Video Indexing and Summarization as a Tool for Privacy Protection

Rubén Heras Evangelio, Tobias Senst, Ivo Keller and Thomas Sikora
Communication Systems Group, Technische Universität Berlin
Einsteinufer 17, 10587 Berlin, Germany
heras,senst,keller,sikora@nue.tu-berlin.de

***Abstract*—The ever increasing number of surveillance camera networks being deployed all over the world has resulted in a high interest in the development of algorithms to automatically analyze the video footage, but has also opened new questions as how to efficiently manage the vast amount of information generated and, more important, how to protect the privacy of the individuals being recorded in their daily life. In this paper, we present a survey on video summarization techniques developed in order to efficiently access to the points of interest in the video footage. Thereby, we emphasize on the links that these techniques show with the task of privacy protection and draw lines of future research directions to incorporate indexing and summarization as tools for privacy protection by design.**

***Keywords*-Video Indexing; Video Summarization; Privacy Protection; Video Surveillance**

I. Introduction

Video surveillance systems have experienced a fast growth in the last decades, especially after the attacks on the 11th of September 2001 in New York, 11th of March 2004 in Madrid and 21st of July 2005 in London, to the point that they have become a part of our daily life. The use of video surveillance systems is not restricted to safety and security applications. Nowadays, video surveillance systems are also being deployed at department stores, highways and even on elderly houses to assist people in a non-invasive manner. This success has been supported by the decaying prices in the sensor industry, which is able to provide higher quality cameras of ever smaller sizes at low prices, the transition to IP camera networks, which allow to monitor large camera networks both local and remotely, the introduction of wireless networks, with the consequent reduction in deployment costs, and the development of automated video analytics, which gave raise to the paradigm of bringing intelligence to the network. A simple search in the Internet makes it easy to realize that this fast growing is expected to follow in the next years.

Nevertheless, as video surveillance systems have become ubiquitous some aspects of the deployed systems have been brought into question. One of the aspects is the effectiveness regarding crime prevention [1]. Moreover, protecting the privacy and security of personal information has gained increasing attention in the recent years. The Telegraph claimed that an individual will appear on average on 300 CCTV cameras during a day [2].

Obviously, the rapid growth of video surveillance systems results in an increasing number of video feeds which should be watched and stored in a control room. This results in a continuously growing workload for CCTV operators, who are overwhelmed by the huge sets of cameras. To alleviate this problem, automatic video analysis techniques aim at understanding actions and human behaviors in video sequences in order to alert CCTV operators upon the occurrence of threatening situations. This scenario corresponds to the proactive side of crime prevention. Besides that, video surveillance systems can also be used for crime investigation and offenders prosecution. Video indexing and summarization can be used in order to effectively accomplish this last task.

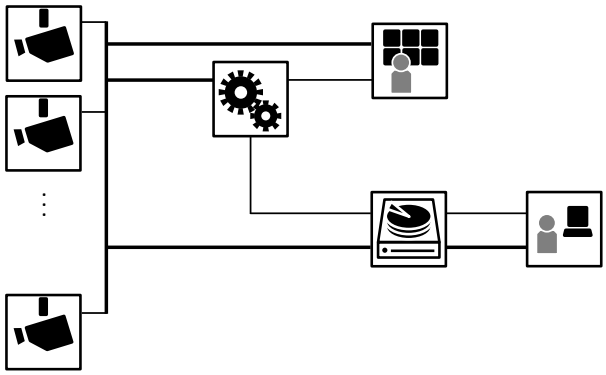


Fig. 1. Automated video surveillance scenario. Extracted content by means of video analysis is used for alerting control room operators in proactive crime prevention (top) and for video summarization for crime investigations (bottom).

Video summarization is a process which aims at providing the viewer with an overview of the content of a video. For that purpose, it is necessary to find the relevant information contained in the video to be summarized (video indexing), and to develop a proper representation method which allows the user to rapidly grasp the extracted information and to navigate through it. Furthermore, as the user is directly driven to the critical points in time, the privacy of the people recorded at irrelevant passages of the video sequences is preserved.

In this paper, we present a survey on state-of-the-art video summarization techniques for the video based surveillance domain, thereby emphasizing on the links that these techniques show with the task of privacy protection. Therefore, in Section II we present the main techniques for both indexing and

978-1-4673-5807-1/13/$31.00 ©2013 IEEE

representation. In Section III, we review some representative approaches of the respective techniques. In Section IV, we analyze the opportunities that the presented techniques offer for the task of privacy protection and draw lines of future research directions to incorporate indexing and summarization as tools for privacy protection by design. Section V concludes our paper.

II. Video Indexing and Summarization Techniques

Summarizing consist in producing a compact representation of a given content so as to provide the user with an idea of the content in a short period of time. Therefore, the usual procedure is to extract the semantic information and represent it in a suitable form. The advent of the digital multimedia era has brought a rich variety of content formats. Associated with the compression and the increasing storage capabilities, the amount of digital multimedia information is rapidly growing. As a consequence, automatic content summarization has attracted the interest of many researchers. Depending on the medium to be summarized, the employed analysis and representation techniques may differ. For multimedia formats involving several kinds of formats, as e.g. video+audio+text, summarization requires the use of multimodal analysis techniques.

Video content has several features, ranging from the colors captured at the individual pixels, over the objects depicted at the successive video frames, to the motion described by the camera capturing the sequence. Moreover, the content of the video sequences is very broad as well, ranging from movies, over news programs, to surveillance videos. In [3], the authors make a distinction between *scripted* and *unscripted* video content. With *scripted* content is meant content which is structured as a series of semantic units as in the case of movies or news. On the contrary, *unscripted* content refers to this type of content which does not follows a predefined structure as in the case of surveillance or sport videos. Depending on the type of content, the techniques employed to extract the semantic information are different; while identifying the changes of scene might be sufficient in order to summarize a news program, this method would not be enough for summarizing a movie and even would fail to summarize a surveillance video sequence. While segmenting the content can be considered a common step towards summarization of scripted video content, the extraction of highlights or relevant information can be considered the equivalent common approach for the case of unscripted content. We denote the process of extracting the relevant information as indexing. Observe that an index can point both to a space-time as well as to a space-lapse-of-time position.

Once the relevant information has been extracted, the next step towards summarization is how to structure and represent the extracted information so as to facilitate the access of the user to the content in a comfortable and efficient manner. Again, depending on the type of content and the application in mind, different kinds of information representation may be more appropriate than others.

In this article we focus on the analysis and representation of surveillance video content. Therefore, we focus on techniques employed in order to extract information out of unstructured video content and to represent it to a user who is potentially carrying out a criminal investigation or needs to rapidly obtain an overview of a certain period of time. We observe three different levels for the extraction of the relevant information:

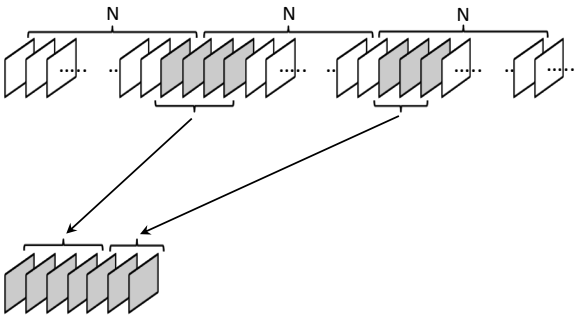
- **Feature** based approaches compute some kind of scoring value based on low-level features as, e.g., number of foreground pixels or frame difference energy, in order to index those frames (or groups of frames) which are supposed to contain the higher amount of information.
- **Object** based approaches look for application-dependent objects of interest as, e.g., persons or cars, and index the frames containing this information.
- **Event** based approaches look for specific events as, e.g., pedestrians crossing the street from left to right or mugging situations, in order to set pointers with a high semantic level.

Event based approaches offer the highest semantic level at the cost of a higher sensitivity to the underlying analysis technique. Therefore, event based approaches are usually more application specific. The more specific the extracted semantic, the more specific the application domain. On the other hand, feature based approaches, which represent the lowest semantic level of analysis, tend to be more application independent but, in the most trivial case, they can only differentiate between segments of activity and segments of inactivity.

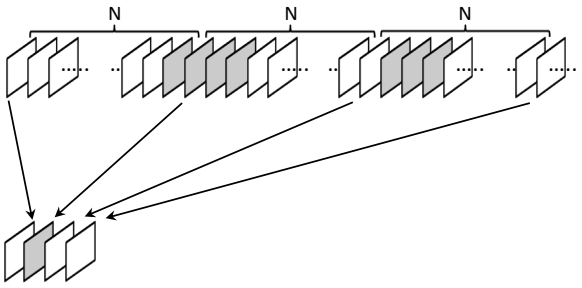
Regarding the representation, we observe three different levels of abstraction:

- **Key frame** based representation relies on the selection of specially relevant frames to depict the content of the whole sequence.
- **Frame-true time compressed video** techniques provide shortened or accelerated versions of the most relevant segments of the whole sequence by selecting a set of frames from the original sequence. Representative for this type of techniques are video editing [4], fast forwarding, and adaptive fast forwarding [5]. Video editing techniques consist in gluing together the parts of a video sequence containing the most relevant information. Fast forward approaches depict only 1 frame out of every group of N frames, therefore, providing an accelerated version of the original video sequence. A more elaborated version of this last approach is adaptive fast forwarding, consisting in increasing the reproduction speed in less interesting parts of the video while slowing down in the parts of interest. Although the mentioned representation techniques were originally formulated for the multimedia domain, they can be applied as well in the surveillance domain. Figure 2 depicts an exemplary frame selection schedule for these three techniques.
- **Frame-free time compressed video** techniques aim at shortening video sequences by eliminating periods of inactivity and, furthermore, by displacing space segments

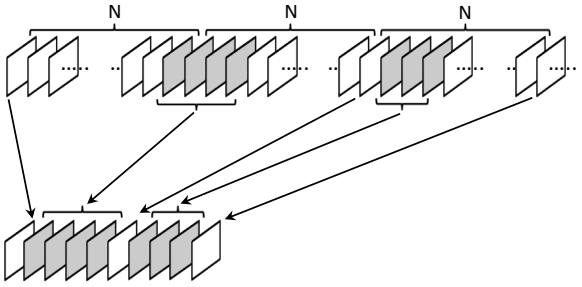
in time so as to present more information at every frame. That means, that some objects may be displaced in space and time and, therefore, represented in other frames that those where they appeared in the original sequence. In this case, the relative timing between activities may change. Examples for this kind of techniques are dynamic video synopsis [6], which condenses video sequences by simultaneously showing several actions even if they occurred at different times, and video condensation by ribbon carving [7], where the temporal warping is explicitly controlled so as to permit avoiding a reversal display order of the activities. Figure 3 depicts an exemplary top view of the space-time trajectories found in a sequence and their corresponding space-time assignment by dynamic video synopsis and video condensation by ribbon carving.



(a) Video Editing.

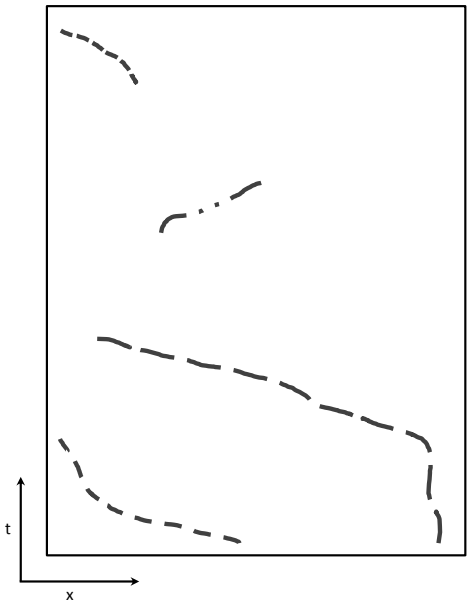


(b) Fast Forwarding.

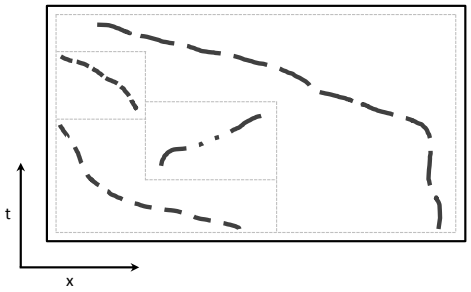


(c) Adaptive Fast Forwarding.

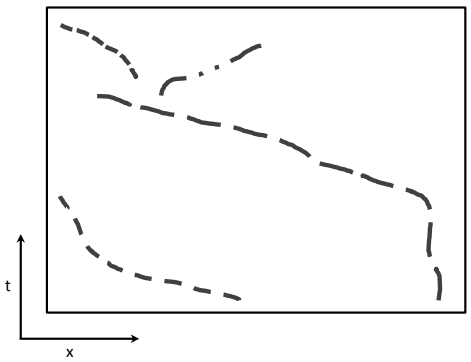
Fig. 2. Frame selection schedule in frame-true video representation. Grey and white are the frames with and without relevant content, respectively.



(a) Top view of space-time object trajectories.



(b) Dynamic video synopsis.



(c) Video condensation by ribbon carving.

Fig. 3. Frame-free video representation techniques.

Key frame based representation techniques allow for the most condensed form of information representation, but contextual information gets lost. Therefore, key frames are often used to provide non-linear access to the segments of video that they represent. Generally, the higher the abstraction, the higher the loss of information. We provide in Table I an overview of the capabilities that the three considered representation techniques allow considering the overall usability of the system. We consider four evaluation criteria. 'Information Compact-

TABLE I
COMPARISON OF THREE LEVELS OF ABSTRACTION FOR THE REPRESENTATION OF INFORMATION EXTRACTED FROM SURVEILLANCE VIDEO SEQUENCES.

	Information Compactness	Context Representation	Information Access Flexibility	Indexation Failure Resilience
Key Frames	High	Low	High	Low
True-Frame Time Compression	Medium	High	Medium	Medium
Frame-Free Time Compression	High	Medium (might be confusing)	Low	Low

ness' refers to the number of frames needed to depict the content of the whole video sequence. 'Context Representation' is the capability of the system to depict the context surrounding the represented video content. 'Information Access Flexibility' is the flexibility that the system provides to the user in order to access specific pieces of the whole video sequences. 'Indexation Failure Resilience' represents the capability of the representation system to provide informative summaries as the quality of the generated indexes decreases.

Depending on the level of the performed video analysis and on the application domain, some representation techniques are more appropriate than others. For instance, while low-level features can be successfully used to detect segments of activity of which a set of key frames can be selected for representation, these same features could not be employed for a frame-free time compressed video representation. On the other hand, while a compact representation as the one provided by frame-free representations can be of interest in order to provide a fast overview the set of objects observed at a given location, such a representation would not be advisable for a crime investigation, where the context and objects interrelations are of crucial importance. We present in the following section some representing approaches for the above mentioned levels of analysis and representation so as to provide an overview of the numerous application scenarios and summarization solutions in the current surveillance arena.

III. REPRESENTATIVE SUMMARIZATION APPROACHES

A quite straightforward summarization approach can be found in [8], where Damnjanovic et al. use the energy of the difference between consecutive frames in a video sequence for indexing, assuming that events of interest are associated with a higher energy. Furthermore, the authors propose to use a normalized cut clustering criterion on the similarity matrix between the frames selected by the energy criterion to select frames for a key frame video representation and to build clusters of frames for a video editing based representation.

Cullen et al. present in [9] an approach which is based on the detection of a set of objects of interest, namely boats, cars and people, which are taken as input for a video condensation algorithm able to remove inactive space-time regions by means of ribbon carving as proposed in [7].

In [10], Li et al. present an event based adaptive fast forwarding summarization approach, where frames depicting the defined event of interest, in this case motion with a certain speed and direction in predefined regions of interest, are played at normal speed and the rest of the frames are played accelerated.

A different approach which also provides adaptive fast forwarding has also been presented by Höferlin et al. in [11]. In this paper, the authors propose to adapt the speed of the videos to the temporal information contained in them. To that aim, they compute the temporal information between consecutive video frames by means of the divergence between the distribution of the absolute frame difference and the distribution of the estimated noise.

An example of a frame-free video representation approach is presented by Rav-Acha et al. in [6], where the authors formulate the video synopsis task as an energy minimization problem. They present two approaches. The first one uses a 3D Markov random field, where each node corresponds to a pixel in the 3D volume of the generated synopsis. The second, consists of first detecting moving objects and then performing the minimization on the detected objects. This second approach has the advantage of being much faster. Pritch et al. propose in [12] to improve video synopsis by clustering activities, and displaying together only similar activities.

Li et al. propose in [7] another frame-free video representation approach which explicitly controls the temporal warping. To that aim, they introduce the concept of a ribbon in the space-time video volume, which allows by means of a flex-parameter to find a trade off between the condensation ratio and the anachronism of the events.

Ji et al. present in [13] an approach based on the depiction of the detected moving objects along with their trajectories. To that aim, they first segment the video sequence based on the difference in foreground pixels detected in equally time-separated frames (a time difference of 10 frames is taken), take the last frame of each segment for video representation and depict the corresponding computed object trajectories. Furthermore, the authors propose to synthesize key frames of a video summary in order to provide even more compact representations of a video sequence.

Porikli presents in [14] an object-based video summarization approach for multi-camera networks which aims at changing the camera-oriented videos into an object-oriented structure so as to allow to respond to semantic queries such as the places where a given object was recorded during a certain period of time. Video representation is provided in form of key

frames, which are selected by minimizing the Semi-Hausdorff distance between the selected set of frames and the set of frames contained in the generated object-specific sequences.

In [15], Babaguchi et al. propose a system to summarize video captured by an omnidirectional surveillance camera by means of event based spatio-temporal indexing. The system displays the contents by using a timeline and a spatial map. Furthermore, video summarization can be provided in form of videos depicting the perspective or panoramic projections of the captured video at the times when events of interest were detected. The rest of the video material is cut-off. The reproduction speed of the generated videos can be controlled by the user.

Li et al. [16] present an interesting approach from a theoretical point of view which aims at finding the optimal summary by formulating the problem as a rate-distortion optimization problem, where the rate can be either the temporal or the bit rate, and the distortion is assumed to be introduced by missing frames and should be measured by an appropriate distortion metric. Nevertheless, as the authors show in the experimental evaluation, this summarization system would not be practicable in the reality, since the computational load grows very fast. A formal computational complexity analysis is not provided, but the authors report 3 and 23 seconds to summarize 100 and 200 frame sequences, respectively.

IV. Main Findings and Future Directions

It seems obvious that incorporating a tool for efficiently getting access to the relevant information in a surveillance system brings the additional advantage of protecting the privacy of persons who might have been recorded by a surveillance system but do not have any relation with a given investigation being carried on. In this sense, the more elaborated the semantic queries that the system is able to process, the higher the privacy protection. This is due to the fact that the results produced by the system are more specific. Put in another words, the system is able to filter out more information.

A. Content extraction

In the ideal case, the user should be able to formulate queries based on events of interest. This means, that the system should be able to extract event information. Nevertheless, a problem of event based indexing and summarization systems is that the detection of the events of interest is mostly defined as a binary problem. This results in the lack of a basis for building up a summary in the case of the absence of event detections, whether because the considered events do not happen in the considered video material or because of failure of the algorithm.

Feature based approaches are more robust to the absence of specific events, but are only able to index points of time where relevant events might happen. Therefore, such approaches are more appropriate either for very restricted scenarios, where the detection of some video features provide a high certainty of the existence of an event, or for very generic scenarios, where the extraction of events is not feasible.

Object based approaches are appropriate for scenarios where the definition and identification of an object of interest is possible (as e.g. cars). Moreover, we have presented an approach aiming to provide an object-oriented structure, which could be considered to have strong links to the privacy protection of individuals, as it provides the possibility of generating video summaries based on objects (as e.g. suspicious persons). In a more developed version, the identity of non-suspicious persons appearing in video segments where the followed person has been recorded could be hidden. Nevertheless, despite the huge challenge posed by multi-camera tracking, the question is how to choose the individuals of interest, since the generated queues are associated to individuals, but not to their actions. Therefore, we consider this system rather of theoretical interest.

We found a lack on experimentation on fusing several queues of content extraction. We consider especially promising the combination of information of different nature as, e.g., low-level features with event detections. Collaborative approaches have been already proposed [17]. Nevertheless, these are more oriented to entertainment applications and the content extraction techniques employed there are not applicable in the surveillance domain.

B. Content representation

The suitability of a given video representation form depends on the application context. Generally speaking, frame-true approaches are more appropriate in scenarios where the relations between objects can be of relevance (as e.g. in security scenarios), whereas frame-free approaches may suit better the requirements of applications where the observation of specific objects is the center of interest, but interactions between the observed objects are not expected.

C. Further remarks

Regarding the protection of privacy, Ding and Marchioni present in [18] a very interesting study on the influence of the representation speed for the tasks of object identification and video recognition. Among their results, the authors observed that an increase in display speed has an earlier effect on the object identification than on the video comprehension task, i.e., the speed limit for successfully carrying out the task of object identification is lower than that for video comprehension. This can be explained by the fact that object identification and video comprehension correspond to different cognitive processes. While object identification requires focused attention, video comprehension implies global attention.

This result could be considered as a motivation for using acceleration techniques for video summarization systems aware of privacy protection. In this sense, bringing video reproduction to the speed where video comprehension is still possible, but object identification is hampered, would not only improve the task of forensic video search, but also protect the privacy of individuals recorded in parts of the video previous to the segments of interest.

Finally, having properly indexed the video content, different access rights can be provided to different kinds of users in order to further protect the privacy of the individuals being depicted in the recorded video material.

V. Conclusions

In this paper we have provided a thorough review of existing video indexing and summarization techniques. Thereby, we have highlighted the strengths and weaknesses that the presented techniques show for different surveillance scenarios. In this sense, this paper can be used as a guide in order to design a suitable summarization system for a given application. Special attention has been paid to the privacy protection abilities provided by some of the presented approaches. Furthermore, we have pointed out some further design opportunities towards developing efficient and robust summarization tools for the domain of security surveillance applications.

Acknowledgment

The research leading to these results has received funding from the European Community FP7 under grant agreement number 261776 (MOSAIC).

References

[1] M. A. Sasse, "Not seeing the crime for the cameras?" *Communications of the ACM*, vol. 53, no. 2, pp. 22–25, Feb. 2010. [Online]. Available: http://doi.acm.org/10.1145/1646353.1646363

[2] R. Gray, 2008. [Online]. Available: http://www.telegraph.co.uk/news/uknews/2571041/How-Big-Brother-watches-your-every-move.html

[3] Z. Xiong, Y. Rui, R. Radhakrishnanz, A. Divakaranz, and T. S. Huang, *A Unified Framework for Video Summarization, Browsing and Retrieval*, 2nd ed. Academic Press, 2005, ch. 9.2, in The Image and Video Porcessing Handbook.

[4] M. A. Smith and T. Kanade, "Video skimming for quick browsing based on audio and image characterization," Tech. Rep., 1995.

[5] N. Petrovic, N. Jojic, and T. S. Huang, "Adaptive video fast forward," *Multimedia Tools Appl.*, vol. 26, no. 3, pp. 327–344, Aug. 2005. [Online]. Available: http://dx.doi.org/10.1007/s11042-005-0895-9

[6] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short: Dynamic video synopsis," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, June 2006, pp. 435–441.

[7] Z. Li, P. Ishwar, and J. Konrad, "Video condensation by ribbon carving," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2572 – 2583, nov. 2009.

[8] U. Damnjanovic, V. Fernandez, E. Izquierdo, and J. Martinez, "Event detection and clustering for surveillance video summarization," in *Proceedings of the Ninth International Workshop on Image Analysis for Multimedia Interactive Services, 2008. WIAMIS '08*, may 2008, pp. 63 –66.

[9] D. Cullen, J. Konrad, and T. Little, "Detection and summarization of salient events in coastal environments," in *Proceedings of the IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, sept. 2012, pp. 7 –12.

[10] J. Li, S. Nikolov, C. Benton, and N. Scott-Samuel, "Adaptive summarisation of surveillance video sequences," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2007)*, sept. 2007, pp. 546 –551.

[11] B. Höferlin, M. Höferlin, D. Weiskopf, and G. Heidemann, "Information-based adaptive fast-forward for visual surveillance," *Multimedia Tools and Applications*, vol. 55, no. 1, pp. 127–150, 2011. [Online]. Available: http://doi.acm.org/10.1007/s11042-010-0606-z

[12] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg, "Clustered synopsis of surveillance video," in *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'09)*, Genoa, Italy, Sept. 2-4 2009.

[13] Z. Ji, Y. Su, R. Qian, and J. Ma, "Surveillance video summarization based on moving object detection and trajectory extraction," in *Signal Processing Systems (ICSPS), 2010 2nd International Conference on*, vol. 2, july 2010, pp. V2–250 –V2–253.

[14] F. Porikli, "Multi-camera surveillance: Object-based summarization approach," 2004.

[15] N. Babaguchi, Y. Fujimoto, K. Yamazawa, and N. Yokoya, "A system for visualization and summarization of omnidirectional surveillance video," in *Proceedings of the 8th International Workshop on Multimedia Information Systems (MIS2002)*, Tempe AZ, Oct. 2002, pp. 18–27.

[16] Z. Li, G. Schuster, and A. Katsaggelos, "Minmax optimal video summarization," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 10, pp. 1245 – 1256, oct. 2005.

[17] E. Dumont, B. Merialdo, S. Essid, W. Bailer, H. Rehatschek, D. Byrne, H. Bredin, N. E. O'Connor, G. J. Jones, A. F. Smeaton, M. Haller, A. Krutz, T. Sikora, and T. Piatrik, "Rushes video summarization using a collaborative approach," in *TRECVID BBC Rushes Summarization Workshop (TVS 2008) at ACM Multimedia 2008, Vancouver, Canada*, A. S. W. K. ACM, Paul Over, Ed., ACM. Vancouver, BC, Canada: National Institute of Standards and Technology (NIST), Washington, DC, USA, Oct. 2008, pp. 90–94, iSBN 978-1-60558-303-7.

[18] W. Ding and G. Marchionini, "A study on video browsing strategies," University of Maryland, College Park, Tech. Rep. CLIS-TR-97-06, 1997. [Online]. Available: http://hcil2.cs.umd.edu/trs/97-11/97-11.html