Video Aesthetic Quality Assessment by Temporal Integration of Photo- and Motion-Based Features

Hsin-Ho Yeh, Chun-Yu Yang, Ming-Sui Lee, and Chu-Song Chen

***Abstract*—This paper presents a new method for accessing the aesthetic quality of videos. It consists of two processes: aesthetic features construction and temporal integration. First, our method combines both photo-based and motion-based visual clues to extract the aesthetic features for each frame in a video. We introduce new motion-based features built from optical flow and salient region extraction, and show their effectiveness to enhance the estimation of aesthetic values. Then, a temporal-order-aware framework that integrates the frame-based features is presented to further improve the evaluation accuracy by taking the time-varying properties into consideration. The experimental results demonstrate that our approach can accomplish remarkable improvement for aesthetic quality assessment of videos.**

***Index Terms*—Aesthetic quality assessment, computational aesthetics, image quality assessment, video quality assessment, visual aesthetics.**

I. Introduction

As high-definition (HD) devices become popular, it is observed that the video qualities are scarcely degraded even undergoing some processes such as compression and restoration. The perceived video quality, therefore, does not heavily rely on measuring how much quantity a video degrades [1]–[3]. Instead, judging a video's quality would be focused more on the *attractiveness*. According to [4], the attractiveness could be affected by both *originality* and *aesthetics*. Originality is hard to interpret since it varies mostly from personal preference; however, many people agree that aesthetics can be judged from professional opinions. For example, continuous shaking of a film, deficient lighting on a target or even an obscure subject might lower the aesthetic appeal. On the other hand, a film carrying high aesthetic characteristics such as bright color or delicate composition catches the audience's eyes, as shown in Fig. 1.

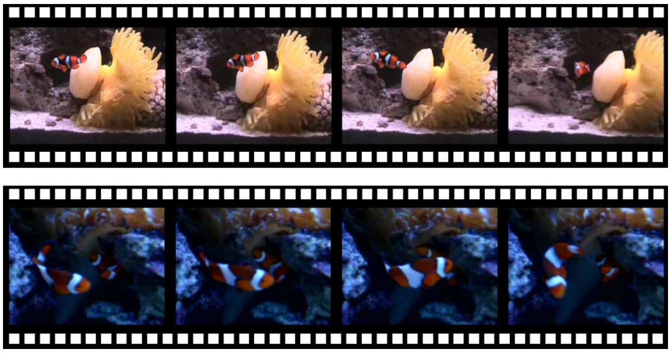


Fig. 1. Two videos from a benchmark dataset [16] for aesthetic quality assessment. In comparison of them, the upper video attracts the audience's interest more than the lower one. Our work seeks to explore the aesthetic features and an integration framework that can automatically reflect the audience's interest of aesthetics.

As an aspect of human nature, aesthetic experiences have been studied in neural science. In Ramachandran and Hirstein [5], a theory of human artistic experience and the associated neural mechanisms has been presented. One of their findings suggests that grouping is a basic principle to human's artistic experience. Different visual areas in the limbic system extract correlations in different domains such as shape and color. Kawabata and Zeki [6] conducted functional-MRI experiments to verify whether there are associated brain areas when viewing beautiful paintings, regardless of the category of painting. Their results show a distinction between two different kinds of brain activities; first, activity associated with a particular stimulus type (eg., color or motion) engages in a specific area of human brain, which reveals their isolation in the early vision process; second, both beautiful and ugly stimuli modulate activity in the same cortical area(s).

Inspired by the results of cognitive neural science, Peters [7] provides six factors for visual aesthetic impression, including color, form (shape), motion, spatial layout, depth, and human body. Since these factors/features could be extracted from images, recent efforts have been put in assessing aesthetics automatically from images via computational or related methods. Among them, learning an aesthetic-value predictor by data mining or machine learning techniques from a large data collection has received much attention in many fields such as cultural analytics [8], computer vision [4], [9]–[11], signal/image processing [12], [13], and multimedia analysis [14], [15]. The learned predictor can help select the video clips and also assist the users to refine their shooting techniques from the judgement feedback.

Learning from professionals is a typical approach to achieve automatic aesthetic-quality (AQ) assessment of photos. Pro-

Manuscript received February 06, 2012; revised July 02, 2012, November 08, 2012, and March 12, 2013; accepted May 06, 2013. Date of publication August 30, 2013; date of current version November 13, 2013. This work was supported in part by the National Science Council of Taiwan under Grant NSC101-2221-E-001-015-MY2. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Weisi Lin.

H.-H. Yeh and C.-Y. Yang are with the Institute of Information Science, Academia Sinica, Taipei, Taiwan (e-mail: hhyeh@iis.sinica.edu.tw; cyyang@iis.sinica.edu.tw).

M.-S. Lee is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan (e-mail: mslee@csie.ntu.edu.tw).

C.-S. Chen is with both the Institute of Information Science and Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan (e-mail: song@iis.sinica.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2280250

fessional photographers adopt special techniques to make their photos appealing from the aesthetic view, e.g., the DOF (depth of field) difference between foreground and background, and the rule-of-third for composition. In a similar way, machine could be able to estimate the AQ of videos (VAQ) based on the rules induced by professionals. However, a challenge for automatic VAQ assessment is that, unlike photos, there is still a lack of well-studied techniques for VAQ enhancement from the professional side.

In this paper, we aim to design an automatic VAQ assessor by integrating the aesthetic features (AFs) in the temporal domain. We design new motion-related AFs and a temporal-variation-aware integration framework. From the experimental results, the proposed assessor significantly improves the prediction accuracy of VAQ. The rest of this paper is organized as follows: The assessment of AQ for photos and videos are reviewed in Section II. In Sections III–V, we introduce the AFs employed and the temporal integration framework developed. We evaluate the proposed method in Sections VI–VII. Finally, conclusions are given in Section VIII.

II. Related Works

AQ assessment for a single photo has been studied for a long time [4], [9]–[11], [14], [15] in either the AFs extraction or the classifier design. Nevertheless, it is not until more recently that some works [16], [17] began to study the automatic VAQ assessment. For completeness, we review the literatures on photo-based AQ in Section II-A and VAQ in Section II-B, respectively.

A. Assessing Photo-Based Aesthetic Quality

There have been numerous photo collections over the Internet since people usually like to share their photos and make them more visible. Hence, photo-sharing websites are the best choice for the researches of photo-based AQ assessment according to [11]. One of the commonly known websites is Photo.net,[1] where more than one million photos are uploaded, and each photo is reviewed in terms of two types of qualities, aesthetic and originality. The qualities are measured by the scores ranging from one to seven, and higher value indicates better quality. Another popular website is DPChallenge,[2] where each photo is rated on the overall quality such as color harmony, composition, and contrast, and the ratings range from one to ten. Photos and ratings in these websites demonstrate the useful shooting techniques and visual properties of professional photographers.

Datta *et al.* [4] analyze certain visual properties which judge the AQ of a photo, and embody them into several kinds of AFs including light, color, exposure, and composition. In this rule, photographers place their subject on one of the four intersecting points of the imaginary lines which equally partition the image into three parts along both of the horizontal and vertical directions. As shown in Fig. 2(a), the cyclist was placed at the lower-left imaginary intersecting point (also shown in the upper-right point of Fig. 9(a)) which leaves space for better imagination of the audience; as a result, the reviewers of photo.net responded by giving a higher score to Fig. 2(a) than Fig. 2(e) on average.

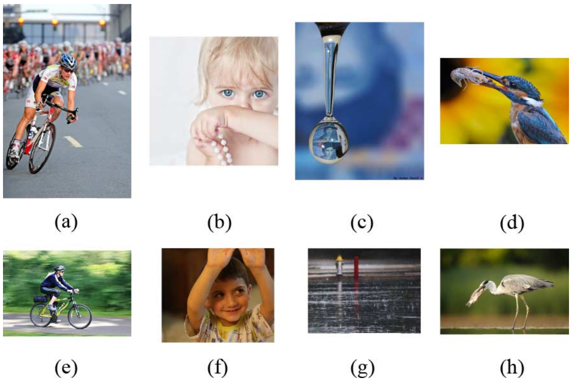


Fig. 2. Eight photos collected from photo.net. In each column, the photos shown on the top have higher aesthetic scores than those shown in the bottom. Among these photos, several special shooting techniques are applied to make them more appealing than the others. These several shooting techniques include rule of third (a), light (b), simplicity (c), and color (d).

Also, Datta *et al.* regard light exposure as a good way to discriminate high quality photographs from poor ones. Figs. 2(b) and 2(f) are different in lighting conditions obtained by tuning the camera aperture. Many photographers discussed about the light settings of these two photos in the forum of photo.net and it turned out that the setting of Fig. 2(b) is more adorable. Besides these visual properties, they also learn a classifier from the paired training data (AFs and ratings) to predict the aesthetic value of a photo.

Ke *et al.* [9] study a set of AFs from color, edge distribution, and composition. In this work, they assume that the professional photographers compose their photos with simplicities including out-of-focus background, color contrast, and lighting contrast. For example, by shooting with background out of focus in Fig. 2(c), the subject, i.e., the water drop, leaves a clear image in audience's brain since the background is blurred. This shooting technique is strongly related to depth of field (DOF). In another example, Fig. 2(g) has no clear separation between the subject and the background, and the raindrops distract the audience. Therefore, Fig. 2(c) has higher simplicity than Fig. 2(g).

Later, Li *et al.* [14] combine face information and the AFs to determine the AQ of images. They believe that face is one of the aesthetic-related factors in consumer photograph albums. Furthermore, they recommend several editing guidelines to make photos more attractive. Yeh *et al.* [15] come up with the AFs similar to previous works but more effective in measuring AQ with the help of the learning-to-rank principle. Instead of judging whether the quality of a photo should be good or bad, learning-to-rank is capable of providing more details such as evaluating the best AQ in a set of photographs. In Yeh *et al.* work, ListNet [18] is applied to rank photographs. ListNet, a learning-to-rank approach which is constructed from the loss in the list level, can better catch personal preference instead of photo-wise preference.

Recently, high level image attributes are also utilized to represent the visual quality of an image. These attributes according to Dhar *et al.* [11] are categorized into composition, content, and

[1] http://photo.net/

[2] http://www.dpchallenge.com

sky illumination attributes. Moreover, each high-level attribute gives a semantic label by a describable attribute classifier. The AQ of an image is assessed by training a support vector machine (SVM) classifier with their high-level attributes. On the other hand, Nishiyama *et al.* [10] observe that the harmony of color in a photo is a key factor to determine the visual quality. For example, Figs. 2(d) and 2(h) show different aesthetic attractiveness toward the audience although they have a similar composition. In detail, Fig. 2(d) presents much more harmony than Fig. 2(h) in terms of color, and Fig. 2(d) gains higher preference than Fig. 2(h) according to the statistics from photo.net. Nishiyama *et al.* think that the existing methods of color harmonization are insufficient to judge the AQ. Therefore, they design *bags-of-color-patterns* to tackle the previous color harmonic limitations and achieve nearly 80% accuracy on classification of images from DPChallenge.

B. Assessing Video-Based Aesthetic Quality (VAQ)

Unlike the photo aesthetics which have a plenty of skills and rules conducted from professionals for amateurs to follow, there are still no such apparent rules applicable for video aesthetics. Strictly speaking, photo-based AFs can be applied to each frame of a video, and by doing so we could construct an assessor reflecting several aspects of the VAQ. However, as a video is composed of successive frames, we are still not aware of how the presentation of photo-based features in the temporal domain affects a video's attractiveness. Unfortunately, there are scarcely any clear rules summarized so far for video aesthetics based on the motion and temporal properties from the professional side.

In addition, unlike photos that have many reference sites for showing shooting skills (such as photo.net and DPChallenge), well-gathered publicly benchmark datasets for video aesthetics remain unavailable until recently [16]. Due to the above reasons, there are yet few works focusing on the automatic VAQ assessment.

In previous studies, Luo *et al.* [17] attempt to use both photo-based AFs and motion features for judging VAQ. They observed that in a professional image, the subject region should always be focused and the background should be out of focus instead. Thus, they follow the principle that the VAQ should be focused on the subject because it gathers most visual attention within the whole image. As a result, the subject-based AFs extracted from these regions serve as a criterion for assessing photo/video AQ. By assuming the subjects being clearly extracted, this approach further introduces two motion-based features: *length of subject region motion* and its *motion stability*. However, this approach is designed to suit videos with clearly focused subjects, indicating the subject has to be tracked using a narrow-DOF camera. Hence, this approach is particularly appropriate for professional films instead of consumer videos, since most videos captured by amateur filmmaker do not necessarily have subjects focused and backgrounds blurred.

In [16], Moorthy *et al.* introduces an approach to classify the aesthetic appeal/unappeal based on three kinds of features, including image-quality features (such as SSIM-based frame-rate estimation, blocking effect), motion features (such as motion ratio and size ratio), and single-photo aesthetic features (such as sharpness/focus, color, luminance, and rule-of-third). They extract these features in the frame level at first. Then, these features can be pooled in the microshot level, where various features can be composed by the combination of the frame-based features and a few pooling operators. Finally the mean or variance of the feature values are computed for all microshots in a video, and a feature-selection process is applied to conduct a low-dimensional feature vector for classification.

Moorthy *et al.* [16] have collected ratings on the aesthetic value for a dataset containing 160 consumer videos, which as we know is the only VAQ-assessment dataset publicly available, referred to as the **Telefonica** dataset in this paper. They have shown that, by designing a hierarchical pooling approach introduced above, 73% classification accuracy can be achieved by selecting 7 most discriminant features. In this paper, we will show that our approach can boost the result to 79% by selecting the same number of discriminant features for the same dataset. Furthermore, since the cross-validation errors are shown in [16] without applying new testing data, we enhance the **Telefonica** dataset by adding high aesthetic-appeal cinema movies and demonstrate also the testing performance of our approach.

Since VAQ is useful for applications such as video broadcasting and recommendation, a recent study [19] has adopted VAQ assessment as a building block for an on-line mobile video sharing system to help construct interesting mashup of live videos.

III. Overview of Our Approach

We give an overview of our approach in this section. Hereafter we follow the notation convention: If $\mathbf{A}$ is a set, $|\mathbf{A}|$ is its cardinality, the number of elements in $\mathbf{A}$.

Let $\mathbf{V} = \{V_i \mid i = 1, \ldots, |\mathbf{V}|\}$ be a set of videos, where each video is associated with a label $y_i \in \{0, 1\}$ representing its aesthetic value, say, '1' for appealing and '0' for unappealing, respectively. In our approach, a video $V \in \mathbf{V}$ is separated into several non-overlapping 'snippets' and each snippet contains consecutive frames in a short period (usually one second) of the video V. Snippet serves as the basic processing unit in our approach because the aesthetic value can usually not be reflected by only a single image frame, but requires the combination of more frames to form a relatively meaningful and reliable measurement.

We conduct a two-stage approach for aesthetic quality assessment of a video. First, for each snippet, we choose several features called *snipped-based features* (SBF). Since a snippet (or microshot) is composed of the video frames in a short period of time, the SBF can reflect the fundamental aesthetic degree or visual clue in a small time interval. Second, we perform a *temporal integration* to combine the SBFs into a video-based feature vector, and perform a binary classification for it by using SVM to predict the video's aesthetic value ('1' or '0').

To construct the SBFs, we will extract AFs for each frame in a snippet at first. In principle, frame-based features only perform transient evaluations of a video's aesthetic value, which could be noisy and unstable. Hence, these frame-based features are aggregated to form the SBFs by some pooling operators (eg., *min*, *median*, and *max*) that capture overall values reflecting the aesthetic appeals in a snippet.

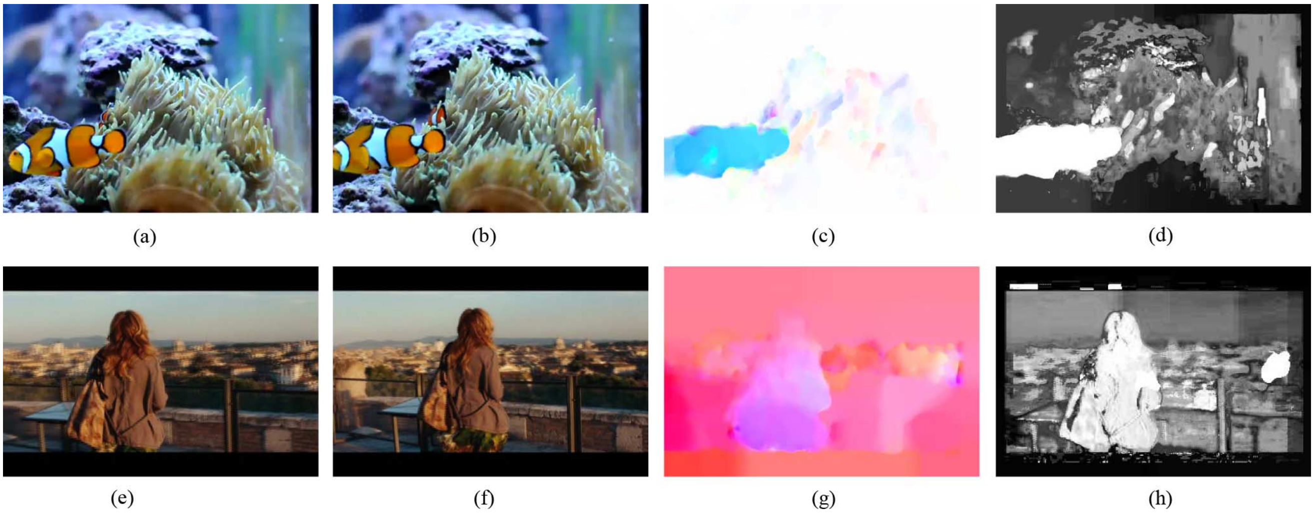


Fig. 3. Examples of the optical flows and saliency maps obtained. (a) and (b) are two adjacent frames in a static-camera video from the Telefonica dataset. The optical flows obtained are plotted in (c), where the color saturation for a pixel is proportional to its motion magnitude. The video-based motion saliency result is shown in (d), where the brighter pixel indicates the higher saliency. In addition, (e) and (f) are two adjacent frames in a moving-camera video from the movie "Eat Pray Love (2010)". The optical flows obtained are shown in (g). Even though the optical flows seem more complex, the video-based saliency detection method [20] can still compute the saliency map well as shown in (h).

In the following, we will present our frame-level features in Section IV, SBF construction method in Section V-A, and the temporal integration method in Section V-B, respectively.

IV. Frame-Based Aesthetic Feature Design

In this section, we focus on the frame-based features deployed in this work. The frame-based features contain two parts. The first part is the *motion-related features*, and the other part is the *photo-related features*. The former is unique for VAQ assessment since motion is the most distinct characteristic making the associated features different from photo-based ones.

A. Preprocessing for Motion-Related Features

In our approach, two pre-processing steps serve as prerequisite procedures to build the proposed motion-based AFs for each frame. Suppose each video V is divided into k snippets, $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$, and each snippet $S \in \mathbf{S}$ is composed of $|S|$ frames $f_1, f_2, \ldots, f_{|S|}$. For each frame, the following preprocessing steps are performed to extract motion and saliency information for the AF construction as follows.

1) Motion-Extraction Preprocessing: We employ the pixel-based motion estimation as a preprocessing-step for constructing the AFs. To compute the motion vector per pixel, we adopt the optical-flow technique developed by Liu *et al.* [21], which measures the motion flows between two adjacent frames under the brightness constancy assumption.

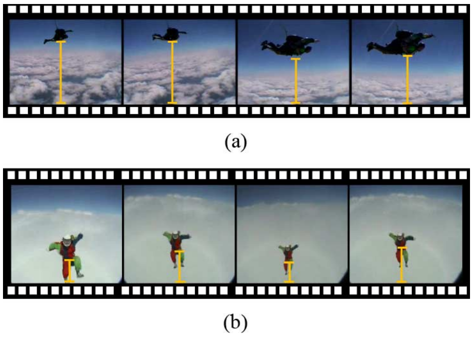
2) Saliency-Based Preprocessing: Visual saliency [20], [22], [23] aims to capture some part of an image which attracts most of our visual attention. For this reason, it is regarded helpful for automatic aesthetic estimation. In the past, single-image-based saliency estimation has been adopted for assessing AQ of photos [14], [24]. In this paper, we employ a video-based saliency region detection method [20] that can provide more reliable results than image-based methods. This method estimates video saliency by using both color and motion information. More specifically, by giving a rectangle composed of a core region and a border region, the saliency of a pixel is estimated as the ratio of the number of color/motion similar pixels in the core region to those in this rectangle in this method.

For example, Figs. 3(a) and 3(b) are two adjacent frames from a static-camera video. The motion flows and the saliency map are shown in Figs. 3(c) and 3(d), respectively. In the other panning-camera example, the motion flows generated are more scattered as shown in Fig. 3(g). Nevertheless, the video-based saliency-detection method [20] we adopted can still detect the saliency region well in Fig. 3(h) by jointly considering color and motion information.

Both the foregrounds (extracted by salient regions) and motion vectors (evaluated by optical flow) serve as essential visual clues for building the motion-based AFs introduced below. In addition, we binarize the video-based saliency map obtained for each frame, and then extract the largest connected component of the binarized map and refer it to as the *foreground* region. The threshold used for binarization is set to preserve the top $1/32$ strength of the saliency map. We do not extract multiple subjects from the saliency map nor separate the foreground region into objects further. This approach simplifies the complexity of segmenting multiple meaningful subjects in an image frame.

B. Design of Motion-Related Features

Among all the aesthetic characters in a video, motion is the most distinctive character for videos compared to photos. In contrast to the previous work, we further propose to use some motion-based features which are of great significance. The first is *motion space*, which reflects the aesthetic quality that tells a camera how to trace a moving object. The second is *motion direction entropy*, which measures the dispersed degree of motion that is also relevant to VAQ. Other motion-related AFs include *hand-shaking* and *shooting type*. Details are given in the following.


(a)

(b)

Fig. 4. Illustration of the situations for more/less imagination in videos. Frames from left to right represent a sequence of videotaping a downward moving object. The yellow segment shows the gap of the subject to the downside frame border, which represents the preserved space. From the frames in (a), the free space preserved for the skydiver gives audience more imagination. On the other hand, the position of the skydiver in (b) gives the audience less imagination due to the lack of space.

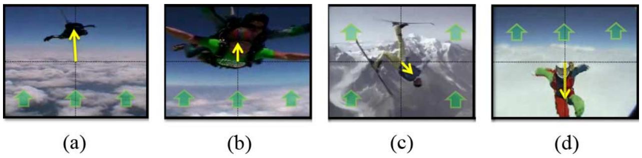

(a) (b) (c) (d)

Fig. 5. 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 values 0.82, 0.47, -0.26, -0.79 respectively.

Motion Space (MS): According to professional skills in film making, as cameramen are videotaping a moving subject, they must beware that the space in front of the moving direction of the subject should be reserved for better VAQ, since the reserved space provides the audience more imagination about the subject motion. For example, when a downward moving object as shown in Fig. 4 is videotaped, the subject looks not moving because the camera moves with the same velocity (and only the background looks moving upward). The gaps (yellow segments) shown in Figs. 4(a) and 4(b) reflect the space reserved when shooting these two videos, where the former that reserves sufficient space will provide audiences better aesthetic appeal.

Although MS is meaningful to the degree of aesthetically pleasing when capturing moving objects, it has not been well employed in previous approaches. Therefore, as illustrated in Fig. 5, we tackle this problem by setting the direction of the averaged optical flow of the entire image as v and the vector between the foreground center and image center as d. The feature of MS is represented as $f^1 =< v, d >$, where $< \cdot, \cdot >$ is the inner product.

Motion Direction Entropy (MDE): Another motion-based feature related to aesthetics is the diversity of motion directions. Variation in motion direction is a characteristic that cannot be reflected only by individual photos; however, this visual clue is neglected in previous studies.

To measure the variation, entropy is used to quantify the uncertainty of the motion direction of every pixel in each frame. In the implementation, we employ the optical flow results obtained in the preprocessing step, where a velocity vector is computed for each pixel. Our approach simply computes the velocity directions of pixels and votes them into five predefined bins that are respectively the upward (bin_1), rightward (bin_2), downward (bin_3), leftward (bin_4), and quasi-steady bins (bin_5), where bin_5 accumulates the pixels with the velocities smaller than a threshold. $bin_1 \sim bin_4$ accumulate the pixels with the velocities larger than the threshold for the velocity directions within $(45^\circ \sim 135^\circ)$, $(-45^\circ \sim 45^\circ)$, $(-45^\circ \sim -135^\circ)$, and $(-135^\circ \sim 135^\circ)$, respectively. In our implementation, the threshold is set as 0.1. A motion histogram is then formed as $f^2 = -\sum_{b=1}^{5} p_b \ln(p_b)$, where $p_b = bin_b / \sum_{k=1}^{5} bin_k$. As we know, this feature has not been used before for VAQ assessment either.

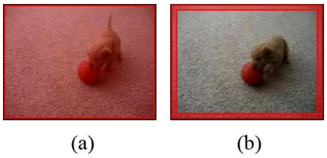

(a) (b)

Fig. 6. If the stability is computed over the whole frame like the region shaded in red of (a), a self-shaking subject such as the ball-playing puppy may cause false alarms. In contrast, the modified detection region such as the area shaded in red in (b) along the borders helps detect frame-based unstableness in a more robust manner.

Hand Shaking (HS): Although many cameras have been equipped with anti-shaking, 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 VAQ. Shaking differs from the above features as it is obtained by computing the change of motion directions based on the optical flow. As there are already many methods ([25], [26]) for HS estimation in video stabilization, we simply implement the following method for computational efficiency. We set shaking detection area at the border which can better distinguish subject's self-shaking from the hand-shaking, as shown in Fig. 6. The horizontal motion of the border region in frame j is defined as mx_j, which is the average of the horizontal projections of the optical flow in the region, and its direction is $I_j = sign(mx_j)$. An exclusive-or operator $(\oplus)$ is adopted to model the change of directions, and the horizontal unstableness feature f^3 is defined as:

$$f^3 = (I_j \oplus I_{j-1}) \times (|mx_j| + |mx_{j-1}|), \tag{1}$$

where $| \cdot |$ denotes the magnitude of motion. In the same way, a vertical unstableness feature f^4 is formed by replacing mx with my. In our implementation, the border is $1/30$ frame-width pixels.

Shooting Type (ST): We further define the border to central unstableness ratio for horizontal motion $f^5 = Ub + \epsilon / Uc + \epsilon$ to show the subtle differences, where ϵ is a small positive constant to avoid the situation of division by zero. Ub is the horizontal motion of the frame border, and Uc is the horizontal motion of the frame except for the border; they are both computed from the optical flow.

In our implementation, the border has $1/30$ frame-width pixels. For border moving faster than the centre type ($f^5 > 1$), it may be the recording type that the shot traces and focuses on the subject as shown in Fig. 7(a). For border moving equally with the centre type ($f^5 = 1$), it may be a panorama shooting

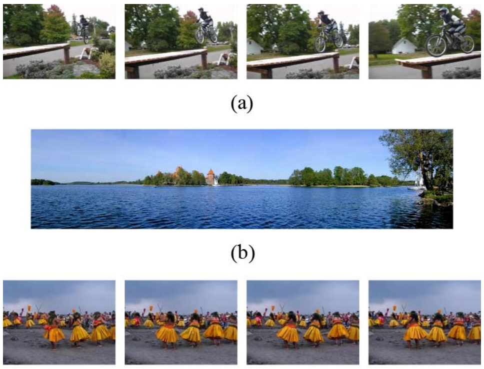


Fig. 7. Examples are given to illustrate the shooting types that we captured. (a) shows the tracking scenario where the border movement is faster than the central one. (b) indicates the panorama shooting that the movements of the central and border are equal. (c) shows the deformable subject in a static shot where the central movement is the most salient.

as shown in Fig. 7(b). Finally for border moving slower than the centre type ($f^5 < 1$), it may be regarded as a static shot on a deformable subject as illustrated in Fig. 7(c).

Likewise, f^6 is defined for the vertical direction. As we know, these features have not been used for VAQ assessment before.

C. Discussion of Our Motion-Based Features

Compared the motion-related features designed in [17] which are only suitable for shots containing clearly focused foreground and out-of-focus background, our motion-based features does not have this limitation and can be applied for a more general class of videos.

In addition, compared to the features designed in [16], we do not employ any image-quality-measurement features such as SSIM-based frame rate and blocking effect estimation; hence, our approach would be also applicable to the datasets containing only high-image-quality videos. Besides, unlike the approach in [16] that finds the foreground by block-based motion estimation, we employ the technique of salient-region estimation for foreground extraction and optical flow for motion evaluation, which can also deal with the case of slow-motion and non-rigid objects. Beyond the simple motion features such as the speed and size of a moving object, our motion-based features including motion space, motion direction entropy, hand-shaking and shooting type are more diverse and versatile. We will show in the experimental section that our designed features can perform better under the same pooling-operators and temporal-integration settings.

D. Adoption of Photo-Based Features

In addition to motion-related features, we also employed conventional photo-based features, categorized into color, composition, and lightness, to reflect the AQ based on a single frame. They are briefly reviewed as follows. As these features are commonly employed for photo aesthetics, we simply give concise introductions of them and the readers could seek to the related references for more details.

Color Harmonization: Human visual perception of aesthetics is strongly affected by color harmonization [10], [11], [27], [28]. We employ the *hue* histogram ($\mathbf{h}$) constructed from HSV color space for each frame to distinguish seven well-arranged color templates $\mathbf{Tp} = [\mathbf{Tp}_1, \ldots, \mathbf{Tp}_7]$ introduced in [28] and shown in Fig. 8.

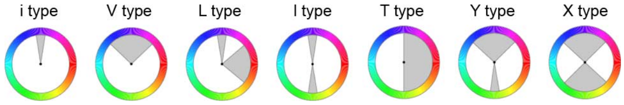


Fig. 8. Seven harmonic templates on the hue channel used in this paper are plotted. The harmonic templates are first defined in Cohen-Or *et al.* [28], we use seven of them to construct our color harmonization AFs. See [28] for more detail.

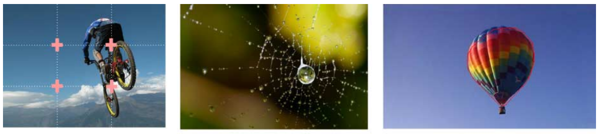


Fig. 9. These pictures selected from photo.net are to expose the importance of (a) rule of third, (b) clarity contrast, and (c) shape convexity in determining the aesthetic quality.

We compute the maximum response to each color template and conduct a seven-dimensional vector defined as:

$$L(\mathbf{h}) = [\max(\mathbf{h} * \mathbf{Tp}_1), \ldots, \max(\mathbf{h} * \mathbf{Tp}_7)]^{\mathrm{T}}, \qquad (2)$$

where $*$ is the convolution operator. Then, the distances between the photo and color templates in the vector space are calculated as our color harmonization AF:

$$f^{6+c} = \|L(\mathbf{h}) - L(\mathbf{Tp_c})\|_2, \forall c \in \{1, \ldots, 7\}, \qquad (3)$$

where $\|.\|_2$ indicates ℓ_2 norm, which defines the AFs from f^7 to f^{13}.

Color Saturation and Value: We compute the average *saturation* and *value* in the HSV space for the whole frame as additional color features (f^{14}, f^{15}). In addition, according to the region of attention, the central part of a picture is distinct to others. Hence we also add two more color features (f^{16}, f^{17}) by averaging the saturation and value of the central part only.

Composition: The composition of a frame is also of great importance for aesthetic, e.g., the rule of third [29], clarity contrast [17], and shape convexity [17]. The rule of third claims that the foreground should be placed in one of four intersecting points as plotted in Fig. 9(a). Clarity contrast for the foreground and a blurred background is notable feature in a professional image. The shape convexity according to [17] hypothesizes that the convex shapes are more likely to please humans. For example, in Fig. 9(c), the peripheral of hot air balloon circles a convex region, which is emphasized by the red line. In computing rule of third feature (f^{18}), clarity contrast feature (f^{19}) and shape convexity feature (f^{20}), we adopt the methods in [17] based on the foreground obtained by the preprocessing step of saliency detection.

Lightness: Here we define one lightness feature as the *lightness* ratio of foreground and background (f^{21}) based on the HSL space. The other feature is described as the lightness ratio of the foreground and whole frame (f^{22}).

V. SBF Construction and Temporal-Order-Aware Integration

In association to our notation, we have $r = 22$ frame-based features to represent the aesthetic property according to motion, color, lightness, and composition. Among them, there are 6 motion-related features, i.e., $f^1 \sim f^6$, and 16 photo-based features, that is, $f^7 \sim f^{22}$. They are denoted as the set $F = \{f^i | i = 1 \ldots 22\}$.

A. Snippet-Based Features

The above features are constructed for each frame. Since our work is video-based, the processing unit is a video snippet containing frames in a short period of time. In our implementation, a video snippet is set as one second. The frame-based features extracted are then pooled for a video snippet, summarizing an overall degree of aesthetic-related visual clues in a moment.

In our work, six pooling operators, $\mathbf{Pool} = \{\min, 1^{st}\mathit{quartile}, \mathit{mean}, \mathit{median}, 3^{rd}\mathit{quartile}, \max\}$ in [16] are also adopted for summarizing the frame-based features into snippet-based features (SBFs). Combining these six pooling operators with the 22 frame-based features, there are thus a total of 132 SBFs allowed to be selected in our approach. For example, the SBF of $(f^2, \min)$ means that the snipped feature is built by extracting f^2 (the Motion Direction Entropy (MDE) feature) for each frame at first, and then pooled by $\min$ (i.e., taking the minimum). We denote the Cartesian product $\mathbf{S} = \mathbf{F} \times \mathbf{Pool}$ as the set consisting of these SBFs.

B. Temporal-Variation-Aware Integration

How to integrate the SBFs is a main issue for VAQ estimation. Unlike previous work [16] that simply computes the mean and variance of the SBFs in a video for integration, we found that the change of SBFs with time will also affect the visual preference for audiences. For example, when seeing a video with constant color harmonic and the other with time-varying harmonics, the audience could be aware which is better, but their overall results obtained by simple pooling (such as mean, median and variance) of the SBFs remain the same. Hence, an effective temporal integration method should be able to distinguish temporal variation of a sequence of SBFs.

We develop a temporal-variation-aware (TVA) method to integrate the SBFs. Unlike the mean and variance which simply reflect the DC value and total energy of a signal, respectively, it is well known that auto-correlation function of a signal can capture more of the changes in the signal shapes. Inspired by this idea, we divide the video into non-overlapping segments and propose to use the correlation between segments to perform the TVA integration. To formulate the approach, let the k snippets in a video be divided into m segments (groups of snippets); each segment contains d successive snippets, and so $k = md$. Hence, each group can be characterized by a d-dimensional feature vector if the snippet feature type has been selected in $\mathbf{S}$.

When the type of SBF is given, there are m feature vectors, $\mathbf{z} = [\mathbf{z}_1^{\mathrm{T}}, \mathbf{z}_2^{\mathrm{T}}, \ldots, \mathbf{z}_m^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{z}_i \in \mathbb{R}^d$ $(i = 1 \ldots m)$. Let $K(\mathbf{z}_i, \mathbf{z}_j)$ be the kernel function that measures the similarity in some implicit mapping space. To reflect the i-th group's relevance, we are particularly interested in the average correlation between the i-th feature vector and the other feature vectors. It is then evaluated as follows:

$$\overline{op}_i(\mathbf{z}) = \frac{1}{m-1} \sum_{j=1\ldots m,\, j \neq i} K(\mathbf{z}_i, \mathbf{z}_j). \tag{4}$$

By doing so, we obtain m integration operators $\overline{op}_i(\mathbf{z})$, $i = 1 \ldots m$.

To construct the kernel, we choose the following function that is efficient to compute and easy to implement:

$$K(\mathbf{z}_i, \mathbf{z}_j) =< sgn(\mathbf{z}_i - \overline{\mathbf{z}}), sgn(\mathbf{z}_j - \overline{\mathbf{z}}) >, \tag{5}$$

where $sgn(.) \in \{-1, 1\}^d$ is the sign function, $< .,. >$ denotes the inner product, and $\overline{\mathbf{z}} = mean(\mathbf{z})$.

Despite the correlation among groups reflects the temporal-variation information, the long-term average information still matters in capturing the overall feature of a signal. Hence, in addition to the m operators defined above, we include further the $(m+1)^{th}$ and $(m+2)^{th}$ operators which are respectively the *mean* and *standard deviation*, denoted as $\overline{op}_{m+1}(\mathbf{z}) = \overline{\mathbf{z}}$ and $\overline{op}_{m+2}(\mathbf{z}) = std(\mathbf{z}) = 1/k \|\mathbf{z} - \overline{\mathbf{z}}\|_2$. The entire set of temporal integration operators employed in our approach is then defined as $\mathbf{OP} = \{\overline{op}_i | i = 1, \ldots, (m+2)\}$.

C. Selection of Key Dimensions in SBF × Operator Space

We have conducted three versions of the proposed methods for comparison. For the notation convenience, let **OP** be separated into two subsets, $\mathbf{TVA} = \{\overline{op}_1, \ldots, \overline{op}_m\}$ and $\mathbf{Sim} = \{\overline{op}_{m+1}, \overline{op}_{m+2}\}$, where $\mathbf{OP} = \mathbf{TVA} \cup \mathbf{Sim}$.

The first version, **AFSim**, selects the features from **S** × **Sim**; i.e., it integrates the SBFs by using only the simple average (or pooling) operators, *mean* and *standard-deviation* .

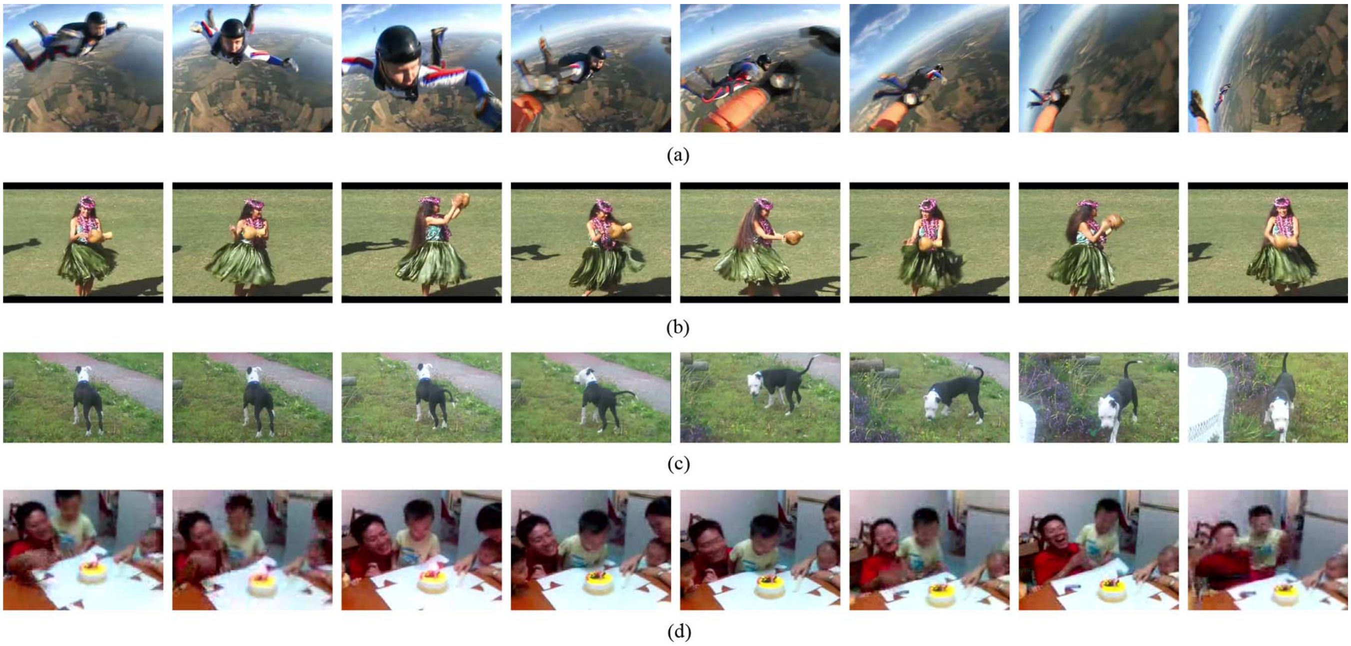
The second version, **AFTVA**, selects the features from **S** ×**TVA**; i.e., it uses the TVA operators to integrate the SBFs.

The third version, **TVASim**, selects the features from **S** ×**OP**; i.e., all operators $\{\overline{op}_1, \ldots, \overline{op}_{m+2}\}$ are allowed to be used for temporal integration.

The combination of SBFs and temporal-integration operators is very high-dimensional. For **AFSim**, there are 132×2 video-based features since there are two choices for the integration ($\overline{op}_{m+1}$ and $\overline{op}_{m+2}$); for **AFTVA** and **TVASim** , there are $132 \times m$ and $132 \times (m+2)$ video-based features, respectively. They require heavy computation-costs for either learning or assessment.

One possibility to reduce the dimension is to use subspace techniques such as principle component analysis. In this way, the reduced dimensions are linear combinations of the original ones, but it is more crucial to find out which SBFs play the most important roles for VAQ assessment.

Hence, instead of using all dimensions, we perform feature selection in the **S** × **Integration_operator** domain. By doing so, we can provide more insights on the influences of the individual feature types or their combinations. We employ the sequential forward-search method that is a greedy approach to pick the combination of SBF and integration-operator in turn. Note that this is also the feature selection approach suggested in [16].


(a)

(b)

(c)

(d)

Fig. 10. Four examples from the Telefonica dataset [16]: snapshots from skydiving (a), Hawaiian hula dancing (b), puppy playing with ball (c), and birthday candles kid (d). The subject rating scores of this dataset is between $[-2, 2]$. Among them, (a) and (b) are rated as high aesthetic quality (0.7591 and 0.8531, respectively), and (c) and (d) are rated as low (-1.1041 and -1.6164, respectively).

The procedure of sequential-forward-selection is as follows: (Here we abbreviate "temporal integration operator" as "operator")

1) Find a SBF-operator pair that best classifies all videos.
2) Pick another SBF-operator pair, which has the best accuracy jointed with the previously selected pairs.
3) Repeat 2) until the number of selections reaches a predefined positive integer q or when the accuracy is converged.

In our implementation, m is selected as 4 and the number of SBF-operator combinations selected are within $q < 18$. In each iteration, we use the SBF-operator pairs selected to build the associated video-based features. For example, combining the SBF of $(f^1, \mathtt{mean})$ with the temporal integration operator $\overline{op}_1$ means that we build the video-based feature by extracting f^1 (the motion-space (MS) feature) for each frame at first, pooling the MS features for the frames in a snippet by *mean* to obtain the SBF, and then perform the temporal operator $\overline{op}_1$ to integrate the values of all SBFs in a video.

Our approach accesses the aesthetic quality as 'low' or 'high' like [16]. The accuracy employed to evaluate the performance is defined as the number of mis-assessed data to that of the total data. We use SVM of radial-basis-function (RBF) kernel to train a classifier in each iteration based on the video-based features selected. The kernel parameters used are $(C, \gamma) = (5, 3.7)$ and are fixed for all experiments.

D. Connection to the Approach of[16]

Our framework shares some similar ideas with the classic framework of Moorthy *et al.* [16]. We now highlight the distinctions between them.

First, by employing the designed AFs and the integration operators in **SIM**, our simplest version of approach, **AFSim**, differs from [16] only in the AF parts. That is, except the distinction of the frame-based features introduces in Section IV, the other parts including pooling into SBF (called the microshot-level features in [16]) and further integration by mean or standard deviation are the same. This, therefore, gives us a nice foundation to compare the designed AFs to those of [16] under the same setting. Hereafter, we name the approach of [16] the **MoorthySim** approach. The comparison of the effectiveness of AFs of these two approaches will be reported in Section IV-A.

Then, our enhanced versions of approaches, **AFTVA** and **TVASim** integrate the SBFs with temporal-variation-aware operators instead of overall averaging (or pooling) operators. They can considerably enhance the performance because the temporal fluctuation of AFs is further employed. Results of them will be reported in Section IV-B.

VI. Experimental Results on Telefonica Dataset

The dataset constructed by Moorthy *et al.* [16], which we call the **Telefonica** dataset, consists of 160 rated videos. This is the only publicly available benchmark for VAQ to our knowledge. Moorthy *et al.* crawled consumer-generated-content from Youtube[3] and downloaded videos from 16 diverse categories, including baby laughing, sky diving, and so on. Each video is cropped as a 15-seconds clip to reduce the potential biases of video length. It is because that, according to the study of Pinson and Wolf [30], the memory effects of human for time-varying quality estimation are probably limited to 15 seconds. Some videos are shown in Fig. 10. Although each video has the same length, the number of frames in each videos could be different due to the different rates of frames per second (fps).

These videos have gone through a controlled user study that each participant (33 people in total) rated the videos on

[3] http://www.youtube.com

two scales: content and aesthetic to produce content-independent ground truth. These people had been involved in a short training session before the study began, and the videos were shown without audio for the elimination of the influence on the perceived audio quality [31].

The rating values of the **Telefonica** dataset range between $[-2, 2]$. Given a video, a mean opinion score (MOS) is generated by normalizing and averaging the rating values of all participants. Follow the setting of [16], we also sort these MOSs and choose the median value as a threshold to separate these 160 videos into two sets of equal size (80), positive (high quality) and negative (low quality). Note that it has also been observed in Datta *et al.* [4] and Joshi *et al.* [32] that the multi-valued aesthetic score is extremely subjective, making the ratings noisy and so predicting it via regression is still challenging. Instead, predicting only high- or low-quality for a video is more tractable.

A. Comparisons of Designed Features

In the first experiment, we demonstrate the effectiveness of the proposed AFs by comparing **AFSim** and **MoorthySim** because they are different only in the AFs as mentioned above. For a fair comparison, we also report the 5-fold cross-validation (CV) accuracy and repeat it 200 times to obtain the averaged accuracy. The number of features selected is $q = 7$ and the snippet period is one second. They are the same as those in [16].

We can see that by applying both methods to the **Telefonica** dataset, the assessment accuracies are $75.2 \pm 1.3\%$ by using **AFSim** and $73.0 \pm 2.0\%$ by using **MoorthySim**, respectively. Hence, not only the mean accuracy is improved but the uncertainty range (standard deviation) is also reduced. It demonstrates that the proposed AFs are more effective than those adopted in Moorthy *et al.* [16]. We further use Student's t-test to identify whether the performance improvement is significant. The statistic t value (t-stat) is computed by $(\overline{x} - \overline{y})/\sqrt{(\sigma_x^2/n_x + \sigma_y^2/n_y)} = 13.0431$, where $\overline{x}$ and $\overline{y}$ are the mean accuracies, σ_x^2 and σ_y^2 are the standard deviations of the accuracies, and $n_x = n_y = 200$ is the number of trials. The degree of freedom used for the t-test is $\min(n_x, n_y) - 1 = 199$. Under the significant level 0.05%, the critical t value (t-crit) is 3.3401 that is considerably smaller than t-stat $= 13.0431$.

Table I shows the 7 SBFs selected in turn. As can be seen, the motion-related features such as ST and HS have been chosen in the first three runs, which means that motion-based features are more critical for estimating the VAQ than photo-based features. Then, by further selecting four photo-based features in the categories of Harmonization, Saturation and Composition, the integration of these selected 7 SBFs achieves the best performance in the **AFSim** approach. To have more insight on the individual SBF selected, we show the accuracy when applying only each single SBF for the assessment as the brown bar in Fig. 11. The green bar shows how much the accuracy is increased when the SBFs are jointed with the previously selected ones. It reveals the significance of integrating the features even when the performance of individual feature is limited.

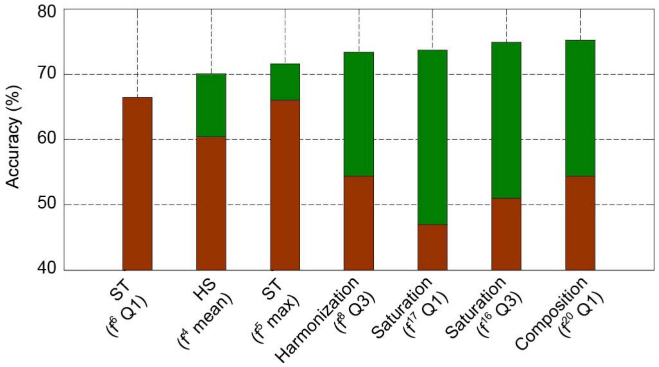


Fig. 11. The performance comparison of the SBFs chosen in AFSim (the green bar) and that of the individual SBF (the brown bar) are plotted. Y-axis indicates the VAQ performance in accuracy. The brown bar in x-axis demonstrates the accuracies of the used features respectively and the green bar shows the accuracies of the features added from left to right. From this result, it clearly shows that forward-selection procedure (the green bar) can enhance the accuracy consistently by selecting useful video-based features, regardless of the individual feature performance (the brown bar) being limited.

TABLE I

The Performance of AFSim, Where the top 7 SBFs Selected are Reported in This Simplify the Representation, we use Q1 and Q3 to Denote 1^{st} Quartile and 3^{rd} Quartile , Respectively. the First Column (#fea) Indicates the Order of the Video-Based Features Selected in AFSim. the Second to Fourth Columns Display the Chosen SBF, its Category, and the CV Accuracy

# fea	SBF		Category	Accuracy (%)
1	f^6	Q1	ST	66.4 ± 1.1
2	f^4	mean	HS	70.1 ± 1.3
3	f^5	max	ST	71.6 ± 1.1
4	f^8	Q3	Harmonization	73.4 ± 1.4
5	f^{17}	Q1	Saturation	73.7 ± 1.3
6	f^{16}	Q3	Saturation	74.9 ± 1.2
7	f^{20}	Q1	Composition	75.2 ± 1.3

B. Comparisons of Temporal Integrations

In the above, we have shown the performance improvement with the designed AFs by comparing the approaches of **AFSim** and **MoorthySim** . In this section, we demonstrate how temporal integration affects the VAQ assessment. Unlike **AFSim** , both **AFTVA** and **TVASim** can distinguish the temporal variation of SBFs. Fig. 12 shows the performance of **MoorthySim**, **Baseline**, **AFSim**, **AFTVA**, and **TVASim** when the number of video-based features are selected from $q = 1$ to $q = 17$, where Baseline indicates the approach always using the mean operator $\overline{op}_{m+1}$ for temporal integration.

It can be seen that, by selecting the same number of features ($q = 7$), **AFTVA** and **TVASim** perform better than **AFSim** and **MoorthySim**, while **Baseline** always achieves the worse results. More specifically, when the number of SBFs is the same ($q = 7$), **AFTVA** (with the accuracy $75.5 \pm 2.0\%$) performs almost the same as **AFSim** ($75.0 \pm 1.6\%$). Table II shows the features selected for **AFTVA** from $q = 1 \sim 7$. However, when the number of selection is increased, **AFTVA** can further boost the performance by keeping finding more discriminating SBFs relevant for VAQ determination, and the accuracy can be raised from $75.5 \pm 2.0\%$ to $80.1 \pm 2.3\%$, as shown in Fig. 12. It demonstrates that **AFTVA** can turn more video-based features into distinctive clues for assessing the VAQ.

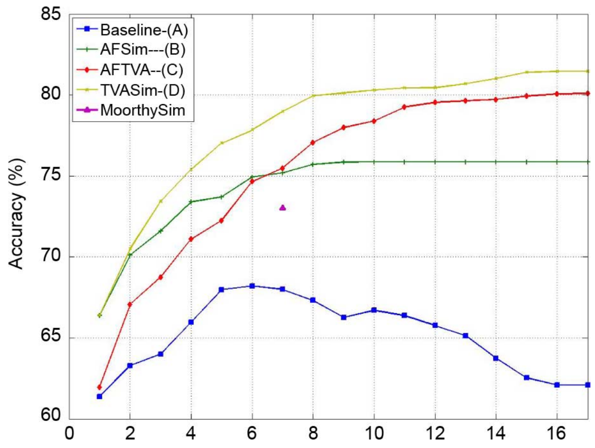


Fig. 12. The accuracies of Baseline, AFSim, AFTVA, TVASim, and MoorthySim obtained by varying the number of the SBFs selected. The x-axis represents the number, and the y-axis indicates the corresponding accuracy.

TABLE II
THE PERFORMANCE OF AFTVA, WHERE THE TOP 7 SBFS SELECTED ARE REPORTED IN THIS TABLE

# fea	SBF		Category	Accuracy (%)
1	f^{22}	min	Lightness	62.0 ± 1.5
2	f^{10}	min	Harmonization	67.1 ± 1.1
3	f^{6}	median	ST	68.8 ± 2.3
4	f^{11}	Q1	Harmonization	71.1 ± 1.8
5	f^{9}	median	Harmonization	72.3 ± 1.8
6	f^{17}	mean	Saturation	74.7 ± 2.0
7	f^{6}	mean	ST	75.5 ± 2.0

In addition, by considering both TVA and the averaged long-term (i.e., *mean*, *std.*) operators, **TVASim** achieves even-better performance, 79% accuracy when $q = 7$, and the best performance, $81.5 \pm 1.9\%$ accuracy when $q = 17$, as shown in Fig. 12. Therefore, it can be seen that our temporal integration operators can improve the performance significantly. Table III lists the video-based features selected by **TVASim**, which is the approach performing the best among all approaches. The top ranked feature types in **TVASim** are ST, MS, and HS, showing again that motion-related features are highly effective in determining VAQ. In addition, it can be observed that many temporal-integration operators of the TVA-type have been selected in the top ranked list, which also reveals that the proposed temporal-variation-aware integration operators are effective. Since the **TVASim** approach performs the best in our study for the **Telefonica** dataset, we recommend it as our main approach.

VII. RESULTS ON ADCCV DATASET

We further enhanced the **Telefonica** dataset by augmenting it with cinema movies.[4] In comparison to the consumer videos in the **Telefonica** dataset, these professional films are 'masterpieces' that would be extremely useful to serve as benchmarks of high AQ. We collected 40 professional movies and label them as 'positive'. These movie include "The Notebook (2004)," "Eat Pray Love (2010)," and so on. Several examples are shown in Fig. 13. By adding them into the original Telefonica dataset, we constructed the joint **A** esthetic **D** ataset of **C** inema and **C** onsumer **V** ideos (ADCCV) dataset. There are 120 positive examples (80 original and 40 professional movies added) and 80 negative examples (original), and so a rough balance between positive and negative examples is maintained. Since the **ADCCV** dataset combines different kinds of videos, it serves as a more general benchmark for VAQ assessment.

TABLE III
THE TOP 17 VIDEO-BASED FEATURES SELECTED BY TVASIM AND THE CORRESPONDING PERFORMANCE ON THE TELEFONICA DATASET

# fea	SBF		op-type	category	accuracy (%)
1	f^{6}	Q1	Sim	ST	66.4 ± 1.1
2	f^{5}	max	TVA	ST	70.5 ± 1.5
3	f^{1}	max	TVA	MS	73.4 ± 1.6
4	f^{4}	median	Sim	HS	75.4 ± 1.4
5	f^{20}	Q1	TVA	Lightness	77.0 ± 1.4
6	f^{16}	min	TVA	Saturation	77.8 ± 1.6
7	f^{10}	max	TVA	Harmonization	79.0 ± 1.5
8	f^{11}	Q1	TVA	Harmonization	79.9 ± 1.8
9	f^{17}	mean	Sim	Saturation	80.1 ± 1.9
10	f^{18}	Q3	Sim	Composition	80.3 ± 1.9
11	f^{17}	Q1	Sim	Saturation	80.4 ± 1.8
12	f^{20}	median	TVA	Composition	80.5 ± 1.9
13	f^{4}	mean	TVA	HS	80.7 ± 1.8
14	f^{8}	Q3	TVA	Harmonization	81.0 ± 2.0
15	f^{15}	mean	Sim	Saturation	81.4 ± 1.9
16	f^{3}	max	TVA	HS	81.5 ± 1.9
17	f^{20}	median	Sim	Composition	81.5 ± 1.9

TABLE IV
THE TOP 7 VIDEO-BASED FEATURES SELECTED BY TVASIM AND THE CORRESPONDING PERFORMANCE ON THE JOINT AESTHETIC DATASET OF CINEMA AND CONSUMER VIDEOS (ADCCV)

# fea	SBF		op-type	category	accuracy (%)
1	f^{1}	min	Sim	MS	70.0 ± 0.9
2	f^{6}	Q1	Sim	ST	74.3 ± 1.3
3	f^{4}	mean	Sim	HS	77.5 ± 1.3
4	f^{14}	max	TVA	Saturation	79.1 ± 1.2
5	f^{18}	max	Sim	Composition	80.1 ± 1.2
6	f^{2}	median	TVA	MDE	80.5 ± 1.3
7	f^{15}	mean	Sim	Saturation	81.1 ± 1.3

In this section, we conduct experimental studies based on the **ADCCV** dataset by using the **TVASim** approach. First, like [16], the CV accuracy is reported in Section VII-A. Then, we split the **ADCCV** dataset into training and testing sets several times and report the testing accuracies in Section VII-B.

A. CV Accuracy

In this experiment, we report the accuracy on the **ADCCV** dataset. As before, 5-fold cross-validation and 200-times repetitions are performed. As shown in Table IV, the accuracy is $81.1 \pm 1.3\%$ when $q = 7$ video-based features are selected.

Even when professional movies have been added as positive data in the **ADCCV** dataset, similar conclusions can be drawn as well: First, the CV accuracy achieved is still close to that of the **Telefonica** dataset. Second, many of the top-ranked types of the selected AFs are the motion-related ones, For example, the top 3 video-based features are also motion-related, and the combination of them attains over 77% accuracy. These results show that our method is effective for VAQ assessment. It is also worth to note that a single MS(Motion Space)-related SBF achieves 70%

[4] The information of the movies are available on http://imp.iis.sinica.edu.tw/IVCLab/research/aesthetictmm/Index.html.

Fig. 13. Five Examples From the Professional Videos in the ADCCV Dataset: Snapshots From “21 (2008)” (a), “The Blind Side (2009)” (b), “Elizabeth: The Golden Age (2007)” (c), “Rounders (1998)” (d), and “The Notebook (2004)” (e).

TABLE V
TESTING ACCURACIES OF RANDOMLY SPLITTING THE ADCCV DATASET INTO TRAINING AND TESTING SETS FOR 5 TIMES

# fea	Test1				Test2			
	SBF	op-type	V-Acc	T-Acc	SBF	op-type	V-Acc	T-Acc
1	f^{2} mean	Sim	69.4	70.0	f^{6} Q1	Sim	70.6	75.0
2	f^{8} max	Sim	75.0	65.0	f^{22} Q3	TVA	75.6	65.0
3	f^{4} mean	Sim	75.6	72.5	f^{4} median	TVA	76.9	70.0
4	f^{8} Q3	Sim	77.5	72.5	f^{6} median	TVA	77.5	67.5
5	f^{18} max	Sim	78.1	72.5	f^{5} max	TVA	80.0	75.0
6	f^{20} median	Sim	78.1	72.5	f^{10} max	Sim	81.9	77.5
7	f^{3} min	Sim	78.1	72.5	f^{4} Q1	TVA	83.1	77.5

# fea	Test3				Test4				Test5			
	SBF	op-type	V-Acc	T-Acc	SBF	op-type	V-Acc	T-Acc	SBF	op-type	V-Acc	T-Acc
1	f^{1} min	Sim	73.1	62.5	f^{4} mean	Sim	73.8	60.0	f^{1} min	Sim	71.3	70.0
2	f^{6} Q1	Sim	78.8	60.0	f^{22} Q3	TVA	76.9	57.5	f^{3} min	TVA	74.4	70.0
3	f^{22} Q3	TVA	81.3	65.0	f^{15} Q1	TVA	78.8	50.0	f^{17} median	Sim	76.3	70.0
4	f^{6} median	TVA	82.5	60.0	f^{22} min	Sim	79.4	50.0	f^{22} Q3	TVA	77.5	70.0
5	f^{18} max	Sim	83.1	60.0	f^{4} median	TVA	80.6	52.5	f^{16} Q3	TVA	80.6	67.5
6	f^{18} Q3	Sim	83.1	60.0	f^{11} Q1	TVA	80.6	45.0	f^{8} Q3	Sim	81.3	72.5
7	f^{20} median	Sim	83.1	60.0	f^{17} mean	Sim	82.5	52.5	f^{20} median	Sim	81.25	72.5

accuracy, indicating that the MS feature is more relevant when professional movies are involved.

B. Testing Accuracies of Random Splits

Though CV accuracy is reported in [16] and used in the comparison protocol for the **Telefonica** dataset, it may not reflect the true error and could sometimes overfit the dataset. In this experiment, we split the ADCCV dataset into training and test sets with 80/20 policy. That is, there are 80% (160 videos) used to train the TVASim approach, and the remaining 20% (40 videos) are reserved for testing, where the above partition is repeated for 5 times. In the training phase, we apply 4-folds[5] CV for selecting the video-based features. The selected video-based features associated with the best 4-fold CV accuracy are then applied to the 20% testing data.

The selected video-based features and accuracies for the five tests are shown in Table V. From the results, it is not surprised that the testing accuracies (T-Acc) are lower than the validation accuracies (V-Acc), because the testing sets are unseen in the training phase. However, most of the tests (such as “Test1”, “Test2”, “Test5”) show that when the validation accuracies (i.e., V-Acc) is increased as more video-based features are adopted, the testing accuracies (i.e., T-Acc) is also roughly increased. In these cases, the reductions from the V-Acc to T-Acc are also lower. The average testing accuracy when selecting $q = 7$ video-based features is $(72.5 + 77.5 + 60.0 + 52.5 + 72.5)/5 = 67\%$. Regarding the comparison protocol of the **Telefonica** dataset which uses the CV accuracy for the performance evaluation, the testing-data performance reflects better the generalization ability of the learned assessor. Although the testing accuracy is not as high as the CV accuracy, it is still within an acceptable level. The

[5] We use 4 folds cross-validation in order to guarantee that the sizes of the validation and testing sets are the same.

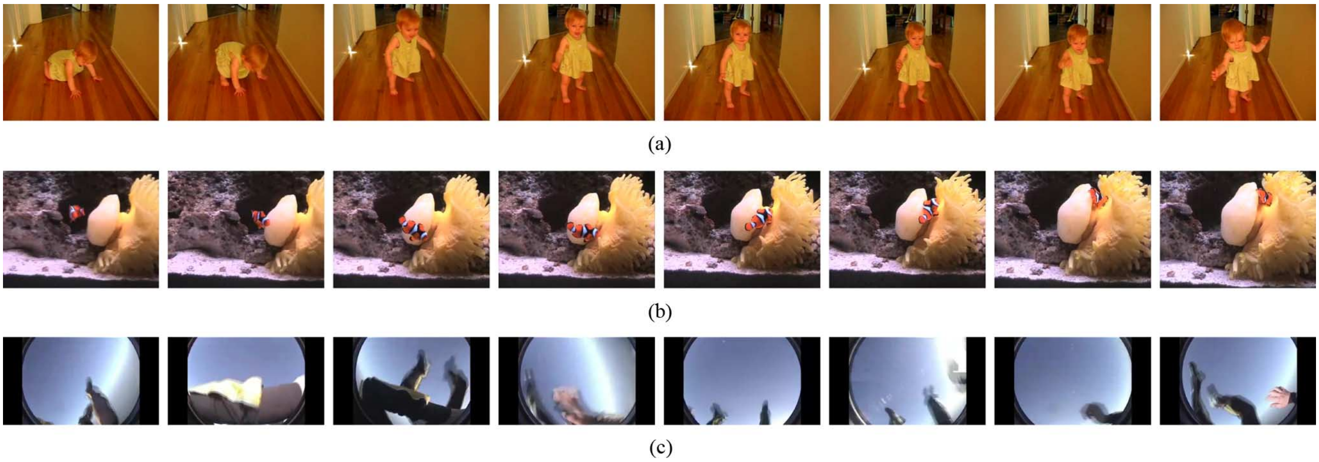


Fig. 14. Three videos are shown. (a) and (b) are labeled as positive, and (c) is labeled as negative. In our method, the motion-related AF, Hand-shaking (HS), can better distinguish these videos.

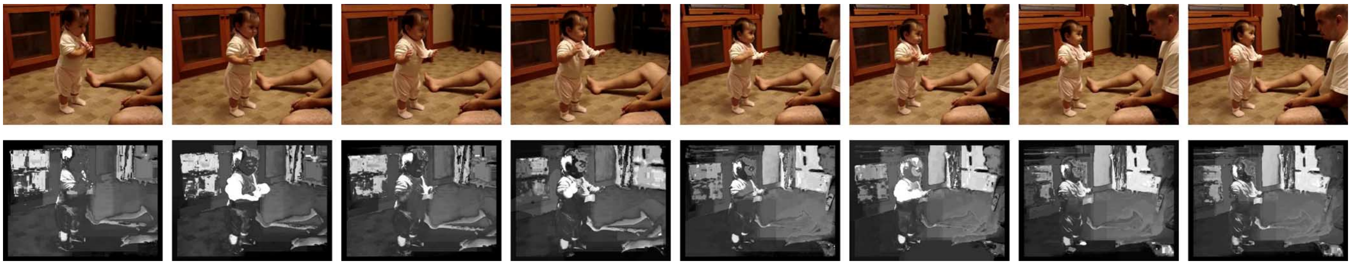


Fig. 15. An example of misclassified video by our method are shown in the first row. The second row shows its saliency maps extracted. The video is labeled as positive in the ADCCV dataset but is classified as negative in Test 1 of Table V.

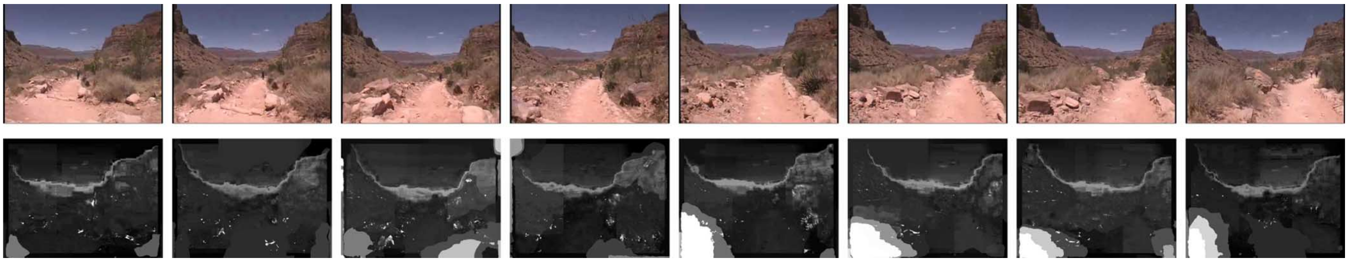


Fig. 16. Another example of misclassified video by our method are shown in the first row. The second row shows its saliency maps extracted. The video is labeled as negative in the ADCCV dataset but is classified as positive in Test 1 of Table V.

result can serve as a starting point for further refinement, and we will release the codes and data for future study.

To provide more insightful information, we discuss an example in the results of Table V. The top two rows of Test 1 in Table V shows that the T-Acc is 65% by using the joint SBFs of $(f^2, \mathtt{mean})$ (motion-direction entropy) and $(f^8, \mathtt{max})$ (color harmonization). However, when the SBF of $(f^4, \mathtt{mean})$ (handing shaking) is further included, the T-Acc is increased to 72.5%. Fig. 14 shows three videos associated with this situation, which are misclassified when the first two ($(f^2, \mathtt{mean})$ and $(f^8, \mathtt{max})$) are selected but is correctly classified when $(f^4, \mathtt{mean})$ is further selected. We found that the motion flows in the videos of Figs. 14(a)–14(b) are relatively stable than that in the video of Fig. 14(c). Hence, it would be the stability of motion flows that makes the hand-shaking (HS) AFs more informative in discriminating these videos.

Finally, we show some misclassified examples in the first rows of Figs. 15 and 16 for discussion. In both figures, the saliency maps obtained are also shown in the second rows, respectively. We find that the background information like color and motion makes the baby less salient in the video of Fig. 15; also, the noisy motion flows result in the incorrect video-based saliency maps in Fig. 16. Hence, the motion-related features extracted for these two videos are inaccurate, which affect the effectiveness of the AFs constructed and cause the wrong classification for VAQ assessment.

VIII. Conclusions and Future Work

In this paper, we presented a classification framework for VAQ assessment by temporal integration of the motion-based and photo-based AFs. We design a new set of the motion-based AFs, and introduce a temporal-aware integration framework for VAQ assessment. Experimental results show that our approaches are effective for video aesthetic-value evaluation. By applying the same comparison protocol in [16], we achieve 81.5 ± 1.9% accuracy on the **Telefonica** dataset that contains 160 rated consumer videos, where the state-of-the-art approach [16] achieves 73.0 ± 2.0%. Besides, we also construct the **ADCCV** dataset by including additional professional movies, and achieve 81.1 ± 1.3% accuracy. These results also reveal that the motion-related features play an important role for VAQ assessment in most experiments.

There are several interesting future directions. First, the AFs are computed from the extracted foreground region. It could be better when the AF is calculated by including semantic subjects or human faces according to some photo-based AQ studies [33], [34]. Second, we only focus on the binary classification (good or bad) for assessment, this could be enhanced by employing learning-to-rank techniques such as [15] for VAQ assessment. In addition, we study the VAQ from the collection of 15-seconds videos currently, it is interesting to study the affection of video lengths to VAQ.

References

[1] H. Sheikh, A. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: Jpeg2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.

[2] T. Mei, X.-S. Hua, C.-Z. Zhu, H.-Q. Zhou, and S. Li, "Home video visual quality assessment with spatiotemporal factors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 6, pp. 699–706, Jun. 2007.

[3] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. Eur. Conf. Comput. Vision (ECCV'06)*, 2006, pp. 7–13.

[5] V. Ramachandran and W. Hirstein, "The science of art: A neurological theory of aesthetic experience," *J. Consciousness Stud.*, vol. 6, no. 6-7, pp. 15–51, 1999.

[6] H. Kawabata and S. Zeki, "Neural correlates of beauty," *J. Neurophysiol.*, vol. 91, pp. 1699–1705, 2004.

[7] G. Peters, "Aesthetic primitives of images for visualization," in *Proc. 11th Int. Conf. Inf. Visualization (IV'07)*, 2007, pp. 316–325.

[8] Cultural Analytics. [Online]. Available: http://en.wikipedia.org/wiki/Cultural_analytics.

[9] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR'06)*, 2006, vol. 1, pp. 419–426.

[10] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR'11)*, 2011, pp. 33–40.

[11] S. Dhar, V. Ordonez, and T. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR'11)*, 2011, pp. 1657–1664.

[12] C. Li and T. Chen, "Aesthetic visual quality assessment of paintings," *IEEE J. Select. Topics Signal Process.*, vol. 3, no. 2, pp. 236–252, Apr. 2009.

[13] J. Li and J. Wang, "Studying digital imagery of ancient paintings by mixtures of stochastic models," *IEEE Trans. Image Process.*, vol. 13, no. 3, pp. 340–353, Mar. 2004.

[14] C. Li, A. C. Loui, and T. Chen, "Towards aesthetics: a photo quality assessment and photo selection system," in *Proc. 18th ACM Int. Conf. Multimedia (ACMMM'10)*, 2010, pp. 827–830.

[15] C.-H. Yeh, Y.-C. Ho, B. A. Barsky, and M. Ouhyoung, "Personalized photograph ranking and selection system," in *Proc. 18th ACM Int. Conf. Multimedia (ACMMM'10)*, 2010, pp. 211–220.

[16] A. K. Moorthy, P. Obrador, and N. Oliver, "Towards computational models of visual aesthetic appeal of consumer videos," in *Proc. Eur. Conf. Comput. Vision (ECCV'06)*, 2010, pp. 1–14.

[17] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. Eur. Conf. Comput. Vision (ECCV'06)*, 2008, pp. 386–399.

[18] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proc. 24th Int. Conf. Mach. Learning (ICML'07)*, 2007, pp. 129–136.

[19] M. Saini, R. Gadde, S. Yan, and W. T. Ooi, "Movimash: Online mobile video mashup," in *Proc. 20th ACM Int. Conf. Multimedia (ACMMM'12)*, 2012, pp. 139–148.

[20] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vision (ECCV'10)*, 2010, pp. 366–379.

[21] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2009.

[22] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.

[23] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. 11th ACM Int. Conf. Multimedia (ACMMM'03)*, 2003, pp. 374–381.

[24] J. You, A. Perkis, M. M. Hannuksela, and M. Gabbouj, "Perceptual quality assessment based on visual attention analysis," in *Proc. 17th ACM Int. Conf. Multimedia (ACMMM'09), 2009*, pp. 561–564.

[25] S. Battiato, G. Puglisi, and A. Bruna, "A robust video stabilization system by adaptive motion vectors filtering," in *Proc. IEEE ICME*, 2008, pp. 373–376.

[26] J. Yang, D. Schonfeld, and M. Mohamed, "Robust video stabilization based on particle filter tracking of projected camera motion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 7, pp. 945–954, Jul. 2009.

[27] P. O'Donovan, A. Agarwala, and A. Hertzmann, "Color compatibility from large datasets," *ACM Trans. Graph.*, vol. 30, no. 4, article no. 63, Aug. 2011.

[28] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y. Xu, "Color harmonization," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 624–630, Jul. 2006.

[29] B. Krages, *Photography: The Art of Composition*. New York, NY, USA: Allworth, 2005.

[30] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Proc. SPIE Video Commun. and Image Processing Conf.*, 2003, pp. 8–11.

[31] F. E. Beerends, G. John, and D. Caluwe, "The influence of video quality on perceived audio quality and vice versa," *J. Audio Eng. Soc.*, vol. 47, no. 5, pp. 355–362, 1999.

[32] D. Joshi, R. Datta, Q.-T. Luong, E. Fedorovskaya, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images: A computational perspective," *IEEE Signal Process. Mag.*, vol. 28, no. 5, pp. 94–115, Sep. 2011.

[33] X. Sun, H. Yao, R. Ji, and S. Liu, "Photo assessment based on computational visual attention model," in *Proc. 17th ACM Int. Conf. Multimedia (ACMMM'09)*, 2009, pp. 541–544.

[34] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. IEEE Int. Conf. Computer Vision (ICCV'11)*, 2011, pp. 2206–2213.

Hsin-Ho Yeh received a B.S. degree in Department of Computer Science and Information Engineering from National Chung Cheng University, Taiwan, in 2007. He received an M.S. degree in 2009 from the Department of Computer Science and Information Engineering, National Cheng-Kung University. He is a research assistant of Institute of Information Science (IIS), Academia Sinica. His research interests include computer vision and data mining.

Chun-Yu Yang received the B.S. degree in Electrical Engineering and Computer Science Honors Program from National Chiao Tung University (NCTU) in 2009 and has been a Research Assistant at the Institute of Information Science (IIS), Academia Sinica. He is currently working toward the M.S. degree in the Department of Computer Science and Information Engineering in National Taiwan University (NTU).

Ming-Sui Lee received her B.S. degree in mathematical sciences from National Cheng-Chi University, Taiwan, in 1999. Then she received the M.S. degree in electrical engineering from the University of California, Los Angeles (UCLA) and the Ph.D. degree in electrical engineering from the University of Southern California (USC) in 2002 and 2006, respectively. Now she serves as an assistant professor at the Department of Computer Science and Information Engineering, National Taiwan University. Her research interests include image/video mosaicking, compressed-domain image/video processing, 2D to 3D image/video conversion and digital signal processing.

Chu-Song Chen received a B.S. degree in Control Engineering from National Chiao-Tung University, Taiwan, in 1989. He received an M.S. degree in 1991 and a Ph.D. degree in 1996, respectively, both from the Department of Computer Science and Information Engineering, National Taiwan University. He is now a deputy director of Research Center for Information Technology Innovation (CITI), Academia Sinica, and a research fellow of Institute of Information Science (IIS), Academia Sinica, Taiwan. Dr. Chen's research interests include pattern recognition, computer vision, signal/image processing, and multimedia analysis. He is on the editorial board of *Journal of Multimedia* (Academy Publisher), *Machine Vision and Applications (Springer), IPSJ Trans. on Computer Vision and Applications*, and *Journal of Information Science and Engineering*.