A System for Video Recommendation using Visual Saliency, Crowdsourced and Automatic Annotations

Andrea Ferracani, Daniele Pezzatini, Marco Bertini
Saverio Meucci, Alberto Del Bimbo
Università degli Studi di Firenze - MICC
[name.surname]@unifi.it

ABSTRACT

In this paper we present a system for content-based video recommendation that exploits visual saliency to better represent video features and content[1]. Visual saliency is used to select relevant frames to be presented in a web-based interface to tag and annotate video frames in a social network; it is also employed to summarize video content to create a more effective video representation used in the recommender system. The system exploits automatic annotations from CNN-based classifiers on salient frames and user generated annotations. We evaluate several baseline approaches and show how the proposed method improves over them.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Design, Experimentation

Keywords

Social video tagging, automatic video tagging, item-based video recommendation, visual saliency

1. INTRODUCTION

A typical item-based video recommender builds its prediction model considering user preferences for videos, expressed as ratings, and suggests potentially interesting videos comparing distributions of such ratings. The proposed system adopts an hybrid approach in which a brief but comprehensive representation of video content, derived from video analysis and concepts extraction from user activity can improve the performance of a standard recommender based on collaborative filtering (i.e. using ratings).

Video recommendation can drive users to watch other videos, much more than direct searching for new videos, as shown in [5]. The approaches presented in the scientific literature are typically based on textual analysis of the metadata that accompany a video, that is sometimes complemented by some multimedia content analysis. The YouTube recommendation system, described in [1], uses two broad classes of data: *i)* content data, such as the raw video streams and video metadata such as title, description etc., and *ii)* user activity data, either explicit (e.g. video rating/liking) and implicit (watching a video for a long time).

[1]Demo video available at http://bit.ly/1FYloeQ

MM'15, October 26–30, 2015, Brisbane, Australia.
ACM 978-1-4503-3459-4/15/10.
DOI: http://dx.doi.org/10.1145/2733373.2807982.

2. THE SYSTEM

The architecture of our system has been implemented in a beta version of a social network that exploits user profiling techniques to propose to the user targeted recommendations of videos, topic of interests and similar users in the network. This is achieved initially by analysing data from other online profiles (i.e. Facebook) and then tracking user's activities on the social network, like number of video views, click-through data, video annotation and rating. Users can share and annotate videos at frame level using concepts derived from Wikipedia. All these concepts are clustered in 54 categories using Fuzzy K-Means in a two-levels taxonomy of interests (12 macro and 42 micro-categories such as Music and Jazz music, inspired by the taxonomy of Vimeo) and classified using a semantic distance [2] with a nearest neighbour approach. All the resources categorised in videos are then used to build a vector describing video content exploited in the recommender.

Video analysis is performed to improve the interaction of users with the system, and to obtain content-based representation of videos, in order to compute the recommendation. Automatic video annotations are extracted using a classifier which exploits convolutional neural networks (CNN) on most salient frames. Automatic annotations are categorised using the semantic relatedness measure weighted according to the confidence returned by the classifier. Fig. 1 shows an overview of the workflow used to collect and annotate videos.

Visual saliency. Visual saliency is used *i)* at the interface level to propose to the users possible frames of interest through a carousel above the video player, to ease the addition of comments and annotations; *ii)* at the automatic annotation level to reduce the computational cost of processing all the frames. Videos are preprocessed to eliminate letterboxing, then visual saliency maps, computed with the iLab Neuromorphic Toolkit [3], are extracted for all the frames of a video. The salient frames to be used in the system interface are selected by identifying the peaks of salience of

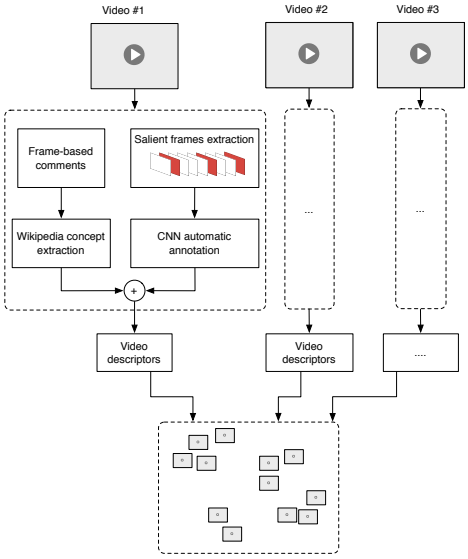

Figure 1: System overview: manual annotation, automatic video content analysis and representation.

the video using the crest detection algorithm proposed in [4]. The frames used in automatic annotation are selected computing the average saliency of the video and choosing those above the average, to have a denser sampling of video content.

Crowdsourced annotations. Users can comment videos at frame level and can add semantic references to Wikipedia entities in comments using an autosuggest widget. A carousel of the most salient frames is also shown above the video player as a video summary (see Fig. 2). Users can click on an image in the carousel and jump directly to the correspondent frame in the video at the exact timecode, to facilitate fast and accurate annotations.

A vector of categories C, with dimension equal to the number of categories of the taxonomy, is used to represent video content according to the comments of the users. Each category is assigned with a weight which indicates the affinity between category and video contents. C is defined by calculating for each category the average of the semantic distance of each annotation to the categories of the taxonomy. This semantic relatedness between the terms is obtained using the Web Link Based Measure [2].

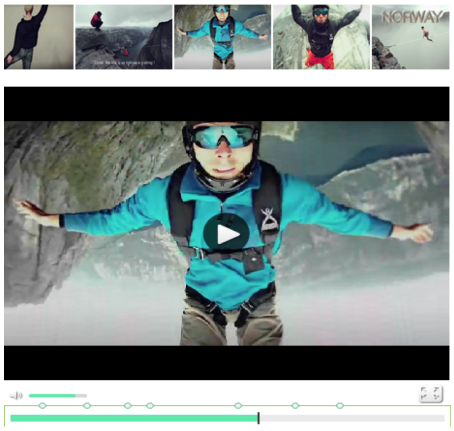

Figure 2: Frame selection from the salient frames carousel.

Visual features. In the proposed system video frames are subsampled according to their visual saliency, considering that visual saliency allows to make a targeted selection of these frames, allowing the system to scale while maintaining a reasonably dense sampling of video content. The convolutional network implemented uses the LibCCV[2] library, and it is trained on the ImageNet ILSVRC 2014 dataset to detect 1000 synsets. Video content is represented using a Bag-of-words approach, applied to the 1,000 synsets, selecting for each video the probabilities that obtained a score above a predefined threshold.

Recommender. The recommender implements an item-based collaborative filtering that builds an item-item matrix determining similarity relationships between pairs of items. Then the recommendation step uses the most similar items to a user's already-rated items to generate a list of recommendations. Videos are represented using a feature vector that concatenates the histogram of the categories of the manual comments and the BoW description obtained using the CNN classifier on most salient frames. User's general rating on a video is built combining explicit and implicit activity on the content itself. Users can explicitly vote a video on a 5 point scale through a visual widget. At the same time, implicit rating is computed taking into account number of visualizations, frame browsing and comments added to every video.

A dataset has been collected by hiring 812 workers from the Microworkers web site, and asking them to use the system to upload their favorite videos, annotate and comment some shots that were more interesting to them and to provide ratings for some videos. The dataset is composed by 632 videos, of which 468 were annotated with 1956 comments and 1802 annotations. 613 videos were rated by 950 of 1059 total network users. Comparing the performance of the system, in terms of Root Mean Square Error (RMSE), with a standard item-based recommender implemented in Apache Mahout shows an improvement of $\sim 26\%$.

3. CONCLUSIONS

In this paper we have presented a system that performs item-based video recommendation using a content based description of videos obtained from crowdsourced and automatic annotations. Visual saliency is exploited to present the most relevant frames to the users and to reduce the number of frames to be processed. Experiments show that the proposed method improves over the standard implementation of an item-based algorithm.

References

[1] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath. The YouTube video recommendation system. In *Proc. of ACM RecSys*, 2010.

[2] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *Proc. of ACM CIKM*, 2008.

[3] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision Research*, 45(2), 2005.

[4] Z. Wang, J. Yu, Y. He, and T. Guan. Affection arousal based highlight extraction for soccer video. *Multimedia Tools and Applications*, 73(1), 2014.

[5] R. Zhou, S. Khemmarat, and L. Gao. The impact of YouTube recommendation system on video views. In *Proc. of ACM SIGCOMM IMC*, 2010.

[2]http://libccv.org/