Video scene segmentation through an early fusion multimodal approach

Rodrigo Mitsuo Kishi
University of São Paulo
Federal University of Mato Grosso do Sul
São Carlos - SP - Brazil
rodrigokishi@usp.br

Rudinei Goularte
University of São Paulo
São Carlos - SP - Brazil
rudinei@icmc.usp.br

ABSTRACT

Temporal segmentation of video into scenes is a prerequisite to various tasks on Multimedia Information Retrieval, like video summarization, content based video retrieval and video recommendation. There isn't, however, a satisfactory method to automatically segment video into scenes. State-of-the-art scene segmentation methods are multimodal, in order to match the multimodal nature of video. Aside from being multimodal, no true early fusion method was found in literature. Early fusion have shown to be useful in related multimedia tasks where potential correlation between data streams of different sources are discovered before the main processing step, improving results. Motivated by this situation, the proposal of this PhD Project is to investigate the impact of a true early fusion multimodal approach on the temporal video scene segmentation task.

Keywords

Multimedia; Content Adaptation and Personalization; Temporal Video Scene Segmentation

1. INTRODUCTION

Nowadays, people became more and more involved with digital video. According to Cisco [1], in 2015, global mobile video traffic accounted for 55 percent of total mobile data traffic and will account for 75 percent in 2020. Youtube [1] cite a growth in watch time of at least 50 percent year by year in the last three years. Time.com[2] reported that Netflix, one of the biggest video on demand services, is responsible by one third of all Internet traffic on the United States [19].

Given these facts, market and the scientific community are working to provide easy and pleasant experience to users, not only of digital video, but of general multimedia content. Unfortunately, the huge amount of available digital video turns search/retrieval task to be difficult. This difficulty can be assigned to the Information Overload Problem, originally addressed by Gross [13] which is a situation where someone cannot timely handle an amount of data in order to find a specific information of interest.

A variety of tasks on Multimedia Information Retrieval require knowledge about temporal structure of the target video. More specifically, video must be divided into semantically coherent pieces like chapters of a film, individual news of a news program and topics of a video lesson. Segmentation of video into smaller parts can improve access and navigation [5], as the smaller parts are more manageable than the entire video.

Video temporal subunits which are of interest to personalization systems are frame, shot and scene. Segmenting video into frames is trivial and segmenting into shots is considered an essentially solved task [11] with recent methods reaching high levels of accuracy [7]. Temporal segmentation of video in scenes, however, is an open problem [11]. The main reason is most definitions of scene are based on subjective concepts, implying that a scene segmentation tool must deal with the semantic gap [23], the difference between the representation of a content in a computer and the human comprehension of the same content.

Available methods for the temporal scene segmentation problem can be classified by which channels, or combinations of channels, they use as source of information [11]. Early works, including the seminal one by Yeung *et al.* [27], were based only on visual features. Over the time, researchers also experimented other channels like aural, textual and its combinations. Techniques based on a single information channel, also called modality, are classified as singlemodal. Techniques employing more than one modality are classified as multimodal. Recent works remark that singlemodal methods are restricted to specific domains [11, 17] and future techniques must employ multimodality to improve temporal scene segmentation results.

Multimodal methods can employ multimodal fusion process [4], where information from different modalities are combined to obtain a final decision. In video scene segmentation, decision is about where scene boundaries must be placed. Multimodal fusion are of two main types [4]: early fusion, where information from different modalities are combined before the final decision step and late fusion, where an individual decision is taken for each modality and this decisions themselves are combined into a final decision.

Other important classification of temporal scene segmentation methods is about the semantic level of information they use: low, mid and high level features. Features are pieces of information that can be extracted from a video with the purpose of characterizing it, enabling comparison. Low

[1] www.youtube.com/yt/press/en-GB/statistics.html
[2] www.time.com

In: Workshop de Teses e Dissertações (WTD), 16., 2016, Teresina. Anais do XXII Simpósio Brasileiro de Sistemas Multimídia e Web. Porto Alegre: Sociedade Brasileira de Computação, 2016. v. 2.

ISBN: 978-85-7669-332-1

level features are extracted directly from raw data, without external knowledge and aren't easily understandable by humans. Mid level features are obtained by processing low level features with the addition of semantic knowledge and aren't also easily understandable by humans. High level features are human interpretable information about a content.

State-of-the-art scene segmentation methods like [22] are multimodal and use mid/high level features. Although these methods are multimodal, none of them apply early fusion on features. As recent literature on other video adaptation and personalization tasks suggests that early fusion may improve accuracy [12, 15], further investigation about its application on temporal scene segmentation task can be fruitful. As no multimodal early fusion temporal scene segmentation method have been found, the main goal of this work is set as the verification of the effectiveness of this approach on the temporal video scene segmentation problem.

This paper is organized as follows: at Section 2, are presented necessary concepts to the comprehension of the proposal. Section 3 presents the state of the art and exposes research gaps. Section 4 is where research hypotheses are stated and experiment design and planning are described. Finally, at Section 5, expected contributions are listed.

2. THEORETICAL FRAMEWORK

At this Section, a theoretical basis to a better understanding of this PhD project is presented.

2.1 Video Hierarquical Structure And The Temporal Video Scene Segmentation Problem

Video can be defined as a sequence of still images such that, when shown in a sufficient fast succession, causes a continuous visual sensation to a viewer [8]. These still images are called frames and to an unbroken sequence of frames taken from one camera, is given the name of shot [16]. This definition is broadly accepted between researchers.

Definition of scene, the next unit above shots in video hierarchical structure, however, isn't consensual. It is commonly defined as a sequence of one or more adjacent shots somehow related, but this relation varies over different works. Many authors proposes their own definitions based on easily computable features or definitions fitted to very specific domains. There are researchers, however, who adopts a definition which is more general and closer to the interests of a domain independent multimedia system user: a scene is a sequence of one or more adjacent shots which are semantically correlated [14]. Concepts of video, frames, shots and scenes are shown in Figure 1, inspired in [11].

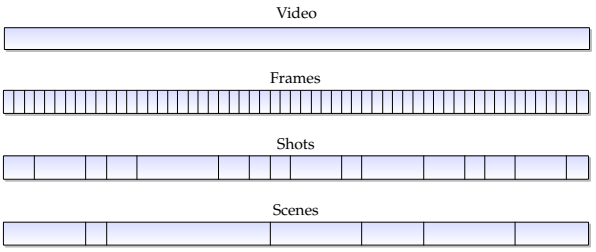

Figure 1: Hierarquical structure of video.

This work is focused on the detection of video temporal units that fits human comprehension of scenes and are also domain independent, the Temporal Video Scene Segmentation (TVSS) task. So, the general scene definition of [14] will be adopted in the definition of the following problem:

> **Temporal Video Scene Segmentation Problem** (TVSSP): Given a video and its ground truth shot boundaries, compute all the sequences of one or more adjacent shots which are semantically correlated and label each one as a scene.

2.2 Video Features

Multimedia systems must handle objects like images, sounds and videos. Comparison between these objects isn't usually performed on the original data representations, but on alternative representations of the original content, in form of metadata. These metadata, normally smaller than the original content, are called feature descriptors, or simply descriptors. In a more concise view, a descriptor is a compact representation of a media object that can be used to compare it to other media object descriptors. Features can be classified regarding their semantic level: low, mid and high.

Low level feature descriptors are numeric or symbolic measurements extracted directly from the data stream without involving external knowledge, learning process or statistical analysis of other documents [20]. To a tool which compute a descriptor from a media stream, is given the name of detector or extractor.

2.2.1 Visual Features

Visual features are divided into two classes, according to its representation scope: global and local. Global features refers the whole image. They are usually compact and indifferent to rotation. One of the most popular global feature for image comparison is the color histogram [29], which is a measure of frequency with which color appears in that image, considering all the pixels.

A local feature, differently from global, corresponds to a subset of an image. This subset can be a point, a region or an edge segment, containing a pattern which differs from its immediate neighbourhood. One of the most popular local feature is the Scale-Invariant Feature Transform (SIFT) [18], which finds corners in images by difference-of-gaussians method and pinpoint corners as keypoints, stored in 128-dimentional vectors. SIFT descriptors are invariant to rotation and scale.

2.2.2 Other Features

Aural features, analog to visual features, represent specific properties of audio signals in compact format, allowing comparison with other audio signals. One of the most well known aural feature extactors, Mel-frequency cepstral coefficients (MFCC) [10] were originally applied on monosyllabic word recognition. Although it wasn't the focus at the time, MFCC was also found to be useful in speaker recognition.

Textual channel of video tends to be a lot smaller than the aural and visual channels and, by this fact, there are some TVSS methods which compare the raw textual data without any processing.

There are, however, lots of textual feature extractors in literature, mainly originated from Natural Language Processing and Data Mining. One of the most popular textual feature extractor is the Bag-of-Words (BoW) which generate a simplified representation of a text [2].

2.3 Mid and High Level Features

High level features are human-interpretable information describing some content [20]. They can be generated by annotation, indexing or generated automatically by analysis of low level features. It is usually text describing the media. For example, the objects contained in a video shot, like a car and a motorcycle. It can also represent sentiment of this media, like a shot that depicts anger of a character.

Mid level features, which are numeric or symbolic measurements obtained after analysis of low level features [20], are placed between low and high level features, as they remain close to media level information, without attempting a high level description of the content. A mid level feature example is the Bag-of-Features (BoF), which is an extension of the BoW model to other data natures like audio and video.

BoF appeared as an analog to the BoW model. Preliminary BoF works, from Computer Vision, replaced texts by images and the words by textons, which are prototype vectors that represent textures. Other authors, when referencing and describing visual oriented BoF models, call them Bag-of-Visual-Words (BoVW) because the words are replaced by some representation of visual features from images. Recently, BoF model have been extended to other natures like the Bag-of-Aural-Words (BoAW).

2.4 Multimodality

Multimodality in video, according to [24] is "The capacity of an author of the video document to express a predefined semantic idea, by combining a layout with a specific content, using at least two information channels".

In multimedia systems, is necessary to define strategies to combine data from the different sources and to exploit correlation between them, if any exists. The combination of features from different media, is called multimodal fusion. According to [4], the fusion of different modalities is generally performed at two levels: early fusion and late fusion.

In an early fusion approach, there is a combination of n feature vectors, producing a new feature vector which contains combined information. Then, the combined feature vector is fed to the decision unit, producing the answer.

In late fusion, there is an individual decision step for each one of n feature vectors, producing n partial answers to the task. These n partial answers are then submitted to a decision level fusion unit, which combines these partial answers, producing a final answer.

2.5 Temporal Video Scene Segmentation Evaluation Metrics

Evaluation of multimedia systems can be objective, when the evaluation doesn't involve user decisions or subjective, when its based on the quality of user experience. Objective evaluation of TVSS methods, however, still have a portion of subjectivity in it because it depends of a subjective concept which is scene definition. Two individuals may disagree on scene boundaries of a video because the semantic correlation of shots is relative to how each person comprehend it.

A TVSS method can be evaluated by comparison of its scene boundary prediction with a ground truth segmentation, which is a previous annotation of the true scene boundaries in a video. Creation of a ground truth, however, is a tricky task, because of the scene definition subjectivity. A way to minimize the possibility of errors, is to consensually combine several human generated annotations of a video.

Given a ground truth segmentation, there are some metrics to compare it to one obtained by a TVSS method. They are precision and recall, coverage and overflow, and the differential edit distance.

Precision and recall measure proportions of scene boundaries that have been correctly identified. Precision is the proportion of correctly identified scene boundaries between all predicted boundaries. Recall is the proportion of correctly identified boundaries between all true boundaries. To simplify comparison between different methods by precision and recall, they are usually combined in the F1 measure, which is the harmonic mean of precision and recall with equal weights to both.

Coverage and overflow are measures proposed by [26] and evaluate a scene prediction according to how much its shots overlap with the ground truth scene segmentation. Coverage of a scene t measures the quantity of shots of t correctly grouped together in the prediction and Overflow measures how many shots of t aren't covered by the prediction.

The Differential Edit Distance (DED) is a TVSS evaluation measure, proposed by [21], to overcome the weakness of existing TVSS evaluation measures when dealing with cases of ambiguity of scene definition. DED measure is based on the well known edit distance. In DED, scene segmentation is seen as a label assignment task, where shots belonging to the same scene must be labeled similarly. DED value between a predicted segmentation and a ground truth is given by the minimum number of shots that must be relabeled to transform the prediction sequence of labels into the ground truth sequence of labels.

3. RELATED WORK

In this Section there is a brief discussion on the main TVSS works, to expose research gaps in this topic. There is also some discussion on early fusion multimodal methods in order to justify our approach.

One of the first methods for the TVSSP was proposed by Yeung *et al.* [27]. It is based on low level visual features only and introduce the Scene Transition Graph (STG), a graph modelling to the TVSS task, which have been broadly adopted as a core module in later TVSS works.

Sundaram and Chang [25] proposed one of the first multimodal TVSS methods. It segments the video individually for each modality based on low level features and using sliding windows and then combines the resulting segmentations with a nearest neighbour algorithm.

Another famous work is the one by Zhai and Shah [28], where the authors present a Markov Chain Monte Carlo technique to determine the scene boundaries. It is also based only on low level visual features.

Chasanis *et al.* in [9] proposed the use of BoVW to represent visual content, as a form of preserving contextual information. Their approach is visual based only, using SIFT descriptors to compute mid level BoVW representations.

In the last five years of research, the main domain independent TVSS works found in the literature were:

- Sidiropoulos *et al.* [22]: multimodal approach, builds four STG's, two for low level features, one aural and one visual, and two for high level features, also one aural and one visual. It uses a probabilistic approach to merge the individually predicted scene boundaries of each STG;

- Lopes *et al.* [17]: multimodal approach, extends the BoVW method of [9] by computing also a BoAW. The two mid level features, BoVW and BoAW are used to segment the video individually and the individual results are combined afterwards;
- Baraldi *et al.* [5]: multimodal approach, based on visual and textual features. In their approach, mid level features for both visual and textual features are computed for each shot and then fed to an deep neural network which computes similarity scores for each adjacent pair of shots.

Appointing state-of-the-art TVSS methods is a delicate matter as available TVSS methods cannot be straightforward compared. It happens because authors adopted different datasets and evaluation strategies to this date [11]. No standard benchmarks are available and only one author recently provided source code and adopted datasets [6, 5].

Although it isn't possible to accurate map the state-of-the-art in TVSS, some trends can be noted with the evolution of TVSS methods. One of these trends seems to be the use of multimodality in attempt to capture all the possible clues to scene boundaries as some information can be expressed only on a modality or the clues to scene boundaries can be only achieved by a combination of information from different natures. Despite multimodality being popular, no early fusion method for the TVSSP with true feature fusion have been found. Some methods claim to be early fusion, but instead of mixing parts of features, they only feed classifiers with different nature features or concatenate feature vectors. Another trend is the use of mid and high level features in attempt to bridge the semantic gap.

Jhuo *et al.* proposed in [15] a feature fusion method for the video event detection task. Their method does something very similar to what is wanted in this work: they build multimodal words using aural and visual information. Their method, based on bipartite graphs, focus on discovering correlations between aural and visual words. They highlight that events are naturally multimodal and have consistent audio-visual patterns. The authors claim that this approach reduces dimensionality of the features, provides strong cues and discriminative power for detecting events.

4. PROPOSAL

Even over twenty years of research, temporal video scene segmentation is still an active research topic because it is useful on diverse tasks and, to the best of our knowledge, there isn't an appropriate domain independent solution for the TVSSP up to this date.

As no satisfactory method have been developed and no true early fusion approach to the TVSSP was found on the literature, this research aims to measure the impact of early fusion in both effectiveness and efficiency on the TVSS task. The following null hypothesis can be used to assess its impact on effectiveness:

Hypothesis 1. An early fusion method cannot obtain better results than a late fusion method when applied to TVSSP.

and its alternative hypothesis as:

Hypothesis 2. An early fusion method can obtain equal or better results than a late fusion method when applied to TVSSP.

A null hypothesis for assessing efficiency can be stated as:

Hypothesis 3. An early fusion temporal scene segmentation approach always demands higher computational cost than a late fusion approach.

and its alternative hypothesis as:

Hypothesis 4. An early fusion temporal scene segmentation approach doesn't always demands higher computational cost than a late fusion approach.

Motivated by the clues found in literature and presented on Section 2, this work will focus on refute Hypothesis 1 and Hypothesis 3 and accept Hypothesis 2 and Hypothesis 4, by experiments with the method described on Section 4.1.

4.1 An Early Fusion Approach to the TVSSP

The multimodal early fusion method illustrated in Figure 2 is proposed to address the TVSSP.

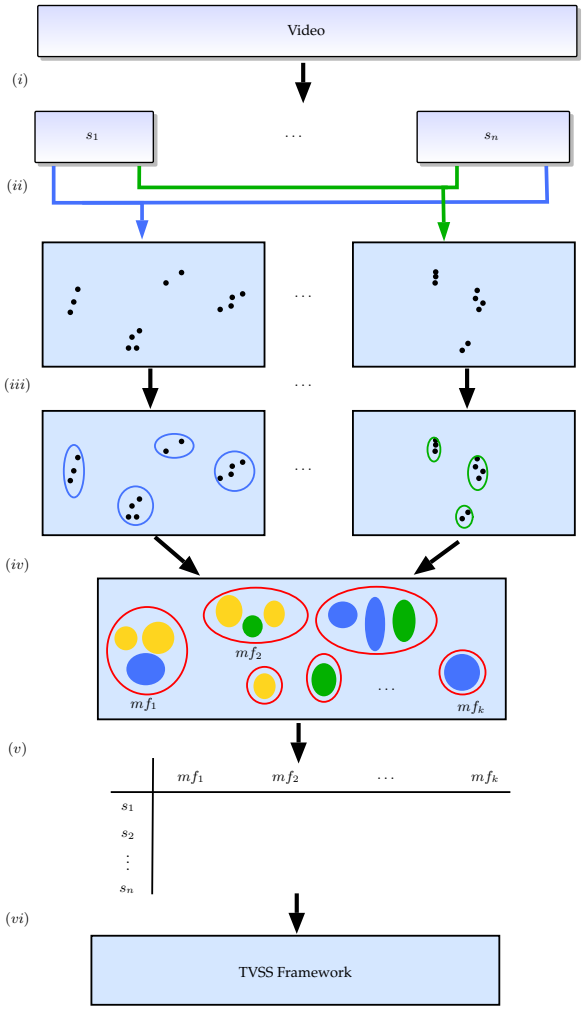

Figure 2: Proposed early fusion method.

From this point, it will be called Early Fusion Scene Segmentation Method (EFSSM). Initially, given a video V as input, is necessary to find its segmentation into shots $S =$

$\{s_1, \ldots, s_n\}$. This action corresponds to (i) in Figure 2. A previously annotated shot segmentation ground truth can be used as a solution to this step. There are three available datasets with annotated shot segmentation ground truths. They are detailed in Section 4.2. Shot segmentation can also be performed by an available state-of-the-art shot segmentation method, like the one in [3].

In (ii), features from different modalities will be extracted from shots. For each shot s_i in S, feature vectors should be extracted for each of J modalities. In the initial version of the EFSSM, two modalities will be explored: visual and aural. For the visual modality, SIFT will be used. Aural modality will be initially explored via MFCC descriptor, as it have been used with success in the TVSSP [17]. There are plans to include other modalities and to experiment different low-level descriptors than SIFT and MFCC in attempt to improve the method. Mid and high level can also be easily coupled with EFSSM and will be experimented.

Step (iii) consists in the application of the BoF model individually on each modality. For each modality, a feature dictionary is created by clustering its feature descriptors using k-means and representing the computed clusters by their centroids. Feature dictionary size of each modality will be defined after experimenting different values in a supervised operation. A distance measure should be carefully defined to be used with each distinct type of feature descriptor. Compute then, for each shot s_i in S, feature histograms $h_{ij} \in H$ for each modality j in J. Each cell $h_{ij}[l]$ of a feature histogram h_{ij} contains a measure of how many times feature l (*i.e* visual or aural word) from modality j appears on s_i.

Consider F as the set of all features computed from all J modalities. Observe the fact that feature histograms H can be used, in a different perspective, to know how many times each feature $f \in F$ happens on each shot. In other words, for each f there is shot pertinence histogram. These shot pertinence histograms are used in (iv) to cluster features $f \in F$ using k-means. The idea behind this step it to capture the co-ocurrence of features from different modalities in shots. Each cluster produced in this step corresponds to one **multimodal feature** mf_k.

The goal of (v) is to compute histograms of multimodal features for each shot. A preliminary strategy to perform this step is to combine the different modalities histogram values from the all the unimodal features composing a multimodal feature from each of the J modalities, by arithmetic mean. Different weights for different modalities and normalization of values can be useful in this calculation.

Finally, in (vi) the multimodal feature histograms are fed to a TVSS framework, which will produce the final scene segmentation. There are many TVSS frameworks with different approaches that can be adopted, like the STG [27], and sequence alignment [9].

Despite the proposed feature fusion technique seems to be simpler than the bipartite graph approach [15], it is also possible to experiment the bipartite graph approach to build the BoMF, possibly extending it to a multipartite graph approach, if the clustering approach fails to exploit correlations between different modalities.

4.2 Experiment Design

EFSSM will be implemented in Java with OpenImaj API[3] and R/JRI[4]. OpenImaj is a set of libraries and tools for multimedia content analysis and generation. R is a statistical computing software, which can be acessed by Java language through JRI, a Java/R interface. With the implementation of EFSSM, is possible to compare it to other approaches and check hypotheses 1, 2, 3 and 4.

To test Hypothesis 1, EFSSM and a Late Fusion Scene Segmentation Method (LFSSM) should be executed over a scene segmentation dataset with ground truths segmentations. There are three options for the dataset and they should all be tested: RAI Dataset [6], BBC Dataset [5] and the Movies Dataset from [17]. BBC Dataset is a set of 11 episodes from a educational TV series called Planet Earth. RAI Dataset is a set of 10 videos including documentaries and talk shows from the Rai Scuola video archive. Movies Dataset is actually composed of five Hollywood movies. All three described datasets contains manually aannotated ground truth scene segmentations. The obtained results will be compared to the respective ground truths segmentations employing the evaluation measures explained on Subsection 2.5. If results from EFSSM are better than results from LFSSM, we can refute Hypothesis 1 and accept Hypothesis 2.

By measuring the running time of both EFSSM and LFSSM is also possible to refute Hypothesis 3 and accept Hypothesis 4 if EFSSM uses less or equal time to run.

4.3 Schedule

The schedule of the research, presented at Table 1, refers to the trimestral execution of the following activities:

(i) Attaining mandatory credits of ICMC-USP PhD program; *(ii)* English proficiency exam; *(iii)* Teaching Improvement Program (PAE); *(iv)* Systematic mapping of TVSS and related work; *(v)* Check for updates on TVSS state-of-the-art; *(vi)* Study, coding and analisys of TVSS methods; *(vii)* Experimental analisys of developed methods; *(viii)* Qualifying exam: writing and examination; *(ix)* Thesis writing; *(x)* Result publication by technical reports, paper submission to the area related events and periodicals; *(xi)* Thesis submission and examination.

Items marked with $\surd$ are finished activities and items marked with ● are upcoming or in progress tasks.

Year	**2015**				**2016**				**2017**				**2018**			
Trimester	1°	2°	3°	4°	1°	2°	3°	4°	1°	2°	3°	4°	1°	2°	3°	4°
(i)	√	√	√	√												
(ii)			√													
(iii)			√	√												
(iv)					√	√	√									
(v)								●	●	●	●	●	●	●	●	
(vi)			√	√	√	√	√	●	●	●	●	●				
(vii)								●	●	●	●	●	●	●		
(viii)						√	√	●								
(ix)													●	●	●	
(x)								√	●	●	●	●	●	●	●	
(xi)															●	●

Table 1: PhD activities schedule.

5. CONCLUSIONS

This paper proposes the execution of a PhD project themed on the use of early fusion in TVSS in order to answer if it improves effectiveness and efficiency on the task. The achievement of this answer is the main expected result from this work. Aside from the main result, there are some other expected results: Human resource development, as a PhD

[3]http://openimaj.org/

[4]https://www.r-project.org/, https://rforge.net/JRI/

and possibly an undergraduate scientific initiation; Publication of results in scientific events and journals; A multimodal early fusion based TVSS system; A dataset and related ground truth scene segmentations of videos.

6. REFERENCES

[1] Cisco visual networking index: Global mobile data traffic forecast update, 20152020. White Paper, Feb. 2016.

[2] C. C. Aggarwal and C. Zhai. *A Survey of Text Classification Algorithms*, pages 163–222. Springer US, Boston, MA, 2012.

[3] E. Apostolidis and V. Mezaris. Fast shot segmentation combining global and local visual descriptors. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6583–6587, May 2014.

[4] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Syst.*, 16(6):345–379, Nov. 2010.

[5] L. Baraldi, C. Grana, and R. Cucchiara. A deep siamese network for scene detection in broadcast videos. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1199–1202, New York, NY, USA, 2015. ACM.

[6] L. Baraldi, C. Grana, and R. Cucchiara. Scene segmentation using temporal clustering for accessing and re-using broadcast video. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[7] T. T. S. Barbieri, T. H. Trojahn, M. P. Ponti-Jr, and R. Goularte. Shot-hr: A video shot representation method based on visual features. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '15, pages 1257–1262, New York, NY, USA, 2015. ACM.

[8] N. Chapman and J. Chapman. *Digital Multimedia*. Wiley Publishing, 3rd edition, 2009.

[9] V. Chasanis, A. Kalogeratos, and A. Likas. Movie segmentation into scenes and chapters using locally weighted bag of visual words. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 35:1–35:7, New York, NY, USA, 2009. ACM.

[10] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, Aug 1980.

[11] M. Del Fabro and L. Böszörmenyi. State-of-the-art and future challenges in video scene detection: a survey. *Multimedia Systems*, 19(5):427–454, 2013.

[12] Y. Dong, S. Gao, K. Tao, J. Liu, and H. Wang. Performance evaluation of early and late fusion methods for generic semantics indexing. *Pattern Analysis and Applications*, 17(1):37–50, 2014.

[13] B. M. Gross. The managing of organizations: The administrative struggle, vols. i and ii. *The ANNALS of the American Academy of Political and Social Science*, 360(1):197–198, 1965.

[14] J. Huang, Z. Liu, and W. Yao. Integration of audio and visual information for content-based video segmentation. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, pages 526–529 vol.3, Oct 1998.

[15] I.-H. Jhuo, G. Ye, S. Gao, D. Liu, Y.-G. Jiang, D. T. Lee, and S.-F. Chang. Discovering joint audio–visual codewords for video event detection. *Machine Vision and Applications*, 25(1):33–47, 2014.

[16] I. Koprinska and S. Carrato. Temporal video segmentation: A survey. In *Signal Processing: Image Communication*, pages 477–500, 2001.

[17] B. L. Lopes, T. H. Trojahn, and R. Goularte. Video Scene Detection by Multimodal Bag of Features. *Journal of Information and Data Management*, 5(2):194, 2014.

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[19] V. Luckerson. Netflix accounts for more than a third of all internet traffic, May 2015.

[20] J. Martinet and I. El Sayad. Mid-level image descriptors. In Z. Ma, editor, *Intelligent Multimedia Databases and Information Retrieval: Advancing Applications and Technologies*, pages 46–60. IGI Global, 2012.

[21] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, and J. Kittler. Differential edit distance: A metric for scene segmentation evaluation. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(6):904–914, June 2012.

[22] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Trans. Cir. and Sys. for Video Technol.*, 21(8):1163–1177, Aug. 2011.

[23] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, Dec. 2000.

[24] C. G. M. Snoek and M. Worring. A review on multimodal video indexing. In *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, volume 2, pages 21–24 vol.2, 2002.

[25] H. Sundaram and S.-F. Chang. Video scene segmentation using video and audio features. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 2, pages 1145–1148 vol.2, 2000.

[26] J. Vendrig and M. Worring. Systematic evaluation of logical story unit segmentation. *IEEE Transactions on Multimedia*, 4(4):492–499, Dec 2002.

[27] M. Yeung, B.-L. Yeo, and B. Liu. Segmentation of video by clustering and graph analysis. *Comput. Vis. Image Underst.*, 71(1):94–109, July 1998.

[28] Y. Zhai and M. Shah. Video scene segmentation using markov chain monte carlo. *IEEE Transactions on Multimedia*, 8(4):686–697, 2006.

[29] H. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, 1993.