UPC at MediaEval 2013 Hyperlinking Task

Carles Ventura
Universitat Politecnica de Catalunya
Barcelona, Catalonia
carles.ventura@upc.edu

Marcel Tella-Amo
Universitat Politecnica de Catalunya
Barcelona, Catalonia
marceltella@gmail.com

Xavier Giro-i-Nieto
Universitat Politecnica de Catalunya
Barcelona, Catalonia
xavier.giro@upc.edu

ABSTRACT

These working notes paper present the contribution of the UPC team to the Hyperlinking sub-task of the Search and Hyperlinking Task in MediaEval 2013. Our contribution explores the potential of a solution based only on visual cues. In particular, every automatically generated shot is represented by a keyframe. The linking between video segments is based on the visual similarity of the keyframes they contain. Visual similarity is assessed with the intersection of bag of features histograms generated with the SURF descriptor.

1. INTRODUCTION

These working notes describe the algorithms tested by the UPC team in the MediaEval 2013 Search and Hyperlinking Task. The reader is referred to the task description [4] for further details about the study case, dataset and metrics. Our team participated only in the Hyperlinking sub-task, discarding the Search one.

The Hyperlinking sub-task aims at linking anchors related to a temporal segment of a video. In this sub-task, as well as the Search one, one of the main challenges is the uncertainty regarding what criteria are to be followed to generate these links. There is ambiguity about what the user expectations are regarding these links, as well as little information about what is considered relevant to the user in the video segment.

In addition, the search within the video collection can consider three different types of media information: visual, audio and textual. Previous editions of the Hyperlinking subtask have especially focused on the textual data, whether in the form of closed captions or audio transcripts. The visual modality, though, has received less attention despite its great potential. In our work, we study the results of the visual cues in trying to capture the user expectations about the links. In particular, we adopt the popular Bag of Features model to generate links to video segments which are visually similar to the provided anchor.

2. RELATED WORK

The generation of links between video segments can be solved with techniques developed for image retrieval. This field was greatly influenced by the Video Google [5] work, which adapted the basic principles of text retrieval to image collections. Video Google considered visual features around sparsely generated interest points as the basic units that describe an image. There are several existing solutions on how to detect and characterise the sparse interest points. In our work, SURF descriptors [2] were adopted.

Once each image is represented by the interest points it contains, Video Google proposes the construction of a visual vocabulary of features (words) by clustering a large amount of them, extracted from a training dataset. After this learning stage, every feature point of the test dataset can be quantised by assigning it to the most similar word within a vocabulary. As a result, every image in the test dataset can be described with a signature, where each bin corresponds to the word occurrences. The representation of each image as a *Bag of Features (BoF)* allows its efficient indexing with inverted indexes. In addition, the resulting signatures tend to be sparse, allowing efficient storage and similarity assessment.

Currently there are several implementations of the algorithms available. We have adopted the solutions included in OpenCV [3], a popular and free software which allows the replication of the experiments.

3. APPROACH

The Hyperlinking sub-task is addressed to temporal video segments, but dealing with all the visual information contained in the collection requires a very large amount of computing power not available to the participants. Nevertheless, the MediaEval organisers published the shot boundaries and associated keyframes generated by [1]. Dealing with the these keyframes, even with a large number of 1,200,000, is nevertheless a reasonable effort, especially if working with sparse BoF. This decision however requires a strategy to firstly convert the provided anchors into query keyframes, and later estimate the linked video segments from the retrieved keyframes.

3.1 From anchors to query keyframes

The provided anchors were defined by a start and end time codes. The temporal boundaries of an anchor typically covered several shots. As every shot was also represented with a keyframe, every anchor corresponded to a collection of keyframes. Each of these keyframes was used to formulate an independent image query, which was compared with all the keyframes of the video collection to generate a ranked list of retrieved keyframes. The similarity assessment was performed with the histogram intersection, an operation which can be efficiently assessed on sparse signatures.

The sub-task organisers required that, in addition to strictly

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

adjusting to the anchor boundaries, one of the runs should consider the context of the general segment. Following the spirit of our exclusively visual solution, we simply expanded the time codes of each anchor to 5 more shots backward and forward. That is, in the context run, 10 more shots were considered as part of the query frames.

3.2 From retrieved keyframes to linked video segments

The retrieved keyframes associated with an anchor require a further processing to generate a ranked list of temporal video segments. Our approach is based on the assumption that the user appreciates diversity within the resulting links. In particular, the implemented solution limits to one the appearance of a video in the retrieved list of video segments.

The first step to create video segments is the merging of all ranked lists of keyframes associated to an anchor into a unified one. If a keyframe appears in more than one ranked list, only the highest score is considered to represent this keyframe.

The algorithm scans in increasing order the ranked list of keyframes and generates a new and sorted entry in the output list whenever a new video is referred to. Taking as a base time code the one of the first appearance of the video in the ranked list, the rest of the unified ranked list is explored searching for more keyframes from the considered video. Whenever a new keyframe is found, it is considered to set a new start or end time code to the entry under process. If expanding the time span does not exceed the limit of 2 minutes set by the organisers, the new time code is accepted. Otherwise, this keyframe is discarded and the rest of the list is explored. When the end of the keyframes ranked is reached, the end time code is expanded so that the final duration of the linked video segments is also 2 minutes. This strategy was adopted because it was assumed that the user would especially appreciate a precise start time, but that it would not be very selective about the end time. On the other hand, expanding the segment may, even by chance, provide the user with some valuable information that they would appreciate.

It must be noticed that our results did not exclude the video source of the anchor from the retrieved video segments. We considered that the results may provide a different segment from the anchor one, although probably containing it.

4. EXPERIMENTS AND RESULTS

The UPC participated in the Hyperlinking sub-task, providing 1,000 video segments for each of the 98 anchors provided. Finally, only 30 of them were assessed by users. The Bag of Features model relied on a vocabulary of 10,000 codewords extracted with the SURF descriptor [2]. This vocabulary is the result of applying the k-means algorithm over the set of SURF descriptors extracted from 2,324 keyframes, one from each dataset video. The results returned by the subtask organisers are shown in Table 1, where the first column corresponds to defining the anchors strictly with the provided time codes, and the second column corresponds to the expansion to 5 shots backward and 5 shots forward.

Results indicate that using an expanded temporal segment as a context is not advisable, probably because the newly considered frames are not relevant for the user. The decreasing precision values point that the upper part of the ranked list contains a higher proportion of relevant targets, which decreases when considering a larger amount of hits.

	No context	**Context**
MAP	0.0282	0.0260
Precision @ 5	0.2600	0.2400
Precision @ 10	0.2000	0.1967
Precision @ 20	0.1233	0.1217

Table 1: UPC results in the Hyperlinking sub-task.

5. CONCLUSIONS

The presented technique has explored the potential of solving for the Hyperlinking sub-task based exclusively on visual similarity. Our proposal has considered videos as collections of shots, each of them represented by a single keyframe. A classic image retrieval solution has been adapted to work with temporal segments, based on the popular Bag of Features approach. In our work we have used a given shot segmentation to map video segments into keyframes, and viceversa.

We consider the obtained results as promising, however, we understand this first participation of our team in the Hyperlinking sub-task as exploratory. The experience has shown has the challenges of dealing with large amounts of visual data, which requires high computational power and an indexing strategy. The results obtained might be improved by considering a larger amount of keyframes to build the vocabulary, as well as allowing multiple targets from the same video.

6. ACKNOWLEDGMENTS

This work has been partially funded by the Spanish project TEC2010-18094 MuViPro.

7. REFERENCES

[1] A. Massoudi, F. Lefebvre, C.H. Demarty, L. Oisel, and B. Chupeau. A video fingerprint based on visual digest and local fingerprints. In *ICIP*, Atlanta, Georgia, USA, October 8-11 2006.

[2] H. Bay, T. Tuytelaars, and L. Gool. Surf: Speeded up robust features. In *ECCV*, Graz, Austria, May 7-13 2006.

[3] G. Bradski and A. Kaehler. *Learning OpencCV: Computer Vision with OpenCV Library*. O'Reilly, 2008.

[4] M. Eskevich, G. J. Jones, S. Chen, R. Aly, and R. Ordelman. The Search and Hyperlinking Task at MediaEval 2013. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[5] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *9th IEEE Intl' Conf. Computer Vision, 2003*, pages 1470–1477 vol.2, 2003.