AVANT, Vol. VIII, Special Issue
ISBN: 978-83-944622-04-6 avant.edu.pl/en
DOI: 10.26913/80s02017.0111.0007

From Computational Aesthetic Prediction for Images to Films and Online Videos

François Lemarchand

Department of Computing, Electronics and Mathematics,
Plymouth University, UK
francois.lemarchand@cognovo.eu

Received 13 May 2017; accepted 26 September 2017; published 21 November 2017.

Abstract

In the last decade, creating and sharing videos online has become a mainstream movement and has led to some creators generating one personal video per day, also called daily vlogging. Although robust solutions exist to suggest photographs based on aesthetic criteria, the rising number of online videos created and watched means that such recommendation systems are required more than ever for videos. The main purpose of this paper is to transfer the skill of computational aesthetic classification of photographs to videos while developing new ways of investigating video creation. Using a dataset of photographs rated on aesthetic criteria by an internet community and recently developed feature extraction algorithms, the computational aesthetic classifier is capable of state-of-the-art photograph classification depending on aesthetic preferences learnt from people's ratings. On a test set of YouTube videos, the same system then displays satisfying aesthetic classification results that consist of an attempt to match the provided human aesthetic quality ratings. Achieving a transfer of skill from photograph to video classification, the computational classifier is used to analyze the evolution of aesthetics in feature films; this highlighted the aesthetic classifier's visual preferences and caused some interesting patterns to emerge that were related to filmmakers' decisions. Aesthetic classification makes it possible to observe the evolution of aesthetics over the careers of daily content creators thanks to their abundant and regular online video content. It can aid the investigation into the impact of aesthetics on the popularity of online videos using the available metadata about the internet audience's appreciation. This can also provide a new tool for video content creators to assess their work and assist them in the production of content of higher aesthetic quality.

Keyword: computational aesthetics; skill transferability; video classification; visual preferences.

Introduction

The popularization of high-speed internet has led to an increase in visual content consumption. While photographs were introduced in the early years of the Internet, high definition videos are part of a trend still subject to fast growth. It has become increasingly complex to select a relevant video among the hundreds of hours of videos uploaded to YouTube every minute. Even though videos are already suggested through textual tags or speech analysis, little has been done to offer aesthetic-based filters for video suggestions due to the limitations of existing datasets. Despite recent efforts to build large datasets of videos, such as YouTube 8M, no video dataset for aesthetic video classification achieves a similar quantity of items as existing datasets for computational aesthetic classification of photographs (Abu-El-Haija et al., 2016). The largest aesthetic dataset of videos known to date is the recently published dataset by Tzelepis et al., which is composed of 700 short videos collected on YouTube and matched with aesthetic ratings (Tzelepis, Mavridaki, Mezaris, & Patras, 2016).

While previous works have focused on computational aesthetic classification of short videos (Niu & Liu, 2012; Tzelepis et al., 2016; Yang, Yeh, & Chen, 2011), "The Colors of Motions" by Charlie Clark illustrated the change in dominant colors over several feature films (Clark, 2014). Moreover, Jason Schulman's "Photographs of Films" offers novel ways of looking into the aesthetics of films as they overlap all frames from a film to obtain a single merged image (Shulman, 2017). The previous computational system was developed and trained to classify photographs depending on their aesthetics (Lemarchand, 2017), whereas this paper introduces the cross-media capabilities of this aesthetic classifier on the video dataset published by Tzelepis et al. To complete tests on Tzelepis et al.'s dataset, the classifier is used on films to observe special aesthetic patterns over time and points out the potential weaknesses and strengths of such classifier on both photographs and videos. At the end of the paper, the classifier is tested using YouTube videos as a resource, particularly videos by Casey Neistat, a filmmaker and daily vlogger. In the form of a case study, potential links between aesthetic prediction and video quality are investigated by looking at the evolution of aesthetics across years of work.

Training of the Aesthetic Classification System

In order to compare divergences in the behavior and performance of aesthetic classification systems in photographs and videos, an artificial intelligent system previously designed to classify images based on aesthetics that achieves state-of-the-art results on different datasets was selected (Lemarchand, 2017). The aesthetic classifier was first trained on a large scale photograph dataset called AVA (Murray, Marchesotti, & Perronnin, 2012). The AVA dataset is superior for learning aesthetic preferences as it provides one rating per image, compared to only one rating per video in Tzelepis et al.'s dataset, which is, on a visual level, a collection of many still

images. In this paper, visual information is defined as aesthetically pleasant if it has the potential to induce a positive response among the average observer, which is represented in existing datasets by the rating community's self-reports. The AVA dataset is also superior to Tzelepis et al.'s dataset in terms of representation of human visual preferences, as every image has received at least a hundred ratings according to aesthetic criteria. Training for aesthetic classification per image (and therefore per frame) allows a deeper understanding of videos as sequences and scenes can be isolated and analyzed. In fact, aesthetic classification systems are usually trained and tested with still images, mainly due to the complexity of collecting aesthetic ratings for video streams.

Previous works have proven to be effective aesthetic classification solutions with, for example, algorithms scoring images based on photography rules (rule of thirds, leading lines, etc.), or more computation-based approaches such as image descriptors and convolutional neural networks linking visual features to expected classifications (Datta, Joshi, Li, & Wang, 2006; Lu, Lin, Jin, Yang, & Wang, 2014; Marchesotti, Perronnin, Larlus, & Csurka, 2011; Romero, Machado, Carballal, & Santos, 2012). The aesthetic classifier used in this paper extracts measures of orientation distribution, curvature distribution, HSB color distribution (Hue, Saturation, Brightness), and reflectional symmetry on cardinal and diagonal axes. A deep neural network composed of 3 hidden layers is then used to learn visual preferences and obtain state-of-the-art results across several datasets such as Datta et al., CUHK and AVA (Datta et al., 2006; Murray et al., 2012; Tang, Luo, & Wang, 2013). This classifier was selected due to its cross-dataset performances and the fact that the low-level visual features extracted illustrate fundamental preferences in the human visual system. Therefore, it is suggested that low-level visual preferences can provide better cross-media performance as they tend to be less influenced than higher level preferences by cultural and personal experiences.

Applying Aesthetic Classification to Videos

In the AVA dataset, aesthetic classes are defined for each photograph by the average rating of all of the human aesthetic ratings provided with the dataset. All further predictions on new images or video frames are considered as a display of the aesthetic preferences of the rating community. In a video stream, the aesthetic classifier categorizes each frame independently as aesthetically low or aesthetically high. The high number of frames per second makes it possible to have several images embedding the same visual content from possibly different points of view. The aesthetic average over time is calculated using a large number of images. This average does not only estimate the aesthetic quality of the visual content over time, but the different points of view observed across frames make it possible to distinguish sequences containing frames with a normal aesthetic level, despite the fact that the binary classification

focuses on low or high levels. Indeed, sequences in which the distribution of the frames' classes is close to chance (50% low, 50% high) implicitly shows that the frames are close to average levels of visual aesthetics, based on the previously learnt visual preferences. This provides additional information regarding the classifier's confidence in its decision; in a binary classification task, this is a significant advantage compared to other existing classifiers. Nonetheless, biases in aesthetic classification may appear due to differences between the norms of photography and videography. Furthermore, in comparison to photographs, videos include additional semantic content due to auditory and motion information. This may mean that even self-reports may not correlate with the aesthetic classifier's predictions, as it focuses purely on visual information.

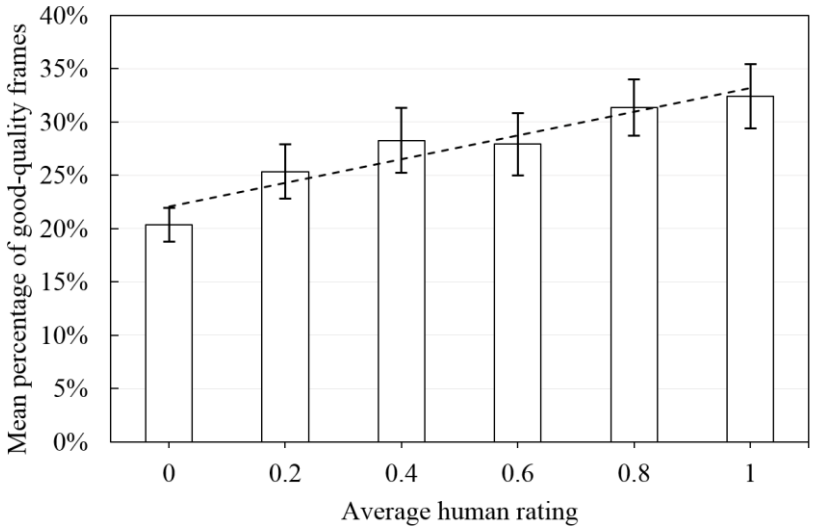

Figure 1. Mean percentage of good-quality frames detected in a video by the aesthetic classifier depending on the average human rating

The dataset of Tzelepis et al. is composed of 700 short videos downloaded from YouTube and rated in terms of aesthetics by 5 people. All frames from each video are then extracted before running the aesthetic classifier previously trained on the AVA dataset on each frame of each video, thus calculating the percentage of good-quality frames per video. The percentage of good-quality frames is then used as an input to a multilayer perceptron that is trained using the videos' aesthetic ratings. The percentage of good-quality frames detected in videos by the aesthetic classifier is shown to be strongly related to the human ratings of aesthetics, as displayed in Figure 1. The linear regression model presents a significant increasing slope of 0.12 ($t(698) = 5.11$, $p < .001$), meaning that a greater number of good-quality frames were detected in the best-rated videos. The method allowed comparison with the existing video aesthetic

classifier designed by Tzelepis et al. For each result presented in Table 1, the average precision out of 1,000 repetitions is calculated and for each repetition the training set (300 videos) and the testing set (400 videos) are randomized. Despite being far from the original results achieved by Tzelepis' solution, the proposed solution achieves results significantly above chance. This transfer of skill from photograph to video classification demonstrates the cross-media capabilities of the aesthetic classifier and that the percentage of good-quality frames in a video can be a relatively efficient predictor of aesthetic pleasantness. This recognition task also makes it possible to test the current binary classifier against another previously designed classifier. While the first version decides between the two aesthetic classes (low and high aesthetics), the second version estimates aesthetics on a scale from 1 to 10, as given in the AVA dataset. However, the percentage of good-quality frames estimated by the second version did not show any relationship with the human ratings, and the classification of videos depending on aesthetics is only slightly above chance, implying that the binary classification version is much more reliable, even though it has a limited scaling range.

Exploiting the Aesthetic Classifier on Films

Table 1: Precision for the top-*n* (5,10,15,20) percent most aesthetic videos with average accuracy and precision (in %).

	(Tzelepis et al., 2016)	Proposed classifier
Top 5%	82.00	64.54
Top 10%	82.00	64.74
Top 15%	83.33	64.89
Top 20%	81.50	64.80
ACC	68.14	54.60
AP	69.97	56.38

Considering that most recent films last between 90 and 180 minutes at a rate of 24 frames per second in average, extracting visual features of all frames involves a substantial amount of processing. The visual quality of the processed films is 720p (1280 × 720 px), which offers a good compromise between a reasonable image size for feature extraction and processing speed. Only one frame per second is extracted in order to limit the number of frames to process. Scoring feature films by different directors using the aesthetic classifier, Wes Anderson's focus on symmetry has resulted in films that achieve good-quality frame ratings, such as 56.16% for *The Grand Budapest Hotel*, 22.0% for *Moonrise Kingdom*, 20.60% for *Fantastic Mr Fox* and 58.83% for *The Royal Tenenbaums*. On another hand, Stanley Kubrick, who is known

for his shots with sophisticated depth of field effects, directed *Full Metal* Jacket, which presents 12.10% of good-quality frames, *A Clockwork Orange* with 17.75%, *The Shining* with 14.12% and *Space Odyssey* with 16.21%. Although percentages of good-quality frames possibly indicate some of the aesthetic classifier's visual preferences, reducing a whole feature film to a single score is highly limiting analyses. Due to having too few films to obtain significant statistics, further investigations on the aesthetic classifier's preferences between the two film directors is difficult, particularly when considering the potential influence of film type or year of release.

Initially, preliminary tests were performed for a pilot psychology experiment investigating potential relationships between physiological data and aesthetic pleasantness of video excerpts used as stimuli. The aesthetic change over time, which is observed by calculating the moving mean of frames' predicted aesthetic classes, shows significant fluctuations in the level of aesthetics depending on film sequences. Following the pilot tests, entire films were processed (for example, by Quentin Tarantino), and some interesting patterns are observed. One film, *The Hateful Eight* (2015), particularly stands out because the aesthetic prediction averages zero in the second part of the film, despite a reasonable number of good-quality frames in the first part. This drop seems to correlate with the switch from outdoor scenes to indoor scenes in the film; this was confirmed by the strong positive correlation (r = .82, p < .001) over the entire film between aesthetic prediction and the feature representing the distribution of pixels with normal brightness values (Figure 2). Considering the number of features involved in the aesthetic prediction processing, it can be seen that the aesthetic classifier trained on photos strongly dislikes darkness in the film. The classifier's results support the assumption that using photographs for training will create a bias in the learnt aesthetic preferences as, for example, high levels of darkness are more acceptable to a human eye watching films due to motion.

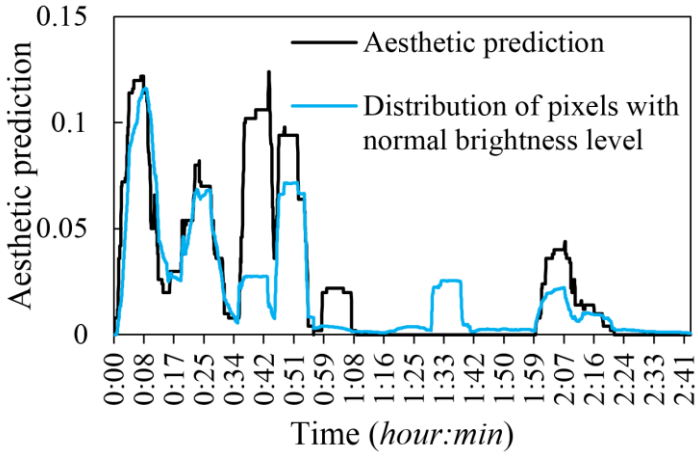

Figure 2. Plot displaying the correlation between brightness and aesthetic prediction in the film, The Hateful Eight (2015).

In another example (Figure 3a), the aesthetic prediction curve of *Django Unchained* (2012) shows a vertical symmetry centered on the middle of the film. The pattern is relevant when it is considered that Quentin Tarantino designed the film in two parts. It can be speculated that Tarantino knowingly wrote the scenario and set up camera shots to generate a symmetry between those two parts. After removing the credits, the axis of symmetry was found in order to compare the two parts using this axis as a splitting point. As shown in Fugure 3b, when mirroring the aesthetic prediction over time of the second part over the y-axis, a strong correlation ($r = .85$, $p < .001$) is found between the two parts, each of which contain a sequence of 4,000 frames. Such a strong correlation score seems to indicate the original intentions of the director. As the curve of aesthetic prediction did not appear to correlate with the different types of shot or the nature of scenes (dialogue, action, etc.), it implies that the apparent pattern must have been generated by an abnormal value in one of the features, similarly to the outdoor–indoor scene observation made for *The Hateful Eight* film. While no definite factor was identified as the origin, such a pattern may have been influenced by easily manipulable features such as symmetry and colors during filming, or altered in postproduction by Robert Richardson, the film's cinematographer.

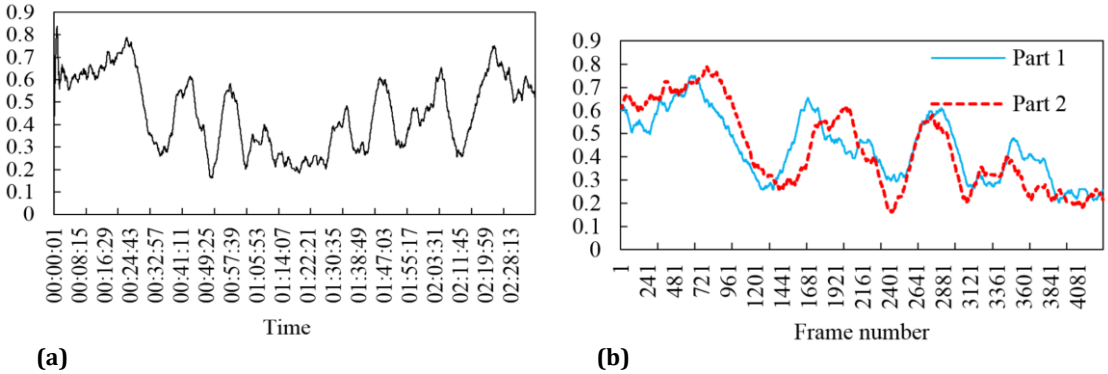

Figure 3. Aesthetic prediction over time, averaged using a sliding window of 500 frames: (a) aesthetic prediction of the film, *Django Unchained* (2012) directed by Quentin Tarantino; (b) comparison of the aesthetic prediction over time (implied by frame number) of the first half of Django Unchained and the mirrored aesthetic prediction of the second part on the y-axis.

Relying on IMDb.com's ratings, films of varying quality were processed such as *Birdemic*, *Batman* and *Robin* or *Kill Bill*. Not all analyzed films displayed interesting curves of aesthetic prediction, but all the curves representing films appeared to be influenced by the different sequences present in the films. Considering that the aesthetic classifier has been trained on photographs, the influence of film sequences on the aesthetic curve may be due to dialogue scenes complying more with photography rules than action scenes. The examples presented expose two advantages of such experiments. First, it makes it possible to test the visual preferences of the trained

classifier and evaluate the extent of the cross-media capabilities of a photograph-trained aesthetic classifier. Second, it allows the emerging patterns and filming styles generated by film directors to be analyzed and investigated.

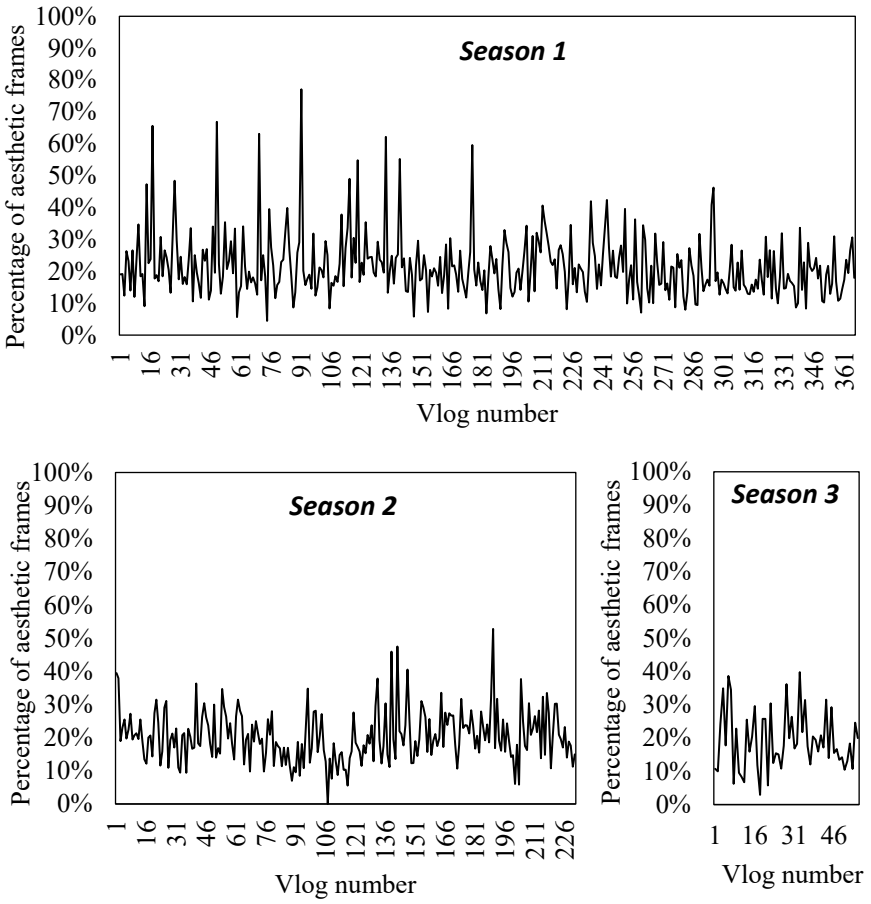

Figure 4. Percentages of good-quality frames in each video of the different seasons of vlogging by Casey Neistat.

Aesthetic Prediction on YouTube Content Creators

While the previous analyses of films are an indirect attempt at investigating patterns related to film directors' creative processes (e.g., Quentin Tarantino), the abundant video content provided by the internet, and especially YouTube, allows this to be studied further. With some video content creators producing up to one video per day, it is now possible to look at the aesthetic prediction of videos for one producer over time, and possibly over a career. For this task, the filmmaker and YouTube video creator Casey Neistat is selected due to the fact that he has a certain interest in videography

and is one of the first vloggers with 600 vlogging videos already online. As shown in Figure 4, percentages of high aesthetic quality frames for each video vary over the different seasons of Casey Neistat's vlogging, but the average aesthetic remains steady across all seasons. Videos achieving 60% to 77% of good-quality frames were all shot in a studio with professional lighting and framing, unlike most of his videos, for which filming was done spontaneously. Surprisingly, no correlation exists between the percentage of good-quality frames predicted by the computational aesthetic classifier and any of the parameters indicating a video's popularity, such as the number of views or the user rating. As previously shown, videos of higher aesthetic quality are classified with more confidence, as is illustrated by the previous statement. It can be suggested that video content creators could use computational aesthetic prediction to improve the quality of their content by selecting particular sequences or shooting their videos in different conditions to match the suggested standards emerging from the aesthetic classifier's training, which mimics people's preferences.

Conclusion

This paper demonstrates that a computational aesthetic classifier trained on photographs can also be used for classification of videos. Despite showing modest performances in this respect, the results demonstrate the cross-media capabilities of the original classifier, particularly when focusing exclusively on visual material. Moreover, processing frames individually and observing them as sequences across films makes it possible to learn more about the content creator's decisions and the aesthetic classifier's preferences, which is a novel approach in the domain of computational aesthetic prediction of videos. The first two tests in this paper respectively focused on the creative product as a whole and its content; the third experiment on vlogging videos establishes a new method to investigate video content creators over time. The approach offered by this paper is novel due to its focus on meaningful decisions of content creators, rather than traditional computational classification tasks. One application of this approach is to support video content creators in aesthetic decisions during the editing and postproduction process, giving them an immediate estimation of their audience's visual appreciation. Furthermore, it could be developed into assisting technology for visually impaired people who are willing to share their experiences and communicate through videos.

Acknowledgments

I would like to thank Agi Haines, Roger Malina, Michael Straeubig and Jane Grant for providing helpful comments and fruitful discussions to improve the paper. This work was completed as part of Marie Curie Initial Training Network FP7-PEOPLE-2013-ITN, CogNovo, grant number 604764.

References

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016, September 27). YouTube-8M: A Large-Scale Video Classification Benchmark. Retrieved from https://arxiv.org/abs/1609.08675

Clark, C. (2014). *The colors of motion*. Retrieved from http://thecolorsofmotion.com/films

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In A. Leonardis, H. Bischof, A. Pinz (Eds.) *Computer Vision – ECCV 2006* (pp. 288–301). Berlin, Germany: Springer-Verlag. doi:10.1007/11744078_23

Lemarchand, F. (2017). *Fundamental visual features for aesthetic classification of photographs across datasets*. Manuscript submitted for publication.

Lu, X., Lin, Z., Jin, H., Yang, J., & Wang, J. Z. (2014). RAPID: Rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia - MM '14* (pp. 457–466). doi:10.1145/2647868.2654927

Marchesotti, L., Perronnin, F., Larlus, D., & Csurka, G. (2011). Assessing the aesthetic quality of photographs using generic image descriptors. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1784–1791). doi:10.1109/ICCV.2011.6126444

Murray, N., Marchesotti, L., & Perronnin, F. (2012). AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 2408–2415).

Niu, Y., & Liu, F. (2012). What makes a professional video? A computational aesthetics approach. *IEEE Transactions on Circuits and Systems for Video Technology, 22*(7), 1037–1049. doi:10.1109/TCSVT.2012.2189689

Romero, J., Machado, P., Carballal, A., & Santos, A. (2012). Using complexity estimates in aesthetic image classification. Journal of Mathematics and the Arts, 6(2–3), 125–136. doi:10.1080/17513472.2012.679514

Shulman, J. (2017). *Photographs of films*. Retrieved from http://www.jasonshulmanstudio.com/photographs-of-films/

Tang, X., Luo, W., & Wang, X. (2013). Content-Based Photo Quality Assessment. *IEEE Transactions on Multimedia, 15*(8), 1930–1943. doi:10.1109/TMM.2013.2269899

Tzelepis, C., Mavridaki, E., Mezaris, V., & Patras, I. (2016). Video aesthetic quality assessment using kernel Support Vector Machine with isotropic Gaussian sample uncertainty (KSVM-IGSU). In *2016 IEEE International Conference on Image Processing: Procedeeings* (pp. 2410–2414). doi:10.1109/ICIP.2016.7532791

Yang, C.-Y., Yeh, H.-H., & Chen, C.-S. (2011). Video aesthetic quality assessment by combining semantically independent and dependent features. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing: Proceedings* (pp. 1165–1168). doi:10.1109/ICASSP.2011.5946616