Content-Based Viewer Estimation Using Image Features for Recommendation of Video Clips

Kohei Yamamoto
Grad. School of Information Science and Technology
The University of Tokyo
yamamoto@nlab.ci.i.u-tokyo.ac.jp

Riku Togashi
Grad. School of Information Science and Technology
The University of Tokyo
togashi@nlab.ci.i.u-tokyo.ac.jp

Hideki Nakayama
Grad. School of Information Science and Technology
The University of Tokyo
nakayama@ci.i.u-tokyo.ac.jp

ABSTRACT

Content-based recommendation is a promising approach to overcome the cold-start problem, which is a fundamental problem facing recommendations on video sharing Web sites and smart television. In this work, we propose and investigate an approach in which viewer profiles are estimated from the image features of video clips. Image features are extracted from representative images such as the key frames and thumbnails of video clips and then used as explanatory variables to construct classifiers for predicting the attributes of audiences having a high probability of watching those clips. We evaluated the proposed method using video clips and corresponding viewer demographics data from YouTube. A thorough investigation of various image features and their combinations demonstrated the effectiveness of the proposed method for this task. Our method is completely content-based and we expect it to enable recommendations and target advertising even for users and videos with little historical data and meta-data.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering

General Terms

Algorithms, Design, Experimentation

Keywords

Video Recommender Systems, Content-Based Recommendation, Image Features, Feature Extraction and Selection

1. INTRODUCTION

With the dramatic spread of consumer generated media (CGM) and video sharing sites in recent years, users now have access to an infinite amount of content on the Web. As the amount of available content explosively increases day by day, both users and content providers have to expend enormous effort for users to find the desired content. To alleviate this problem, a variety of recommendation techniques that estimate and display content matching user interests and preferences have been actively studied [1]. Collaborative filtering [6][15], which calculates the similarities between items or users from the historical data of users, is utilized in many services including electronic commerce (EC) sites such as Amazon and social networking services (SNSs) such as Facebook [10].

However, since collaborative filtering requires the historical data of users in order to make recommendations, problems emerge due to the so-called cold-start problem, a fundamental difficulty in recommendation systems regarding users and content with little historical data [12]. Obviously, since a huge number of new video clips are uploaded every day on video sharing sites — about 100 hours of video per minute on YouTube, according to [20] — there is always a large amount of content with little or no historical data. Moreover, unlike SNSs and EC sites, with video sharing sites, many users browse content without logging in, and so historical data is largely absent on the user side, too. Due to the cold-start problem caused by these factors, many video sharing sites cannot provide satisfactory recommendations or accurately targeted advertising at present. According to the statistical data of YouTube [20], only 14% of YouTube views can be monetized [17]. In this study, to overcome the cold-start problem and provide more effective recommendations, we propose a content-based viewer estimation method that estimates the viewership from the image features of video clips.

2. RELATED WORK

Conventional video recommendation techniques from earlier days relied on meta-data to retrieve suitable videos for users [5]. Szomszor et al. [16] proposed a method to create user profiles based on the tendency of tags annotated to videos that users rated highly or poorly in the past. By comparing the tags given to content with the user profile, they can predict the score of new content. However, in practical terms, the meta-data of content on video sharing sites is very limited if not non-existent, and it is obviously difficult to make recommendations for such content. To tackle this problem, Ulges et al. [17] estimated the viewership from video clips based on the content itself without meta-data. In their method, various pre-defined semantic concepts (objects, locations, activities, etc.) are detected using the image features of video clips that are then used as enhanced meta-

data to estimate the viewership. The success of their method demonstrates the effectiveness of concept detection to estimate video viewership. However, since there could be an infinite number of concepts appearing in real-world videos, it is unrealistic to annotate all of them to form a training corpus. Furthermore, since pre-defined concepts are not directly related to the user evaluation of content, the tagging process itself might drop important information for recommendation that the original content features contain. Yang et al. [19] proposed a method to perform content-based recommendation of video clips in combination with the features of various modalities. They calculated the similarity of video clips using features such as colors and movements and the average value and standard deviation of sound tempo. They showed that the video recommendation accuracy improved by combining the features of other modalities with video clip meta-data. In the present study, we focus more on the image features of video clips for content-based recommendation. We thoroughly investigate state-of-the-art image features in the field of computer vision to connect content directly to high level user profiles without the help of pre-defined visual concepts.

3. CONTENT-BASED ESTIMATION OF VIEWER DEMOGRAPHICS

3.1 Outline

The outline of our proposed method is shown in Fig. 1. This method can be roughly divided into two steps:

Step 1. Feature extraction: Various image features are extracted from the representative images (i.e., key frames and thumbnails) of video clips that have viewer demographics data and a feature vector for every single image is computed.

Step 2. Classification: The feature vectors that were obtained in step 1 serve as explanatory variables. Each class, which is defined according to viewership (e.g., gender), is assumed to be an objective variable on top of which we train a linear logistic regression classifier. The target viewership of the video clips is classified from the posterior probability of the objective variable.

By performing these two steps, the classifier for estimating the viewership can be built.

3.2 Feature extraction

Our proposed method extracts image features from each of the representative images (described in detail in Section 4) available from video clips. We investigate several standard visual features as follows.

Gist [13]: Gist is a global image feature that is commonly used for describing the scene of an image. The Gist method divides an image into 4×4 regions and computes 20-directional responses from filter banks for each region. We applied this process to each color in an image to obtain 960-dimensional feature vectors for each image.

Fisher Vector [14]: Fisher Vector is a state-of-the-art method of bag-of-visual-words [3] based global image representation. While bag-of-visual-words represents the distribution of local descriptors through vector quantization, Fisher Vector does it by means of a Gaussian mixture model (GMM). Fisher Vector enables us to obtain more expressive global feature vectors with smaller codebooks. For an implementation, we used SIFT [11], C-SIFT [2], Opponent SIFT [18], and RGB-SIFT [18] as local descriptors and then applied PCA to each descriptor for compression into 64 dimensions. Finally, we encoded the compressed descriptors into a global feature vector using the Fisher Vector framework [14]. We used 64 Gaussians to estimate GMM. To include rough spatial information in an image, we calculated the Fisher Vector from its upper, middle, and lower horizontal regions as well as the image as a whole and then concatenated them to obtain a 32,768-dimensional feature vector for each image.

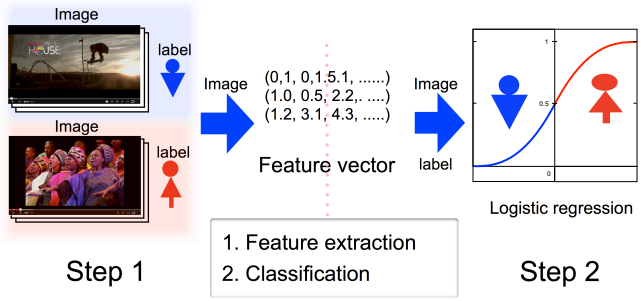

Figure 1: Overview of our proposed method.

CNN (Convolutional neural network) [9]: CNN is an instance of deep neural networks that has been shown to achieve surprisingly high performance in visual recognition. It extracts features through repeated convolution and pooling layers. Krizhevsky et al. [8] trained CNN using the ImageNet [4] dataset, a large-scale labeled image dataset for generic object recognition constructed via crowdsourcing, and won the ILSVRC2012 competition by a large margin. In this study, we use the responses of the last layer of this CNN as image features using Yangqing's software, Caffe [7], which provides the pre-trained CNN model of Krizhevsky's architecture [8]. Consequently, we obtained 4096-dimensional feature vectors for each image. Although this CNN model is tuned for the ImageNet dataset, it is equally valid for other generic image recognition datasets [7].

3.3 Classification

We apply linear logistic regression to build a model that directly links the image features obtained from video clips to user profiles. To combine scores (i.e., posterior probability) obtained from multiple key images of a video, we followed the late-fusion strategy. Specifically, the final classification score of a video clip is defined as the average of the scores of individual images. Further, scores of multiple image features are also integrated in late-fusion. We search for the best combination of features in this way.

4. EXPERIMENT

4.1 Outline

Among the many possible user attributes, in this experiment, we focus on estimating the gender of viewers for each video clip. We use the thumbnail and key-frame images of popular video clips for each gender obtained from the YouTube Trends Map [21]. In the YouTube Trends Map, it is possible to retrieve the 10 most viewed video clips from the past 24 hours specific to the region, gender, and age group of viewers. In this experiment, we searched for and collected

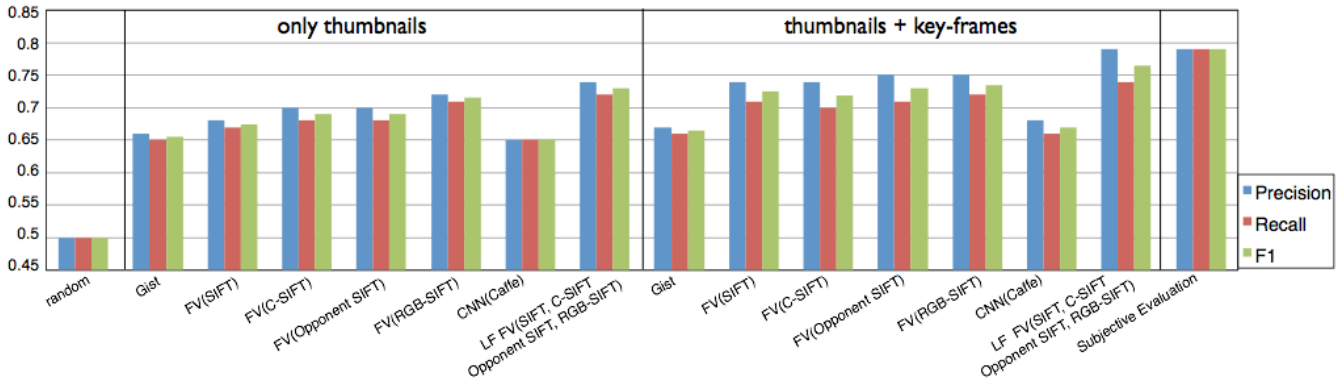

Figure 2: Comparison of precision, recall, and F_1 value of each classifier made from various image features. FV = Fisher Vector, LF = Late-fusion.

video clips confined to the region of Japan and then created a dataset of thumbnail and key-frame images of the most frequently watched video clips by users of each gender. We extract various image features from those images and classify them into one of two genders by applying L2-regularized linear logistic regression. Finally, we randomly selected five persons to classify the target viewership of videos subjectively into male or female in order to qualitatively evaluate human performance at this task.

4.2 Experiment setup

4.2.1 Dataset

We made a URL list of video clips popular with males and females using the YouTube Trends Map, retrieved one thumbnail and four key-frame images (defined by YouTube in advance) for each of them using a public API, and created a dataset from the images. The thumbnail images are 480 × 360 pixels and the key-frame images are 120 × 90 pixels. For a training set, we prepared 5,000 male and 5,000 female video clips (10,000 thumbnail and 40,000 key-frame images) from the list. This dataset contains images that overlap within and between each gender. For the test set, we prepared 500 male and 500 female video clips (1,000 thumbnails and 4,000 key-frame images). Of course, no duplications are included in this test set in order to guarantee the validity of the test.

4.2.2 Image features

As stated in Chapter 3, in this study, we compared Gist, Fisher Vectors that encode local descriptors (SIFT, C-SIFT, opponent SIFT, and RGB-SIFT), and CNN (Caffe) as the image features of each image. As for the local descriptors, we extracted them from each of the images by 8 pixels for one thumbnail and by 5 pixels for one key-frame image. Also, it was necessary to resize the images to 256 × 256 pixels to use CNN (Caffe).

4.2.3 Subjective evaluation

We randomly selected 100 thumbnail images of video clips from the test set used in this experiment and asked our five participants to subjectively classify these images into male or female according to the estimated viewership. They observed the images of the training set in advance to keep things fair. The final classification score was defined as the average of the scores of each participant.

4.3 Result

Figure 2 summarizes the precision, recall, and F_1 score of the classifiers obtained from each image feature to evaluate the classifier accuracy. These scores were calculated as the average of the classification scores of each gender. Further, we compared the recall-precision curve in Figure 3 to evaluate the retrieval accuracy. This figure illustrates the retrieval performance of males in the test set. From examining Figs. 2 and 3, it is clear that the Fisher Vector (RGB-SIFT) achieves the best classification accuracy among the compared image features, probably because the approximate object position in an image and local features such as shape and color are important for this task. Figure 2 shows that the classification accuracy is improved when the thumbnail and key-frame images are integrated. This result suggests that extracting features from more frames is a reasonable way to improve the accuracy of viewer estimation. Moreover, the classification accuracy is improved by combining image features. The combination of Fisher Vectors is the most effective, for the reason discussed above. Figure 4 shows the top 15 samples of video clips that have the highest posterior probability for each gender output by the logistic regression classifier based on Fisher Vector (RGB-SIFT). As shown, there are many video clips of games or sports among males and many of pets or singers among females. It is reasonable to suppose that our estimation model reflects the tendency of video clips that each gender often watches.
Considering the noisy nature of the data set (i.e., the overlapping images between male and female) and the intrinsic ambiguity of the task of estimating viewership from just images, the accuracy of our method is acceptable. In addition, the results of the subjective assessment show that human performance at this task was approximately 79% while our best model achieved 74%, which is also satisfactory.

5. CONCLUSION

We have presented a novel methodology to estimate the viewer demographics of video clips using image features. This approach is expected to alleviate the cold-start problem, which is currently a challenging problem hampering video recommendations on video sharing Web sites and sm-

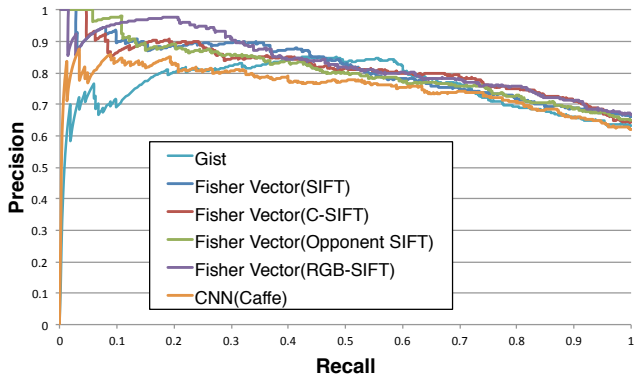

Figure 3: Comparison of recall-precision curve of male classifiers of each feature (integrated thumbnail and key-frame images).

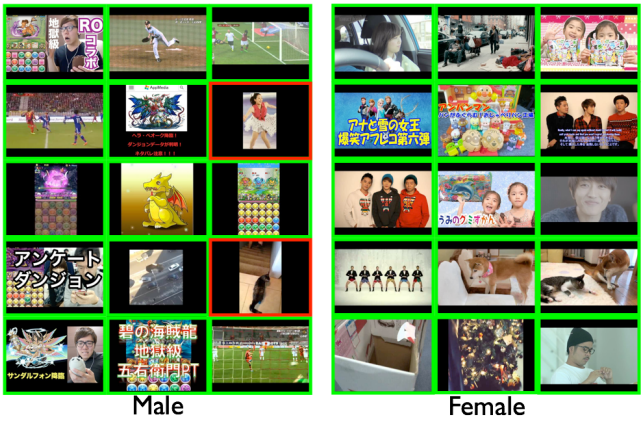

Figure 4: Top 15 samples of high posterior probability applying linear logistic regression classifier to Fisher Vector (RGB-SIFT). Green frames indicate "true" and red frames indicate "false".

art TV. The results of an experiment to classify viewer gender from the image features extracted from video clips demonstrated the effectiveness of our proposed method. Integrating the various image features of video clips and considering the key frames of video clips led to relative improvement of the classification accuracy.

6. ACKNOWLEDGEMENTS

This work is partially supported by JST, CREST.

7. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generate-on of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE TKDE*, 17(6):734–749, 2005.

[2] G. J. Burghouts and J. M. Geusebroek. Performance evaluation of local color invariants. *Computer Vision and Image Understanding*, 113(1):48–62, 2009.

[3] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categoriza- tion with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms. *ACM TOIS*, 22(1):143–177, 2004.

[6] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative ïňĄltering to weave an information tapestry. *CACM*, 35(12):61–70, 1992.

[7] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding, 2013. `http://caffe.berkeleyvision.org/`.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classiiňĄcation with deep convolutional neural networks. In *NIPS*, 2012.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *IEEE*, 1998.

[10] G. Linden, B. Smith, and J. York. Amazon.com recommend-ations: item-to-item collaborative ïňĄltering. *IEEE IC*, 7(1):76–80, 2003.

[11] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[12] D. Maltz and K. Ehrlich. the way: Active collaborative filtering. In *SIGCHI*, 1995.

[13] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[14] F. Perronnin, J. S ´ anchez, and T. Mensink. Improving the fisher kernel for large-scale image classiiňĄcation. In *ECCV*, 2010.

[15] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative ïňĄltering of netnews. In *ACM CSCW*, 1994.

[16] M. Szomszor, C. Cattuto, H. Alani, K. O ' Hara, A. Baldassarri, V. Loreto, and V. P. Servedio. Folksonomies, the semantic web, and movie recommendation. In *ESWC*, 2007.

[17] A. Ulges, M. Koch, and D. Borth. Linking visual concept detection with viewer demographics. In *ICMR*, 2012.

[18] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[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 *ACM CIVR*, 2007.

[20] YouTube. Press statistics. `http://www.youtube.com/t/press_statistics`.

[21] YouTube. Trends map. `http://www.youtube.com/trendsmap`.