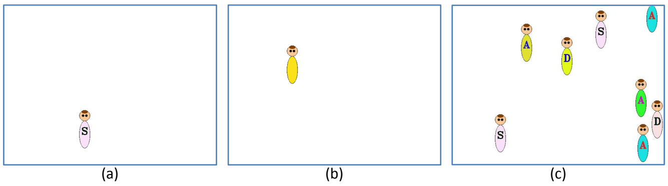

Figure 5. Task Interface. (a) An underloaded video with a single object. (b) Alerts are given by hiding objects label. (c) A sample frame from an overloaded period of expected workload case with lots of moving targets. See supplementary-2 for sample frames of our synthetic test scenarios.)

In the experiment, there are three different levels of workloads determined by the number of moving targets. For underloaded case (Fig. 5(a)), test videos consist of 12000 frames which is the repetition of the first 3000 frames four times. In expected-load case (Fig. 5(c)), 3000 frames of underloaded case and 3000 frames of overloaded case is used twice. In adaptive workload, we start with a small number of target objects and adjust the number by one every 30 seconds according to the operator's attention level. Average results are given in Table 1.

The table 1 shows that underloaded volunteers have better scores than the others. Their catch per alert rate is over 85% which is over 30% better than the best performance of the expected-loaded operators. Expected-loaded operators did nearly 90 percent of the misses when they are overloaded. From table 1, we concluded that although being underloaded is good for mistakes done, the total work done by underload operator is minimal compared to the others (Fig. 6). The operators track nearly five times more moving target than the underloaded operators.

We had several interesting observation about the gaze metrics and task scores. We observed that the operators who blink more have less score, which is an indication of boredness. We also observed that operator saccade rates start increasing when there are more target objects, which

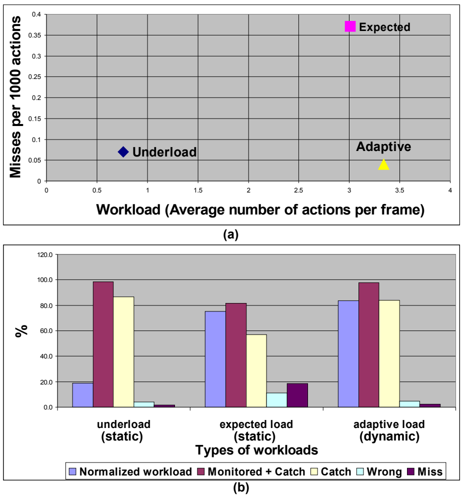

Figure 6. Graphical representation of Table 1. (a) Relation between workload and operator success is shown. Operators under adaptive workload do the best miss rates and they are nearly five times more efficient that the underloaded operators. (b) Bar representation of task performances for three workload types. Higher rates are better for the first three bars and the lowers are better for the last two. Workload of the adaptive operators are similar to the operators under expected-load but the success of operators under the adaptive load is nearly as good as the underloaded operators.

is an indication of overloading. Therefore, we use the time derivative of the saccade rates as a measure of overloading. If the derivative of saccade rates increases over 10% from the previous interval, we decrease the number of objects. We also decrease the number of objects if the blink rate increases over 5%, otherwise the attention level of the operator is increased.

When the results of adaptive-load experiment are compared with the results of the non-adaptive experiments, we see that adaptive-loaded operators clearly have better performance in terms of monitored, missed, and wrong alerts compared to expected-loaded case. The numbers for the proposed method and the underloaded cases are very similar. However, the underloaded operators perform at a much lower production rate as indicated by the average number of objects per frame. Therefore, we can claim that the proposed system is both more reliable and more efficient than the other methods. (Please, see supplementary-3 for the results of the tractability adjustment system on real surveillance data.)

Table 1. Quantitative Results of Psychological Experiment.

	Underload	Expected-Load	Adaptive-Load
#participants	3	3	5
Avg. #Actions per frame	0.76	3.01	3.34
Avg. #Alerts	40	74	87
Catch (%)	86.7	56.8	83.9
Wrong Alerts (%)	4.17	11.2	4.75
Missed Alerts (%)	1.67	18.4	2.30
Monitored + Catch Alerts (%)	98.3	81.6	97.7
Misses per 1000 actions	0.07	0.37	0.04

We tested our video tractability adjustment system on real surveillance data from PETS dataset [1]. In order to test the tractability adjustment system with this dataset we simulate the operator attention level. The results of a sparse and a dense output video examples are shown in Fig. 4. We also show the 2D space-time action trajectories with their initial line models. The optimized action location spaces are also shown. Our tractability adjustment running times are asymptotically $O(a.l)$, which takes 5 ms per frame on a standard Pentium 4 PC. Please see the supplementary material for the sample outputs of our system.

5. Conclusions

We introduced a surveillance monitoring method that places human operators into an active feedback loop within the video synthesis system. The system measures the attention levels of human operators and adjusts the video tractability to match the operator attention levels. The results of experiments show that proposed method increases the efficiency of operators while preserving the reliability of the system. The proposed system could be used for balancing the workloads among different operators, evaluating the operator performance, measuring the quality of the monitoring operation, and determining the shift periods of operators.

The proposed system defines the video tractability adjustment task as a facility location problem, which manipulates actions for better tractability and it produces more viewable surveillance video outputs than the classical video summarization methods. The proposed method can also be used for video summarization and it can produce more compact summaries by manipulating speed and direction of objects.

We plan to extend our video tractability adjustment system to work with multiple operators and multiple video streams. We also plan to perform object tracking using static and dynamic features extracted from experiments on operator object tracking behaviour.

Acknowledgements

This work is supported by TUBITAK Project 110E033.

References

[1] *PETS 2001 Benchmark Data*. Online, 2001.

[2] A. R. Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *IEEE Computer Vision and Pattern Recognition or CVPR*, pages I: 435–441, 2006.

[3] R. Allen, P. Mcgeorge, D. Pearso, and A. B. Milne. Attention and expertise in multiple target tracking. *Applied Cognitive Psychology*, 18:337–347, 2004.

[4] G. Alvarez and S. Franconeri. How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision*, 7(13), 2007.

[5] K. Arrington. Viewpoint eye tracker. *Arrington Research*, 1997.

[6] P. K. Atrey, A. El Saddik, and M. S. Kankanhalli. Effective multimedia surveillance using a human-centric approach. *Multimedia Tools Appl.*, 51:697–721, January 2011.

[7] E. Byrne and R. Parasuraman. Psychophysiology and adaptive automation. *Biological Psychology*, 42(3):249–268, 1996.

[8] P. Cavanagh and G. Alvarez. Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, 9(7):349–354, 2005.

[9] A. R. Dick and M. J. Brooks. Issues in automated visual surveillance. In *In International Conference on Digital Image Computing: Techniques and Applications*, pages 195–204, 2003.

[10] D. Douglas and T. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10(2):112–122, 1973.

[11] H. Fehd and A. Seiffert. Looking at the center of the targets helps multiple object tracking. *Journal of vision*, 10(4), 2010.

[12] H. M. Fehd and A. E. Seiffert. Eye movements during multiple object tracking: Where do participants look. *Cognition*, 108(1):201–209, July 2008.

[13] S. Franconeri, S. Jonathan, and J. Scimeca. Tracking multiple objects is limited only by object spacing, not by speed, time, or capacity. *Psychological Science*, 21(7):920, 2010.

[14] S. L. Franconeri, G. A. Alvarez, and J. T. Enns. How many locations can be selected at once? *J Exp Psychol Hum Percept Perform*, 33(5):1003–1012, October 2007.

[15] E. Gu, C. Stocker, and N. Badler. Do you see what eyes see? Implementing Inattentional Blindness. In *Intelligent Virtual Agents*, pages 178–190. Springer, 2005.

[16] T. Horowitz, A. Holcombe, J. Wolfe, H. Arsenio, and J. DiMase. Attentional pursuit is faster than attentional saccade. *Journal of Vision*, 4(7), 2004.

[17] H. Keval and M. A. Sasse. Man or gorilla? performance issues with cctv technology in security control rooms. *16th World Congress on Ergonomics Conference, International Ergonomics Association*, 2006.

[18] C. Kim and J.-N. Hwang. An integrated scheme for object-based video abstraction. In *MULTIMEDIA '00: Proceedings of the eighth ACM international conference on Multimedia*, pages 303–311, New York, NY, USA, 2000. ACM.

[19] H. Koskela. The gaze without eyes: Video-surveillance and the nature of urban space. *Progress in Human Geography*, 24(2):243–265, June 2000.

[20] F. C. Li, A. Gupta, E. Sanocki, L. wei He, and Y. Rui. Browsing digital video. In *CHI '00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 169–176, New York, NY, USA, 2000. ACM.

[21] B. Liu. Facility Location Problem. *Theory and Practice of Uncertain Programming*, pages 157–165, 2009.

[22] W. Ma and W. Huang. No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. *Journal of Vision*, 9(11), 2009.

[23] Y. Ma, L. Lu, H. Zhang, and M. Li. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 533–542. ACM, 2002.

[24] T. Oron-Gilad, A. Ronen, Y. Cassuto, and D. Shinar. Alertness maintaining tasks while driving. In *Human Factors and Ergonomics Society Annual Meeting Proceedings*, volume 46, pages 1839–1843. Human Factors and Ergonomics Society, 2002.

[25] R. Parasuraman. Adaptive automation for human-robot teaming in future command and control systems. Technical report, ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING DIRECTORATE, 2007.

[26] Z. W. Pylyshyn and R. W. Storm. Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial vision*, 3(3):179–197, 1988.

[27] R. Ren, P. Swamy, J. Jose, and J. Urban. Attention-based video summarisation in rushes collection. In *Proceedings of the international workshop on TRECVID video summarization*, pages 89–93. ACM, 2007.

[28] B. Scholl, Z. Pylyshyn, and J. Feldman. What is a visual object? Evidence from target merging in multiple object tracking. *Cognition*, 80(1-2):159–177, 2001.

[29] C. R. Sears and Z. W. Pylyshyn. Multiple object tracking and attentional processing. *Canadian Journal of Experimental Psychology*, 54(1):1–14, 2000.

[30] T. Stutzle. Iterated local search for the quadratic assignment problem. *European Journal of Operational Research*, 174(3):1519–1539, 2006.

[31] P. Thiffault and J. Bergeron. Monotony of road environment and driver fatigue: a simulator study. *Accident Analysis & Prevention*, 35(3):381–391, 2003.

[32] U. Vural and Y. Akgul. A parallel non-linear surveillance video synopsis system with operator eye-gaze input. In *Video Surveillance*. InTech., 2011.

[33] U. Vural and Y. S. Akgul. Eye-gaze based real-time surveillance video synopsis. *Pattern Recogn. Lett.*, 30(12):1151–1159, 2009.

[34] E. Wallace, D. Diffley, E. Baines, and J. Aldridge. Ergonomic design considerations for public area CCTV safety and security applications. In *Proceedings of the 13th Triennial Congress of the International Ergonomics Association, Tampere, Finland*, 1997.

[35] J. Warm, W. Dember, and P. Hancock. Vigilance and workload in automated systems. *Automation and human performance: Theory and applications*, pages 183–200, 1996.

[36] J. Wolfe, S. Place, and T. Horowitz. Multiple object juggling: Changing what is tracked during extended multiple object tracking. *Psychonomic bulletin & review*, 14(2):344–349, 2007.