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

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

ABSTRACT

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.

Keywords

Video tagging, crowdsourcing, automatic video tagging, item-based video recommendation, user profiles, visual saliency

1. INTRODUCTION AND RELATED WORK

Collaborative Filtering (CF) is a technique often used by Recommender Systems (RSs) which aims at predicting interesting items to a user based on the preferences, explicit and implicit, of other users. A standard item-based video RS builds its prediction model considering user preferences for videos, expressed according to ratings, and suggests potential videos of interests comparing their distributions. Hybrid approaches in RSs have been proved to give best results [5]. These approaches combine CF with content-based techniques and reduce issues related to the large amount of data to be annotated and data sparsity. Recommending relevant videos can help users to find the most pertinent content according to their view habits or preferences. As shown in [21], recommendation is a powerful force in driving users to watch other videos, much more than direct search of new videos. Hybrid approaches presented in the literature typically exploit textual video metadata, sometimes complemented by multimedia content analysis [19], user profiling, social features and User-Generated Content (UGC) [1, 4, 8]. Crowdsourced data is usable information that can be leveraged to improve different online services. In [3] crowdsourced annotations are used to create video previews that are more related to the queries of the users, to improve video retrieval. In [14] the performance of a video retrieval system based on crowdsourced annotations of sport videos shows that despite the heterogeneity and poor quality of the annotations, they are close to ground-truth. Time accurate annotations of social videos, based on user comments and temporal, personalised topic modelling, has been proposed in [17]. In [18] a large crowdsourcing experiment has been carried out to analyse the differences between "timed" tags (i.e. added to a specific timecode in a video) *versus* "timeless" tags. The authors observed that most of the visually-related tags are relevant for short segments of the video, i.e. people tend to tag when something is "flashed" in the video.

We build on top of these studies and propose the adoption of an hybrid approach in which a brief and comprehensive representation of video content can improve the performance of a standard recommender based on CF (i.e. using only ratings). The approach relies on content-based features gathered both through crowdsourced and CNN-based classifiers annotations. The dataset has been collected through a prototype of a Social Network (SN). Annotations collection is improved exploiting two human factors: *i)* user profile interfaces and *ii)* video frames visual saliency. The main goals are: *i)* to increase the number of crowdsourced annotations, that provide an enrichment of automatic video annotations; *ii)* to improve the quality of video recommenders through video content analysis.

The paper is organised as follows: in Sect. 2 the social network architecture and modules are described. Experimental results are presented in Sect. 3 to show the influence of user profiles and visual saliency on the collection of user annotations of videos. Evidence is given that systems featuring a user profile interface stimulates user activity, increasing the number of annotations. We also show that frames with an high visual saliency are more likely to be annotated; this can be used as a criterion *i)* to suggest to users relevant frames; *ii)* to filter relevant frames for automatic annotation. The recommender is evaluated in Sect. 3.3.

2. THE SYSTEM

The item-based RS has been implemented in a prototype of a SN[1]. The idea behind the SN is to exploit user profiling techniques to propose to the user targeted recommendations of videos, exploiting suggestions of topics of interest

[1] `http://fiona.micc.unifi.it/intime`

ICMR'16, June 06-09, 2016, New York, NY, USA

DOI: http://dx.doi.org/XXXXXXXX

and similar users. This is achieved tracking user's activities on the SN, such as comments, number of video views, click-through data and video ratings. Users can comment videos at frame level tagging concepts derived from Wikipedia. All the concepts manually added are clustered in 54 categories using Fuzzy K-Means and classified using a semantic distance [9] with a kNN approach. Categorised resources in videos are used to build a vector describing video content, then exploited in the RS. The SN also allows users to build a public personal profile of resources of interest from those extracted from comments or added by the SN users. The profile module is exposed relying on the hypothesis that self-expression and *self-esteem* can be exploited to engage the user in the annotation process, in this way easing the collection of crowdsourced annotations. Salient frames of each video are extracted and related users activity on them is monitored in order to verify if visual saliency can affect user engagement with the system. The positive correlation, verified in Sec. 3.2, is exploited at the interface level for easing the annotation process proposing a widget of most salient frames above each video. Automatic video annotations are then extracted using a CNN-classifier on the more salient frames.

User profile interface. As noted in [20], profile curation is inherent to the use of SNs since management of personal content is integrated with its generation. The content that people choose to share online has to do with how they curate their self-image and present themselves to others. In a 2013 survey, participants ranked their relational identities as most important to them when sharing content on social media [11]. SNs such as Facebook and LinkedIn, for example, are commonly regarded as a space for personal self-expression and self-promotion [15]: users shape their identities in order to gain popularity and reach more and more recognition and connectedness. 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, as shown in Fig. 1.

Many factors influence users' continued intention to use SNs such as social interactions, knowledge expansion and targeted recommendations [7]. The users' desire of social interactions has been demonstrated to increase the number of *likes* and comments in [6]. The assessment of users engagement with content gives the opportunity to improve targeted services and recommendation. User propulsion at showing knowledge for self-promotion is used, for example, by platforms such as LinkedIn (Q&A) and StackOverflow as a mean to increase the quality and number of crowdsourced annotations but, to the best of our knowledge, there is not a study in the literature which confirms that user profile interfaces affect and improve user activity in SNs. In 3.1 we have conducted a controlled experiment to show how the user's effort to shape his public identity can be exploited

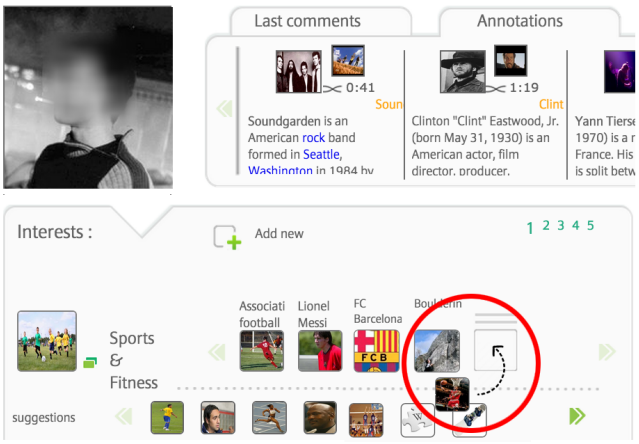

Figure 1: User profile interface: the user can publish resources of interest dragging suggestions from the below to the above carousel.

to increase user's activity and production of crowdsourced annotations.

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. Videos are preprocessed to eliminate letterboxing (i.e. black bars in videos). Then, visual saliency maps are extracted for all the video frames. Maps are defined by a visual attention model which uses a dynamic neural network on multiscale image features computed with the iLab Neuromorphic Toolkit [12]. Salient frames in the video carousel are selected by identifying the peaks of saliency using the crest detection algorithm proposed in [16]. Automatic annotation is performed on frames selected computing the average saliency of the video and choosing those above the average, to have a dense sampling of video content.

Crowdsourced annotations. Users can comment videos at frame level and add semantic references to Wikipedia entities using an autosuggest widget, as shown in Fig. 2. Wikipedia entities are also extracted automatically using entities detection.

A carousel of the most salient frames is also shown above the video player as a video summary. This facilitates fast and accurate annotations at exact timecodes, since users are more likely to interact with salient frames rather than with the less visually interesting ones. A vector of categories C, with the same dimensionality of the SN categories taxonomy, is used to represent video content. Each category in C is assigned with a weight defined by the average of the semantic distance of each annotation to the categories' taxonomy. This semantic relatedness between the terms is obtained using the Wikipedia Link-based Measure [9].

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

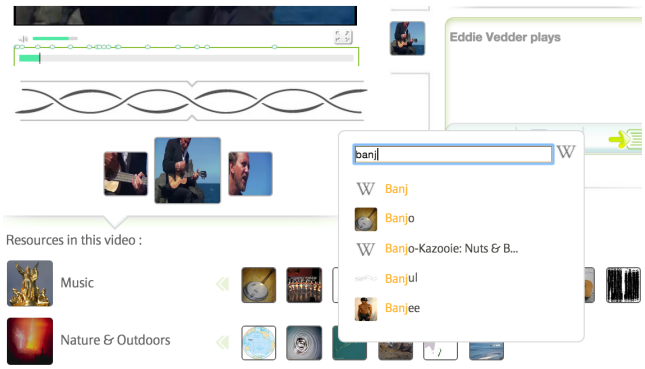

Figure 2: Wikipedia annotation in video frame-level comments.

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 [2] 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. Compared to user-based CF approaches, item-based recommenders minimise the sparse item ratings issue, are scalable and in general perform better than user-based recommenders [13]. 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, namely:

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

The set of video $R_u \cup F_u$ is used to create the item-item matrix used for recommendation. This set is significantly smaller than the whole collection of videos contained in the system. The pre-filtering step uses several approaches in order to infer the top-N similar videos. These approaches, reported in Sect. 3.3, exploit automatic and crowdsourced annotations as well as visual saliency, and use distance measures to compute the overlap between histograms distributions.

3. EXPERIMENTAL RESULTS

Recommendation is a prediction problem: the system should be able to predict the user's level of interest in specific items (e.g. videos) and rank these according to their predicted values [10]. In order to evaluate the accuracy of the prediction, a percentage of the collected data, represented by users ratings on videos, is extracted and used as test data, not used to train the RS. The RS produces rating predictions for the missing test data, that are compared to the actual values in order to evaluate the accuracy. The performance is evaluated using Root Mean Square Error (RMSE). The more accurately the RS predicts user ratings, the lower the RMSE will result. 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.

3.1 User profile interface

An A/B test experiment was conducted at the interface level to test the user profile influence on users' comments activity. The experiment was run on all the active users of the prototype SN for three months. Users were exposed to one of two variants of the SN, featuring (i.e. the variant) or not (i.e. the control) the profile curation interface. The variant was introduced in the third month, so that the number of users exposed to the variant is smaller than that of the control interface. Users who logged into the system and commented on videos since the third month were assigned to the variant (group B), whilst the others were assigned to the control group (group A). In this period of time 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. The result was statistically significant and validated by a t-test that gave a t-difference = -2.684. Minimum sample size for the evaluation criterion validity was calculated and resulted in 127 for both group A and group B with an optimum allocation ratio of 3.42. Results show a positive correlation between the use of the user profile interface and the increment in user annotations, and suggest that modules for profile curation can be effective in improving conversion rate in user online activity (e.g. videos annotations).

3.2 Visual saliency and manual annotations

The impact of the visual saliency of video frames on user comment activity has also been tested. 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%. Therefore, salient frames receive more attention by users, although not considerably. Results improve consistently considering also carousel driven annotations as in case *ii*: in fact the percentage of comments increases to 65.24%. Percentage of comments carousel driven is 24.01% of the overall dataset, showing that one out of four comments are added using the carousel: it is an high percentage considering threaded comments, added by users in response to others. So, it can be said that salient frames suggestion can be useful if proposed in a web interface as to

visually capture the user's attention and help in the annotation tasks.

3.3 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 (no BoW content description); *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.

Results are reported in Fig. 3 and show how the proposed *v)* approach results in a lower RMSE value than all the other approaches. In particular, it can be observed that video representation using salient frames improves over random selection, and that the addition of semantics extracted from manual annotations provides another improvement. In this experiment the threshold used to select the confidence scores of the classifiers is 0.85. In a second experiment we have evaluated the effect of the confidence of the classifier, using a threshold of 1. In this case the RMSE is further improved from 0.97 to 0.86.

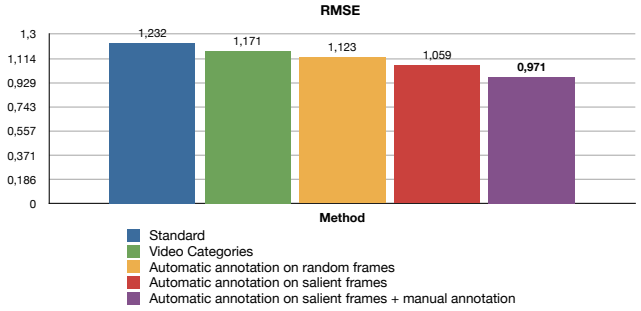

Figure 3: Comparison of the proposed recommender (rightmost result) w.r.t. baselines in terms of RMSE.

4. CONCLUSIONS

In this paper we presented a system to improve an item-based video RS. The RS uses a reduced item-item matrix, computed from content based description of videos obtained from crowdsourced and automatic annotations. User engagement through profile curation and visual saliency has been used *i)* to increase the number of crowdsourced annotations, presenting the most relevant frames to users, and *ii)* to address system scalability in terms of automatic annotation, reducing the number of frames to be processed. The effectiveness of exploiting human factors for user engagement (i.e. self-esteem in user profile interfaces and visual saliency) is evaluated by user experiments on a SN prototype. Experiments show also that the proposed RS improves over the standard implementation of an item-based algorithm, and that the combination of manual and automatic annotations is more effective than the use of a single type of annotations. A positive correlation of the two human factors with the performance of the RS is not yet fully demonstrated, but it can be hypothesized and it is worth of further investigation.

Acknowledgments. This work is partially supported by the "Social Museum and Smart Tourism" MIUR project (CTN01_00034_231545).

References

[1] M. Bertini, A. Del Bimbo, A. Ferracani, F. Gelli, D. Maddaluno, and D. Pezzatini. Socially-aware video recommendation using users' profiles and crowdsourced annotations. In *Proc. of WSAM*, 2013.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. of BMVC*, 2014.

[3] B. Craggs, M. Kilgallon Scott, and J. Alexander. ThumbReels: Query sensitive web video previews based on temporal, crowdsourced, semantic tagging. In *Proc. of CHI*, 2014.

[4] P. Cui, Z. Wang, and Z. Su. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *Proc. of ACM MM*, 2014.

[5] A. Felfernig, M. Jeran, G. Ninaus, F. Reinfrank, S. Reiterer, and M. Stettinger. Basic approaches in recommendation systems. In *Recommendation Systems in Software Engineering*, pages 15–37. Springer, 2014.

[6] F.-H. Huang. Motivations of Facebook users for responding to posts on a community page. In *Proc. of OCSC*, 2013.

[7] K.-Y. Lin and H.-P. Lu. Intention to continue using Facebook fan pages from the perspective of social capital theory. *Cyberpsychology, Behavior, and Social Networking*, 14(10):565–570, 2011.

[8] X. Ma, H. Wang, H. Li, J. Liu, and H. Jiang. Exploring sharing patterns for video recommendation on YouTube-like social media. *Multimedia Systems*, 2013.

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

[10] B. Mobasher. Data mining for web personalization. *The adaptive web*, 2007.

[11] R. Morris. Identity salience and identity importance in identity theory. *Current Research in Social Psychology*, 21(8):23–36, 2013.

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

[13] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proc. of WWW*, 2001.

[14] F. Sulser, I. Giangreco, and H. Schuldt. Crowd-based semantic event detection and video annotation for sports videos. In *Proc. of CrowdMM*, 2014.

[15] J. Van Dijck. 'You have one identity': performing the self on Facebook and LinkedIn. *Media, Culture & Society*, 35(2):199–215, 2013.

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

[17] B. Wu, E. Zhong, B. Tan, A. Horner, and Q. Yang. Crowdsourced time-sync video tagging using temporal and personalized topic modeling. In *Proc. of ACM KDD*, 2014.

[18] P. Xu and M. Larson. Users tagging visual moments: Timed tags in social video. In *Proc. of CrowdMM*, 2014.

[19] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li. Online video recommendation based on multimodal fusion and relevance feedback. In *Proc. of ACM CIVR*, 2007.

[20] X. Zhao and S. E. Lindley. Curation through use: Understanding the personal value of social media. In *Proc. of CHI*, 2014.

[21] R. Zhou, S. Khemmarat, L. Gao, and H. Wang. Boosting video popularity through recommendation systems. In *Proc. of DBSocial*, 2011.