Video summarization based on shot boundary detection with penalized contrasts

Paschalina Medentzidou and Constantine Kotropoulos
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, GREECE
Email: pmedentz@aiia.csd.auth.gr, costas@aiia.csd.auth.gr

***Abstract*—In this paper, we propose a novel technique for shot boundary detection and video summarization that is based on change point detection with penalized contrasts. The proposed method extracts a proper time series from the video, calculating the mean level of the hue component of consecutive video frames in the Hue Saturation Value color space. Change point detection is applied to the time series, estimating the number of shot transitions and their location. The posterior distribution of the change point sequence is defined, that splits the video into homogenous temporal segments. A representative frame is selected in each segment and similar or meaningless keyframes are deleted from the summary. The resulting summary is of comparable quality to that of the state of the art techniques.**

***Keywords*—Change Point Detection; Shot Boundary Detection; Bayesian Information Criterion; Keyframe Extraction.**

I. Introduction

The amount of video data being recorded and distributed every day has been increased due to the low storage cost, the high speed of internet connections, and the ability of individuals to capture videos using smart devices and cameras. As a consequence, the need for a short representation of video data has emerged. Video summarization aims to tackle this need. It is defined as the process of finding a sequence of still or moving pictures (with or without audio), representing the content of a video in a concise manner. This way the essential message of the original video is preserved [1]. As set in [2], the video summary must contain high priority entities and events from the video, exhibit reasonable degrees of continuity, and be free of repetition.

There are two types of video summarization, namely static keyframe extraction and dynamic video skimming. A keyframe is a frame that represents the content of a logical unit, like a shot or scene, for example. This content must be as representative as possible [3]. A dynamic video skimming consists of a collection of associated audio-video sub-clips selected from the original sequence, but with a much shorter length [4]. One advantage of a video skim over a keyframe set is its ability to include audio and motion elements that enhance both the expressiveness and information of the summary. On the other hand, keyframe sets are not restricted by any timing or synchronization issues [5].

This paper addresses the video summarization problem as a static keyframe extraction challenge. A novel method for change point detection in the mean is applied to a video time series extracted by calculating the mean of the hue component in Hue Saturation Value (HSV) color space in consecutive video frames. Change point detection in the mean of a signal is the process of finding changes in the average level of the signal parameters. Some of its applications include fault detection in chemical processes, detection and identification of seismic waves, biomedical signal processing, music restoration, changes in DNA sequences, speech and speaker segmentation [6].

The novelty of the proposed approach for shot boundary detection and video summarization, is in extracting a proper video time series capturing the video structure and its modeling as a linear model with additive noise. Noise accounts for small changes in the mean hue of consecutive frames that are attributed to camera motion, entering of objects, or exiting a shot, etc. Akaike (AIC) and Bayesian (BIC) information criteria are frequently employed in order to find a parsimonious solution to this model. In a Bayesian framework, the conditional distribution of the change point sequence is constructed. A Markov Chain Monte Carlo (MCMC) is employed to sample the posterior distribution. A Stochastic Approximation of Expectation Maximization (SAEM) estimates the distribution parameters, as detailed in Section III-A-III-C.

The detected change points correspond to the candidate shot cuts that split the video into homogenous temporal segments. Representative keyframes are chosen from each segment, while meaningless keyframes and keyframes that are similar to others are rejected, as being redundant. The keyframes that are derived, create the video summary describing the main video content.

The remainder of the paper is organized as follows. Section II reviews existing video summarization methods. Section III describes the proposed technique for shot cut detection and video summarization. Section IV reports and discusses experimental results on a benchmark dataset. Finally, some concluding remarks are given and future research objectives are identified in Section V.

II. Related work

Many static summarization methods were proposed. Mundur et al. [7] developed a method based on Delaunay Triangulation, which was applied to cluster video frames. The method samples the frames of the original video and represented them by the color histogram in the HSV color space. The Delaunay diagram was built and clusters were

obtained by separating the diagram edges. For each cluster, the frame nearest to its center was selected as keyframe.

Furini et al. introduced the STIll and MOving Video Storyboard, (STIMO) [8]. The STIMO was designed to produce video storyboards on the fly. It was composed of three main phases. First, the video was analyzed and the color description was extracted in the HSV color space. Second, a clustering algorithm was applied to color description. Third, meaningless video frames were removed from the produced summary.

de Avila et al. proposed a video summarization technique, called VSUMM based on color feature extraction from the color histograms in the HSV color space [9]. The video frames were sampled at 1 frame per second (fps) and then clustered by k-means algorithm, while redundant frames were removed.

Ngo et al. proposed a unified approach for video summarization based on the analysis of video structures and video highlights. The main components of this approach were scene modeling and highlight detection. A video was represented as an undirected graph and the normalized cut algorithm partitioned the graph into video clusters. Video summaries were generated from a temporal graph [10]

A temporal segmentation into semantically consistent segments, delimited by shot boundaries and general change points was performed in [11]. Importance scores were assigned to each segment, using a Support Vector Machine classifier. The resulting video summary assembles the sequence of segments with the highest scores.

Detection of attention-invoking audiovisual segments in movies was proposed on the basis of saliency models for the audio, visual, and textual information conveyed in a video stream [12].

Video summarization using shot boundary detection based on the mutual information and the joint entropy between frames was developed in [13].

III. Proposed Video Summarization Method

The proposed video summarization method executes the following steps: (a) A time series is extracted, which describes the video structure by calculating the average hue per frame in consecutive video frames. (b) Change point detection is applied to the aforementioned time series in non overlapping sliding windows of 500 frames. When the number of frames is not integer multiple of the window size then an overlap is allowed between the last two window instances. That is, if a video consists of 1350 frames, then the method is applied to frames 1-500, 501-1001, and 850 to 1350. Duplicate detected changes are deleted. Change points are simultaneously detected by minimizing a penalized contrast defined in Section III-B and III-C. (c) Each video segment defined by pairs of consecutive change points is represented by the temporal middle frame. For example, the middle frame for the segment defined by the $10th$ and $81th$ frame is the $46th$ frame (d) Meaningless frames (i.e., monochromatic ones) characterized by a low ratio of edge pixels are deleted. (e) Highly correlated frames, frames with similar RGB color histogram as well as similar texture are rejected.

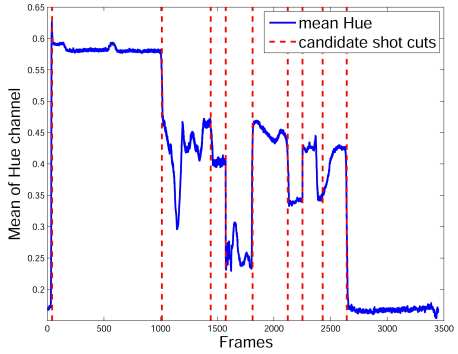

Fig. 1. Mean of the hue channel in video v57.mpg and detected shot cuts.

A. Time series extraction

A time series from each video is extracted that models the video structure. It is defined as the mean of the hue component of consecutive video frames. The HSV color space is used, because it provides an intuitive representation of color that matches the color perception by humans. Frame averages are employed instead of frame differences that are commonly used in the literature, because the proposed method for change point detection identifies changes in the mean level of a signal.

The time series in Fig. 1 depicts the average hue value per frame for the video v57.mpg in the dataset employed in [9]. It exhibits a change every time a shot cut occurs. Let the time series be modeled as $d_i = \mu_i + \sigma\, e_i,\ i = 1, 2, ...N$ where $\{e_i\}$ is a sequence of zero-mean random variables with variance equal to one and N is the number of video frames.

The method proposed by M. Lavielle [14][1] is applied for change point detection. This method is proved efficient for both abrupt and gradual transition detection. The method estimates the number of change points (i.e., shot cuts) and their location using penalized contrasts. Formally, the posterior distribution of the change point sequence is defined as a function of the penalized contrast in a Bayesian framework, as is detailed next.

B. Shot cut detection

In most cases, the abrupt changes in a signal $\{d_i\}$ are attributed to a parameter $\theta = \{\mu\}$, where μ is the mean level of the time series. This assumption is used to define the contrast function $J(\boldsymbol{m}, \boldsymbol{d})$. Let K be an integer indicating the number of video segments (shots) and $\boldsymbol{m} = (m_1, m_2, ..., m_{K-1})$ be a sequence of integers defining the shot boundaries. The sequence is satisfying the following ordering $0 < m_1 < m_2 < ... < m_{K-1} < N$. For any k, such that $1 \leq k \leq K$, let $U(d_{m_{k-1}+1}, ..., d_{m_k} \mid \mu)$ be a contrast function useful for the estimation of the unknown value of parameter μ in shot k. A Gaussian log-likelihood can be used as a contrast function even if $\{e_i\}$ is not a sequence of Gaussian random variables, i.e.:

$$U(d_{m_{k-1}+1}, ..., d_{m_k} \mid \mu) = \sum_{i=m_{k-1}+1}^{m_k} (d_i - \mu)^2. \qquad (1)$$

[1] Matlab code is available at http://www.math.u-psud.fr/∼lavielle/programmes_lavielle.html

Let G be defined as

$$G(d_{m_{k-1}+1},...,d_{m_k}) = U(d_{m_{k-1}+1},...,d_{m_k} | \hat{\mu}\,(d_{m_{k-1}+1},...,d_{m_k})). \quad (2)$$

The minimum contrast estimate $\hat{\mu}\,(d_{m_{k-1}+1},...,d_{m_k})$ on segment k satisfies

$$G(d_{m_{k-1}+1},...,d_{m_k}) \leq U(d_{m_{k-1}+1},...,d_{m_k} \mid \mu). \quad (3)$$

Then,

$$G(d_{m_{k-1}+1},...,d_{m_k}) = \sum_{i=m_{k-1}+1}^{m_k} (d_i - \overline{d}_{m_{k-1}+1:m_k})^2 \quad (4)$$

where $\overline{d}_{m_{k-1}+1:m_k}$ is the sample mean in the shot, having boundaries at $m_{k-1}+1$ and m_k, i.e., $\overline{d}_{m_{k-1}+1:m_k} = \frac{1}{(m_k - m_{k-1})} \sum_{i=m_{k-1}+1}^{m_k} d_i$.

A proper contrast function is defined as [14]:

$$J(\boldsymbol{m}, \boldsymbol{d}) = \frac{1}{N} \sum_{k=1}^{K} G(d_{m_{k-1}+1},...,d_{m_k}) \quad (5)$$

where $\boldsymbol{d} = (d_1,...,d_N)$, $m_0 = 0$ and $m_K = N$. The objective is to find some instants $m_1^* < m_2^* < ... < m_{K^*-1}^*$, such that, $\mu_{m_{K-1}^*+1} = \mu_{m_{K-1}^*+2} = \cdots = \mu_{m_K^*}$.

When the number of shot cuts is unknown, it can be estimated by minimizing a penalized version of $J(\boldsymbol{m}, \boldsymbol{d})$. For any sequence of shot cuts $\boldsymbol{m}$, let $pen(\boldsymbol{m})$ be a function of $\boldsymbol{m}$ that increases with the number of shots $K(\boldsymbol{m})$ and resembles the penalty term in BIC added to the goodness of fit term. The sequence of shot cuts $\hat{\boldsymbol{m}}$ minimizes

$$H(\boldsymbol{m}) = J(\boldsymbol{m}, \boldsymbol{d}) + \frac{2\sigma^2}{N}\, pen(\boldsymbol{m}). \quad (6)$$

If the second term in (6) is a function of N that goes to 0 at an appropriate rate as N goes to infinity, the estimated number of segments $K(\hat{\boldsymbol{m}})$ converges in probability to K^*.

Following the most popular information criteria, such as the AIC and the BIC, the simplest penalty function is $pen(\boldsymbol{m}) = K(\boldsymbol{m})$. To defend this choice, let $d_i = \mu_i + \sigma\, e_i,\ 1 \leq i \leq N$, where μ_i is constant in each particular segment defined by the boundaries $\boldsymbol{m} = (m_1, m_2, ..., m_{K-1})$, i.e., $\mu_{(1<i<m_1)} = \mu_1, ..., \mu_{(m_{K-1}<i<N)} = \mu_K$. If the sequence $\{e_i\}$ is a sequence of Gaussian white noise with variance 1, a penalized least-squares estimate obtained by minimizing:

$$H(\boldsymbol{m}, \boldsymbol{d}) = \frac{1}{N} \sum_{k=1}^{K(\boldsymbol{m})} \sum_{i=m_{k-1}+1}^{m_k} (d_i - \overline{d}_k)^2) + \frac{2\sigma^2}{N}\, pen(\boldsymbol{m}) \quad (7)$$

where the penalty function takes the form

$$pen(\boldsymbol{m}) = K(\boldsymbol{m}) \left(1 + c \log \tfrac{N}{K(\boldsymbol{m})}\right) \quad (8)$$

minimizes $E\{|\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}^*|^2\}$, where $\hat{\boldsymbol{\mu}}_m$ is the estimated sequence of the time series means and $\boldsymbol{\mu}^*$ is the real sequence of the time series means. In [14], c was set to 2.5.

The number of homogenous time series segments (shots) K can be estimated as follows [14]:

1) For $1 \leq K \leq K_{MAX}$, let

$$\widetilde{J}_K = \frac{J_{K_{MAX}} - J_K}{J_{K_{MAX}} - J_1}(K_{MAX} - 1) + 1. \quad (9)$$

The new sequence $\widetilde{J}_K$ is normalized such that $\widetilde{J}_1 = K_{MAX}$ and $\widetilde{J}_{K_{MAX}} = 1$. The aforementioned sequence decreases with an average slope equal to -1.

2) For $2 \leq K \leq K_{MAX} - 1$, let $D_K = \widetilde{J}_{K-1} - 2\widetilde{J}_K + \widetilde{J}_{K+1}$ and $D_1 = \infty$. Then, the minimum penalized contrast (MPC) estimate of K is given by:

$$\widetilde{K}_{MPC} = \max\{1 \leq K \leq K_{MAX} - 1 \mid D_K > S\} \quad (10)$$

where S is a threshold. $\widetilde{K}_{MPC}$ is defined as the greatest number of shots K, such that the second derivative of J is greater than S. If no second derivative is greater than S, then there are no shot changes and $\widetilde{K}_{MPC} = 1$.

Many different numerical experiments were concluded in order to find that the best performing threshold value is $S = 0.75$.

C. Bayesian approach

Let us express $H(\boldsymbol{m})$ in (6) as $H(\boldsymbol{m}) = J(\boldsymbol{m}, \boldsymbol{d}) + \beta\ pen(\boldsymbol{m})$ with $\beta > 0$. For a fixed value of β, the procedure described in Section III-B yields only one solution $\hat{\boldsymbol{m}}_{\hat{K}(\beta)}$. In general, assume the following posterior distribution:

$$p(\boldsymbol{m} \mid \boldsymbol{d}; \alpha, \beta) = D\,(\boldsymbol{d}\,;\, \alpha, \beta) \exp\left(-\alpha(J(\boldsymbol{m}, \boldsymbol{d}) + \beta pen(\boldsymbol{m}))\right) \quad (11)$$

where $D\,(\boldsymbol{d}\,;\, \alpha, \beta)$ is a normalizing constant, and $\alpha > 0$. The mode of this posterior distribution is the minimum contrast estimate of shot cuts $\boldsymbol{m}$. Obviously, the mode depends on α, β that have to be estimated.

First, the parameters of the posterior distribution are estimated by the SAEM algorithm, which yields the maximum likelihood estimates of α and β [15]. Second, an MCMC procedure is used to estimate the posterior distribution $p(\boldsymbol{m}|\boldsymbol{d}; \alpha, \beta)$. Let T be a temperature parameter, which controls how the posterior distribution is concentrated around its mode. For different values of $T \in (0, 1]$:

1) Use the MCMC algorithm to sample the conditional distribution $p(\boldsymbol{m} \mid \boldsymbol{d}; \hat{\alpha}/T, \hat{\beta})$
   a) estimate the marginal conditional probability $P\{\exists$ a changepoint at $d_i \mid \boldsymbol{d};\ \hat{\alpha}/T,\ \hat{\beta}\}$ for $1 \leq i \leq N-1$, where d_i is the i-th observation.
   b) estimate the conditional probability $P\{K(\boldsymbol{m}) = K \mid \boldsymbol{d};\ \hat{\alpha}/T\ \hat{\beta}\}$.
2) Compute the maximum a posteriori estimate of $\boldsymbol{m}$ by minimizing $J(\boldsymbol{m}, \boldsymbol{d}) + \hat{\beta} pen(\boldsymbol{m})$.

The parameter T should be chosen small enough to neglect the models $\boldsymbol{m}$ admitting a low posterior probability and to increase the probability of the most likely models. Here, the MCMC algorithm creates a homogeneous Markov chain, since the temperature parameter T remains constant. Accordingly, the maximization of the conditional distribution is achieved by using a simulated annealing procedure.

D. Elimination of similar and meaningless candidate keyframes

After the extraction of videos' time series and the shot boundary detection, we draw the temporal middle frame of each segment as representative keyframe. Any detected monochromatic keyframes are discarded. Such keyframes occur primarily due to fade-in or fade-out effects. As suggested in [9], meaningless keyframes can be identified by the standard deviation of the hue color histogram. Zero or sufficiently small values of standard deviation indicate such meaningless keyframes. Here, the meaningless keyframes are deleted by thresholding the ratio of pixels in edges at each frame. That is, if

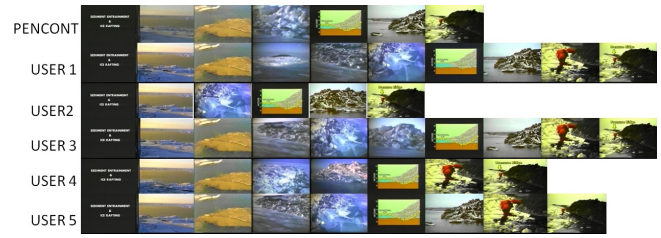

Fig. 2. Sample video summarization obtained by the proposed method PENCONT and ground truth provided by 5 users.

the ratio of the pixels in edges is less that 1% then the keyframe is deleted. In monochromatic keyframes, edges are not detected, so such keyframes can easily be identified. To extract the edges, the Prewitt edge detector is used.

Redundant keyframes are eliminated based on the RGB color histogram. A similarity matrix is calculated for the keyframes that are candidate for the summary. The 16 bin color histograms of the red, green, and blue channel are extracted. The three histograms are merged in a single color vector of size 48. Color vectors are compared using the Manhattan distance. If the distance is less than $\tau_{HIST} = 0.3$, then among the two compared keyframes the first is deleted. During the computation of the similarity matrix, if the distance between two keyframes is less than τ_{HIST}, then calculations stop. There is no need to calculate the distance between this frame and the others since the rejection condition has been reached. This way, computational complexity decreases.

To measure if two candidate keyframes are similar with respect to their texture, a similarity matrix is built, computing the Manhattan distance between their GIST descriptors [16]. If the distance is less than the threshold $\tau_{GIST} = 0.2$, then one of them is deleted. SURF features that detect and describe points of interest in a keyframe are extracted as well. If two candidate keyframes have more than $\tau_{SURF} = 4$ matching points, they are considered similar and they are also deleted [17]. Finally, the correlation between candidate keyframes is calculated. If it exceeds $\tau_{CORR} = 0.9$, then the candidate keyframes are considered similar and one of them is deleted from the summary. Note that the main body of the duplicate keyframes is discarded by the RGB color histogram, which is the computationally fastest method to reject redundant frames. Therefore, the similarity matrices that are calculated for the SURF and GIST features, and the correlation between the keyframes are created using a much smaller number of frames than the number of frames used to build the similarity matrix of RGB color histograms. In Fig.2, the video summarization results and the ground truth of 5 users are shown for a video of the database.

IV. Experimental Evaluation

The publicly available database[2], which consists of 50 videos chosen from Open Video Project[3] has been used. Each video is about $100s$ long and has 3036 frames on average. This database was also used for video summarization in [9] and [18]. To evaluate the quality of video summaries, the method proposed in [9], called Comparison of User Summaries (CUS), was employed. In the CUS, the video summary is built manually by a number of users from videos sampled at 1 fps. The user summaries (US) are taken as reference ground truth and are compared with the automatic summaries (AS) derived by the method. A frame from the US is considered as matched with a frame from AS, if the Manhattan distance of their hue histograms is less than $\tau_H = 0.5$ [9]. Note that only one frame from the AS can be matched with a specific frame from the US. This ensures that there will not be number of matches exceeding the number of frames in the US.

[2]https://sites.google.com/site/vsummsite/download

[3]http://www.open-video.org

TABLE I. Performance evaluation of video summarization methods.

	Precision	Recall	F-score	CUS error	Average K
OV	0.67	0.7	0.71	0.57	9.66
DT	0.78	0.53	0.68	0.29	6.2
STIMO	0.63	0.72	0.69	0.58	9.96
VSUMM1	0.75	0.85	0.8	0.38	9.62
VSUMM2	0.81	0.7	0.77	0.27	9.62
PENCONT	0.75	0.7	0.72	0.33	8.42

The AS generated by different algorithms are compared with all the US to obtain quantitative assessment. Evaluation metrics used in [9], [18] to measure summarization quality of each algorithm include the precision (P), the recall (R), the $F-score$, and the CUS_{error} defined next:

$$P = \frac{n_{matched}}{n_{AS}} \tag{12}$$

$$R = \frac{n_{matched}}{n_{US}} \tag{13}$$

$$F - score = \frac{2\ P\ R}{P + R} \tag{14}$$

$$CUS_{error} = \frac{n_{AS} - n_{matched}}{n_{US}} \tag{15}$$

where $n_{matched}$ is the number of matched keyframes between the AS and US, n_{AS} is the number of keyframes in the AS, n_{US} is the number of keyframes in a US. The proposed method, coined as video summarization with penalized contrasts (PENCONT), is compared with the DT [7], the STIMO [8], the VSUMM1 and the VSUMM2 [9], and the OV [19], whose summaries can be found at the website hosting the database. The results are summarized in the Table I. PENCONT has achieved results comparable to those of the state of the art techniques. Specifically, it has achieved the same precision as the VSUMM1, the same Recall as VSUMM2 and OV, and a competitive CUS error of 0.33. The average number of keyframes per summary extracted by the PENCONT is 8.42, which is closer to the average number of keyframes extracted by users (i.e., 8.64).

PENCONT is preferred among the other video summarization techniques, when the number of shots in a video is unknown and when real-time operation is needed. The feature extracted from each video is simple and can be computed fast. Moreover, the complexity of PENCONT can be adjusted by performing change point detection on sub-sampled time series.

V. Conclusions

In this paper, we have addressed shot boundary detection in video as a change point detection problem, using penalized contrasts on an appropriate time series measuring the average hue value per frame over the video duration. Shot boundaries have been detected effectively even in cases of gradual transitions. The shot boundaries have been used to define homogenous temporal video segments. The middle frame from each segment was extracted and considered as a keyframe. Meaningless frames have been rejected by observing the ratio of frame pixels in image edges. Similar frames were detected and removed based on their color captured by the RGB histogram, their texture measured by the GIST and SURF features, and their correlation. The resulting summary

is satisfactory and comparable to that of state of the art techniques with respect to commonly used objective figures of merit. Future work could employ autoregressive data models in the description of the video structure and study the impact of different noise distributions on the efficiency of the method.

VI. Acknowledgements

This research has been co-financed by the European Union (European Regional Development Fund - ERDF) and Greek national funds through the Operation Program Competitiveness-Cooperation 2011 - Research Funding Program: 11SYN-10-1730-ATLAS.

References

[1] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg, "Abstracting digital movies automatically," *Journal Visual Communication and Image Representation*, vol. 7, no. 4, pp. 345 – 353, 1996.

[2] M. Albanese, C. Cesarano, M. Fayzullin, A. Picariello, and V.S. Subrahmanian, *Encyclopedia of Multimedia*, Springer US, 2006.

[3] A. Hanjalic and H. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1280–1289, 1999.

[4] J. Nam and A. Tewfik, "Video abstract of video," in *Proc. 3rd IEEE Workshop Multimedia Signal Processing*, 1999, pp. 117–122.

[5] B. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 1, Feb. 2007.

[6] M. Basseville and I. Nikiforov, *Detection of Abrupt Changes: Theory and Application*, vol. 104, Prentice Hall Englewood Cliffs, 1993.

[7] P. Mundur, Y. Rao, and Y. Yesha, "Keyframe-based video summarization using delaunay clustering," *Journal Digital Libraries*, vol. 6, no. 2, pp. 219–232, 2006.

[8] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: Still and moving video storyboard for the web scenario," *Multimedia Tools and Applications*, vol. 46, no. 1, pp. 47–69, 2010.

[9] S. de Avila, A. da Luz, and A.A. De Araujo, "VSUMM: A simple and efficient approach for automatic video summarization," in *Proc. 15th Int. Conf. Systems, Signals and Image Processing*, June 2008, pp. 449–452.

[10] C. Ngo, Y. Ma, and H. Zhang, "Video summarization and scene detection by graph modeling," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 296–305, 2005.

[11] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. European Conf. Computer Vision*, pp. 540–555. Springer, 2014.

[12] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov 2013.

[13] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 82–91, Jan 2006.

[14] M. Lavielle, "Using penalized contrasts for the change-point problem," *Signal Processing*, vol. 85, no. 8, pp. 1501 – 1510, 2005.

[15] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a stochastic approximation version of the em algorithm," *The Annals of Statistics*, vol. 27, no. 1, pp. 94–128, March 1999.

[16] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Journal Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[17] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. European Conf. Computer Vision*, pp. 404–417. Springer, 2006.

[18] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. Feng, "Video summarization via minimum sparse reconstruction," *Pattern Recognition*, vol. 48, no. 2, pp. 522 – 533, 2015.

[19] D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," in *Proc. 6th ACM Int. Conf. Multimedia*, pp. 211–218. 1998.