SELECTING ATTENTIVE FRAMES FROM VISUALLY COHERENT VIDEO CHUNKS FOR SURVEILLANCE VIDEO SUMMARIZATION

Wenzhong Wang Qiaoqiao Zhang Bin Luo Jin Tang Rui Ruan Chenglong Li *

Department of Computer Science and Technology, Anhui University
No.111 Jiulong Road, Hefei 230601, China
{wenzhong,tj,luobin}@ahu.edu.cn, ahu_qq@163.com, rr_ahu@126.com, lcl1314@foxmail.com

ABSTRACT

This paper investigates how to extract key-frames from surveillance video while maximizing their diversity and representational ability. We solve this problem by two steps, i.e., video partition and frame selection. The first step is to partition a surveillance video into visually coherent video chunks, which have high intra-chunk similarity and inter-chunk dissimilarity. In particular, we propose an object-based frame metric to measure the relevance of two frames, and apply the Normalized Cut algorithm to achieve video partition. The second step is to select the attentive frames from the partitioned video chunks. We propose an attention score based on the content completeness and the visual satisfaction for each frame, and select most attentive frame with highest attention score in each chunk. Extensive experiments on both public and our newly created datasets suggest that our approach significantly outperforms other video summarization methods.

***Index Terms*—** video summarization, video partition, normalized cut, frame selection, attention function

1. INTRODUCTION

The task of surveillance video summarization is to extract meaningful and representative information, a small set of key-frames, from a surveillance video. It has attracted much attention recently due to its importance in browsing and storing huge surveillance videos. The main difficulty of large scale surveillance video summarization arises from the contradiction between the high-degree spatiotemporal redundancies and the limited storage budget.

Recent techniques on surveillance video summarization can be categorized into two groups: 1) Frame-based methods, that extract the feature representations from the whole frame, and then summarize the surveillance video [1, 2, 3, 4]. Avila et al. [1] employed the k-means algorithm on the color features extracted from video frames to perform video summarization, and Yang et al. [2] proposed an edge histogram descriptor to describe each frame for summarizing lengthy surveillance videos. A context-aware surveillance video summarization framework is proposed by Zhang et al. [4] that adopt the generalized sparse group lasso to learn a dictionary of video features and a dictionary of spatio-temporal feature correlation graphs. These methods, however, utilize the whole frame information to summarize videos, which usually includes many redundant noises (e.g., static background). 2) Object-based methods, that first extract the objects or moving objects from the input videos, and then employ the object-based frame representation to perform surveillance video summarization [5, 2, 6]. Ji [5] summarized the video using object trajectories depicted in key-frames. The key-frame selection is arbitrary, and they use the total areas of objects in each frame as frame feature to segment video into chunks. Yang et al. [2] used localised foreground entropy of foreground objects to extract the key-frames. An event-based approach is proposed by Song et al. [6] that extract the trajectories of vehicles and pedestrians in a tracking-by-detection manner, and then detects the abnormal events using the trajectories to generate a summarized sequence. Although focus on foreground objects to remove redundant information, these methods ignore some important cues or priors, such as high-level features of foreground objects and visual attention mechanism for key-frame selection, and thus may decrease the performance in challenging scenes.

To handle these problems, we propose an effective approach in this paper for surveillance video summarization, which selects a set of attentive frames from visually coherent video chunks. More specifically, we perform surveillance video summarization in two steps:

- We partition a video into several chunks that are visually coherent. We first employ the ViBe algorithm [7] to detect moving objects for each frame, and the difference between two frames is then calculated by the distance between their respective moving object set. In

This work was supported in part by the National Natural Science Foundation of China under Grant 61671018 and Grant 61472002, in part by the Natural Science Foundation of Anhui Higher Education Institution of China under Grant KJ2017A017 and in part by the Co-Innovation Center for Information Supply & Assurance Technology of Anhui University under Grant Y01002449.

particular, we use several object's features, which include appearance and geometric feature and class labels assigned by Support Vector Machine (SVM) [8] to match objects between two frames, and define a new frame affinity based on the matching results. Then, the original video is partitioned into several video chunks by applying the Normalized Cut (NCut) algorithm [9] on the computed affinity matrix.

- We select an attentive frame from each video chunk. Specifically, we define an attentive function for each frame based on the content completeness and the visual satisfaction [10]. The former indicates that a key-frame should include all objects appeared in its scope, and the latter denotes that a key-frame should present all objects in a visually pleasing way. Therefore, we can select most attentive frame by maximizing the defined score in one video chunk. Finally, we obtain a set of key-frames as the summarization of the original video.

This paper makes the following contributions. First, we propose an effective approach for surveillance video summarization. Extensive experiments show that our method can effectively suppress the redundant information and preferably select representative key-frames. Second, we design a new distance metrics between two frames based on the location, appearance and semantic information of their respective objects. The designed distance can guarantee that the partitioned video chunks are visually coherent. Third, we present an attentive function based on the content completeness and the visual satisfaction, and selects a most informative frame from each video chunk by maximizing the attentive score. Fourth, we create a new dataset for surveillance video summarization, and will release it online for free academic usage[1].

2. APPROACH

In this paper, we solve the problem of surveillance video summarization by two steps. The first step is to find visually coherent video chunks (Section 2.1), and the second one is to select the most representative frame from each chunk (Section 2.2). These two steps can guarantee the selected key-frame set with high diversity and representative ability. The pipeline of our approach is presented in Fig. 1.

2.1. Visually Coherent Video Partition

Given the input video with N frames, we first employ the ViBe algorithm [7] to detect the moving objects, and denote O^{f_A} as the moving object set of the frame f_A with containing m moving objects, i.e., $O^{f_A} = \{O_1^{f_A}, O_2^{f_A}, \dots, O_m^{f_A}\}$. In this paper, each moving object $O_j^{f_A}$ is described using a set of low-level and high-level features:

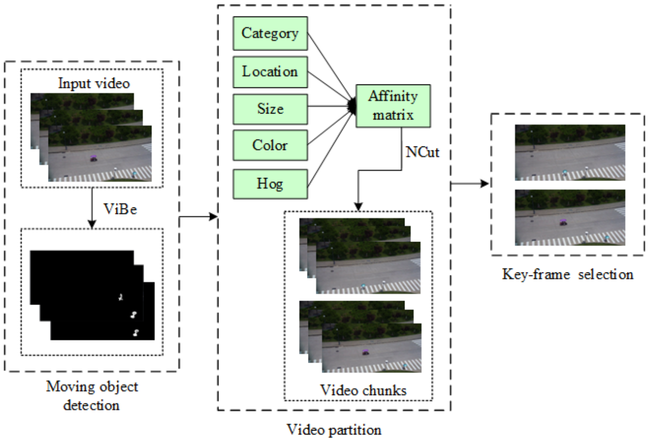

Fig. 1. The pipeline of our framework.

$$O_j^{f_A} = \langle \mathbf{p}(O_j^{f_A}), \mathbf{s}(O_j^{f_A}), \mathbf{a}(O_j^{f_A}), l(O_j^{f_A}) \rangle, \quad (1)$$

where $\mathbf{p}(O_j^{f_A})$ and $\mathbf{s}(O_j^{f_A})$ are the position and size of the object $O_j^{f_A}$, respectively. $\mathbf{a}(O_j^{f_A})$ denotes the appearance features of $O_j^{f_A}$, including the RGB histogram and the HOG feature, and $l(O_j^{f_A})$ is the class label of $O_j^{f_A}$. In particular, we use the multi-class linear SVM algorithm [8] to classify the objects into five categories, single pedestrian, multiple pedestrians, bicycle, car, and the unknown class (some rare objects). By this way, we combine the low-level and high-level features to represent each moving object effectively.

Next, we utilize the detected moving objects to measure the relevance between two video frames. Let O^{f_A} and O^{f_B} denote the object set of the frame f_A and f_B with the size of m and n, respectively, and we employ the class labels of moving objects to generate some matching candidates. Let $\mathbf{X}^0$ denote the candidate match matrix constructed by the consistency of class labels, and we have

$$\mathbf{X}_{ij}^0 = \begin{cases} 1 & if\ l(O_i^{f_A}) = l(O_j^{f_B}) \\ 0 & else \end{cases} \quad (2)$$

with $i = 1, ..., m$, and $j = 1, ..., n$. Given the candidate match matrix $\mathbf{X}^0$, we solve the final matching matrix $\mathbf{X}$ by optimizing the following program:

$$\begin{aligned} &\mathbf{X}^* = \arg\min_{\mathbf{X}} \sum_{i,j} \mathbf{X}_{ij}^0 \mathbf{D}_{ij}^{AB} \mathbf{X}_{ij} \\ &s.t. \forall_i, \sum_{j=1}^{n} \mathbf{X}_{ij} \le 1,\ \forall_j, \sum_{i=1}^{m} \mathbf{X}_{ij} \le 1,\ \mathbf{X}_{ij} \in \{0,1\}, \end{aligned} \quad (3)$$

where $\mathbf{D}_{ij}^{AB}$ is the feature distance of $O_i^{f_A}$ and $O_j^{f_B}$:

$$\begin{aligned} \mathbf{D}_{ij}^{AB} =& \parallel \mathbf{p}(O_i^{f_A}) - \mathbf{p}(O_j^{f_B}) \parallel + \\ & \parallel \mathbf{s}(O_i^{f_A}) - \mathbf{s}(O_j^{f_B}) \parallel + \parallel \mathbf{a}(O_i^{f_A}) - \mathbf{a}(O_j^{f_B}) \parallel, \end{aligned} \quad (4)$$

[1] Dataset webpage: http://chenglongli.cn/people/lcl/confs.html.

The problem (3) can be efficiently solved by using the integer programming algorithm [11].

Given the optimized match matrix $\mathbf{X}^*$, we define the distance between two frames f_A and f_B:

$$d(f_A, f_B) = \frac{\sum_{i,j} \mathbf{X}^*_{ij} \mathbf{D}^{AB}_{ij} + (n+m-2c)d_0}{n+m-c}, \quad (5)$$

where d_0 is a constant that penalties the similarity between mismatching objects, which is fixed to be $10 \max\{\mathbf{D}^{AB}_{ij}, A, B \in \{1,2,...,N\}, i \in \{1,2,...,m\}, j \in \{1,2,...,n\}\}$ in this paper. $c = \sum_{i,j} x_{ij}$ is the number of the matched objects.

From (5) we can see that if $O^{f_A}_i$ and $O^{f_B}_j$ have sufficiently similar visual features as well as the same class label, their visual distance $\mathbf{D}^{AB}_{ij}$ would contribute a small quantity to the inter-frame distance $d(f_A, f_B)$. On the other hand, each unmatched object in O^{f_A} or O^{f_B} will add a large quantity d_0 to $d(f_A, f_B)$, indicating a significant content change between frame f_A and f_B resulted from this unmatched objects.

Finally, we partition the input video into several chunks based on the defined distance in (5) with two criteria: (1) Coherence, each chunk present a visually coherent content, i.e., all of frames in the same chunk should contain the objects with strong correlation. (2) Separation, different chunks should be visually separable in terms of visual content, i.e., their object sets should exhibit different characteristics. The first criterion can be expressed as finding the chunk $\pi_l \in \{\pi_1, \pi_2, \ldots, \pi_s\}$, that minimizes the total intra-chunk distance $\sum_{l=1}^{s} d(\pi_l, \pi_l)$, and the second one is to maximize the inter-chunk distance to encourage the distinction between different chunks $\sum_{p \neq q} d(\pi_p, \pi_q)$, where the distance between two chunks π_p and π_q is $d(\pi_p, \pi_q) = \sum_{f_A \in \pi_p} \sum_{f_B \in \pi_q} d(f_A, f_B)$.

In this paper, we employ the NCut algorithm [9, 12] to achieve the above goals. Specifically, we construct a weighted graph $G = (V, E)$, where V is a node set(video frame), and E is an edge set (connecting each pair of frames). The weight of an edge between f_A and f_B is defined as:

$$\mathbf{W}_{AB} = e^{-\frac{d^2(f_A, f_B)}{\sigma_f^2}} e^{-\frac{|A-B|^2}{\sigma_\tau^2}}, \quad (6)$$

where $|A-B|$ is normalization distance of frame number, σ_f and σ_τ are weight of $d(f_A, f_B)$ and $|A-B|$, respectively. We apply the NCut algorithm on the normalization of the graph affinity matrix $\mathbf{W}$ to partition the input video into the chunks with high intra-chunk similarity and inter-chunk dissimilarity.

2.2. Attentive Frame Selection

Given the partitioned visually coherent video chunks, we now proceed to select one attentive key-frame from each chunk. We hope that a good key-frame would convey visual contents as much as possible, and be visually appealing as well. Therefore, two criteria are taken into account for these purposes. The first one is content completeness, which means that a key-frame should present all objects appeared in its scope. The second one is visual satisfaction [10], meaning that a key-frame should present the objects in a visually satisfying way, and we use closeness and uniformity around the image center to represent, see Figure 2. We fuse these two criteria to define a attention function, and select the most representative frames by maximizing the attention score.

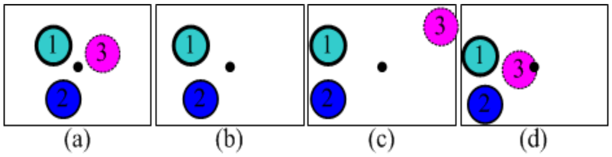

Fig. 2. Illustration of key-frame selection. In an individual chunk, our algorithm favors (a) since it contains all of the three objects and their configuration is the most visually satisfying one. (b) missed one object, (c) objects are spread too widely and (d) objects are far away from the image center.

The Content Completeness of frame f is quantified as proportion of the number of objects appeared in f with respect to the maximum count of different objects in its scope:

$$Completeness(f) = \frac{|O^f|}{\max_{k \in \pi} |O^k|} \quad (7)$$

where π is the representation scope of f, and $|O^k|$ is the cardinality of object set O^k.

Our definition of the visual satisfaction is motivated by the "center bias" phenomena discovered in eye-tracking studies [13, 14, 15]. These studies have shown that human gaze have a bias towards the center of natural scenes, as shown in Figure 2 (a). More specifically, the objects should spread closely and uniformly around the image center. The closeness of frame is defined as the radius of the object set O^f:

$$Closeness(f) = 1 - \frac{1}{\sqrt{2}} \max_j \| (\frac{x^f_j}{w}, \frac{y^f_j}{h}) - (\frac{x^f}{w}, \frac{y^f}{h}) \|,$$
$$s.t. \quad x^f = \frac{1}{|O^f|} \sum_{j=1}^{|O^f|} x^f_j, y^f = \frac{1}{|O^f|} \sum_{j=1}^{|O^f|} y^f_j. \quad (8)$$

where (x^f_j, y^f_j) is the position of object $O^f_j \in O^f$, (x^f, y^f) is the centre of the object set O^f, and w, h are the image width and height.

The uniformity of frame f is defined using the distance of centroid of the object set O^f to the image center:

$$Uniformity(f) = 1 - \sqrt{2} \, \| (\frac{x^f}{w} - \frac{1}{2}, \frac{y^f}{h} - \frac{1}{2}) \| \quad (9)$$

Finally, we fuse the above three terms to define the attention function of the frame f:

$$\begin{aligned} H(f) = \alpha \, Completeness(f) + \beta \, Closeness(f) \\ + \gamma \, Uniformity(f) \end{aligned} \quad (10)$$

where α, β, and γ indicate the weights of these terms. Hence, the best key-frame is identified as the most visually satisfying frame by $f^* = \arg\max_f \ H(f)$.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. Evaluation Setup

Datasets. Our study focuses on the surveillance videos recorded under still cameras, and thus we select 9 suitable videos from the public available surveillance video datasets, CAVIAR [16], the Change Detection Dataset [17], and PETS [16]. The scene images are referred to the **supplementary file**. For more comprehensive comparison, we also create a new dataset, which includes 8 videos captured on a college campus. Each video contains one or more moving objects with different categories, and we annotate the groundtruths of all videos using the protocol of [1]. The new dataset with annotated ground truths will be available online, where the download site are presented in Section 1.

Parameters. For fair evaluations, we fix all parameters in our experiments as follows. σ_f and σ_τ are set to be the 20 percent of the maximum values of the normalization of $d(f_A, f_B)$ and $|A-B|$, respectively. Besides, we set $[\alpha,\ \beta,\ \lambda]$ = [0.6, 0.2, 0.2].

Baseline methods. To demonstrate the effectiveness of our method, we implement 5 baseline methods, including 2 frame-based methods, Color [1], EHD [2], and 3 object-based representation methods, Kim [18], LFE [2], Ji [5].

3.2. Comparsion results

We report the performance on both the newly created and public datasets. The quantitative comparison results of the proposed approach on the newly created dataset with other 5 baselines (including 3 object-based methods and 2 frame-based methods) are presented in Table 1 and Table 2.

From Table 1 and Table 2, we can see that our method achieves superior performance than other methods. It demonstrates the effectiveness of the proposed feature representations of moving objects that integrate low-level and high-level features, distance metric between two frames and attention function that reflects the representational ability of video frame.

Specifically, we analyze the details of different methods as follows. Color [1] and EHD [2] these two frame-based representation methods utilize the holistic information of a frame, and thus includes much redundant information. Such feature representations are not effective to discriminate different objects in the scenes. In addition, LFE [2] extracts localised foreground entropy feature from foreground masks while Kim [18] focuses on object geometrical features. Ji [5] extracts the key-frame is arbitrary with using the total areas of objects in each frame as frame feature to segment video into chunks. They all can not reveal the rich information (such as the appearance of individual object, the relationship among each object, etc.) conveyed by each object. In contrary, we make best use of the low-level and high-level features of moving objects and their relationship to represent a frame, which is far more informative than these methods. And we select key-frames using two attention criteria other than the arbitrary selection in [5]. As a result, our method extracts objects and calculates the distance of two frames effectively.

Table 1. Comparison results of our method against other object-based baseline methods, where P, R, and F denote $precision$, $recall$ and $F - measure$, respectively.

Videos	Ours			LFE [2]			Kim [18]			Ji [5]		
	P	R	F	P	R	F	P	R	F	P	R	F
001	**0.64**	**0.32**	**0.43**	0.22	0.25	0.23	0.29	0.32	0.30	0.33	0.12	0.18
002	**0.75**	**0.37**	**0.50**	0.25	0.28	0.27	0.13	0.12	0.12	0.50	0.25	0.33
003	**0.64**	**0.59**	**0.61**	0.39	0.36	0.37	0.40	0.46	0.43	0.00	0.00	0.00
004	**0.61**	**0.54**	**0.57**	0.29	0.30	0.27	0.11	0.28	0.15	0.32	0.28	0.30
005	**1.00**	0.51	**0.67**	0.62	**0.70**	0.66	0.67	0.50	0.57	0.39	0.20	0.27
006	**1.00**	**1.00**	**1.00**	0.67	**1.00**	0.80	**1.00**	**1.00**	**1.00**	0.00	0.00	0.00
007	**0.68**	**0.68**	**0.68**	0.64	0.64	0.64	0.54	**0.68**	0.60	0.00	0.00	0.00
008	**0.50**	0.43	**0.46**	0.38	0.49	0.43	0.23	**0.56**	0.33	0.24	0.15	0.18
Ave	**0.73**	**0.56**	**0.62**	0.43	0.50	0.46	0.42	0.49	0.44	0.22	0.13	0.16

Table 2. Comparison results of our method against other frame-based baseline methods, where P, R, and F denote $precision$, $recall$ and $F - measure$, respectively.

Videos	Ours			Color [1]			EHD [2]		
	P	R	F	P	R	F	P	R	F
001	**1.00**	**0.49**	**0.66**	0.31	0.26	0.28	0.19	0.21	0.20
002	**0.75**	**0.37**	**0.50**	0.14	0.12	0.13	0.37	0.40	0.38
003	**0.79**	0.71	**0.75**	0.27	0.42	0.33	0.52	**0.94**	0.67
004	**0.61**	0.54	**0.57**	0.38	**0.75**	0.50	0.29	0.50	0.36
005	**1.00**	0.51	**0.67**	0.69	**0.52**	0.60	0.05	0.05	0.05
006	**1.00**	**1.00**	**1.00**	0.50	**1.00**	0.67	0.29	**1.00**	0.44
007	**0.75**	**0.75**	**0.75**	0.04	0.07	0.05	0.25	0.25	0.25
008	**0.50**	**0.43**	**0.46**	0.19	0.24	0.21	0.07	0.19	0.10
Ave	**0.81**	**0.60**	**0.67**	0.32	0.42	0.35	0.25	0.44	0.31

In addition, we also present the qualitative results on our dataset in the **supplementary file**. The effectiveness of the proposed approach is also verified in public videos, and we present the qualitative comparison results in the **supplementary file** due to space limitation.

4. CONCLUSION

In this paper, we have proposed an effective approach for surveillance video summarization via attentive frame selection from visually coherent video chunks. We have integrated the low-level and high-level features of moving objects to refine the affinities among video frames, and then employed the NCut algorithm to partition the input video into a set of visually coherent video chunks. For each video chunk, we have selected one most informative key-frame by considering the content completeness and the visual satisfaction. Extensive experiments on both public and newly created datasets have demonstrated the effectiveness of the proposed approach. In future work, we will study the online or streaming algorithm in our framework to improve the practicality.

5. REFERENCES

[1] S.E.F. de Avila and A. de Albuquerque Araújo, "Vsumm: an approach based on color features for automatic summarization and a subjective evaluation method," in *proceedings of the XXII Brazilian symposium on computer graphics and image processing*, 2009. 1, 4

[2] Y. Yang, D. Farhad, S. Conrad, and B.C. Lovell, "Summarisation of surveillance videos by key-frame selection," in *proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras*, 2011. 1, 4

[3] G. Liu, Y. Zhou, X. Li, and P. Yan, "Unsupervised, efficient and scalable key-frame selection for automatic summarization of surveillance videos," *Multimedia Tools and Applications*, pp. 1–23, 2016. 1

[4] S. Zhang, Y. Zhu, and Amit K. Roy-Chowdhury, "Context-aware surveillance video summarization," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5469–5478, 2016. 1

[5] Z. Ji, Y. Su, R. Qian, and J. Ma, "Surveillance video summarization based on moving object detection and trajectory extraction," in *proceedings of the International Conference on Signal Processing Systems*, 2010. 1, 4

[6] X. Song, L. Sun, J. Lei, D. Tao, G. Yuan, and M. Song, "Event-based large scale surveillance video summarization," *Neurocomputing*, vol. 187, pp. 66–74, 2016. 1

[7] B. Olivier and V.D. Marc, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image processing*, vol. 20, no. 6, pp. 1709–1724, 2011. 1, 2

[8] S. Christian, L. Ivan, and C. Barbara, "Recognizing human actions: a local svm approach," in *proceedings of the International Conference on Pattern Recognition*, 2004. 2

[9] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000. 2, 3

[10] P. Tseng, R. Carmi, I. Cameron, D. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of vision*, vol. 9, no. 7, pp. 4–4, 2009. 2, 3

[11] G. Nemhauser and L. Wolsey, "Integer programming and combinatorial optimization," *Wiley, Chichester. GL Nemhauser, MWP Savelsbergh, GS Sigismondi (1992). Constraint Classification for Mixed Integer Programming Formulations. COAL Bulletin*, vol. 20, pp. 8–12, 1988. 3

[12] C. Li, L. Lin, W. Zuo, S. Yan, and J. Tang, "Sold: Suboptimal low-rank decomposition for efficient video segmentation," in *Computer Vision and Pattern Recognition*, 2015. 3

[13] T. Foulsham and G. Underwood, "What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition," *Journal of Vision*, vol. 8, no. 2, pp. 6–6, 2008. 3

[14] B. Markus, "Scene and screen center bias early eye movements in scene viewing," *Vision research*, vol. 50, no. 23, pp. 2577–2587, 2010. 3

[15] N. Ejaz, I. Mehmood, and S.W. Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Processing Image Communication*, vol. 28, no. 1, pp. 34–44, 2013. 3

[16] V. Roberto and C. Rita, "Video surveillance online repository: an integrated framework," *Multimedia Tools and Applications*, vol. 50, no. 2, pp. 359–380, 2010. 4

[17] G. Nil, J. Pierre-Marc, P. Fatih, K. Janusz, and I. Prakash, "Changedetection. net: A new change detection benchmark dataset," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012. 4

[18] C. Kim and J.N. Hwang, "Object-based video abstraction for video surveillance systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 12, pp. 1128–1138, 2002. 4