An Automated Estimator of Image Visual Realism Based on Human Cognition

Shaojing Fan[1,2], Tian-Tsong Ng[2], Jonathan S. Herberg[4], Bryan L. Koenig[3], Cheston Y. -C. Tan[2], and Rangding Wang*[1]

[1]Ningbo University
[2]Institute for Infocomm Research
[3]Lindenwood University
[4]Institute of High Performance Computing

Abstract

Assessing the visual realism of images is increasingly becoming an essential aspect of fields ranging from computer graphics (CG) rendering to photo manipulation. In this paper we systematically evaluate factors underlying human perception of visual realism and use that information to create an automated assessment of visual realism. We make the following unique contributions. First, we established a benchmark dataset of images with empirically determined visual realism scores. Second, we identified attributes potentially related to image realism, and used correlational techniques to determine that realism was most related to image naturalness, familiarity, aesthetics, and semantics. Third, we created an attributes-motivated, automated computational model that estimated image visual realism quantitatively. Using human assessment as a benchmark, the model was below human performance, but outperformed other state-of-the-art algorithms.

1. Introduction

Visual realism is defined as the degree an image appears to people to be a photo rather than computer generated. Predicting image visual realism is a challenging yet important task for the visualization and CG communities. For instance, image realism could be used as a metric for CG image quality evaluation or during manipulation of the realism level of computer games. Image realism could also be integrated into content-based image retrieval and image forensics.

Over the last decade, some noteworthy research has provided a base for understanding visual realism. In the CG community, scholars have analyzed the impact of rendering parameters like illumination and shadow on how similar a CG image is to reality, *i.e.*, its *CG fidelity* [17, 21]. In the computer vision field, much research has been devoted to detecting and improving the realism of composite images [14, 28]. However, we are unaware of any research that has systematically analyzed the perceptual factors relevant to the visual realism of images of general scenes, and how these perceptual factors could be turned into a quantitative realism estimation problem. Current datasets in related fields only contain labels of image type, with no ground truth on realism score. Therefore they are not suitable for quantified realism assessment. Besides, there is no set of unified evaluation criteria for such a quantitative estimation.

Our research differs from previous work in computer vision on image-type classification. Our method is realism-centric, focusing on estimating the realism level of individual images regardless of their types (Fig. 1). In this paper, we

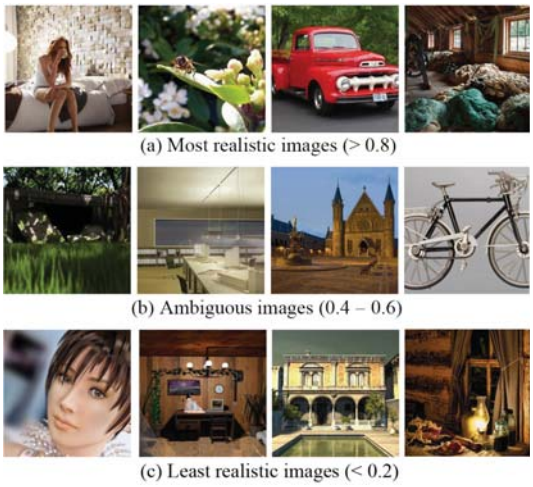

Figure 1. Image type may not indicate visual realism – photos may appear unrealistic whereas CG images can appear very real. Above are images of different realism levels from our visual realism dataset. Half of the images in each row are CGs, half are photos. The number in parentheses represents the realism score (the proportion of participants who rated the image as a photo rather than as CG).

*Corresponding author. E-mail:wangrangding@nbu.edu.cn

Figure 2. Sample CG images (top) and photos (bottom) from our dataset distributed based on degree of realism. The numbers on the bar represent realism score (the proportion of participants who rated the image as a photo rather than as CG).

develop a computational approach to realism estimation that incorporates factors empirically related to human realism assessment. The paper has three goals. First, to construct a unified benchmark dataset for quantitative realism estimation (Sec. 2). Second, to explore the high level attributes related to visual realism of images (Sec. 3). Third, to develop a model rooted in machine-learning for automatically inferring realism of images using their visual content, and to assess model performance in terms of the degree to which it matches human performance (Sec. 4).

1.1. Related work

CG fidelity: Since the early 1980's, research has explored CG fidelity [17, 21]. A common approach has been controlled experiments in which participants judge between a real scene and its CG replica generated with different parameter settings. Recent work [7] showed that realism perception of face images is related to intrinsic image components such as shading and reflectance as well as cognitive factors such as viewers' expertise and ethnicity. However, these studies were conducted using datasets limited by specific scenes and small sample sizes. The current work included additional visual factors and other image attributes important to visual realism. We based our study on a large-scale dataset with a variety of scenes.

Image type classification: Computer vision researchers tend to focus on how to classify images as photos or CG based on various image characteristics, such as higher-order wavelet statistics [16], physics-motivated geometry features [19], and physical noise rendered by cameras [5]. However, these methods are not directly rooted in human perception, which is an essential contrast to our approach. Although these algorithms developed apart from considering human perception can often reach high classification accuracy, the features used are usually sensitive to such image manipulations as compression and post-processing. Our work on visual realism differs fundamentally from the previous photo-vs-CG classifiers in three ways: first, visual realism is perceptual; image type is not. Second, realism scores range from 0 to 1, whereas photo-vs-CG is a binary distinction. Image type does not necessarily indicate image realism level, and vice versa (Fig. 1). Third, our study included matte paintings, which are hybrid images characterized naturally by visual realism but not by image type as photo or CG.

Image composites evaluation: Some studies have focused on understanding and assessing realism of composite images [14, 28]. Evaluation of various image statistical measures has indicated that the most important factors for realism of composite images are illumination, color, and saturation.

2. Visual realism benchmark dataset

We established a benchmark dataset based on quantitative measures of the visual realism of each image. Visual realism scores were collected from a large-scale psychophysics study on Amazon Mechanical Turk (MTurk). The following section describes the assembly of the dataset and the study.

2.1. Dataset construction

The dataset consists of 2520 images, among which half are photos and half are CG images. Sample images and dataset statistics are shown in Fig. 1, 2 and 3.

A good dataset should reflect the types of images we encounter in daily life. However, digital technology has advanced sufficiently that hybrid images whose image type is difficult to determine have become common. For instance, digital matte painting (MP) images are now common in movies. Digital MP images are often composed of a CG image superimposed on a base plate (a photo or moving footage; Fig. 4). Our dataset differs from others in related fields in that it includes digital MP images. We selected those for which over $1/3$ of their area was CG.

We excluded CG images with unrealistic content, like spaceships flying in a city. Obviously unrealistic CG images like cartoons were also excluded. All images were scaled and cropped about their centers to be 256×256 pixels.

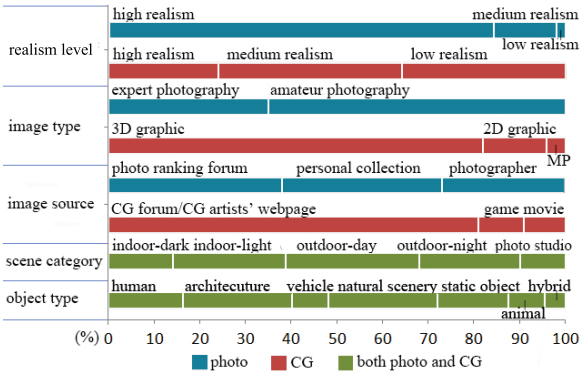

Figure 3. A depiction of the variation across our visual realism dataset. Each bar is labeled by variety-category (leftmost labels). Within each category, different specific features, labeled above the sub-bars, apply to CG images versus photos (except the last two rows). Sub-bar lengths represent proportions. In the first category, high realism, medium realism, low realism indicate realism scores between the ranges of $(.67,\ 1]$, $(.33,\ .67]$, $[0,\ .33]$, respectively.

Figure 4. An example of digital matte painting. Left: Final matte painting. Right: Original image before applying matte painting. Courtesy of Matte World Digital, CA.

2.2. Psychophysics study I: perceptual realism

Study design: We had workers on MTurk view a sequence of images and judge each as "CG" or "photo". We defined CG images as entirely or in part created using computer software. To estimate how diverse our participants were regarding prior familiarity to CG images and photography, we asked participants to select one or more options that best fit their background from "have jobs related to graphic design", "keen computer game players", "photographers or photography enthusiasts", and "laypersons". We paid workers \$1.00 for completing the task, and to encourage participants to try their best we paid a \$0.20 bonus to workers whose accuracy exceeded 90%.

Empirical realism score: We performed a pilot study to determine how many participants are necessary to provide sufficient reliability for visual realism assessment. Split-half correlations and root mean square error analysis suggested that 30 judgments per image is enough (for details see supplementary material [1]). Based on the pilot study, we recruited 1292 participants from MTurk (for all studies we required workers to have $> 95\%$ approval rating in Amazon's system). Each image was judged by a mean of 31 participants. We calculated a *realism score* (ranges from 0 to 1) for each image as the proportion equal to the number of judgments indicating that the image is a photo over the total number of judgments for that image. The distribution of realism scores and sample images of different realism levels are shown in Fig. 3, and 2, respectively.

2.3. Dataset statistics

We wanted our findings to generalize to the various types of images people often see. We also hoped the computational model built on the dataset could be generalizable in terms of image type, realism level, and image content. So significant diversity in images is important. Fig. 3 summarizes the statistics of our dataset. For more detailed information of our dataset, users could refer to our project website [1].

3. Measuring attributes and visual realism

In order to assess images' visual realism by constructing a computational model similar to human perception, we first investigated image attributes relevant to people's visual realism perception and modeled visual realism empirically.

3.1. Psychophysics study II: attributes annotation

We recruited a new group of 3794 MTurk participants to annotate the images (Table 1; for complete questionnaire see supplementary material [1]). On average, 10 participants annotated each image. We also had images labeled via LabelMe [23], an online annotation tool (a_{31-32} in Table 1). Due to budget constraints, these tasks were done for half of the entire image set. These 1260 images, which we refer to as the *annotated subset*, were selected so they had realism scores distributed as uniformly as possible for both photos and CG images, over the entire realism score range.

3.2. Correlation of attributes and visual realism

We measured and investigated the relationships between image attributes and visual realism by using the realism scores we got in Study I (Sec. 2.2) as ground truth. We used the Spearman's rank-order correlation (ρ) and one-way ANOVA [22] to assess such relations (see Table 1 and Fig. 5).

Realism ratings: We asked participants to rate the degree to which images appeared to be a photograph versus computer generated (a_1) on a five-point scale (1 = computer generated, 5 = photograph). These ratings strongly correlated with the human realism scores we got in Study I ($\rho = .80$). The participants for the two tasks were different, so this demonstrates the stability of human perception of visual realism over both measurements.

Familiarity: Familiarity attributes (a_2, a_{4-5}) correlated substantially with realism (ρs = .23, $-.33$, $-.36$, respectively). This might be because people obtain greater capacity for assessing image realism from prior exposure to similar scenes. Consistent with this, previous research suggests that people have specific memories of common objects entities such as

Table 1. Image attributes (Attr), related survey item, attributes category, and their Spearman's rank correlations (ρ) with ground truth image realism scores (from Study I). Meaningful and statistically significant correlations ($|\rho| > .15$, $p < .05$) are highlighted in bold. Numbers in parentheses are participants' mean ratings for each attribute standardized to a scale of 0 to 1.

Attr	Survey item	Category	ρ	Attr	Survey item	Category	ρ
a_1	Appears to be a photograph? (.68)	Realism	**.80***	a_{21}	Clean scene and objects? (.83)	Layout	.07*
a_2	Familiar with the scene? (.60)	Familiarity	**.23***	a_{22}	Makes you happy? (.60)	Emotions	.08
a_3	Familiar with the objects? (.76)	Familiarity	.15*	a_{23}	Makes you sad? (.08)	Emotions	-.10
a_4	Unusual or strange? (.28)	Familiarity	**-.33***	a_{24}	Exciting? (.56)	Emotions	**-.16***
a_5	Mysterious? (.32)	Familiarity	**-.36***	a_{25}	Contain fine details? (.58)	Texture	-.03
a_6	Lighting effect natural? (.74)	Illumination	**.49***	a_{26}	Dynamic scene? (.33)	Semantics	-.15*
a_7	Shadows in the image? (.60)	Illumination	-.15*	a_{27}	Is there a storyline? (.43)	Semantics	**-.25***
a_8	How sharp are the shadows? (.37)	Illumination	-.07*	a_{28}	Contain living objects? (.36)	Semantics	.06
a_9	Color appearance natural? (.82)	Color	**.47***	a_{29}	Naturalness of objects? (.77)	Semantics	**.36***
a_{10}	Colors go well together? (.88)	Color	.15*	a_{30}	Object combinations natural? (.76)	Semantics	**.20***
a_{11}	Colorful? (.53)	Color	.05	a_{31}	Number of unique objects (.60)	Semantics	-.09*
a_{12}	Image quality (.69)	Quality	.04	a_{32}	Total number of objects (.72)	Semantics	-.06
a_{13}	Image sharpness (.72)	Quality	.10*	a_{33}	Number of people (.49)	Human semantics	-.08
a_{14}	Expert photography? (.57)	Aesthetics	**.33***	a_{34}	Face visible? (.18)	Human semantics	.24*
a_{15}	Attractive to you? (.69)	Aesthetics	.03	a_{35}	Is the person attractive? (.35)	Human semantics	-.12
a_{16}	Close-range or distant-view? (.63)	Layout	.04	a_{36}	Making eye contact with viewer? (.12)	Human semantics	.13
a_{17}	Have objects of focus? (.71)	Layout	.00	a_{37}	Posing for the image? (.22)	Human semantics	-.10
a_{18}	Neat space? (.70)	Layout	.10*	a_{38}	Human activities (.48)	Human semantics	.01
a_{19}	Empty space? (.48)	Layout	.01	a_{39}	Human expressions (.40)	Human semantics	.03
a_{20}	Perspective natural? (.75)	Layout	**.33***	a_{40}	Expression genuine? (.43)	Human semantics	.38*

* $p < .05$.

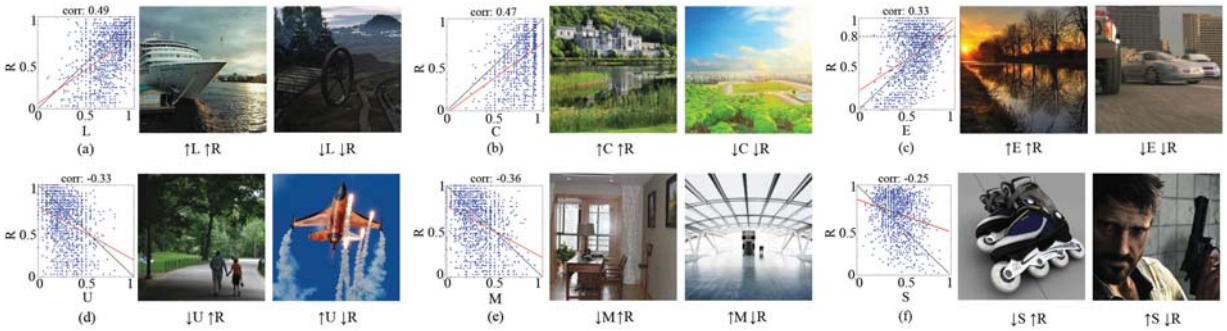

Figure 5. Distribution of realism R of images with respect to lighting naturalness L (a), color naturalness C (b), degree of expert photography E (c), unusualness U (d), mysteriousness M (e), and degree of having a storyline S (f). Also shown are example images that demonstrate such correlations (*e.g.* the left image in (f) shows an image that does not seem to have a storyline, but is more realistic). In the left graph of each set, the black line stands for $y = x$ (first row) or $y = 1 - x$ (second row), the red line is the linear regression line of all image points.

the sky or skin. Therefore an image may look more natural or realistic if the coloring of image entities coheres with memory representations [14, 3].

Color: Color naturalness (a_9) moderately correlated with realism (ρ = .47), which is consistent with previous findings on image composites [14, 28]. However, there was no significant correlation between colorfulness (a_{11}) and realism, which contrasts with [3], who found that colorfulness was a key attribute to image naturalness. This may imply that image naturalness and image visual realism are not based on entirely the same perceptual processes. In previous studies, *naturalness* was defined as the degree of correspondence between an image presented on an imaging device and memories of real-life scenes [3], whereas we define *visual realism* as the degree to which an image appears to be a photograph versus computer generated. As seen from their distinct definitions, naturalness and visual realism have intrinsic differences in terms of evaluation criteria and perceptual process.

Illumination: The naturalness of lighting (a_6) correlated moderately with realism (ρ = .49), suggesting the importance of illumination for realism. This accords with previous research which suggests that image properties like illumination, shadow, and surface roughness are important factors for CG fidelity [21, 7]. However, we did not observe meaningful correlation between shadow characteristics (a_{7-8}) and realism. This contrasts with prior research suggesting that shadow softness is an important factor for CG fidelity [21].

This difference might be because [21] used images of simple objects, while our images consisted of varied scenes, entailing more complex and varied shadowing effects. Alternatively, whereas [21] used a fixed viewing environment and rendering parameters, ours were uncontrolled.

Aesthetics: The degree to which an image appeared to be a work of expert photography (a_{14}), an aesthetics attribute, moderately correlated with realism ($\rho = .33$). Interestingly, this correlation was negative ($\rho = -.23$) for images with realism scores greater than 0.8. This might suggest that more aesthetics in a highly realistic image can lower its realism. This is consistent with prior research on human skin rendering [8], which suggests that maximal attractiveness and extreme realism were opposing perceptions. Despite the somewhat non-linear relationship between aesthetics and realism, we still used linear regression for simplicity. Modeling the non-linear relationship is left to future work.

Spatial layout: We found that the naturalness of perspective (a_{20}) influences realism ($\rho = .33$). This has been noted in the CG community by [6], who investigated the impact of viewpoint on apparent realism of virtual crowds.

Semantics: The naturalness of object appearances (a_{29}) and of object combinations (a_{30}) both correlated moderately with realism ($\rho = .36, .20$, respectively). However object statistics (a_{31-32}) did not appear to influence realism. This accords with [21], who showed that the number and diversity of objects have minimal influence on realism. The amount of semantic information an image conveys (a_{27}) negatively correlated with realism ($\rho = -.25$), suggesting that explicitly dramatic scenes appear less realistic. We performed one-way ANOVAs to investigate the effect of scene and object type on visual realism (for detailed categories see Fig. 3). Results suggested a significant effect of scene and object types on realism, $Fs(12, 2507) > 4.81$, $ps < .05$.

3.3. Empirical visual realism model

We used feature selection and multiple regression to determine which factors most influenced visual realism. Finally, image visual realism was modeled by the major factors based on the psychophysical data.

Feature selection: We used the attributes as features for training support vector regressor (SVR) [2] to predict image realism. We used grid search to select cost, RBF kernel parameter γ, and ϵ hyperparameters. We split the 1260 images from the annotated subset into 80% as a training set and 20% as a test set. We performed a greedy feature selection. The prediction performance was evaluated using Spearman's rank coefficient between predicted realism scores and human realism scores (from Study I). As shown in Fig. 6, performance improved with more attributes, but improved little with more than 10 attributes. Therefor we selected the 10 top attributes for modeling visual realism. Some attributes with small correlations with realism individually had stronger correlations jointly (Fig. 6), such as attractiveness (a_{15}), image quality (a_{12}), and presence of living objects (a_{28}).

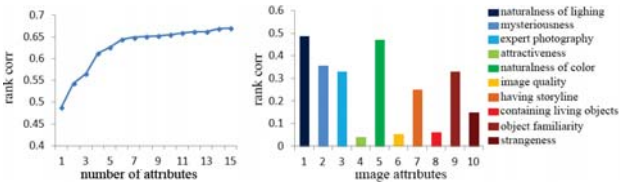

Figure 6. Feature selection results. Left: Spearman's rank correlation between predicted realism score and human realism scores as a function of the number of predictor attributes. Right: Independent prediction performance of top 10 attributes.

Principal component factor analysis: Several attributes from feature selection were correlated, such as mysteriousness and strangeness. We performed a principal component (PC) factor analysis with varimax rotation [22] to remove the high inter-correlations and identify a compact set of attributes related to realism. The 10 attributes from feature selection were grouped into 4 major PCs based on factor analysis, which most strongly correlated with naturalness, aesthetics, familiarity, and semantics, respectively (Table 2). The "Cumulative variability" row shows that the 4 PCs accounted for nearly 65% of the variability in the 10 attributes.

Multiple regression: Finally, PC scores were computed as a weighted average of the 10 attributes (with factor loadings as weights). We predicted realism scores with these PC scores using multiple regression, adjusted $R^2 = .44$, $p < .001$. Seen from Table 3, naturalness strongly predicted realism, while aesthetics, familiarity and semantics weakly but significantly predicted realism. The relative predictive ability of this statistical model is consistent with the computational performance of each component presented in Sec. 4.2 (Table 4).

Table 3. Principal components, their standardized coefficients (β), t value, and significance (p) in multiple regression with realism.

Component	β	t	p
Naturalness	.63	29.98	.000
Aesthetics	.14	6.53	.000
Familiarity	-.11	-5.19	.000
Semantics	-.12	-5.54	.000

4. Computational visual realism

We designed features motivated by image attributes relevant to visual realism. We built a computational model for quantitative realism assessment based on these features. We also compared our model with state-of-the-art algorithms using human realism scores as a benchmark.

4.1. Image features for visual realism

Based on our psychophysics studies, visual realism correlated strongest with naturalness, aesthetics, familiarity, and

Table 2. The loadings of 10 selected attributes on the 4 major principal components (PC). Bold numbers are the strongest loading of each attribute on one of the PCs. The "Cumulative variability" row shows how each PC cumulatively explains the variability of 10 attributes in presented sequence.

Attributes \ Principal components	1 (Naturalness)	2 (Aesthetics)	3 (Familiarity)	4 (Semantics)
Naturalness of color appearance (a_9)	**.88**	.13	-.19	.03
Naturalness of lighting effect (a_6)	**.87**	.20	-.17	.02
Image quality (a_{12})	.06	**.79**	-.03	-.05
Attractiveness (a_{15})	.10	**.74**	-.20	.18
Expert photography (a_{14})	.44	**.67**	-.02	-.10
Unusualness/strangeness (a_4)	-.07	-.07	**.71**	.03
Mysteriousness (a_5)	-.10	-.03	**.71**	.19
Objects familiarity (a_3)	.27	.17	**-.64**	.21
Containing living objects(a_{28})	.10	.00	-.16	**.78**
Having storyline (a_{27})	-.08	.03	.31	**.71**
Cumulative variability explained (%)	18.41	35.53	52.03	64.36

semantics. We identified automated methods to determine feature values that corresponding to these attributes. Instead of simple concatenation, we applied kernel sum to fuse the features for support vector regression.

Naturalness: We modeled image naturalness in three ways. First, [7] suggested that shading and reflectance affect visual realism differently. This inspired us to model naturalness using intrinsic image components. We first decomposed each image into intrinsic components by extending Retinex algorithm into RGB space [9]. We then computed three 256-bin histograms for each image, to represent shading and reflectance components as well as original image. We further calculated the histogram difference between the intrinsic components and original image. Second, based on [25] we calculated the image naturalness statistics derived from the local patch (3×3) structures and image power spectrum. Finally, unnaturalness was modeled by using the method in [10], who identified simple and uniform color, and strong edges, as characteristics of CG.

Aesthetics: We applied Ke's method [11] for extracting aesthetics features, which considers image properties like edges distributions, blur, and contrast. We also used local self-similarity geometric patterns (SSIM [24]) to represent content symmetry, which is often regarded as a measure of aesthetics. We densely sampled the SSIM descriptors with a grid spacing of 4 and learned a dictionary of size 100. We used 2-level spatial pyramid pooling on the descriptors.

Familiarity: First, we defined a measure for semantic familiarity using the content-based similarity measure commonly used in image retrieval. We used $10,000$ images from the SIMPLIcity dataset [27] as a pre-determined *anchor* database of images with common scenes and objects. We then computed the image similarity by using color, illumination and texture information [13], and performed a robust content-based matching with the anchor database. Primarily meant for image retrieval applications, we used it here to quantify familiarity. The familiarity measure was denoted by the distances of the top 50 matches. Second, [14, 3] suggested that an image may look more realistic if its coloring coheres with memory representations. We included color compatibility [14] as a measure for color familiarity. We also included color name features learned from real-world images [26] to better represent daily color compositions. We densely sampled the feature with a grid spacing of 4 and learned a dictionary of size 256. We then applied 2-level spatial pyramid pooling to obtain the color descriptors.

Semantics: We applied GIST [20] to model scene structure using 4×4 image block. We used automatic Object Bank (OB) [15] to model presence of a pre-defined set of objects. In OB, an image is represented as a collection of response maps of a large number of pre-trained generic object detectors. We used max pooling on OB features.

4.2. Results and evaluation

Evaluation methods: We evaluated our method by its ability to predict realism scores. As a simple application, our model was also evaluated in classifying images as photos or CG. For prediction, we used human realism scores from Study I as ground truth, and Spearman's rank correlation to evaluate the prediction performance from SVR. For classification, image-type labels were used as ground truth, and area under ROC curve as evaluating measure (where realism scores from SVR were treated as image-type probability, high realism scores correspond to photo and low correspond to CG). The SVR settings are as described in Sec. 3.3.

Evaluation results: In Table 4 and Fig. 7, we compared performance of various computational methods, as well as human judgment and attributes annotation that motivated our features. For human judgment, we treated human realism scores from Study I (Sec. 2.2) as image-type probability in

Table 4. Experimental results of realism prediction and image classification. ρ_1 and A_1 are respectively the Spearman's rank correlation, and area under ROC curve on annotated subset, ρ_2 and A_2 are those on whole dataset[1]. The best result from computational features on each evaluation metric is highlighted in bold.

Category	Feature type	Prediction		Classification	
		ρ_1	ρ_2	A_1	A_2
Human	Human	.65[2]	n.a.	.79	.88
Attributes annotation	Naturalness	.52	n.a.	.62	n.a.
	Aesthetics	.39	n.a.	.64	n.a.
	Familiarity	.39	n.a.	.57	n.a.
	Semantics	.30	n.a.	.61	n.a.
	All combined	.66	n.a.	.67	n.a.
Our method	Naturalness	.38	.45	.66	.74
	Aesthetics	.34	.42	.65	.73
	Familiarity	.33	.42	.64	.74
	Semantics	.28	.37	.61	.67
	All combined	**.41**	**.51**	**.68**	**.77**
Signal feature	Wavelet [16]	.16	.20	.56	.63
	Geometry feature [19]	.31	.47	.64	.74
	Camera noise [5]	.04	.06	.53	.50
	Color compatibility [14]	.20	.23	.57	.61
Object & scene feature	SIFT [12]	.28	.34	.61	.66
	GIST [12]	.16	.23	.58	.61
	HOG2x2 [12]	.28	.33	.58	.66
	LBP [12]	.25	.30	.59	.64
Feature learning	K-means encoding [4]	.28	.37	.63	.71

[1] Results are consistently better on whole dataset than annotated subset. It might be because the images in the subset was purposefully selected to make a uniform distribution on realism scores. Thus these images are intrinsically harder to be distinguished.

[2] This result is the split half consistency among participants for study II.

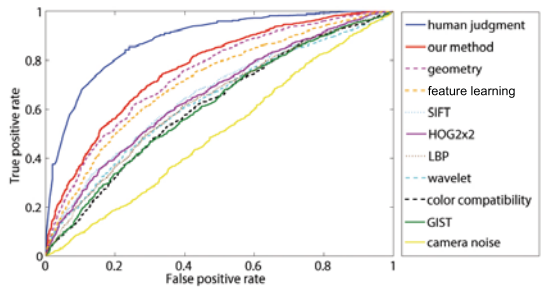

Figure 7. ROC curve of binary image classification on whole dataset. Our method outperforms other computer algorithms, yet is still far below human performance.

classification tasks. For attributes annotation, we grouped the 10 selected attributes in Sec. 3.3 into 4 components (defined in Table 2) and used them as training features for SVR. The attributes annotation was only on the annotated subset, so for comparison we tested all computer methods on both the whole dataset as well as the annotated subset.

We also compared our method with the signal processing features commonly used in CG and photo classification, which include high-order correlations of wavelet coefficients [16], physics-motivated geometry structure [19], camera noise [5], and color compatibility for evaluating the realism of image composites [14]. We further tested some well known object and scene features like SIFT, GIST, HOG2x2, and LBP, computed from an open-source library [12]. Finally, we investigated unsupervised feature learning. We adopted the unsupervised feature learning framework with a single-layer triangular K-means encoding [4] on image patches preprocessed by local intensity and contrast normalization, as well as whitening. During test, we scan an image with 16-by-16 pixel receptive field and 1 pixel stride, before mapping the preprocessed image patches to 256-dimensional feature vectors. The details on feature computation can be found in our project website [1].

Our results suggest the following three things:

First, both attributes annotation and our features predicted image realism moderately well (ρs $> .28$; Table 4). Among the four factors, naturalness predicted best, which is consistent with our regression model (Table 3), indicating naturalness is the most important factor among the 4 components.

Second, although the performance of our method was lower than that of attributes annotation in prediction task, our method slightly outperformed attributes annotation in classification task (Table 4). This suggests that our attributes-motivated features represent human annotation to a certain degree.

Third, our combined features outperformed other computer algorithms in all evaluation metrics, suggesting not only that our method is the most similar to human perception, but also that understanding human perception helps create better computational models. The low performance of the camera noise feature might be due to its sensitivity to image compression and post-processing. Unsupervised learning features were among the best, but humans performed the best on both tasks.

Limitation: As seen from Fig. 8, our method overpredicted realism for images with unusual scenes (including CG persons), whereas it underpredicted realism for images that are common scenes but with unusual illumination or image quality. This might indicate that one limitation of our method is on scene understanding. Investigating scene semantics might be fruitful. For example, we could fully utilize the data collected from LabelMe or explore image context.

5. Conclusion

In this paper we have shown that predicting image visual realism is a task that can be addressed with current computer vision techniques. We constructed an image realism benchmark dataset and designed a realism predictor motivated by human-annotated image attributes. To the best of our knowledge, this work is a first realism-centric study that

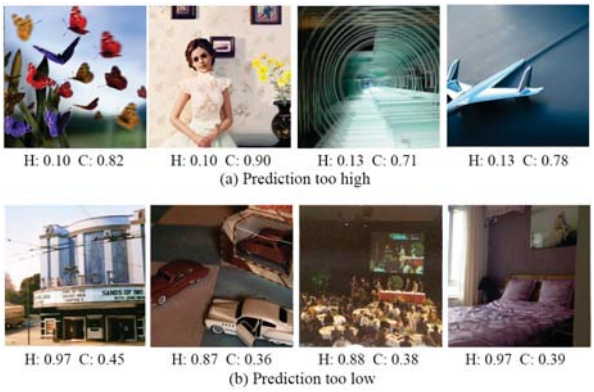

Figure 8. Samples of poorly predicted images by our method. The number on the left under each image is the ground truth realism score evaluated by humans (H), the number on the right is computer predicted realism score by our method (C).

attempted to quantify visual realism of individual images. We have shown a simple application of our realism predictor on image classification. For future work, we will incorporate our realism predictor for perception-based image retrieval and computer graphics rendering. We also plan to develop a web service for image realism prediction [18].

Acknowledgements

We thank Karianto Leman, Miao Jie and Zhang Fan for their help in this research. This work is partially supported by Open-end Fund in Information and Communication Engineering, Zhejiang, China (No. XKXL1313).

References

[1] Visual realism project. http://www1.i2r.a-star.edu.sg/~ttng/VisualRealism/index.html. 3, 7

[2] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2011. 5

[3] S. Y. Choi, M. Luo, M. Pointer, and P. Rhodes. Investigation of large display color image appearance-lll: Modeling image naturalness. *JIST*, 2009. 4, 6

[4] A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In *Conference on Artificial Intelligence and Statistics*, 2011. 7

[5] E. Dirik, S. Bayram, H. Sencar, and N. Memon. New features to identify computer generated images. In *ICIP, 2007*. 2, 7

[6] C. Ennis, C. Peters, and C. O'Sullivan. Perceptual effects of scene context and viewpoint for virtual pedestrian crowds. *ACM Transactions on Applied Perception (TAP)*, 2011. 5

[7] S. Fan, T.-T. Ng, J. Herberg, B. Koenig, and S. Xin. Real or fake?: Human judgments about photographs and computer-generated images of faces. In *Technical Briefs, ACM SIGGRAPH Asia*, 2012. 2, 4, 6

[8] F. Giard and M. J. Guitton. Beauty or realism: The dimensions of skin from cognitive sciences to computer graphics. *Computers in Human Behavior*, 2010. 5

[9] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009. 6

[10] T. I. Ianeva, A. P. de Vries, and H. Rohrig. Detecting cartoons: A case study in automatic video-genre classification. In *International Conference on Multimedia and Expo, 2003*. 6

[11] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR, 2006*. 6

[12] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In *NIPS*, 2012. 7

[13] Kirk. Content based image retrieval. https://github.com/kirk86/ImageRetrieval, 2013. 6

[14] J. Lalonde and A. Efros. Using color compatibility for assessing image realism. In *ICCV*, 2007. 1, 2, 4, 6, 7

[15] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010. 6

[16] S. Lyu and H. Farid. How realistic is photorealistic? *IEEE Transactions on Signal Processing*, 2005. 2, 7

[17] G. Meyer, H. Rushmeier, M. Cohen, D. Greenberg, and K. Torrance. An experimental evaluation of computer graphics imagery. *ACM Transactions on Graphics*, 1986. 1, 2

[18] T.-T. Ng and S.-F. Chang. An online system for classifying computer graphics images from natural photographs. In *Electronic Imaging*, 2006. 8

[19] T.-T. Ng and S.-F. Chang. Discrimination of computer synthesized or recaptured images from real images. In *Digital Image Forensics*. 2013. 2, 7

[20] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 2001. 6

[21] P. Rademacher, J. Lengyel, E. Cutrell, and T. Whitted. Measuring the perception of visual realism in images. In *Rendering Techniques 2001*. 1, 2, 4, 5

[22] J. A. Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2007. 3, 5

[23] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 2008. 3

[24] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR, 2007*. 6

[25] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, 2003. 6

[26] J. Van De Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. In *CVPR, 2007*. 6

[27] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *TPAMI*, 2001. 6

[28] S. Xue, A. Agarwala, J. Dorsey, and H. Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on Graphics, 2012*. 1, 2, 4