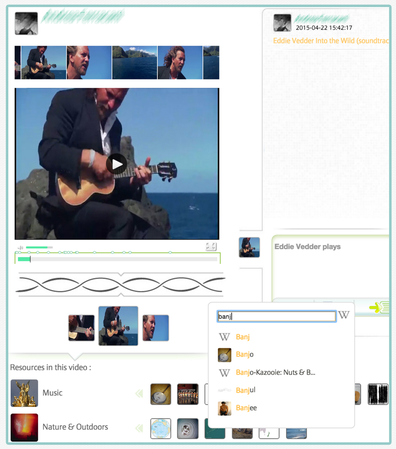
Item-Based Video Recommendation: an Hybrid Approach considering Human Factors

Andrea Ferracani, Daniele Pezzatini, Marco Bertini, Alberto Del Bimbo
MICC - University of Florence, Italy

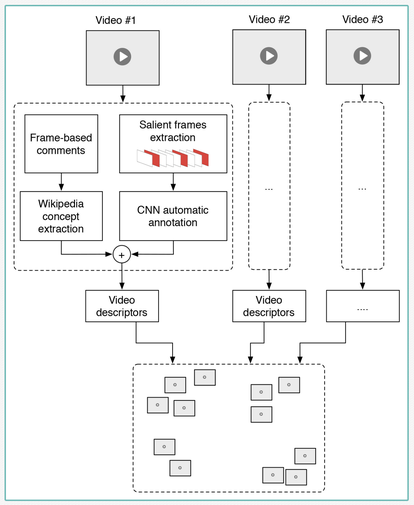
The Project

In this paper we propose a method for **video recommendation in Social Networks** based on crowdsourced and automatic video annotations of salient frames.

We show how two human factors, **users' self-expression** in user profiles and perception of **visual saliency** in videos, can be exploited in order to stimulate annotations and to obtain an efficient representation of video content features. Results are assessed through experiments conducted on a prototype of social network for video sharing. Several baseline approaches are evaluated and we show how the proposed method improves over them.



The system



User profile interface: our prototype system provides users with a public profile that can be curated in a semiautomatic way. The profile shows user's last comments and annotations as well as annotated video frames and tagged Wikipedia resources with thumbnails. A **profiling algorithm** categorizes annotations and automatically proposes inferred user interests. Each user can present himself with a set of categories that are visually shown on his profile. Resources annotated by SN users, automatically categorised, are suggested as items that users can drag and promote in their public profile for each detected user interest.

Visual saliency: we propose the use of visual saliency in SN systems and interfaces at two levels: i) at the automatic annotation level to reduce the computational cost of processing all the frames; ii) at the interface level to propose to the users possible frames of interest. The SN prototype also features a salient frames carousel above each video to ease the addition of crowdsourced comments. **Saliency maps** are defined by a visual attention model which uses a dynamic neural network on multiscale image features computed with the iLab Neuromorphic Toolkit.

Visual features: automatic annotation of all the frames of the videos in a SN is a time-consuming task which requires a lot of resources. In the proposed SN video frames are subsampled according to their visual saliency, allowing the system to scale while maintaining a reasonably dense sampling of video content. The convolutional network used was trained on the **ImageNet ILSVRC 2014 dataset** to detect 1000 synsets. A **very deep CNN with 16 layers** was used to extract the final output layer for each frame, containing 1,000 object probabilities. Video content is represented using a Bag-of-Words (BoW) approach. The features vector is computed using the frequency of occurrence of detected concepts with a probability above a threshold, then also complemented by crowdsourced annotations.

The recommender: the proposed hybrid RS adopts a solution that combines **a semantic pre-filtering of content with an item-based algorithm**. Videos are represented using a feature vector that concatenates the histogram of the categories of the crowdsourced comments and the BoW description obtained using the CNN classifier.

User's rating on a video is computed combining explicit and implicit activity. Users can explicitly vote a video on a 5 point scale with a visual widget. Number of visualizations, frame browsing and annotations are also taken into account.

In order to reduce the dimensionality of the item-item matrix used by the algorithm, a pre-filtering on the set of possible videos to suggest is performed. Given a user *u*, we extract a set F_u of videos for which *u* generated a rating.]

For each video v_i contained in F_u, the system selects the *top-N* similar videos creating a subset of similar videos S_i. The set of videos that will be used for the item-based recommender for user *u* is then composed by the union of all the subsets S_i.

The set of video R_u U F_u is used to create the item-item matrix used for recommendation.

$$R_u = \bigcup_{i=1}^{|F_u|} S_i .$$

Evaluation

The SN dataset is composed by 632 videos, of which 468 have been annotated with 1956 comments and 1802 annotations. 613 videos were rated by 950 of the 1108 users of the prototype SN.

User profile interface - A/B test: users were exposed to one of two variants of the SN, featuring (group B) or not (group A) the profile curation interface. There were 464 active users (321 in group A and 143 in group B) with a conversion rate of 3.75 and 5.81 average comments. User annotation average increased by a factor of 2.06.

Visual saliency and manual annotations: in the experiment were considered: i) the number of comments added without using the most salient frames carousel and ii) all the comments, i.e. adding also those coming from a click in the carousel. Results of case i show that 53.5% of user comments are on frames with a saliency above the average saliency of the videos, and that the percentage of frames above the average saliency is 46.5%.

Recommendation: the RS is evaluated, in terms of RMSE, comparing it to several baselines: i) standard item-based RS, that considers users ratings of all the videos; ii) RS working on a selection of videos, based on similarity computed using system categories only; iii) RS working on a selection of videos, based on content similarity (i.e. automatic annotations) computed on n randomly selected frames; iv) RS working on a selection of videos, based on content similarity computed on n frames with visual saliency above the average; v) RS working on a selection of videos, based on content similarity computed on a) frames with visual saliency score above the average and b) crowdsourced annotations.