Hardness Prediction for Object Detection Inspired by Human Vision

Yuwei Qiu, Huimin Ma, Lei Gao

Department of Electronic Engineering, Tsinghua University
{qyw14@mails., mhmpub@,gao-l15@mails.}tsinghua.edu.cn

Abstract. We introduce eye tracking features including existing features like 1) scan path and 2) heat map, and novel features including 1) components of scan path and 2) peaks of heat to better define and understand human vision. In this paper, these features are used to describe the eye movements of a person when he/she is watching an image and looking for target object in it. Based on these features we define eye tracking complexity. Eye tracking complexity can be computed either by carrying out eye tracking experiments and extracting eye tracking features or through a convolutional neural network (CNN), which is introduced in this paper. This CNN computes eye tracking complexity directly from images. It has been validated that eye tracking complexity of an image corresponds to the detection algorithms average precision over an image. Thus, eye tracking complexity predicts the hardness of object detection, which can yield guidelines for the hierarchical algorithms design.

1 Introduction

Prediction of bad results of object detection is quite useful, with which algorithms can be designed as hierarchical to improve both efficiency and accuracy. When a person is watching a image, he/she knows whether it will be a tough task to find a target in the image. However, it is challenging for computers to predict how difficult it will be to detect objects before carrying out automatic detection algorithms.

Generally, object detection in human vision is transparent and all human beings carry out such task without any difficulty but with high accuracy and efficiency. However, automatic algorithms like neural networks or feature extraction are designed on a basis of math without taking any human perception into consideration. There is rare study combining *human factor* with automatic algorithms in the task of object detection. Inspired by the process of object detection in human vision, this paper aims at characterizing several critical features in human vision driven by the task of object detection, and further combines human factors into hierarchical algorithms design.

It is hard to reasonably describe human vision since the eye movements of a person during his/her observation is complex. According to some psychological studies [1, 2, 3, 4], eye tracking is one of the most prevalent methods used for analyzing eye movements. However, the raw eye tracking data are enormous two or three dimensional coordinates of gazes. Such forms are too massive to indicate any meaningful pattern of human vision. Valid features are required to be extracted.

In this paper, equipped with eye trackers, eye movements are recorded when people are watching images and looking for target object. The characteristics in eye tracking data like 1) the number of gazes, 2) the distance between gazes, 3) the density of gazes or the 4) angle of two neighboring scan lines is the basis of the eye tracking feature definition. Existing eye tracking features including 1) scan path and 2) heat map [5] are introduced. Novel features like 1) components of scan path and 2) peaks of heat are newly defined. These features are used to describe the eye movements.

Based on these features, a complexity metric is further defined: *eye tracking complexity (ETC)*. Eye tracking complexity can be computed either by carrying out eye tracking experiments and extracting eye tracking features, or through a convolutional neural network (CNN). Replacing time-consuming eye tracking experiments, the CNN can directly compute eye tracking complexity from raw images. Note that our purpose is to build direct connection between images and eye tracking complexity, or in other words, to classify images into categories according to the eye tracking complexity. CNN is adopted due to its well-known performance in image classification [6].

The effectiveness of the proposed complexity definition has been validated through numerical experiments. According to the results, eye tracking complexity corresponds to the average precision of automatic detection algorithms over an image. Namely, the eye tracking complexity predicts how difficult it is for the object to be detected by automatic algorithms. Based on the hardness prediction with eye tracking complexity, adaptations (e.g., changing object detection algorithms used for the image or slightly increasing or decreasing the iterations of the detection networks) can be carried out to improve the efficiency.

The contribution of this paper is above all the **novel method for description of human visual process**. Eye tracking features including 1) scan path, 2) heat map [5], 3) components of scan path and 4) peaks of heat are either introduced or newly defined. Moreover, a metric of image complexity (ETC) is defined based on these eye tracking features. It has been validated that eye tracking complexity corresponds to the average precision of automatic detection algorithms over the image and therefore can be used to predict the hardness for object detection. With eye tracking complexity, hierarchical algorithms can be designed to improve both efficiency and accuracy. We note that eye tracking complexity can be computed either by carrying out eye tracking experiments and extracting eye tracking features, or through a CNN.

The rest of this paper will be organized as follows. Section 2 is a review of related works including various definitions of image complexity, the application of CNN in computer vision and the studies using eye tracking. In Section 3, implementation details of eye tracking experiments are described. In Section 4, eye tracking features including existing ones (scan path and heat map) and novel ones (components of scan path and peaks of heat) are introduced. In Section

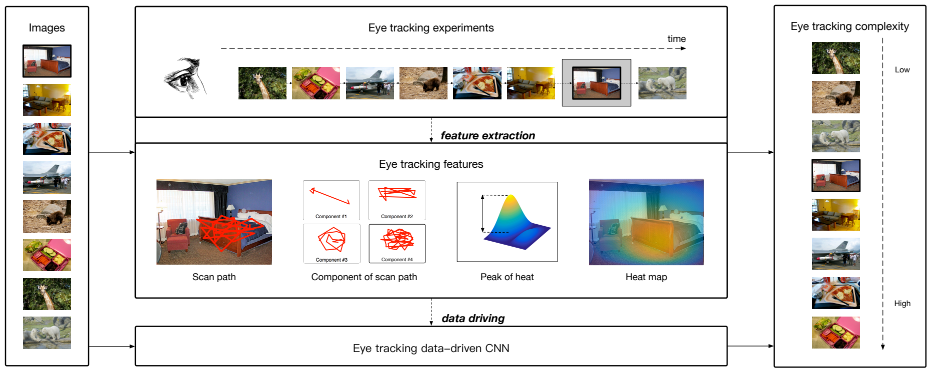

Fig. 1. Briefly summarize, eye tracking experiments are carried out and data are collected. Next, eye tracking features are extracted from the data and eye tracking complexity is computed. Also, a CNN is trained to compute eye tracking complexity directly from the images. Eye tracking complexity can be either computed from this CNN or eye tracking experiments.

5, the ETC is defined based on eye tracking features introduced in Section 4. In Section 6, the typical process of hardness prediction is built up. We draw conclusions and discuss about the numerical results in Section 7.

2 Related Works

There have been a lot of studies about image complexity and we discuss pros and cons of some of the previous definitions of image complexity. Meanwhile, the using of eye tracking data or human visual system (like saliency) is also introduced in this part.

2.1 Image Complexity

Digital images can be analyzed at wide range of levels ranging from pixel arrangement to the level of human understanding [7] and there have been numerous studies.

In the level of image analysis by machines[8, 9, 10, 11, 9, 12, 13]. The method adopting neural network from three aspects including texture, edge information and significant area to describe the complexity of the image has been raised in [8]. Also in [9], the researcher used color similarity. However, the features like texture, edges, or significant areas can only represent the local complexity and lose some global information.

As for studies using human perception over the content to describe image complexity[14, 15, 16], a measurement method based on SIFT & K-means algorithm, namely the estimation of the mismatch between the target and the interesting points has been introduced by Juan [16]. He carried out memory experiments, which asked people to memorize the content like objects, style or context of the images then analyzed the difference to study the complexity of the images. However, this work did not give a quantized description of peoples memory over images, which is one of the most significant limitation since automatic computation cannot be carried out if the complexity is involved with human behavior experiments.

2.2 Eye Tracking

In the field of psychology, eye tracking studies are useful for evaluating usability or analyzing human attention and more importantly understanding underlying cognitive processes based on the eye-mind hypothesis[17].

One of eye tracking studies is to identify key words of interests in a web page [18]. The visual experiences of the users are recorded and the analyzed data in the form of heat maps of the concern keywords explains the user eye behavior over the screen.

2.3 Convolutional Neural Network

A CNN contains one or more layers containing neurons with neighboring interactions, which has various implementations like face or object detection, text detection or recognition or image classification. The application of CNN in image classification is introduced in [6]. This architecture of CNN has been adapted to various forms for printed or handwritten character detection or recognition, face detection or recognition.

3 Eye Tracking Experiments

In this section, implementation details are shown for eye tracking experiments.

Participants: All of the 1280 times of eye tracking experiments are carried out by adults (aged above 18, with an educational level above college and without any mental diseases or ophthalmic diseases) and supervised by our group.

Eye tracker: Participants are equipped with *Tobii* eye trackers in the experiments and all of the eye movements are recorded. The average time for each time of experiment is around 6 minutes and the interval between two gazes recorded is 500 ms.

Task for participants: In the experiments, people are asked to watch a series of images switching over the computer screen. Participants are able to control the computer to switch from the first image toward the second one. Each of the participants has a different target object to seek out in one section of the experiment and this target appears somewhere in the images.

Images set: 1280 natural scene images used for participants to observe in the experiments are natural scenes from ImageNet database [19]. These images are chosen from 20 different classes in ImageNet. And for each class, 64 different images are selected randomly. Each experiment are divided into two sections of 12 images (6 images in one section). Each section contains 6 images from one

Classes	airplane	bear	chair	furniture	giraffe	glasses	helmet	horse	ladle	monkey
Num.	140	128	118	124	152	124	128	162	128	130
Classes	person	purse	piano	rabbit	sheep	sofa	table	tiger	whale	zebra
Num.	136	134	124	156	138	136	128	126	132	132

Table 1. The images used in the eye tracking images are chosen from the classes in this form. And for each classes, 6 different images are selected randomly. Times of valid experiments are also listed here. We note that for each time of experiment, two sections are contained with 12 images of two different classes. Therefore, there will be approximately 2560 groups of eye tracking data and each group is recorded for 6 images of the same class.

class, therefore, eye tracking data for one image is collected 12 times repeatedly in case of unexpected errors.

Note that: 1) Participants are required to take off their glasses during the process of the experiments; 2) In one experiment, the participants are not allowed to take any rest and all of their eyes movements are recorded; 3) All of the data collected from eye tracking experiments will be firstly filtered in case of malfunction of eye trackers. The number of valid experiments are listed in Table 1.

4 Features Extraction from Eye Tracking Data

Generally, the eye tracking features including 1) scan path and 2) heat map are existing ideas for eye tracking data analysis [5]. Apart from these, novel features are newly defined to further illustrate characteristics including 1) the number of gazes, 2) the distance between gazes, 3) the density of gazes or the 4) angle of two neighboring scan lines.

The eye tracking data contains massive points which denote gazes of human eye balls. Furthermore, the gazes are cut up in groups for each image and each group contains all of the gazes of a person when he/she is observing one image. In the following, detailed definitions of eye tracking features are introduced:

1) **Scan path**: A broken line connecting every two gazes which have neighboring time stamps. It illustrates the moving path of human eye balls and can be directly computed by raw eye tracking data.

2) **Heat map**: A map presenting the density of gazes over an image when people are watching it. Note that heat maps have various forms. Specifically, in this paper a gray level form is chosen. Basically, the denser the gazes located in an area, the higher the grey level (the darker) of the pixels. Besides, values of pixels in the heat map are normalized, ranging from 0 to 1.

Both the above two features can be easily extracted by simple signal processing techniques. Two more novel features are introduced in this paper:

3) **Components of scan path**: Massive gazes are connected in scan path, which is composed of basically four types of components: 1) one-way lines, 2) back-and-forth straight lines, 3) typical polygon and 4) dense polygon. These types of components can successfully form up any scan path (See Figure 2).

Moreover, judging from the 1) density of gazes and 2) the number of lines involved in each type of components, one-way lines, back-and-forth straight lines, typical polygon and dense polygon are scored 1,2,3 and 4. This score will be used for complexity definition in the next section.

We note that the judgment is made by detailed analyzing 1) the distance between points, 2) the angle of two neighboring lines and 3) the density of points. The main process of this judgment goes as follows:

- Given that two gazes at time t and time $t+1$ are adjacent, compute the distance between them. Carry out the same procedure to gazes at time $t+1$ and time $t+2$.
- Compute the distances between gaze t and $t+1$, and gaze $t+1$ and $t+2$.
- Compute the angle between line(t, $t+1$) and line($t+1$, $t+2$).
- By asking yes/no questions like "Is the distance between current gaze and last gaze larger than threshold?" or "Is the angle between current-next and current-last smaller than 90 degrees?", complex scan paths will be cut up into components (See Figure 4 to see the whole procedure).

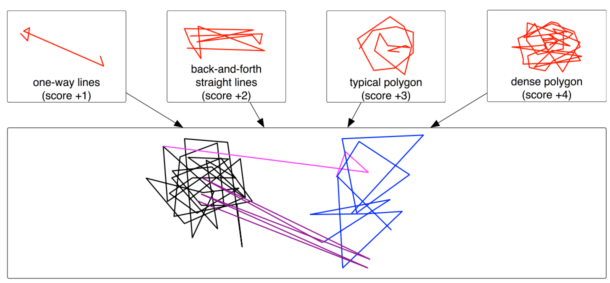

Fig. 2. Four types of components are shown in this figure. Any kind of scan path can be formed up by these components roughly.

4) **Peaks of heat**: In the heat map, which illustrates the density of gazes locally, local maximum value of density is picked as local peak of heat. Note that since local maximum values are not unique in a heat map, there will be more than one peak of heat in a heat map as well. Peaks will be picked by *steepest ascent*.

We refer the reader to Figure 3 to see an illustration.

5 Eye tracking complexity

In this section, the newly introduced eye tracking complexity is defined based on novel eye tracking features (components of scan path and peaks of heat).

Generally, the components of scan path illustrate how a person distributes his/her attention, or the time of observation. More complex components means

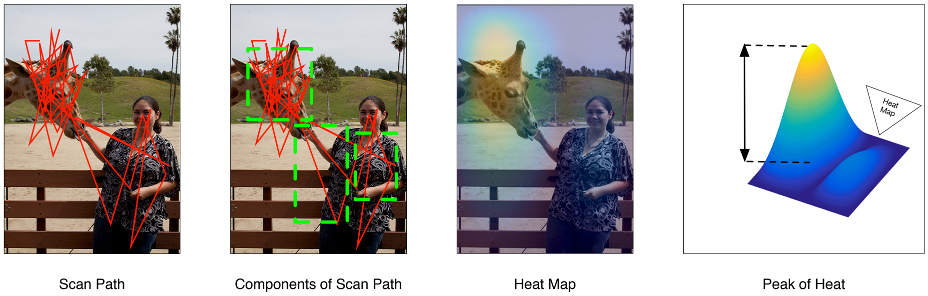

Fig. 3. Existing eye tracking features include scan path and heat map and novel features including scan path component and peak of heat.

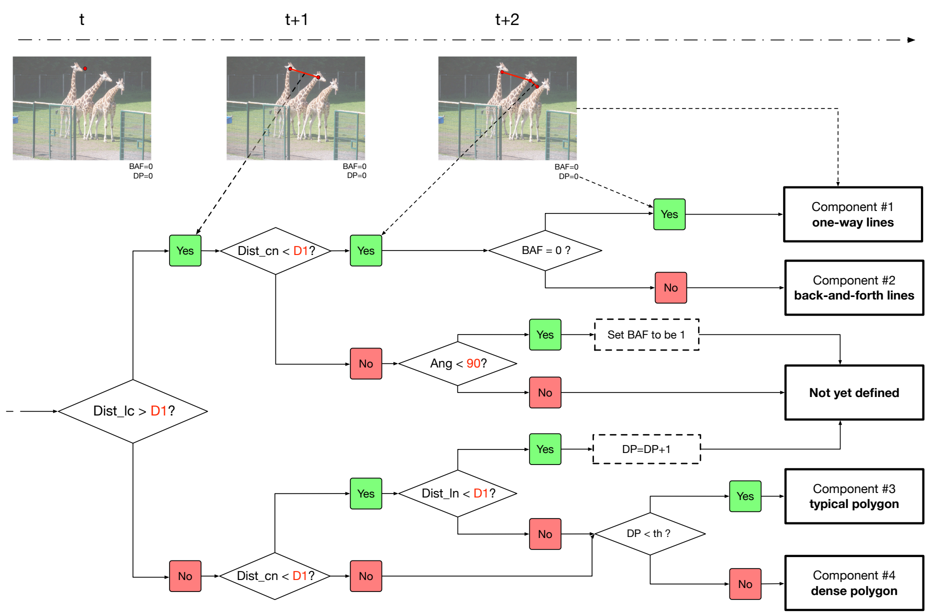

Fig. 4. Scan path segmentation can be accomplished by asking yes/no questions as it is shown in this figure. We note that $Dist - lc$ ($Dist - cn$) is the distance between last and current (current and next) gaze. Ang is the angle of the lines between current-last and current-next.

a higher complexity. The peak of heat describes the level of how much attention this person has paid to specific areas. A high value means the person may be confused or interested in that area. Human attention over a whole image should be measured as the sum of the distributions of attention in each part of the image.

Inspired by this, eye tracking complexity (ETC) of an image is defined as follows:

$$\mathrm{ETC} = \left\lfloor \sum_{c_i \in \mathcal{C}} s_i \times p_i + \frac{1}{2} \right\rfloor \tag{1}$$

where $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ is the set of scan path components extracted from the image, s_i and p_i is the corresponding score and the peak value of heat in local areas of the component c_i, respectively. Note that for each of the images, eye tracking complexity is an integer, which serves as labels in the classifier.

6 Hardness Prediction

In this section, the model for hardness prediction is presented.

High-level overview: Briefly summarize, eye tracking experiments are carried out for data collection at first. Then eye tracking features like scan path, heat map and novel features including components of scan path and peaks of heat are extracted from raw eye tracking data. With these features, the ETC is computed as it has been illustrated in Section 4 to evaluate the image complexity. For the next step, a classifier (CNN) is used to classify *gray level images* into categories. This CNN is trained with images labeled by eye tracking complexity. Eye tracking complexity of an image can be computed either by eye tracking features extraction or without time-consuming experiments with this fully trained CNN. Hierarchical algorithms can be designed according to this complexity to improve the detection efficiency.

6.1 Training the CNN

Training set: The eye tracking features, which is stated in Section 4, is slightly updated for network training in this part. The the number of input images is multiplied by 5 times with adaptations including 1) mirroring, 2) vertical clockwise rotation, 3) vertical anti-clockwise rotation and 4) horizontal rotation to be 6400. 80% of these images (5120) are used as training set and the rest (1280) is for testing.

Architecture: We adjust the architecture of *LeNet* [20] to the CNN used for hardness prediction because this network functions well in gray level images classification. Briefly summarize, this CNN contains two convolutional layers, two pooling layers and two fully connected layers. The input gray level images are re-sized to 32 × 32.

Fig. 5. All of the three CNNs performed equally on precision over images in **Category 1** (Bird, Tiger and Zebra) but quite different on **Category 12** (Lamp).

6.2 Hardness prediction

Images for test: To validate the effectiveness of eye tracking complexity as hardness prediction for object detection, 72 images are chosen from ImageNet. These 72 images are from 8 categories (each category has 9 images, randomly selected). 4 of the 8 categories are the hardest (with the highest possibility for algorithms to fail) and the rest of them are the easiest (with the lowest possibility for algorithms to fail) classes for object detection according to the result of ILSVRC 2016.

Note that the so-called “hardest” and “easiest” classes are illustrated by ImageNet in the purpose of warning participants of ILSVRC worldwide. Generally, hardest classes attain the lowest mean Average Precision (mAP).

Eye tracking complexity can also be computed by experiments. However, in our experiment, the classifier CNN replaces eye tracking experiments for the computation of eye tracking complexity of these images and the output is shown in Table 2.

Results: Generally, images of the hardest classes (Backpack, Ladle, Lamp and Microphone) cluster in **Category 11** (with eye tracking complexity of 15) and **12** (with eye tracking complexity of 16) while the easiest classes (Bird, Dog, Tiger and Zebra) in **Category 1** (with eye tracking complexity of 5). We note that the only one image of Class Lamp clustered in Category 11 might be caused by unsteadiness of attention distraction in eye tracking experiments.

This result indicates that the output of this CNN, namely the eye tracking complexity, corresponds to the performance of automatic algorithms. Eye

Classes	Bird	Dog	Tiger	Zebra	Backpack	Lamp	Ladle	Microphone
Output Ca.	9 in Ca.1	9 in Ca.1	9 in Ca.1	9 in Ca.1	9 in Ca.12	9 in Ca.12	8 in Ca.12 1 in Ca.11	9 in Ca.12

Table 2. Generally, images of the hardest classes (Backpack, Ladle, Lamp and Microphone) cluster in **Category 11** (with eye tracking complexity of 15) and **12** (with eye tracking complexity of 16) while the easiest classes (Bird, Dog, Tiger and Zebra) in **Category 1** (with eye tracking complexity of 5). We note that the only one image of Class Lamp clustered in Category 11 might be caused by unsteadiness of attention distraction in eye tracking experiments.

tracking complexity can predict the hardness of object detection for automatic algorithms.And the CNN adopted in our experiment can replace eye tracking experiments for eye tracking complexity automated computation.

Classes	Bird	Dog	Tiger	Zebra	Backpack	Lamp	Ladle	Microphone	AETC
Set 1	12	12	12	12	0	0	0	0	**4.8**
Set 2	6	6	6	6	6	6	6	6	**10.1**
Set 3	0	0	0	0	12	12	12	12	**14.8**

Table 3. Three image datasets are shown in this table. The numbers of images from each class and the average eye tracking complexity (AECT) are also listed here.

mAP	Set 1	Set 2	Set 3
CNN-100	**0.632**	0.142	0.017
CNN-1000	**0.637**	0.254	0.086
CNN-10000	**0.637**	0.365	0.118

Table 4. As is shown in this table, increasing iterations does improve the performance of automatic algorithms over datasets that consist of images with higher eye tracking complexity while a dataset with lower eye tracking complexity, a larger number of iterations is invalid for performance improvement.

6.3 Hierarchical algorithms design

According to the eye tracking complexity of an image, hardness of detection can be predicted in advance, which means hierarchical algorithms can be designed accordingly.

One of possible design for hierarchical algorithms based on eye tracking complexity is presented here.

Networks: Three CNNs with different iterations (CNN-100, CNN-1000 and CNN-10000) are all trained with the same starting point and the same training sets including 50,000 images for training and 10,000 for validation. All of the images used for training are from ImageNet.

Test datasets: Three test datasets are built up and the contents are shown in Table 3. Average eye tracking complexity (AETC) of each dataset is also listed here. CNN-100, CNN-1000 and CNN-10000 are tested on all of these three datasets and the mean Average Precision (mAP) of each dataset is shown in Table 4.

Results: The results of hierarchical, non-hierarchical algorithms and bounding boxes are shown in Figure 5.

What has been shown from the results is that: 1) All of the three CNNs performed equally on precision over images in **Category 1** (Bird, Tiger and Zebra) but quite different on **Category 12** (Lamp). 2) Increasing iterations does improve the performance of automatic algorithms over datasets that consist of images with higher average eye tracking complexity. However, for a dataset with lower eye tracking complexity, a large number of iterations is invalid for performance improvement.

Therefore, in computation of eye tracking complexity before training detection network or designing algorithms, efforts in vain can be avoided by choosing suitable iteration times.

Further studies about eye tracking complexity will be focused on performance of networks with different 1) depth, 2) architecture or 3) kernels, on images with different eye tracking complexity, which yields more guidelines for hierarchical detection algorithms design.

7 Conclusion and discussion

To sum up, eye tracking features including existing features including 1) scan path and 2) heat map, and novel features including 1) components of scan path and 2) the peak of heat are introduced in this paper. Based on these features, a kind of image complexity, eye tracking complexity, and a CNN for its computation in place of eye tracking experiments are introduced in this paper. Generally, this complexity corresponds to the hardness of object detection. Based on that, hierarchical algorithms can be designed accordingly. .

References

[1] K. Kiili, H. Ketamo, and M. D. Kickmeier-Rust, "Evaluating the usefulness of eye tracking in game-based learning," *International Journal of Serious Games*, vol. 1, no. 2, 2014.

[2] C. S. Longman, A. Lavric, and S. Monsell, "More attention to attention? an eye-tracking investigation of selection of perceptual attributes during a task switch." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 39, no. 4, p. 1142, 2013.

[3] M. Kassner, W. Patera, and A. Bulling, "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication.* ACM, 2014, pp. 1151–1160.

[4] A. Duque and C. Vázquez, "Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 46, pp. 107–114, 2015.

[5] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "State-of-the-art of visualization for eye tracking data," in *Proceedings of EuroVis*, vol. 2014, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[7] M. Cardaci, V. Di Gesù, M. Petrou, and M. E. Tabacchi, "A fuzzy approach to the evaluation of image complexity," *Fuzzy Sets and Systems*, vol. 160, no. 10, pp. 1474–1484, 2009.

[8] Y.-Q. Chen, J. Duan, Y. Zhu, X.-F. Qian, and B. Xiao, "Research on the image complexity based on neural network," in *Machine Learning and Cybernetics (ICMLC), 2015 International Conference on*, vol. 1. IEEE, 2015, pp. 295–300.

[9] S. Yang, P. Gao, F. Meng, X. Jiang, and H. Liu, "Objective image quality assessment based on image complexity and color similarity," in *Software Engineering (WCSE), 2013 Fourth World Congress on*, vol. 1. IEEE, 2013, pp. 5–9.

[10] J. Rigau, M. Feixas, and M. Sbert, "An information-theoretic framework for image complexity," in *Proceedings of the First Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging*. Eurographics Association, 2005, pp. 177–184.

[11] R. A. Peters and R. N. Strickland, "Image complexity metrics for automatic target recognizers," in *Automatic Target Recognizer System and Technology Conference*. Citeseer, 1990, pp. 1–17.

[12] H. Yu and S. Winkler, "Image complexity and spatial information," in *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*. IEEE, 2013, pp. 12–17.

[13] F. Yaghmaee and M. Jamzad, "Introducing a new method for estimation image complexity according to calculate watermark capacity," in *Intelligent Information Hiding and Multimedia Signal Processing, 2008. IIHMSP'08 International Conference on*. IEEE, 2008, pp. 981–984.

[14] M. P. Da Silva, V. Courboulay, and P. Estraillier, "Image complexity measure based on visual attention," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 3281–3284.

[15] I. Mario, M. Chacon, D. Alma, and S. Corral, "Image complexity measure: a human criterion free approach," in *NAFIPS 2005-2005 Annual Meeting of the North American Fuzzy Information Processing Society*. IEEE, 2005, pp. 241–246.

[16] J. Huo, "An image complexity measurement algorithm with visual memory capacity and an eeg study," in *SAI Computing Conference (SAI), 2016*. IEEE, 2016, pp. 264–268.

[17] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognitive psychology*, vol. 8, no. 4, pp. 441–480, 1976.

[18] N. Kumar, V. Maheshwari, and J. Kumar, "A comparative study of user experience in online social media branding web pages using eye tracker," in *2016 International Conference on Advances in Human Machine Interaction (HMI)*. IEEE, 2016, pp. 1–8.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[20] Y. LeCun, C. Cortes, and C. J. Burges, "The mnist database of handwritten digits," 1998.