Towards a Comprehensive Computational Model for Aesthetic Assessment of Videos

Subhabrata Bhattacharya
subh@cs.ucf.edu

Behnaz Nojavanasghari
behnaz@eecs.ucf.edu

Tao Chen
taochen@ee.columbia.edu

Dong Liu
dongliu@ee.columbia.edu

Shih-Fu Chang
sfchang@ee.columbia.edu

Mubarak Shah
shah@cs.ucf.edu

ABSTRACT

In this paper we propose a novel aesthetic model emphasizing psychovisual statistics extracted from multiple levels in contrast to earlier approaches that rely only on descriptors suited for image recognition or based on photographic principles. At the lowest level, we determine dark-channel, sharpness and eye-sensitivity statistics over rectangular cells within a frame. At the next level, we extract Sentibank features ($1,200$ pre-trained visual classifiers) on a given frame, that invoke specific sentiments such as "colorful clouds", "smiling face" etc. and collect the classifier responses as frame-level statistics. At the topmost level, we extract trajectories from video shots. Using viewer's fixation priors, the trajectories are labeled as foreground, and background/camera on which statistics are computed. Additionally, spatio-temporal local binary patterns are computed that capture texture variations in a given shot. Classifiers are trained on individual feature representations independently. On thorough evaluation of 9 different types of features, we select the best features from each level – dark channel, affect and camera motion statistics. Next, corresponding classifier scores are integrated in a sophisticated low-rank fusion framework to improve the final prediction scores. Our approach demonstrates strong correlation with human prediction on $1,000$ broadcast quality videos released by NHK as an aesthetic evaluation dataset.

Category and Subject Descriptors: H.4 [Information Systems Applications] : Miscellaneous

Keywords: Video Aesthetics, Affect features, Camera motion features, Cinematography, Low Rank late fusion

1. INTRODUCTION

Automatic aesthetic ranking of images or videos is an extremely challenging problem as it is very difficult to quantify beauty. That said, computational video aesthetics [4, 7–11, 13, 14] has received significant attention in recent years. With the deluge of multimedia sharing websites, research in this direction is expected to gain more impetus in future, apart from the obvious intellectual challenge in scientific formulation of a concept as abstract as *beauty*.

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

http://dx.doi.org/10.1145/2502081.2508119.

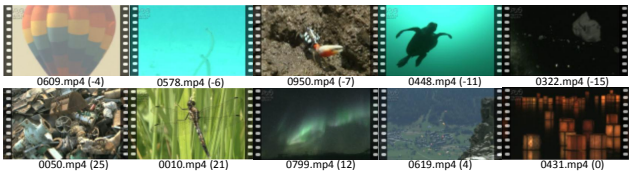


Figure 1: Video aesthetic assessment: Top row shows frames from videos ranked from $1-5$ and bottom row shows the bottom 5 ranked videos in the NHK video aesthetic dataset, using the proposed approach. Numbers in parentheses indicate deviation from ground truth rankings. Our approach correctly predicts video 0431 (bottom right) to be the lowest ranked.

Since there is no well-defined set of rules following which any image or video can be deemed as beautiful, the problem can be posed appropriately in machine learning paradigm. In this vein, the authors of [4, 7] propose the use of low-level statistical features e.g., hue count, blur, saturation to predict aesthetic appeal. [9] insinuated using these properties for foreground regions, which intuitively is a more reasonable approach. However it involves segmentation. Later these are circumvented using visual saliency maps [22] or dedicated content based classifiers [8].

In contrast to these customized statistics that are believed to capture some aspects of aesthetics, the authors of [10] propose the use of image descriptors popular in object recognition literature to assess beauty of photographs. Except for a few earlier efforts [9, 11, 13], most research efforts are concentrated on modeling aesthetic in still images. While features proposed in [4, 7, 8, 10, 14] demonstrate some success for images, we argue that these do not necessarily apply to videos. For example, rule of thirds – a composition guideline used to assess photography, requires foreground to be aligned to 4 specific locations in an image frame [8, 9]. Videos captured with such constraints, tend to have non-moving foreground and lead to viewer dissatisfaction [13]. Also, cinematographic techniques – camera motion, lighting changes etc. are often introduced in produced videos to increase their aesthetic appeal, which cannot be captured using aesthetics based on single images.

Motivated by the above, we propose a hierarchical framework encapsulating aesthetics at multiple levels which can be used independently or jointly to model beauty in videos. Our contributions are: (1) We extract motion statistics that latently encode cinematographic principles, specific to foreground and background or camera, superior to approaches proposed in [9, 11, 13], (2) We introduce application of human sentiment classifiers on image frames that capture vital affective cues which are directly correlated with visual aesthetics and are semantically more interpretable than [8], (3) We employ a relatively small set of low-level psychovisual [15, 18] features as opposed to [4, 7, 9, 10, 14] and encode them efficiently into descriptors that capture spatial variations within video frames, and finally, (4) We exploit a more sophisticated fusion scheme [21] that

show consistent improvement in overall ranking when compared to methods used in [10, 11, 13]. Section 2 discusses computation of these features in detail.

2. VIDEO AESTHETICS

We analyze a video at multiple granularities employing three strategies, each catering to specific aspects of aesthetics. Each video is divided into shots and for each shot, keyframes are selected uniformly. Following this, features are selected at three levels – cell, frame and shot.

2.1 Cell-level

At this level, each keyframe in a shot is divided into $m \times m$ grids of rectangular cells. Each cell ($\{A_j\}_{j=1}^{m \times m}$) is further described using the following statistics, thus providing a mean to analyze low-level aesthetics in different spatial regions: (a) The **Dark-channel** statistics, proposed in [5] and used in [8], is essentially a minimum filter on RGB channels which reflects local clarity, saturation, and hue composition in a given image (I) subjected to a neighborhood, $\Omega(i)$ around i-th pixel. For j-th, cell the dark channel feature is computed in Eqn. (1) as:

$$F_{dc}(j) = \sum_{(i) \in A_j} \frac{\min\limits_{c \in R,G,B} (\min\limits_{i' \in \Omega(i)} I_c(i'))}{\sum\limits_{c \in R,G,B} I_c(i)}, \quad (1)$$

with each real valued numbers normalized by the cell area. (b) **Sharpness** statistics are derived from squared root of the product of spectral map (α_{x_j}) and sharpness map computed over 8-pixel row (x_j^r) and column neighborhoods (x_j^c) of the gray-scale equivalent (x_j) of a cell from a color frame (A_j). The simplified formulation from [19] as given in Eqn. (2)

$$F_{sp}(j) = \left[\left(1 - \frac{1}{1 + e^{-3(\alpha_{x_j} - 2)}} \right) \left(\frac{1}{4} \max_{r,c} \frac{\sum\limits_{r,c} |x_j^r - x_j^c|}{255} \right) \right]^{\frac{1}{2}}, \quad (2)$$

captures the perceived sharpness of j-th cell, indicating the amount of detail in each cell. (c) **Eye-sensitivity** is one of the oldest known aesthetic attribute known to humans, which is associated with the visual cortex's sensitivity to colors of certain wavelength in the visible spectrum. We obtain this by building a weighted color histogram for a given cell using the corresponding color sensitivity scores predefined in [18]. The peak of the histogram represents the sensitivity score with respect to the human eye, for a given cell and is used as our final cell level statistics ($\mathbf{F_{es}}$).

2.2 Frame-level

Global frame-level descriptors as used in [10] are more suited in for recognition of scenes or objects, primarily due to their invariance towards pose, viewpoints, illuminance condition under which images are captured etc. We hypothesize that such properties are not desired in the current problem context and resort to detection of a more abstract concept – *human affect* present in a given video frame. We apply a large set of affect based classifiers from a recent work by Borth and colleagues [2], hereafter being referred to as SentiBank Detectors. These detectors are founded on a psychological research known as Plutchik's Wheel of Emotions [15].

To train this detector library, a Visual Sentiment Ontology is firstly built by data-driven discovery from Internet: adjective and noun tags are extracted from Flickr and Youtube based on their co-occurrence with each of the 24 emotions defined in Plutchik's theory. These tags are then assigned with sentiment values and used to form Adjective Noun Pairs (ANP) such as "cold winter" or "happy baby". More than 3,000 ANPs are then selected to form the ontology based on their frequency ranking on Flickr and semantic coverage. For each ANP of the ontology a detector is trained from Flickr images that are tagged with this ANP and only 1,200 ANP concept detectors are selected according to the detector performance to form the SentiBank detectors library, which we apply on video frames to generate a 1,200 dimensional ANP detector response, hereafter referred to as $\mathbf{F_{af}}$. Fig. 2 demonstrates the results of applying a small subset of these detectors on 4 different keyframes from 1,000 videos in the NHK video aesthetics dataset. In each case, we observe direct correlation between the sentiment invoked in a viewer with the individual detector's confidence, which argues in favor of their applicability in this context.

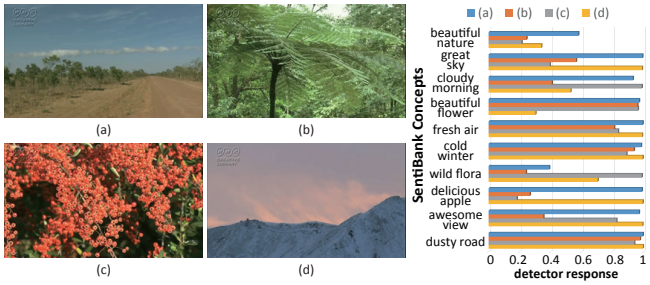


Figure 2: Results of applying sentiment classifiers on 4 different keyframes from the NHK dataset. A subset of 10 unique classifiers listed on the right of the figure, are applied on the keyframes shown above. Detection confidence for each of these classifiers for all 4 images are shown as a barchart.

2.3 Shot-level

This is the final level of our feature hierarchy where we compute three different statistics from a video shot based on (a) **Foreground** and (b) **Background motion** and (c) **Texture dynamics**. To accomplish the former two tasks, we track distinctive points [16] from a sequence of frames in a shot, to get meaningful trajectories. Since trajectories can either belong to moving foreground objects and either non-moving background or moving camera, we further apply a Bayesian spatio-temporal saliency detection algorithm [22] based on natural image statistics, employing priors from viewer's eye fixation. This generates a coarse map for every frame with regions corresponding to foreground with high saliency and those to background with low saliency. Using this coarse map, trajectories are labeled whether they are emanated from either foreground or background.

For both types of trajectories, we independently perform a low-rank matrix decomposition using [3] to find dominant trajectories belonging to both motion classes. The low-rank assumption is based on the observation that: all trajectories belonging to either foreground or background, when stacked together, form a matrix which can be factorized into a low-rank component and a noise component. This step ensures further refining of trajectories obtained initially. First k-components of the low-rank matrix identify the dominant trajectories. Finally each dominant trajectory is described using statistically invariant features listed as: mean, standard deviation, range, minimum, maximum, l_1 and l_2 norm. Thus, in this case, a 2 dimensional x, y trajectory of arbitrary length, is reduced to a 2×7 dimensional vector, denoted as $\mathbf{F_{fg}}$ or $\mathbf{F_{bg}}$ depending on its source (foreground or background/camera).

In addition, we also represent a shot based on its texture dynamics. For this purpose, we employ a spatio-temporal local-binary pattern algorithm [23], for describing volumetric texture varia-

tions in videos. The spatio-temporal LBP descriptor is computed by binning LBP codes along a small spatio-temporal neighborhood in three orthogonal planes: $x-y$, $x-t$ and $y-t$. Hence the final descriptor for a shot is given by a histogram ($\mathbf{F_{tx}}$) of 3×2^n, n indicating the spatial-pixel neighborhood in x, y plane.

All descriptors are further clustered independently using a fast k-means clustering algorithm [12] to create vocabularies of different sizes. Thereafter, each video is described using separate bag-of-X representations obtained by standard vector quantization. Here X denotes a particular feature type from our pool of features: $\mathbf{F_{dc}}$ (dark channel), $\mathbf{F_{sp}}$ (sharpness), $\mathbf{F_{es}}$ (eye-sensitivity), $\mathbf{F_{af}}$ (affect), $\mathbf{F_{fg}}$ (foreground motion), $\mathbf{F_{bg}}$ (background motion), and $\mathbf{F_{tx}}$ (texture dynamics). Finally, bag-of-X representations for a particular feature type, are used in a ranking SVM framework [6] to learn an aesthetic model for the given feature type. In the next section, we discuss how individual models generated from complementary features, can be efficiently fused to produce a final aesthetic model.

2.4 Rank Fusion

We employ the Low-Rank Late Fusion (LRLF) scheme [21] to fuse the ranking scores from m ranking SVM models, each of which is trained with one specific aesthetic feature. The advantage of this fusion scheme is that it is not only isotonic to the numeric scales of scores from different models but also removes the prediction errors from each model. Given a ranking score vector $\mathbf{s}^i = \{s_1^i, \ldots, s_n^i\}$ from the i-th model, where n is the number of videos and each s_j^i denotes the ranking score of the j-th video predicted by the i-th model. We convert it into an comparative relationship matrix T^i such that $T_{jk}^i = 1$ if $s_j^i > s_k^i$, $T_{jk}^i = -1$ if $s_j^i < s_k^i$ and $T_{jk}^i = 1$ if $s_j^i = s_k^i$. In this way, the real-valued score vector is encoded as an integer matrix, which gets rid of the numeric scale variances among scores from different models. Taking the comparative relationship matrices $\{T^i, i = 1, \ldots, m\}$ as input, LRLF scheme seeks a shared rank-2 matrix through which each T^i can be decomposed into a common rank-2 matrix and a sparse error matrix. Our hypothesis behind the rank-2 matrix is that if ideal score vector $\mathbf{s}^*$ exists, the real-valued comparative relation matrix T^* is a rank-2 matrix formed by $T^* = \mathbf{s}^*\mathbf{e}^\top - \mathbf{e}\mathbf{s}^{*\top}$, where $\mathbf{e}$ is a all-one vector. LRLF considers the reverse problem: Based on the inferred rank-2 matrix $\hat{T}$, it performs rank-2 factorization of $\hat{T}$ such that $\hat{T} = \hat{\mathbf{s}}\mathbf{e}^\top - \mathbf{e}\hat{\mathbf{s}}^\top$. Finally, the recovered $\mathbf{s}^\top$ can be used as final fused score for video aesthetic ranking.

3. EXPERIMENTS

We evaluate our algorithm on the recently released "NHK - Where is beauty?" dataset, consisting of $1,000$ broadcast quality videos, aesthetically ranked from best (1) to worst ($1,000$) by humans. In our setup, each keyframe from a video is divided into 3×3 grids containing a total number of 9 cells. Empirically, $m = 3$ implicitly mandates one of the composition rules proposed in [1, 4, 9] – the rule of thirds, hence a natural choice. Sharpness for a cell, described using minimum, maximum, mean and standard deviation of the S_3 map discussed in Eqn. (2) leads to a 9×4 dimensional vector i.e, $\mathbf{F_{sp}} \in \mathbb{R}^{36}$. Dark channel and eye-sensitivity are measured described using the mean and maximum values for a given cell, respectively, implying $\mathbf{F_{dc}} \in \mathbb{R}^9$ and $\mathbf{F_{es}} \in \mathbb{R}^9$. For frame level features, we replicate the experimental setup in [2]. For shot-level features, k is empirically set to 5 to select top 5 dominant trajectories from each of the two motion classes. Finally, while computing the texture dynamics descriptor based on [23], we consider a $5 \times 5 \times 5$ neighborhood yielding $\mathbf{F_{tx}} \in \mathbb{R}^{96}$ ($3 \times 2^5 = 96$).

In addition, we implemented two independent baselines inspired from [10] whose performance are recorded in top two rows in Tab. 1. The two baselines are generated on generic image [17] (BL-I) and video [20] (BL-V) descriptors, widely used in recognition tasks. All descriptors are randomly sampled from 200 videos to construct vocabularies with different sizes ($128, 256, 512, 1024$) and subsequently for every vocabulary-feature combination, bag-of-X representations are generated for all videos. We use SVM rank [6] on linear kernels using the 1-slack (dual) with constraint cache optimization algorithm, to train separate models for each setting. We employ a 5-fold cross validation scheme to select the cost parameter for each SVM. Once parameters are selected, we use the regular query setting with $80\% - 20\%$ (training - testing) split, ensuring testing samples do not appear during the training phase for a particular split. Thus with 5 splits, we can generate the ranks for all videos in the dataset. Finally, evaluation is performed using the Kendall (τ) and Spearman's (ρ) rank correlation coefficients:

$$\tau = \frac{N_c - N_d}{\frac{1}{2}N(N-1)}, \text{and} \quad \rho = 1 - \frac{6\sum d_i^2}{N(N^2-1)}, \tag{3}$$

where N_c and N_d denote the number of concordant and discordant pairs respectively, N being total number of samples $= 1000$, and d_i being the difference of between the ground truth and predicted ranks. Both τ and ρ can take values between $[-1, 1]$, where -1, 1 indicating ground truth rank and predicted ranks are completely reverse, and same, respectively, while 0 signifies no-correlation between the two rank sets.

Feature	Kendall's coefficient (τ)				Spearman's coefficient (ρ)			
	128	256	512	$1,024$	128	256	512	$1,024$
BL-I	-0.011	0.001	0.006	-0.002	-0.041	0.014	-0.013	0.012
BL-V	0.001	-0.021	0.012	0.001	0.001	0.001	-0.002	-0.049
$\mathbf{F_{dc}}$	0.014	**0.017**	0.016	0.013	0.039	**0.025**	0.045	0.044
$\mathbf{F_{sp}}$	-0.018	0.017	0.014	0.023	-0.027	-0.003	0.026	0.031
$\mathbf{F_{es}}$	0.003	0.004	0.006	0.001	-0.004	-0.004	0.004	-0.002
$\mathbf{F_{af}}$	0.028	0.021	0.034	**0.030**	0.035	0.041	0.045	**0.046**
$\mathbf{F_{fg}}$	0.009	**0.005**	0.011	0.013	0.015	**0.008**	0.013	0.021
$\mathbf{F_{bg}}$	**0.027**	0.041	0.039	-0.016	**0.039**	0.044	0.047	-0.018
$\mathbf{F_{tx}}$	-0.001	-0.001	-0.002	**-0.000**	0.001	-0.004	-0.003	**-0.003**
FUSED	0.0861				0.1266			

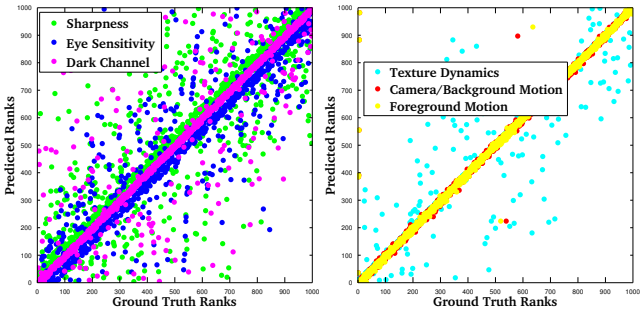
Table 1: **Summary of experiments using classifiers trained on different Bag-of-X representations of baseline features from [17] (BL-I) and [20] (BL-V), followed by our proposed features (labeled as $\mathbf{F_{xx}}$). The row labeled FUSED shows the best fusion results obtained after combining 5 different feature combinations (indicated in bold) using [21].**

Tab. 1 summarizes all our experiments. For every feature - vocabulary combination, both Kendall(τ) and Spearman (ρ) are reported. Our proposed psychovisual features perform significantly better than the baselines reported as BL-I and BL-V in top two rows. Thus, classifiers trained on generic image level descriptors such as colorSIFT [17] and video level descriptors such as MBH [20] are not suited for quantifying video aesthetics. Intuitively, this can be explained from the following observations: (a) colorSIFT descriptors on high-quality images capture more detail (Flickr images as shown in [10]) as compared to low-resolution 640×360 video frames from the NHK dataset, and (b) MBH [20] descriptors are unable to differentiate between motion originating from background or foreground, which is necessary for describing motion based aesthetics, resulting into generation of suboptimal aesthetic models.

The last row in Tab. 1 (FUSED), shows fusion [21] results obtained from a two-step process: first we use cross validation to select top 5 features (which may use different vocabulary sizes) based on high ρ and τ values. Next we use various late fusion methods

(score averaging, normalized rank, and low-rank fusion [21]) to fuse these independent 5 scoring models. Thus selecting an optimal algorithm for fusing knowledge from individual models, boosts overall performance. More detailed analysis of fusion is provided in Fig. 4. It is evident that models built on shot-level features: camera/background motion perform well with smaller vocabulary sizes, in comparison to the ones that are trained using texture descriptors and sentibank features - which perform better with larger vocabularies.

We also observe that, shot level features : camera/background and foreground motion statistics outperform all other features in most cases. It is also interesting to note that frame-level sentibank features also perform equally well. Among cell-level features, dark channel based statistics demonstrate best performance. This is further supported by the scatter plot in Fig. 3(a). Ideally, the dots in the figures are expected to be along the line that connects $(1, 1)$ to $(1000, 1000)$, indicating no discordance between ground truth rank and predicted rank. We notice that magenta dots, corresponding to dark channel features are more aligned to the line, than those belonging to sharpness or eye-sensitivity.



(a) (b)

Figure 3: Scatter plot showing label disagreement between (a) Cell-level (Dark channel – Magenta , Sharpness – Green, Eye-sensitivity – Blue), and, (b) Shot-level (Foreground motion – Yellow, Camera/Background motion – Cyan, Texture dynamics – Red) feature based classifiers (Best viewed when zoomed). Each dot is plotted according to Cartesian pairs (Ground Truth Rank, Predicted Rank).

However, in case of shot level features, we observe a larger degree of agreement between the ground truth and the prediction, in comparison to the cell-level features. This is reflected in Fig. 3(b). Finally, in Fig. 4, we provide some results using classifiers trained on frame-level affect based features (Orange). Except for a few ($< 5\%$), we see strong label agreement between ground truth and prediction - similar to the shot-level features. This encourages us to fuse the individual classifier outcomes in 3 different settings. These results are also shown in Fig. 4. Fusion of classifiers trained on affect and dark channel based features are indicated in Turquoise. These results are further improved after adding classifiers trained on camera/background motion (Lime green). Ultimately fusion results from classifiers trained on top 5 features is shown in Pink. We notice strong level of concordance in this setting.

4. CONCLUSION

We proposed a novel model for assessing beauty of broadcast quality videos, emphasizing on statistics from multiple granularities i.e. cells within keyframes, entire frame and shots. Specifically, we introduced three novel features – affect statistics at frame level, and motion statistics of foreground and background at shot level. Through extensive experiments, we demonstrated that using only a handful of carefully selected features, and efficiently fusing models learned on top them, can greatly increase the performance of aesthetic evaluation over previously published approaches. Although

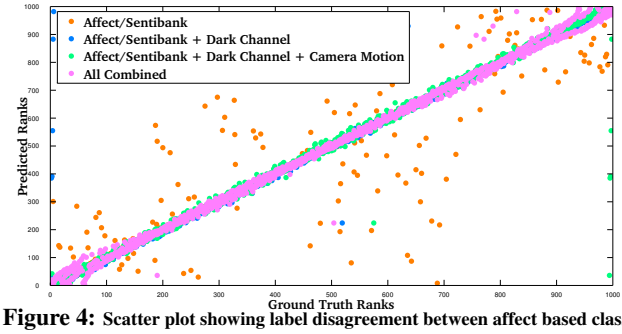


Figure 4: Scatter plot showing label disagreement between affect based classifier results (Orange) versus ground truth. Fusion results are also shown : affect + dark channel (Turquoise), affect + dark channel + camera/background motion (Lime green), and top 5 features combined (Pink). Refer to Fig. 3 for details on interpretation.

our results indicate significant performance gain over prior work, we conjecture that there is still plenty of scopes for further improvement. Since research in aesthetic quality assessment of videos is currently at an inchoate stage, we intend to explore how knowledge from different domains images, audio, text etc. can be exploited in this direction, as part of our future work.

5. REFERENCES

[1] S. Bhattacharya, R. Sukthankar, and M. Shah. A holistic approach to aesthetic enhancement of photographs. *TOMCCAP*, 7(Supplement):21, 2011.

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.

[3] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, pages 288–301, 2006.

[5] K. He, J. Sun, and X. Tang. Single image haze removal using dark channel prior. *TPAMI*, 33(12):2341–2353, 2011.

[6] T. Joachims. Optimizing search engines using clickthrough data. In *ACM KDD*, pages 133–142, 2002.

[7] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, pages 419–426, 2006.

[8] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *ICCV*, pages 2206–2213, 2011.

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

[10] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, pages 1784–1791, 2011.

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

[12] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, pages 331–340. INSTICC Press, 2009.

[13] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, pages 2408–2415, 2012.

[14] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato. Aesthetic quality classification of photographs based on color harmony. In *CVPR*, pages 33–40, 2011.

[15] R. Plutchik. Harper & Row Publishers, 1980.

[16] J. Shi and C. Tomasi. Good features to track. In *CVPR*, pages 593–600, 1994.

[17] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 32(9):1582–1596, 2010.

[18] J. J. Vos. Colorimetric and photometric properties of a 2-deg fundamental observer. *Color Research and Application*, 1978.

[19] C. Vu, T. Phan, and D. Chandler. S3: A spectral and spatial measure of local perceived sharpness in natural images. *TIP*, 21(3):934–945, 2012.

[20] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.

[21] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang. Robust late fusion with rank minimization. In *CVPR*, pages 3021–3028, 2012.

[22] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A bayesian framework for saliency using natural statistics. *J Vis*, 8(7):32.1–20, 2008.

[23] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *TPAMI*, 29(6):915–928, June 2007.