Beauty is Here: Evaluating Aesthetics in Videos Using Multimodal Features and Free Training Data

Yanran Wang, Qi Dai, Rui Feng, Yu-Gang Jiang
School of Computer Science, Fudan University, Shanghai, China
{11210240061, daiqi, fengrui, ygj}@fudan.edu.cn

ABSTRACT

The aesthetics of videos can be used as a useful clue to improve user satisfaction in many applications such as search and recommendation. In this paper, we demonstrate a computational approach to automatically evaluate the aesthetics of videos, with particular emphasis on identifying beautiful scenes. Using a standard classification pipeline, we analyze the effectiveness of a comprehensive set of features, ranging from low-level visual features, mid-level semantic attributes, to style descriptors. In addition, since there is limited public training data with manual labels of video aesthetics, we explore freely available resources with a simple assumption that people tend to share more aesthetically appealing works than unappealing ones. Specifically, we use images from DPChallenge and videos from Flickr as positive training data and the Dutch documentary videos as negative data, where the latter contain mostly old materials of low visual quality. Our extensive evaluations show that combining multiple features is helpful, and very promising results can be obtained using the noisy but annotation-free training data. On the NHK Multimedia Challenge dataset, we attain a Spearman's rank correlation coefficient of 0.41.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

Keywords

Video aesthetics, beautiful scenes, free training data, multimodal features, attributes.

1. INTRODUCTION

Estimating the aesthetic quality of videos has many practical applications. In video search or recommendation, among videos with similar relevance scores, people would prefer to view aesthetically more appealing ones. It also can help broadcasting organizations to quickly select better materials out of normally very huge databases (see example clips in Figure 1).

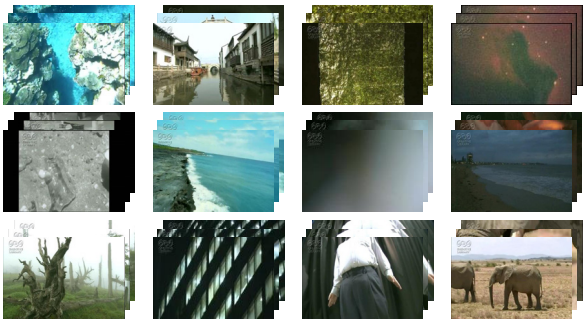

Figure 1: Example clips of the NHK broadcast video footage. This paper presents an approach to automatically identify aesthetically more appealing clips.

There have been several studies focusing on the estimation of image aesthetics [3, 5, 8, 4]. Various features and classifiers have been evaluated in this context, with a general conclusion that using multiple features is helpful. A few researchers have also exploited video aesthetics [7, 9, 1, 11]. In [7], Moorthy et al. evaluated several low-level features on a small dataset of 160 consumer videos, and observed promising results by combining seven features. Niu and Liu [9] designed a few features to distinguish professionally edited videos from amateur raw videos, assuming that professional videos have better visual quality. In [1], Chan et al. adopted cinemagraphs to determine beautiful scenes using basically motion information. Redi and Merialdo [11] assumed that beauty is highly related to interestingness. They trained models using images on Flickr, which provides a ranking criterion called interestingness.

In this paper, we present a computational approach to evaluate video aesthetics, which is different from the above works in two aspects. First, we consider a large variety of features including not only low-level features that have been widely used, but also mid-level semantic attributes and style descriptors. Second, as there is very limited training data publicly available, we propose to directly make use of resources from several domains without doing any manual annotation. This is based on our observations that images and videos on a few media-sharing websites such as DPChallenge and Flickr are mostly aesthetically appealing, while some old materials like documentary videos are much less

MM 13, Oct. 21˜25, Barcelona, Catalunya, Spain
ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2508121.

Figure 2: *Free* training data from different sources, where the DPChallenge images and Flickr videos are used as positive samples (aesthetically more appealing) and the Dutch documentary videos are used as negative samples.

pleasing. Different from [11] where Flickr images with low interestingness scores were used as negative training samples, we observe that even some of those images are beautiful and therefore do not use any Flickr images as negative data. Note that this is critical as the selection of training data affects the final results significantly. Through extensive experiments using the *free* training data, we also provide valuable insights on the selection of useful features for evaluating video aesthetics.

In the following we first introduce our training data, and then describe the evaluated features and discuss results.

2. FREE TRAINING DATA

One of the reasons that video aesthetics has not been extensively studied is that very few training samples with manual labels are publicly available. We therefore propose to construct two annotation-free training datasets by assuming that images/videos on certain websites are mostly beautiful, particularly those highly rated ones.

The first training set uses images from DPChallenge.com as positive samples. From the AVA dataset released in [8], we select 60,000 images with high ratings as indicated by DPChallenge. 50,000 frames from 1,400 Dutch documentary videos are uniformly extracted as negative examples. These videos were obtained from the past U.S. NIST TRECVID evaluations[1].

The second training set uses Flickr videos as positive samples and the 1,400 Dutch documentary videos as negative data. The Flickr videos are collected using 10 interestingness-enabled searches (keywords: animal, beach, flower, food, mountain, nature, night, ocean, street, and sunset), and the top 200 videos are downloaded from each search, leading to a set of 2,000 pseudo-positive samples.

[1]http://trecvid.nist.gov/

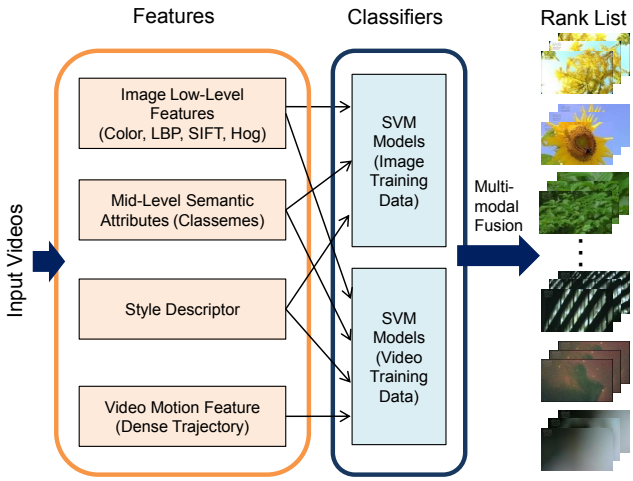

Figure 3: Overview of our system framework for evaluating the aesthetic quality of videos.

Figure 2 gives a few examples of our training data. As can be seen, the DPChallenge and Flickr data are mostly beautiful, while the documentary videos are much less appealing aesthetically.

3. THE COMPUTATIONAL APPROACH

Figure 3 shows our framework, where we adopt SVM classifiers to predict aesthetic quality due to their overwhelming performances in many applications. A large number of features are evaluated. We briefly describe each of them below.

Color Histogram in HSV space is the first feature to be tested. We compute a histogram on each sampled frame (from every 2 second window) and use the averaged histogram to represent a video.

LBP (Local Binary Pattern [10]) is a popular texture feature that describes the local pattern of one pixel to its neighboring pixels. A standard setting of 8 neighbors are used, leading to a representation of 256 dimensions.

SIFT (Scale Invariant Feature Transform [6]) is a very popular local descriptor. We follow the traditional process described in [6], where salient local patches are detected by DoG (Differences of Gaussian). We adopt the bag-of-words representation using a spatial pyramid with a codebook of 500 words.

HOG (Histogram of Oriented Gradients [2]) feature has been popular in several applications such as human detection. We compute HOG descriptors over densely sampled patches, and convert the descriptors to a bag-of-words representation with a codebook of 4,000 words.

Dense Trajectory feature [13]) is a state-of-art motion-based representation. Densely sampled local patches are tracked over time and several descriptors are used to describe the local volume around each trajectory. These descriptors are converted into bag-of-words representations. We use the source codes provided by the authors of [13].

Classemes [12] is a mid-level attribute feature, where each dimension is the prediction score of a semantic class. 2,659 semantic concept classifiers are used, leading to a representation of 2,659 dimensions. The classifiers were trained using external data on the Web [12]. Classemes is one of

the most popular attribute features, and it is interesting to study whether such mid-level representations are useful for estimating aesthetics.

Style Descriptor [8] is formed by concatenating 14 photographic style predicting scores, including complementary colors, image grain, motion blur, and vanishing point, etc. These have been proved effective for evaluating the aesthetics of images. We use SIFT, LBP and Color Histogram as features to train 14 models, using the training data provided by [8].

Among the above features, the first five are low-level visual features. The Classemes is a mid-level semantic representation as mentioned earlier, and the Style Descriptor was specially designed for predicting visual aesthetics. Note that audio features are not considered here since the targeted NHK test videos focus more on visual scenes and most of them do not have audio soundtracks. However, in a more general setup of evaluating video aesthetics we conjecture that audio features are also useful.

We use χ^2 kernel for all the features except that Classemes uses RBF kernel since it has negative feature values. For the combination (fusion) of multiple features, we adopt kernel-level fusion with equal fusion weights, where kernels computed from each of the features are averaged to form a fused kernel. Note that using adaptive fusion weights may lead to (normally marginally) better performance.

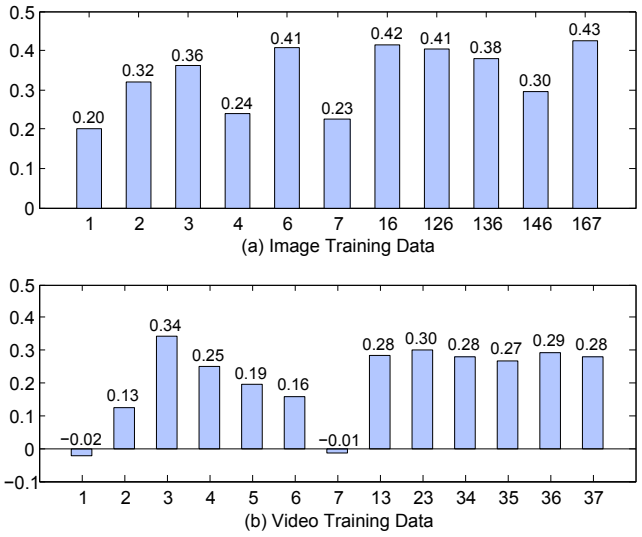

Figure 4: Evaluation results on 500 NHK clips using our annotations, measured by Spearman's rank correlation coefficients. We report results of both individual features and their fusion (1. Color Histogram; 2. LBP; 3. SIFT; 4. HOG; 5. Dense Trajectory; 6. Classemes; 7. Style). See texts for more explanations on the feature fusion setups.

4. EXPERIMENTS

4.1 Test Data and Evaluation Measure

The test dataset consists of 1,000 clips of broadcast video footage provided by NHK, with an average duration of around 1 minute. These clips cover a wide range of contents such as landscapes and creatures. With models trained by both our image- and video-based training sets, rank lists of the test clips can be generated according to aesthetic quality. Spearman's rank correlation coefficient is used as the evaluation measure, which is a value between -1 and 1, indicating the degree of correlation between the ground-truth ranking and the predicted ranking.

Since the official ground-truth ranking of the NHK test dataset is not publicly available for self evaluation[2], we manually annotated half of the test clips to assign an integer score between 0 and 10 to each clip. This allows us to perform more evaluations, so that it is possible to provide more detailed result analysis and insights in this paper. As will be discussed later, we find that the expected best result submitted to NHK is not from the actual optimal feature choices.

4.2 Results and Analysis

Let us now evaluate the results of all the features and their fusion. Results on the 500 NHK clips evaluated based on our annotations are summarized in Figure 4. Since there can be many feature combinations, we adopt a simple strategy to reduce the number of evaluated results: Starting from the best individual feature, we incrementally add in other features and a feature is dropped immediately if it does not help in a fusion experiment.

We first discuss results using the image-based training data. As shown in Figure 4(a), among all the low-level visual features, SIFT is the best performer, achieving a Spearman's coefficient of 0.36. Color has the worst performance, indicating that color information is less useful in predicting visual aesthetics. Overall, features focusing more on texture structures like the SIFT and LBP tend to be more effective. The mid-level semantic attribute feature Classemes offers the best single-feature result (0.41). This indicates that the semantic contents of images play an important role in aesthetic quality evaluation, which is an interesting observation. The style descriptor, which was observed to be very useful for evaluating image aesthetics, is not as good as expected. However, it may be used in combination with other features for improved performance, as discussed in the following feature fusion experiments.

The fusion of multiple features can further improve results in the image-based training experiment. As shown on the right side of Figure 4(a), the best result is attained by fusing Color Histogram, the Classemes, and the Style Descriptor (Spearman's coefficient 0.43). Adding Color over Classemes (the bar marked by "16" in the figure) also improves results marginally. Although LBP, HOG and SIFT perform well individually, they are not complementary when combined with Color or Classemes.

Next we discuss results obtained by the video-based training data. As shown in Figure 4(b), overall the results are much worse than that from the image-based training. This is not surprising as the NHK video clips focus more on scenes, for which it is intuitively sufficient to purely use images for model training. We believe that video based training is necessary and effective if the targeted application domain is broader. Among the individual features, SIFT is the best, with a Spearman's coefficient of 0.34. The Dense Trajectory motion feature, which is very powerful in human action recognition, performs poorly on this task. This is due to the

[2]Only a maximum of five result runs can be submitted to NHK for evaluation.

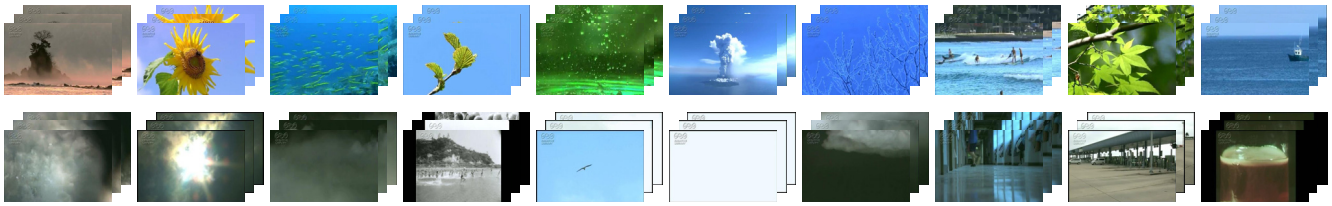

Figure 5: The top and bottom rows show the most and the least beautiful 10 videos respectively, identified by our best submitted run (image-based training data with Color Histogram and Classemes features).

Table 1: Evaluation results of our five submitted runs using the NHK ground-truth. "Image+Video" means the late fusion (average prediction scores) of the two runs using the image- and video-based training data respectively.

Training Data	Image	Video	Image+Video
Color+Classemes	**0.41**	0.03	0.39
Color+Classemes+SIFT	0.37	0.19	—

fact that very few foreground object motions exist in the NHK footage, and the aesthetically quality of scenes is intuitively not highly related to motion. In the feature fusion experiments using the video-based training data, we do not observe any performance gain, probably because the results of the other features are much worse than that of the SIFT.

Finally, we report the performances of our five submitted result runs, which were measured over the entire test set using NHK's official ground-truth labels. As shown in Table 1, the best result is also from the image-based training data, using Color Histogram and Classemes. This is similar to that reported in Figure 4(a) using half of the test data and our own labels. According to our analysis earlier, we expect that slightly better performance may be obtained on this entire test set if the Style Descriptor is further added. Figure 5 shows the top and bottom 10 clips identified by our best submitted run.

5. CONCLUSIONS

We have presented an approach for evaluating aesthetics in videos. Since there is limited labeled training data available, we constructed two annotation-free training datasets by making use of image and video data from various sources (e.g., Flickr and DPChallenge). Our results suggest that the image-based training is more suitable for the scenario of the NHK Challenge, where the focus is mainly on identifying beautiful scenes. However, for a more broader task of evaluating video aesthetics, this conclusion may not hold.

We also evaluated several off-the-shelf features, including not only low-level visual features, but also mid-level semantic attribute features and a photographic style descriptor. Results show that all these three types of features are useful, and among them the semantic attribute feature gives superior performance.

6. ACKNOWLEDGEMENT

This work was supported in part by two grants from the National Natural Science Foundation of China (#61201387 and #61228205), two grants from the Science and Technology Commission of Shanghai Municipality (#13PJ1400400 and #12511501602), a National 863 Program (#2011AA010604), and a New Teachers' Fund for Doctoral Stations, Ministry of Education (#20120071120026), China.

7. REFERENCES

[1] Y.-T. Chan, H.-C. Hsu, P.-Y. Li, and M.-C. Yeh. Automatic cinemagraphs for ranking beautiful scenes. In *Proceedings of ACM International Conference on Multimedia*, 2012.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of European Conference on Computer Vision*, 2006.

[4] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[5] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[7] A. K. Moorthy, P. Obrador, and N. Oliver. Towards computational models of visual aesthetic appeal of consumer videos. In *Proceedings of European Conference on Computer Vision*, 2010.

[8] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[9] Y. Niu and F. Liu. What makes a professional video? a computational aesthetics approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(7):1037–1049, 2012.

[10] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[11] M. Redi and B. Merialdo. Where is the interestingness? retrieving appealing video scenes by learning flickr-based graded judgments. In *Proceedings of ACM International Conference on Multimedia*, 2012.

[12] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *Proceedings of European Conference on Computer Vision*. 2010.

[13] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.