A Video Similarity Measure Combining Alignment, Graphical and Speech Features

Daniel Fuentes
ETS Ing. Informatica
Univ. de Sevilla
dfuentes@us.es

R. Bardeli
Department Netmedia
Fraunhofer IAIS
rolf.bardeli@iais.fraunhofer.de

Juan A. Ortega
ETS Ing. Informatica
Univ. de Sevilla
jortega@us.es

Luis Gonzalez-Abril
EU Estudios Empresariales
Univ. de Sevilla
luisgon@us.es

Abstract

A large volume of video content on the web is available today, which demands efficient management. To effectively manage, search, retrieve and copy detection, similarity methods play a critical role. In this paper, a novel video similarity measure using visual features, alignment distances and speech transcripts is proposed. Video files are represented by a sequence of segments set where each segment contains color histograms, start time and a set of syllables extracted from the speech in the audio track. In a first step, textual, alignment and visual features are extracted. They complement each other and can be further combined to boost the segment similarity. The second step describes how the Maximum Bipartite Matching and some statistical features are applied to find segments correspondences and calculate a global similarity value respectively. Experiments for video similarity were performed on a dataset and promising results were achieved to demonstrate the effectiveness of this method.

1 Introduction

Since the popularity of social media, the amount of digital content on the World Wide Web has grown enormously in the last decade. Consequently, the volume of professional or user video is increasing exponentially and a large number of video clips are generated and added everyday. To facilitate effective search, retrieval, browsing, or copy detection, an automatic similarity measure is an essential tool. Frame-level features computations demand a large storage space and a high computational complexity in video similarity measure. In our approach, we propose to extract the color histograms values at segment level to accelerate the calculation of the image similarity.

For other hand, video similarity metrics are usually referred to measure visual content. But, nowadays in the web there are thousands of videos copies, especially movie fragments, where users only change the speech but keep equal the image. Furthermore, when in a video appears one or more speakers, speech content results another factor to include in a video similarity computation. Usually, the video sound and image are studied separately [1]. The employment of either textual or visual concepts alone may not be sufficient since either content can appear differently over time. To address this problem, in this paper the quality of correspondence between segments is jointly measured by three factors: visual similarity, alignment distortion and speech similarity, either in a weighted fashion. Hence, the main contribution of this work is the combination of different and simple metrics to measure the similarity in a efficient way. Firstly, speech, distance and image features for every two segments of two videos are extracted. After that, we apply the Maximum Bipartite Matching technique to find the most similarity segments pairs. Finally, statistical metrics are applied to calculate a concrete similarity score. In the remainder of the paper, in Section 2 we provide background for other video similarity and video copy detection approaches. Section 3 describes how we represent a video and our proposed method for efficient video similarity measure. Section 4 describes the experimental results obtained by applying our method to a video dataset. Finally, we conclude the paper with a summary of our contributions and propose ideas for future research based on these concepts.

2 Related Work

Shot and clip similarity have been extensively addressed for retrieval and clustering. Previously, clip and video copy detection (e.g. [2, 3]) were investigated by using image similarity measure with low-level global features. Global signatures are suitable for matching clips with almost identical content. Bipartite graph based algorithms were proposed in [3] to compare the similarity of two clips. Clip similarity ranking [4] was built on top of shot similarity and combines temporal order, granularity and so on. However, shot similarity detection built on global features is not robust enough for clip similarity measure due to the complicated variations of keyframes [5]. Moreover, cross-lingual video similarity measure remains a challenging problem that has seen little exploration. Signature-based methods (e.g. [2]) were proposed to identify similar clips, which use a global statistic of the low-level features. They can achieve rapid detection but its effectiveness is limited to detecting almost identical or superficially edited videos [6]. Frame-level similarity [7] is slower but capable of handling matchings of videos with a substantial degree of editing. Video copies with variations

in background, color and lighting, content modification are studied but they need a high cost of time complexity. In this respect, in our approach we represent a video like a set of segments, when a segment represents a scene where the camera stays still. This decision reduces the computation time regard to frame-level methods [8].

3 Video Similarity Detection

In this section we describe our approach to measure the similarity between two clips in detail.

3.1 Framework

A framework scheme of our similarity method is shown in Figure 1. To do the comparison, we start with two videos files, which include audio and visual contents. The first step is to divide the video into segments. Next, for each segment three features are extracted: the start time, the speech transcription and the color histograms (Figure 2). The second stage consists of the calculation of segment similarity. The start time and length values are used to measure the alignment distance between two segments. For the segment speech similarity, we decide to apply the Levenshtein distance, a simple metric for measuring the amount of difference between two video sequences. And finally, the distance between histograms is expressed by a metric based on the Bhattacharyya distance. With these features, the position, speech and image of both segments are studied. The weighted combination of these results provides a measure of the segment-level similarity. All the segment similarity values are incorporated to a Segment Similarity Matrix, which is the result of this stage. The next goal is to find the best correspondences between the segments using this matrix. For this task, we apply the Maximum Bipartite Matching technique because it directly maps the segments. After this stage, a pairs set which includes the most similar segments in two videos are computed. Now, with the selected segments, we calculate three statistical metrics to obtain the final similarity score. With the first formula, we measure the distance between the similar segments in both videos. With the second one, we compare the length of the similar area with the total length of the two videos. And with the last feature, using the Segment Similarity Matrix, the similarity level of the segment pairs with regard to whole video is compared. When these metrics values are weighted, we obtain a concrete number (between 0 to 1) which determines the similarity level of the two clips. Next, we described in detail all of these steps.

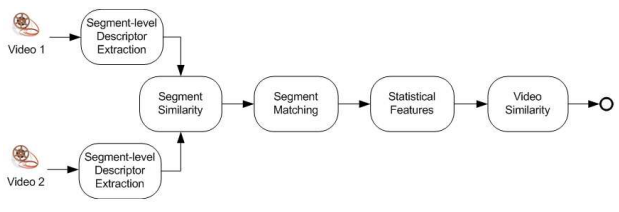

Figure 1: Video Similarity Flowchart

	Alignment	Histograms	Speech
1%	0.07%	0.23%	0.24%
5%	0.20%	0.28%	0.46%
10%	0.30%	0.40%	0.85%

Table 1: Impact of different features in the similarity value

4 Experimentation

To evaluate the effectiveness of the proposed method in detecting video similarity a dataset of 52 videos from Beijing Olympic Games was used. Mainly, this dataset consist of videos about different olympic sport competitions but also are referred to interviews and Chinese culture. Many of them contain German speech from an narrator, interviewer or interviewee and the extracted speech transcriptions are used for the similarity computation. For the experiments, 100 compound videos were formed using the original videos. A compound video is considered the joint from one to four videos (chosen randomly) in only one. The experiments consist on applying the similarity method with every two original and compound videos for different similarity thresholds. Some of the results are shown in Table 1.

5 Conclusion

In this paper, we have presented a novel method to measure the video similarity by analyzing properties of segments in two clips. Multiple features are extracted to model the appearance of the segments, including color and speech descriptors. Utilizing the proposed video similarity framework, we have achieved very promising and competitive performance in video similarity for video comparison. Firstly, our method focuses on the value of spatiotemporal, histograms and speech features to match the segments in pairs. We feel that the particular combination of these descriptors can be crucial for different comparisons. Secondly, we have applied different statistical metrics to compare the similar segments to whole videos to obtain a final similarity score. By using this method we have demonstrated the particular application identifying highly similar video sequences in a large set of web videos.

Acknowledgments

This research is supported by the Spanish Ministry of Science and Innovation I+D project ARTEMISA (TIN2009-14378-C02-01).

References

[1] J. Foote. An overview of audio information retrieval. In *ACM Multimedia Systems*, vol. 7, pp. 2-10, 1999.

[2] S. C. Cheung and A. Zakhor. Efficient Video Similarity Measurement with Video Signature . In *IEEE TCSVT*, 2003.

[3] Y. Zhang, J. Callan and T. Minka. Clip-based Similarity Measure for Query-Dependent Clip Retrieval and Video Summarization. In *IEEE Trans. on CSVT*, vol. 16, no. 5,pp. 612- 627, 2006.

[4] A. K. Jain, A. Vailaya and W. Xiong. Query by Video Clip. In *ACM Multimedia Syst*, vol. 7, pp. 369-384, 1999.

[5] D-Q. Zhang and S-F. Chang. Detecting Image Near-Duplicate by Stochastic Attributed Relational Graph Matching with Learning. In *ACM MM'04*, USA, Oct. 2004.

[6] A. Hampapur and R. Bolle. Comparison of sequence matching techniques for video copy detection. In *Conf. on Storage and Retrieval for Media Databases* , 2002.

[7] W.-L. Zhao, C.-W. Ngo, H.-K. Tan and X. Wu. Near-duplicate keyframe identification with interest point matching and pattern learning. In *IEEE Trans. on Multimedia* , vol. 9, no. 5,pp. 1037-1048, 2007.

[8] C-W. Ngo, W-L. Zhao and Y-G. Jiang. Fast Tracking of Near-Duplicate Keyframes in Broadcast Domain with Transitivity Propagation. In *ACM MM'06* , USA, Oct. 2006.