A user-centric approach for event-driven summarization of surveillance videos

Anastasios Dimou*, Dimitra Matsiki*, Apostolos Axenopoulos*, Petros Daras*

*Information Technologies Institute, Centre for Research and Technology Hellas,Thessaloniki, Greece

Keywords: video summary, key frames, event.

Abstract

In this paper, a user-centric approach for video summarization is introduced. The method produces meaningful video summaries, by fusing low-level visual information, extracted by processing consecutive frames, with high-level information derived from detected events. The video summaries are presented to the user in the form of most representative frames, while an intuitive user interface allows the user to adjust the level of granularity of the presented summaries.

1 Introduction

In recent years, the rapid development of the visual media technology has led to an impressive increase of the produced video data. Surveillance system applications is one of the areas facing this problem. Video surveillance systems typically consist of a high number of cameras with overlapping or not field of views (FOVs) [5]. To avoid exhaustive manual inspection of the vast amounts of video data, which is definitely a time-consuming task, video summarization has been utilized to facilitate browsing and navigation in video repositories, attracting the attention of the research community.

Video summary has been defined in [13] as “a sequence of still or moving images presenting the content of a video in a way that the respective target group is rapidly provided with concise information about the content while the essential message of the original is preserved”. A number of methods and techniques have been proposed for the automatic extraction of video summaries; however, no robust answer has been given to overcome all challenges. Two are the main aspects of video summarization: the selection of frames that can represent the content of the video and the visualization of those frames in a user intuitive manner.

Representative frames, also known as key-frames, are extracted from the video source [15]. A common approach for key-frame selection is clustering. Color histograms or color features are extracted from the video frames; then Delaunay triangulation [10] ,modified hierarchical clustering [8], or line Gaussian Mixture Model clustering [12] are applied to produce clusters. The cluster centroids are selected as keyframes. The authors in [16] present a summarization technique based on robust low-rank subspace segmentation. A series of video frame subspaces are segmented based on the Normalized Cuts algorithm and the key frames are chosen from the significant subspaces.

Different approaches have been also proposed for key-frame selection. In [3], the Heterogeneity Image Patch index is introduced, where the level of heterogeneity of the video frames is measured, and is used to select a set of candidate key frames. In [6], the visual content change of a video sequence is calculated by frame to frame differences using color and edge direction histograms, and wavelets statistics. The final key frames are curvature points in the cumulative frame differences curve. Finally, in [4], inter-frame differences are calculated based on the correlation of RGB colour channels, colour histogram and moments of inertia, and then, an aggregation mechanism is employed to combine these difference measures and to extract key frames.

The visualization of the video summary aims to create an appealing and at the same time informative presentation of the video. In [1], a multi-level storyline visualization method is presented, where a still image is created featuring a number of sub-stories summaries. In [11], the authors present a 3D visualization of a video sequence which represents the video as a space-time cube, using volume rendering techniques. In [14], two visualization methods are proposed for the generation of arbitrary length summaries of large sports video archives: a compressed video clip, and a video poster, as a 2D plane of image key frames. The authors in [9] extract Regions Of Interest (ROIs) from the frames, and arrange them on a given canvas, preserving the temporal structure of the video content. In [2], the authors propose to collect the most significant moving objects to construct a compact video, where the temporal coordinates of the moving objects are rearranged, but the appearing order is preserved.

A common drawback of the existing video summarization systems is that they cannot adapt the amount of image/video content to be presented in the summary to the varying importance of the content itself as well as to the specific user needs. They are based on a number of heuristic assumptions regarding thresholds and content importance, offering no control to the actual viewer. This can result to either information redundancy or omission. While this can be an adequate assumption for some applications, in surveillance systems the user may need different levels of summarization details depending on the nature of the respective investigation.

In an attempt to address this issue, we propose a user-centric approach for video summarization and visualisation, which produces video summaries with a user-defined variable granularity. The method depicts event-centred shots in a temporal order. The contribution of the paper is twofold. First, it proposes a method for assessment of the importance of each frame of the video sequences exploiting multiple information queues, including low-level characteristics of the frames, event-based semantic information about the content of the videos and reasoning of the semantic importance of each event.

The second contribution of the paper is the proposal of a visualization scheme for effective surveillance video summaries, improving the users experience and simplifying the information presented to him/her. The idea behind the proposed method is, in contrast to the literature, to present a number of key-frames that is not fixed but varies according to the needs to the user, taking under account the importance of each frame. The results are presented in a single timeline, highlighting information on the camera source, for multi-camera scenarios, and the time details. Moreover, the proposed methodology enables a meaningful interaction between the user and the system, where the former can adjust the amount of information provided.

The rest of the paper is organized as follows: The process for keyframe selection is described in Section 2, while the visualization strategy is defined in Section 3. Experimental results are presented in Section 4 and conclusions in Section 5.

2 Key frame Selection

The proposed key-frame extraction methodology is a two-stage process (Figure 1). At the first stage, the video is segmented into multiple meaningful fragments using an event-based schema. Each fragment constitutes an interesting event, i.e. an action performed by either people or objects. During the second stage, for each meaningful fragment, the most important frames are selected by analysing their inter-frame differences, concerning motion and color content. By fusing both event-based and inter-frame information, an importance score is assigned to selected key-frames, which is exploited, during visualisation, for adjustment of the amount of information to be presented to the user.

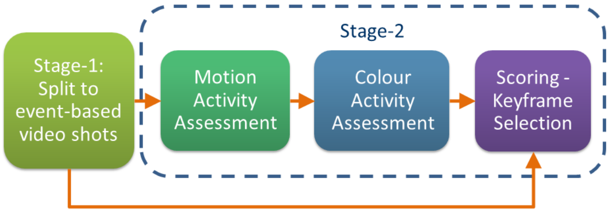

Stage-2
Stage-1: Split to event-based video shots
Motion Activity Assessment
Colour Activity Assessment
Scoring - Keyframe Selection

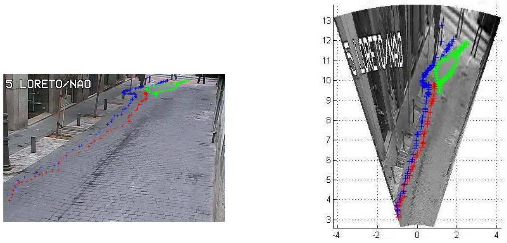
Figure 1: key-frame extraction methodology

2.1 Event-Driven Video Segmentation

The proposed event-based segmentation into meaningful fragments exploits the semantic information derived from the low-level actions, which are performed by individuals appearing on the video sequences. These low-level actions are used to split the video sequence into a series of action shots, separately for each individual. Thus, instead of focusing on the video sequence as a whole, the behavior of each individual is analyzed independently. Thus, the video summary becomes more thorough and detailed, targeted around the events taking place in the video and allowing the user to have a more analytical view.

In order to achieve elementary action recognition, in our framework, a pedestrian tracker [7] is utilised, which identifies the trajectories of the people involved. Then, elementary actions (e.g. walking, running, loitering) are modelled exploiting the displacement of the examined person in a pre-defined time window. However, the perspective of the camera can significantly affect the perception of the action and the detection models.

To alleviate this effect, a methodology is applied to remedy the perspective distortion. More specifically, the Inverse Perspective Mapping (IPM) transform is applied on the extracted trajectories and the top-down view coordinates of the tracked trajectories are calculated. An example of the approach is depicted in Figure 2. Using the real-world coordinates, it is possible to estimate the velocity of a tracked object and thus, moderate the perspective effect error and produce more accurate action characterizations. It is worth mentioning that, due to image quantization errors, IPM is not robust in producing accurate estimates of the velocity, especially for distant objects. However, an approximation of velocity is still feasible, which is adequate for distinguishing between generic actions, such as "walk" and "run".



(a) (b)

Figure 2: Example of Inverse Perspective Mapping method

Summarising, each video sequence is segmented into a series of video shots, where each video shot consists of the frames representing an action performed by a specific individual. Thus, all participants that appear in the video sequence are presented as separate actions (video shots) to the user.

2.2 Low-Level-Feature-Based Reasoning

After the event-based segmentation of the video sequence into multiple shots, the importance of the frames for each shot is assessed. The assessment is driven by the intuition that, in video sequences with static background, important information is present in segments where significant motion activity is observed. Moreover, changes in the color distribution of a frame can also be an indication of importance.

Based on the above facts, we use two types of image features in order to calculate the importance of each frame in a video sequence: motion activity and color histograms. To quantify the significance of the frames, we introduce a new scoring function, which effectively measures the level of changes between consecutive frames.

Motion activity in a frame refers to the amount of motion detected between the frame and the previous one. Since we are not interested in identifying the nature or source of the motion, frame differencing is utilised to produce a rough estimation of the overall motion activity. The objective is to quantitatively detect the activity in the frame from its pixel-based difference with the reference frame.

However, the accuracy of frame differencing can be easily affected by compression artefacts and background slight motion that produce pixel-level noise. To address these issues, two pre-processing steps are applied: image transformation to grey-scale and image smoothing by applying a Gaussian filter. Then, the pixel-wise difference of the two consecutive frames is calculated, in the form of a mask. Assuming that I_t and I_{t+1} are two consecutive frames, then the mask is defined as $I_m = I_{t+1} - I_t$. Subsequently, the sum of the pixel values of the mask is calculated to represent the overall motion activity of the frame under investigation. The sum of the elements of the mask is calculated as follows:

$$Sum_{mf} = \frac{1}{255}\left(\frac{\sum_{p=0}^{N}\sum_{q=0}^{M} I_m(p,q)}{NM}\right), \tag{1}$$

where p, q are the pixel coordinates, N is the width and M is the height of the mask I_m.

Colour histograms are statistics that represent the distribution of colours in an image. More specifically, they accumulate the number of pixels that have colours within each of a predefined list of colour ranges. They are frequently used to compare images, because they are simple and fast to compute. Image histograms can be calculated in various colour spaces. In our framework, the input frames are transformed to the HSV colour space first, and then their histograms are built. The choice of the HSV colour space lies to the fact that unlike RGB, it separates the image intensity from the colour information. This separation proves to be valuable in many applications.

For the comparison of two colour histograms, an appropriate metric that expresses how well the histograms match is required. Eventually, the Bhattacharyya distance has been selected among several distance metrics. Bhattacharyya distance has been widely used in statistics to measure the similarity of two discrete or continuous probability distributions. It is closely related to the Bhattacharyya coefficient, which is a measure of the overlap between two statistical samples of populations, and it is described with the following function:

$$d_{Bh}(H_1, H_2) = \sqrt{1 - \frac{1}{\sqrt{\overline{H_1} \cdot \overline{H_2} N_2}} \sum_{i=0}^{N} \sqrt{H_1(i) \cdot H_2(i)}} \tag{2}$$

where

$$\overline{H_k} = \frac{1}{N}\sum_{i=0}^{N} H_k(i) \tag{3}$$

H_1, H_2 are the two histograms to be compared and N is the total number of histogram bins. Based on the equation above, low values of d_{Bh} correspond to high similarity between histograms, while high values of d_{Bh} correspond to high dissimilarity. The latter indicates high alteration in the visual content of the frame, which may imply the occurrence of an interesting event. Thus, the histogram comparison score can be used as a measure for the frame importance, where most important frames are those with the highest score d_{Bh}.

2.3 Selection of the most Important Frames

The approach followed for assessing the importance of each frame is depicted in the second processing stage in 1. For the frames of an event-based shot motion differencing is initially applied as a filtering stage. The motion activity on a frame to frame base is calculated using the sum of elements Sum_{mf}. Frames with very low Sum_{mf} are not regarded as important, thus, they are discarded before the next step.

Then, changes in the colour distribution of the remaining frames are assessed and the importance score d_{Bh} (2) is calculated. Instead of keeping only one representative keyframe for each segmented video shot, to be presented in visualisation, we propose to maintain a number of important frames, presenting a variable number k of them. As it will be explained in the next section, this number k will not be fixed but will vary depending on the user requirements and the action itself.

Apart from the basic selection criteria, two additional parameters are taken into account. The first is a temporal constraint that needs to be applied to avoid selection of multiple neighbouring frames with very high score. This is achieved by selecting frames that fulfil the following criterion $Frm_k - Frm_n > fps$, where Frm_k and Frm_n are two frames assessed as important and fps is the frames per second rate of the video sequence. In case the criterion is not met, the next higher ranked frame is considered and so on.

The second factor that influences the importance of a frame is the type of low-level action that is identified in the specific video shot. We argue that actions related to important incidents should be favoured against those that are parts of less interesting events. This can be achieved by assigning different weights to the score of each specific action. In this paper, we propose a new weighting function that is based on the following assumption: actions that have lower probability to occur should have higher weights, since they usually correspond to abnormal events.

Now, let a video frame, where N persons are involved and each one performs a specific action. The total weight of the frame is given by the following equation:

$$w = \frac{1}{N}\sum_{i=1}^{N}(1 - p(i)) \tag{4}$$

where $p(i)$ is the probability of occurrence of the action performed by the i^{th} individual. The probabilities $p(i)$ are computed using a predefined training dataset of known actions. It is obvious that less probable actions are assigned higher weights. Weights assigned are in the [0,1] range.

3 Visualization

Visualization is an important aspect of the summarization functionality. In order to provide to the user an intuitive interface the results of the summarization are presented in a timeline. Multiple cameras are depicted in different rows of the interface to better signal the spatio-temporal correlation of the footage. The user can also adjust the granularity of keyframes in both time and importance. For the former functionality zoom in and out in the time domain is provided. For the latter, the user is provided with a summarization granularity control that defines the minimum importance of a frame to be presented.

In Figures 3 and 4, two snapshots of the proposed visualisation scheme are presented. All keyframes are shown in a timeline. In the Figure 3, keyframes are shown to the user when they exceed importance level of 0.6, while in Figure 4, the user selects the threshold of importance to be 0.8. It is obvious that the proposed scheme can adapt the amount of content to be presented in the summary to the varying importance of the content itself as well as to the specific user needs.

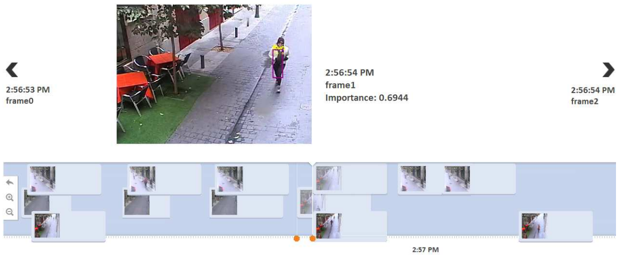


Figure 3: Example of the proposed visualisation scheme; here, keyframes are shown to the user when they exceed importance level 0.6.

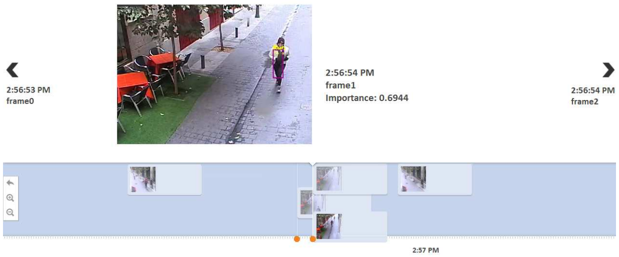


Figure 4: Example of the proposed visualisation scheme; here, keyframes are shown to the user when they exceed importance level 0.8.

4 Experimental Results

Experiments have been performed on a dataset of real-world surveillance videos provided by the London Metropolitan Police in the context of the LASIE research project. For the purposes of the experiments, three types of low-level actions are identified: run, $walk$ and $loiter$.

A person is considered loitering if one of the following two criteria is fulfilled; the person remains in a small area for more than 6 seconds or the ratio of the first to last position distance to the overall covered distance is very small. To distinguish $walk$ from run for a moving person, a threshold on the velocity is employed to characterize the action. If the calculated velocity exceeds the defined threshold, the action is labelled as run, otherwise as $walk$. In Figure 5, examples of video shots are presented, where several keyframes of those actions and their combinations are shown, along with their importance scores.

According to Figure 5, it is obvious that keyframes that involve $walk$ are assigned lower scores than those involving $loiter$ or run, which is due to the fact that, following equation 4, $walk$ is an action with higher probability of occurrence compared with the other two. This makes sense taking into account that in a surveillance video, a low level action such as run can be an indication of an important incident, while actions as $walk$ or $loiter$ are considered more normal.

5 Conclusions

In this paper, a user-centric approach for video summarization and visualization is presented. During video summarization, a number of frames is selected and their importance is assessed. This is achieved by appropriately fusing semantic information with low-level features. The semantic information is used to segment the event into a series of action shots, enabling the independent behaviour analysis of each individual. For each of the shots, inter-frame differences are extracted, using the motion and color frame content.

At the keyframe selection stage, a score is assigned to each of the selected key frames, which represents the importance of the specific frame. The final number of the selected frames is variable, and it is closely related to the video events. Video events containing a lot of visual information produce a higher number of key-frames, in contrast to video events containing a low amount of information.

A novel visualisation scheme is proposed that allows the interaction between the user and the system. The user is presented with a set of keyframes on a timeline that takes under consideration the camera setup, providing a spatiotemporally meaningful overview of the events. Moreover, it allows the user to adjust the granularity of the results in time and importance. This enables the user control the amount of information that s/he is demonstrated.

In conclusion, it is worth mentioning that successful video summarization very subjective. Due to the lack of an objective ground truth, it is almost impossible to evaluate the correctness of a video summary, while it is difficult even for humans to decide whether a video summary is better than another [15]. The

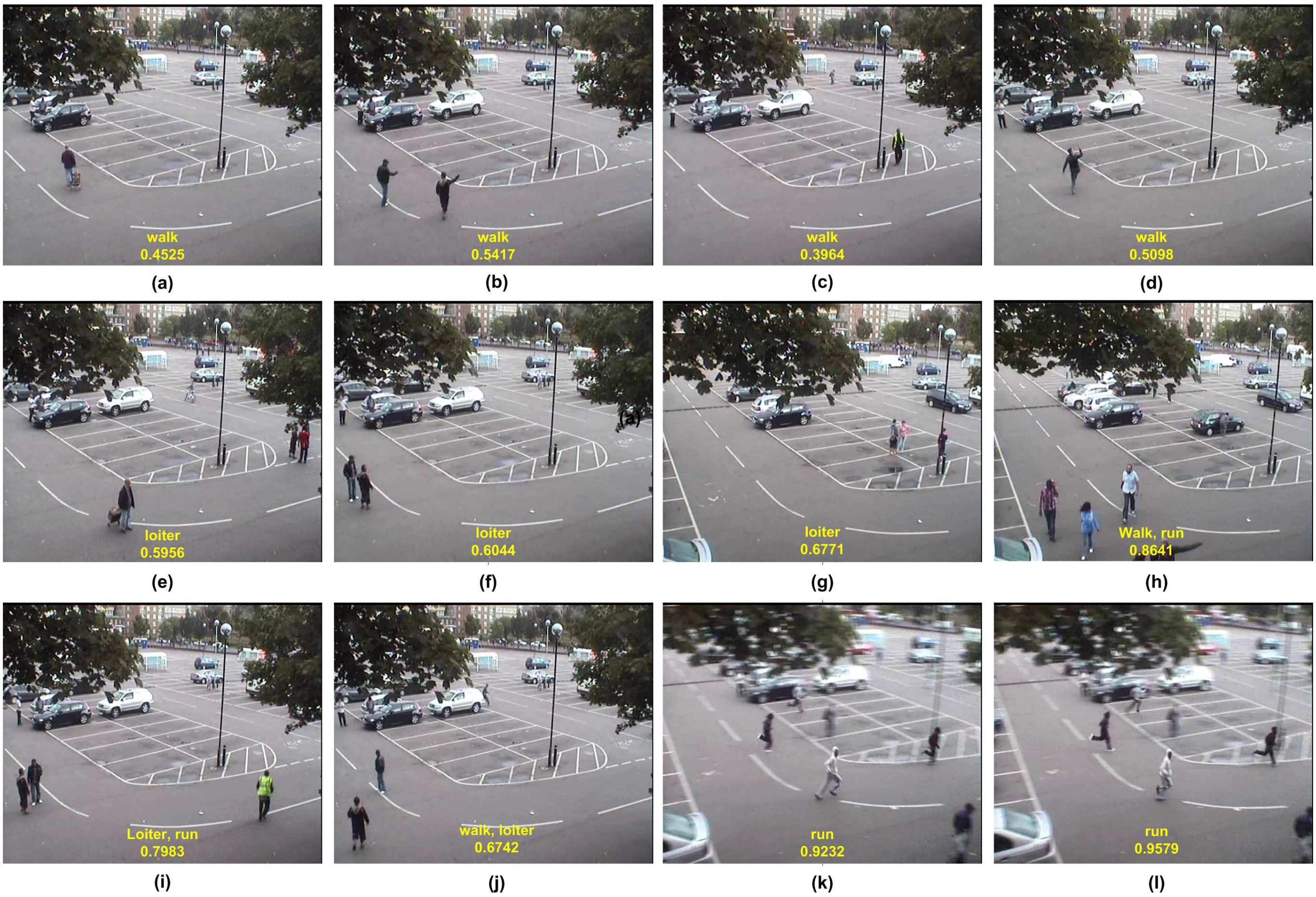


Figure 5: Example of keyframes for the actions $walk$, $loiter$ and run in MET dataset. A person is considered loitering if s/he remains in a small area for more than 6 seconds or the ratio of the first to last position distance to the overall covered distance is very small. To distinguish $walk$ from run, a threshold on the velocity is employed to characterize the action. Since $walk$ is an action with higher probability of occurrence compared with $loiter$ or run, keyframes that involve $walk$ are assigned lower scores. Additionally, run gets the highest scores since it can be an indication of an important incident.

goal of this framework is to overcome these difficulties, by producing a user-centric approach that enables users interact with the system and select a variable number of key-frames for the video summary. This framework can be generalised so as to address other scenarios and other types of actions. Of course, the outcomes of action recognition and key-frame selection depend on the methods that will be selected on each step.

Acknowledgments

This work was supported by the European Communitys funded project LASIE (www.lasie-project.eu) under Grant Agreement no. 607480.

References

[1] Tao Chen, Aidong Lu, and Shi-Min Hu. Visual storylines: Semantic visualization of movie sequence. *Computers and Graphics*, 36(4):241 – 249, 2012. Applications of Geometry Processing.

[2] Cheng-Chieh Chiang and Huei-Fang Yang. Quick browsing and retrieval for surveillance videos. *Multimedia Tools and Applications*, pages 1–17, 2013.

[3] C.T. Dang and H. Radha. Heterogeneity image patch index and its application to consumer video summarization. *Image Processing, IEEE Transactions on*, 23(6):2704–2718, June 2014.

[4] Naveed Ejaz, Tayyab Bin Tariq, and Sung Wook Baik. Adaptive key frame extraction for video summarization using an aggregation mechanism. *J. Vis. Comun. Image Represent.*, 23(7):1031–1040, October 2012.

[5] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. Multi-view video summarization. *Trans. Multi.*, 12(7):717–729, November 2010.

[6] Ciocca Gianluigi and Schettini Raimondo. An innovative algorithm for key frame extraction in video summarization. *Journal of Real-Time Image Processing*, 1(1):69–88, 2006.

[7] Vasileios Lovatsis, Anastasios Dimou, and Petros Daras. Introducing context awareness in multi-target tracking using re-identification methodologies. In *Imaging for Crime Detection and Prevention*, London, UK, December 2013.

[8] K.M. Mahmoud, N.M. Ghanem, and M.A. Ismail. Unsupervised video summarization via dynamic modeling-based hierarchical clustering. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 2, pages 303–308, Dec 2013.

[9] Tao Mei, Bo Yang, Shi-Qiang Yang, and Xian-Sheng Hua. Video collage: presenting a video sequence using a single image. *The Visual Computer*, 25(1):39–51, 2009.

[10] Padmavathi Mundur, Yong Rao, and Yelena Yesha. Keyframe-based video summarization using delaunay clustering. *International Journal on Digital Libraries*, 6(2):219–232, 2006.

[11] Cuong Nguyen, Yuzhen Niu, and Feng Liu. Video summagator: An interface for video summarization and navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 647–650, New York, NY, USA, 2012. ACM.

[12] Shun-Hsing Ou, Chia-Han Lee, V.S. Somayazulu, Yen-Kuang Chen, and Shao-Yi Chien. Low complexity online video summarization with gaussian mixture model based clustering. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1260–1264, May 2014.

[13] Silvia Pfeiffer, Rainer Lienhart, Stephan Fischer, and Wolfgang Effelsberg. Abstracting digital movies automatically. *Journal of Visual Communication and Image Representation*, 7(4):345 – 353, 1996.

[14] Y. Takahashi, N. Nitta, and N. Babaguchi. Video summarization for large sports video archives. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1170–1173, July 2005.

[15] Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1), February 2007.

[16] Zhengzheng Tu, Dengdi Sun, and Bin Luo. Video summarization by robust low-rank subspace segmentation. In Zhixiang Yin, Linqiang Pan, and Xianwen Fang, editors, *Proceedings of The Eighth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 2013*, volume 212 of *Advances in Intelligent Systems and Computing*, pages 929–937. Springer Berlin Heidelberg, 2013.