A Machine Learning System For Human-in-the-loop 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
{uvural, akgul}@bilmuh.gyte.edu.tr

Abstract

We propose a novel human-in-the-loop surveillance system that continuously learns the properties of objects that are interesting for a human operator. The interesting objects are automatically learned by tracking the eye gaze positions of the operator while he or she monitors the surveillance video. The system automatically detects interesting objects in the surveillance video and forms a new synthetic video that contains interesting objects at earlier positions in the time dimension. The operator always views this synthetically formed video which makes manual video retrieval tasks more convenient. Sensitivity to operator interests and interest changes are other major advantages. We tested our system both on synthetic and real videos, which are provided as supplementary materials [1]. The results show the effectiveness of the proposed system.

1 Introduction

Digital surveillance technologies have become very popular with decreasing initial set-up costs [11]. In many cases, customers demand larger number of cameras for a better coverage. Although an extended coverage is desired, the amount of visual data increases by the number of cameras. Recorded surveillance videos and live video streams should be examined for events and suspicious actions. There are considerable number of studies for automatic inspection of these data [15]. Fully automated video analysis approaches [4] are fast but they only work for a limited number of models and their success is far from the desired levels [8]. Human operators generally make better decisions than the advanced surveillance systems at critical times[13] but manual processing of the information is a hard and time-consuming task. Therefore, we argue that a human-in-the loop approach should be adopted for increasing the efficiency of the human operators for the surveillance task.

Video synopsis is one popular approach for increasing the operator efficiency [9]. Synopsis methods show all actions in the video archive in a shorter time. Key-frame based video synopsis methods drop the frames with negligible amount of actions but these methods can easily discard some important actions when they are forced to produce shorter synopsis videos. Non-linear video synopsis methods [2] preserve actions better by changing the video time positions of the actions instead of discarding frames. One main problem of non-linear synopsis methods is that they generally produce complex output videos which are full of actions regardless they are interesting or not.

Human operators have limited tracking capabilities [3] and they can easily overlook important actions when the scenes are crowded. Human psychology should be taken into account for increasing surveillance system reliability because the number of objects can be tracked by an operator highly depends on the operator's experience, fatigue, and workload [13]. Recently, eye-gaze based metrics were used in surveillance systems for measuring the performance of the operators. In [12], we track the operators' eye-gaze positions to determine if an action in the scene is monitored or not. This system then gives a second chance to the operator by showing a non-linear synopsis of overlooked parts again. Another eye-gaze based method addresses the problem of crowded synopsis output videos [13]. This method adjusts the trackability of surveillance video according to the operator attention level. These methods support operators and increase the system reliability by preventing overlooks. Eye-gaze based metrics are also used for measuring the human interest especially for designing shop windows or internet sites [6] but none of the human-in-the-loop surveillance systems analyses the features that the operators find interesting.

In this paper, we propose a novel method that synthesizes new surveillance videos according to the operator's interest. Surveillance operators generally have special

interests on some types of objects or actions. These objects and actions vary according to the surveillance task at hand and operator's motivation. Furthermore, a special action or object seen in the video may change the interest of the operator. Thus, an online learning algorithm is a vital part of our system. We define the surveillance video synthesis problem as an online learning problem and introduce a novel learning scheme which classifies each action as interesting or not. Image and motion based features of an action are extracted and they are classified as interesting if eye-gaze positions of the operator match the action for some duration. For unseen video sections, the dynamically learned features are used to determine the interesting actions or objects and the chronological positions of these actions or objects are changed in the synthesized video so that the operator can see them earlier. This makes any manual video retrieval tasks more convenient for the operators. Similar video synthesis systems[10] classify the observed actions into several different classes or clusters without a dynamic operator model. As a result, they cannot easily adapt themselves for different operators and changing operator interests. Our method, on the other hand, continuously trains the learning scheme and if the operator shows change of interest, the system automatically adjusts itself. Furthermore, our synthesised video includes every action or object in the video but only changes their chronology so there are no missed action problems.

2 Method

Our system uses methods from the field of computer vision (CV), human-computer interaction (HCI), and machine learning (ML). An overview of the proposed method is shown in Figure 1.

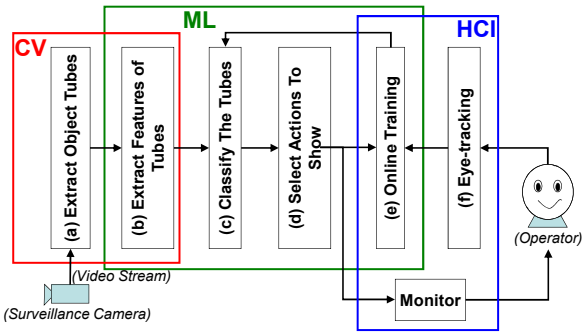


Figure 1. Overview of our method.

Computer vision algorithms are used to extract the set of all actions A from the input video V (Fig. 1-(a)). An action $a_i \in A$ is represented with an object tube $t_i \in T$ which is composed of a sequence of related bounding boxes t_i [10]. We use a mixture-of-Gaussian based foreground/background subtraction algorithm for finding the moving objects and their rectangular bounding boxes.

After obtaining an action set A, image and motion based low level features $\mathcal{X}$ of these actions are extracted (Fig. 1-(b)). Histograms of local binary patterns (LBP) are used for representing the texture of objects [14]. LBP based features are computed for each bounding box for which an eight-bin histogram is constructed. RGB color histogram of a bounding box provides another image based feature vector. We also use the sizes of bounding boxes, their aspect ratios, and center positions as other image based features. The direction of movement and speed are used as motion features. The motion based features are not computed for each bounding box and are only computed for each action once. The total number of features in a feature vector x_i for an action a_i contains 48 elements.

An online boosting method based on the gentleboost algorithm[7] is proposed to learn interesting actions in a binary classification setting. Our binary boosting algorithm takes any input $\mathcal{X} \in \mathbb{R}^D$ where D is the dimension of a feature vector x and maps it to a class label $\mathcal{Y} = \{+1, -1\}$. A binary boosting algorithm is continuously trained by using a data set $\{x_i, y_i\}, i = 1..N$ where $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$, and N is the number of training samples. The class label y_i of a training sample action a_i is either interesting ($y_i = 1$) or uninteresting ($y_i = -1$). If an action a_i is eye tracked (Fig. 1-(f)) by the operator longer than a given threshold, the system marks the action as interesting. The system determines the tracking time by measuring the duration of the overlap time between the eye gaze position and the bounding boxes of a_i. Note that our sample training set is continuously updated with the recently viewed actions by the operator, which can be achieved only by online training (Fig. 1-(e)). The training results are used in classifier C to assign class labels for the features x_i of the actions a_i detected from the input video (Fig. 1-(c)). An action queue Q is used to store all unmonitored actions a_i and their class labels $y_i = C(x_i)$. A selection priority $P(a_i)$ of each action a_i is calculated by

$$P(a_i) = C(x_i) + \lambda R(a_i), \qquad (1)$$

where λ is a weighting constant and $R(a_i) \in [0, 1]$ assigns a random number depending on the time of arrival of a_i so that $P(a_i)$ is larger for both interesting actions and actions that happened earlier. Actions that have higher selection priorities are shown to the operator first (Fig. 1-(d)) using the video synthesis method of [13]. Psychological studies indicate that a human operator can only track at most 4 independently moving

items at a time [3]. Therefore, the maximum number of active actions visible can be at most 4 at a given time.

3 Experiments

We tested our system on both synthetic and real world videos. First, we built a synthetic test video of length 120 minutes to evaluate our system in a controlled setting. The synthetic test video includes 30 different types of objects which have 5 different patterns and 6 different colors (Fig. 2). These objects move in random directions at random speeds. There are total of 1200 actions included in the test video. Figure 4-(a) shows the graph of cumulative object counts for each object type at a given time position, which shows that the count values increase regularly for all 30 object types.

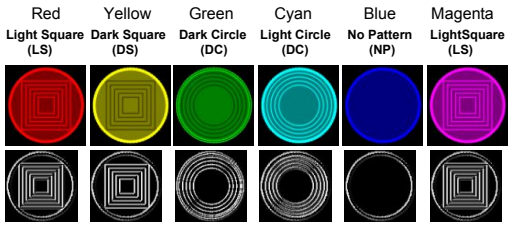


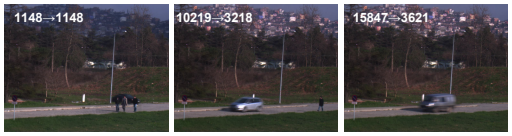
Figure 2. Sample synthetic objects.

This test video is provided to our system as the video stream from a surveillance camera (see Fig 1). We used the eye-gaze tracker of Arrington Research [5] for obtaining the operator eye gaze positions. We run our system 5 times on the test video. For each of the tests, we asked the volunteer operators to track a specific object or objects for the first 10 minutes and then asked them to track another object or objects for the next 10 minutes. Table 1 shows the object property values used for each test.

Table 1. Properties of interesting objects for the 5 synthetic test cases

	The first 10 minutes		The second 10 minutes	
Test	Color	Pattern	Color	Pattern
1	Red	All	Magenta	DC
2	Red	DC	Blue	DC
3	Blue	NP	All	LS
4	Magenta	LC	Magenta	NP,DS
5	Yellow	LS	Yellow,Cyan	LC

For each experiment, we kept the cumulative counts of each object type that the operator monitors. Figure 4-(b) shows the graph of cumulative object counts for experiment 1. The counts of all red objects increase sharply for the first 10 minutes because we asked the operator to track red objects of all patterns for the first 10 minutes. For the next 10 minutes, the counts of magenta dark circle objects increase sharply as expected. All the other (uninteresting) object counts increase slowly for the whole experiment. We only include two uninteresting object types in Figure 4 for brevity. We see similar results for the other experiments (Fig. 4-(c,d,e,f)). Note that a very big percentage of the interesting objects are viewed in the first 20 minutes of the experiment which shows the suitability of the proposed system for a manual video retrieval/search task. Please see the supplemental materials for the complete videos of these experiments[1].



(a) All gray automobiles are interesting.

(b) Humans are interesting, cars are not.

Figure 3. Sample frames from real dataset.

We performed 2 experiments on the real data using a surveillance video from a university campus setting. The original video was 120 minutes long. For the first experiment, we asked the operator to track gray automobiles. The synthesized video frames are shown in Figure 3-a. We show the original and synthetic video frame numbers of the actions on top left corners of the frames, which shows that the gray colored automobile actions are moved to earlier times in the synthesized videos. For the second experiment, we asked the operator to track humans instead of automobiles. Figure 3-b shows a few sample frames from the resulting videos. Note that there are objects other than automobiles that are shown at earlier times, which is expected because the selection priority formula (Eq. 1) includes a term for randomness and favor for earlier objects. Note also that, the initial frames of the produced videos may include uninteresting objects because at this phase, the online learning algorithm does not have enough training data to make a good decision. Please see the supplemental materials for the complete videos of the real data experiments[1]. Both experiments on real and synthetic data showed that our system can successfully adapt to operator interests. The final system makes it very convenient for the operator to go through a lengthy video to search for desired objects and actions without giving explicit instructions to the system.

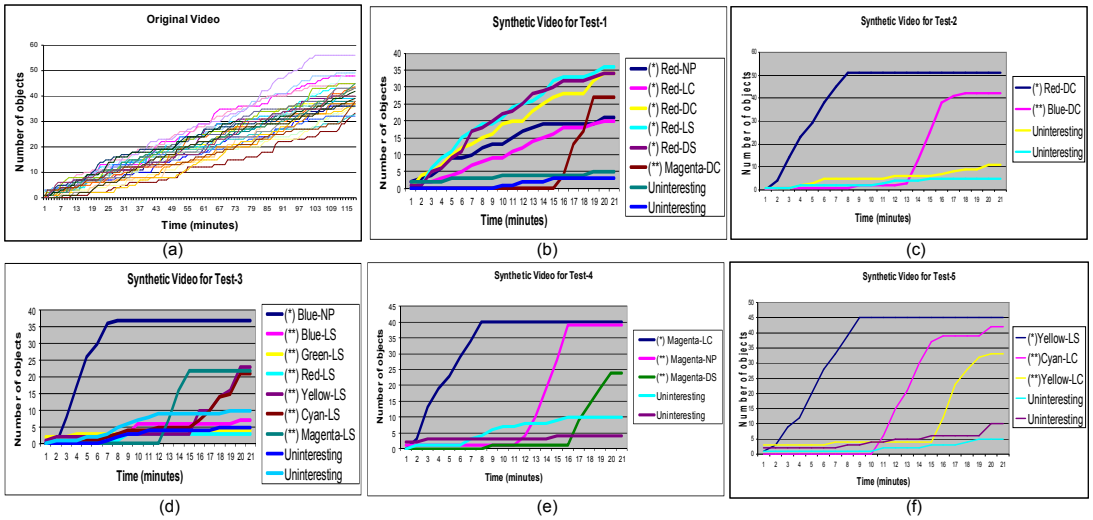


Figure 4. Graphical results of the synthetic experiments. (*): Interesting object types for the 1st 10 minutes, (**): Interesting object types for the 2nd 10 minutes.

4 Conclusions

A novel surveillance video synthesis method that uses online operator interest is introduced. The method uses eye-gaze measures of the operator to find interesting actions and synthesizes the interesting actions first. The results of the experiments show that the system adapts itself well on interest changes. In addition, the time required for monitoring the interesting actions is decreased. The proposed system could be used for real-time surveillance monitoring and offline retrieval.

The proposed system combines methods from human-computer interaction, computer vision, and machine learning. Operator interest based binary classification of objects is novel. The system is robust against classification errors because it changes the times of actions instead of filtering uninteresting classes. We plan to extend our system to work with multiple operators and multiple video streams where each operator monitors the surveillance video with different interests. This method could also be used for analysing the monitoring strategies, subjective interests and overall performance of the surveillance operators.

Acknowledgements

This work is supported by TUBITAK Project 110E033.

References

[1] Project site: http://vision.gyte.edu.tr/projects.php?id=7.

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

[3] 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.

[4] B. Antic and B. Ommer. Video parsing for abnormality detection. In *ICCV*, pages 2415–2422, 2011.

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

[6] S. Castagnos and P. Pu. Consumer decision patterns through eye gaze analysis. In *Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 2010.

[7] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The annals of statistics*, 28(2):337–407, 2000.

[8] 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.

[9] Y. Li, Y. Li, T. Zhang, T. Zhang, D. Tretter, and D. Tretter. An overview of video abstraction techniques. Technical report, HP Laboraties, Palo Alto, 2001.

[10] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg. Clustered synopsis of surveillance video. In *AVSS*, pages 195–200, Washington, DC, USA, 2009.

[11] M. A. Sasse. Not seeing the crime for the cameras? *Commun. ACM*, 53(2):22–25, 2010.

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

[13] U. Vural and Y. S. Akgul. Operator attention based video surveillance. In *ICCV Workshops*, pages 1955–1962, 2011.

[14] X. Wang, T. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV*, pages 32–39, 2009.

[15] C. Yu, X. Zheng, Y. Zhao, G. Liu, and N. Li. Review of intelligent video surveillance technology research. In *(EMEIT*, volume 1, pages 230–233. IEEE, 2011.