Study of naturalness in tone-mapped images

Quyet Tien Le, Patricia Ladret, Huu-Tuan Nguyen, Alice Caplier

Study of naturalness in tone-mapped images

Quyet-Tien Le[1,2], Patricia Ladret[1], Huu-Tuan Nguyen[2] and Alice Caplier[1]

[1]GIPSA Lab, Grenoble Alpes University
11 rue des Mathématiques, Grenoble Campus BP 46, F-38402 Saint Martin d'Hères Cedex FRANCE
[2]Faculty of Information Technology, Vietnam Maritime University
484 Lach Tray, Le Chan, Hai Phong, VIETNAM
*Corresponding author: Quyet-Tien LE

***Abstract*—Nowadays, images can be obtained in various ways such as capturing photos in single-exposure mode, applying Multiple Exposure Fusion algorithms to generate an image from multiple shoots of the same scene, mapping High Dynamic Range images to Standard Dynamic Range (SDR) images, converting raw formats to displayable formats, or applying post-processing techniques to enhance image quality, aesthetic quality,. . . When looking at some photos, one might have a feeling of unnaturalness. This paper deals with the problem of developing a model firstly to estimate if an image looks natural or not to humans and the second purpose is to try to understand how the unnaturalness feeling is induced by a photo: Are there specific unnaturalness clues or is unnaturalness a general feeling when looking at a photo? The study focuses on SDR images, especially on tone-mapped images. The first contribution of the paper is the setting of an experiment gathering human naturalness opinions on 1,900 SDR images mainly obtained from tone mapping operators. Based on the collected data, the second contribution of the paper is to study the efficiency of different feature types including handcrafted features and learned features for image naturalness analysis. A binary classification model is then developed based on the determined features to classify if an image looks natural or unnatural.**

Keywords: image unnaturalness study, high dynamic range image, standard dynamic range image, tone mapping operator, multiple exposure fusion, handcrafted features, learned features, transfer learning, convolutional neural network

I. Introduction

In recent years, more and more new camera models, photography techniques and image processing applications have been introduced to consumers. Three emphasises should be mentioned: High Dynamic Range (HDR) images, Multiple Exposure Fusion (MEF) algorithms and Tone Mapping Operators (TMOs). The dynamic range of images is the ratio between the highest and lowest luminance values. The dynamic range of irradiance in real scenes possibly reaches 1,00,000,000:1. The human eye can perceive the dynamic ranges from 10,000:1 to 1000,000:1 (depending on circumstances) while a normal display is able to present a low dynamic range (LDR - in recent years, LDR and SDR are considered as the same concept) from 100:1 to 300:1 [1], [2], [3]. As a consequence, the luminance range of scenes displayed on standard screens is narrower than that of real scenes and it is also lower than the dynamic range perception of human eyes. In the past, the problem of high dynamic range was caused by the camera sensors and the display devices. The camera sensors were not able to cover the whole irradiance range of real scenes. Nowadays, the capability of professional camera sensors has increased and those sensors can capture high dynamic range of almost normal scenes (14 stops of dynamic range - the dynamic range is 2^{14}:1). And when the dynamic range is too high to be covered (for example, 20 stops of dynamic range - the dynamic range is 2^{20}:1) or with a none-professional camera, Multiple Exposure Fusion (MEF) algorithms can be used to help covering the whole range. MEF is a technique generating an image from multiple shoots taken under different exposures for a given scene [4], [5], [6] by using fusion algorithms (see examples in Fig. 1). The MEF technique helps an image having a higher dynamic range than that of an image taken with a fixed exposure.

On the side of display devices, the work is in progress. Some of new commercial devices are able to present irradiance peaks around 1,000 $cd\ /\ m^2$ and black levels less than 0.05 $cd\ /\ m^2$ (the dynamic range is 20,000:1). Especially, some special models used in research can reach the highest luminance value of 10,000 $cd\ /\ m^2$. Although the dynamic range of new display devices is quite high, it is still quite modest when compared to the dynamic range of real scenes and the perception range of human eyes [7]. Thus, nowadays the problem of high dynamic range images is mainly related to display devices.

8 bit data is currently used to present images displayed on standard screens. Although there is no direct relation between bit-depth and dynamic range, it is necessary to use more steps (more bits) to present a higher dynamic range. A pixel of any HDR image is represented by 3 colors and each color is coded by 10 bits, 12 bits, 16 bits or 32 bits. Although some new monitor models (HDR monitors) are able to display a high dynamic range content (20,000:1), most of the popular display devices are SDR screens that are able to display only SDRs of irradiance. Thus, it is necessary to map HDR images to SDR format before display on SDR screens. To perform this task, many Tone Mapping Operators (TMOs) have been proposed [8], [3], [9], [10], [11], [12], [13]. Generally, TMOs map colors of HDR images from a high dynamic range (from 10,000:1 to 1000,000:1) to a low range (from 100:1 to 300:1), this process can be considered as a range compression process.

tienlqcnt@vimaru.edu.vn,quyet-tien.le@gipsa-lab.grenoble-inp.fr
patricia.ladret@gipsa-lab.grenoble-inp.fr
huu-tuan.nguyen@vimaru.edu.vn
alice.caplier@gipsa-lab.grenoble-inp.fr

Beside this, in order to correct the colors of images or to create special effects, some post-processing algorithms can be applied on SDR images. For example, in the first row of Fig. 2, post-production colors and contrast enhancements have been used to produce the image on the right. In the post-processed image of the second row, orange sky and sun-rays have been created by Adobe CameraRaw and Photoshop respectively. Additionally, an exposure enhancement algorithm has been used in that photo. In the last row, Nik Color Efex Pro and Photoshop have been used to enhance the colors and to create dodge and burning effects.

One problem of SDR images obtained by using those algorithms might be the loss of naturalness or the appearance of unnaturalness (see examples in Fig. 3). Those processing methods are like double-edge swords since they can significantly improve the image quality but they also can provoke the unnaturalness and decrease the image quality [14].

In this research, the naturalness concept is focused and it is defined on two sides. On one side, an image is considered as natural if the appearance of the image looks familiar for a human observer (it makes the observer have the feeling that the photo is a faithful representation of the scene). On the other side, if the observer has the feeling that something in the photo is wrong (due to color appearances, abnormal details or more subtil changes) so that the appearance of the photo does not look faithful, the photo is considered as unnatural. In our work, the naturalness concept is not supposed to be related to the image content itself (see Fig. 7). For example, augmented images are considered as natural in this study. The research focuses on collecting naturalness opinions from viewers to design features representing naturalness and unnaturalness. This study is not about image aesthetic assessment [15] or image quality assessment [16].

In this paper, there are two main contributions. The first one is an experiment of subjective image naturalness assessment without references. The experiment is conducted thoroughly at the laboratory with a set of SDR images obtained in various ways. The second contribution is the study of different features including handcrafted, shallow learned and deep learned features (features learned from shallow and deep convolutional neural networks respectively) for the image naturalness assessment task with the hope to define the best features representing naturalness / unnaturalness.

The paper is organized as follows. Section 2 presents the state of the art of image naturalness. Section 3 introduces the experiment of subjective image naturalness assessment and the dataset that has been collected. In section 4, feature definition and feature selection for image naturalness assessment are described. Section 5 presents the results of automatic natural / unnatural SDR image classification. Conclusions and some future works are presented in the last part.

II. Image naturalness studies: state of the art

In the literature, different definitions of image naturalness have been given. In [20], image naturalness is defined as the degree of correspondence between a photo displayed on a device and the memories about the corresponding real-life scene. An experiment is conducted with 13 observers, 8 color images and 22 manipulations of them to gather the perceived naturalness. The perceived naturalness is then compared with a naturalness index based on sharpness, colorfulness and reproduction of shadow details, memory colors of skin, grass and sky. In [17] and [18], image naturalness is defined as the same as in [20] but high quality images should be considered as natural. By analyzing the chromatic, hue, saturation and lightness variations, they point out the significant roles of those factors in image quality and image naturalness. But those studies only focus on evaluating the impacts of some factors on naturalness instead of finding factors affecting photo naturalness. In [19], image naturalness is defined as the degree of correspondence between a scene (seen directly) and the corresponding scenes in photos based

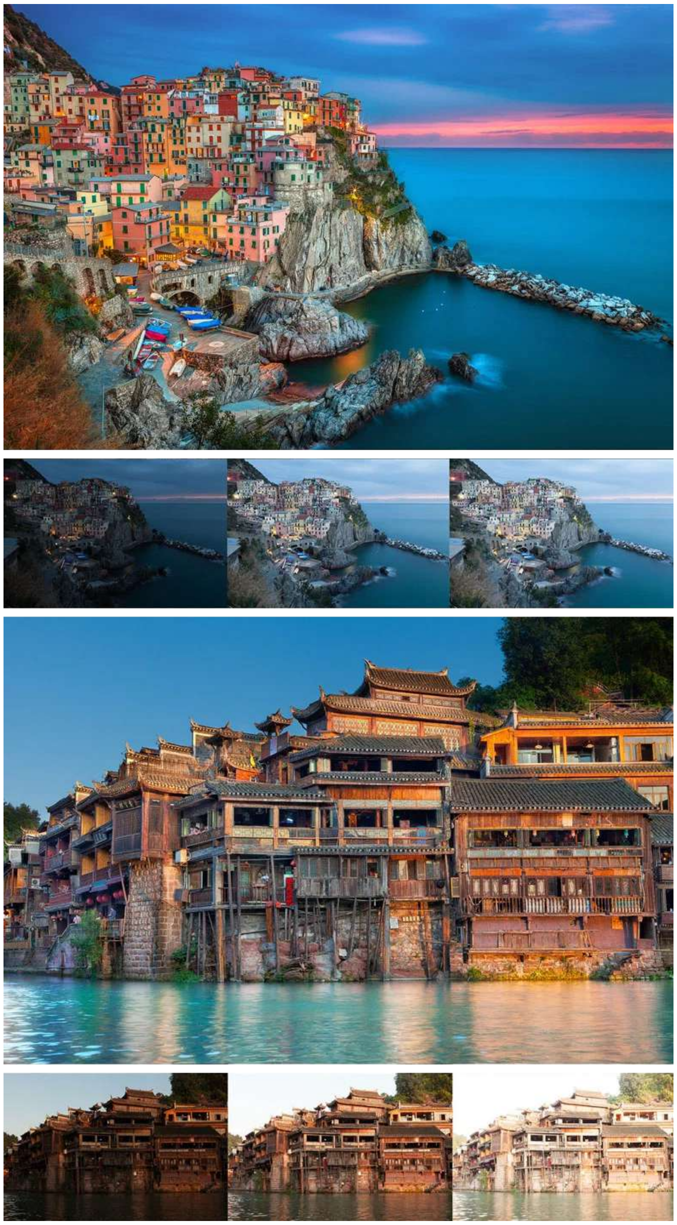

Fig. 1. Examples of MEF. The first and the third rows present images generated with the multi-exposure images of the second and the fourth rows respectively. (image source: https://petapixel.com).

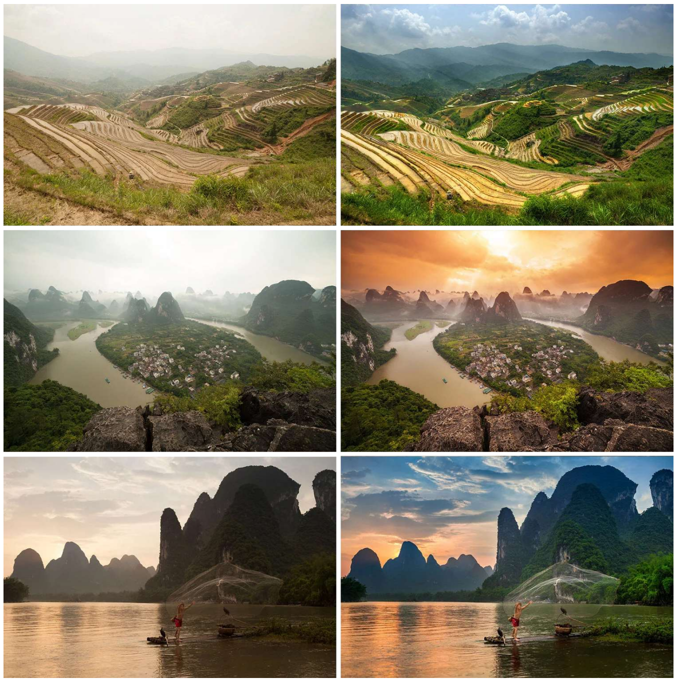

Fig. 2. Examples of post processing methods. The first column contains original images (produced directly by cameras) while the second one presents the corresponding post-processed images. (image source: https://petapixel.com).

TABLE I
OVERVIEW OF NATURALNESS INDEXES AND NATURALNESS FEATURES IN PREVIOUS STUDIES.

De's studies [17], [18]
Image naturalness is defined as the degree of correspondence with (memorized) reality. High quality images should be considered as natural.
Naturalness indexes: - Chromatic variation. - Hue variation. - Saturation variation. - Lightness variation.
Cadik's study [19]
Image naturalness is defined as the degree of correspondence between a scene (seen directly) and the corresponding scenes in photos based on some criteria: brightness, contrast, colour reproduction, reproduction of details, simulation of glare, visual acuity and artifacts.
Naturalness indexes: - Brightness. - Contrast. - Colour reproduction. - Reproduction of details. - Reproduction of shadow details. - Simulation of glare. - Visual acuity. - Artifacts.
Choi's study [20]
Image naturalness is defined as the degree of correspondence between a photo displayed on a device and the memories about the real-life scene.
Naturalness features: - Memory colors of skin, grass and sky. - Sharpness. - Colorfulness. - Reproduction of shadow details.
Gu's study [21] and Yaacoub's study [22]
Images obtained from a camera (including pictures of man-made objects as well as forest / natural environments) are considered as natural images. (No definition for unnatural images)
Naturalness feature is calculated based on standard deviation and mean of pixel values and a statistic of natural images.
Jiang's study [23]
Image naturalness definition is based on exposure of images. Over or under exposure images are considered as unnatural images while normal exposure images are considered as natural.
Naturalness features are calculated based on luminance and yellow intensities.

on some criteria: brightness, contrast, colour reproduction, reproduction of details, simulation of glare, visual acuity and artifacts. An experiment is conducted to evaluate naturalness of SDR images generated by 14 different TMOs with a human naturalness assessment experiment with references. The real scene and the tone-mapped version of an HDR image of the same scene are shown to the observers. The observers have to give a subjective score (in range [0,. . . 10]) for the 5 criteria including brightness, contrast, visibility, reproduction of details and reproduction of colors. Based on the subjective scores, the TMOs are compared. In [14], a similar issue was discussed. The influences of the width and magnitude parameters of countershading in image quality enhancement and in provoking artifacts are studied. A subjective experiment was conducted in which observers were asked to adjust the magnitude parameter to the maximum level without artifacts under different settings of the width parameter. Based on the experimental results, some existing methods for image quality enhancement and tone mapping are improved to avoid noise and artifacts.

Besides, some naturalness features have been proposed for tone-mapped Images Quality Assessment (IQA) in few studies. In [21], [22], naturalness is mentioned as a factor to assess quality of images since it is considered as a feature in a feature set for IQA. In those researches, images obtained from a camera (including pictures of man-made objects as well as forest / natural environments) are considered as natural images. The naturalness in this work is computed based on statistics with 3,000 natural images. It is considered as the fitness of the standard deviation and the mean of pixel values to a Gaussian function and a Beta probability density function. In another study, Jiang et al.[23] define naturalness features based on the differences of normal exposure images and abnormal (over or under) exposure images. Simply, over or under exposure images are considered as unnatural images in that research. Naturalness features computed based on the luminance and yellow values. Those features are then used with details features and aesthetic features for tone-mapped image quality assessment. In those studies, it is concluded that naturalness plays a role in tone-mapped image quality assessment but naturalness is mentioned as a factor and there is no clear definition, conclusion or evaluation about the consistency of the naturalness. Table I presents an overview of naturalness features in previous studies.

In the state of the art about image naturalness, it is worthy to notice that none of those studies has the same definition of naturalness and their purposes are different from the purpose of this work. Moreover most of the naturalness features mentioned in previous researches are handcrafted features. But in the naturalness concept, we think that there is also an abstract part related to individual memories which cannot be precisely described to be inferred by handcrafted features. As a consequence, there is probably a need about naturalness features learning and the respective influences of handcrafted and learned features on image naturalness assessment is still an open question.

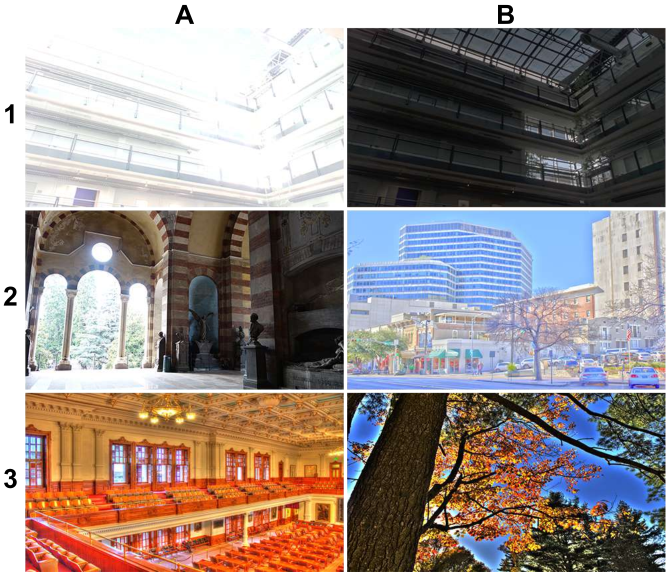

Fig. 3. Examples of artifacts. 1A: Over exposure, lost details. 1B: Under exposure, lost details. 2A: Too high contrast. 2B: Too low contrast, incorrect color reproduction. 3A: Bloom effect, incorrect color reproduction. 3B: Hallow effect, incorrect color reproduction.

III. Experiment of subjective image naturalness assessment

There are few research about image naturalness and one important challenge is that labelled natural / unnatural datasets are not available. Thus, there is a need of collecting such data so the first step before studying image naturalness is to organize an experiment of subjective image naturalness assessment without references. The description of the experiment includes the image sources, the experiment design, the experiment process, the observers, the experiment results and the naturalness dataset built from the data collected from the experiment.

A. Image sources

The dataset contains 2,727 SDR images coming from 3 main sources. The first one is 624 SDR images mapped from 208 HDR images coming from Debevec's [4], Fairchild's [24], Cadik's [25], Narwaria's [26], Yeganeh's [27], Korshunov's [28] and Krasula's [29] datasets. HDR images are not easy to collect and the number of images in each dataset is often small, so 7 HDR datasets are used in this research. Those HDR images are mapped to SDR images by using different TMOs including Reinhard's [9] (based on global contrast), Ashikhmin's [10] (using local contrast) and Khan's [8] (based on histogram and human visual system) algorithms. In order to focus on both naturalness and unnaturalness, there is a need of considering not only a well known TMO like Reinhard's TMO but also an TMO generating artifacts like Ashikhmin's TMO and an TMO generating both natural and unnatural images like Khan's TMO.

The second image source includes 1,811 SDR images of ESPL-LIVE dataset [30]. It includes 747 images tone-mapped from HDR images by using 4 TMOs [9], [13], [11], [12], 710 images grenerated directly from multi-exposure images by using 5 MEFs [30], [31], [32], [33] and 354 images gained after applying 2 post-processing algorithms [30].

The last part of the dataset contains 292 images including single-exposure, tone-mapped and post-processed images downloaded from Flickr website. The contents of the images are real world scenes including landscape, building, objects, people,. . . and they are taken under indoor, outdoor, day time, night time conditions.

B. Experiment setup

1) Experiment design: The experiment was conducted at GIPSA Lab, France where the experimental conditions are controlled according to the ITU BT-500 for a subjective experiment. Every observer performed the experiment by interacting with an interface displayed on a 24 inch (16:10) Samsung display (see Fig. 4). The resolution and color profile of the display has been set to 1920 × 1200 pixels and sRGB respectively. The peak brightness of the display is 250 $cd\ /\ m^2$. It is connected to a computer exporting a 32 bit color signal. The display and the computer are put in an experimental room where the light conditions are controlled thoroughly. The distance from observers to the display was fixed to 0.7 meter. Although the number of observers in the laboratory experiment is lower than that of some online crowd surveys, the thorough control of experimental conditions is the compensation ensuring the reliability of the experiment results.

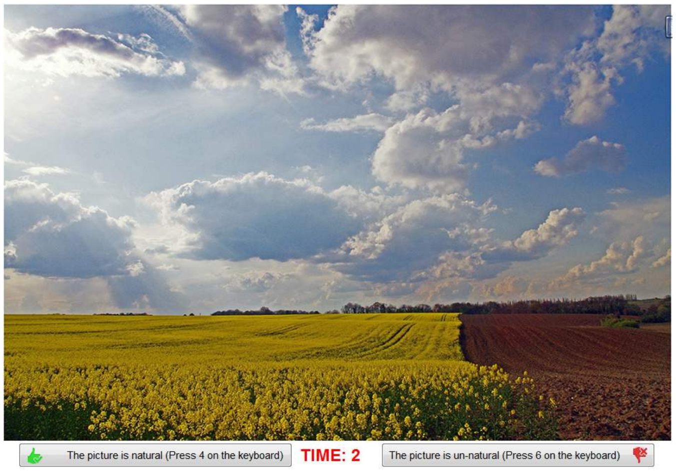
The picture is natural (Press 4 on the keyboard)
TIME: 2
The picture is un-natural (Press 6 on the keyboard)

Fig. 4. The interface for assessing image naturalness.

2) Experiment process: The process of the experiment for an observer is described in Fig. 5. Before starting the experiment, the observer performs an eye sight and a color sensation tests. The observer then reads the instructions, views some examples and performs a trial experiment to understand the experiment precisely and to be familiar with the interface of the experiment. The observer is instructed to focus on image naturalness rather than image quality or image aesthetic. In the trial phase, the observer has to evaluate the naturalness of five photos covering different causes of unnaturalness (they were pre-evaluated by the authors) and not belonging to the official experiment phase. Each turn, only one photo is showed to the observer during a short time. As explained previously, naturalness is an abstract concept not so easy to define precisely to each observer. So in our opinion, defining a scale of naturalness (from 1 to 10 from example) would not have been meaningful. As a consequence, the observers have been asked to quote each image in a binary way: natural or unnatural so that we could trust more the provided quotation. Therefore, there are only two choices: the photo looks natural or it looks unnatural to the observer (see Fig. 4). The observer can click a button on the interface or use the keyboard to enter his/her decision. Although the maximum time for evaluating an image is 7 seconds, the actual time in the experiment ranges from 3 to 5 seconds per photo. After giving the subjective evaluation, an uniform gray background is displayed for 1 second and the next image is then presented automatically to the observer. In the next step, the official experiment is performed in the same way as the trial experiment but the number of photos is higher. In the official phase, the number of assessed photos per observer is 380. The total performing time per subject ranges from 25 to 30 minutes.

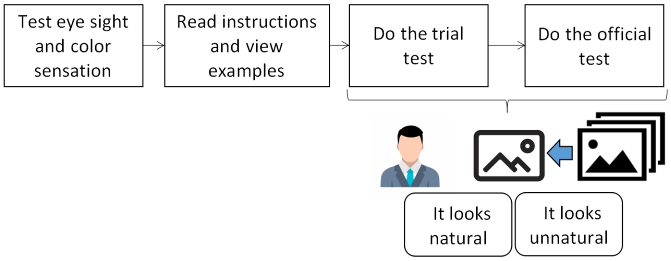

Fig. 5. The process of the experiment for an observer. There are 4 main steps including testing eyes, reading instructions, doing the trial test and doing the official test.

3) Observers: There were 45 people participating in the experiment which have quoted 1,900 images among the 2,727 available images. The number of men and women are 33 and 12 respectively. Among the 45 observers, 33 observers are familiar with image processing. The observers' ages range from 18 to 57. The average and the standard deviation of their ages are 26.2 and 7.53 respectively. The results show that 100 percent of them have normal or corrected to normal vision at that time.

C. Experiment results and the naturalness dataset

17,100 no reference subjective evaluations of photo naturalness were collected from 45 observers for 1,900 SDR images. Each SDR image has been assessed by the 9 observers. The distributions of the evaluations wrt each transformation method are presented in Table II. The distributions of the different groups are various. Some transformation methods receive a significant difference between the number of positive evaluations (assessing an image as natural) and the number of negative evaluations (assessing an image as unnatural) such as Durand's method (179 versus 1,063), Surreal effect (117 versus 1,062), Grunge effect (47 versus 835), Ashikhmin's TMO (230 versus 1,372). In contrast, the difference in the numbers of positive evaluations and the number of negative evaluations is in-significant for Pece's method (422 versus 377), Khan's method (732 versus 870). In other cases, there is a slight difference between the number of positive evaluations and the number of negative evaluations: Reinhard's method (1,549 against 725), Fattal's method (260 against 415), Larson's method (730 against 413), Flickr dataset (1,631 against 997).

TABLE II
THE DISTRIBUTION OF THE NATURALNESS EVALUATIONS WRT EACH TRANSFORMATION METHOD (OR IMAGE SOURCE) FOR THE WHOLE DATASET. NI: NUMBER OF IMAGES, PV: NUMBER OF POSITIVE VOTES (EVALUATING IMAGES AS NATURAL IMAGES), NV: NUMBER OF NEGATIVE VOTES (EVALUATING IMAGES AS UNNATURAL IMAGES).

Transformation method (or image source)	NI	PV	NV
Khan's TMO [8]	178	732	870
Ashikhmin's TMO [10]	178	230	1,372
Durand's TMO [12]	138	179	1,063
Fattal's TMO [11]	75	260	415
Reinhard's TMO [9]	253	1,549	725
Larson's TMO [13]	127	730	413
Paul's MEF [33]	97	582	291
Pece's MEF [32]	91	422	377
Raman's MEF [31]	133	945	252
Local Adjustment for MEF	50	57	393
Global Adjustment for MEF	59	409	122
Surreal effect (post processing)	131	117	1,062
Grunge effect (post processing)	98	47	835
Flickr dataset	292	1,631	997

The images are categorized into 10 groups based on the number of positive and negative evaluations they got. The results are showed in Fig. 6. A group is represented by a column in the chart. For example, the first left column in the chart corresponds to the 301 images that have been assessed as unnatural by the 9 observers (no one assessed them as natural) while the right last column shows that 143 images have been evaluated as natural by the 9 observers (no one evaluated them as unnatural).

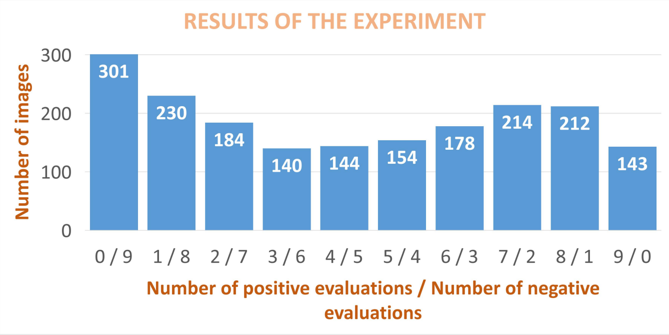

Fig. 6. Results of the subjective naturalness experiment.

Because the purpose of the research is to study naturalness and unnaturalness signs, there is a need of relevant data. Thus, only the images with a significant difference between the number of positive evaluations and the number of negative evaluations have been considered. Based on the results of the

TABLE III
THE DISTRIBUTION OF THE NATURALNESS EVALUATIONS FOR THE SELECTED IMAGES (IMAGES WITH AT LEAST 8 POSITIVE VOTES OR 8 NEGATIVE VOTES). NI: NUMBER OF IMAGES, PV: NUMBER OF POSITIVE VOTES (EVALUATING IMAGES AS NATURAL IMAGES), NV: NUMBER OF NEGATIVE VOTES (EVALUATING IMAGES AS UNNATURAL IMAGES).

Transformation method (or image source)	NI	PV	NV
Khan's TMO [8]	59	231	300
Ashikhmin's TMO [10]	123	73	1,034
Durand's TMO [12]	98	61	821
Fattal's TMO [11]	21	31	128
Reinhard's TMO [9]	93	696	141
Larson's TMO [13]	45	317	88
Paul's MEF [33]	33	257	40
Pece's MEF [32]	13	84	33
Raman's MEF [31]	67	560	43
Local Adjustment for MEF	36	18	306
Global Adjustment for MEF	31	265	14
Surreal effect (post processing)	101	28	881
Grunge effect (post processing)	87	18	765
Flickr dataset	79	544	167

experiment, an image in this study is considered as natural if there are at least 8 positive evaluations (in total 9 evaluations). Similarly, if there are at least 8 negative evaluations (in total 9 evaluations), it is considered as unnatural. The others are considered as uncertain images because related to controversial evaluations. In this experiment, a binary classification approach is chosen instead of a regression approach because for our first experiment with the notion of unnaturalness, we would like to focus on very contrasted cases of natural and unnatural images to be able to learn a bit more about unnaturalness. Thus, the question for the observers is a binary question "does the image look natural or unnatural?" (not a regression question). The regression problem will be considered later in the future.

After discarding the uncertain images, 531 unnatural images and 355 natural images are kept. The details of the evaluation distribution of the reduced version are described in Table III. Obviously, natural images and unnatural images have been generated by different transformation methods. Some methods generate mainly natural images (Reinhard's method) or unnatural images (Ashikhmin's method). And some methods generate both natural and unnatural images such as Khan's method with 231 positive votes versus 300 negative votes, and Pece's method with 84 positive votes against 33 negative votes.

In order to balance the dataset 176 unnatural images are removed randomly. Then, the ground-truth of the image naturalness dataset is built from 355 unnatural and 355 natural photos. After applying data augmentation including re-scaling, shifting, flipping, cropping and padding, 200 modified versions of size 224×244 are generated from every original photo (See examples in Fig. 7) and the labels of the augmented versions are set the same as the label of the original one. Totally, there are 142,000 images in the naturalness dataset in which half of them are natural and the others are labelled as unnatural. The dataset is available at http://www.gipsa-lab.fr/∼quyettien.le/projets_en.html.

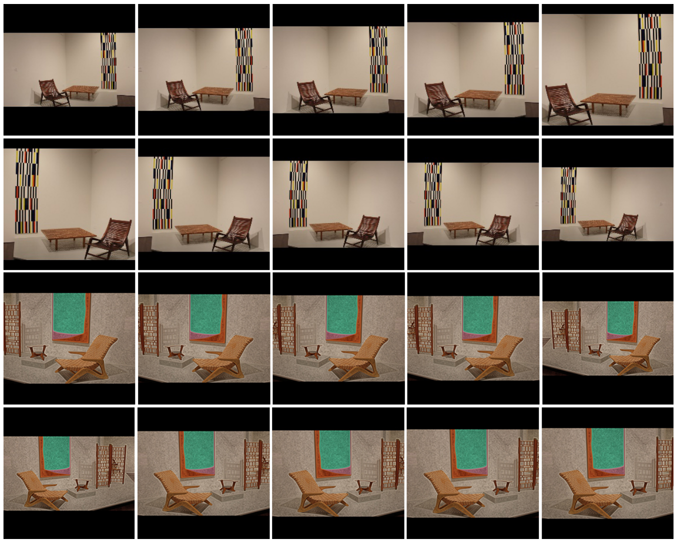

Fig. 7. Examples of data augmentation including re-scaling, shifting, flipping, cropping and padding (the black padding parts in those image are not presented to the observers). The two first rows present augmented versions of a natural image while the two last rows present augmented versions of an unnatural image (based on observers' evaluations). The data augmentation operations do not change the feeling of naturalness or unnaturalness so that the same label is kept.

IV. FEATURE DEFINITION AND FEATURE SELECTION

There is a lot of factors responsible for the unnaturalness of an image. Some of them can be described and defined by looking at the images while it is not easy to explain and modelize the others (see examples in Fig. 8). As a consequence, in this study, the considered features for the purpose of image naturalness assessment are built based on the one side on hancrafted features designed to take into account some a priori about unnaturalness and on the other side on features learned directly either from Convolutional Neural Networks (CNNs) or from pre-trained models (in order to access to non priori, indescribable information). The proposed handcrafted features are designed to focus on the popular artifacts induced by TMO, MEF and post-processing methods such as the feeling of perceived luminance, contrast, reproduction of detail and colors, bloom, halo and dark band effects [19]. In contrast, learned features are used to detect the abstract factors causing an unnaturalness feeling about photos.

A. *Handcrafted features*

Based on the ideas mentioned in [19] the considered handcrafted features are:

1) Brightness features: SDR images generated by TMO, MEF or post-processing algorithms sometimes look unnatural because of the perceived brightness. The brightness channel is one of the 3 channels of the HSV (or HSB: Hue, Saturation and Value or Brightness) color space. The brightness of a pixel is also calculated as the maximum value of the red, green, blue values. By analyzing the brightness histogram of the photos in the dataset, some artifact signs related to brightness could be detected. As an example, in Fig. 9, according to the results of the experiment, the top left image looks more natural to the

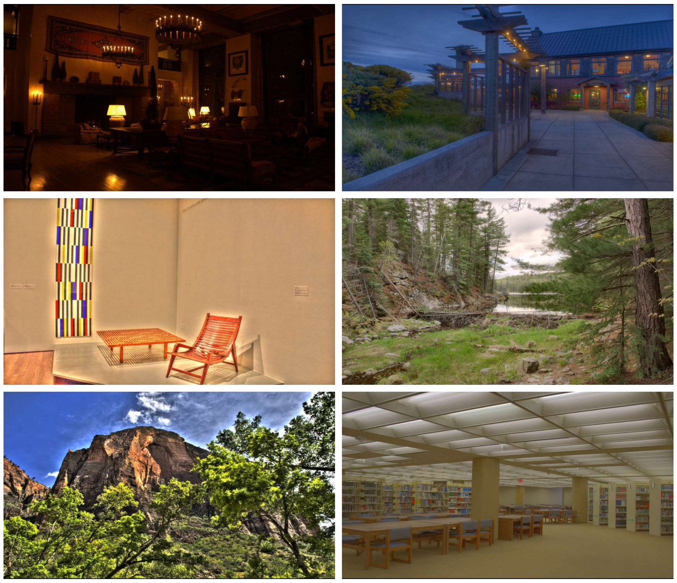

Fig. 8. Examples of unnatural images. In the left column, the images have clear artifact signs. The brightness in the first image is too low, there are halos surrounding the objects in the second image, the color saturation in the last one is too high. In contrast, when the observers look at the images of the right column, they have the feeling that the images are unnatural without being able to explain clearly why.

observers than those of the other color images. Looking at the brightness histogram in the last row, it appears that the density of medium brightness values in the natural image seems to be denser than those of the other images. In contrast, the two other images look too bright or too dark which can be detected on the brightness histograms that are distributed more in high or low values. The features representing the brightness histogram including mean (f_1), standard deviation (f_2), skewness (f_3), kurtosis (f_4) and continuity (f_5) of brightness are the first handcrafted features for image naturalness assessment. In which, the continuity of brightness is defined as:

$$f_5 = \sum |H_{br}(i) - H_{br}(i+1)| \quad (1)$$

where $H_{br}(i)$ and $H_{br}(i+1)$ are the values of the i^{th} and $i+1^{th}$ bins in the brightness histogram.

Another important factor affecting the image naturalness is the brightness contrast. Obviously, the global brightness contrast of an over-exposed image or an under-exposed image is often low. However, a photo with a too high global brightness contrast could also look unnatural. In this study, the features representing the global brightness contrast of an image are defined as:

$$f_6 = \frac{\mu_{br}^h}{\mu_{br}} \quad (2)$$

$$f_7 = \frac{\mu_{br}^l}{\mu_{br}} \quad (3)$$

$$f_8 = \frac{f_6 - f_7}{f_6 + f_7} \quad (4)$$

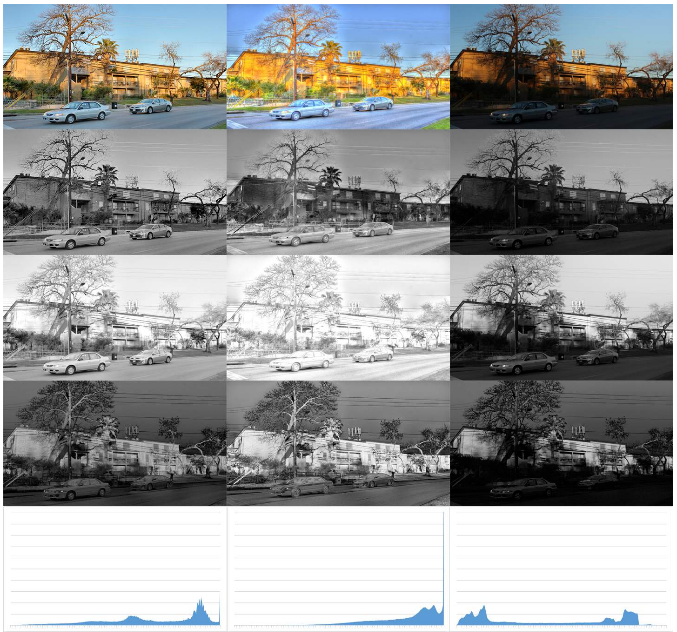

Fig. 9. The first row presents the color images. The second and the third rows illustrate the corresponding darkness and brightness channels (Eq. 6 and Eq. 7) of them. The fourth row shows the absolute difference (Eq. 8) between the darkness and the brightness channels. The brightness histograms of the color images are presented in the last row.

where f_6 and f_7 represent the highest and the lowest brightness values normalized by the brightness mean (μ_{br}). μ_{br}^h and μ_{br}^l are the means of the brightness values of top 5 percent pixels having the highest and the lowest brightness values respectively. The global brightness contrast (f_8) represents the relation between the highest and lowest brightness values in the photo. An unnatural photo often has a too high or a too low contrast or it also could have low highest brightness values (under-exposed images) or high lowest brightness values (over-exposed images).

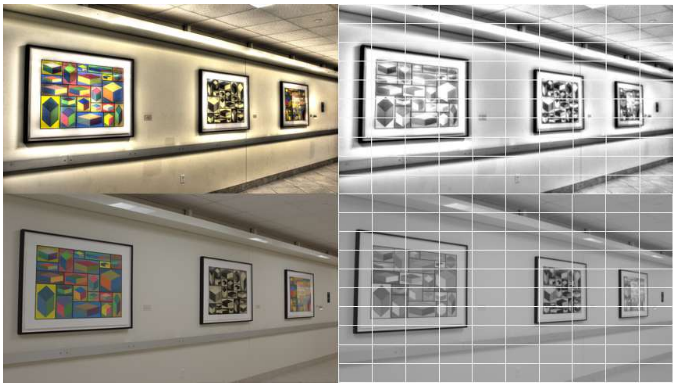

Fig. 10. The left column presents the color images. The right one illustrates the corresponding brightness channels of them. The first left image labelled as unnatural contains artifact signs: halo, dark band and bloom effects while the second color image is assessed as natural by the observers.

Additionally, images mapped by some TMOs have artifact signs such as dark bands, halos or blooms (see examples in Fig. 3 and Fig. 10). Halos and dark bands surrounding details increase the contrast of local parts as in Fig. 10 (halos have high brightness values while dark bands have low brightness values). In order to detect the high contrast of local parts caused by halo and dark band effects, an image is divided into M parts and the local brightness contrast of the image is defined as the mean of brightness contrasts of the M parts:

$$f_9 = \frac{1}{M} \times \sum_{i=1}^{M} \frac{max_{br}^i - min_{br}^i}{max_{br}^i + min_{br}^i} \tag{5}$$

where max_{br}^i and min_{br}^i are the maximum and the minimum brightness values respectively in the i^{th} part. In this study, M is set to 100 (10 × 10 as in the right column of Fig. 10).

2) Saturation features: The impression of colors is not only caused by brightness factors but is also affected by saturation factors. Thus saturation factors have significant influences on naturalness perception of images. Similarly to the brightness features, 9 features (f_{10} to f_{18}) are defined based on saturation information (extracted from the channels of the HSV color space) to present the saturation distribution and the saturation contrast of an image.

3) The darkness channel and its relation with the brightness channel: Analyzing the darkness channel and its relation with the brightness channel is an effective way to classify over-exposed, under-exposed and well exposed images. Considering an image in the RGB color space, the darkness channel (I_{da}) and the brightness channel (I_{br}) are defined based on the RGB channels as:

$$I_{da}(x,y) = \min\big(R(x,y), G(x,y), B(x,y)\big) \tag{6}$$

$$I_{br}(x,y) = \max\big(R(x,y), G(x,y), B(x,y)\big) \tag{7}$$

where (x,y) are the coordinates of a pixel. $R(x,y)$, $G(x,y)$ and $B(x,y)$ are red, green and blue levels at point (x,y) respectively. The difference between the two channels is defined as:

$$I_{di}(x,y) = I_{br}(x,y) - I_{da}(x,y) \tag{8}$$

In Fig. 9, it appears that the pixel values of darkness channel in the under-exposed image (the last column) are very low while those of the brightness channel in the over-exposed image (the second column) are too high. Beside this, the difference between the brightness and darkness channels of the over-exposed image is higher than that of the well exposed image. In contrast, this difference for the under-exposed image is less significant than that of the natural one. Therefore the information of the darkness channel and its relation with the brightness channel is an important clue to evaluate the naturalness of a photo. The 8 next features (f_{19} to f_{26}) for image naturalness assessment are the mean, standard deviation, kurtosis, skewness of I_{da} and I_{di} respectively.

Obviously, some details in the darkness and brightness channels of the images in the 2 last columns (Fig. 9) are lost. By comparing the details of the original image in gray scale and the details of the darkness and brightness channels, the reproduction of details and the balance between the darkness and the brightness channels can be evaluated. Thus the 2 last handcrafted features are defined as:

$$f_{27} = \frac{\sum |G_{I_g} - G_{I_{da}}|}{\sum G_{I_g}} \tag{9}$$

$$f_{28} = \frac{\sum |G_{I_g} - G_{I_{br}}|}{\sum G_{I_g}} \tag{10}$$

where $G_{I_g}, G_{I_{da}}, G_{I_{br}}$ are the gradient images [34] of the original image in gray scale, the darkness and the brightness channels respectively. $\sum G$ is the sum of pixel values of the image G. Note that the black padding regions (generated by the data augmentation methods) of images are discarded before calculating the handcrafted features.

To sum up, the overview of the considered handcrafted features is presented in Table IV. The features are presented in 3 groups including brightness features (9 features), saturation features (9 features) and darkness features (10 features).

B. Learned features

In some cases, it is possible to explain why an image looks unnatural to an observer but in general, it is a tough task. No direct relation appears between the unnatural feeling and the image clues such as color, brightness, saturation and so on. As a result, besides being handcrafted, features have also be learned directly from images by using CNNs [35]. Because of the modest image number of the naturalness dataset, the 2 approaches used for learning features in this study are shallow CNNs and transfer learning [36] using deep features learned from deep CNNs.

1) Shallow learned features: In the first approach, shallow learned features are learned from shallow CNNs (models with a low number of convolutional layers and a shallow architecture). The general structure of the 4 considered models (see Fig. 11) includes a convolutional layer receiving input color images of size 224×224, a global average pooling layer transforming 3D outputs from the convolutional layer into 1D outputs, a batch normalization layer normalizing the outputs from the global pooling layer and a fully connected layer on the top for predicting the input images as natural or unnatural. The size and the number of kernels in the convolutional layer are designed according to the number of samples in the dataset (142,000 samples of size 224×224). In order to learn various types of features, different models using different kernel sizes and different kernel numbers (490 kernels of size 5×5, 229 kernels of size 9×9, 65 kernels of size 17×17 and 65 kernels of size (2×17)×(2×17) - an average pooling layer is used to resize the input image by 50 percent) are designed as in Fig. 11. After the training phase, the models without the prediction layer are considered as feature extractors computing the learned features from the input images. The 4 feature extractors calculate 65, 65, 229 and 490 shallow learned features (features learned from shallow CNNs) for the purpose of image naturalness assessment. In the training process, the Adam optimizer and a binary cross-entropy loss function are used and the batch size is assigned to 128. The learning rate and the number of iterations are set to 10^{-6} and 3,000 respectively.

TABLE IV

OVERVIEW OF THE PROPOSED HANDCRAFTED FEATURES FOR IMAGE NATURALNESS ASSESSMENT.

Features	**Formula**
Brightness features	$f_1 = \frac{\sum_{i=1}^{N} I_{br}(i)}{N}$ $f_2 = \sqrt{\frac{\sum_{i=1}^{N} (I_{br}(i) - f_1)^2}{N-1}}$ $f_3 = \frac{\sum_{i=1}^{N} (I_{br}(i) - f_1)^3}{N \times f_2^3}$ $f_4 = \frac{\sum_{i=1}^{N} (I_{br}(i) - f_1)^4}{N \times f_2^4}$ $f_5 = \sum \lvert H_{br}(i) - H_{br}(i+1) \rvert$ $f_6 = \frac{\mu_{br}^{h}}{\mu_{br}}$ $f_7 = \frac{\mu_{br}^{l}}{\mu_{br}}$ $f_8 = \frac{f_6 - f_7}{f_6 + f_7}$ $f_9 = \frac{1}{M} \times \sum_{i=1}^{M} \frac{max_{br}^{i} - min_{br}^{i}}{max_{br}^{i} + min_{br}^{i}}$ I_{br} is the brightness channel.
Saturation features	$f_{10} = \frac{\sum_{i=1}^{N} I_{sa}(i)}{N}$ $f_{11} = \sqrt{\frac{\sum_{i=1}^{N} (I_{sa}(i) - f_{10})^2}{N-1}}$ $f_{12} = \frac{\sum_{i=1}^{N} (I_{sa}(i) - f_{10})^3}{N \times f_{11}^3}$ $f_{13} = \frac{\sum_{i=1}^{N} (I_{sa}(i) - f_{10})^4}{N \times f_{11}^4}$ $f_{14} = \sum \lvert H_{sa}(i) - H_{sa}(i+1) \rvert$ $f_{15} = \frac{\mu_{sa}^{h}}{\mu_{sa}}$ $f_{16} = \frac{\mu_{sa}^{l}}{\mu_{sa}}$ $f_{17} = \frac{f_{15} - f_{16}}{f_{15} + f_{16}}$ $f_{18} = \frac{1}{M} \times \sum_{i=1}^{M} \frac{max_{sa}^{i} - min_{sa}^{i}}{max_{sa}^{i} + min_{sa}^{i}}$ I_{sa} is the saturation channel.
Darkness features	$f_{19} = \frac{\sum_{i=1}^{N} I_{da}(i)}{N}$ $f_{20} = \sqrt{\frac{\sum_{i=1}^{N} (I_{da}(i) - f_{19})^2}{N-1}}$ $f_{21} = \frac{\sum_{i=1}^{N} (I_{da}(i) - f_{19})^3}{N \times f_{20}^3}$ $f_{22} = \frac{\sum_{i=1}^{N} (I_{da}(i) - f_{19})^4}{N \times f_{20}^4}$ $f_{23} = \frac{\sum_{i=1}^{N} I_{di}(i)}{N}$ $f_{24} = \sqrt{\frac{\sum_{i=1}^{N} (I_{di}(i) - f_{23})^2}{N-1}}$ $f_{25} = \frac{\sum_{i=1}^{N} (I_{di}(i) - f_{23})^3}{N \times f_{24}^3}$ $f_{26} = \frac{\sum_{i=1}^{N} (I_{di}(i) - f_{23})^4}{N \times f_{24}^4}$ $f_{27} = \frac{\sum \lvert G_{I_g} - G_{I_{da}} \rvert}{\sum G_{I_g}}$ $f_{28} = \frac{\sum \lvert G_{I_g} - G_{I_{br}} \rvert}{\sum G_{I_g}}$ I_{da} is the darkness channel. I_{di} is the differences between the brightness and darkness channels. $I_{di} = I_{br} - I_{da}$

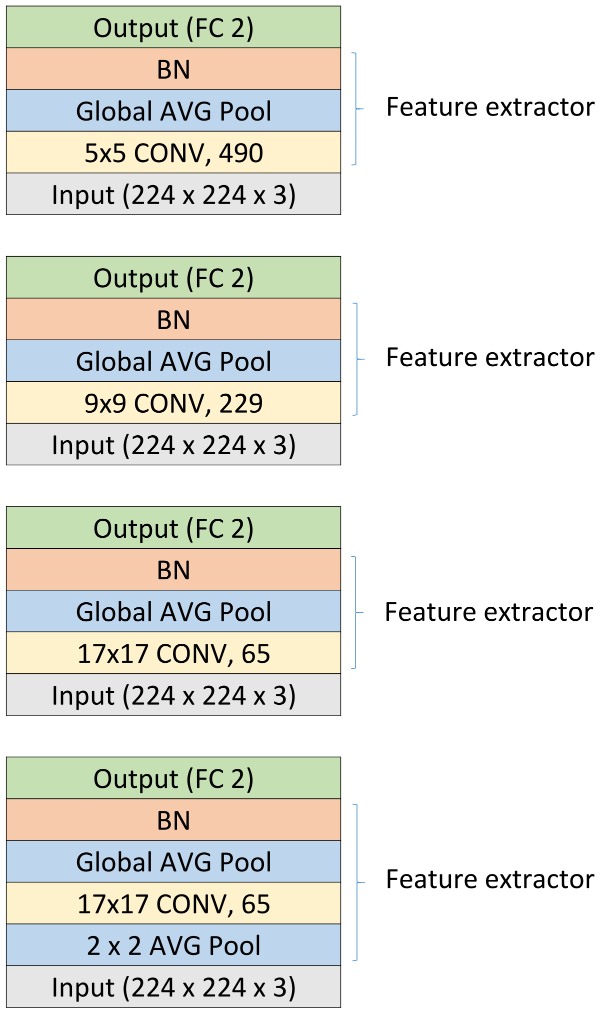

Fig. 11. Four different architectures of the shallow CNN. 2×2 AVG Pool: Average pooling layer with the pooling of size 2×2 that reduces the size of the input image by 50 percent. W×W CONV, N: N kernels of size W×W of the convolutional layer. Global AVG Pool: global average pooling layer. BN: Batch normalization layer. FC 2: The fully connected layer containing 2 output neurons (the prediction layer).

2) Deep learned features: Deep learned features are learned from deep CNNs (models with a high number of convolutional layers and a deep architecture). To learn deep features directly, there is a need of a very high number of images and it is impossible for the case of this study. Considering deep features learned by pre-trained models could be a good solution. Although the deep learned features (features learned from deep CNNs) have been learned for a given task, they can be considered to be used for different tasks [36]. The general structure of deep CNNs includes convolution layers at the bottom and fully connected layers on the top. The convolution layers are responsible for learning features while the fully connected layers are in charge of combining features learned from the convolution layers to solve the task. In other words, after removing the fully connected layers of a pre-trained deep neural network, the model can be considered as a feature extractor. In our study, several deep models including VGG16 [37], Xception [38], ResNet [39], NASNet large and NASNet mobile [40], MobileNet [41], Inception [42], DenseNet [43], Inception ResNet [44] pre-trained on the ImageNet dataset for the task of image classification are transferred to the new purpose of image naturalness assessment by keeping the convolution layers and replacing the top fully connected layers and training them for the new task. Instead of using all pre-learned features for the new task, there is a feature selection process to extract relevant features. The feature selection algorithm is presented in the next part.

3) Feature selection: When using transfer learning, features are primarily learned for a different task. Some features can be transfered well to perform a new task while the remaining are not relevant. Additionally, combining several feature sets could increase the performance but it also increases the number of features. This increase makes the computation complicated and sometimes it also increases the requirement of data. Therefore, simplifying a feature set by selecting the most relevant features is a good solution. Thus, there is a need of selecting and keeping the most relevant features and discarding the irrelevant features from a feature set. The relevance of each feature for the purpose of INA needs to be evaluated. This is done by using the relief method [45]. The training set S containing 113,600 images (56,800 natural images and 56,800 unnatural images) selected randomly from the dataset is considered as the evaluation set to compute the relevance of the features. All features of each image in S are calculated and normalized to range [0,... 1]. The relevance of a given feature f is calculated as:

$$r(f) = dif(f,N,U) - dif(f,N,N) - dif(f,U,U) \qquad (11)$$

$$dif(f,X,Y) = \frac{\sum_{i=1}^{\|X\|}\sum_{j=1}^{\|Y\|}(d(f,X_i,Y_j))}{\parallel X \parallel \times \parallel Y \parallel} \qquad (12)$$

where N and U are the natural images and the unnatural images of the evaluation set S respectively. The number of images in set X is presented as $\parallel X \parallel$. X_i is the i^{th} image of the set X while the absolute difference between f values of the 2 images x and y is represented as $d(f,x,y)$. The feature relevances are then normalized to range [0,... 1]. The highest $r(f)$ values illustrate the most relevant features.

In order to reduce the number of features and keep the most relevant features F_T, it is necessary to find a threshold T to be applied on feature relevance R to discard irrelevant features (features having the relevance smaller than the threshold T). To find the threshold, an algorithm based on the feature relevance and the binary search algorithm is applied [46]. The details of the algorithm are described in Fig. 12. The first step of the algorithm is to initialize a lower threshold T_1 and an upper threshold T_2 to 0 and 1 respectively. T_1 and T_2 are then considered as the thresholds to select 2 feature sets F_{T_1} and F_{T_2} ($F_{T_j} = \{f_x | r_x \geq T_j\}$). F_{T_1} and F_{T_2} are then applied to classify natural and unnatural images by using 2 SVM models. Comparing the 2 models trained on S_1 with 85,200 images selected from S and tested on S_2 containing 28,400 images coming from S can point out which threshold is the best (T_1 or T_2). The better threshold is kept while the worse threshold is updated to reduce the distance between the 2 thresholds. After performing K iterations, the final threshold T is computed as the average of the 2 thresholds T_1 and T_2. The algorithm is then applied on the feature sets to keep the most relevant features only. To evaluate the feature selection method, look at the improvement of the classification performance in Table VII and Table VIII in the experimental section. After selecting the most relevant features, the accuracy of the classification based on shallow learned features increases from 0.789 (with 849 features) to 0.808 (with 731 features) while the accuracy and the loss of the classification based on the deep learned features improves from 0.858, 0.299 (with 2,048 features) to 0.865, 0.139 (with 425 features) respectively. The feature reduction helps simplifying the feature sets and increasing the performance.

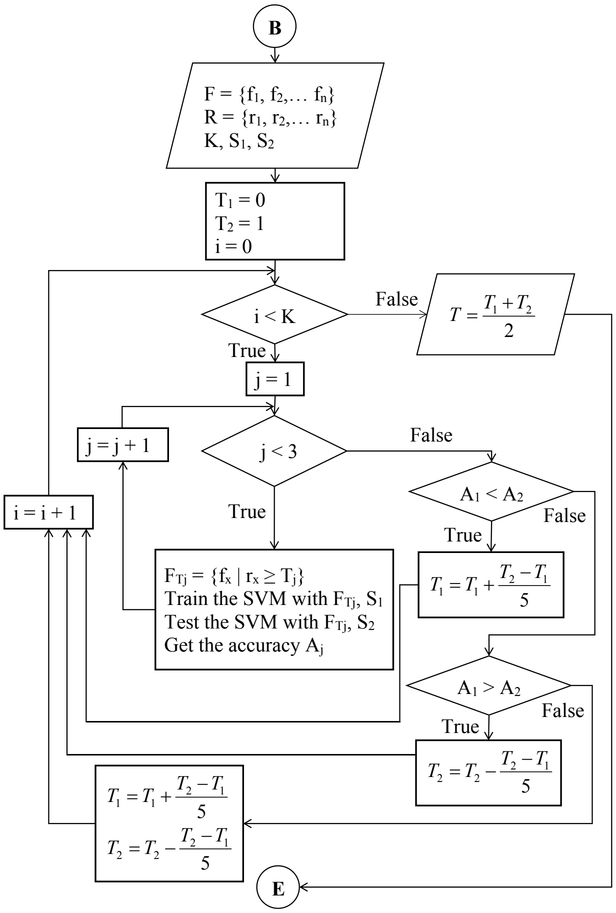
B
F = {f1, f2,... fn}
R = {r1, r2,... rn}
K, S1, S2
T1 = 0
T2 = 1
i = 0
i < K
False
$T = \frac{T_1 + T_2}{2}$
True
j = 1
j = j + 1
j < 3
False
A1 < A2
i = i + 1
True
False
True
F_Tj = {fx | rx ≥ Tj}
Train the SVM with F_Tj, S1
Test the SVM with F_Tj, S2
Get the accuracy Aj
$T_1 = T_1 + \frac{T_2 - T_1}{5}$
A1 > A2
False
True
$T_2 = T_2 - \frac{T_2 - T_1}{5}$
$T_1 = T_1 + \frac{T_2 - T_1}{5}$
$T_2 = T_2 - \frac{T_2 - T_1}{5}$
E

Fig. 12. Flowchart of the algorithm finding the optimal threshold. The input includes the feature set F, the feature relevance set R, the number of iterations K, the training set S_1 and the testing set S_2. T_1, T_2 are the lower and upper thresholds respectively. F_{Tj} is the reduced feature set selected with the threshold T_j. A_j is the accuracy of the SVM classifier trained and tested with S_1 and S_2 respectively with the feature set F_{Tj}. The output of the algorithm is the optimal threshold T.

V. Experiments and results

There are several purposes in this part. The first goal is to automatically answer the question “does an image look natural or not?”. The second one is to evaluate the efficiency of each feature set (handcrafted features, shallow learned features and deep learned features towards naturalness assessment). The last goal is to try to define more precisely the most relevant features describing unnaturalness.

Image naturalness assessment is considered here as a binary classification problem (natural / unnatural image) reflecting the fact that an observer might or might not feel that the image is natural. In order to evaluate the performance of each feature set, the classification is performed separately with the handcrafted features, the shallow learned features and the deep learned features.

A. Dataset and setup

The system having the general structure as in Fig. 13 is trained and tested to evaluate the classification performances of each feature set. The general structure includes an input layer, an output layer and P hidden blocks (in this study, P is set to 4). Each block contains a fully connected layer, a batch normalization layer and a dropout layer. The output layer contains 2 neurons corresponding to the 2 classes (natural and unnatural). The model is designed to learn how to combine the computed features for the classification task. Only the fully connected layers are trained in the training process so the convergence is fast. In the experiment, the number of iterations is set to 150. The Adam optimizer is used and the loss function is the binary cross-entropy loss. The learning rate and the mini-batch size are set to 10^{-3}, 512 respectively. Regarding the feature extraction block, each of the 3 feature sets is tested alone.

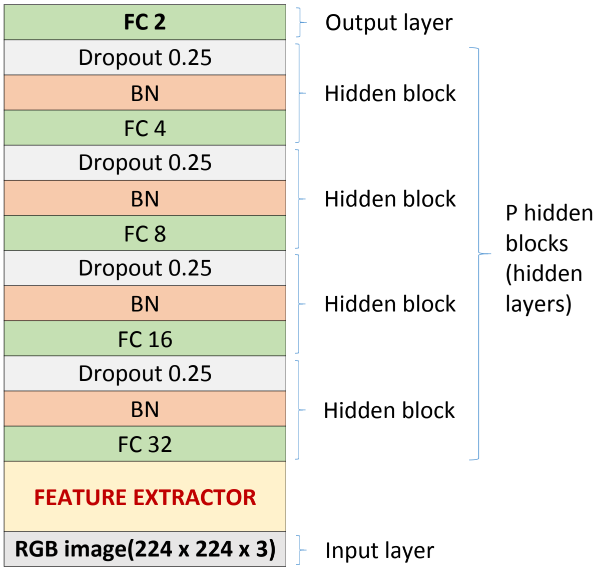

Fig. 13. General structure of the network designed for natural / unnatural image classification. Features extracted from an RGB input image of size 224×224×3 by the feature extractor are passed through the layers to classify the image as natural or unnatural. There are 4 hidden blocks with a fully connected layer, a batch normalization layer and a dropout layer in each block.

TABLE V
OVERVIEW OF EVALUATION CRITERIA.

Evaluation criteria	**Formula**
Accuracy	$A = \frac{TP+TN}{TP+FP+TN+FN}$
Lower accuracy	$A_l = A - I_a$
Upper accuracy	$A_u = A + I_a$
Loss	$L = \frac{\sum_{i=1}^{n} \lvert y_i - o_i \rvert}{n}$
Lower loss	$L_l = L - I_l$
Upper loss	$L_u = L + I_l$

The model is trained on the 113,600 images of the training set S and tested on a testing set S'_1 including 28,400 images (14,200 natural images and 14,200 unnatural images) generated from the 142 remaining original images by applying the data augmentations. There is no overlapping images (images generated from the same original images) between the training set and the testing set. Additionally, the classifier is also tested on a testing set S'_2 containing 142 images (71 natural images and 71 unnatural images) obtained from 142 original images by re-scaling and padding (just to convert images to the format of size 224×224 without cropping) to evaluate the influence of data augmentation on performances. It helps to demonstrate the validity of the data augmentation process regarding the labelling in particular. This is mainly because we focus on the naturalness of the image and not the naturalness of the image content. The model is evaluated based on the Accuracy (A) depending on TP, TN, FP, FN (true positive, true negative, false positive and false negative expressed as a number of images) and on the Loss (L) described in Table V. The lower accuracy (A_l) and the upper accuracy (A_u) reflect the range of accuracy within the 95% confidence interval [47], [48]. In general, the accuracy (or overall accuracy) is the most popular metric for evaluating classification performance while the loss (or mean absolute error) reflects the classification certainty. In Table. V, n is the class number (in this case $n = 2$), y and o are the target and the output (prediction) values respectively. The lower loss (L_l) and the upper loss (L_u) present the range of loss within the 95% confidence interval [47], [48]. I_a and I_l are the accuracy interval and the loss interval. They are calculated as:

$$I_a = z \times \sqrt{\frac{A \times (1 - A)}{N}} \tag{13}$$

$$I_l = z \times \sqrt{\frac{L \times (1 - L)}{N}} \tag{14}$$

where N is the number of testing samples (in this study, N is 28,400 for S'_1 and 142 for S'_2) and z is the number of standard deviations from the Gaussian distribution (z = 1.96 for 95% confidence interval).

The experiments have been performed on a PC equipped with an Intel(R) Xeon(R) CPU X5650 2.67 GHz (12 CPUs) and 24 GB memory. The feature computational time T_F (the time for computing features from images directly), the classification time T_C (the time for classifying images based on computed features), the classification accuracy and the loss with each feature set are estimated in two cases with and without a graphic card (NVIDIA GeForce GTX 1080 Ti) to compare the impacts of the feature set choice on the classification and to evaluate the balance between the computational cost and the classification accuracy.

B. Results and discussions

1) Handcrafted features based classification: In this first case, the feature extractor box in Fig. 13 computes the handcrafted features defined in section IV-A. Table VI shows the performances of the classification based on handcrafted features. The impact of each handcrafted feature subset is also estimated and is showed in the table. It appears that the overall accuracy of classification based on the separate feature subsets is quite low (0.712, 0.640, 0.716 for the brightness, saturation and darkness channels features respectively) and the

TABLE VI
IMAGE NATURALNESS ASSESSMENT BASED ON THE 28 HANDCRAFTED FEATURES AND IMPACT OF EACH HANDCRAFTED FEATURE GROUP ON THE ASSESSMENT.

Classification based on handcrafted feature subsets performed on the testing set S_1'			
9 Brightness features		A = 0.712	L = 0.389
9 Saturation features		A = 0.640	L = 0.495
10 Darkness channel features		A = 0.716	L = 0.401
Classification based on all the handcrafted features (28 features) performed on the testing set S_1'			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 12,680	FN = 1,520
	Unnatural	FP = 3,832	TN = 10,368
A = 0.812	I_a = 0.005	A_l = 0.807	A_u = 0.817
L = 0.321	I_l = 0.005	L_l = 0.316	L_u = 0.326
Classification based on all the handcrafted features (28 features) performed on the testing set S_2'			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 66	FN = 5
	Unnatural	FP = 19	TN = 52
A = 0.831	I_a = 0.062	A_l = 0.769	A_u = 0.893
L = 0.326	I_l = 0.077	L_l = 0.249	L_u = 0.403
Processing time without the graphic card			
T_F	63 ms	T_C	2 ms
Processing time with the graphic card			
T_F	59 ms	T_C	2 ms

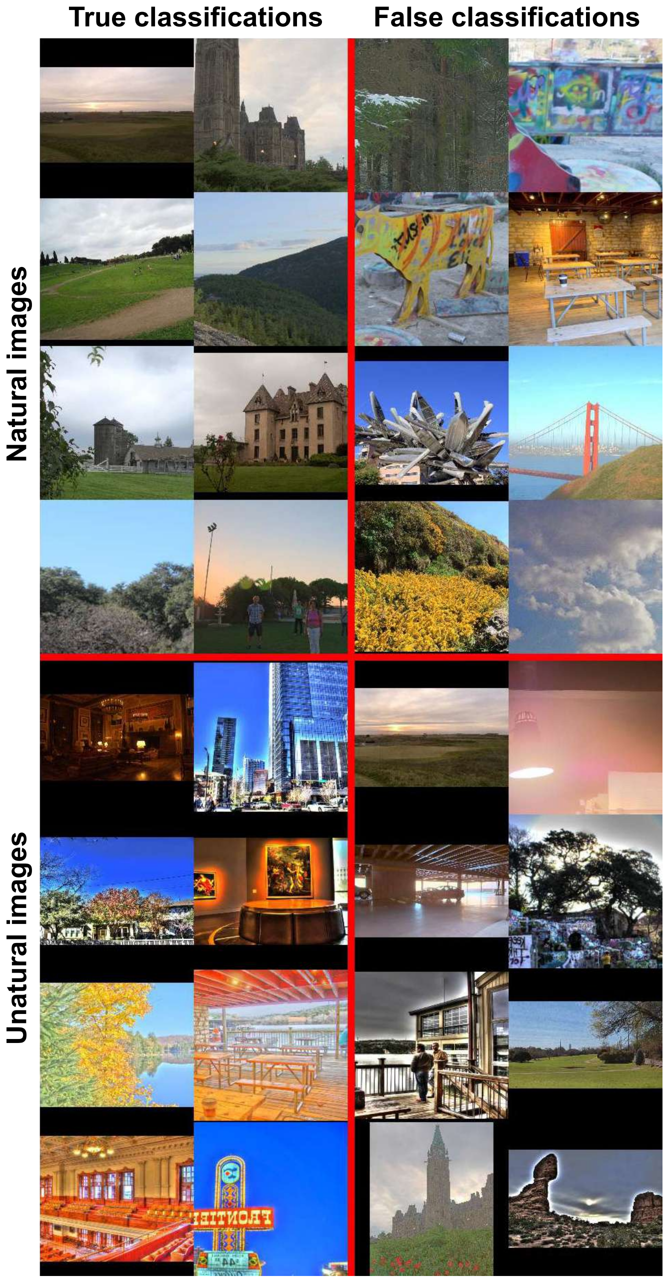

Fig. 14. Classification examples with handcrafted features. The four first rows and the four last rows show natural and unnatural images respectively. The two left columns contain well classified images associated to a very low loss value while the two right columns contain misclassified images associated to a very high loss value.

loss is high (0.389, 0.495, 0.401 for the brightness, saturation and darkness channels features respectively). By combining them, the overall accuracy increases to 0.812 and the loss decreases to 0.321. Beside this, it appears that the FP value is much higher than the FN value (3,832 versus 1,520), the handcrafted features appear to be more sensitive to unnatural images in this case.

Additionally, the feature computation of the handcrafted features is performed quite fast and the difference between computations with and without the graphic card is insignificant (63 ms, without the graphic card and 59 ms, with the graphic card).

Classification examples based on handcrafted features are shown in Fig. 14 (because of the size reduction and the PDF transformation, the perception of contrast and therefore the impression of image naturalness might have been changed a little in the figure). As expected, the unnatural images caused by low saturation are well classified with those features. Although halos around details and contrast factors have been considered during the feature design stage, there are some images with those artifacts in the misclassified unnatural images. Additionally, the classifier based on the handcrafted features appears to be too sensitive to colorfulness since most of the well classified unnatural images and the misclassified natural images are colorful while the well classified natural images and the misclassified unnatural images are less colorful. It seems that the handcrafted features are not able to detect all the cases and sometimes they are too sensitive to some factors. So some discriminant features are not taken into account with the considered handcrafted features.

2) Shallow learned features based classification: In Fig. 13, the feature extractors are now made of the shallow CNNs described in section IV-B1. Beside the classifications based on separate feature sets, the classification with the combination of all the shallow learned features is also performed. The details of classification using features learned from the 4 shallow CNNs are shown in Table VII. Obviously, the classification based on the combination of shallow learned features has the best overall accuracy (0.786) and the best loss (0.269) but the number of features is also the highest (849 features) among the shallow learned feature sets (65, 65, 229 and 490 features). In addition to this, the computational time for the shallow learned features without the graphic card is high (114 ms per photo) but it decreases a lot (to 64 ms per photo) when using the graphic card.

TABLE VII

IMAGE NATURALNESS ASSESSMENT BASED ON THE SHALLOW LEARNED FEATURES AND IMPACT OF EACH SHALLOW LEARNED FEATURE GROUP ON THE ASSESSMENT.

Classification based on the separate feature subsets learned from the shallow convolutional networks performed on the testing set S_1'			
Features learned from the model with 490 5×5 kernels		A = 0.756	L = 0.337
Features learned from the model with 299 9×9 kernels		A = 0.766	L = 0.305
Features learned from the model with 65 17×17 kernels		A = 0.753	L = 0.329
Features learned from the model with 65 17×17 kernels and an average pooling layer		A = 0.741	L = 0.332
Classification based on all the shallow learned features (849 features) performed on the testing set S_1'			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 11,504	FN = 2,669
	Unnatural	FP = 3,376	TN = 10,824
A = 0.786	I_a = 0.005	A_l = 0.781	A_u = 0.791
L = 0.269	I_l = 0.005	L_l = 0.264	L_u = 0.274
Classification based on the reduced shallow learned feature set (731 features) performed on the testing set S_1'			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 11,131	FN = 3,069
	Unnatural	FP = 2,390	TN = 11,810
A = 0.808	I_a = 0.005	A_l = 0.803	A_u = 0.813
L = 0.274	I_l = 0.005	L_l = 0.269	L_u = 0.279
Classification based on the reduced shallow learned feature set (731 features) performed on the testing set S_2'			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 65	FN = 6
	Unnatural	FP = 19	TN = 52
A = 0.824	I_a = 0.063	A_l = 0.761	A_u = 0.887
L = 0.277	I_l = 0.074	L_l = 0.203	L_u = 0.351
Processing time without the graphic card			
T_F	114 ms	T_C	2 ms
Processing time with the graphic card			
T_F	64 ms	T_C	2 ms

In order to study the compromise between the number of features and the accuracy, the feature reduction algorithm based on the Relief method (see Fig. 12) is applied on the combined feature set to reduce the feature number from 849 to 731. Although the number of features decreases, the feature computational time does not change because the two feature sets are computed by the same CNNs. By keeping the most relevant features only for the classification, the overall classification accuracy increases from 0.786 to 0.808 but the loss increases slightly from 0.269 to 0.274.

Fig. 15 shows classification examples based on the combination of the shallow learned feature sets. Focusing on the true classification samples of unnatural images, it appears that the filters in the shallow models are efficient to detect unnatural images caused by halos around details. Contrary to the classification based on handcrafted features, the shallow learned features based classifier is not efficient to detect color saturation artifacts since the color saturation of 5 (of the 8) misclassifed unnatural images is low. With the handcrafted features, the classifier focuses on the characteristics of the whole image while the shallow learned features based classifier focuses on each sub region of the image (the size of sub regions depends on the size of kernels). It explains the differences between the classification results based on the two feature sets.

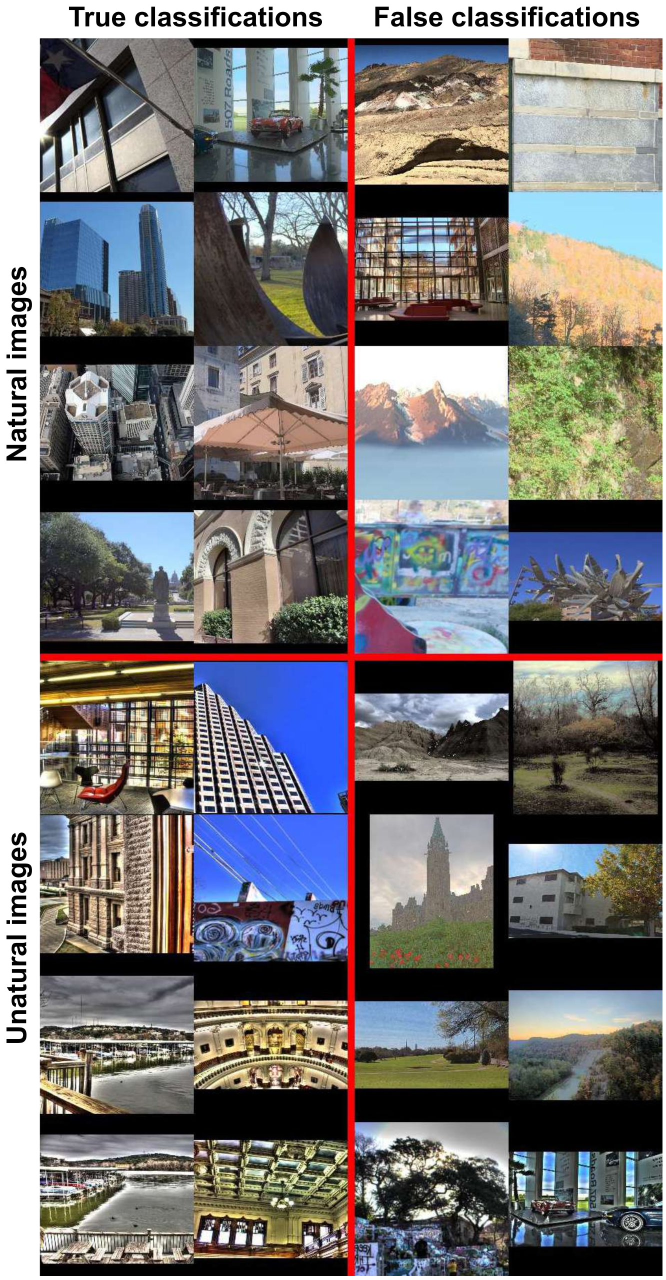

Fig. 15. Classification examples with shallow learned features. The four first rows and the four last rows show natural and unnatural images respectively. The two left columns contain well classified images associated with a low loss value while the two right columns contain misclassified images associated with a high loss value.

3) Deep learned features based classification: The feature extractor is successively made of the nine pre-trained deep models described in section IV-B2 followed by the feature selection process described in section IV-B3. After training and testing the models using the 9 reduced feature sets, the highest overall accuracy (0.865) is obtained with the model using the features learned from the ResNet extractor. The ResNet model was pre-trained on ImageNet dataset using an SGD optimizer, a batch size of 256, a momentum of 0.9. The model was trained for 60 × 10^4 iterations with the learning

TABLE VIII
IMAGE NATURALNESS ASSESSMENT BASED ON THE FEATURES LEARNED FROM THE RESNET MODEL PRE-TRAINED ON THE IMAGENET DATASET.

Classification based on the ResNet feature set (2,048 features) performed on the testing set S_1'			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 12,539	FN = 1,661
	Unnatural	FP = 2,359	TN = 11,841
A = 0.858	I_a = 0.004	A_l = 0.854	A_u = 0.862
L = 0.299	I_l = 0.005	L_l = 0.294	L_u = 0.304
Classification based on the reduced ResNet feature set (425 features) performed on the testing set S_1'			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 12,709	FN = 1,491
	Unnatural	FP = 2,336	TN = 11,864
A = 0.865	I_a = 0.004	A_l = 0.861	A_u = 0.869
L = 0.139	I_l = 0.004	L_l = 0.135	L_u = 0.143
Classification based on the reduced ResNet feature set (425 features) performed on the testing set S_2'			
		Prediction	
		Natural	Unnatural
Ground truth	Natural	TP = 66	FN = 5
	Unnatural	FP = 11	TN = 60
A = 0.887	I_a = 0.052	A_l = 0.835	A_u = 0.939
L = 0.132	I_l = 0.056	L_l = 0.076	L_u = 0.188
Processing time without the graphic card			
T_F	93 ms	T_C	2 ms
Processing time with the graphic card			
T_F	47 ms	T_C	2 ms

rate starting at 0.1 and divided by 10 when the error reaches a plateaus [39]. In this case, there is no re-trained ResNet layers. The model without the last layer (the fully connected layer) is considered as the feature extractor for the proposed model as in Fig. 13. Specifically, 425 learned features are extracted from the 2,048 ResNet features by applying the Relief based feature reduction algorithm. The details of the best classification are showed in Table VIII. The overall accuracy and the loss of the classification are quite good at 0.865 and 0.139 respectively. The number of features used in the classification process is high (425 features) but the computation of the features with the graphic card is quite fast, 47 ms (the computational time without the graphic card is 93 ms).

Classification examples based on the ResNet features are presented in Fig. 16. It appears that some of the well classified unnatural images have halos around details. Secondly, brightness factors are not detected well since there are some misclassified unnatural images having a too low brightness. Beside this, it is similar to the handcrafted features based classification since most of the well classified unnatural images are colorful. There are some overlapping images (4 of 8) between the misclassified natural images based on the shallow learned features and the ones based on the deep learned features. It demonstrates that some similar characteristics are learned from the training samples by both deep and shallow CNNs.

4) Discussions: The general purpose of the study is the naturalness of images (not naturalness of scenes). There are several images of the same scene generated from the same original image but in various ways and they might look totally different (See the first row of Fig. 9 where 3 images of the same scene are generated in 3 different ways). Table IX reflects that the classifications performed on S_1' and S_2' are similar since the differences in classification accuracy and

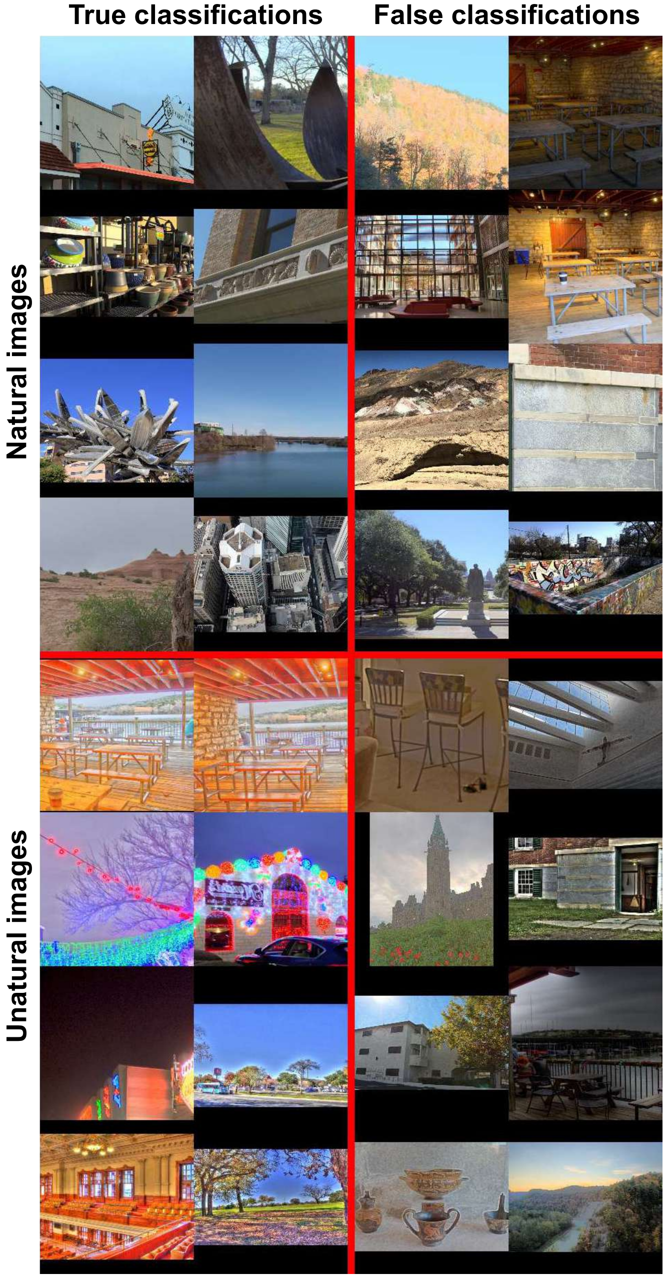

Fig. 16. Classification examples with deep learned features. The four first rows and the four last rows show natural and unnatural images respectively. The two left columns contain well classified images associated with a low loss value while the two right columns contain misclassified images associated with a high loss value.

TABLE IX
IMAGE NATURALNESS ASSESSMENT BASED ON THE 3 FEATURE SETS PERFORMED ON THE TESTING SETS S_1' AND S_2'.

Feature set	$A \pm I_a$ (testing on S_1')	$L \pm I_l$ (testing on S_1')	$A \pm I_a$ (testing on S_2')	$L \pm I_l$ (testing on S_2')
Handcrafted features	0.812 ± 0.005	0.321 ± 0.005	0.831 ± 0.062	0.326 ± 0.077
Shallow learned features	0.808 ± 0.005	0.274 ± 0.005	0.824 ± 0.063	0.277 ± 0.074
Deep learned features	0.865 ± 0.004	0.139 ± 0.004	0.887 ± 0.052	0.132 ± 0.056

classification loss are insignificant. This means that the image naturalness assessment is not affected by data augmentation. That is why we kept the same natural or unnatural label after data augmentation. According to (13) and (14), the intervals of accuracy and loss depend on the number of testing samples. Indeed, the accuracy and loss intervals of the classification performed on S'_1 (from 0.004 to 0.005) is much smaller than those of the classification executed on S'_2 (from 0.052 to 0.077) because the number of samples in S'_1 is much bigger than that of S'_2 (28,400 versus 142). The highest confidence of the classification results is obtained with the biggest testing set S'_1.

In Fig. 17, it is seen that the classifications based on different feature sets act differently since some images are classified well with a feature set but they are misclassified with the others. Table X presents a general comparison between the classifications based on the 3 feature and sets. By using deep learned features, the classification accuracy and the classification loss reach the best values (Accuracy: 0.865 compared to 0.812, 0.808 and Loss: 0.139 versus 0.321, 0.274 for the handcrafted, shallow learned features based classifications respectively). Additionally, the classification accuracy with the handcrafted features and the one with the shallow learned features are almost equal. Looking at Fig. 18, it appears that the evolution of the loss function is quite different between the feature sets. With handcrafted features and shallow learned features, the loss values constantly increase from low values to high values (even not reaching 0 or 1 in the case of handcrafted features). In contrast, with deep learned features, the loss values in true classifications are nearly zero and they are nearly one in false classifications. This makes the decision more reliable. Moreover, the average loss of the classification based on deep learned features is much smaller than that of the others. Additionally, it appears that the handcrafted features (when well designed) are quite efficient since an overall accuracy of 0.812 is obtained with only 28 features. The classification accuracy with handcrafted features is similar to the accuracy with shallow learned features (accuracy: 0.812 versus 0.808) but the classification certainty for shallow learned features is better (loss: 0.274 against 0.321). Although the number of handcrafted features is smaller than that of the learned features (28 versus 731 and 425), the feature computational time of the handcrafted features and the learned features with the graphic card are almost the same (59 ms, 64 ms and 47 ms for the handcrafted features, the shallow learned features and the deep learned features respectively). Deep learned features are quite efficient since the classification accuracy and the loss are better than those of handcrafted features and shallow learned features (accuracy: 0.865, loss: 0.139). It is seen that the problem of naturalness is abstract and too complicated for shallow CNN architectures to learn features reflecting this problem. Using simple and shallow CNN architectures could not be a good choice for this problem (the classification performance with shallow learned features is even lower than that with handcrafted features).

The last discussion of this part is "did the models learn to recognize the signatures of the transformation methods or did they learned to access to naturalness / unnaturalness?"

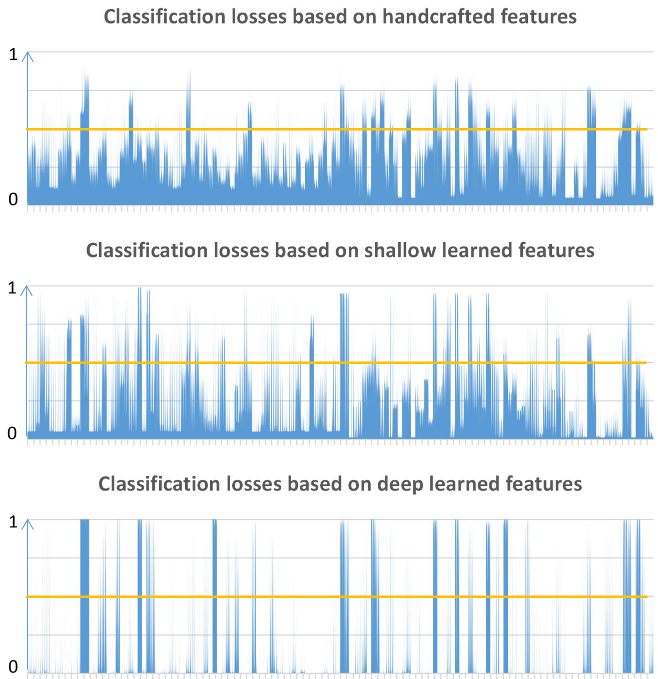

Fig. 17. Classification losses based on the 3 feature sets. Y axis represents the loss values while X axis represents the images. Each horizontal line is the border between true classifications (loss $<$ 0.5) and false classifications (loss $>$ 0.5).

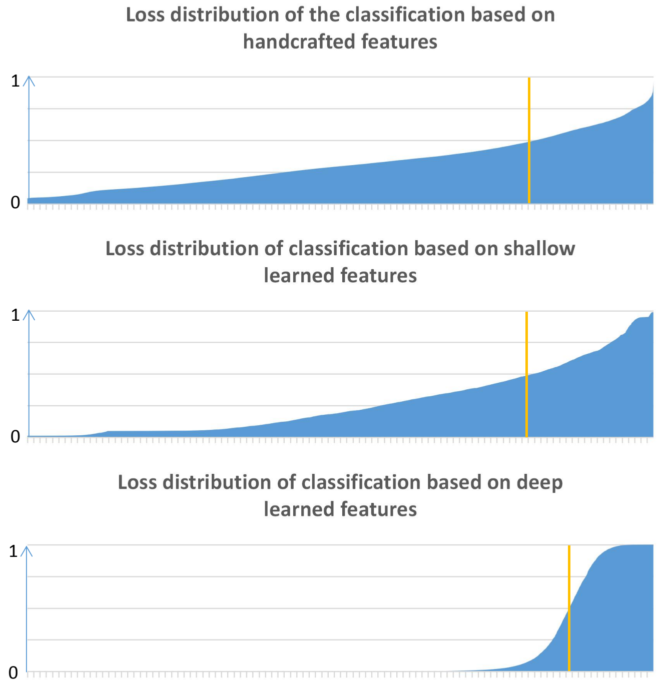

Fig. 18. Loss distribution of classification based on the 3 feature sets. Y axis represents the loss values while X axis represents the images (sorted based on loss values). Each vertical line is the border between true classifications (loss $<$ 0.5) and false classifications (loss $>$ 0.5).

TABLE X
IMAGE NATURALNESS ASSESSMENT BASED ON THE HANDCRAFTED, THE SHALLOW, THE DEEP LEARNED FEATURES. (N_F AND T_F ARE THE NUMBER OF FEATURES AND THE FEATURE COMPUTATIONAL TIME RESPECTIVELY)

Feature set	N_F	T_F (ms)	$A \pm I_a$	$L \pm I_l$
Handcrafted features	28	59	0.812 ±0.005	0.321 ±0.005
Shallow learned features	731	64	0.808 ±0.005	0.274 ±0.005
Deep learned features	425	47	0.865 ±0.004	0.139 ±0.004

TABLE XI
CROSS VALIDATION OF THE MODEL USING THE REDUCED RESNET FEATURE SET (425 FEATURES). EACH GROUP OF IMAGES IS CONSIDERED AS THE TESTING SET WHILE THE REMAINING GROUPS ARE CONSIDERED AS THE TRAINING SET.

Transformation method (or image source) of the testing set	$A \pm I_a$	$L \pm I_l$
Khan's TMO [8]	0.756±0.008	0.282±0.008
Ashikhmin's TMO [10]	0.795±0.005	0.214±0.005
Durand's TMO [12]	0.833±0.005	0.187±0.005
Fattal's TMO [11]	0.679±0.014	0.340±0.014
Reinhard's TMO [9]	0.694±0.007	0.306±0.007
Larson's TMO [13]	0.906±0.006	0.119±0.007
Paul's MEF [33]	0.755±0.010	0.253±0.010
Pece's MEF [32]	0.906±0.011	0.169±0.014
Raman's MEF [31]	0.879±0.006	0.164±0.006
Local Adjustment for MEF	0.583±0.012	0.489±0.012
Global Adjustment for MEF	0.951±0.005	0.061±0.006
Surreal effect (post processing)	0.718±0.006	0.285±0.006
Grunge effect (post processing)	0.959±0.003	0.085±0.004
Flickr dataset	0.812±0.006	0.340±0.007
Total	0.801±0.002	0.221±0.002

because unnaturalness signs come from transformation methods. Looking at Table III, it appears that different sources of images have been considered. And negative and positive evaluations are not coming all from the same source of images. Additionally, there are few images per group, so it is unlikely that the signatures of the transformation methods are learned in this case. In order to verify this assumption, an additional experiment has been performed. The images are categorized in 14 groups as in Table III. The experiment is performed 14 times, each time the augmented images from only one group are considered as the testing set while the classifier using the reduced ResNet feature set (425 features) is trained with the augmented images from the remaining ones, so the signatures of the transformation method in the testing set can not be learned. The initialization of the training process is similar to that of the previous one with the reduced ResNet feature set. The results are showed in Table XI. Although the accuracy and the loss change a little bit compared to the accuracy and the loss of the model trained in the previous way (from 0.865 to 0.801 for accuracy and from 0.139 to 0.221 for loss), the accuracy and the loss values are quite good at 0.801, 0.221 respectively. The differences between the 2 tests are insignificant so it can be concluded that the extracted features are for image naturalness assessment and the solved problem here is not the classification of transformation methods.

5) Towards unnatural image understanding: Although this part sounds a little speculative, the authors want to share with the readers experiences obtained after working with the human subjective image naturalness experiments and analyzing the obtained results. This might be helpful for other studies.

It seems that the feeling of unnaturalness comes from 2 main causes: visible unnaturalness clues and viewers' experience.

The first unnaturalness clue is color. It includes brightness, color saturation and hue. In general, it is impossible for a camera to cover the whole range of brightness of real scenes. By applying algorithms (TMOs, MEFs), the brightness range of a real scene is compressed and it leads to the fact that the brightness distribution, the brightness range and the brightness contrast of photos and those of real scenes are different. For instance, in the first row of Fig. 3, the left photo is too bright (over exposure) while the right one is too dark (under exposure). Additionally, the left image of the second row has a too high brightness contrast since some regions are too bright while some regions are too dark. Beside being compressed, the brightness also could be affected by using post processing algorithms. For example, when a photo has been taken under dark conditions, if the photographer wants to make it brighter, he / she might post-process the photo to increase the brightness. Generally, if the difference is insignificant, it might not be detected by viewers' eyes but if the difference is important, it becomes an artifact sign. Besides this, choosing parameters for transforming algorithms is a very important task. If parameters of a method are chosen correctly for a given photo, the photo quality can increase significantly but a wrong chosen parameter can make the photo horrible [14] and of course unnatural. TMOs, MEFs and post-processing re-produce brightness, color saturation and hue of images. An abnormal color saturation (too high or too low) could be detected by human eyes. Unusual hues in photos make photos unnatural to viewers. For instance, it is impossible to have orange sky as in Fig. 2 or dark blue sky as in Fig. 3.

Beside color, the second visible unnaturalness clue is the reproduction of details. In order to reproduce lost details, to enhance sharpness or to reduce noise in photos, some post-processing algorithms modify photo details. Those changes could lead to artifact signs such as blurriness, graininess, halo, dark band effects. Additionally, when combining multiple shoots taken under different exposures, MEFs try to preserve details coming from different images. Sometimes, the details are not combined well and some artifact clues such as motion blur, ghost effects are produced. TMOs generate unnatural details in a different way. When compressing color range, some TMOs try to preserve local contrast and global contrast in photos. The reduction of the dynamic range might produce artifact details (too sharp, contrast details, halo bands) or some details could be lost after mapping.

When the unnaturalness feeling comes from viewers' experience, unnaturalness clues are not obvious. As a matter of fact, the observers compare scenes in photos to scenes retrieved from their memory (what they have seen) [49] to find differences and similarities, so assessment results depend on individual factors [50]. For example, some viewers think dark photos and bright photos are unnatural because they are not familiar with those scenes while some people disagree because they have seen similar scenes in few cloudy days or few sunny days (see examples in Fig. 19). Another example is about the tree colors. Green colors of trees are not the same and they vary under different light conditions. However, some viewers fix a range of green colors for plants in their mind. Except those colors, they consider that other green colors are unnatural. In this case, when evaluating the naturalness of scenes with trees in photos, viewers often focus on 3 questions "What trees are they?", "What are their colors?" and "Do they and their colors match?". As a result, it is not easy to design handcrafted features in this case because naturalness appears to be an individual feeling while deep learning helps us to learn others features that are not always explained by building on a combination of different handcrafted features.

How to explain that during our naturalness assessment experiment, people all agree on the evaluation of some images and completely disagree about some other images? Looking at Fig. 6, the images are categorized in 10 image groups based on the corresponding evaluations. In order to analyze the confidence of naturalness labels of those images, the 10 groups are merged into 5 categories in which each category is presented by a pair of values (X, Y). X is the number of observers having the same opinion about the naturalness of an image while Y is the number of observers having the opposite opinion. According this definition, there are 5 categories: (9,0), (8,1), (7,2), (6,3) and (5,4). The naturalness label of each image is decided by the majority of the observers so it appears that the confidence of the labels in category (9,0) is the highest and the confidence in category (5,4) is the lowest. Analyzing each category, it appears that images with the highest confidence labels generally present obvious visible artifact. Thus, the answer here is that the unnaturalness clue is clearer to viewers and it is easier for them to make the decision in the cases with clear unnaturalness signs. On the contrary, images with the lowest confidence labels are images on which the naturalness / unnaturalness feeling is more related to the viewer's experience. The feeling of unnaturalness based on viewers' experience sometimes is not the same and the image naturalness in those cases could be controversial.

Overall, the naturalness concept can definitely be defined based on 2 terms. The first one is memory color that reflects the typical color of an object that a beholder acquires through viewers' experience with that object [49]. The second term is obvious artifacts that can be recognized by eyes such as very high or very low contrast, too sharp details, loss details, artifact details (see Fig. 3). That is why the problem of naturalness assessment is so tricky.

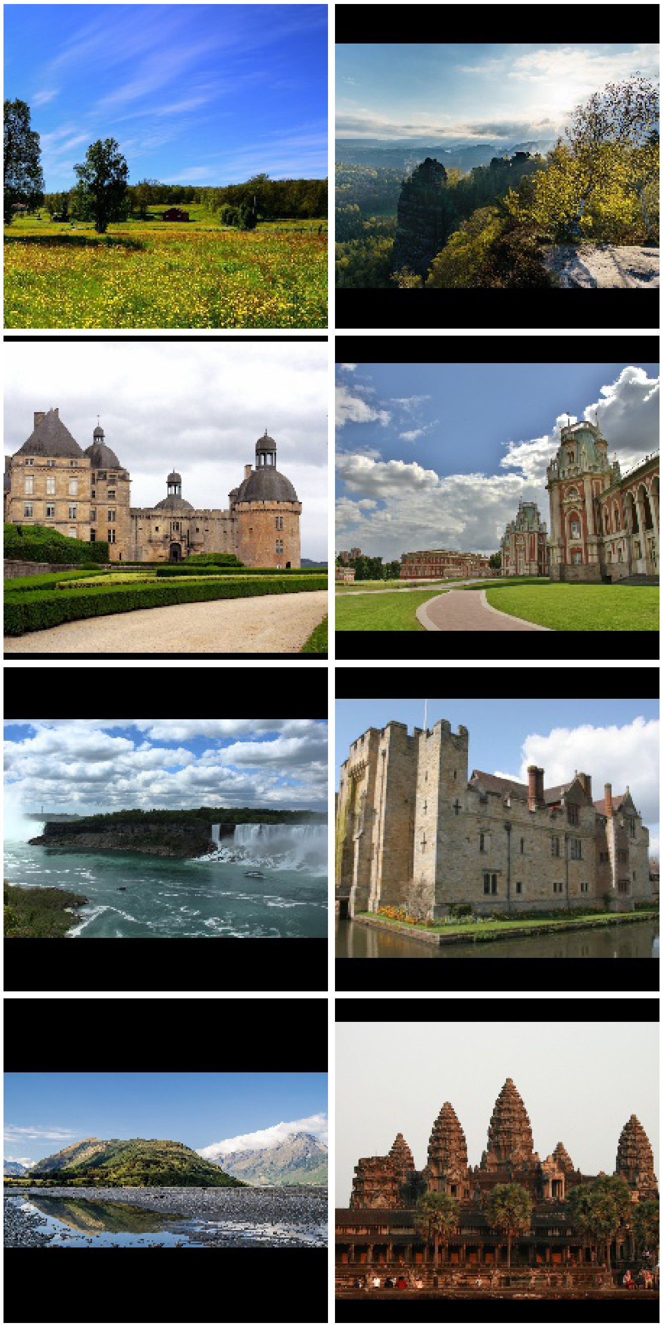

Fig. 19. Examples of controversial images. The images in the left column are assessed as natural by 5 observers and as unnatural by 4 observers. The images in the right column are assessed as unnatural by 5 observers and as natural by 4 observers.

VI. Conclusions

In this paper, 2 main contributions have been presented. Firstly, an experiment of subjective image naturalness classification without references was organized. It was performed under strict experimental conditions. From 45 observers, over 17,000 subjective naturalness evaluations for 1,900 SDR images have been obtained to establish a naturalness dataset for the purpose of analyzing photo naturalness automatically. Secondly, the image naturalness is evaluated in different ways using handcrafted features, features learned directly from CNN and transferred learned features. The experiments on the naturalness dataset point out the roles of the different feature types in the task of image naturalness evaluation. Handcrafted features are simple and quite efficient while deep learned features are complicated but get higher a performance and shallow learned features are not a good choice for analyzing image naturalness.

According to the current results, the direction of our research in the future is to organize an experiment with an HDR screen firstly to answer the question "are TMOs introducing unnaturalness in images or are the HDR images with unnatural artifacts even when displayed on an HDR screen?" and secondly to analyze the similar points of naturalness features for HDR images and SDR images. The second future direction is to develop a system giving naturalness score to images (regression problem). Finally, detecting unnatural images will be considered as the first step before developing methods in order to restore them.

References

[1] H. Seetzen, W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A. Ghosh, and S. Technologies, "High dynamic range display systems," *ACM Transactions on Graphics*, vol. 23, 05 2004.

[2] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High dynamic range imaging: acquisition, display, and image-based lighting, ISBN: 9780123749147, 9780080957111*, 2nd ed. Morgan Kaufmann publisher, 2010.

[3] K. Kim, J. Bae, and J. Kim, "Natural hdr image tone mapping based on retinex," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 4, pp. 1807–1814, November 2011.

[4] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '97. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 369–378.

[5] A. A. Goshtasby, "Fusion of multi-exposure images," *Image and Vision Computing*, vol. 23, no. 6, pp. 611–618, 2005.

[6] T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," in *Computer graphics forum*, vol. 28, no. 1. Wiley Online Library, 2009, pp. 161–171.

[7] T. Bashford-Rogers, M. Melo, D. Marnerides, M. Bessa, K. Debattista, and A. Chalmers, "Learning preferential perceptual exposure for hdr displays," *IEEE Access*, vol. 7, pp. 36 800–36 809, 2019.

[8] I. R. Khan, S. Rahardja, M. M. Khan, M. M. Movania, and F. Abed, "A tone-mapping technique based on histogram using a sensitivity model of the human visual system," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 4, pp. 3469–3479, April 2018.

[9] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, Jul. 2002.

[10] M. Ashikhmin, "A tone mapping algorithm for high contrast images," in *Proceedings of the 13th Eurographics Workshop on Rendering*, ser. EGRW '02. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2002, pp. 145–156.

[11] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," in *ACM transactions on graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 249–256.

[12] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *ACM transactions on graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 257–266.

[13] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 3, no. 4, pp. 291–306, 1997.

[14] M. Trentacoste, R. Mantiuk, W. Heidrich, and F. Dufrot, "Unsharp masking, countershading and halos: Enhancements or artifacts?" in *Computer Graphics Forum*, vol. 31, no. 2pt3. Wiley Online Library, 2012, pp. 555–564.

[15] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

[16] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: A survey," *CoRR*, vol. abs/1406.7799, 2014.

[17] H. de Ridder, F. J. Blommaert, and E. A. Fedorovskaya, "Naturalness and image quality: chroma and hue variation in color images of natural scenes," in *Human Vision, Visual Processing, and Digital Display VI*, vol. 2411. International Society for Optics and Photonics, 1995, pp. 51–62.

[18] H. de Ridder, "Naturalness and image quality: saturation and lightness variation in color images of natural scenes," *Journal of imaging science and technology*, vol. 40, no. 6, pp. 487–493, 1996.

[19] M. Cadik and P. Slavik, "The naturalness of reproduced high dynamic range images," in *Proceedings of the Ninth International Conference on Information Visualisation*, ser. IV '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 920–925.

[20] S. Y. Choi, M. Luo, M. Pointer, and P. Rhodes, "Investigation of large display color image appearance–iii: Modeling image naturalness," *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 31 104–1, 2009.

[21] K. Gu, S. Wang, G. Zhai, S. Ma, X. Yang, W. Lin, W. Zhang, and W. Gao, "Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 432–443, 2016.

[22] C. Yaacoub, J. Melhem, and P. Bilane, "A no-reference metric for quality assessment of tone-mapped high dynamic range images," *International Journal of Applied Engineering Research*, vol. 12, pp. 2598–2603, 06 2017.

[23] G. Jiang, H. Song, M. Yu, Y. Song, and Z. Peng, "Blind tone-mapped image quality assessment based on brightest/darkest regions, naturalness and aesthetics," *IEEE Access*, vol. 6, pp. 2231–2240, 2018.

[24] M. D. Fairchild, "The hdr photographic survey," in *Color Imaging Conference*, 01 2007, pp. 233–238.

[25] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of hdr tone mapping methods using essential perceptual attributes," *Computers & Graphics*, vol. 32, pp. 330–349, 2008.

[26] M. Narwaria, M. Perreira Da Silva, P. Le Callet, and R. Pépion, "Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality," *Optical Engineering*, vol. 52, no. 10, Oct 2013.

[27] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 657–667, Feb 2013.

[28] P. Korshunov, H. Nemoto, A. Skodras, and T. Ebrahimi, "Crowdsourcing-based evaluation of privacy in hdr images," *Optics, Photonics, And Digital Technologies For Multimedia Applications Iii*, vol. 9138, p. 11, 2014.

[29] L. Krasula, M. Narwaria, K. Fliegel, and P. Le Callet, "Preference of experience in image tone-mapping: Dataset and framework for objective measures comparison," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 64–74, Feb 2017.

[30] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "Large-scale crowdsourced study for tone-mapped hdr pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4725–4740, Oct 2017.

[31] S. Raman and S. Chaudhuri, "Bilateral filter based compositing for variable exposure photography." in *Eurographics (short papers)*, 2009, pp. 1–4.

[32] F. Pece and J. Kautz, "Bitmap movement detection: Hdr for dynamic scenes," in *2010 Conference on Visual Media Production*. IEEE, 2010, pp. 1–8.

[33] S. Paul, I. S. Sevcenco, and P. Agathoklis, "Multi-exposure and multi-focus image fusion in gradient domain," *Journal of Circuits, Systems and Computers*, vol. 25, no. 10, p. 1650123, 2016.

[34] P. P. Acharjya, R. Das, and D. Ghoshal, "A study on image edge detection using the gradients," *International Journal of Scientific and Research Publications*, vol. 2, no. 12, pp. 1–5, 2012.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[36] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct 2010.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computing Research Repository*, vol. abs/1409.1556, 2015.

[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Computing Research Repository*, vol. abs/1610.02357, 2016.

[39] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Computing Research Repository*, vol. abs/1512.03385, 2015.

[40] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning transferable architectures for scalable image recognition,” *Computing Research Repository*, vol. abs/1707.07012, 2017.

[41] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *Computing Research Repository*, vol. abs/1704.04861, 2017.

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” *Computing Research Repository*, vol. abs/1512.00567, 2015.

[43] G. Huang, Z. Liu, and K. Q. Weinberger, “Densely connected convolutional networks,” *Computing Research Repository*, vol. abs/1608.06993, 2016.

[44] C. Szegedy, S. Ioffe, and V. Vanhoucke, “Inception-v4, inception-resnet and the impact of residual connections on learning,” *Computing Research Repository*, vol. abs/1602.07261, 2016.

[45] K. Kira and L. A. Rendell, “A practical approach to feature selection,” in *Proceedings of the Ninth International Workshop on Machine Learning*, ser. ML92. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992, pp. 249–256.

[46] Q. T. Le, P. Ladret, H.-T. Nguyen, and A. Caplier, “Large Field/Close-Up Image Classification: From Simple to Very Complex Features,” in *Computer Analysis of Images and Patterns*, ser. Lecture Notes in Computer Science, M. Vento and G. Percannella, Eds., vol. 11679. Springer, 2019, pp. 532–543, proceedings of the 18th International Conference, CAIP 2019, Salerno, Italy, September 3–5, 2019, Part II. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02368500

[47] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

[48] T. J. DiCiccio and B. Efron, “Bootstrap confidence intervals,” *Statistical science*, pp. 189–212, 1996.

[49] C. Witzel and K. Gegenfurtner, *Memory Color*. New York, NY: Springer New York, 2014, pp. 1–7.

[50] J. Granzier and K. Gegenfurtner, “Effects of memory colour on colour constancy for unknown coloured objects,” *i-Perception*, vol. 3, pp. 190–215, 04 2012.