Video Synopsis for IR Imagery Considering Video as a 3D Data Cuboid

Nikhil Kumar, Ashish Kumar and Neeta Kandpal

Abstract Video synopsis is a way to transform a recorded video into a temporal compact representation. Surveillance videos generally contain huge amount of recorded data as there are a lot of inherent spatio-temporal redundancies in the form of segments having no activities; browsing and retrieval of such huge data has always remained an inconvenient job. We present an approach to video synopsis for IR imagery in which considered video is mapped into a temporal compact and chronologically analogous way by removing these inherent spatio-temporal redundancies significantly. A group of frames of video sequence is taken to form a 3D data cuboid with X, Y and T axes, this cuboid is re-represented as stack of contiguous $X-T$ slices. With the help of Canny's edge detection and Hough transform-based line detection, contents of these slices are analysed and segments having spatio-temporal redundancy are eliminated. Hence, recorded video is dynamically summarized on the basis of its content.

Keywords Video synopsis · Video summarization · IR · MWIR · Canny's edge detection · Hough transform-based line detection · Spatio-temporal redundancy

1 Introduction

Popularity of thermal imaging systems in surveillance technology has drawn a lot of attention from vision community in the past few decades. Increasing population of such systems is generating vast amount of data in the form of recorded videos; with help of video summarization a compact but informative representation of video

N. Kumar (✉) · A. Kumar · N. Kandpal
Instruments Research and Development Establishment, Dehradun, India
e-mail: nikhilkumar@irde.drdo.in

A. Kumar
e-mail: ashishkumar@irde.drdo.in

N. Kandpal
e-mail: neeta@irde.drdo.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 459,
DOI 10.1007/978-981-10-2104-6_21

sequence may be provided. Since for surveillance purpose timing information of events is important, chronology of events is also maintained in compact representation. Generally, IR (infra-red) signatures of targets are more prominent than background and clutter; this contrast is commonly used as a clue for change detection. We have also decided contrast-based clue for detecting representative segments with motion but in place of processing video sequence in $X - Y$ plane, we have chosen $X - T$ plane. Spatio-temporal regularity [1] is utilized for labelling representative segments with motion.

2 Related Work

The goal of this section is to review and classify the state-of-the-art video synopsis generation methods and identify new trends. Our aim is to extract information from unscripted and unstructured data obtained from recorder of surveillance system. Ding [2] categorized video synopsis techniques in the following three levels:

- Feature-Based Extraction: In such approaches low level features like number of foreground pixels and distance between histograms are used to identify frames with higher information content.
- Object-Based Extraction: In such approaches objects of interest like vehicle, pedestrian are used for labelling frames with higher information content.
- Event-Based Extraction: In such approaches events like entrance of a vehicle, pedestrian in field of view are used for setting pointers with high semantic level. Such approaches are more application specific.

Li et al. [3] presented an optical flow based approach for surveillance video summarization. It is a motion analysis-based video skimming scheme in which play-back speed depends upon motion behaviour.

Ji et al. [4] presented an approach based on motion detection and trajectory extraction. Video is segmented based on the moving objects detection and trajectories are extracted from each moving object. Then, only key frames along with the trajectories are selected to represent the video summarization.

Cullen et al. [5] presented an approach to detect boats, cars and people at coastal area. For this, the region of interest is decided and validated. It is taken as input for video condensation algorithm to remove inactive time space.

Rav-Acha et al. [6] presented a method for dynamic video synopsis in which several activities were compressed into a shorter time, where the density of activities were much higher. For better summarization of video event, chronology is not maintained as several events are merged in few frames.

Petrovic et al. [7] presented an approach for adaptive video fast forward. A likelihood function based upon content of video is formulated and playback speed is modelled accordingly.

Hoferlin et al. [8] presented an information based adaptive fast forward approach in which the playback speed depends on the density of temporal information in the

video. The temporal information between two frames is computed by the divergence between the absolute frame difference and noise distribution.

Porikli [9] presented multiple camera surveillance and tracking system based on object based summarization approach. For this, only the video sequence for each object is stored in place of storing video for each camera. Then, object is tracked by background subtraction and mean shift analysis.

Most of the approaches discussed above rely on motion detection-based techniques in $X - Y$ plane for video summarization but in case of IR sequences with poor SNR and targets limited in very small fraction of $X - Y$ plane it becomes challenging to detect targets, to tackle with such scenarios a novel approach of video summerization is presented in subsequent sections. In place of detecting targets in $X - Y$ plane, trajectory of motion is extracted from $X - T$ slices which covers a relatively larger fraction of $X - T$ slice.

3 Methodology

3.1 Overview of the Approach

Problem of video synopsis can be defined as a mapping generation problem between a video sequence and its temporal compact version. In the present approach for mapping generation, considered video sequence is analysed in $X - T$ plane and spatio-temporal redundant segments are eliminated. Trajectory of moving objects is utilized for this job. First a video sequence is represented as a 3D data cuboid V_{XYT} then this cuboid is chopped in contiguous $X - T$ slices and a set of slices $I(y)_{XT}$ is generated. Set of binary edge maps $E_{XT}(y)$ is obtained from $I(y)_{XT}$ with help of Canny's edge detection. A consolidated edge map ξ_{XT} is generated by registration of all elements of $E_{XT}(y)$. Using Hough transform-based line detection with a number of constraints representative segments with motion are labelled in ξ'_{XT} and ζ_{XT} is generated where ξ'_{XT} is binary edge map obtained from ξ_{XT}. For transformation from V_{XYT} to Ψ_{XYT} a transformation matrix τ_T is needed which is extracted from τ_{XT} where τ_{XT} is formed by subtracting ξ'_{XT} from ζ_{XT} and Ψ_{XYT} is 3D data cuboid representation of temporal compact version of video sequence.

3.2 Data Cuboid Representation of a Video Sequence

As in Fig. 1a, b group of frames of video sequence is taken to form a 3D data cuboid V_{XYT} with X, Y and T as axes. V_{XYT} can be expressed [10, 11] as following:

$$V_{XYT} = \left\{ I(t)_{XY}, \forall t \in \{1, 2, 3, \ldots \ldots \ldots \ldots, p\} \right\} \tag{1}$$

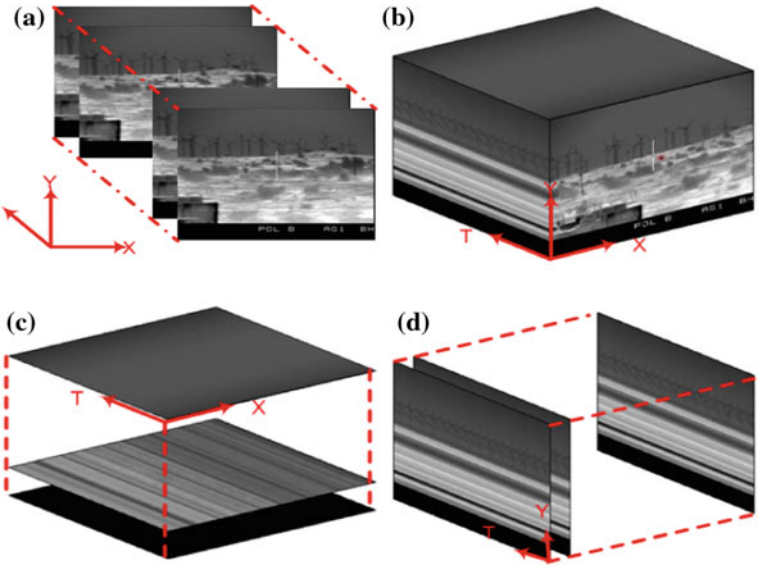

Fig. 1 Data cuboid representation of a video sequence **a** Frames from video sequence *Windmill*, **b** Video sequence *Windmill* represented as a 3D data cuboid, **c** Data cuboid chopped in contiguous $X-T$ slices and **d** Data cuboid chopped in contiguous $Y-T$ slices

Where $I(t)_{XY}$ is a frame of video sequence with X and Y axes at any particular time t and p is number of such frames of size mXn.

As shown in Fig. 1c data cuboid V_{XYT} has an alternative representation [10, 11] as an stack of m number of contiguous $X-T$ slices $I(y)_{XT}$ with size nXp.

$$V_{XYT} = \left\{ I(y)_{XT}, \forall y \in \{1, 2, 3, \ldots\ldots\ldots\ldots, m\} \right\} \tag{2}$$

Yet another way to represent [10, 11] the same data cuboid V_{XYT} is suggested in Fig. 1d by stacking n number of contiguous $Y-T$ slices $I(x)_{YT}$.

$$V_{XYT} = \left\{ I(x)_{YT}, \forall x \in \{1, 2, 3, \ldots\ldots\ldots\ldots, n\} \right\} \tag{3}$$

3.3 Mapping Between Contents of $X-Y$ and $X-T$ Planes of Data Cuboid

If content of $X-Y$ frame is stationary then in $X-T$ slices there will be a number of horizontal features parallel to T axes. Since present approach assumes that video is recorded from a stationary IR system, such horizontal features are most likely content

of $X-T$ slices. Most important conclusion related to present work is that if there are pixels with local motion in $X-Y$ frame then trajectory of motion appears in features of $X-T$ slices having those pixels. Geometry of this trajectory can be approximated by combining a number of inclined line segments. If there is any acceleration in motion, then there will be a number of curves in trajectory but any curve can be approximated by combining a number of small inclined line segments. This fact is utilized for labelling of segments with motion. In Fig. 2, an $X-T$ slice is shown which corresponds to $X-Y$ frame containing stationary as well as moving objects, hence combination of corresponding features is appearing in figure.

Formation of a Set of Binary Edge Maps. From Eq. 2 a set $\{I(y)_{XT}, \forall y \in \{1, 2, 3, ..., m\}\}$ is obtained from V_{XYT}. In this section we obtain $E_{XT}(y)$ from $\{I(y)_{XT}, \forall y \in \{1, 2, 3, ..., m\}\}$ using Canny's edge detection [12], which is one of the most widely used edge detection algorithms. Even though it is quite old, it has become one of the standard edge detection methods and it is still used in research [13]. Canny redefined edge detection problem as a signal processing optimization problem and defined an objective function with following

- Detection: Probability of detecting real edge points should be maximized while the probability of falsely detecting non-edge points should be minimized.
- Localization: Amount of error between detected edge and real edge should be minimum.
- Number of responses: For one real edge there should be one detected edge though this point is implicit in first point yet important.

Consolidated Edge Map Generation. Since we are mapping video sequence into a temporal compact representation, information carried along Y axes of V_{XYT} is redundant atleast for labelling of representative segments with motion; therefore for further processing we are using a consolidated edge map formed by utilizing all elements of $E_{XT}(y)$. As $E_{XT}(y)$ is generated from $I(y)_{XT}$ whose elements are contiguous slices, all elements of $E_{XT}(y)$ are already registered in spatial domain; hence consolidated edge map ξ_{XT} of V_{XYT} is generated by using logical OR operation over all elements of $E_{XT}(y)$.

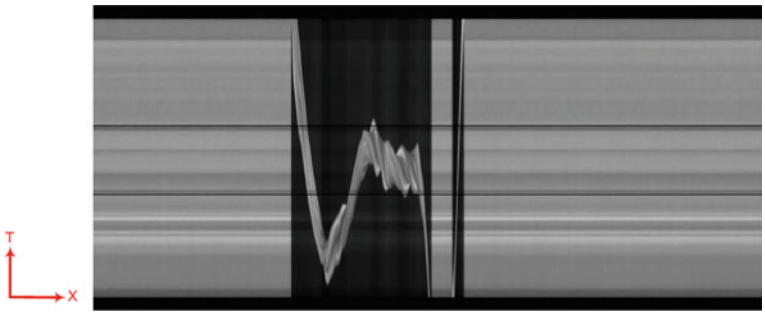

Fig. 2 A typical $X-T$ slice representing features of stationary as well as moving objects in corresponding $X-Y$ plane of video sequence, Moving objects are appearing in curved trajectory

3.4 Extraction of Representative Segments with Motion from Consolidated Binary Edge Map

As discussed earlier, to extract segments having motion we have to extract inclined line segments from $X-T$ slices, hence our goal is to find out set of inclined Y_{XT}. But as $Y_{XT} \subset L_{XT}$ where L_{XT} is set of lines with cardinality r in any $X-T$ slice, elements of Y_{XT} are obtained from L_{XT} with imposed constraints. Hough transform-based line detection is used to find out elements of L_{XT}.

Hough Transform-Based Line Detection. Now we have to explore a set of line segments $L_{XT}(y)$ from binary edge map ξ'_{XT} of consolidated edge map ξ_{XT} which is mathematically a set of points in any V_{XYT}. Hough transform [14] based line detection is a very popular, accurate, easy, and voting-based approach for such kind of operations [15]. Hough transform is based upon line point duality between $X-Y$ and $M-C$ domains, where $y = mx + c$ is equation of line. By quantizing the $M-C$ space appropriately a two-dimensional matrix H is initialized with zeros. A voting-based method is used for finding out elements of H matrix $H(m_i, c_i)$, showing the frequency of edge points corresponding to certain (m, c) values.

Considered Constraints. Following are assumed constraints while implementing Hough transform-based line detection:

- Slope constraint: If $L_{XT} = \{l_{iXT}(y), \forall i \in \{1, 2, 3....r\}\}$ where $l_{iXT}, \forall i \in \{1, 2, 3....r\}$ are line segments with slopes $\{\theta_{iXT}, \forall i \in \{1, 2, 3....r\}\}$ in any $I_{XT}(y), \forall y \in \{1, 2, 3....m\}\}$ then $l_{iXT} \in Y_{XT}$ if $\theta_{low} < \theta_{iXT} < \theta_{high}, \forall i \in \{1, 2, 3....r\}$
  Where θ_{high} is dependent upon global motion in $I(t)_{XY} \forall t \in \{1, 2, 3...., p\}$ and θ_{low} is dependent upon velocity of moving object and clutter in scene.
- Maximum Length constraint: In present approach we are using Hough transform based line detection for labelling representative segments having motion, so few constraints have been imposed on this method. It will increase temporal redundancy if an object is with motion with similar pose is part of for more than few frames. Since ξ_{XT} is generating transformation matrix between V_{XYT} and Ψ_{XYT}, inclined line segments of a fixed slope with more than a threshold length are replaced with inclined line segments of a fixed slope with threshold length. By setting an upper threshold on H matrix of Hough transform line segments more than certain length can be avoided.
- Minimum Length constraint: As it is obvious in real time scenarios that there will be a substantial amount of clutter available in captured scenes in form of unwanted motions due to various causes, e.g., motion in leaves due to wind, it becomes necessary to tackle such scenarios for robustness of proposed approach. By analysing such unwanted motions we can conclude that such motions will also generate incline trajectories in $X-T$ slices but shorter in length, hence by selecting a lower length threshold in H matrix of Hough transform these can be eliminated.

3.5 Labelling of Representative Segments with Motion

From Eq. 4 set of representative segments with motion τ_{XT} is difference of ζ_{XT} as in Fig. 3b and ξ'_{XT} as in Fig. 3a.

$$\tau_{XT} = \zeta_{XT} \quad - \quad \xi'_{XT} \tag{4}$$

3.6 Extraction of Representative Segments with Motion

A sparse set τ_T is generated from τ_{XT} with unity entries corresponding to frame numbers with representative motion segments. This set is used as transformation matrix for obtaining Ψ_{XYT} from V_{XYT}.

4 Results

Results of video synopsis along with video sequences are presented based on our approach on two datasets. As there are very limited datasets available for such sequences, we have tried to generate a robust test bed of thermal imaging sequences captured in different environmental conditions using a 640×512 detecting elements-based MWIR imaging system. Number of frames in temporal compact representation are dependent upon motion content of considered video sequence. There are also few false alarms in form of frames containing no motion information, due to outlier line segments during Hough transform-based line detection. As in Fig. 4a there is *Room* dataset containing an IR video sequence of 1393 frames out of which 939 frames contain object(s) with motion, in its temporal summarized representation there are 525 frames out of which 55 frames are false positives; this implies that we are getting almost 2.65 times compressed sequence. There are three randomly moving persons in this sequence, it can be concluded that almost all important events related with motion are captured in its compact representation. Analysis is also done for *Road* dataset containing a MWIR video sequence *Road*-2 of 2440 frames out of which 1425 frames contain object(s) with motion, it is transformed in a compact representation containing 1084 frames with almost 2.25 times compression out of which 243 frames are false positives. Similarly as in Fig. 4b for *Road*-1 sequence containing a thermal video of 823 frames we are getting almost two times temporal compact representation with 411 frames.

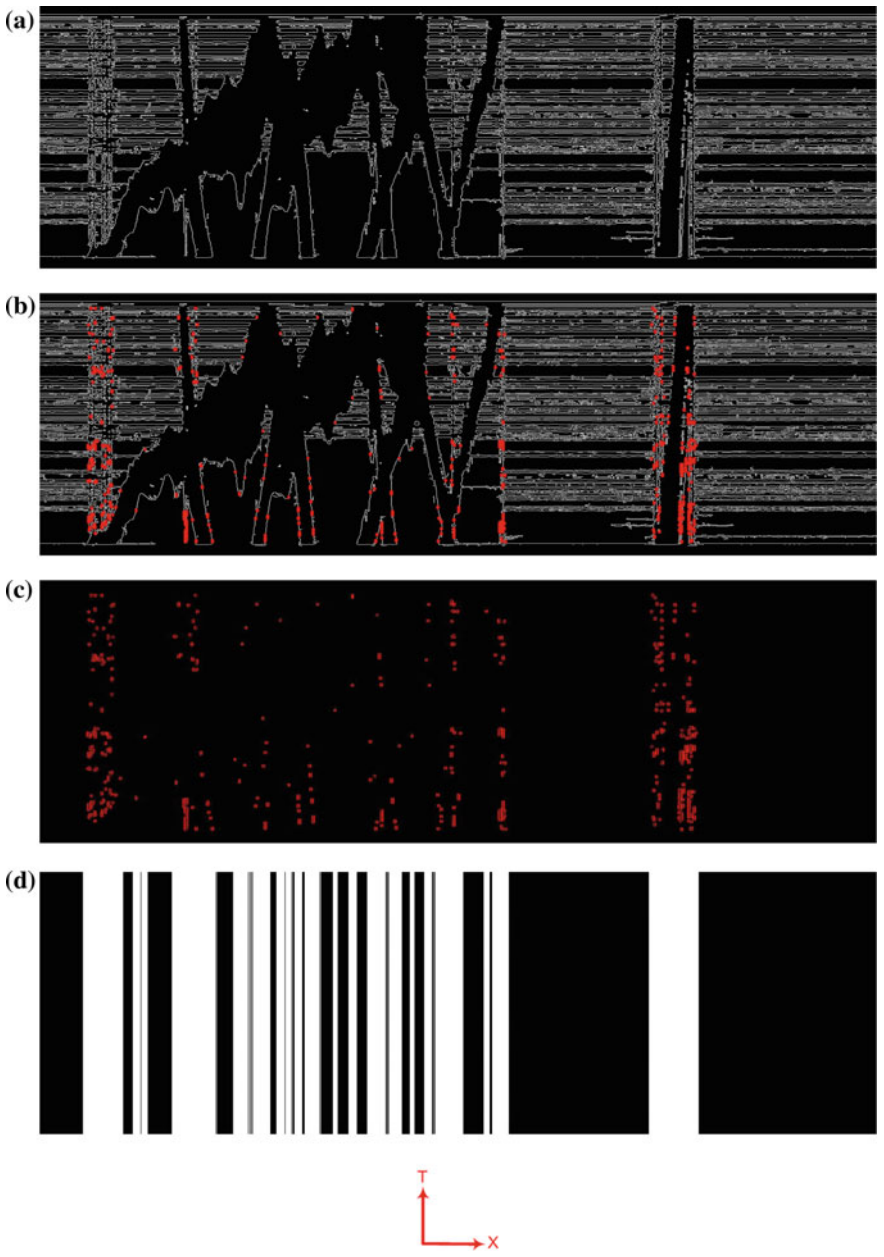

Fig. 3 For *Room* video sequence of 1393 frames X axes representing T (number of frames) and Y axes representing X **a** ξ'_{XT} Binary edge map obtained from Canny's edge detection of consolidated binary edge map ξ_{XT}, **b** ζ_{XT} Result of Hough-based line with imposed constraints (in *red*), **c** τ_{XT} representative segments with motion (in *red*) and **d** Selected frame nos. for compact representation (in *white*)

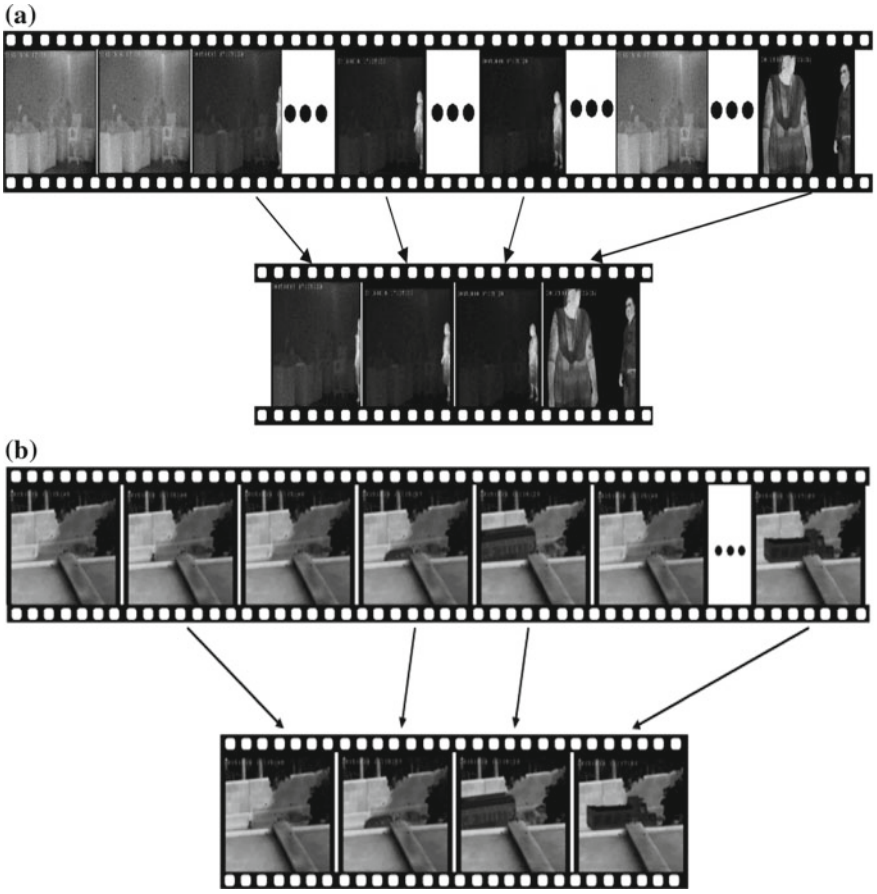

Fig. 4 For both figures **a** and **b** sequence shown above is considered video and sequence shown below is temporal compact representation of considered video **a** Synopsis Generation for *Room* dataset: In its compact representation, there are four non-contiguous frames, first frame corresponds to entrance of person-1, second and third frames correspond to pose change and fourth frame corresponds to entrance of person-2 and **b** Synopsis Generation for *Road*-1 sequence: In its compact representation, there are four non-contiguous frames representing entrance of pedestrian, car, bus, and truck, respectively

5 Limitations

Although we have obtained very promising results from present approach there are certain limitations. As Hough transform is a voting-based mechanism for detecting geometries from a set of points and we are using it with some imposed constraints, hence it is obvious that there will be a number of outliers and missing segments. When transformation matrix is generated using these outliers then there are a few frames which unnecessarily become part of temporal compact representation Ψ_{XYT}

and hence we are unable to completely eliminate spatio-temporal redundancy. On the other hand, if the missing segments are part of τ_{XT} then few of important events may be missing from Ψ_{XYT}. Number of such outlier or missing segments can be reduced by adjusting upper and lower thresholds of Canny's edge detection.

6 Conclusion

We considered a novel approach for video synopsis in IR imagery. Although there are a number of approaches suggested in literature yet Hough transform based line detection has barely been used to solve such kind of problems. We are making use of Canny's edge detection and Hough transform based line detection, fortunately both are very old and well established algorithms. This makes implementation aspect of present model very simple. The results are promising barring limitations and model is extremely simple.

Acknowledgements We take this opportunity to express our sincere gratitude to Dr. S.S. Negi, OS and Sc 'H', Director, IRDE, Dehradun for his encouragement. As good things cannot proceed without good company, we would like to thank Mrs Meenakshi Massey, Sc 'C' for not only bearing with us and our problems but also for her support in generating datasets.

References

1. Alatas, Orkun, Pingkun Yan, and Mubarak Shah. "Spatiotemporal regularity flow (SPREF): Its Estimation and applications." Circuits and Systems for Video Technology, IEEE Transactions on 17.5 (2007): 584–589.
2. Ding, Wei, and Gary Marchionini. "A study on video browsing strategies." (1998).
3. Li, Jian, et al. "Adaptive summarisation of surveillance video sequences." Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on. IEEE, 2007.
4. Ji, Zhong, et al. "Surveillance video summarization based on moving object detection and trajectory extraction." Signal Processing Systems (ICSPS), 2010 2nd International Conference on. Vol. 2. IEEE, 2010.
5. Cullen, Daniel, Janusz Konrad, and T. D. C. Little. "Detection and summarization of salient events in coastal environments." Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on. IEEE, 2012.
6. Rav-Acha, Alex, Yael Pritch, and Shmuel Peleg. "Making a long video short: Dynamic video synopsis." Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 1. IEEE, 2006.
7. Petrovic, Nemanja, Nebojsa Jojic, and Thomas S. Huang. "Adaptive video fast forward." Multimedia Tools and Applications 26.3 (2005): 327–344.
8. Hoferlin, Benjamin, et al. "Information-based adaptive fast-forward for visual surveillance." Multimedia Tools and Applications 55.1 (2011): 127–150.
9. Porikli, Fatih. "Multi-camera surveillance: object-based summarization approach." Mitsubishi Electric Research Laboratories, Inc., https://www.merl.com/reports/docs/TR2003-145.pdf (Mar. 2004) (2004).

10. Paul, Manoranjan, and Weisi Lin. “Efficient video coding considering a video as a 3D data cube.” Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on. IEEE, 2011.
11. Liu, Anmin, et al. “Optimal compression plane for efficient video coding.” Image Processing, IEEE Transactions on 20.10 (2011): 2788–2799.
12. Canny, John. “A computational approach to edge detection.” Pattern Analysis and Machine Intelligence, IEEE Transactions on 6 (1986): 679–698.
13. Azernikov, Sergei. “Sweeping solids on manifolds.” Proceedings of the 2008 ACM symposium on Solid and physical modeling. ACM, 2008.
14. VC, Hough Paul. “Method and means for recognizing complex patterns.” U.S. Patent No. 3,069,654. 18 Dec. 1962.
15. Illingworth, John, and Josef Kittler. “A survey of the Hough transform.” Computer vision, graphics, and image processing 44.1 (1988): 87–116.