Multimodal Video Copy Detection using local features

Xavier Anguera, Tomasz Adamek***
**Telefónica Research, Torre Telefonica-Diagonal 00, 08019 Barcelona, Spain*
***Catchoom, Llacuna 162-168, Barcelona Activa, 08018 Barcelona*

xanguera@tid.es

1. Introduction

Content-based video copy detection (CBCD) systems aim at finding video segments that are identical or transformed versions of segments in a known video. Joly et al. [8] propose a definition of video copy based on a subjective notion of tolerated transformations. A tolerated transformation is a function that creates a new version of a document where the original document "remains recognizable". CBCD systems perform the detection by processing visual and/or audio content of videos, ignoring any metadata and avoiding the embedding of watermarks into the original videos.

The detection of video duplicates in a video database is one of the key technologies in multimedia management. Its main applications include, among others: (a) storage optimization; (b) copyright enforcement; (c) improved web search and (d) concept tracking. In storage optimization, the goal is to eliminate exact duplicate videos from a database and to replace them with links to a single copy or, alternatively, link together similar videos (near duplicate) for fast retrieval and enhanced navigation. Copyright enforcement strives at avoiding the use and sharing of illegal copies of a copyright protected video. In the context of web search, the goal is to increase novelty in the video search result list, by eliminating copies of the same video that may clutter the results and hinder users from finding the desired content [9]. Finally, concept tracking in video feeds [10] focuses on finding the relationship between segments in several video feeds, in order to understand the temporal evolution of, for example, news stories.

Traditionally, CBCD systems have relied only on the visual information in the videos, driving research towards new features that allowed for scalable systems and fast retrieval. In recent years systems started incorporating features derived from the audio modality. Through the fusion of multimodal information, CBCD systems can obtain improved performance levels and be more robust to errors (e.g. when one of the modalities is severely damaged or missing).

In this letter we briefly describe the CBCD system we have developed over the last 3 years at Telefónica Research, which has achieved outstanding results in the 2010 and 2011 NIST-Trecvid Video Copy Detection evaluations. The main characteristics of our system are the use of multimodal (i.e. audio and video) information to detect plausible video copies in each modality and an effective late fusion of results.
The Individual modalities we use are both based on local features (DART [4] for video and MASK [6] for audio) which are able to robustly encode the multimodal content, thus allowing for successful retrieval of video copies.
The late fusion algorithm [7] takes into account the scores and ranking of each result in each modality to combine them most effectively.
The currently system has been developed into a pseudo-commercial application by using a scalable database architecture and feature extraction parameters optimized for speed.

2. Multimodal Video Copy Detection System

Figure 1 shows the main blocks that conform Telefónica Research CBCD system. The system is composed of two parallel feature streams (one for audio and one for video) that are processed in parallel, obtaining each one an individual set of results (for NIST-Trecvid experiments we retrieve 20 results from each modality). Then a late fusion is used to merge these results into a multimodal output. The choice of a late fusion versus an earlier one was done so that the system gains in flexibility to be applied for different applications like music IR, where one of the modalities is non existent or relevant.
Next we describe each of the steps in the system.

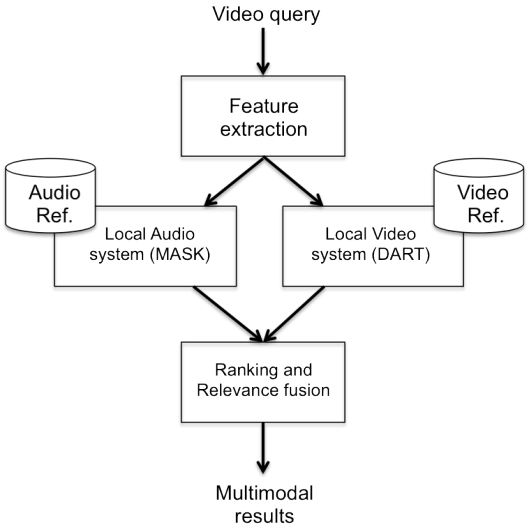

Figure 1: Multimodal Video Copy Detection system

2.1 Local Video System

The local visual features processing module [1][5] compares all query keyframes with all reference keyframes using a state-of-the-art image retrieval engine relying on local features [4] and then combines the obtained ranked lists of matched keyframes into copied video segments by performing a temporal consistency post-processing. The module is divided in three tasks: Feature Extraction (which samples every video with one frame per second and extracts novel local features called DART [4]), Keyframe Matching (by using hierarchical dictionaries of visual words and inverted files to locate matching frames, which are then refined through a spatial verification stage), and Temporal Consistency (which computes the time differences between matches, and returns a list copy candidates each one with a location within the video, and a score).

2.2 Local Audio System

The audio local features module [5] is based on the recently proposed MASK features [6]. These are local features computed over the spectral domain of the audio signal by encoding in binary form the energy differences of predefined regions around spectral maxima. The MASK features therefore consist on an asynchronous stream of fingerprints that later are matched, using a similar technique to the one used in the video system, with the reference features. Instead of a spatial consistency step, in here we perform an alignment step between all MASK points in the reference and query matching segments to obtain an accurate matching score.

2.3 Ranking and relevance fusion

Each of the modalities in the presented system results in a list of 20-best possible matches between segments in the provided query video and the reference database. The fusion module is in charge of merging these monomodal decisions into a multimodal output as described in [7]. To do so, it makes use of both the ranking of each matching segment within its modality and the normalized score. As each modality might have different score distributions and sometimes not even output all 20 matches, an L1 normalization and flooring preprocessing is performed before the results are merged.

In [5] we show the suitability of this fusion algorithm for any number of monomodal inputs. We show we are able to lower the best score obtained in the NIST-Trecvid evaluation by a blind combination of several of the submitted system outputs without.

2.4 Scalable implementation

Ultimately, a CBCD system is useful if it can be scaled to index large multimodal databases and it is able to retrieve matching segments for a given query in little time. In order to achieve the first goal we use a disk-based inverted file index over SSD drives that structures the information at indexing time in a way that it is faster to retrieve similar reference matches at retrieval time. Also, given the locality of both fingerprints we use, we are able to adapt the amount of information we store per second of content depending on the application and the relation between accuracy and speed we want to achieve.

Currently our system is flexible to accommodate various application use-cases as neither we impose constraints on the length of the query or reference, nor we predefine where the matching start-end points might be found.

References

[1] E. Younessian, X. Anguera, T. Adamek, N. Oliver, and D. Marimon, “Telefonica research at trecvid 2010 content-based copy detection,” in Proc. NIST-TRECVID Workshop, 2010.

[2] T. Adamek and D. Marimon, “Large-scale visual search based on voting in reduced pose space with application to mobile search and video collections,” in Multimedia and Expo (ICME), 2011 IEEE International Conference on,

july 2011, pp. 1 –4.

[3] X. Anguera, J. M. Barrios, T. Adamek, and N. Oliver, "Multimodal fusion for video copy detection," in Proc. ACM Multimedia, 2011.

[4] D. Marimon, A. Bonnin, T. Adamek and R. Gimeno, "DARTs: Efficient scale-space extraction of daisy keypoints," in Proc. Computer Vision and Pattern Recognition (CVPR), June 2009.

[5] X. Anguera, T. Adamek, D. Xu and J.-M. Barrios, "Telefonica Research at TRECVID 2011 Content-Based Copy Detection", in Proc. NIST-TRECVID Workshop, 2011

[6] X. Anguera, A Garzon and T. Adamek, "MASK: Robust Local Features for Audio Fingerprinting", in Proc. ICME 2012, Melbourne, Australia

[7] X. Anguera and J-M. Barrios, T. Adamek and N. Oliver, "Multimodal Fusion for Video Copy Detection", in Proc. ACM Multimedia 2011.

[8] A. Joly, O. Buisson, and C. Fr´elicot. Content-based copy retrieval using distortion-based probabilistic similarity search. IEEE Trans. on Multimedia, 9(2):293–306, 2007.

[9] X. Wu and A. G. Hauptmann and Ch.-W. Ngo, "Novelty Detection for Cross-Lingual News Stories with Visual Duplicates and Speech Transcripts", in Proc. ACM Multimedia 2007

[10] Y. Zhai and M. Shah, "Tracking News Stories Across Different Sources", In Proc. ACM Multimedia 2005.

Xavier Anguera Ing. [MS] 2001 by UPC (Barcelona, Spain), [MS] 2001 European Masters in Language and Speech, Dr. [PhD] 2006 UPC University. From 2001 to 2003 he worked for Panasonic Speech Technology Lab in Santa Barbara, CA. From 2004 to 2006 he was a visiting researcher at the International Computer Science Institute (ICSI) in Berkeley, CA. Since 2007 he is with Telefonica Research in Barcelona, pursuing research on multimedia analysis. His main research interests involve speech/speaker processing and multimedia processing.

Tomasz Adamek is the cofounder and CTO of Catchoom, a startup that provides visual recognition technology designed for real live applications. He received his Ph.D. degree from Dublin City University (DCU), Ireland, in 2006. Between 2006 and 2007 he worked as a postdoctoral researcher at CDVP, DCU. In 2008 he joined Telefónica R&D, Barcelona, Spain

were he was responsible for technological aspects of several research projects in the area of large-scale multimedia indexing and retrieval. His work led to several patents and over 20 academic papers in high impact journals and conferences.