Video Scene Detection by Multimodal Bag of Features

Bruno Lorenço Lopes, Tiago Henrique Trojahn, Rudinei Goularte

Universidade de São Paulo, Brazil
{blopes, ttrojahn, rudinei}@icmc.usp.br

Abstract. Recent advances in technology have increased the availability of video data, creating a strong demand for efficient systems to manage this kind of content. To make efficient use of video information, first, the data have to be automatically segmented into smaller, manageable and understandable units, like scenes. This article presents a new multimodal video scene segmentation technique. The proposed approach combines Bag of Features based techniques (visual and aural) in order to explore the latent semantics obtained by them in a complementary way, improving the scene segmentation. The achieved results showed to be promising.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Scene detection; I.2.10 [**Vision and Scene Understanding**]: Video analysis; I.4.6 [**Image Processing and Computer Vision**]: Segmentation; H.3.3 [**Information Search and Retrieval**]: Shot Detection

Keywords: audio descriptors, bag of features, multimedia, scene detection, visual descriptors

1. INTRODUCTION

Sales of digital devices capable of capturing digital images and videos, like video cameras and smartphones, are increasing every year. This fact led to a significant rise on the digital content (photos, videos and audio) generation and sharing. Systems like YouTube[1] and Instagram[2], which provide services allowing people to share digital content, have become very popular nowadays mainly because users can choose when they want to access an specific content.

However, it is hard to find relevant content due to the huge volume of data provided by systems. This problem, associated with the current demand of multimidia applications for interactivity, has stimulated recent research initiatives aiming to provide easy and transparent content access, specially in areas like Information Retrieval, Content-Based Retrieval, Summarization and Content Personalization. All these areas share a common requirement: knowledge about the content. This knowledge is known as metadata and, generally, is used as a way to provide semantic information capable of reducing the semantic gap [Smeulders et al. 2000]. In digital videos, the first stage to acquire semantic information is to segment the video into smaller units, computationally easier to deal with: frames, shots and scenes.

There are many proposed methods on the literature that segment video into shots, exploring techniques ranging from histograms and space-frequency transforms to multimodality. The shot concept is defined as a series of consecutive frames continually captured by a single camera, representing a continuous action in time and space [Chianese et al. 2008]. The state-of-the-art in this area shows methods that ensure a high precision [Smeaton et al. 2010]. However, in general, shot detection

[1] http://www.youtube.com
[2] http://www.instagram.com

This work was supported by grants #2011/05238-0 and #2011/00422-7, São Paulo Research Foundation (FAPESP), and by National Council for Scientific and Technological Development (CNPq).

methods generate a large number of segments, not necessarily related with how users commonly understand events, actions or stories. Due to these facts, video segmentation in human semantically comprehensible units is challenging, and opens space to the scene segmentation.

Scene segmentation is challenging mainly because it relies on the subjectivity of the scene concept: a set of consecutive shots semantically related [Wang et al. 2006]. In spite of the advances made in the field, related work (see Section 2) focuses on finding video segments (scenes) containing single objects (persons) or events, not exploring video segments containing a closer relation with the common users' understanding of scene, which should be characterized by the semantic relation between the shots that compose it. We call those segments semantically complex segments. A way to establish the semantic relation is through the shot subject: if the subject of two consecutive shots changes, then, we should have a scene transition. For instance, considering a TV news program, for a video segment between two appearances of the same anchorman (the presenter of the news program), composed of three visually different shots, but that are semantically related, the most appropriate segmentation would find only one scene. However, most of related work, using narrower scene definitions, will find three scenes.

Since semantically complex segments are semantically richer than common video segments, it is harder to detect scenes belonging to them, and, as a consequence, new approaches must be developed. An interesting approach to face that problem, as far as we know not explored yet, is a Multimodal Bag of Features. Bag of Features (BoF) has been used with success to capture latent semantics in the Image Retrieval domain. This technique extracts features - local descriptors - from images, clustering together the similar ones and building dictionaries (codebooks) of "visual words" [Yang et al. 2007], which can be used later for retrieval purposes. During the retrieval phase, the visual word from an input query image is compared with those in the dictionary. Due to the high discriminative power of the local descriptors, BoF is very efficient in discovering the group (cluster) that each object belongs to, although it is not able to classify the groups [Yang et al. 2007]. This latent semantics can be used to find out which video shots belong to the same scene (group).

In this article we explore BoF in a multimodal fashion. First, we use only visual features to segment a video into scenes. Then, we use only audio features. Finally, we combine their results in a process called late fusion, improving segmentation. The different media nature (aural and visual) can capture intrinsic complementary semantics present in the video shots, allowing us to identify scenes in semantically complex segments. This article is organized as follows. Section 2 presents a review on scene segmentation techniques. Section 3 details our scene detection techniques based on the BoF model. In Section 4, we report our experiments and their results. Finally, Section 5 presents some final remarks about our work.

2. SCENE SEGMENTATION

Scene segmentation is a very important step in content indexing, retrieval and personalization, since scene is a concept that is semantically meaningful for the user. In this section, some important related work in this area is presented, discussing how it is related to semantically complex video segments.

Regarding scene boundary detection, most approaches relies on news videos. Chaisorn et al. [2002] demonstrate that TV news structure can be organized as: opening, composed of the most important news summary, and the main part, in which the news are organized in accordance with geographic interests and categories such as politics, entertainment and sports. Moreover, usually a scene begins with the anchorman introducing the subject that will be presented. This highly structured composition of news videos have been used as a clue to detect scene boundaries. Chaisorn et al. [2002] propose a two level method. The first level identifies and classifies shots in thirteen pre-defined news categories, using a decision tree. The second identifies scenes, by the means of Hidden Markov Models. In another common approach, Liu et al. [2008] use anchorman detection to segment scenes. They developed a content personalization and adaptation system for three-screen services based on users profiles, where the acquired content is submitted to analysis modules - shot and anchorperson detection - before the multimodal scene segmentation. This kind of systems needs to find the anchorman in order to detect scene boundaries because they assume that a scene will be between two appearances of the anchorman. Those approaches take advantage of the news domain structure or use a narrow scene definition, making difficult to apply them to other video domains or to handle semantically complex segments.

Efforts have been made towards domain-independent segmentation techniques. Yeung et al. [1998] proposed a video representation using scene transition graphs. In this graph each shot is a node and the edges are the transitions between shots. The graph can be divided using a similarity restriction in order to represent scenes. Although the presented technique is based on visual features, which are domain-independent, the authors agree that using semantics could enhance their technique. Hanjalic et al. [1999] used a similar technique for MPEG videos, identifying what they call logical story units (LSU), in the movies domain. They defined LSU as being "the series of temporally contiguous shots, which is characterized by overlapping links that connect shots with similar visual content elements". Sakarya and Telatar [2007] developed a graph partition based method to detect scenes transitions, using multiple features. A unidimensional signal is constructed for each shot similarity feature. Next, the signal is filtered, and any unnecessary information is eliminated. Finally, the K-Means algorithm is used to group visually similar shots and obtain scene transitions candidates. These approaches exclude semantically complex segments, which are characterized by subject - visually dissimilar shots may have the same subject, so, should be part of the same scene.

Still towards a domain independent approach, a remarkable technique is the one proposed by Rasheed and Shah [2003]. Initially the video is segmented into shots. Then, one or more keyframes are extracted from each shot and their histogram's dissimilarities and motion vectors are calculated. Next, a metric called Backward Shot Coherence (BSC) is applied to identify shots that are candidate to be in the same scene. The last step consists in comparing the the amount of motion in each scene (based on the MPEG motion vectors), and, if they are similar, these scenes are clustered as a unique scene. The technique was experimented using different video domains (like Hollywood movies and talk shows). The obvious drawback is that the technique is dependent of the video file format - it works only with MPEG-1 compressed video files, from where it extracts the used motion vectors.

Regarding the use of semantics in scene detection, most of related work makes use of Bag of Features (BoF) [Zen et al. 2012] technique, since it can exploit latent semantics. However, the word *scene* is used as synonymous with *action*. For example, Kläser et al. [2010] proposed a method for scene (action) detection in videos using BoF. Person and upper body detectors are used to create boxes that limit the actors and their closest neighbouring areas. Only features contained in the limiting boxes are used. The code book is generated with 4000 features, sampled in a random way. According to the authors, this is a faster way than using K-Means to generate the code book, with similar performance. Three databases were chosen: KHT, UCF and Hollywood, containing respectively normal daily actions, sport actions and common actions in movies. This work differs from our proposal, since it consists in determining previous established actions, not semantically related segments. But it demonstrates the power of BoF to recognize the visual semantics present in the video. As another example, Yang et al. [2007] presented a scene classification technique based on BoF. In their work, features were extracted from keyframes and they applied techniques used in text categorization, such as stop words removal, term weighting and feature selection. This work relates to our in the sense that BoF is used in videos, but their aim is to classify scenes, while ours is to segment the video. In a different, and rare, approach, Chasanis et al. [2009] proposed a method that uses locally weighted Bag of Visual Words to detect scenes and chapters in movies. Although the method presented good results, it understands a scene as a set of shots representing a continuous action or taking place in the same location, but do not addressing semantically related segments.

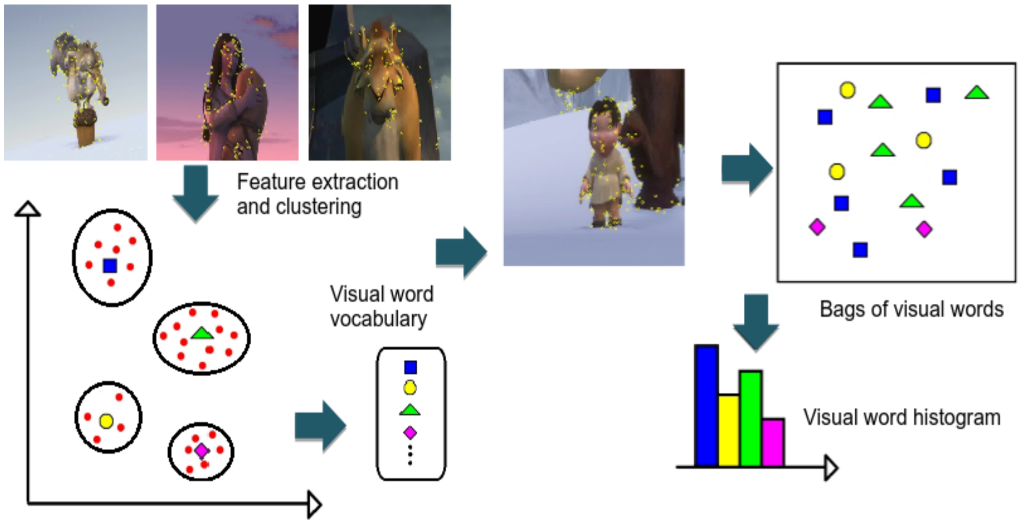

Fig. 1: Bag of Features technique (image adapted from [Yang et al. 2007]).

In summary, most of the techniques to segment videos in scenes presents one or more of these limitations: analyze only the video stream, not addressing complex segments with low visual similarities [Rasheed and Shah 2003; Chasanis et al. 2009]; have video file format dependency [Rasheed and Shah 2003], or video domain restrictions; use a very strict scene definition, different from the common user understanding. These limitations make difficult to apply these techniques to detect semantically complex segments. For example, a technique developed with a specific scene segmentation could use a particular feature to determine the scene transition. Another problem is that complex segments cannot be fully detected when using only the video stream: sometimes, even when the visual information remains the same, the subject of the scene changes. In this sense, we propose a technique which aims to overcome some of these problems. The technique, which is described in Section 3, uses the BoF in order to capture latent semantics and then segment videos into scenes - including scenes in semantically complex segments.

3. SCENE SEGMENTATION USING BAG OF FEATURES

In this article, we adopt the same scene definition presented by Zhai and Shah [2006]: “A scene or a story is a group of semantically related shots, which are coherent to a certain subject or theme”. This definition links scene with the subject, instead of the place or the objects presented in each shot. Moreover it is independent of the video domain. Based on this definition, our three step technique detects scenes by combining two BoF based approaches and a fusion method. Since BoF is a key concept to understand our technique, it is briefly explained in Subsection 3.1. Subsection 3.2 details the first module of our technique, which uses visual descriptors in order to detect scenes. The second module, based on audio descriptors, is described in Subsection 3.3. Finally, Subsection 3.4 presents our strategy to combine the results from the first two modules in a late fusion way.

3.1 Bag of Features

Bag of Features (BoF) is based on the successful Bag of Words technique, widely used for document classification and document retrieval. The main idea behind Bag of Words is to represent a text by computing the frequency of the words from its vocabulary. Similarly, BoF can be usually applied to image databases, such as the set of three images illustrated in Figure 1. A visual feature extractor[3],

[3] An algorithm to detect and describe visual features.

like SIFT (Scale-Invariant Feature Transform) [Lowe 2004], can be used to identify Points of Interest (PoIs) in every image (keyframes from shots previously segmented) and to compute a multidimensional vector for each PoI. These vectors are called visual descriptors. PoIs are represented by the yellow points in the images and, for illustration purposes, descriptors are represented by the red 2d points.

In a multidimensional space, those descriptors can be organized in clusters, such that close (similar) vectors tend to be in the same cluster. A clustering algorithm, like K-Means [MacQueen 1967], can be applied to find K clusters and their K centroids, i.e., elements that represent the general behavior of each cluster (illustrated in Figure 1 by the blue, yellow, green and purple polygons). These centroids are called visual words, and the set of visual words defines a codebook. The size K of the codebook is an important parameter for the BoF performance. Smaller values reduce computational requirements, as memory and processing power, but it also may reduce the discriminative power of the technique, since significantly different descriptors may be clustered together and thus be represented by the same visual word. On the other hand, high values increase computational requirements, and very similar descriptors may be represented by different visual words. Therefore, it is important to determine a size K that fulfills both precision and computational requirements, considering the characteristics of each kind of application. Although there are methods to automatically determine K [Chiang and Mirkin 2010], the most common approach is testing a range of K values to find the best result.

With the codebook, a new image can be represented by a histogram with the amount of descriptors that are part of every cluster found. That histogram is obtained by finding, for each descriptor of the image, the closest visual word in the vocabulary, according to a distance function. The histograms computed in that way are a compact image representation and may be used to find (latent) semantically related images [Cai et al. 2012]. So, images with similar histograms tend to be semantically similar in visual terms.

3.2 Scene Segmentation Technique with Visual Descriptors

As stated before, BoF is a technique that can extract latent semantics from images. It is a desirable property in the context of scene segmentation, since shots that belong to the same scene tend to preserve some visual elements, like actors or objects [Kumar et al. 2011]. Thereby, scene segmentation can be seen as the act of grouping semantic related shots. So, the BoF technique could be used to represent every shot in the video, and then, similar shots could be clustered.

However, applying BoF to all frames from one video shot, in order to create a representation of it, can be computationally expensive since a shot can be composed of a large number of frames. One possible strategy is to select a reduced set of frames that can properly represent each shot. In our technique we employed the keyframe selection method presented in Section 3.2.1. After selecting representative frames from each shot, we extract SIFT visual descriptors (we use the library developed by the SIFT author [Lowe 2004] and available at http://www.cs.ubc.ca/~lowe/keypoints) from those frames and apply BoF to compute a histogram of each shot. Each histogram is then smoothed and the scene cuts are detected. These two steps are detailed in Section 3.2.2.

3.2.1 *Keyframes Detection.* The method described here was adapted from [Chasanis et al. 2008]: a 3D HSV normalized histogram is used to represent each frame from one shot. Each histogram has 8 bins for hue, 4 bins for saturation and 4 bins for value. The spectral clustering method [Ng et al. 2001] organizes all the histograms of the shot in k groups. The number of groups k is determined automatically and can vary for each shot. For each group, the medoid (the histogram with maximal average similarity to all the others) is selected and its corresponding frame is considered one of the shot's key frames. This method reaches high precision of shot representation [Chasanis et al. 2008] and is able to define automatically the number of keyframes per shot - two desirable properties in the domain.

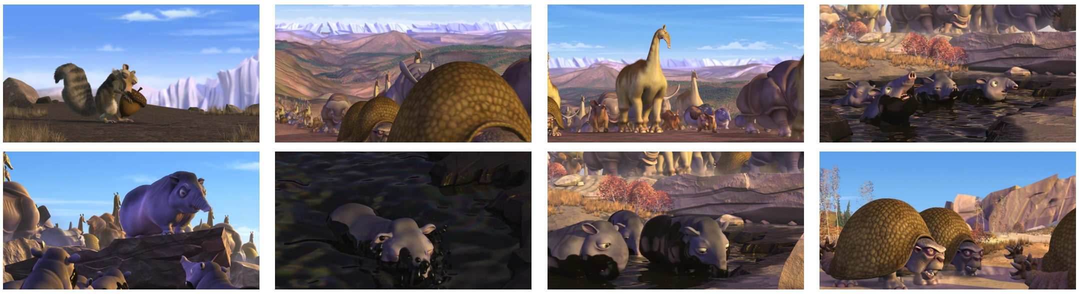

Fig. 2: Key frames from shots of one scene of the Ice Age movie.

3.2.2 *Scene Cut Detection.* Although BoF histograms can preserve the semantics of similar shots, sometimes consecutive shots may be visually different, while belonging to the same scene. One example is a phone dialogue scene between two characters. They may be calling from completely different places, and the histograms of these scene shots can be very different, but these shots appear intercalated. In this case, a visual technique could segment the scene in many shots, although they are semantically related. Therefore, it is desirable to indicate that neighboring shots can influence a specific shot. Near shots have stronger influence than the farther ones. This situation can be seen in Figure 2, that shows keyframes from a sequence of shots that are in the same scene. In this scene, characters from a movie (Ice Age) are talking while migrating to escape from a new ice age. It is important to notice that some actions occur along the migration, such as the children having fun in shots 4 to 7, which are visually dissimilar when compared with other shots of the same scene.

In text retrieval, the idea of neighboring elements influencing each element has been explored by using a local smoothing kernel [Lebanon et al. 2007], which determines how much one word will interfere with their neighboring words. It is possible to apply the same idea to the video domain [Chasanis et al. 2009]. Words of a text can be translated to shots of a video as well as paragraphs to scenes, since each paragraph is a piece of the whole subject of the text and each scene represents one of the subjects of the video. Shot normalized histograms can be smoothed with the use of a Gaussian kernel. Considering a shot t, its smoothed histogram SH is computed from its normalized histogram as follows:

$$SH_t = \sum_{n=1}^{N} H_{t-n}.K_\sigma(t-n), \tag{1}$$

K is a Gaussian kernel with zero mean and standard deviation σ. N is the number of shots. The number of neighboring shot histograms used to compute the smoothed histogram depends on the σ value. Therefore, σ will determine the number and the level of preserved contextual information. Higher σ values increase the number of shot histograms, increasing shot grouping into one scene. We have tried a range of integer numbers varying from 4 to 12 and, as the value 8 presented best results, we chose set $\sigma = 8$ in our experiments. In order to determine scene cuts we computed the Euclidean Difference between consecutive histograms to find the local maximum.

3.3 Scene Segmentation Technique with Audio Descriptors

Audio descriptors have been used in many works addressing problems like speaker recognition [Md. Rashidul Hasan 2004] and video copy detection [Liu et al. 2010]. Three of the best known audio descriptors are MFCC (Mel-Frequency Cepstrum Coefficients), LPCC (Linear Prediction Cepstral Coefficients) and Chroma.

The power of these descriptors to represent audio is very useful to solve many problems, including video segmentation. Figure 3 shows a scene cut that could be detected using those audio descriptors.

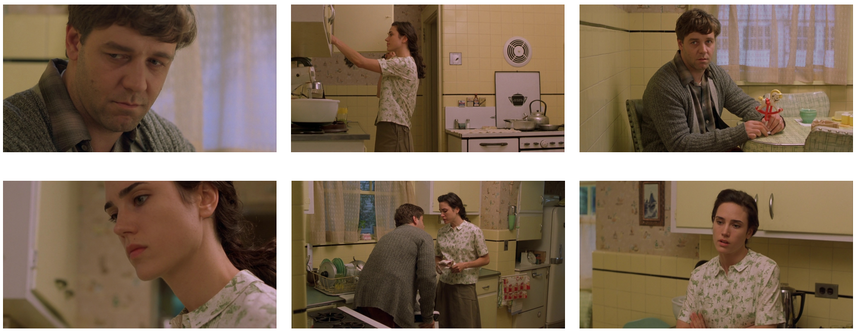

Fig. 3: Key frames from shots of A Beautiful Mind scene.

In that figure the first three shots, represented by one of their keyframes, are from a Beautiful Mind movie scene, where John and Alicia talk about John is doing home tasks. In the next three shots, Alicia tells John she thought that the garbage men didn't work at night. Although the shots are visually very similar (same place and same actors), from an audio perspective, they are different, since Alicia began to do the dishes, making a characteristic noise, and she changes the subject of the conversation.

Many works in audio and video segmentation areas makes use of audio descriptors [Sundaram and Chang 2000]. In general, these works need to a training base in order to be able to detect events or to determine the rules to classify scenes. Furthermore, the precision and recall of such techniques tend to decrease if they are used to segment videos or audios very dissimilar from the ones present in the base. For this reason, it would be good to use a technique that does not depend on a previous training. Thus, we decided to adapt our previous technique, presented in Subsection 3.2, to use audio descriptors. The details are described at Subsection 3.3.1.

3.3.1 *BoF and Audio Descriptors.* Aural and visual video feature extraction are non symmetric processes. Visual descriptors can be extracted from each video frame, or from each video keyframe. For a particular video frame many points of interest are detected and a respective visual descriptor is computed. So, each frame is represented by a set of visual descriptors. For audio it is a different situation. An audio segment can be sampled in a way that the number of samples is equal to the number of video frames. But while a frame represents one instant of the video, an audio sample represents a period.

Despite that difference, a sample could be seen as an audio frame. However, each audio sample generates only one audio descriptor, while each video frame generates many visual descriptors. So, if we directly apply BoF, each audio sample will be represented by a histogram with only one audio word. In this way, it would be unfeasible to calculate the level of similarity between segments, because they will be considered 1) equal, if they are represented by the same word or 2) completely different, if they are represented by different words. A possible way to overcome this problem is to reduce the sample period, increasing the number of audio descriptors computed between two video frames. In this way, one audio frame would be represented not only by one audio descriptor, but by a set of descriptors, which is equivalent to the approach that we have with visual descriptors. Details on how these adjustments were applied in our technique are presented at Subsection 3.3.2.

3.3.2 *Audio Descriptors Extraction.* We decided to use the well known MFCCs descriptors, which have been successfully used in tasks like speaker recognition [Md. Rashidul Hasan 2004]. In order to use these descriptors, we first generate WAV audio files from the selected videos with FFMPEG[4], with a sample rate of 44100 Hz. Then, we use MARSYAS[5] to compute audio descriptors with non

[4] https://www.ffmpeg.org/

[5] http://www.marsyasweb.appspot.com/

overlapped windows of 256 audio samples. Since videos from the database present 24 visual frames per second, we obtained approximately 1837 audio samples between 2 visual frames. This number, divided by the 256 samples used to generate one audio descriptor, results into 7 audio descriptors in each interval. These 7 audio descriptors are used to represent each audio frame (in the time interval between 2 video frames).

We chose to use all the audio frames for each shot, instead of selecting key audio frames, like in the visual features technique. Therefore, it increments the number of audio descriptors used to represent one shot, enhancing the technique representativeness for each shot. This is useful since audio frames from the same shot tend to be very different. Therefore, without a good representation strategy, it will not be feasible to measure similarity between consecutive shots.

3.4 Late Fusion Scene Segmentation Technique

The final step of our multimodal BoF-based technique is a strategy to fuse the aural and the visual segmentations. Essentially, there are two possible ways: early fusion and late fusion. In the early fusion approach, the aural and visual features are combined at an early stage, trying to explore the correlation between them. One difficulty that arises is the synchronization of features [Atrey et al. 2010]. The late fusion approach consists in combining local decisions, based on individual features, in order to produce a final decision [Atrey et al. 2010]. We chose the second approach to combine the results of visual and audio techniques. As discussed at sections 3.2 and 3.3, the visual and aural features techniques generate the same type of result. In other words, the two individual techniques return scene cuts, enabling the fusion. Moreover, this approach preserves the semantics behind the decision of each approach and, since the decisions are not necessarily overlapped, they can be complementary, improving the segmentation efficiency.

The main idea behind our technique is the following assumption: two scene cuts, one detected by the visual technique and the other detected by the aural technique, that are close enough to each other, tend to be the same cut. For instance, the visual technique considers that shot 45 is a scene cut and the aural technique considers 47. Even though the detected cuts are at different shots, it is not likely to have a whole scene between shots 45 and 47, so, we can assume they are the same cut. As mentioned before, there are situations in which a specific media may not be enough to detect a cut, being necessary to use another media to detect this cut. In this way, it is interesting to keep cuts detected by only one technique, let us say visual, since this can complement the cuts detected by the other technique, let us say aural.

With these two lines of thought, we determine the final decision, i.e., the result from our multimodal scene segmentation technique, using the following heuristic: If two detected scene cuts are less than 3 shots away, they are fused in only one, calculated by their mean; otherwise, each cut is held. With this approach, we ensure that near false positives from visual and audio techniques will be fused, what will reduce the total number of false positives. This can lead to new true positives, since the average of them may be correct. Moreover, distant true positives will be held, increasing the number of them, and consequently increasing the technique's recall. The drawback is that distant false positives will be kept, what could affect precision. Those situations are analyzed in the next section.

4. EXPERIMENTS AND RESULTS

In this section we describe the experiments we conducted in order to evaluate the efficiency of our proposed technique. We start discussing the dataset. Video datasets are, in general, an open challenge in the video segmentation domain [Smeaton et al. 2010]. This is particularly true for video scene detection. We did not find any video scene segmentation database considering semantically complex

Table I: Movies in ground truth

Movie	Length (m)	Shots	Scenes
A Beautiful Mind	135	1656	95
Back to the Future	116	1354	121
Gone in 60 Seconds	117	2737	148
Ice Age	81	1384	64
Pirates of the Caribbean	133	2660	237

Table II: Results of video techniques applied in the movies "A Beautiful Mind", "Gone in 60 Seconds" and "Ice Age".

	A Beautiful Mind						Gone in 60 Seconds					
Technique	T_p	F_p	F_n	P	R	F_1	T_p	F_p	F_n	P	R	F_1
VT_1	49	30	45	62.25%	52.13%	56.65%	94	44	53	68.12%	63.95%	65.96%
VT_2	47	32	47	59.49%	50.00%	54.34%	87	52	60	62.59%	59.18%	60.84%
VT_3	58	54	36	51.79%	61.70%	56.31%	78	136	69	36.45%	53.06%	43.21%

	Ice Age					
Technique	T_p	F_p	F_n	P	R	F_1
VT_1	34	30	29	53.13%	53.97%	53.54%
VT_2	30	20	33	60.00%	47.62%	53.10%
VT_3	44	110	19	28.57%	69.84%	40.55%

segments. Even TRECVID[6], a well known video database, is not adequate to our needs, since it is composed of short length videos, with few scenes and very few semantically complex segments. So, we decided to develop our own ground truth (a database composed by the set of manually identified scene cuts) using a collection of well known movies with a representative number of semantically complex segments.

We chose five Hollywood movies: A Beautiful Mind, Back to the Future, Gone in 60 Seconds, Ice Age and Pirates of the Caribbean: the Curse of the Black Pearl. Those movies belong to different genres (variability) and are also evaluated in related work on scene segmentation. The ground truth for this dataset was built using human observers who manually identified the first and the last frame for each shot and the first and last shot for every scene. The last frame of a shot is considered a shot cut. In the same way, the last shot from a scene is considered a scene cut. The length of each movie in minutes as well as the number of shots and scenes are presented in Table I. To evaluate the performance of our technique with regard to other techniques found in literature, we chose three well established measures: Precision (P), Recall (R) and F-Measure (F_1).

In our first experiment, we compared only the visual module of our own technique (VT_1) with other two fully visual techniques: a BoF-based technique [Chasanis et al. 2009] (VT_2) and a Backward Shot Coherence (BSC) based technique [Trojahn and Goularte 2013] (VT_3), all using the same scene definition and dataset. Our goals are to demonstrate that these techniques cannot handle semantically complex segments and that BoF can present better results. This was demonstrated by showing that their performance decreases when they are applied to our ground truth. We implemented the two competing techniques and used three movies from our database: "A Beautiful Mind", "Gone in 60 Seconds" and "Ice Age". The results are shown, for each movie, in Table II, where T_p is true positive, F_p is false positive and F_n is false negative.

Techniques VT_1 and VT_2, which are BoF-based, presented better results for precision in all the three movies, compared to technique VT_3. Technique VT_3 showed better results in terms of recall for two movies, but our technique was the best for all the movies tested in respect to F_1. This fact demonstrates that our technique was able to handle scene detection in complex semantic segments better than the other techniques. Also that BoF can be used to detect semantic relationships between

[6] http://trecvid.nist.gov/

Table III: Results of the audio, video and multimodal techniques applied to the five movies in the database.

	A Beautiful Mind			**Back to the Future**			**Gone in 60 Seconds**		
	Audio	**Video**	**Multimodal**	**Audio**	**Video**	**Multimodal**	**Audio**	**Video**	**Multimodal**
P	61.54%	62.25%	58.18%	79.10%	77.94%	79.28%	52.00%	68.12%	58.42%
R	51.64%	52.13%	68.09%	44.17%	44.17%	73.33%	44.22%	69.95%	80.27%
F1	55.81%	56.65%	62.75%	56.68%	56.38%	76.19%	47.79%	56.96	67.62%

	Ice Age			**Pirates of the Caribbean**		
	Audio	**Video**	**Multimodal**	**Audio**	**Video**	**Multimodal**
P	50.00%	53.13%	48.51%	78.86%	81.89%	80.56%
R	53.97%	53.97%	77.78%	41.10%	44.04%	73.73%
F1	51.91%	53.54%	59.76%	54.04%	57.30%	76.99%

Table IV: Comparison of Multimodal Techniques

Technique	P	R	F_1
MT_1	66.62%	74.70%	70.43%
MT_2	73.22%	86.97%	79.50%
MT_3	72.40%	67.10%	69.65%

shots, showing good results. Other interesting point is that, as we mentioned earlier, using only one media present in the video, as the visual one, it is difficult to capture all the semantics present in shots - this explains the low F_1 values. Moreover, as we will demonstrate latter, the use of multimodality helps to capture more semantics, improving the technique efficiency. It is important to mention that we used vocabularies with 500 words for the audio and video techniques. This value has proven to get the best results within the range of 100 to 600 words (with steps of 100 words).

Our second experiment was the evaluation of both, the audio module alone and the whole multimodal technique. Table III presents the results of the audio, video and multimodal approaches for the movies: A Beautiful Mind, Back to the Future, Gone in 60 Seconds, Ice Age and Pirates of the Caribbean. The analysis of the results shows that, when compared with the audio technique, the visual technique presented better performance in precision and recall. Therefore, it is possible to notice that images can capture an important part of the video semantics. On the other hand, the audio technique may present a similar performance to the visual one using a considerably lower computational cost. In addition, the audio technique was able to identify scene cuts not detected by the visual technique, especially when consecutive shots were visually similar but the subject between them changed. This reinforces our point about multimodality. This fact explains the better recall results obtained by the multimodal technique, which merges the true positives found by the two techniques. However, the multimodal technique had a reduction in precision compared to the pure visual and aural techniques, since the false positives from the two techniques are incorporated into the final result.

Nevertheless, the increase in the recall obtained by the multimodal technique is higher than the precision decrease, so, the F_1 is higher than those from the individual modules in all evaluated movies. Different from related work, our technique was able to handle semantically complex segments, as the ones shown in Figures 2 and 3. This fact proves the benefits of our combination of distinct techniques, as the visual and aural ones, since the latent semantics extracted by the BoF technique is different in each technique. Moreover, they are complementary and, in this way, they can be successfully combined in order to improve efficiency. We also compared our multimodal technique (MT1) results with two other multimodal techniques: AuViFuse (MT2) [Kyperountas et al. 2004] and a statistical video scene segmentation approach (MT3) [Parshin and Chen 2006]. In this case, we have not implemented the other two multimodal techniques, and the results were obtained from their original paper. Although their techniques were not designed to handle semantically complex segments, we show that our results are comparable to MT2 and MT3, even handling a more complex problem. The results are presented in Table IV are obtained by the average of the results from every tested movie.

5. FINAL REMARKS

In this article we presented a multimodal technique designed to detect scenes in semantically complex segments. These kind of segments are relevant due to both: they are present in many genres of videos, like news, talk shows and movies; they are closer to the common users understanding of scene than to those approaches employed by related work. The proposed technique uses the Bag of Features (BoF) approach in a multimodal fashion, capturing the latent semantics in each shot and generating a dictionary with representative multimodal features. Then, the technique merges the shots into meaningful scenes, exploring different and complementary features (aural and visual) in a flexible way.

We analyzed the proposed technique in the movie domain, comparing the proposed multimodal approach with pure visual and pure aural approaches. We also made comparisons with related work in the domain. The results showed to be promising, indicating that BoF can retain the semantics of shots and, in this way, can help to detect scenes in semantically complex segments. They also showed that the aural module is able to identify scene cuts not detected by the visual one, especially when consecutive shots are visually similar but the subject between them changes. As a consequence, the technique merges the true positives found by the two modules, improving the recall. The drawback is that the precision is reduced, since the false positives from the two modules are incorporated into the final result. However, the increase in the recall obtained by the multimodal technique is higher than the precision decrease. As a consequence, the achieved F_1 value is higher than those from the individual modules, for all evaluated movies.

Therefore, we can prove that our proposed approach, using a combination of features of distinct nature, as the visual and aural ones, is appropriate to handle scene detection in semantically complex segments. This is due the fact that latent semantics extracted by the BoF technique is different in each technique. Moreover, they are complementary and, as a consequence, they can be successfully combined in order to improve the scene detection efficiency. In this article we chose to combine visual and aural media present in videos. But the technique is flexible enough to afford other kind of media, like closed caption. This should be explored in future work. Future work may also investigate: ways to reduce false positives, since they are a side effect of the adopted fusion method; and how to apply the proposed technique to other video domains.

REFERENCES

Atrey, P., Hossain, M., El Saddik, A., and Kankanhalli, M. Multimodal Fusion for Multimedia Analysis: a survey. *Multimedia Systems* 16 (6): 345–379, 2010.

Cai, Y., Tong, W., Yang, L., and Hauptmann, A. G. Constrained keypoint quantization: towards better bag-of-words model for large-scale multimedia retrieval. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. Hong Kong, China, pp. 16:1–16:8, 2012.

Chaisorn, L., Chua, T.-S., and Lee, C.-H. The Segmentation of News Video into Story Units. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. Lausanne, Switzerland, pp. 73–76, 2002.

Chasanis, V., Kalogeratos, A., and Likas, A. 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*. Santorini, Greece, pp. 35:1–35:7, 2009.

Chasanis, V., Likas, A., and Galatsanos, N. Efficient Video Shot Summarization Using an Enhanced Spectral Clustering Approach. In *Proceedings of the International Conference on Artificial Neural Networks*. Prague, Czech Republic, pp. 847–856, 2008.

Chianese, A., Moscato, V., Penta, A., and Picariello, A. Scene Detection using Visual and Audio Attention. In *Proceedings of the Ambi-Sys Workshop on Ambient Media Delivery and Interactive Television*. Quebec, Canada, pp. 4:1–4:7, 2008.

Chiang, M.-T. and Mirkin, B. Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. *Journal of Classification* 27 (1): 3—40, 2010.

Hanjalic, A., Lagendijk, R., and Biemond, J. Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems. *IEEE Transactions on Circuits and Systems for Video Technology* 9 (4): 580–588, 1999.

Klaser, A., Marszalek, M., Laptev, I., and Schmid, C. Will person detection help bag-of-features action recognition? Research Report RR-7373, INRIA. Sept., 2010.

Kumar, N., Rai, P., Pulla, C., and Jawahar, C. V. Video Scene Segmentation with a Semantic Similarity. In *Proceedings of the Indian International Conference on Artificial Intelligence*. Tumkur, India, pp. 970–981, 2011.

Kyperountas, M., Cernekova, Z., Kotropoulos, C., and Gavrielides, M. Scene Change Detection using Audiovisual Clues. In *Proceedings of the Norwegian Conference on Image Processing and Patter Recognition*. Stavanger, Norway, pp. 36–40, 2004.

Lebanon, G., Mao, Y., and Dillon, J. The Locally Weighted Bag of Words Framework for Document Representation. *The Journal of Machine Learning Research* vol. 8, pp. 2405–2441, 2007.

Liu, Y., Zhao, W.-L., Ngo, C.-W., Xu, C.-S., and Lu, H.-Q. Coherent Bag-of Audio Words Model for Efficient Large-Scale Video Copy Detection. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. Xi'an, China, pp. 89–96, 2010.

Liu, Z., Gibbon, D. C., Drucker, H., and Basso, A. Content Personalization and Adaptation for three-screen Services. In *Proceedings of the International Conference on Content-based image and Video Retrieval*. Niagara Falls, Canada, pp. 635–644, 2008.

Lowe, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60 (2): 91–110, 2004.

MacQueen, J. B. Some Methods for Classification and Analysis of MultiVariate Observations. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, USA, pp. 281–297, 1967.

Md. Rashidul Hasan, M. J. Speaker Identification using MEL Frequency Cepstral Coefficients. In *Proceedings of the International Conference on Electrical & Computer Engineer*. Bangladesh, India, pp. 565–568, 2004.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On Spectral Clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*. MIT Press, pp. 849–856, 2001.

Parshin, V. and Chen, L. Statistical Audio-Visual Data Fusion for Video Scene Segmentation. In P. Y. Zhang (Ed.), *Semantic-Based Visual Information Retrieval*. Idea Group Inc., pp. 68–88, 2006.

Rasheed, Z. and Shah, M. Scene Detection in Hollywood Movies and TV Shows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Madison, USA, pp. 343–348, 2003.

Sakarya, U. and Telatar, Z. Graph Partition Based Scene Boundary Detection. In *Proceedings of the International Symposium on Image and Signal Processing and Analysis*. Istanbul, Turkey, pp. 544–549, 2007.

Smeaton, A. F., Over, P., and Doherty, A. R. Video Shot Boundary Detection: Seven years of TRECVid activity. *Computer Vision and Image Understanding* 114 (4): 411–418, 2010.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (12): 1349–1380, 2000.

Sundaram, H. and Chang, S. F. Video Scene Segmentation using Video and Audio Features. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. New York, USA, pp. 1145–1148, 2000.

Trojahn, T. H. and Goularte, R. Video Scene Segmentation by Improved Visual Shot Coherence. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*. Salvador, Brazil, pp. 23–30, 2013.

Wang, J., Duan, L., Lu, H., Jin, J., and Xu, C. A Mid-Level Scene Change Representation Via Audiovisual Alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Toulouse, France, pp. 409–412, 2006.

Yang, J., Jiang, Y.-G., Hauptmann, A. G., and Ngo, C.-W. Evaluating Bag-of-visual-words Representations in Scene Classification. In *Proceedings of the International Workshop on Multimedia Information Retrieval*. Augsburg, Germany, pp. 197–206, 2007.

Yeung, M., Yeo, B.-L., and Liu, B. Segmentation of Video by Clustering and Graph Analysis. *Computer Vision and Image Understanding* 71 (1): 94–109, 1998.

Zen, G., Rostamzadeh, N., Staiano, J., Ricci, E., and Sebe, N. Enhanced Semantic Descriptors for Functional Scene Categorization. In *Proceedings of the International Conference on Pattern Recognition*. Tsukuba Science City, Japan, pp. 1985–1988, 2012.

Zhai, Y. and Shah, M. Video Scene Segmentation using Markov Chain Monte Carlo. *IEEE Transactions on Multimedia* 8 (4): 686–697, 2006.