Estimating Beauty Ratings of Videos using Supervoxels

Gökhan Yildirim
School of Computer and Communication Sciences
EPFL, Switzerland
gokhan.yildirim@epfl.ch

Appu Shaji
School of Computer and Communication Sciences
EPFL, Switzerland
appu.shaji@epfl.ch

Sabine Süsstrunk
School of Computer and Communication Sciences
EPFL, Switzerland
sabine.susstrunk@epfl.ch

ABSTRACT

The major low-level perceptual components that influence the beauty ratings of video are color, contrast, and motion. To estimate the beauty ratings of the NHK dataset, we propose to extract these features based on *supervoxels*, which are a group of pixels that share similar color and spatial information through the temporal domain. Recent beauty methods use frame-level processing for visual features and disregard the spatio-temporal aspect of beauty. In this paper, we explicitly model this property by introducing supervoxel-based visual and motion features.

In order to create a beauty estimator, we first identify 60 videos (either beautiful or not beautiful) in the NHK dataset. We then train a neural network regressor using the supervoxel-based features and binary beauty ratings. We rate the 1000 videos in the NHK dataset and rank them according to their ratings. When comparing our rankings with the actual rankings of the NHK dataset, we obtain a Spearman correlation coefficient of 0.42.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*; I.4.6 [**Image Processing and Computer Vision**]: Segmentation—*Region growing, partitioning*

Keywords

Video beauty; supervoxel; video ranking

1. INTRODUCTION

Rating any visual stimulus as "beautiful" is highly subjective and very personal. What some person might find beautiful another might be indifferent to or even find ugly. Yet, there are many images and videos that most of us find beautiful, which follow some perceptual arrangement that is pleasing to the majority of us. Thus, estimating beauty ratings does not always require high-level perceptual cues, such as objects, people, and action recognition, but can often be modeled by low-level features, such as color, contrast, and motion.

In this paper, we attempt to estimate the beauty ratings of the videos in the Japan Broadcasting Corporation (NHK) dataset[1]. In the dataset, there are 1000 videos with an average duration of one minute and a resolution of 640×360 pixels. Most of them are very simple videos, are composed of only a few shots and contain little action. A video *shot* is a stack of frames where the video recording is not interrupted. As each shot can have its own content, the less shots there are, the simpler it is to understand a video. The videos in the NHK dataset have three shots on average, are short and thus much simpler compared to a regular movie or even a YouTube video. We thus believe that a low-level approach will be sufficient to achieve a good rating estimation.

We select our features based on recent research in image and video beauty, which has shown that low-level visual features, such as composition, colorfulness, lightness, and contrast [3, 8, 9, 10], are successful in estimating image aesthetics. Video beauty is also related to camera motion stability [8, 9, 10]. While the low-level features are calculated for each video frame, motion stability is measured by matching image patches or keypoint descriptors from one frame to the next, which might fail on uniform areas.

The main drawback of the above methods are that they calculate the low-level visual features and the motion features with frame-level processing. We claim that there is a combined, spatio-temporal dimension to beauty. Thus, in our method, we jointly extract low-level visual and motion features from **supervoxels**. A supervoxel, which is a 3D extension of a superpixel, is a group of pixels sharing similar color and spatial information along the temporal axis. Supervoxels allow us to summarize and simplify the content of a video through its color/texture and motion components. They have been used in biological applications, such as motion estimation on microscopy data [2] and mitochondria segmentation [7], as well as in video retrieval [6]. We use Achanta et al.'s method [1] to compute the supervoxels of every video shot. In comparison to the state-of-the-art methods, our visual features are able to represent the video on a shot level. In addition, our motion features are based on the velocity of the supervoxels throughout a video shot and the initial and final positions of the supervoxels. They successfully measure the motion profile and the spatial composition of both uniform and textured image regions.

[1]http://acmmm13.org/submissions/call-for-multimedia-grand-challenge-solutions/task-where-is-beauty/

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2508125.

In our rating algorithm, supervoxel features are used to train a neural network-based regressor, which estimates the "beauty" of a given video. In order to establish a ground truth, we select 60 videos from the NHK dataset, which are either beautiful or not in a collective sense. Our neural network is regressed over these binary ratings so that it can learn the distinctive properties of video beauty. We test our algorithm on 1000 videos in the NHK dataset and obtain a Spearman correlation coefficient of 0.42 with respect to the actual NHK challenge ground truth.

2. SUPERVOXELS AND TRAJECTORIES

We summarize the visual and motion features of a video with supervoxels. As different parts of a video can have different contents or viewpoints, we need to detect the individual shots that are visually consistent for proper supervoxel extraction.

2.1 Shot Detection

A video, $\mathbf{V}$, contains several shots. We can extract individual shots by detecting the discontinuities in optical flow in a video. In our paper, we use Horn-Shunck's optical flow method [5]. For shot detection, we only need to identify the sudden changes in the magnitude of the optical flow.

$$\mathbf{d}(t) = \frac{1}{Z}\sum_{x}\sum_{y}\sqrt{\mathbf{O_u}(t)^2 + \mathbf{O_v}(t)^2} \qquad (1)$$
$$\mu = \alpha \times \text{median}(\mathbf{d})$$

Here, $\mathbf{O_u}(t)$ and $\mathbf{O_v}(t)$ are the optical flow vector components in the horizontal and the vertical directions at frame t, respectively. $\mathbf{d}(t)$ is the average optical flow for frame t, Z is the total number of pixels in one video frame, x and y are the spatial coordinates, μ is the threshold for shot detection, and α is a constant to modify the threshold. A shot is detected if the optical flow is greater than this threshold. The median provides an average motion value that is resistant to extreme values. We found α equal to 10 provides the best shot detection for the NHK dataset.

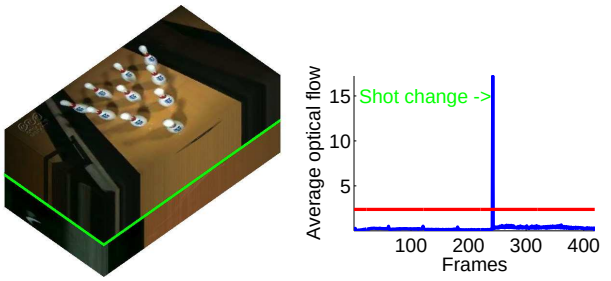
An example of shot detection on a bowling video (0113.mp4 in the NHK dataset) is given in Figure 1. As we can see from Figure 1(a) and 1(b), the video includes two shots illustrated with a sharp optical flow jump at frame #242.

After shot detection, we have visually consistent stacks of video frames, such as Figure 1(c) and 1(d), that have the visual consistency we require to extract supervoxels and their properties.

2.2 Supervoxel Extraction

Oversegmenting a single object into multiple supervoxels is acceptable, provided that the segment size is large enough to benefit from pixel grouping, i.e. the resultant supervoxels have similar visual and motion features. However, undersegmentation might cause mixtures between irrelevant objects, which is not desirable. Thus, we select a spatial size for supervoxels by assuming that the minimum size for an object of interest is equal to 20 pixels (recall from Section 1 that videos have a size of 640×360 pixels). We choose the longest possible temporal size to capture the motion of an object throughout the video shot.

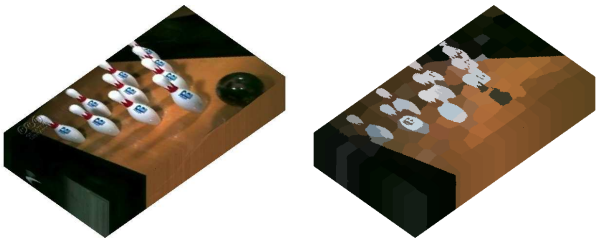
We can see an example of supervoxel extraction in Figure 2. We represent each supervoxel with their average color (averaged in L*a*b* space) in Figure 2(b).



(a) Original video (b) Average optical flow ($\mathbf{d}$, in blue), shot detection threshold (μ, in red)

(c) First shot (d) Second shot

Figure 1: Shot detection on a video with two shots. Shot change is depicted with a green line in (a)



(a) Second shot (b) Supervoxels of second shot

Figure 2: Supervoxel extraction.

2.3 Feature Extraction

In order to *learn* the relationship between video features and beauty, we use the visual and motion features of the supervoxels and their trajectories.

2.3.1 *Visual Features*

We employ visual features similar to the ones in the state-of-the-art methods that judge video beauty [3, 8, 9, 10], with one important difference: we compute them from the supervoxels. The explanation of the ten visual features ($f_1 - f_{10}$) we use are given in Table 1.

2.3.2 *Motion Features*

In order to express the motion inside a supervoxel, we calculate the supervoxel trajectories through a video shot by computing the center of mass of a supervoxel on each frame. Supervoxel $\mathbf{X}_{jk}$ represents the k^{th} supervoxel of j^{th} shot.

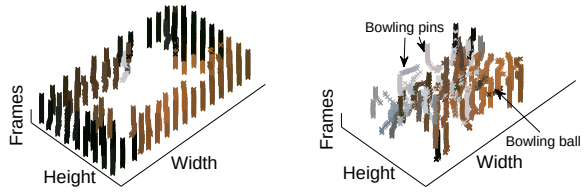
$$\mathbf{r}^x_{jk}(t) = \frac{1}{|x \in \mathbf{X}_{jk}(t)|}\sum_{x \in \mathbf{X}_{jk}(t)} x$$
$$\mathbf{r}^y_{jk}(t) = \frac{1}{|y \in \mathbf{X}_{jk}(t)|}\sum_{y \in \mathbf{X}_{jk}(t)} y \qquad (2)$$

Table 1: Visual features computed over supervoxels.

Feature	Comment
Brightness	Average value of the Y channel of the YCbCr color space
Contrast	Average brightness difference between a supervoxel and the rest of the supervoxels
Saturation	Average value of the S channel of the HSV color space
Saturation Contrast	Average saturation difference between a supervoxel and the rest of the supervoxels
Colorfulness	Colorfulness measure in [4] between a supervoxel and the rest of the supervoxels
Average Color	Average value of L*, a* and b* channels of the CIELab color space (3 features)
Saliency	Average color difference between a supervoxel and the rest of the supervoxels (over L*a*b*)
Normalized Voxel Size	The number of pixels in a supervoxel normalized by the size of the shot cube

Here, $\mathbf{r}^x_{jk}(t)$ and $\mathbf{r}^y_{jk}(t)$ are the (x, y) coordinates of the center of mass of the supervoxel $\mathbf{X}_{jk}$ at frame t. The computed coordinates form the trajectories. The trajectories of the supervoxels in Figure 2(b) are illustrated in Figure 3(a) and 3(b). For didactic purposes, we illustrate stationary and moving trajectories separately.

We can observe from Figure 3(a), for this video, the supervoxels that are close to the frame border have negligible motion. The progression of the bowling pins and the bowling ball are illustrated in Figure 3(b).



(a) Stationary trajectories (b) Moving trajectories

Figure 3: Supervoxel trajectories.

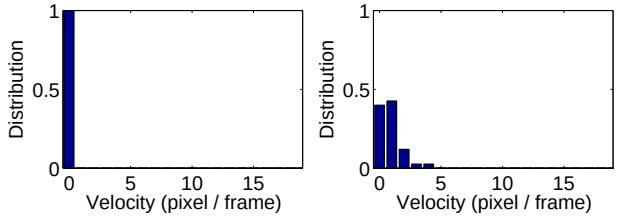
To compute our motion features, we calculate the velocity of the supervoxel trajectories at each frame as follows:

$$\begin{aligned} \mathbf{l}^x_{jk}(t) &= \mathbf{r}^x_{jk}(t+1) - \mathbf{r}^x_{jk}(t) \\ \mathbf{l}^y_{jk}(t) &= \mathbf{r}^y_{jk}(t+1) - \mathbf{r}^y_{jk}(t) \\ \mathbf{l}_{jk}(t) &= \sqrt{\left[\mathbf{l}^x_{jk}(t)\right]^2 + \left[\mathbf{l}^y_{jk}(t)\right]^2} \end{aligned} \tag{3}$$

Here, $\mathbf{l}^x_{jk}(t)$ and $\mathbf{l}^y_{jk}(t)$ are horizontal and vertical velocity components of the trajectory at frame t, respectively, $\mathbf{l}_{jk}(t)$ is the magnitude of the trajectory velocity for supervoxel $\mathbf{X}_{jk}$ at frame t. In order to profile the motion of a supervoxel, we create a 20-bin histogram using $\mathbf{l}_{jk}$. The maximum velocity is equal to 19 pixels per frame, because the expected spatial size for supervoxels is 20 pixels (see Section 2.2). After normalizing with $||\mathbf{l}_{jk}||_1$, a trajectory gives us 20 features in total $(f_{11} - f_{30})$.

The distribution represents how smooth a supervoxel is advancing inside a shot. For example, a concentrated histogram as illustrated in Figure 4(a) corresponds to a stationary supervoxel. Conversely, a distributed histogram as shown in Figure 4(b) corresponds to a moving supervoxel, in this case a bowling pin (see Figure 3(b)).

In addition to a velocity histogram, we also use the normalized initial and final spatial trajectory positions (4 features in total, $f_{31} - f_{34}$) of the supervoxels as motion-based features. The normalization is performed using the dimensions of the video frame, so that the feature values vary in the interval $[-0.5, 0.5]$. These positions are related to the composition of supervoxels in the shot.



(a) Velocity histogram of a stationary supervoxel (b) Velocity histogram of a bowling pin

Figure 4: Velocity histograms of two supervoxel trajectories in the bowling video.

3. VIDEO RANKING RESULTS

It is possible to define heuristic rules to estimate the beauty of a video. For example, people might consider colorful videos with high contrast and smooth movements as beautiful. However, we choose to discover those rules, if they exist, by regressing our features over video ratings.

We select 60 videos from the NHK dataset, half of which can be considered as collectively "beautiful" (with rating = 1) and the rest as "not beautiful" (with rating = 0). Instead of having a continuous rating, we create a binary ground truth to properly learn the separation between good and bad videos. We then train a neural network-based regressor (with one hidden layer of 10 neurons) for rating estimation. As video beauty is a subjective concept, parametrizing the joint distribution of the input features and the beauty is prone to errors. Thus, we choose a discriminative regression model instead of a generative model. The input of the network is the supervoxel feature vector $(f_1 - f_{34})$ of all supervoxels in all of the training videos and the ground truth is the binary beauty ratings.

In order to estimate the rating of a video, we extract its supervoxel features and pass them through the neural network. The rating of a video is calculated by averaging the ratings of its supervoxels as shown in (4).

$$R_{\mathbf{V}} = \frac{1}{N} \sum_{\forall \mathbf{F_{kj}} \in \mathbf{V}} \psi(\mathbf{F_{kj}}) \tag{4}$$

Here, $R_{\mathbf{V}}$ is the final rating of the video $\mathbf{V}$, N is the number of supervoxels in $\mathbf{V}$, $\psi(.)$ is the neural network function, and $\mathbf{F}_{jk}$ is the feature vector of supervoxel $\mathbf{X}_{jk}$.

We rank the videos in the NHK dataset with respect to their estimated final ratings. The correlation coefficients of

our ranking and the user study-based ranking obtained by the NHK challenge is given in Table 2.

Table 2: Correlation between estimated and ground truth rankings.

Feature Type	Correlation Type	Value
Motion only	Spearman	0.052
Visual only	Spearman	0.387
Both	Spearman	**0.424**
Motion only	Kendall	0.036
Visual only	Kendall	0.264
Both	Kendall	**0.290**

As we can see from Table 2, the main contribution is provided by visual features. Another observation is that the sum of the individual performances of visual and motion features is very similar to their combined performances. This shows that our supervoxel-based features are successful to decouple the visual and motion properties of a video.

In Figure 5, we show example frames from the top- and bottom-ranked videos in the NHK dataset. These ranking results are obtained using both visual and motion-based supervoxel features. We can observe that the top-ranked videos are more colorful than the bottom-ranked ones. We refer the reader to the NHK video website for judging motion related performance.

4. CONCLUSION

We show that low-level supervoxel-based features are successful in estimating the beauty rating of short and simple videos in the NHK dataset. The supervoxels accurately represent and separate the coupling between the visual and motion aspects in a video. Because, different than the state-of-the-art techniques, our method regards the spatio-temporal aspect of the beauty by involving shot-level processing.

We build a neural network regressor to learn the relationship between the perceptual components of a video and its beauty. We then use this regressor to automatically rate and rank the videos in the NHK dataset and obtain a moderate amount of correlation.

We collect binary ground truth ratings for only 60 videos of the NHK dataset. A larger training set will significantly enhance the ranking performance. Moreover, the potential of additional supervoxel-based features can be investigated not only in the context of video beauty but also, in general, video processing. For practical purposes, we might need a scalable learning method, such as a stochastic neural network, in order to perform online learning. In addition, new features can be enabled or disabled by activating or deactivating corresponding neurons.

5. ACKNOWLEDGEMENTS

This work was supported by the Swiss National Science Foundation under grant number 200021_143406 / 1.

6. REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on PAMI*, 34(11):2274–2282, 2012.

[2] F. Amat, E. W. Myers, and P. J. Keller. Fast and robust optical flow for time-lapse microscopy using super-voxels. *Bioinformatics*, 29(3):373–380, 2013.

[3] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proc. of the International Conference on Multimedia*, pages 271–280, 2010.

[4] D. Hasler and S. Süsstrunk. Measuring colorfulness in natural images. In *Proc. SPIE Human Vision and Electronic Imaging*, volume 5007, pages 87–95, 2003.

[5] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[6] R. Hu, S. James, T. Wang, and J. Collomosse. Markov random fields for sketch based video retrieval. In *Proc. of ACM Multimedia Retrieval*, pages 279–286, 2013.

[7] A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE Transactions on Medical Imaging*, 31(2):474–486, 2012.

[8] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *Proc. of ECCV*, volume 3, pages 386–399, 2008.

[9] A. Moorthy, P. Obrador, and N. Oliver. Towards computational models of the visual aesthetic appeal of consumer videos. In *Proc. of ECCV*, volume 6315, pages 1–14, 2010.

[10] 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.

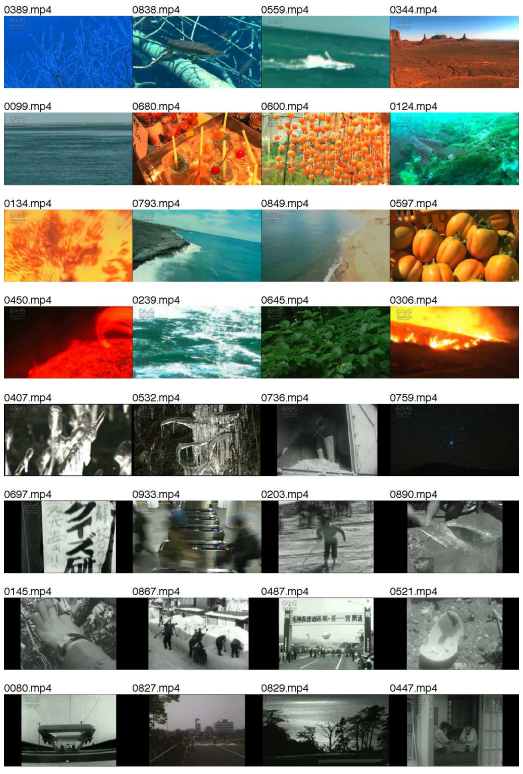


Figure 5: Frames from the videos in the NHK dataset that are top-ranked (16 images above) and bottom-ranked (16 images below) by our algorithm.