Human-centric Image Rendering for Natural and Comfortable Viewing—Image Optimization Based on Human Visual Information Processing Models

Taiki Fukiage

Abstract

As display technology and devices advance, using any surface or space as a display screen is becoming possible. However, emerging technologies that use projectors and see-through displays face challenges in maintaining consistent image quality, as the appearance of the displayed image can vary significantly depending on factors such as ambient light and background patterns. The key to solving this problem is understanding how the human visual system works. In this article, I introduce an approach that addresses this issue by modeling the visual information processing of the human brain. This model enables us to optimize displayed images to ensure they are perceived as intended despite environmental variations.

Keywords: media display technology, human information science, visual information processing model

1. Media technology based on understanding of human vision

Visual media, which are media for transmitting and sharing visual information, have evolved in various forms from paintings and photographs to televisions, projectors, smartphones, and head-mounted displays (HMDs)[*1]. These media have become indispensable in our daily lives. As technology advances, it is expected that information will be seamlessly presented in every space, effectively turning our surroundings into displays in the near future. How can we ensure that visual information is conveyed as intended across these diverse media? Ideally, reproducing a real scene would involve capturing and playing back all the information from the physical space. However, such ultimate media devices do not yet exist, and the degree of reproduction is constrained by the physical limitations of each device, such as the intensity, wavelength, and resolution of the light they can display. To convey information as intended within these physical constraints, it is crucial to understand how humans process, perceive, and recognize visual information.

Let us take an example of the technology behind color monitors. Human retinas have cells that respond to light in specific wavelength ranges, corresponding to red, green, and blue. Our perception of color arises from the combination of these responses—a phenomenon known as trichromatic vision. Leveraging this knowledge, modern displays recreate a vast spectrum of colors by blending red, green, and blue light. Similarly, three-dimensional (3D) televisions and HMDs convey 3D depth information on the basis of an understanding of human stereoscopic vision. Our brains perceive depth through binocular disparity—subtle differences in the images seen by each eye.

*1 HMD: A display device worn on the head. By projecting images directly in front of the eyes, it provides a highly immersive visual experience.

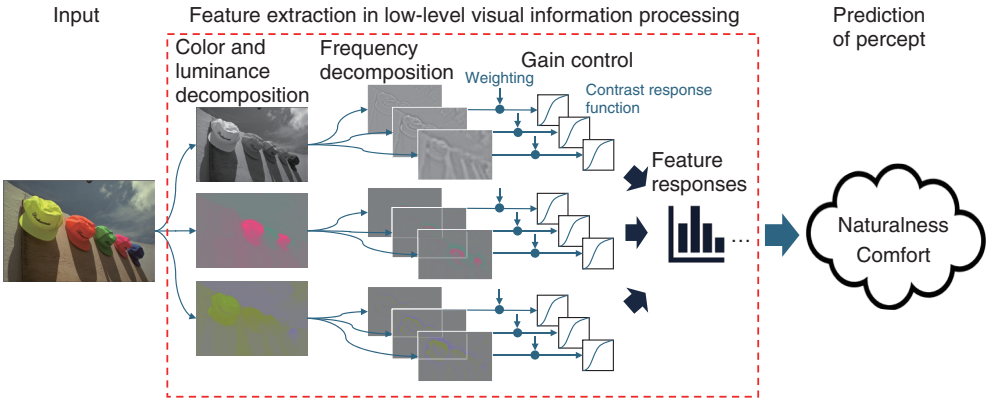


Fig. 1. An overview of visual information processing model.

Using this principle, 3D televisions and HMDs present different images to each eye, enabling viewers to experience a sense of three-dimensionality without the need for a physical 3D space. Thus, understanding and exploiting the characteristics of human vision allows for the efficient reproduction of perceived realities without fully replicating the physical world.

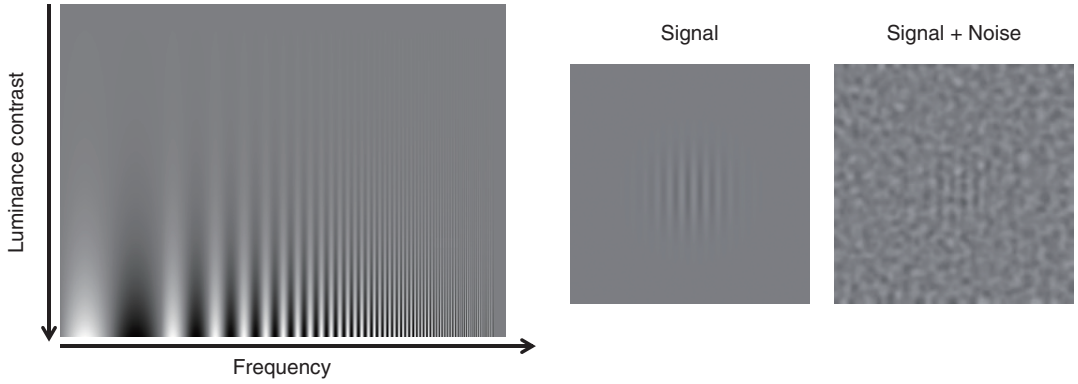
While the examples discussed thus far focus on designing display devices to align with human visual characteristics, the future of information presentation technology poses new challenges. In emerging technologies with which real and virtual information coexist, the appearance of displayed content is expected to change dynamically across different viewing environments. In such scenarios, predesigned devices alone will not be sufficient for optimal results. Instead, we will need to optimize the content itself in real time for each specific situation. To achieve this, an effective approach is to use a *visual information processing model* capable of quantitatively predicting perception for any given image and optimize the presented visuals on the basis of these predictions.

2. Visual information processing model

A visual information processing model is a mathematical representation of how the brain processes visual information. **Figure 1** illustrates the processing flow of a visual information processing model, which is discussed in this article. This model takes any image as input and extracts features we use when recognizing the input. It then predicts the intensity of our sensory response to these features (feature responses). Finally, the model estimates important indicators for visual presentation, such as naturalness of appearance and visual comfort, on the basis of these extracted features.

What exactly are these "features"? Our visual system extracts and uses various features from the information that enters the retina to recognize the world and guide actions. This feature extraction process is hierarchical. It begins with simple features such as color and luminance contrast (differences in luminance) in localized areas. It then progresses by integrating these features to detect more complex and global characteristics such as orientation, shape, texture, and eventually faces, objects, and landscapes. However, only a limited portion of this feature extraction process has been established as concrete, practical computational models. In the following sections, I focus on explaining low-level visual information processing, which has been used in the research examples covered in this article.

The specific process of feature extraction with a low-level visual information processing model is illustrated within the dashed box in Fig. 1. Let us begin by explaining *color and luminance decomposition*. Our retinas have cone cells, which are sensors corresponding to three wavelength bands: red, green, and blue. The light information received with these sensors is converted into a format called opponent colors, which emphasizes color differences while efficiently transmitting color information for subsequent processing. The color and luminance decomposition process mimics this color processing mechanism of the human visual system. It decomposes the input image by adding and subtracting the red, green, and blue color channels. This results in three components: one representing luminance and two opponent



(a) Contrast sensitivity for each frequency (b) Contrast masking effect

Fig. 2. Demonstration of contrast perception.

color components expressing the differences between red and green and between blue and yellow.

Next, the images corresponding to each color component undergo *frequency decomposition*. Frequency represents the spatial fineness of patterns. The human visual system has neurons that selectively respond to various levels of fineness, and these responses represent the frequency characteristics within the retinal image. The low-level visual information processing model uses image processing called *convolution* to reproduce this frequency-based information representation. Convolution yields images that represent contrast at various frequency scales. Finally, by applying weights to each frequency component, the model reflects the varying sensitivities of the human visual system to different frequencies [1]. **Figure 2(a)** illustrates this difference in sensitivity across frequencies. In this image, frequency increases (patterns become finer) from left to right, while physical contrast decreases from bottom to top. Although the physical contrast is constant at the same vertical level regardless of frequency, the boundary between visible and invisible stripe patterns appears as an upward curved line. This curve illustrates the visual system's varying sensitivity to different frequencies. Specifically, the visual system is most sensitive to patterns of intermediate fineness and less sensitive to very coarse or very fine patterns.

Finally, let us discuss *gain control*. This process is closely related to the perceptual strength of contrast. The visual system adjusts the gain of neural responses to accommodate a wide range of contrasts. Initially, the response increases rapidly with physical contrast, but it gradually levels off in high-contrast regions [2]. This behavior is illustrated by the contrast response function shown in Fig. 1, where the horizontal axis represents physical contrast and the vertical axis represents neural response.

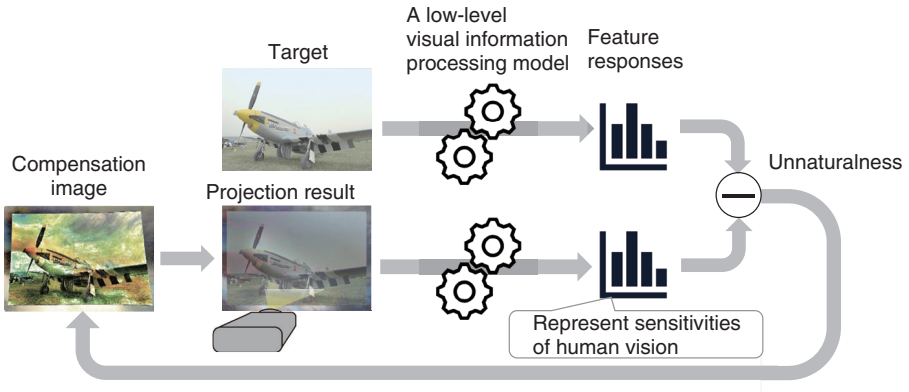
The contrast masking effect is a specific example that supports the presence of the gain control mechanism. In **Fig. 2(b)**, both left and right images have stripe patterns embedded at the same contrast. However, the stripes on the right, superimposed on background noise, appear much less visible. This can be explained by the strong neural response already triggered by the background noise, which makes the additional response to the stripes relatively small. In the low-level visual information processing model, these gain control mechanisms are mathematically expressed to quantitatively predict the perceptual magnitude of the visual system's response to each feature.

3. Optimizing display images using visual information processing models

As described above, a visual information processing model converts arbitrary images into features that reflect the sensitivities of the visual system. I will now explain what can be achieved using the model on the basis of research we have conducted.

3.1 Natural appearance manipulation of real object surfaces

First, let us look at research on spatial augmented reality (AR) using projectors. This technology, also known as projection mapping, allows manipulation of the appearance of real object surfaces. While it is



(a) Optimization of a compensation image with a visual information processing model

(b) Comparison of projection results

Fig. 3. Natural appearance manipulation of real object surfaces.

predominantly used for large-scale shows and demonstrations today, it has the potential for various information displays in more everyday settings. One technical challenge that needs to be addressed in such scenarios is the problem of interference between the object's own patterns and the projected image. A solution to this problem is a technique called radiometric compensation. This technique captures the projection surface with a camera and modifies the projected image to cancel out the surface patterns [3]. Since projectors cannot output negative light to cancel out light, for example, if the projection surface has a red pattern, cyan light is projected to neutralize the color, then the desired color is added to create the final projected image. However, in bright ambient light, the contrast of the surface pattern increases, requiring much stronger light to cancel it out. Typical projectors may not be able to output such strong light, making it impossible to fully compensate for the patterns.

Using the sensitivity characteristics of the vision system can be very effective in solving this problem. By prioritizing the reproduction of features to which humans are highly sensitive, while sacrificing features with lower sensitivity, it is possible to achieve perceptually natural results even if physical compensation is not perfect. We used a low-level visual information processing model to achieve this [4]. The specific procedure is shown in **Fig. 3(a)**. First, the target image and camera-captured image of the projection result are input into the model and converted into perceptual feature representations. Since these features represent the sensitivities of the visual system, the magnitude of the difference between these features can be regarded as the perceptual unnaturalness of the projection result. We then optimize the compensation image to minimize this unnaturalness. This automatically produces projection results that, while not physically identical to the target, are perceptually natural. Examples of actual optimization results are shown in **Fig. 3(b)**. While the physics-based method barely compensates for the surface

pattern, the perception-based compensation using a visual information processing model achieves a result that is perceptually much closer to the target image.

A similar method was also used to address the challenges of the projection technique called "HenGenTou" we previously developed. HenGenTou creates an illusion of motion in stationary real objects by projecting black and white dynamic patterns that express object motion [5]. However, there was a limit to the size of movement that could appear natural, and fine manual adjustments were previously necessary. To address this issue, we developed a method that uses a visual information processing model to predict the naturalness of the projection result and automatically optimize motion information [6]. This enables us to achieve maximum movement within a range that does not feel unnatural, enabling effective use of HenGenTou in interactive applications, such as moving the expressions of paintings to match user expressions.

3.2 Comfortable semi-transparent visualization on real-world scenes

In media technologies such as virtual reality (VR)*2 and AR*3, which are expected to cover the entire field of view, information is often displayed semi-transparently to avoid obstructing the view. However, in situations where the background real scene is constantly changing, it is generally difficult to maintain consistent visibility of the overlaid content. This is because visibility is greatly affected by the contrast of the background, as illustrated with the example of the contrast masking effect mentioned earlier. However, using a visual information processing model, it is possible to quantitatively predict changes in the visibility of such semi-transparent images. We previously proposed a technique that automatically adjusts transparency using a visibility prediction model that is based on a visual information processing model [7]. As shown in **Fig. 4(a)**, this method enables users to specify the target visibility rather than the physical transparency. When the content and background are given, the visibility prediction model predicts the visibility of the blended transparent image. The transparency map is then optimized to minimize the difference between the target visibility and predicted visibility. **Figure 4(b)** shows example results. The same content is displayed transparently over two different backgrounds. In the results of standard blending, even with the same transparency settings, the visibility of the content image varies greatly depending on the background. With the proposed method, however, the transparency is optimized in accordance with the target visibility map, resulting in consistent content visibility across different backgrounds. Therefore, our proposed method enables users to directly manipulate perceptual attributes such as visibility, resulting in more intuitive and precise control over transparent compositing. This method opens up exciting possibilities for applications in interactive media such as VR and AR, where it could enable semi-transparent displays that consistently maintain comfortable visibility across varying backgrounds.

4. Future challenges and prospects

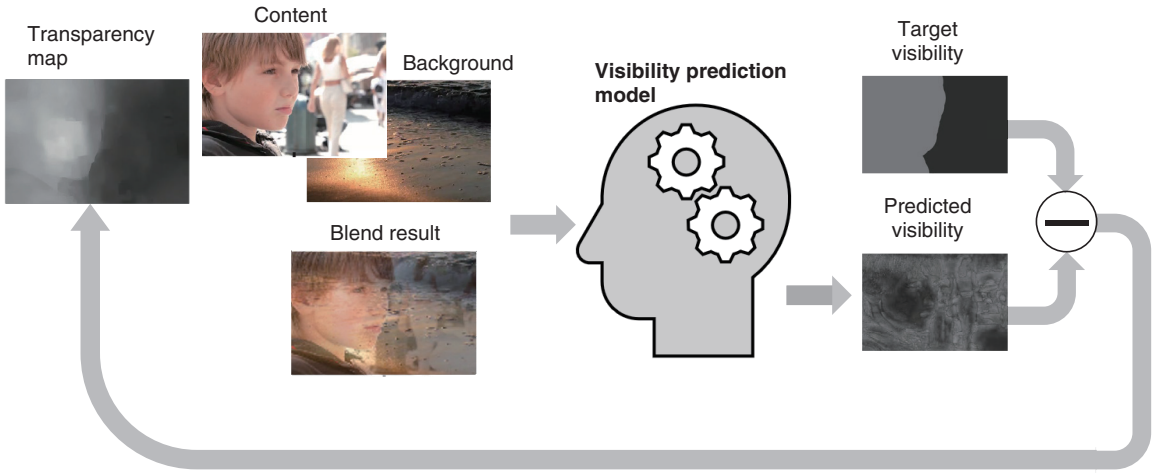
The optimization of content using visual information processing models is expected to become increasingly important in future media technologies. However, there are still many challenges to be addressed with this approach. First, the current visual information processing models for image optimization only cover a very small part of the complex visual information processing occurring in the human brain, corresponding to the initial stages. To advance future research, we need to progress towards modeling middle to higher-level information processing. For example, by enabling the prediction of texture, depth, motion, and material perception, it will be possible to adapt the presentation images more flexibly without changing these impressions.

However, the construction of higher-level processing models faces limitations when using the component-based approach classically used in low-level visual modeling, which involves understanding and assembling visual information processing in small sub-processes. Deep learning models are considered promising to address this issue. By training deep learning models on tasks such as object recognition, they learn to execute complex information processing tasks autonomously, from analyzing input images to generating task-specific outputs. It is noteworthy that the similarity between deep learning models trained for object recognition and human brain information processing has been revealed from various perspectives [8].

Nevertheless, these models do not quantitatively match human perception, thus cannot be directly used for image optimization. There are also reports

*2 VR: A technology that immerses users in a virtual visual world created using computers.

*3 AR: A technology that overlays virtual information on the real world, making information delivery more intuitive and convenient.



(a) Optimization of transparency map by visibility prediction using a visual information processing model

(b) Comparison of blend results

Fig. 4. Comfortable semi-transparent visualization using a visual information processing model.

suggesting that as performance improves, the divergence from human perception increases [9]. In the future, it will be necessary to develop methods to train deep learning models while enhancing their alignment with human perception.

Along with advancing the modeling of visual information processing, it is crucial to clarify the necessary conditions for naturalness and comfort from a human perspective. As seen in examples such as Escher's impossible staircase, humans can perceive physically impossible situations as natural at first glance. Therefore, the distribution of images that humans perceive as natural is thought to have a broader range than the distribution of images faithfully reproduced according to physics. By accurately estimating the spread of this distribution, we can expect to further expand the range of visual expression within various environmental and physical constraints.

Some of the results introduced in this article are from joint research with the University of Tokyo.

References

[1] F. W. Campbell and J. G. Robson, "Application of Fourier Analysis to the Visibility of Gratings," The Journal of Physiology, Vol. 197, No. 3, pp. 551–566, 1968. https://doi.org/10.1113/jphysiol.1968.sp008574

[2] D. J. Heeger, "Normalization of Cell Responses in Cat Striate Cortex," Visual Neuroscience, Vol. 9, No. 2, pp. 181–197, 1992. https://doi.org/10.1017/s0952523800009640

[3] M. D. Grossberg, H. Peri, S. K. Nayar, and P. N. Belhumeur, "Making One Object Look Like Another: Controlling Appearance Using a Projector-camera System," Proc. of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), Washington, DC, USA, June/July 2004. https://doi.org/10.1109/CVPR.2004.1315067

[4] R. Akiyama, T. Fukiage, and S. Nishida, "Perceptually-based Optimization for Radiometric Projector Compensation," Proc. of the 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW 2022), Christchurch, New Zealand, Mar. 2022. https://doi.org/10.1109/VRW55335.2022.00226

[5] T. Kawabe, T. Fukiage, M. Sawayama, and S. Nishida, "Deformation Lamps: A Projection Technique to Make Static Objects Perceptually Dynamic," ACM Transactions on Applied Perception, Vol. 13, No. 2, pp. 1–17, 2016. https://doi.org/10.1145/2874358

[6] T. Fukiage, T. Kawabe, and S. Nishida, "Perceptually Based Adaptive Motion Retargeting to Animate Real Objects by Light Projection," IEEE Transaction on Visualization and Computer Graphics, Vol. 25, No. 5, pp. 2061–2071, 2019. https://doi.org/10.1109/TVCG.2019.2898738

[7] T. Fukiage and T. Oishi, "A Content-adaptive Visibility Predictor for Perceptually Optimized Image Blending," ACM Transaction on Applied Perception, Vol. 20, No. 1, Article no. 3, pp. 1–29, 2023. https://doi.org/10.1145/3565972

[8] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex," PNAS, Vol. 111, No. 23, pp. 8619–8624, 2014. https://doi.org/10.1073/pnas.1403112111

[9] M. Kumar, N. Houlsby, N. Kalchbrenner, and E. D. Cubuk, "Do Better ImageNet Classifiers Assess Perceptual Similarity Better?", TMLR, 2022.

Taiki Fukiage

Senior Research Scientist, Sensory Representation Group, Human Information Science Laboratory, NTT Communication Science Laboratories.

He received a Ph.D. in interdisciplinary information studies from the University of Tokyo in 2015. He joined NTT Communication Science Laboratories in 2015, where he studies media technologies based on scientific knowledge about visual perception. He is a member of the Vision Sciences Society and the Vision Society of Japan.