VIDEO AESTHETIC QUALITY ASSESSMENT BY COMBINING SEMANTICALLY INDEPENDENT AND DEPENDENT FEATURES

Chun-Yu Yang[1], *Hsin-Ho Yeh*[1] *and Chu-Song Chen*[1,2]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan.
[2] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan.
{cyyang, hhyeh, song}@iis.sinica.edu.tw

ABSTRACT

This paper aims to accomplish the work of assessing the aesthetic quality of a video. Unlike previous assessing works focusing mainly on the extraction of aesthetic features in a film, we further study the features, discover their semantic property on videos and then come up with more useful video-based features such as motion space and motion direction entropy. In the experiment, we compare the assessing accuracy between two different semantic types of features and find that the semantic-independent feature is more reliable from the results. By combining all features, our method learned a more robust and accurate assessment model.

Index Terms— Image Quality Assessment, Video Quality Assessment, Aesthetic Quality.

1. INTRODUCTION

For decades, high quality videos have been the pursuit of human beings. In an HD-video-widely-spread era, to judge the quality of a video would mostly depend on the aesthetics or the comfort degree of the video. For example, continuous shaking of a film or deficient lighting on a target or even an obscure subject might agitate the audience and cause their uncomfortableness on watching. On the other hand, a film carrying high aesthetic characteristics for instance bright colour or delicate composition impresses the audience better. To achieve high aesthetic characteristics, professionals adopt special techniques to make their photos perfect in an aesthetic view, say, the DOF (depth of field) difference between foreground and background, and the well-known rule of third [1]. In a similar way, depending on the technique, human beings can pursue a higher quality video with the aids of the automation in assessing the aesthetic quality (AQ) of videos.

Automatic aesthetic assessment has many applications. For example, imagine that you are learning to record something professionally. Instead of criticizing the bad compositions, over/under-exposure on lighting, or hand shaking afterwards, it is more helpful if the camcorder report the AQ to the user in advance. In addition, AQ assessment can also help key frame selection, seeing that key frames should consider not only the contents but the aesthetics as well.

From the advantages above, the problem of measuring the AQ of videos becomes important. In the past, AQ assessment had been studied thoroughly in photos [2][3][4]. They focus on extracting AFs to represent the artistic feeling a person perceive toward photos. These feelings are widely discussed and embodied into colour, exposure, composition, or more. Then they model the assessing behaviour similar to the human beings as a classification problem. Real-AdaBoost and SVM are used in their works for classification respectively. In later work, Luo et al. [5] attempt to assess video quality using these AFs and propose an insightful view that the AQ should be focused on the subject because it gathers most attention within the whole image; thus, subject-based AFs are extracted from subject regions as a criteria for assessing photo/video quality. However, Luo et al. claim that, in a professional image, the subject region should always be in focused and background should be out of focused. Hence, the subject extracted from the above definition may fail since most of videos captured by the consumer are not professional enough. Moreover, without considering temporal property, the photo-based AFs are insufficient in judging the AQ of videos. Recently, Moorthy et al. [6] propose a hierarchical pooling method by combining photo-based AFs and temporal property to model the AQ of videos. However, the discriminating ability of AFs still needs improvements for lacking of consideration on motion property.

To sum up the past works [2][3][4][5][6], many focus on extracting the features but miss consider their semantic property. To further explain, some features have the property that aesthetic criteria measured from them vary with the video content, for example, lightness. We call them semantic-dependent features. For example, when we evaluate the lightness feature of videos containing starlight and sunlight, lightness is its innate property, and therefore their AQs should not be assessed by the same criteria. However, some AFs have an independent property. For example, the severe vibration of the scene always arouse the uncomfortableness of the audience no matter what video contents are. Therefore, the aesthetic criteria measured from them remain the same even

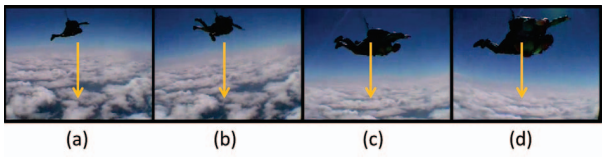

Fig. 1. Frames from (a) to (d) represent the sequence of videotaping skydiving and yellow arrow indicates the imaginary space. From these frames, the motions of skydiver and camera give audience more imagination due to more free space.

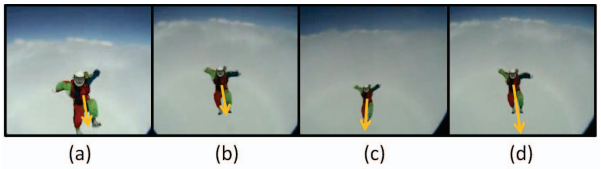

Fig. 2. The position of the skydiver and the camera motion give the audience less imagination due to lack space.

with different video content. Here we call this kind of AFs semantic-independent features. In fact, both semantically dependent and independent features are important since they can be applied to different situations: dependent features are more useful in similar scenes while independent features are more distinctive in diverse datasets. In this paper, we study features such as motion, colour, and composition, and investigate their assessing ability concerning semantic properties. Later we conduct experiments to explore the influences of the semantic property on AQ. By combining all features, our method assesses video quality more accurately

2. OUR APPROACH

2.1. Problem Definition

For i^{th} video V_i, we define notations to illustrate our work: a video-pooled feature X_i is extracted from video V_i with m frames, $V_i = \{F_1, F_2, ...F_m\}$, F_j for the features of the j^{th} frame. Similarly, for the frame j with n features, $F_j = \{f_1, f_2, f_3, ..., f_n\}$, where $f_k \in \mathbb{R}$ and f_k represents the k^{th} type of feature. With all these features provided, we use a pooling method[1] to summarize the features within a short duration. The pooling method (P) includes several operators, eg., mean, median. To be more specific, we have frame features within one second, $A_j = [F_j, F_{j+1}, ..., F_{j+N}]^T$ where N is frame per second, pooled together, and obtain a one-second-pooled feature vector, $B_j = P_1(A_j)$. Then we apply another pooling to summarize the one-second-pooled features in one video to form a video-pooled feature $X_i = P_2(B_{j;\forall j})$[2].

2.2. Aesthetic Feature Computation

Among all the aesthetic characters in a video, motion is the most salient character in the so-called video. Whereas

[1]For more detail on feature pooling, refer to [6].

[2]According to [6] : $P_1 = \{\texttt{mean}, \texttt{median}, \texttt{min}, \texttt{max}, 1^{st}\ \texttt{quartile}, 3^{rd}\ \texttt{quartile}\}$ and $P_2 = \{\texttt{mean}, \texttt{std}\}$; hence, $X_i \in \mathbb{R}^{n \times 12}$.

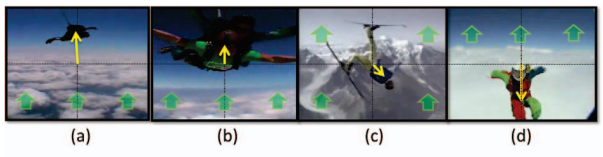

Fig. 3. In these frames, v is the green arrow representing the average optical flow and d is the yellow arrow. Frames from (a) to (d) get f_1 equals 0.82, 0.47, -0.26, -0.79 respectively.

previous work considers only the motion in finding hand-shaking [7], we further introduce some important features: motion space and motion direction entropy, which are also of great significance. Seeing that all AFs can be classified into two types: semantic-independent and semantic-dependent features, we introduce them respectively in Sec. 2.2.1 and Sec. 2.2.2.

2.2.1. *Semantic-Independent Feature*

Motion Space (MS): Based on the professional skills in film making: as users are videotaping a moving object, they must beware that the space in front of the moving direction of the object should be reserved for the better AQ of the video, since the reserved space gives the audience more imagination about the subject. Moreover, different scenes make no difference on the effect of motion space since what this feature concerns about is the imagination space; therefore this is a semantic independent feature. Fig. 1 and Fig. 2 illustrate the situations for more/less-imagination, respectively. We tackle this problem by setting the direction of optical flow as v and the vector between subjects and center of frame as d, which are shown in Fig. 3. Both vectors (v and d) are normalized by their size individually. The feature of MS is represented as $f_1 = v \cdot d$, where $\cdot$ is inner product and $f_1 \in [-1, 1]$.

Hand-shaking (HS): Hand-shaking occurs occasionally and has often been disturbing when the audience tries to concentrate in a video, thus making it significant to distinguish the high AQ videos from low; obviously HS is semantic-independent. Shaking differs from other features for it is gained by computing the change of motion direction between the current frame and the previous one, instead of only the direction of its own. Apart from all the other related works, we set shaking detection area at the border to distinguish subject's self-shaking from the hand-shaking, as shown in Fig. 4.

To achieve this, the motion indicator I_j in frame j is defined as an unit step function $I_j = u(mx_j)$ where mx_j represents the border motion vector in frame j along the horizontal direction. An exclusive-or operator ($\oplus$) is adopted to model the direction change within two adjacent motion indicators, that is I_j and I_{j-1}. Moreover, the magnitude of motion indicates the degree of unstableness; thus horizontal unstableness feature f_2 is defined as:

$$f_2 = (I_j \oplus I_{j-1}) \times (|mx_j| + |mx_{j-1}|).$$

Fig. 4. The unstableness detection area (a) defined using the whole frame may cause false alarm in a self-shaking subject such as this ball-playing puppy shown above. Therefore, the modified detection set (b) along the borders helps detect a more robust frame-based unstableness.

In the same way, a vertical unstableness feature f_3 is formed by replacing mx with my.

We further define the border(Ub) to central unstableness (Uc) ratio for horizontal $f_4 = \frac{Ub}{Uc}$to show the subtle differences: for border moves more than the centre type ($f_4 > 1$), it may be the recording type that its shot traces and focuses on the subject; for border moves equally with the centre type ($f_4 = 1$), it may be a panorama shooting type. Finally for the border moves less than the centre type ($f_4 < 1$), it may be regarded as a static shot on a subject. Similarly f_5 for vertical.

Colour Harmonic: The human visual perception of aesthetics is strongly related to colour harmonization [9] and it is semantically independent since it aims at colour arrangement. Here we adopt HSV colour space and construct a hue histogram (h) for each frame to distinguish between seven well arranged colour types [9]. The seven templates for each types of colour histogram are listed as $Tp = [Tp_1, Tp_2, ..., Tp_7]^T$.

$$f_{5+c} = |h *' Tp - Tp_c *' Tp|^2, \forall c \in [1, 2, ..., 7], \quad (1)$$

where $*'$ means to convolute and choose the maximum value of the resulted histogram.

Composition: The composition of a frame is also of great importance in an aesthetic view, eg., the rule of third [1], clarity contrast, and shape convexity. The rule of third claims that subjects should be placed in one of the four intersections of the lines that divide the images equally in three parts horizontally and vertically for better aesthetic appeal. Clarity contrast for subject and background also plays a role in a professional image since the focused subject with a blurred background is of great aesthetic appeal. The shape convexity is a consideration according to [5] as well. In computing rule of third feature (f_{13}), clarity contrast feature (f_{14}) and shape convexity feature (f_{15}), we adopt the methods in Luo et al. [5]. And these rules are semantically independent since photographers seek for better composition in all kinds of scene.

2.2.2. Semantic-Dependent Feature

Motion Direction Entropy (MDE): Entropy is commonly treated as the amount of uncertainty, whereas the concept of the entropy is used to quantify the uncertainty of the motion direction of every pixel in each frame. Our approach is to categorize the velocity of each pixel to five bins indicating individually the upward (bin_1), rightward (bin_2), downward (bin_3), leftward velocity bins (bin_4), and the quasi-steady bin (bin_5). Thus, a motion histogram with five bins is obtained. Using the histogram, the feature of MDE is formed as $f_{16} = \sum_{b=1}^{5} p_b \ln(p_b)$, where $p_b = \frac{bin_b}{\sum_{k=1}^{5} bin_k}$. Moreover, for the same MDE value in different scenes, the aesthetic quality could be diverse. For example, when shooting videos for a sport game or kids playing, the motion entropy could be larger than that of shooting far mountains. Therefore, it is semantic dependent.

Table 1. This table shows the number of semantic features among motion, colour, lightness, and composition, respectively.

Type	Motion	Colour	Lightness	Comp.
Dependent	1	4	2	0
Independent	5	7	0	3

Colour Saturation and Value: In HSV space, there are saturation and value to be considered; thus we compute the average saturation and value for each frame as additional colour features (f_{17}, f_{18}). According to the region of attention, the centre block of picture is distinct to others, and thus we add two more colour features (f_{19}, f_{20}) by averaging the saturation and value of the centre block.

Lightness: Here we define one lightness feature as the lightness ratio (f_{21}) of subject and background without subject. The other feature is described as the lightness ratio (f_{22}) of subject and whole frame. For f_{21}, exclusion of the subject in computing background lightness is needed since the subject lightness may be so strong as to influence the whole frame. However, f_{22} used in [5] is still necessary for video without subjects or with too many subjects because the exclusion for subject lightness in such a video may leave only noise.

Altogether, the size of each category is listed in Table. 1, and, in total, we have 22 scalar features that can be selected in each frame, i.e. $n = 22$.

3. EXPERIMENTAL RESULTS

The dataset collected by [6] is used for evaluation. This dataset consists of 160 videos with 15 seconds short-segment, and each video was rated by two authors on a 5-point scale. To evaluate our work fairly, we follow the same experimental settings of [6]: using 5-fold cross-validation and repeating it 200 times to obtain assessment accuracy. According to their work, seven most discriminative features are selected to avoid overfitting, and support vector machine (SVM) with radial basis function (RBF) kernel is also adopted. Furthermore, the

Table 2. The assessing accuracy (in %) of each paired-categories among four different feature types where I. and D. indicate independent and dependent respectively.

D. \ I.	Motion	Colour	Lightness	Comp.
Motion	72 ± 1.3	63 ± 2.1	$\mathbf{60 \pm 2.3}$	$\mathbf{66 \pm 1.5}$
Colour	$\mathbf{74 \pm 1.4}$	64 ± 2.5	56 ± 2.2	60 ± 2.2
Lightness	71 ± 1.7	$\mathbf{70 \pm 2.1}$	59 ± 2.4	65 ± 2.4
Comp.	72 ± 1.3	63 ± 2.7	N/A	65 ± 2.5

salient objects segmented by [10] with optical flow [8] are used as our subject. The criteria of the most discriminative features are listed below:

1. Find a feature that best classifies all videos.

2. Pick another feature which has the best classification accuracy jointed with the previously selected features.

3. Repeat step 2 until the number of selected feature is reached to M ($M = 7$ as described above).

To investigate the usability of semantic-independent features in conjunction with semantic-dependent ones, we show the experimental results in Table 2 where each of its entry is the accuracy obtained by combining the two types of features. Eg., the I-Motion/D-Motion entry is obtained by the features $\{f_1, f_2, f_3, f_4, f_5\} \cup \{f_{16}\}$, and the I-Lightness/D-Motion entry is obtained by the features $\{\ \} \cup \{f_{16}\}$. Since the previous work did not separate features with regard to their semantic property, they only consider part of the values in Table 2, i.e. the diagonal values. Further from this table we discover some complementary features, eg., I-motion with D-colour, since their jointed accuracy outperforms motion features itself. We conclude that instead of generating many but redundant features, we make features complementary to enhance the performance of assessment.

In order to distinguish the assessment performance between semantically dependent and independent features, we report the experimental results in Table. 3. It can be seen that, the semantically independent feature exceeds the semantically dependent feature in terms of aesthetic quality assessment. Finally, we use the total 22 features for assessment. To show the improvement using our method, two works are carried out for comparison: Moorthy's method [6] and Luo's method [5]. The assessment performances are $75 \pm 1.1\%$, $73 \pm 2.0\%$ and 54% in our method, Moorthy's method and Luo's method, respectively,where we report the experimental results as shown in [6]. As the result shows, our method learns a more robust performance, achieving the state-of-the-art assessment performance.

4. CONCLUSIONS

In this paper, we present a classification framework to tackle the problem of assessing video quality using the features: motion, colour, lightness, and composition. By addressing motion information and temporal property, we come up with effective motion features: MS, MDE, and HS. Combining the newly explored features with the existing features to learn an assessment model by SVM, we achieve the state-of-the-art performance which has $75 \pm 1.1\%$ accuracy. By investigating the properties of our features, we find out that some features are content-independent. Hence, we categorize semantically dependent and independent features by separating the influence of video contents on AQ. Finally, from the experimental results, semantic-independent features show more promising performance than that of semantic-dependent features.

Table 3. The assessing accuracy (in %) of semantic dependent/independent features.

semantic-dependent	semantic-independent
$69 \pm 2.2\%$	$\mathbf{74 \pm 1.5\%}$

5. ACKNOWLEDGEMENTS

This paper is supported in part by the National Science Council, Taiwan, R.O.C. under the grants NSC98-2221-E-001-012-MY3 and NSC100-2631-H-001-013.

6. REFERENCES

[1] B. Krages, *Photography: The Art of Composition*, Allworth Press, 2005.

[2] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *CVPR*, 2006.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*, 2006.

[4] C.D. Cerosaletti and A.C. Loui, "Measuring the perceived aesthetic quality of photographic images," in *QoMEx*, 2009.

[5] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *ECCV*, 2008.

[6] A. K. Moorthy, P. Obrador, and N. Oliver, "Towards computational models of visual aesthetic appeal of consumer videos," in *ECCV*, 2010.

[7] W. Yan and M. Kankanhalli, "Detection and removal of lighting & shaking artifacts in home videos," in *ACM Multimedia*, 2002.

[8] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," *Doctoral Thesis. Massachusetts Institute of Technology*, 2009.

[9] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.Q. Xu, "Color harmonization," *ACM Trans. Graph.*, vol. 25, no. 3, 2006.

[10] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *ECCV*, 2010.