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

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

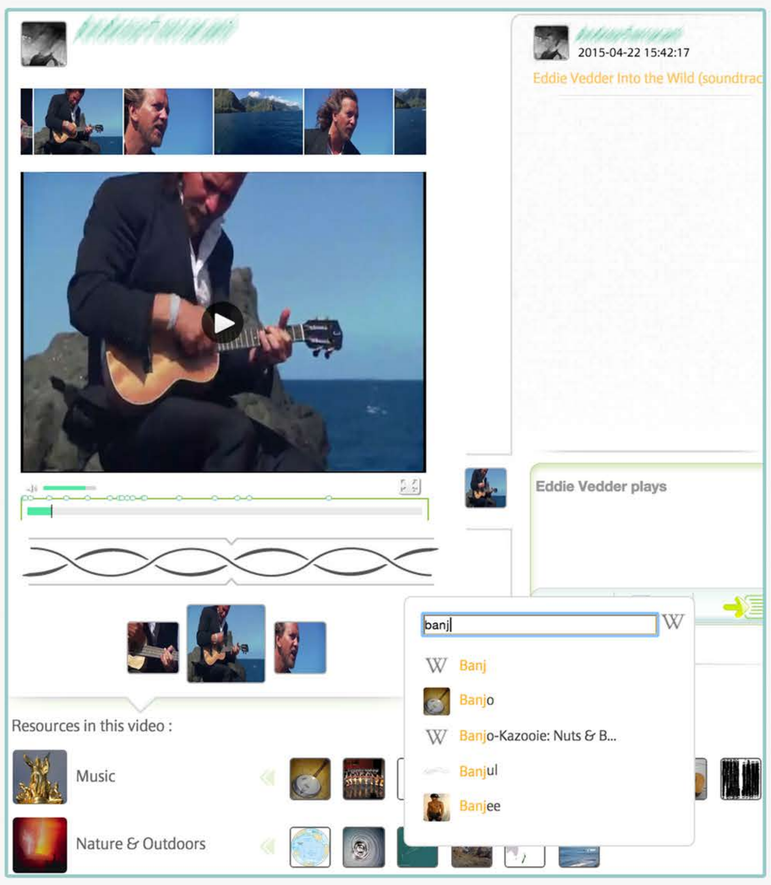
The Project

In this demo we present a system for **content-based video recommendation** that exploits **visual saliency** to better represent video features and content.

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.

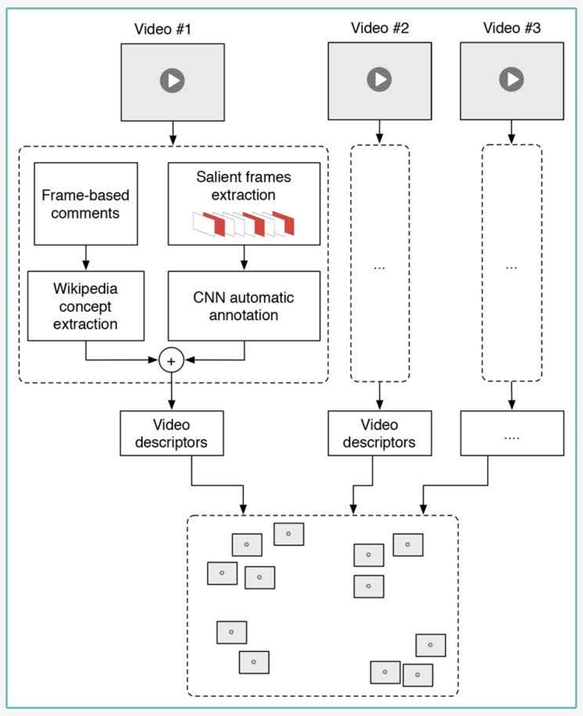
The system exploits **automatic annotations** from CNN-based classifiers on salient frames and **user generated annotations**.

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 and classified using a semantic distance with a nearest neighbour approach.



The system

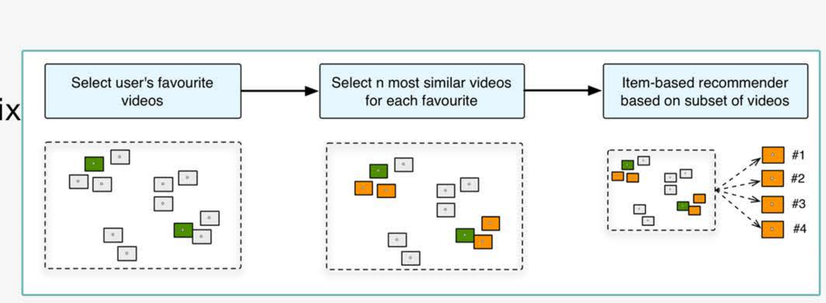
Visual saliency is used at the interface level to propose to the users possible frames of interest through a carousel above the video player, to ease the addition of comments and annotations.
At the automatic annotation level, visual saliency is used to reduce the computational cost of processing all the frames. The salient frames to be used in the system interface are selected by identifying the peaks of salience of the video using a **crest detection algorithm**.



Crowdsourced annotations. Users can comment videos at frame level and can add semantic references to **Wikipedia entities** in comments using an autosuggest widget.
A vector of categories is used to represent video content according to the comments of the users. The vector is defined by calculating for each category the semantic distance of each annotation to the categories of the taxonomy. Annotation can have been added manually or **extracted automatically** using Wikification. This semantic relatedness between the terms is obtained using the **Web Link Based Measure**.

Visual features. 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 LibCCV2 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 threshold.

The **recommender** implements an **item-based collaborative filtering** that builds an item-item matrix determining similarity relationships between pairs of items. The recommendation step uses the **most similar items** to a user's already-rated items per category to generate the 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.

Evaluation

A **dataset** has been collected by hiring 812 workers from the **Microworkers** web site.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.

We evaluate the performance of the proposed recommender, in terms of **RMSE**, comparing it to several baselines:

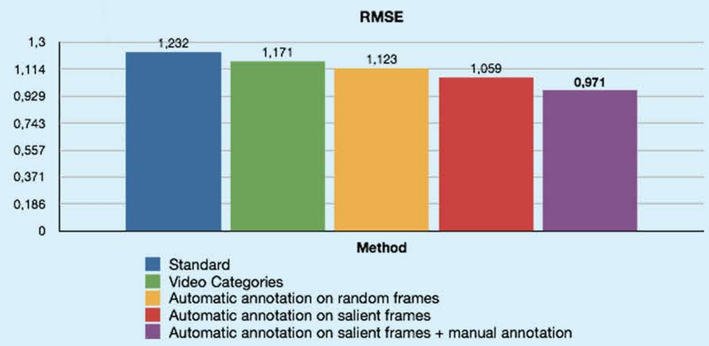
1) **standard item-based** recommender, that consider user ratings for all the items of the system;

2) recommender working on a selection of items, based on similarity computed using **system categories only** (no BoW content description);

3) recommender working on a selection of items, based on content similarity (i.e. automatic annotations) computed on **randomly selected frames**

4) recommender working on a selection of items, based on content similarity computed on **n frames with visual saliency** score above the average;

5) recommender working on a selection of items, based on content similarity computed on a) n frames with **visual saliency** score above the average and b) crowdsourced **manual annotations**



Watch the demo video of the system online