2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance

Multiple Cue Indexing and Summarization of Surveillance Video

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

Abstract

In this paper we propose a system for the summarization of safety and security surveillance video. By combining the information provided by multiple analysis cues, we improve the quality of the information extracted out of the analyzed video sequences with respect to the state-of-the-art approaches, therefore, being able to generate summaries that better align with the content of the original video. The proposed system has been tested using an extensive set of surveillance sequences, showing compression ratios ranging from 11 to 114, depending on the video content and the configuration of the system.

1. Introduction

Video summarization is a process which aims at providing the user with an overview of the content of a video. To that aim, it is necessary to find the relevant information contained in the video to be summarized, and to properly represent it in order to allow the user to rapidly grasp the extracted information and to navigate through it. Depending on the type of video content being analyzed, the techniques used to that aims may differ. In [14], 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 sports videos. In this paper, we focus on the extraction and representation of the information contained in unscripted content, namely surveillance videos.

Extracting the relevant information can be done by means of low-level features, objects or events of interest. Low-level feature based approaches compute some kind of scoring value based on features such as the energy of the difference frame between consecutive frames as in [2], or more elaborated temporal information density measures as proposed in [7]. Object based approaches look for application-specific objects of interest such as persons, cars or boats, as in, e.g., [1]. Finally, event based approaches look for predefined events of interest as, e.g., targets moving in a given direction as in [8]. Event based approaches provide the highest semantic level. Nevertheless, the summaries provided based on this kind of information are very sensitive to the quality of the performed analysis. On the absence of event detections, either because the searched events do not happen in the considered video material or because of failure of the event detection algorithm, there is no basis for building up a summary. Furthermore, it is often the case that there is little information on a given event, which needs to be investigated. This requires the inspection of large hours of video data. In these cases, low-level features based approaches may be useful in driving the user to the potential points of interest. Surveys on state-of-the-art video summarization approaches can be found in [5, 10]. Although the one presented in [10] is actually focused on multimedia, it still provides some ideas which can be assimilated in the surveillance domain.

Information representation can be done by means of key frames or shortened video sequences, which can be generated by means of video edition [13] and video acceleration [11], or by means of displacing objects in time and eliminating periods of inactivity [9, 12]. Depending on the application of interest, a kind of representation might be more appropriate than other. Key frame representations provide a very compact representation of the information but are not able to depict the context. Shortened video sequences are more indicated when context should be considered in order to assess the depicted scenes, but their more compact versions, which introduce object displacements in time, might result confusing in environments where interactions between objects are expected.

One of the common issues, which is easy to observe in the state-of-the-art summarization approaches is that the information of interest is extracted by means of a unique level of analysis, i.e., either low-level feature extraction or mid-level object detection/classification or high-level event detection. Therefore, the quality of the generated summaries

978-1-4799-0703-8/13/$31.00 ©2013 IEEE

is limited by the kind of the analysis tool used. In this paper, we propose a system that combines multiple cues of different kinds of analysis. Therefore, we introduce a diversity factor in the content gathering process, which turns out in summaries that better align with the content of the original video sequences. The proposed system is described in Section 2. In Section 3 we present experimental results. Section 4 concludes this paper.

2. Video Summarization

The proposed system provides indexes and summaries for security video investigations. Therefore, it is important to preserve the context surrounding the objects and events of interest in the generated summaries. Furthermore, the system should be easy to operate by a non-expert user and provide a flexible and rapid access to the gathered information. To that aim, we provide both non-linear access to the segments of the input videos containing events of interest, and accelerated versions of the original videos, which speed is adapted to their content. Video segments containing few relevant information are displayed at a high speed, while those with important content at a lower. The speed of the generated videos is computed by combining multiple video analysis cues. Figure 1 provides an overview of the proposed system. The input video is analyzed by several kinds of analysis tools which, respectively, generate an index and compute an associated speed according to their detections. The speed v_t of the generated summary at time t is computed as the minimum of the set of speeds $V_t = \{v_{c,t}\}_{c=1}^{C}$, where C is the number of cues used. We use the variable t to refer to discrete points in time associated to the consecutive frames of the analyzed video sequence and are, therefore, meant to be members of the set of natural numbers excluding zero ($\mathbb{N}^+$). The generated indexes can be used both for providing non-linear access to the set of events detected and for generating additional summaries according to different combinations of analyses and their respective computed associated speed.

In this paper we assume that the input video has been recorded by fixed cameras and demonstrate the proposed system by combining two video analysis cues: one provided by a dynamic foreground analyzer and the other by a new static objects detector. The dynamic foreground analyzer computes an associated speed $v_{f,t}$ based on low-level features extracted by means of background subtraction. The events triggered by the new static objects detector are used in order to compute an associated speed $v_{s,t}$ on an event basis. Therefore, the system combines two different levels of video analysis. Furthermore, the maximal speed of the generated summary is limited by v_{max}. The speed of the generated video is computed as:

$$v_t = min\{v_{f,t}, v_{s,t}, v_{max}\} \quad (1)$$

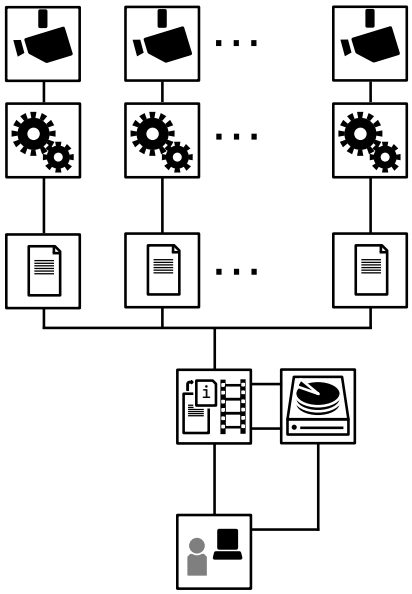

Figure 1. Overview of the proposed system.

2.1. Low-level Features Analysis

The dynamic foreground analyzer takes for every frame the foreground mask corresponding to the input frame and computes an associated speed $v_{f,t}$ based on the absolute difference of the portion of foreground pixels $F_{diff,t}$ between consecutive frames.

We use this cue in order to rapidly direct the user to those parts of the video where the dynamics of the scene change. Thereby, we assume that dynamics changes are more relevant from a summarization point of view than the amount of foreground pixels itself. This can be intuitively illustrated by using the example of a crowded commercial street, where there is a large amount of moving objects, which, nevertheless, do not reveal any relevant information for a summarization system. On the contrary, the entrance of a single moving object into an empty scene can be considered as relevant. Figure 2 depicts graphically the analysis of the foreground masks obtained for a sequence reproducing this last example. In the analyzed sequence, a person enters an empty room at frame number 900, remains staying in the foreground for a while and then leaves the room again at frame number 4200. It is easy to see that the profile obtained by considering the difference of the portion of foreground (bottom) can be used to efficiently bring the user to the events of entering and leaving the room, while conveniently accelerating the rest of the sequence.

The foreground masks are obtained by means of background subtraction. We use a Gaussian mixture model as described in [4], which is able to autonomously choose an

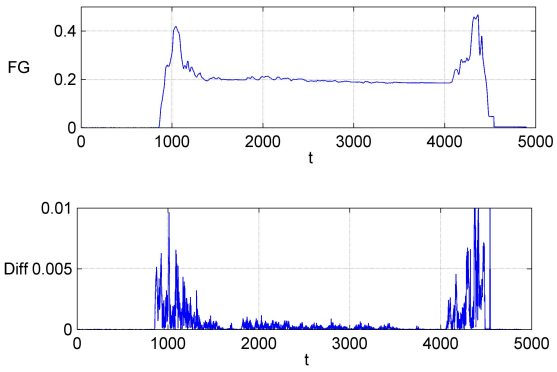

Figure 2. Analysis of the foreground masks for an exemplary sequence. Top: Foreground portion. Bottom: Difference of foreground portion.

initialization value for the variance of new created modes, therefore better adapting to the characteristics of the scene. For every frame, we compute the amount of foreground pixels normalized to the size of the frame $\bar{F}_t$ and follow the difference of this value $\bar{F}_{diff,t} = \bar{F}_t - \bar{F}_{t-1}$ along the sequence. After processing each frame, a scaled version of $\bar{F}_{diff,t}$ is added to the score value D_t, which triggers a frame marker when $D_t > 1$. D_t is computed as:

$$D_t = \alpha D_{t-1} + \beta \bar{F}_{diff,t} \tag{2}$$

where $\alpha \leq 1$ is a retaining factor and β is a weighting factor controlling the influence of the foreground difference into the speed of the summary. Upon the triggering of the frame marker, the value of D_t is set to zero.

$v_{f,t}$ can then be easily computed as:

$$v_{f,t} = (t - t_d)v_i \tag{3}$$

where t is the current point in time, t_d is the previous point in time in which a frame marker was triggered by the dynamic foreground analyzer and v_i is the speed of the input video.

In this way, the associated speed to the dynamic foreground analyzer gently adapts to the changes in the dynamic of the scene, associating high acceleration values to the segments of the sequence where the amount of foreground remains stable, while decreasing the acceleration for segments with high differences. By using the score value D_t we indeed filter out noise which can be contained in the foreground masks.

For every time t, the value of $\bar{F}_{diff,t}$ is logged into a file which can be used in order to generate alternative summaries of the analyzed video as we show later.

2.2. High-level Events

The second cue of our proposed system computes an associated speed $v_{s,t}$ based on the events triggered by a new static objects detector. To that aim, we use the system proposed in [6], which analyzes each video frame at two levels: at the pixel level, pixels are classified as background, dynamic foreground or static foreground; static foreground pixels are grouped by means of connectivity and analyzed at the region level in order to classify them as new static objects or uncovered background.

The detection of new static objects is a very important cue in safety and security applications as it advices for the presence of objects which might imperil the security of people in public spaces. Furthermore, by analyzing large archives of security video data, we observed that most of the events of interest were preceded by the occurrence of a new static object as, e.g., a car parked by the subjects committing an offense. Therefore, the speed associated to the static objects detector $v_{s,t}$ is set to a low value for a given number of frames N upon the detection of new static objects, and set to a high value otherwise:

$$v_{s,t} = \begin{cases} a_{s,l} v_i, & \text{for } t_e \leq t < t_e + N,\ \forall e \in \{1 \dots E\} \\ a_{s,h} v_i, & \text{otherwise} \end{cases} \tag{4}$$

where $\{a_{s,l}, a_{s,h}\} \in \mathbb{N}^+$ are the low and high acceleration factors, respectively, t_e is the time of detection of the event e, and E is the total number of events detected.

The speed associated to the static objects detector $v_{s,t}$ on the event of removal of the detected new static objects is also computed as per Equation 4.

Furthermore, the events raised by the occurrence of new static objects are logged into a file containing the number of frame of the detection and the bounding box associated to the object. This log-file can be used in order to provide non-linear access to the segments of the video where the new static objects appear and to generate alternative summaries.

2.3. Further Analysis Cues

Further analysis cues can be easily added to the proposed system by properly defining the associated speed of the video output depending on the performed analysis and feeding this value into the output speed computation as in Eq. 1. For practical reasons (see Section 2.4), the speed associated to each of the analysis cues must result from the multiplication of the input video speed with a natural number other than zero.

2.4. Summary Generation

The information gathered by the proposed system is provided to the user by means of two kinds of representation: a list of the detected events, which provides non-linear access

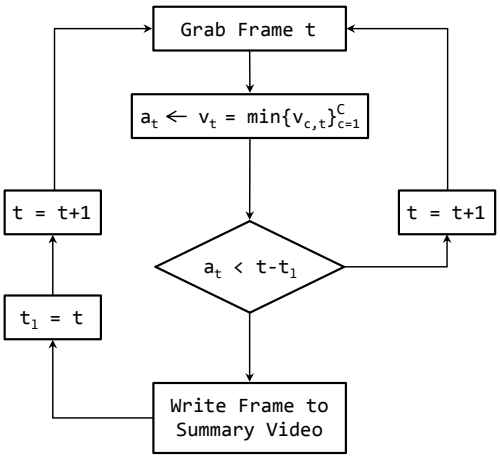

Figure 3. Summary video generation.

Sequence	Frame Nr.	Event Description
AB-Easy	2600	a piece of baggage is abandoned
	4600	abandoned baggage removal
AB-Medium	2250	a piece of baggage is abandoned
	4290	abandoned baggage removal
AB-Hard	2300	a piece of baggage is abandoned
	4450	abandoned baggage removal
library	900	a person enters an empty room
	4200	the person leaves the room
office	600	a person enters an empty office
	2000	the person leaves the office
tramstop	1000	tram starts moving
	1270	tram leaves scene
	1400	an object is abandoned

Table 1. Main events of the summarized sequences.

to the segments of the video containing events of interest, and adaptively accelerated versions of the input videos.

The list of the events of interest (index) is generated by fusing the log-files generated by the individual analysis cues. This list can be visualized in text or in image form, by using the frame at which the event was first detected. Furthermore, the user can filter events by type, time of occurrence, and so on.

The accelerated versions of the input video are generated by using the speeds associated to each of the analysis cues. For each time t, the current output speed is computed as in Eq. 1. This speed is a multiple number of the input video speed, with an acceleration factor $a_t = v_{max}/v_i$, being $a_t \in \mathbb{N}^+$. The summary video generator keeps a register of the point in time corresponding to the last frame recorded t_l. If the difference between time corresponding to the current input frame t and t_l is bigger or equal than a_t, the current frame is recorded into the summary video. If not, it is skipped. Figure 3 depicts graphically the described procedure.

Observe that, although we have described the indexation and the summary video generation separately, these processes can be run together. In fact, the described system has been implemented for online generation of indexes and video summaries immediately afterwards of the video analysis with negligible processing time for the indexing and summary video generation tasks.

Furthermore, by decoupling the tasks of indexing from the summary video generation, custom summaries can be easily generated in order to better fit individual user preferences.

3. Experimental Results

The proposed system has been tested using an extensive set of surveillance sequences comprising both public and private datasets. From the i-Lids dataset for AVSS 2007, we took the abandoned baggage scenario, which consist of three video sequences recorded in a subway station where a piece of baggage is abandoned. Furthermore, several subways arrive and depart from the station, occasionally producing increased flows of passengers on the platform. From the CDnet dataset [3], we took the sequences 'library', 'office' and 'tramstop'. The two first sequences depict scenes in which a person enters an empty room, remains for a while, and then leaves the room. The sequence 'tramstop' depicts a more intricate situation involving the departure of a tram from a stop position and the abandonment of a box on a sidewalk. Table 1 summarizes the most important events of the described sequences and the approximated frame number of their occurrence. The sequences provided by our client depict hours of surveillance video recorded in outdoor environments. The scenes show most of the time people walking and cars driving through. The most relevant events are cars parking in an out and a very reduced set of events as mugging and a housebreak.

The system was configured with the same parameters for all test sequences. The background subtraction system and the static objects detection were configured with the parameters used by the authors in their respective papers. The dynamic foreground analyzer was configured with a retaining factor α equal to one and a weighting factor β equal to 25. That means, we use the difference score D_t as a pure accumulator. For the cue associated to the new static objects detector we set the low acceleration factor $a_{s,l}$ to one and the high acceleration factor $a_{s,h}$ to 32.

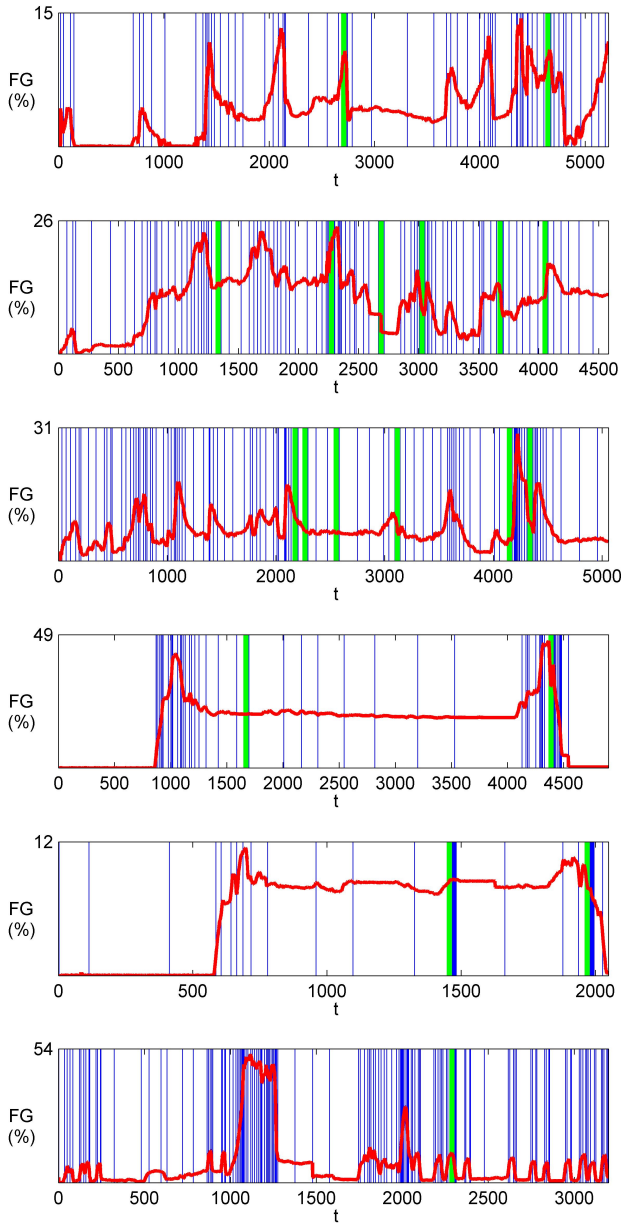

Figure 4. Analysis results for the summarization of the sequences iLids AVSS2007 AB-Easy, AB-Medium and AB-Hard, and CD-net library, office and tramstop. Blue: frames to be added to the video summary. Green: detected events. Red: percentage of pixels classified as foreground.

Figure 4 depicts the analysis results for the summarization of the sequences corresponding to the public datasets. The blue vertical lines correspond to the frames of the input video that were recorded in the summary video. Therefore, segments of time with a high density of blue lines correspond to low accelerated parts of the summary video, while segments with a low density correspond to high accelerated

Sequence	Compression Rate		
	standard	no v_{max}	only events
AB-Easy	23.4170	43.8824	96.7037
AB-Medium	14.9251	18.4758	29
AB-Hard	15.2840	19.3092	32.0190
library	21.6770	46.6571	90.7222
office	17.6638	28.8592	37.9444
tramstop	15.0896	18.4913	114.2500
priv-01	17.5185	26.7819	37.7566
priv-02	14.1031	17.9504	25.2680
priv-03	19.0496	26.4841	67.4032
priv-04	11.6267	14.4018	17.7608

Table 2. Compression rate of the generated summary videos for the test sequences by using three different configurations.

ones. For the sake of depiction clarity, we have depicted only the frames triggered by the foreground analyzer and by the new static objects detector, but not those by v_{max}. The green vertical lines correspond to the detected events. The red curve represents the portion of foreground pixels that were detected for each input frame. It is easy to appreciate that non-complex scenes, which are indeed very usual in the security surveillance, as the 'library' and the 'office' sequences can be very accurately segmented by means of low-level features. In fact, we were able to decelerate the generated summary at the events of a person entering the room thanks to this analysis cue. On the other hand, more involved sequences as the i-Lids and the 'tramstop' sequences needed the information provided by the new static objects detector cue in order to decelerate the video at the segments containing the events of interest. Furthermore, it is also easy to observe that sequences with a higher level of semantic information also show higher differences in the amount of foreground and are, therefore, summarized with a higher number of frames. In overall, it can be said that by combining analysis cues of different types increases the diversity of the information gathering process, which leads to a better alignment of the generated video summaries with the content of the original video.

A very useful functionality of our proposed system is that, thanks to the explicit decoupling of the indexing and summarization tasks, customized summary videos can be easily generated and displayed. In fact, based on the indexes generated by the individual video analysis tools, the reproduction speed of the analyzed videos can be computed online. Furthermore, the user can preview the amount of time needed for watching a summary generated by a given configuration. In this way, the more appropriate configuration, given the length of the video and the time available to the video operator, can be chosen. Table 2 shows the compres-

sion rates, computed as the number of frames in the generated summary video divided by the number of frames in the input video, achieved for the whole set of sequences by using different summarization configurations. The configurations used are 'standard', which is the one explained in this paper, 'no v_{max}', which corresponds to the speed computed by both analysis cues without using an upper limit, and 'only events', which corresponds to the summary video generated by using only the events cue. Sequences with a higher visual semantic content as, e.g., 'tramstop' achieve lower compression rates than sequences with a lower content as, e.g., 'library'.

4. Conclusions

In this paper we have proposed a system for the summarization of surveillance video in the context of safety and security applications. The system is able to combine multiple video analysis cues, therefore accounting with a richer amount of information, which is used to generate indexes and video summaries that better align with the content of the original video. Furthermore, thanks to the explicit separation of the indexing and the visualization tasks, the system is able to generate on-line customized summaries adapted to the user requirements. The proposed system has been tested using an extensive set of surveillance sequences, showing compression ratios ranging from 11 to 114, depending on the video content and on the configuration of the system.

Acknowledgment

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

References

[1] 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)*, pages 7 –12, sept. 2012.

[2] 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*, pages 63 –66, may 2008.

[3] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar. changedetection.net: A new change detection benchmark dataset. In *Proceedings of the IEEE Workshop on Change Detection (CDW'12) at CVPR'12*, Providence, RI, 16-21 Jun. 2012.

[4] R. Heras Evangelio, M. Pätzold, and T. Sikora. Splitting gaussians in mixture models. In *9th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, Beijing, China, September 2012.

[5] R. Heras Evangelio, T. Senst, I. Keller, and T. Sikora. Video indexing and summarization as a tool for privacy protection. In *IEEE International Conference on Digital Signal Processing (DSP 2013)*, Greece, Santorini, July 2013.

[6] R. Heras Evangelio, T. Senst, and T. Sikora. Detection of static objects for the task of video surveillance. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 534–540, January 2011.

[7] B. Höferlin, M. Höferlin, D. Weiskopf, and G. Heidemann. Information-based adaptive fast-forward for visual surveillance. *Multimedia Tools and Applications*, 55(1):127–150, 2011.

[8] 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)*, pages 546 –551, sept. 2007.

[9] Z. Li, P. Ishwar, and J. Konrad. Video condensation by ribbon carving. *IEEE Transactions on Image Processing*, 18(11):2572 –2583, nov. 2009.

[10] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121 – 143, 2008.

[11] N. Petrovic, N. Jojic, and T. S. Huang. Adaptive video fast forward. *Multimedia Tools Appl.*, 26(3):327–344, Aug. 2005.

[12] 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)*, pages 435–441, June 2006.

[13] M. A. Smith and T. Kanade. Video skimming for quick browsing based on audio and image characterization. Technical report, Carnegie Mellon University, 1995.

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