The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Understanding Image Impressiveness Inspired by Instantaneous Human Perceptual Cues

Jufeng Yang,[1] Yan Sun,[1] Jie Liang,[1] Yong-Liang Yang,[2] Ming-Ming Cheng[1]
[1]College of Computer and Control Engineering, Nankai University, No.38 Tongyan Road, Tianjin, China
[2]Department of Computer Science, University of Bath, Claverton Down, Bath, United Kingdom

Abstract

With the explosion of visual information nowadays, millions of digital images are available to the users. How to efficiently explore a large set of images and retrieve useful information thus becomes extremely important. Unfortunately only some of the images can impress the user at first glance. Others that make little sense in human perception are often discarded, while still costing valuable time and space. Therefore, it is significant to identify these two kinds of images for relieving the load of online repositories and accelerating information retrieval process. However, most of the existing image properties, *e.g*., memorability and popularity, are based on repeated human interactions, which limit the research and application of evaluating image quality in terms of instantaneous impression. In this paper, we propose a novel image property, called *impressiveness*, that measures how images impress people with a short-term contact. This is based on an impression-driven model inspired by a number of important human perceptual cues. To achieve this, we first collect three datasets in various domains, which are labeled according to the instantaneous sensation of the annotators. Then we investigate the impressiveness property via six established human perceptual cues as well as the corresponding features from pixel to semantic levels. Sequentially, we verify the consistency of the impressiveness which can be quantitatively measured by multiple visual representations, and evaluate their latent relationships. Finally, we apply the proposed impressiveness property to rank the images for an efficient image recommendation system.

Introduction

In recent years, we have witnessed an explosive proliferation of online digital images, a large proportion of which provide little significance yet cost considerable repository resources. As a result, how to effectively identify meaningful images from miscellaneous visual sources is highly important for efficient computation and storage, benefiting a number of applications such as image exploration and retrieval.

In the psychology community, there are two types of models evaluating the significance of the given image, *i.e*., the memory-based models and the impression-driven models (Lodge, McGraw, and Stroh 1989; Newell 1972). The former depends on the sustained retrieval of specific information from long-term memories (Fiske and Taylor 1984), which are increasingly difficult to access with the rapid growth of the visual data. Also, the memory-based information is hard to be measured quantitatively, which restricts the development and application of the corresponding model. As illustrated in (Lodge, McGraw, and Stroh 1989), for evaluating the significance of images, people rarely rely on long-term memories but only require short-term perceptions with consistent logics. In contrast with the memory-based models, impression-driven processes investigate the instantaneous perceptual information when the viewers encounter an image (Wyer and Srull 1986; Hastie and Park 1986). As the longer exposures do not significantly alter the prior sensation of viewers (Willis and Todorov 2006), the instantaneous impression also preserves the consistency of the quantitative measurement and enhances the identification performance regarding both efficiency and effectiveness. Therefore, the impression-driven processes enable efficient judgment of image significance compared with long-term memory based models.

Figure 1: Impressive images (**top**) *vs.* non-impressive images (**bottom**). The investigation shows that the impressiveness is mainly affected by the six human perceptual cues, *i.e*., the variety between the foreground and the background, the aesthetic, the contrast or the clarity, the strong arousal, the spectacularity and the unusualness. Each column corresponds to a specific property of the impressiveness, *e.g*., for the first column, the top image exhibits strong variety between foreground and background, thus is impressive, while the bottom one looks otherwise and thus unimpressive. The label of the impressiveness is generated by annotators of which the details are elaborated in the Dataset section.

To investigate impression-driven image evaluation, we first construct three datasets from three popular image exploration contexts, *i.e*., photo sharing, news propagation and commercial popularization. All the selected contexts suffer from a common problem that massive redundant images prevent effective information access. We then generate binary labels of image impression based on the instantaneous perceptions of the annotators.

Both the visual characteristics of the datasets and the quantifiable psychological findings demonstrate that the judgment on image impression is influenced by multiple human visual perception cues (Gilron and Gutchess 2012). Inspired by that, we propose a novel image property, *i.e*., the image *impressiveness*, to evaluate the human impression when encountering an image.

Definition 1 (Image Impressiveness) *The* impressiveness *of an image evaluates the influence on human perception which is formed with the instantaneous encounter between the viewer and the image (Willis and Todorov 2006). cues motivated by various image attributes (Geng and Hamilton 2006) are listed below.*

- *Foreground: the distinction between the foreground and the background of an image;*
- *Aesthetic: the pleasure or satisfaction with the image quality;*
- *Contrast: the resolution or the clarity of an image;*
- *Arousal: the emotional influence of an image to viewers;*
- *Spectacularity: the openness of an image;*
- *Unusualness: the degree of the strangeness or mysteriousness of an image.*

We consider the images with high impressiveness as impressive images, while the others are termed as non-impressive images. Figure 1 shows the contrast between impressive and non-impressive images arranged by the aforementioned human perceptual cues. In the investigation of interestingness, Gygli *et al*. (Gygli et al. 2013) denotes the aspects/variables affecting the image property as 'cues'. Sequentially, the six primary cues can be considered as significant attributes to determine whether an image is impressive or not. Specifically, the **Foreground** property is widely studied in the cognitive community which helps the process of detection, classification and content analysis by bridging the semantic gap (Sethi 2000). The **Aesthetic** property of an image is associated with both the principle of nature and the appreciation of beauty (Datta et al. 2006). The common experiences of photography verify that specific aspects *e.g*., the color space, sharpness and texture, are critical for measuring the aesthetics. The **Contrast** property is determined by the variation of luminosity, color or brightness which induces a distinguishable visual representation. The **Arousal** is an affective factor of an image relevant to human emotions, of which different level evokes different feeling of the viewers. The emotion is an essential semantic pattern which complements the traditional object modeling (Machajdik and Hanbury 2010; Zhao et al. 2017). The **Spectacularity** relies on the attribute of openness which is correlated with intuitions about the natural scenes (Lehman and Stanley 2012). The **Unusualness** turns out to be an important cue in the property of interestingness (Geng and Hamilton 2006), following which we capture the unusualness information in artificial scenes. Particularly, there are existing theories in either psychological or visual community validate the observation on cues of impressiveness. Kumamoto *et al*. (Kumamoto and Tanaka 2005) analyze impressions of articles with various emotions. Lodge *et al*. (Lodge, McGraw, and Stroh 1989) propose the intuition influence of unusualness (contrast to the familiarity) in the formation of impressiveness. Aesthetics has been an important aspects in the investigation of image quality (Bhattacharya, Sukthankar, and Shah 2010). As for contrast and foreground, they have played important role in almost all image properties.

We also note that due to the multi-modal nature of the six cues *e.g*., statistical, semantical, emotional, they are quantitatively measured based on image features at multiple levels, *i.e*., pixel-level information (*e.g*., pixel values in the HSV color space), low-level features (*e.g*., texture, gradient), and higher-level semantic representations (*e.g*., object, emotion). We select proper feature combinations based on the classic works which are verified to be effective on measuring single perception cues (Khosla et al. 2015a; Borth et al. 2013). Inspired by mutual influence of the various cues contributed to different image properties, we further propose a comparison method to quantify the relationship among different properties. Sequentially, we find the cues with high influence on the determination of the impressiveness.

In summary, our work makes three major contributions: First, we propose impressiveness as a novel image property, and the corresponding instantaneous human perceptual cues for measurement. We evaluate the quality of impressiveness via a top-down strategy: impressiveness $\rightarrow$ cues $\rightarrow$ visual representations following classic works. Second, we construct three datasets in different contexts with extensive statistics, *e.g*., the consistency of the labels. Then, we conduct experiments to quantify and demonstrate the proposed impressiveness cues. Further comparisons are provided to distinguish the proposed impressiveness from existing properties.

Related Work

In this section, we discuss related work by summarizing the existing image properties investigated in the field, and reviewing how image properties are identified using psychological and computational approaches.

Image Properties

The recent development of information technology and social networks in particular has resulted in a rapid growth of online digital images, which significantly influences human life. How to identify high-level quality/meaningfulness of an image attracts many researchers (Parikh et al. 2012). In recent years, several high-level image properties, *e.g*., popularity (Khosla, Sarma, and Hamid 2014), virality (Deza and Parikh 2015; Guerini and Staiano 2015) and memorability (Khosla et al. 2015b; Danescu-Niculescu-Mizil et al.

2012), are presented to describe the natural influence of images, most of which are based on the long-term memory with repeated human interactions. For instance, (Khosla, Sarma, and Hamid 2014) formulates image popularity on the number of views for each image. Similarly, (Deza and Parikh 2015) calculates image virality based on the number of the up-votes and resubmissions attached to the image. Besides, image memorability is estimated through a 'Memory Game' in which the viewers watch a long stream of images repeatedly (Isola et al. 2011; Khosla et al. 2012). Most of the existing properties rely on comprehensive human interactions, *e.g*., a long-term observation of one person or a common reaction of a group, which are sequentially time-consuming and hard to access. In this paper, we present a novel image property called *impressiveness* for subjective image quality evaluation. This is based on an impression-driven model using instantaneous human perception cues other than costly human interactions. Differ from these properties, impressiveness refers to the degree of an image impresses human at the first glance. While the smoothing picture to the eyes are easier to be remembered or resubmitted, images with negative information can stick into human impressions as well as the positive. The assessment of impressiveness quality is extremely subjective and our approach only relies on visual information that is intrinsic to an image. It does not require any textural annotations which are knowledge-dependent, and often not available for online images.

The Identification of the Properties

Psychological Approach. In the field of psychology, there are extensive works studying human impression in different contexts. For example, researchers investigate the management organization under the consideration of impression (Gardner and Martinko 1988). The human-computer experiments based on visual and vocabulary features attempt to shape the first impression of the users (Tuch et al. 2012). In (Kumamoto and Tanaka 2005), the impressions derived from the news articles are exploited. Besides, in information retrieval, the impression also serves as a clue of the document (Hirabayashi, Matoba, and Kasahara 1988). While human impressions have been studied in the field of psychology, no prior work has investigated impressiveness as an intrinsic image property in computer vision. To the best of our knowledge, we are the first to introduce image impressiveness, and to utilize it for evaluating images in real application domains.

Computational Approach. To computationally identify high-level image property, a wide range of elements should be considered. Specifically, in the study of the memorability property, attributes including simple features, object statistics and global elements are measured to confirm the predictability of memorability (Borkin et al. 2013). (Khosla, Sarma, and Hamid 2014) exploits various features on texture, gradient, *etc*. which are highly correlated with the popularity property of an image. (Gygli et al. 2013) studies the interestingness of each image regarding the unusualness, aesthetics, *etc*. Besides, the image content and the social context are also adopted in the modeling of popularity (Khosla, Sarma, and Hamid 2014), virality (Deza and Parikh 2015), and specificity (Jas and Parikh 2015).

Since the impressiveness is modeled instantaneously, we investigate the intrinsic features without the influence of any social affairs. Inspired by the previous works which exploit various attributes of an image, we utilize basic visual attributes in the investigation of impressiveness. Nonetheless, no previous work takes into account the human affection, which actually serves as an essential pattern for image impressiveness (Hanjalic 2006).. We introduce novel emotion factor for the first time to measure image impressiveness, complementing the basic visual features which are widely investigated in the tasks of image properties (Machajdik and Hanbury 2010).

Dataset

For the quantitative research of the image impressiveness, we construct three benchmark datasets from different image repositories.

Collection

We collect images from diverse platforms to create three datasets: Flickr_Imp, News_Imp and Trip_Imp, each of which contains over $10,000$ images in the same context, *i.e*., photo sharing, news propagation and commercial popularization, respectively. We also verify that the three datasets suffer from the inefficiency problem caused by non-impressive images.

Flickr_Imp consists of $10,258$ images from Flickr which famous for sharing daily life photos. Part of Flickr images are uploaded by expert photographers or bloggers who are good at capturing attractive images. These images can be sequentially impressive with high contrast, aesthetic or unusualness, *etc*. In contrast, images from common users are more likely to be non-impressive. **News_Imp** comprises $10,315$ images from the major news sites, *e.g*., CNN, USA Today, BBC, *etc*. As the illustrations of news reports, these images are required to effectively represent real events. While some of them are concerned by the audience due to the implied objects or emotions. **Trip_Imp** gathers images provided by several famous travel sites, *e.g*., TripAdvisor. There are $10,400$ images in total. Since the goal of the travel sites is mainly for advertisement and promotion, images with relatively high aesthetics or spectacularity are with high priorities against the images captured by ordinary people.

The example images from the three datasets are shown in the supplemental material. We then design a voting scheme to generate binary labels (impressive or not) for the collected images, which is illustrated in the next subsection.

Voting Scheme

We recruit a group of 15 annotators (7 males, 8 females with different backgrounds) to generate the label of impressiveness for the collected images (5 annotators are allocated to each dataset). All annotators are informed with the following before they start: First, the three datasets are constructed under the context of photo sharing, news propagation and commercial popularization, respectively. Second, the weak suggestions of the six human perceptual cues as described

in Definition 1. Moreover, 1.5 seconds are allowed at most for annotation (including the time for manipulating the annotation software).

The annotation process is detailed as follows. As an undeveloped property the impressiveness is, we ask annotators to vote without training. The raw votes should serve for human consistency and will be used to discuss the predictability of impressiveness following (Jas and Parikh 2015) (Section). Every annotator takes 100 milliseconds to observe one image and form an impression, followed by another 1.4 seconds to manipulate the annotation software (Isola et al. 2011). We obligate enough time for human action, since longer contacts do not significantly alter the impression (Gilron and Gutchess 2012). Annotators are free to score impressiveness based on their own feelings beyond the weak suggestion of primary cues. For each image, we record the impressiveness score from the annotator according to the Likert Scale method in (Jamieson 2005). Specifically, we employ the six-point scale ($0 \sim 5$), in which higher score indicates image with higher impressiveness. For each dataset, we ask the annotators to generate annotations independently. Then we take the average and truncate it to the nearest integer as the final impressiveness score.

Statistics and Analysis

To give a clear dichotomy for exploring the image impressiveness, in each dataset, we only keep the images with score lower than 2 or greater than 3 (Deza and Parikh 2015; Gygli et al. 2013). We then discard all the images with intermediate scores, and only keep the subset of images with low/high impressiveness. As a result, Flickr_Imp consists of $3,230$ images with low impressiveness and $2,651$ images with high impressiveness, News_Imp contains $2,964$ images with low impressiveness and $2,559$ images with high impressiveness, and Trip_Imp includes $2,777$ images with low impressiveness and $2,512$ images with high impressiveness.

To do so, we calculate the consistency correlations among annotators following (Jas and Parikh 2015). For each image annotated by five participates, we split these votes into two parts. One part contains vote(s) from one or two participates, the rest are in the other part. Then we measure the impressiveness correlations between these two parts. We employ the $Spearman's$ rank correlation (Gygli et al. 2013) and the average correlation coefficient from $\binom{5}{2}$ combinations to demonstrate the consistency of impressiveness. The generated correlations on (the subset of) Flickr_Imp, News_Imp and Trip_Imp are 0.67, 0.65 and 0.65 respectively, which indicate a high consistency of human perception in terms of image impressiveness.

Predicting Image Impressiveness

In this section, we investigate various visual representations which identify the impressiveness according to the six primary cues introduced in Definition 1. Inspired by existing works which develop measurable image features to realize perceptual cues, we map the six cues into representations at three levels, *i.e.*, the pixel statistics, the low-level and mid-level features, according to the modal of the cues. We also incorporate the deep-level semantic features to further improve the prediction performance. Table 1 summarizes the mapping from cues to the corresponding representations. The details are elaborated in the following subsections.

Table 1: The mapping from six human perceptual cues to the combination of visual representations, *e.g.*, the aesthetics of an image can mainly be measured by the representations of texture, color and gradient. Then, the mappings are summarized into three levels according to the accepted taxonomy of the visual representations.

Levels	Cues	Visual Representations
Pixel-	Contrast	luminosity & brightness
Low-	Aesthetics	texture & color & gradient
	Spectacularity	texture & scene
	Unusualness	color & gradient & scene
Mid-, Deep-	Arousal	emotions
	Foreground	object semantics

Pixel-Level Statistics

Pixels are the rawest components of an image which reflect the basic visual information. We utilize pixel-level statistics including luminosity and brightness to measure impressiveness based on the *contrast* cue. We represent each image in the HSV color space and employ the extractor proposed by (Bhattacharya et al. 2013) to model the luminosity. The brightness is computed by the arithmetic mean of the red, green, blue channels in RGB color space. In addition, we exploit the Low Depth of Field (LDoF) (Luo, Wang, and Tang 2011) which is a well-known characteristic for image quality.

Low-Level Features

In this section, we consider four common visual representations as the low-level features for impressiveness measurement, namely texture, color, gradient and scene.

Texture. The texture feature is a descriptive component effective for distinguishing sharp and blurred images. The *aesthetics* and *spectacularity* cues are taken into account using the texture features. We employ Local Binary Pattern (LBP) (Ojala, Pietikinen, and Menp 2002) to capture the texture information. A uniform-LBP of 59-dimensions is adopted as the final representation.

Color. Colors play an important role in the vision system of human, and its combinations interpret the cultural and anthropological backgrounds of artists (Colombo, Del Bimbo, and Pala 1999). We employ the color histogram defined by (Siersdorfer et al. 2010) to include both global (GCH) and local (LCH) statistics. The GCH features are represented by RGB histogram of 64-dimensions. For LCH features, images are split into $4 * 4$ blocks, resulting in a 1,024 dimensional feature representation.

Gradients. Image gradient is an effective tool for various visual understanding tasks. We capture the gradient features using the Histogram of Oriented Gradients (HOG). Bag-of-words (BOW) is adopted to encode the HOG features of 300 dimensions.

Table 2: Prediction performance for impressiveness on three proposed datasets with the representations of multiple levels. The reported results are obtained from the best features for respective visual representations. We evaluate the results with common metrics including the accuracy, precision, recall and F1_measure. The classification performance by the representations of three levels, *i.e*., the low-, mid- and deep-level, are reported. For each level, we conduct the experiments on the fused representations. We also fuse the six features, *i.e*., texture, color, gradient, scene, emotion and object, of which the prediction results are shown in the last column. As shown, the performance is successively improved from the low-level to deep-level representations.

Data-sets	Metrics	Low-Level								Mid-Level			Deep-Level			All Fusion
		Texture (T)	Color (C)	Gradient (G)	Scene (S)	T&C&G	T&S	C&G&S	T&C& G&S	Emotion (E)	Object (O)	E&O	EmoNet	CaffeNet	Fusion	
Flickr_Imp	Accuracy	0.539	0.636	0.636	0.612	0.660	0.686	0.715	0.734	0.742	0.751	0.780	0.818	0.814	0.845	**0.852**
	Precision	0.647	0.515	0.509	0.485	0.664	0.673	0.711	0.563	0.600	0.617	0.740	0.713	0.711	0.732	**0.834**
	Recall	0.408	0.548	0.548	0.516	0.335	0.419	0.485	0.712	0.709	0.720	0.695	0.773	0.759	0.780	**0.784**
	F1_measure	0.500	0.531	0.528	0.500	0.441	0.517	0.577	0.629	0.649	0.665	0.716	0.741	0.734	0.755	**0.809**
News_Imp	Accuracy	0.553	0.657	0.689	0.614	0.692	0.732	0.729	0.778	0.749	0.741	0.761	0.811	0.821	0.830	**0.842**
	Precision	0.420	0.548	0.669	0.548	0.651	0.739	0.696	0.764	0.682	0.752	0.723	0.803	0.809	0.811	**0.840**
	Recall	0.508	0.642	0.652	0.551	0.688	0.631	0.713	0.750	0.743	0.728	0.764	0.712	0.724	0.751	**0.803**
	F1_measure	0.460	0.591	0.660	0.550	0.669	0.680	0.704	0.757	0.711	0.740	0.743	0.755	0.760	0.780	**0.821**
Trip_Imp	Accuracy	0.542	0.681	0.663	0.616	0.734	0.697	0.715	0.728	0.768	0.789	0.796	0.799	0.802	0.808	**0.820**
	Precision	0.592	0.678	0.592	0.520	0.692	0.665	0.681	0.829	0.803	0.816	0.795	0.774	0.834	0.792	**0.805**
	Recall	0.511	0.656	0.660	0.608	0.783	0.717	0.743	0.667	0.731	0.756	0.763	0.809	0.743	0.803	**0.816**
	F1_measure	0.549	0.667	0.623	0.560	0.735	0.690	0.711	0.739	0.765	0.785	0.778	0.791	0.787	0.797	**0.811**

Scene. The scene of an image is generally recognized from the global configuration. The well-known feature GIST (Oliva and Torralba 2001) encodes naturalness, openness, roughness, expansion and ruggedness into a 512-dimensional vector, which is utilized in this paper.

Inspired by prior works, we employ a compound feature scheme to realize the human perceptual cues. For estimating the *aesthetics*, (Dhar, Ordonez, and Berg 2011) employs not only the pixel statistics, *e.g*., luminosity and LDoF, but also the low-level representations. Similarly, we utilize a set of descriptors include GCH, LCH, HoG and GIST to cope with the color signals, human presence and the scene semantics. The *spectacularity* is an important cue for impressiveness which is relevant to the degree of openness and novelty (Lehman and Stanley 2012). We employ the GIST feature to represent the openness and capture the novelty using LBP feature. For estimating the *unusualness*, (Gygli et al. 2013) investigates the global outliers and compositions of parts, where the fusion of features including color, gradient and scene are used, so as in this paper.

Implementation Details To systematically describe the primary cues with corresponding features (as shown in Table 1), we fuse features via concatenation and multi-kernel learning (MKL) (Wang et al. 2017). In the framework of MKL, both Gaussian and polynomial kernels are adopted for selection. We examine the gaussian kernel with variances [1 3 5 7 10 12 15 17 20] and the polynomial with degrees [1 2 3 4] for efficient performance. We train classifiers on 10 random splits for each of the three datasets. Then we evaluate the performance with four common metrics, *i.e*., accuracy, precision, recall, F1_measure. We show classification results on low-level features and their combinations in separate columns of Table 2. As shown, single attribute is insufficient for the prediction of impressiveness. For instance, the results from considering only texture or scene just slightly outperform the random classification (50% accuracy). The combinations of low-level features based on the human perceptual cues achieve higher performance on all metrics. For example, on Flickr_Imp dataset, the unusualness (a combination of color(C), gradient (G) and scene(S), noted as C&G&S) is superior to aesthetics and spectacularity for over 3% to 5%. As Flickr_Imp images are collected from photo sharing websites, we find that images in unusual fashion take relatively big proportion of highly impressive images. This coincides with the fact that the unusualness property makes the photo more impressive than others in the social networks. We also conduct the classification experiment by fusing all low-level features, *i.e*., T&C&G&S. It outperforms any low-level cues on all three datasets. Therefore, the impressiveness can not be replaced by any specific cues.

Mid-Level Semantics

The mid-level object and emotion factors are higher semantic representations than the low-level image features. Both semantics are verified to be significant for impressing the viewers. For example, content based image retrieval (CBIR) and emotional semantic image retrieval (ESIR) both demonstrate the capability of involving object and emotion to mimic the human intuitions in creative applications. In addition, a survey about "Why do you embed an image in a tweet?" (Chen et al. 2015) reflects that 66.6% participants vote for the emotion enhancing while the visual relations account for 29.4%. Therefore, we investigate both visual content and emotional components for representing the proposed image impressiveness, of which the implementation details are presented as follows.

Emotion. The emotion factor in images is a relatively novel aspect when discussing image perception. It potentially influences the impression of the viewers according to the *arousal* cue as mentioned in Definition 1. We employ the emotion detector SentiBank in (Borth et al. 2013) which shows great ability in affective models. We extract the emotional representations in $1,200$ adjective noun pairs (ANP) which correspond to different levels of emotions in the Sen-

tiBank.

Object. The object factor reflects the identification of the visual content, especially for the foreground. Therefore, it is of high correlation with the *foreground* cue. We investigate the object factor by conducting the ObjectBank in (Li et al. 2010), which encodes the semantic and spatial information of objects in $44,604$-dimensional representations.

Compared to the low-level features, mid-level semantics approach is close to human perception. As shown in Table 2, Emotion (SentiBank) and Object (ObjectBank) features achieve about 75.3% and 76.0% accuracy in the prediction of impressiveness, which are far better than the four individual low-level features, and are comparable to the fusion of them. We also conduct multi-kernel learning on the combination of SentiBank and ObjectBank, which further enhances the prediction performance.

Deep-Level Incorporation

While mid-level features and their fusion achieve acceptable performance in reflecting the human perception, the prediction performance can be further improved by involving deep-level features. We extract the cognitive and affective information with deep-level representations. For the object factor, we employ the CaffeNet (Krizhevsky, Sutskever, and Hinton 2012) pre-trained on the ImageNet dataset. For the emotion factor, we fine-tune the CaffeNet on the large-scale emotion benchmark proposed by (You et al. 2016), namely EmoNet. We evaluate the distinguishing capability of deep features derived from the CaffeNet and EmoNet in the last three columns of Table 2. As shown, the deep-features of both object or emotion show the accuracies of more than 80%. In most cases, fusing all features based on multi-kernel learning further improves the performance on all metrics.

While impressiveness is induced from six cues, we propose to integrate all cues through multi-level features except contrast (as pixel statistics present trivial correlations with impressiveness) . Table 2 highlights final results under the column 'All Fusion'. Although arousal and foreground cues have shown great performance via deep features, low-level cues allow further improvement. This is due to the fact that impressiveness is a multi-modal image property affected by complex human perception, and diverse attributes give rise to the formation of impressiveness. Developing multi-level features enables the comprehensive representation of impressiveness, which is otherwise not possible using single level features. Some mid-level features (*i.e.*, emotion and object) also demonstrate the effectiveness of reflecting impressiveness, a more comprehensive representation with complementary features provides a better description.

Comparison with Other Properties

In the previous subsections, we map the human perceptual cues of the proposed impressiveness to the measurable attributes *i.e.*, texture, color, gradient, scene, object and emotion. Note that most of the existing image properties, *e.g.*, the interestingness (Gygli et al. 2013) and memorability (Isola et al. 2011), are also quantified based on visual representations. In this section, we compare the proposed impressiveness with other image properties according to their fundamental visual component.

Table 3 shows the comparison among various image properties based on their corresponding intrinsic visual representations. For each property, we show the rank of individual components according to reported performances of the corresponding paper (cited in the first column). As shown, the proposed impressiveness is the first property considering the emotion as a major component. We also propose to distinguish these properties by calculating the correlations between the impressiveness and others. Specifically, assume two properties A and B (always being the impressiveness), we set X_A to be the ranked components of A, *e.g.*, $X_{Memorability} = \{5, 4, 2, 3, 1, -\}$ where '-' means the component is not considered. $|X_A|$ reflects the number of components for the property A, *e.g.*, $|X_{Memorability}| = 5$. Then, we compute the correlation in three steps. First, we calculate the intersection-over-union (IoU) score s_{IoU} according to the overlapped components: $s_{IoU} = |X_A \cap X_B|/|X_A \cup X_B|$, where $|\cdot|$ denotes the number of the components. Second, to quantify rank relations of two properties, we employ the evaluation of mean average precision (mAP). For each property X_A, we calculate the mAP score regarding the impressiveness as follows: $s_{mAP} = \frac{\sum_{k=1}^{|X_B|} AP(k)}{|X_B|}$, where X_B denotes the impressiveness and $AP(k)$ is calculated based on the ranked components of X_A. Third, we introduce the penalty term s_{inv} on the inversion pairs, *i.e.*, the order of two components are opposite, which is defined as: $s_{inv} = \log_2(1 + \exp(-\frac{n_{inv}}{n_{reg}}))$. Here, n_{inv} and n_{reg} denote the number of inversion and regular pairs, respectively. The final correlation score $s_{correlation}$ between properties is then defined as: $s_{correlation} = s_{IoU} * s_{mAP} * s_{inv}$.

We report the correlation score $s_{correlation}$ between impressiveness and other properties in the last column of Table 3. The highest $s_{correlation}$ (0.61) is with the memorability. While the impressiveness costs less cognitive tax in the image evaluation, the memorability relies on the knowledge learned from long term memory. In addition, it turns out that there is a weak correlation between interestingness and impressiveness ($s_{correlation} = 0.30$). As verified in (Geng and Hamilton 2006), interestingness is often tied to the time-dependence and novelty, while impressiveness usually does not alter with longer exposure.

Application: Image Recommendation

In this section, we apply the proposed impressiveness property to image recommendation application to demonstrate its usefulness. The ever increasing number of digital images on the Internet poses new challenges for image recommendation, since a large proportion of online images are not impressive and suggesting these images can largely affect user experience. We consider the proposed impressiveness as a prominent image property to diminish redundant information caused by unimpressive images, thus elevate the performance of image recommendation.

We first gather four sets of images from different topics in the News_Imp dataset, *i.e.*, technology, politics, sport, travel,

Table 3: The comparison among various image properties against the corresponding intrinsic visual representations. The "$\surd$" indicates that the visual representation is required for represent the corresponding image property. The ranks in each row behind the "$\surd$" are ordered by their experimental performance according to the previous works. The last column reflects the correlations between the specific image property and the proposed impressiveness based on the related visual components. As shown, the proposed impressiveness property is the most comprehensive measurement for which the novel emotion factor is considered as an primary component. Also, it is distinctive to the existing properties.

Image Properties	Texture	Color	Gradient	Scene	Object	Emotion	$s_{correlation}$
Memorability (Isola et al. 2011)	$\surd$ (4)	$\surd$ (5)	$\surd$ (2)	$\surd$ (3)	$\surd$ (1)	-	0.61
Importance (Yamaguchi et al. 2012)	-	-	-	$\surd$(2)	$\surd$(1)	-	0.28
Interestingness (Gygli et al. 2013)	$\surd$ (4)	$\surd$ (3)	$\surd$(2)	$\surd$(1)	-	-	0.30
Popularity (Khosla, Sarma, and Hamid 2014)	$\surd$ (4)	$\surd$(2)	$\surd$(1)	$\surd$(5)	$\surd$(3)	-	0.43
Virality (Deza and Parikh 2015)	$\surd$(5)	$\surd$(4)	$\surd$(3)	$\surd$(1)	$\surd$(2)	-	0.55
Impressiveness	$\surd$(6)	$\surd$ (5)	$\surd$ (3)	$\surd$ (4)	$\surd$ (1)	$\surd$ (2)	1.00

Table 4: The recommendation performance on four topics based on different representations, *i.e.*, SIFT encoded by BoW, the concatenation of low-level features including texture, color, gradient and scene, the proposed impressiveness. Three common metrics are employed for evaluation.

Metrics	Method	Technology	Politic	Sport	Travel
MAP	SIFT&BoW	0.515	0.664	0.495	0.475
	Low-level	0.544	0.786	0.603	0.590
	Ours	**0.630**	**0.807**	**0.763**	**0.601**
DCG	SIFT&BoW	0.633	0.758	0.563	0.569
	Low-level	0.663	0.799	0.678	0.685
	Ours	**0.733**	**0.842**	**0.831**	**0.698**
MRR	SIFT&BoW	0.698	0.842	0.698	0.543
	Low-level	**0.768**	0.939	0.777	0.800
	Ours	0.759	**0.959**	**0.857**	**0.835**

which consist of 125, 97, 160 and 117 images, respectively. We then conduct experiments to evaluate the recommendation ability using different image representations/properties. To verify the effectiveness of the proposed impressiveness, we introduce two baseline image representations. First, we extract the SIFT features inspired by the work in content-based image retrieval. Second, we represent impressiveness using only low-level features, *i.e.*, texture, color, gradient and scene, which describe fundamental content information of images. We concatenate the low-level features to form the second baseline representation. Finally, we add the cues considering objects and emotions into the second baseline representation to form the proposed impressiveness property. For image recommendation, we take an arbitrary image with high-impressiveness as the query to retrieve relevant images in the corresponding set. To rank result images, we calculate the Euclidean distance from the query to all images based on the image representation/property.

We evaluate the performance of the recommendation with three common measurements, *i.e.*, the Mean Average Precision (MAP), the Discounted Cumulative Gain (DCG) and the Mean Reciprocal Rank (MRR). MAP is the mean of the average precision scores for each query. DCG measures the ranking quality of an item based on its relevance and position. MRR is the multiplicative inverse of the rank of the first correct answer. All results of the three metrics range from 0 to 1 and higher value indicates better performance.

Table 4 demonstrate the recommendation performance of the two baselines and impressiveness-based approach. It is easy to see that the proposed approach outperforms the baselines. For example, it achieves higher MAP of 0.807 on the topic of politics while the SIFT&BoW achieves only 0.664. We see that the proposed impressiveness is very effective in image recommendation for suggesting impressive images for the convenience of the users. And we believe it can also benefit other vision/graphics applications which rely on efficient image exploration and retrieval.

Conclusion

In this paper, we propose a novel image property, called *impressiveness*, that measures short-term impression of an image to the viewers. This is realized by visually qualifying six instantaneous human perceptual cues. We first construct three benchmark datasets from different application contexts. Then we quantify image impressiveness using measurable visual features via pre-labeled images with different level of impressiveness. We map multi-modal perceptual cues to visual representations of different levels, which are then fused to boost the performance on impressiveness prediction. We also verify the distinctiveness of the proposed impressiveness from existing image properties. Finally, we apply the impressiveness property to image recommendation to demonstrate its usefulness. All analysis exhibit that the impressiveness is predictable and effective for evaluating the image impression.

Acknowledgments

This research was sponsored by NSFC (61620106008, 61572264), CAST (YESS20150117), Huawei Innovation Research Program (HIRP), and IBM Global SUR award.

References

Bhattacharya, S.; Nojavanasghari, B.; Chen, T.; Liu, D.; Chang, S. F.; and Shah, M. 2013. Towards a comprehensive computational model foraesthetic assessment of videos. In *ACM MM*.

Bhattacharya, S.; Sukthankar, R.; and Shah, M. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM MM*.

Borkin, M. A.; Vo, A. A.; Bylinskii, Z.; and Isola, P. 2013. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*.

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

Chen, T.; Salaheldeen, H. M.; He, X.; Kan, M. Y.; and Lu, D. 2015. Velda: Relating an image tweet's text and images. In *AAAI*.

Colombo, C.; Del Bimbo, A.; and Pala, P. 1999. Semantics in visual information retrieval. *IEEE Multimedia*.

Danescu-Niculescu-Mizil, C.; Cheng, J.; Kleinberg, J.; and Lee, L. 2012. You had me at hello: how phrasing affects memorability. *Social Science Electronic Publishing* 1:892–901.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *ECCV*.

Deza, A., and Parikh, D. 2015. Understanding image virality. In *CVPR*.

Dhar, S.; Ordonez, V.; and Berg, T. L. 2011. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*.

Fiske, S. T., and Taylor, S. E. 1984. *Social cognition*. Addison-Wesley Pub.

Gardner, W. L., and Martinko, M. J. 1988. Impression management in organizations. *Journal of Management: Official Journal of the Southern Management Association*.

Geng, L., and Hamilton, H. J. 2006. Interestingness measures for data mining:a survey. *ACM Computing Surveys*.

Gilron, R., and Gutchess, A. H. 2012. Remembering first impressions: Effects of intentionality and diagnosticity on subsequent memory. *Cognitive Affective & Behavioral Neuroscience*.

Guerini, M., and Staiano, J. 2015. Deep feelings: A massive cross-lingual study on the relation between emotions and virality. In *WWW*.

Gygli, M.; Grabner, H.; Riemenschneider, H.; Nater, F.; and Gool, L. V. 2013. The interestingness of images. In *ICCV*.

Hanjalic, A. 2006. Extracting moods from pictures and sounds: towards truly personalized tv. *IEEE Signal Processing Magazine*.

Hastie, R., and Park, B. 1986. The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review*.

Hirabayashi, F.; Matoba, H.; and Kasahara, Y. 1988. Information retrieval using impression of documents as a clue. In *ACM SIGIR*.

Isola, P.; Xiao, J.; Torralba, A.; and Oliva, A. 2011. What makes an image memorable? In *CVPR*.

Jamieson, S. 2005. Likert scale: how to (ab)use them. *Medical Education*.

Jas, M., and Parikh, D. 2015. Image specificity. In *CVPR*.

Khosla, A.; Xiao, J.; Isola, P.; Torralba, A.; and Oliva, A. 2012. Image memorability and visual inception. In *SIGGRAPH Asia*.

Khosla, A.; Raju, A. S.; Torralba, A.; and Oliva, A. 2015a. Understanding and predicting image memorability at a large scale. In *ICCV*.

Khosla, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2015b. Memorability of image regions. In *NIPS*.

Khosla, A.; Sarma, A. D.; and Hamid, R. 2014. What makes an image popular? In *WWW*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Kumamoto, T., and Tanaka, K. 2005. *Proposal of Impression Mining from News Articles*. Springer Berlin Heidelberg.

Lehman, J., and Stanley, K. O. 2012. Beyond open-endedness: Quantifying impressiveness. In *Simulation and Synthesis of Living Systems*.

Li, L. J.; Su, H.; Xing, E. P.; and Li, F. F. 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*.

Lodge, M.; McGraw, K. M.; and Stroh, P. 1989. An impression-driven model of candidate evaluation. *American Political Science Review*.

Luo, W.; Wang, X.; and Tang, X. 2011. Content-based photo quality assessment. In *ICCV*.

Machajdik, J., and Hanbury, A. 2010. Affective image classification using features inspired by psychology and art theory. In *ACM MM*.

Newell, A. 1972. *Human Problem Solving*. Prentice-Hall.

Ojala, T.; Pietikinen, M.; and Menp, T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*.

Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*.

Parikh, T.; Isola, P.; Parikh, D.; Torralba, A.; and Oliva, A. 2012. Isola understanding the intrinsic memorability of images. *Journal of Vision*.

Sethi, I. K. 2000. Content-based multimedia information retrieval. *New Technology of Library & Information Service*.

Siersdorfer, S.; Minack, E.; Deng, F.; and Hare, J. 2010. Analyzing and predicting sentiment of images on the social web. In *ACM MM*.

Tuch, A. N.; Presslaber, E. E.; Stcklin, M.; Opwis, K.; and Bargas-Avila, J. A. 2012. The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International Journal of Human-Computer Studies*.

Wang, W.; Wang, H.; Zhang, C.; and Gao, Y. 2017. Fredholm multiple kernel learning for semi-supervised domain adaptation. In *AAAI*.

Willis, J., and Todorov, A. 2006. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*.

Wyer, R. S., and Srull, T. K. 1986. Human cognition in its social context. *Psychological Review*.

Yamaguchi, K.; Stratos, K.; Berg, A. C.; Sood, A.; Mitchell, M.; Mensch, A.; Goyal, A.; Han, X.; Dodge, J.; and Daume, H. 2012. Understanding and predicting importance in images. In *CVPR*.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*.

Zhao, S.; Yao, H.; Gao, Y.; Ji, R.; and Ding, G. 2017. Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Transactions on Multimedia*.