A Browsing and Retrieval System for Broadcast Videos using Scene Detection and Automatic Annotation

Lorenzo Baraldi[1], Costantino Grana[1], Alberto Messina[2] and Rita Cucchiara[1]

[1]Università degli Studi di Modena e Reggio Emilia
name.surname@unimore.it

[2]RAI - Radiotelevisione Italiana - Centre for Research and Technological Innovation
alberto.messina@rai.it

ABSTRACT

This paper presents a novel video access and retrieval system for edited videos[1]. The key element of the proposal is that videos are automatically decomposed into semantically coherent parts (called *scenes*) to provide a more manageable unit for browsing, tagging and searching. The system features an automatic annotation pipeline, with which videos are tagged by exploiting both the transcript and the video itself. Scenes can also be retrieved with textual queries; the best thumbnail for a query is selected according to both semantics and aesthetics criteria.

Keywords

Video Browsing; Temporal Video Segmentation; Retrieval

1. INTRODUCTION

Video sharing and browsing platforms like Youtube allow to upload, search and access user-generated videos. Retrieval is usually performed using tags and descriptions given by the content provider, and thumbnails are static, meaning that the same thumbnail is shown when a video is searched with different queries. This paradigm fits well for short user-generated videos, in which the content is quite uniform and it is easy to get a glimpse of it. The same does not apply to broadcast videos, which are usually longer and more complex. Beside the length, those have indeed more complex story lines and may contain different topics.

The system we present in this paper addresses these limitations with three novelties: first, it lets the user browse and retrieve videos with a finer granularity level, by decomposing them into sets of *scenes*, which are parts of video with uniform semantic content. Secondly, it does not require manually generated tags or descriptions, as it extract annotations using visual content and transcripts. Lastly, given a textual query it retrieves the most significant scenes by presenting them with the most appropriate thumbnail, according to both semantics and aesthetics.

[1]Demo is available at http://imagelab.ing.unimore.it/neuralstory

MM '16 October 15-19, 2016, Amsterdam, Netherlands

ACM ISBN 978-1-4503-3603-1/16/10.

DOI: http://dx.doi.org/10.1145/2964284.2973825

2. THE SYSTEM

Our system comprises three main components: a scene detection module, a concept detection algorithm, which visually detects the presence of concepts expressed in the transcript, and a retrieval algorithm with which users can search for scenes inside a video collection.

Scene detection To segment a video into a set of scenes, we apply the algorithm described in [1]. Given a ground-truth temporal segmentation of a set of videos, it runs a shot detector and trains a Siamese Neural network to predict whether two shots should belong to the same scene. Each branch of the Siamese network is composed by a Convolutional Neural Network (CNN), whose penultimate layer is concatenated with features extracted from the transcript of the video. Training is done using a contrastive loss function, which computes the distance between two input shots. At test time, distances between shots given by the Siamese network are arranged into a similarity matrix, which is then used with Spectral Clustering to obtain the final scene limits.

Automatic concept detection Sentences in the transcript of a video are parsed with the Stanford CoreNLP tagger [3]. Each unigram tagged as noun, proper noun or foreign word is converted to its lemmatized form, and then matched with the most semantically similar class in the Imagenet database [4], which contains more than 40.000 categories. This is achieved with a Word2Vec model [5], which assigns each word to a vector in a semantic space.

Having mapped each unigram to an Imagenet class, we build a classifier to detect the presence of a visual concept in a shot. Images from Imagenet are represented using feature activations from pre-trained CNNs; a linear probabilistic SVM is then trained for each concept, using randomly sampled negative images from other classes. The probability output of each classifier is used as an indicator of the presence of a concept in a shot. Given the temporal coherency of a video, it is unlikely for a visual concept to appear in a shot which is far from the point in which the concept found in the transcript. For this reason, we run a classifier only on shots which are temporally near to its corresponding term, and apply a Gaussian weight to the probability of each term, based on the temporal distance.

Retrieval Given a textual query, we build a retrieval

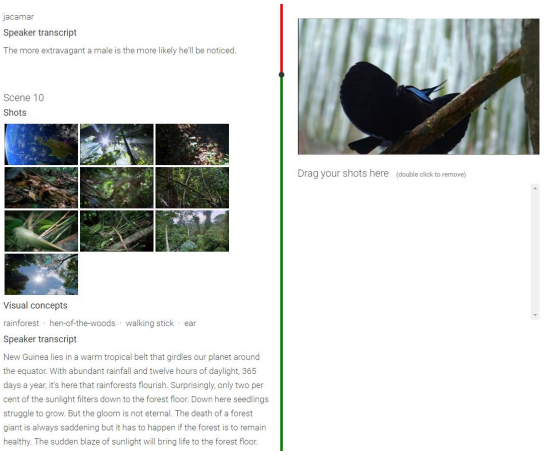

Figure 1: Browsing interface.

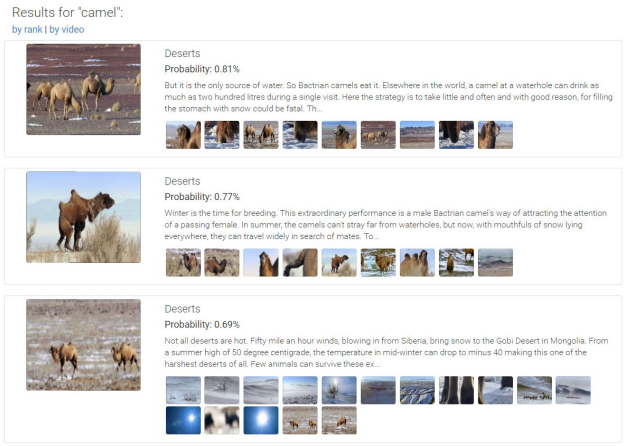

Figure 2: Retrieval interface (for query *camel*).

strategy which exploits automatically detected concepts. Details can be found in [2].

We first match the query with the most similar detected concept, using the Word2Vec embedding. If the query is composed by more than one word, the mean of the embedded vectors is used. Given a query q, each scene inside the video collection is then assigned a score according to the following function:

$$R_{scene}(q) = \max_{s \in scene} \left(\alpha P(s,u) + (1-\alpha) \max_{d \in s} E(d) \right) \quad (1)$$

where s is a shot inside the given scene, u is the most similar concept found in transcripts to query q, according to the Word2Vec embedding, and d represent all keyframes extracted from a given shot. $P(s,u)$ is the probability that shot s contains concept u, computed as described above. $E(d)$ is the aesthetic ranking of key-frame d, computed according to [2]. Parameter α tunes the relative importance of semantic representativeness and aesthetic beauty. The final retrieval results is a collection of scenes, ranked according to $R_{scene}(q)$, each one represented with the keyframe that maximizes the second term of the score.

It is worth to notice that most of the proposed pipeline is executed off-line. Scenes boundaries are detected only once when a video is uploaded, while probabilities $P(s,u)$ are computed offline for each unigram found in the transcript of a video and stored in an inverted index. $E(d)$, as well, are computed in advance for each key-frame, thus greatly reducing the computational needs in the on-line stage. The off-line stage requires on average the duration of the video.

3. INTERFACE

Given automatically extracted scenes, a video is presented in a time-line fashion (see Fig. 1), where users can get an insight of the content of the video by scrolling the timeline. For each scene, a carousel of shots is also presented, along with the speaker transcript and automatically detected concepts. On the right-hand side of the timeline, a video player is put, and users can navigate from one point to another of the video by clicking on shot thumbnails. A drag-and-drop zone is finally placed under the video player, in which users can drop shots or entire scenes, to build new videos from watched ones.

Figure 2 depicts the retrieval interface. Given the query, a ranked list of scenes is reported: for each scene, the title of the corresponding video and the transcript of the scene is reported, along with a confidence score. The full set of shots in the scene is also shown, as well as the thumbnail which maximized the conceptual and aesthetic criterion.

The system is presented with a collection of videos from the BBC educational series *Planet Earth* as well as a portion of the first season of the TV-Series *Ally McBeal*.

4. CONCLUSIONS

We described a video browsing and retrieval system specifically designed for collections of edited videos. Videos are presented as sequences of scenes to enhance browsing, and a retrieval strategy based on automatically detected concepts is used to let the user search inside the collection.

Acknowledgments This work has been partially founded by the project "Città educante" (CTN01_00034_393801) of the National Technological Cluster on Smart Communities (cofunded by the Italian Ministry of Education, University and Research - MIUR).

5. REFERENCES

[1] L. Baraldi, C. Grana, and R. Cucchiara. A deep siamese network for scene detection in broadcast videos. In *ACM MM*, pages 1199–1202. ACM, 2015.

[2] L. Baraldi, C. Grana, and R. Cucchiara. Scene-driven retrieval in edited videos using aesthetic and semantic deep features. In *ACM ICMR*, 2016.

[3] M.-C. De Marneffe, B. MacCartney, C. D. Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *ANIPS*, pages 3111–3119, 2013.