Serendipity-driven Celebrity Video Hyperlinking

Shujun Yang[1], Lei Pang[1], Chong-Wah Ngo[1], Benoit Huet[2]
[1]Department of Computer Science, City University of Hong Kong, Hong Kong, China
[2]EURECOM, Sophia Antipolis, France
{xingguinvwu, cactuslei}@gmail.com, cscwngo@cityu.edu.hk, Benoit.Huet@eurecom.fr

ABSTRACT

This demo showcases the utility of video hyperlinks with celebrities as the link anchors and their social circles as targets, aiming to help users quickly explore the *aboutness* of a celebrity by link traversal. Through content analysis, our system embeds hyperlinks into videos such that users can click-and-jump between celebrity faces in different videos to get-to-know their social circles. One peculiar feature is the ability of the system in providing links that maximize users' chance encounter, or serendipitous experience, beyond information need. Our system is enabled by two key components, name-face association and diversity-based ranking, for the *aboutness* and *serendipity* features respectively in hyperlinking. The former component assigns names video faces while mining celebrity social networks. The latter performs topic modeling so as to rank videos based on topic diversity.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Design, Human Factors

Keywords

Video hyperlinking; Face naming; Serendipity

1. INTRODUCTION

Making searchers happy beyond supplying relevant information is a problem seldom being addressed in the field of multimedia retrieval. Video hyperlinking is one such effort initiated by MediaEval [9] and TRECVID [15] for embedding links into videos so as to enable efficient and on-the-fly content navigation between video segments. One usage scenario is detail-on-demand [1, 8, 12, 14], where similar to HTML web pages, a user clicks on an anchor text and the browser leads user to a new page with additional information about the subject of interest.

This paper presents a system for hyperlinking celebrities in videos. Celebrity is picked as the subject of exploration in this demo, due to the fact that the majority of queries and significant amount of video uploads on social sharing platforms concern famous people. According to YouTube trends map, around 80% of popular videos are people related, and among them around 75% are about celebrities. The application scenario is that a user may want to know the social circle of a celebrity. Hyperlinks allow the user quickly learns the circle by listing socially connected celebrities along with their related videos for exploration. Connecting videos based on featuring celebrities will enhance the accessibility of web videos for nonlinear interactive browsing/navigation hence improving user experience.

Our system emphasizes *aboutness* [1] and *serendipity* [13] when recommending videos for hyperlinking. Different from *relevancy* which seeks for similar videos, *aboutness* provides summary along with links to other videos for detailing or explaining the subject of interest. *Serendipity* refers to unexpected but delightful chance encounters, by referring user to additional information that is previously unknown and could potentially "surprise" viewers. Our system models *serendipity* as content diversity, by recommending videos based on topic diversity or novelty. Ideally, the first few ranked videos should cover different perspectives of a subject. By doing so, users quickly grasp new knowledge and hence enhance the chance of experiencing serendipitous browsing. Figure 1 illustrates our system. While a video showing Barack Obama is playing, the user may hover with the mouse or tap on him to expose his social circle composed of Michelle Obama, Jinping Xi, David Cameron, etc. Each celebrity in the circle is linked to a list of videos ranked based on diversity. When selecting Jinping Xi from Obama's social circle, the top-3 videos shown on the right end side highlight different content topic-wise. Specifically, both v_1 and v_3 report meetings between Barack Obama and Jinping Xi in different occasions, and v_2 is about the nuclear problem in North Korea that involves both political leaders.

Existing research efforts mostly treat video hyperlinking as a retrieval task. Concretely, given a query (or link anchor), the task is to recommend videos (or link targets) that will potentially interest users based on relevancy between anchor and target. Examples of approaches include ranking videos by low-level textual-visual features [3, 4, 7, 11], visual concept classifiers [17] and query expansion through linked data [10]. A major drawback of these approaches is that similar or sometimes redundant videos are recommended, rather than the videos that could supplement the

Figure 1: An example of linking Obama to his social circle. Each celebrity (e.g., Jinping Xi) in the circle is further linked to diverse videos elaborating different perspectives of the relationship with Obama.

anchor for detail-on-demand browsing. Furthermore, these approaches use to return a long list of candidate videos, resulting in the problem of "link explosion" as mentioned by [1]. One exception is the work by [18], which emphasizes serendipity in hyperlinking, by constructing a deep topical structure of 10 hierarchies for weighting the contribution of videos based on topic specificity. Therefore, videos of different granularities in topics could be selectively recommended, which effectively controls the number of videos for serendipitous hyperlinking. Our demo is similar in spirit to [18], specifically to recommend limited but diverse videos useful for exploration, but operates in the domain of social circles. Potentially there are rich information (e.g., anecdote, event, rumor, opinion) to be exploited for hyperlinking in this domain, which are yet to be studied. Particularly, concerning the utility of hyperlinks for helping users navigate in large scale media collections.

2. SYSTEM ARCHITECTURE

We developed two key modules for enabling celebrity hyperlinking; face naming and diversity-based ranking. For practical consideration, face-name corresponding is learned from weakly tagged images crawled from the Web (sections 2.1-2.2). LDA (Latent Dirichlet Allocation) is performed to discover fine-grained topics underlying within the video corpus for novelty ranking (sections 2.3-2.4).

2.1 Name-face association

We employ CRF (Conditional Random Field) for name-face association based on our prior work [16]. CRF considers a rich set of relationships including name-to-face unary energy and multiple face-to-face pairwise energies. This results in a network of faces and names as vertices and energies as edges for name inferencing. Based on person names given in the video metadata, weakly tagged web images are crawled and used for face model training. By these models, unary potential estimates the likelihood of a detected face belonging to a mentioned name in the video metadata. Pairwise potential considers similarities between faces appearing in different shots. In addition, multiple constraints including spatial-temporal proximity of faces and restriction that no two faces in a frame should be labeled the same name are incorporated for pairwise energy modeling. The inference of names is performed by LBP (loopy belief propagation) algorithm, allowing null assignment of names.

2.2 Deep face feature

As in [16], CRF is sensitive to the underlying facial feature representation, especially when the faces undergo moderate to large changes in visual appearance. Different from [16] which represents face with local features extracted from 13 facial regions, deep face feature is employed in our system. We employ the deep model in [2] due to its greater representational efficiency, which achieves state-of-art face recognition performance using only 128-bytes per face. The input to the model is a cropped frontal face which is rectified by Dlib's real-time pose estimation with OpenCV's affine transformation. The deep network is based on NN4 of GoogLeNet style Inception models where the input face resolution is 96×96. The facial feature is extracted from the layer corresponding to face embedding, which is the layer after the deep CNN and L2 normalization. The feature is encoded as a vector in 128 dimension. Different from [16] which employs GMM (Gaussian Mixture Model) for learning, SVM classifiers are trained for each celebrity face model.

2.3 Corpus-wise topic modeling

As in [18], topic modeling is performed so as to measure video similarity based on topic distribution rather than BoW (bag-of-words) representation. As empirically studied in [18], using latent topics has a better chance to retrieve videos with diverse content potentially useful for serendipitous hyperlinking. This is in opposition to BoW modeling where very often only highly similar videos are returned. In the implementation, LDA is performed on the video metadata, or specifically words pooled from title, tag and user description. The number of latent topics is empirically set to be 50, and each topic is modeled as a probability distribution over a set of words. In such a way, each video is represented as a vector of 50 dimensions, where each dimension specifies the likelihood of a video belonging to a topic. Similarity between two videos can thus be straightforwardly measured by cosine similarity.

2.4 Novelty Ranking

As videos are recommended based on celebrity relationship, the first step is to model celebrity co-occurrence in the video corpus. Our system performs person name extraction from metadata and verifies each name against Wikipedia. Basically names without the category tag "birth year" in the Wikipedia pages are removed from processing. The relationship between two names is then calculated by Dice coefficient based on the frequency of two names being tagged in videos. By doing so, the system maintains a sparse graph with celebrities as vertices and edges as closeness between two vertices. This sparse graph is represented as an inverted index. Each entry in the index corresponds to a celebrity linking to his or her social circle, followed by the sets of videos where both celebrities appear.

Using this inverted index, each named face in a video can be quickly hyperlinked with the related celebrities based on co-occurrence. Furthermore, the set of relevant videos can also be easily retrieved for hyperlinking. To limit the number of videos while not linking to redundant videos, we experiment two different methods, based on greedy search and clustering respectively, for diversifying the recommended videos for hyperlinking.

2.4.1 MMR-based greedy search

Inspired by MMR [5], maximal diversity (MD) is used as

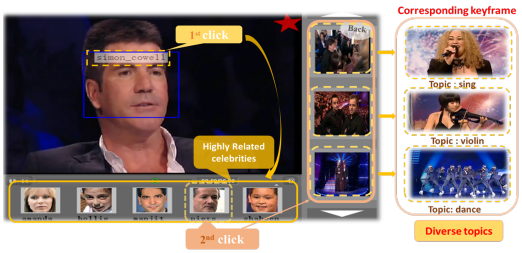

Figure 2: System interface showing how users can explore "Simon Cowell" and his social circle through interacting with hyperlinks for video navigation.

the criterion for selecting diverse videos. Denote R as the set of all videos having both celebrities, and $S \subset R$ is the subset already selected by the algorithm. Thus, $R \setminus S$ is the subset of as yet unselected videos. The next video $v_i \in R \setminus S$ to be included in S is selected based on MD, defined as

$$Arg \min_{v_i \in R \setminus S} \max_{v_j \in S} sim(v_i, v_j) \quad (1)$$

At each iteration, $v_i \in R \setminus S$ is compared against every video in S, and its similarity with S is determined based on the highest similarity score. The video to be included into S is the most dissimilar one, i.e., with minimal relevance score, so as to maximally diversify the topics of existing videos in S. The process starts by picking a video from $R \setminus S$ as the first video. The second selected video is the one most diverse from the first video. The remaining videos are then greedily included into S until the desired number of videos is reached. Here, we only select $K = \min(5, N * 0.5)$ videos for hyperlinking, where N is number of videos. We repeat this process by picking every video $v_i \in R$ as the first video. This ends up with N different numbers of lists. The list which has the minimum average similarity of K videos is picked as the recommended list for hyperlinking. Note that if the value of N is large, the process is only repeated for a subset of videos randomly picked. Finally, we match the K selected videos with the remaining $N - K$ videos based on their similarities. In other words, each of the remaining videos is matched to the closest one in the K selected videos. The K videos in the recommended list is ranked based on the number of matches. Similarity between videos is computed using cosine similarity between their 50 dimensional topic vectors (section 2.3).

2.4.2 *K-medoids clustering*

The second method is based on video clustering by K-medoids, which is a variant of Kmeans algorithm and is more robust to outliers. By clustering on 50 dimension vector resulting from LDA, similar videos are grouped and the idea is to pick one video per group for ranking. We set the number of medoids as $K = \min(5, N * 0.5)$, where N equals to number of videos, and select the most centrally located video (i.e., medoids) of each cluster for hyperlinking. The ranking of videos is based on the size of their clusters.

3. DEMONSTRATION

The demonstration focuses on showcasing the utility of hyperlinks on a dataset consisting of 2,583 videos crawled from YouTube where 141 celebrities have been identified.

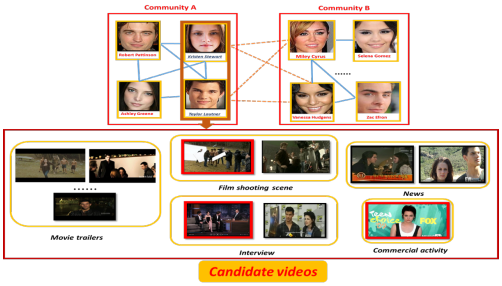

Figure 3: Example of candidate videos (manually grouped into clusters) to be selected and linked for elaborating the relationship between "Kristen Stewart" and "Taylor Lautner". Potentially serendipitous targets are highlighted with red boxes.

The average length of a video is about 4 minutes. Figure 2 shows our system's interface which offers three panels users can use for efficient navigation and exploration of celebrity relationships in video collections. The first panel displays a video, and a "star" on the top right corner hints about the availability of a hyperlink. By mouse hovering on the face, the corresponding name "Simon Cowell" is prompted. Further clicking the name shows a list of frequently co-occurring celebrities in the second panel. In Figure 2, the celebrity "Piers Morgan" is selected and the recommended videos are displayed on the third panel based on K-medoids clustering. Each of the videos has a different theme, respectively, "singing", "violin show" and "dancing", where "Simon Cowell" and "Piers Morgan" are both judges for these TV shows. Through these videos, users quickly know the kinds of performances that both celebrities once judged. The ranking of videos also inherently gives the clue that most frequent performances are about "singing".

4. EVALUATION

This section gives insights about the performance of our system. We first introduce the video dataset used for demonstration, followed by the experimental results for celebrity naming and diversity ranking.

4.1 Dataset

The video corpus is WebV-Cele [6], which consists of 2,583 YouTube videos. A total of 141 celebrities are labeled for 19,240 frontal faces in the dataset. The celebrities altogether form 12 social communities. Figure 3 shows a snapshot of few celebrities in two networks. Take "Kristen Stewart" and "Taylor Lautner" as examples, our system retrieves 12 videos which are relevant to them. When asking a user watching these videos, three videos were picked and regarded as "surprise". Interestingly, while most candidate videos are movie trailers, the selected videos are a raw footage not screening in the film "twilight" and an interview that both celebrities talked about their personal lives. For this example, our system is able to pick two out of the three "surprise" videos as link targets.

4.2 Celebrity Naming

Around 4,000 web images are crawled for training face models for 141 celebrities. Images with more than one face detected are excluded from learning. We employ accuracy,

which counts the number of faces correctly labeled, to measure the performance of name-face association. Using deep features for training SVM classifiers, we achieve an accuracy of 52.1%. Further using CRF for probabilistic labeling, the accuracy is boosted to 59.3%, versus 58.4% as reported in [16] using hand-crafted local features and GMM.

4.3 Serendipitous Hyperlinking

We measure serendipity based on video diversity. In the dataset, there are 199 celebrity pairs detected by our system. We rank the celebrity pairs in descending order of their video frequency, and select the top 13 pairs for evaluation. The number of candidates for hyperlinking video to these pairs ranges from 6 to 53. The evaluation is to measure the degree of diversity for top K videos recommended by the system, where $K = \min(5, N * 0.5)$ and N is the number of candidate videos. To generate ground-truth for experiment, an evaluator was recruited to manually group topic-wise similar videos into clusters. The evaluator took 2 days to complete the process, and the number of clusters ranges from 3 to 5 for 13 celebrity pairs. In addition, we asked the evaluator to pick serendipitous videos from each celebrity pair. A total of 26 videos are picked for 13 pairs. We employ cluster recall (CR) to measure the diversity, defined as

$$CR(K) = \frac{clusters(K)}{tc} \quad (2)$$

where $clusters(K)$ is the number of different clusters encountered up to and including the K^{th} ranked video, and tc is the total number of clusters in the ground-truth. The perfect value for CR is 1.0 when the top K videos cover all the clusters. Our result shows that K-medoids (CR=0.74) outperforms MMR (CR=0.66). Using K-medoids, for 6 out of 13 celebrity pairs, our system is able to recommend serendipitous videos picked by the evaluator.

5. CONCLUSION

We have presented a system allowing users to navigate in video collections by following hyperlinks created based on celebrity circles. Our analysis, shows that our approach, using face-naming and topic diversity modeling is able to effectively select serendipitous videos. It was noticed that the evaluator favored amusing or unexpected video (e.g., two celebrities introducing a cute cat). Recommending such videos is beyond the capability of the current system. Future work will include online learning of personal interest and analysis of emotional and semantic cues in videos for serendipitous videos detection.

6. ACKNOWLEDGEMENT

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11210514), and a grant from the National Hi-Tech Research and Development Program (863 Program) of China under Grant 2014AA015102.

7. REFERENCES

[1] R. Aly, R. J. Ordelman, M. Eskevich, G. J. Jones, and S. Chen. Linking inside a video collection: What and how to measure? In *WWW*, 2013.

[2] B. Amos, B. Ludwiczuk, J. Harkes, P. Pillai, K. Elgazzar, and M. Satyanarayanan. OpenFace: Face Recognition with Deep Neural Networks. http://github.com/cmusatyalab/openface.

[3] W. Bailer, M. Lokaj, and H. Stiegler. Context in video search: Is close-by good enough when using linking? In *ICMR*, 2014.

[4] C. A. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. Multimodal reranking of content-based recommendations for hyperlinking video snippets. In *ICMR*, 2014.

[5] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

[6] Z. Chen, C.-W. Ngo, W. Zhang, J. Cao, and Y.-G. Jiang. Name-face association in web videos: A large-scale dataset, baselines, and open issues. *JCST*, 29(5):785–798, 2014.

[7] Z. Cheng, X. Li, J. Shen, and A. Hauptmann. Cmu-smu@ trecvid 2015: Video hyperlinking. 2015.

[8] J. Doherty, A. Girgensohn, J. Helfman, F. Shipman, and L. Wilcox. Detail-on-demand hypervideo. In *ACM MM*. ACM, 2003.

[9] M. Eskevich, R. Aly, D. Racca, R. Ordelman, S. Chen, and G. Jones. The search and hyperlinking task at mediaeval 2014. In *Working Notes Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, 2014.

[10] M. Eskevich, G. J. Jones, R. Aly, R. J. Ordelman, S. Chen, D. Nadeem, C. Guinaudeau, G. Gravier, P. Sébillot, T. de Nies, P. Debevere, R. Van de Walle, P. Galuscakova, P. Pecina, and M. Larson. Multimedia information seeking through search and hyperlinking. In *ICMR*, 2013.

[11] M. Eskevich, G. J. F. Jones, C. Wartena, M. Larson, R. Aly, T. Verschoor, and R. Ordelman. Comparing retrieval effectiveness of alternative content segmentation methods for internet video search. In *CBMI*, 2012.

[12] M. Eskevich, H. Nguyen, M. Sahuguet, and B. Huet. Hyper video browser: Search and hyperlinking in broadcast media. In *ACM MM*, 2015.

[13] A. Foster and N. Ford. Serendipity and information seeking: an empirical study. *Journal of Documentation*, 59(3):321–340, 2003.

[14] A. Girgensohn, L. Wilcox, F. Shipman, and S. Bly. Designing affordances for the navigation of detail-on-demand hypervideo. In *AVI*, 2004.

[15] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Quénot, and R. Ordelman. Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2015.

[16] L. Pang and C.-W. Ngo. Unsupervised celebrity face naming in web videos. *IEEE Transactions on Multimedia*, 17(6):854–866, 2015.

[17] B. Safadi, M. Sahuguet, and B. Huet. When textual and visual information join forces for multimedia retrieval. In *ICMR*, 2014.

[18] A.-R. Simon, R. Bois, G. Gravier, P. Sébillot, E. Morin, and S. Moens. Hierarchical topic models for language-based video hyperlinking. In *Workshop on Speech, Language and Audio in ACM Multimedia*, 2015.