Computer Science and Information Systems 20(2):879–892 https://doi.org/10.2298/CSIS220919017S

The Duration Threshold of Video Content Observation: An Experimental Investigation of Visual Perception Efficiency

Jianping Song[1], Tianran Tang[2], and Guosheng Hu[3*]

[1] School of Art and Design, Zhejiang A&F University,
Hangzhou 311300, China
[2] Department of Computer Engineering, Dongguan Polytechnic,
Dongguan 523808, China
[3] Design Innovation Center, China Academy of Art,
Hangzhou 310002, China
huguosheng@msn.com

Abstract. Visual perception principle of watching video is crucial in ensuring video works accurately and effectively grasped by audience. This article proposes an investigation into the efficiency of human visual perception on video clips considering exposure duration. The study focused on the correlation between the video shot duration and the subject's perception of visual content. The subjects' performances were captured as perceptual scores on the testing videos by watching time-regulated clips and taking questionnaire. The statistical results show that three-second duration for each video shot is necessary for audience to grasp the main visual information. The data also indicate gender differences in perceptual procedure and attention focus. The findings can help for manipulating clip length in video editing, both via AI tools and manually, maintaining perception efficiency as possible in limited duration. This method is significant for its structured experiment involving subjects' quantified performances, which is different from AI methods of unaccountable.

Keywords: Video Editing, Human Visual System, Perceived Efficiency, Footage Duration, Video Information.

1. Introduction

Along with the booming of computer vision technology in recent years, artificial intelligence (AI) techniques have been extensively applied in visual arts. Since IBM's Watson created the movie trailer for *Morgan* in 2016, AI application in video editing has become dramatically widespread and in-depth. However, video editing always requires an understanding of the perception mechanisms and laws of human vision system (HVS), no matter how advanced the editing technology is. The perceptual laws and principles of HVS is a key knowledge and prerequisite for ensuring that video works are correctly and effectively understood by audience. In fact, scientific researchers have

[*] Corresponding author

been focusing on the application of HVS mechanism in image processing since 1980s, and video artists are definitely the peer that started to focus on such issues earlier.

Video artists, especially directors and film editors, need to effectively manipulate the duration of each shots when dealing with the relationship between the narrative flow of videos and the audience's perception. Traditionally, artists rely on their visual experience and professional skills to fulfill their intentions. While, due to the widespread usage of computer vision tools of AI technology, it is urgently in need for definite quantification on such factors that directly affect the legible, definable and understandable of video works. Undoubtedly, clear quantitative laws can improve our understanding of watching behavior and be directly applied to AI-driven or manual video creation. Besides, these laws may benefit AI technology in recognition of video contents [1-3].

Technically, physiological and psychological fluctuation could rise on viewing the video contents, which intricately affects the audience's mastery and understanding of the video contents. Among them, the duration of the shots can directly affect the audience's perception of the video information. Video content is often made up of consecutive shots, each shot needs a reasonable length of exposure to meet the visual perception requirements. The audience will not have enough time to grasp the main content of the shot in inadequate duration, while lengthy duration will occupy the audience’s attention to the minor details of the scenes, thus distracting their attention from the main theme. In addition, audience’s visual perception of video is also influenced by the content of the images per se, and it should not be ignored that information density of video is also an important factor for cognitive efficiency.

This visual perception issue not only determines whether AI technology can meet the needs of audience-oriented video creation, but also is a technical problem that has long troubled video artists. To reveal this issue, it is necessary to analyze audience’s perceptual efficiency on various scenes through duration counted experiments. The experiment of this project statistically analyzed the efficiency and accuracy of audience's visual perception of video contents from the video editing perspective. With focusing on the perceptual perceptions of exposure durations and scene contents, the experiment aims to clarify the necessary shot duration for the audience to perceive and grasp the main information in the video.

Six video clips of different perspectives and contents were collected, and composed into five video clips of different durations as controlled experimental materials. Then, the subjects were divided into five groups according to the very video clip they watched and finished a questionnaire. Finally, the census data were analyzed and discussed according to the questionnaire results. Despite the fact that the experiment faced subjective factors, the results still shows a clear tendency. However, unexpected interference of the experimental results caused by factors, such as certain contents and gender differences, was also found.

This paper consists of six sections. Cognitive mechanism of video and multimedia are reviewed in section 2. Section 3 is about conducting the experiment that involves volunteers’ participation. The data of the experimental results are measured and analyzed in section 4, and discussion is conducted in section 5. Section 6 shows the summary of the conclusions.

2. Related Research

Rather than a complex optical imaging system, HVS is influenced by a variety of operating mechanisms that is often explained as psychological and neurophysiological issues, and is highly intelligent for information perception and processing. Marr, an expert in neurology and psychology, studied the visual system from the information processing perspective and proposed, for the first time, a sophisticated explanation of the basic construction of the visual system, which laid the theoretical foundation for the application of human visual cognitive mechanisms in image processing [4,5]. Since then, researchers have continued to improve the vision theory from multi-aspects, such as information acquisition, visual sensitivity, content perception, and gradually applied them in computer vision and image processing fields.

The work of this project concerned with visual attention and perception of video contents in terms of exposure duration, for the scene duration of video directly affects human access to the visual information. Johansson raised questions about the relationship between human perceptual speed and the moving objects [6]. He also studied perceptual organization on stimulus patterns, by simulating human motion as walking, running, etc., with ten moving bright spots, and tested human perceptual speed through different exposure durations of images [7]. Albright et al. made a further research on the visual perception of moving images, and suggested the importance of context factors in visual perception [8]. Chun investigated the interrelationship between context and visual attention, and the mechanism of contextual information learning and its guidance of visual attention deployment, suggesting that context cueing facilitates the efficiency of visual search and recognition [9]. With the booming of digital image technology and creation, the research on visual perception of images has extended in-depth. By investigation of the relationship between spatial frequency and image exposure duration, Watt affirmed the general features of human vision – turning from rough perception to detail perception as exposure duration increases [10-12].

The image perception efficiency of HVS is not only determined by the context, but also related to the familiarity of the theme and visual objects. Thorpe et al. investigated the influence of audience's familiarity with the theme and contents on perceptual efficiency, and showed negative significance of familiarity with the content when faced with complex perceptual objects [13]. However, Bülthoff and Newell's study showed that object recognition is directly related to familiarity with the contents. He claimed that recognition decisions are largely driven by familiarity [14]. Fabre-Thorpe et al. experimentally demonstrated that familiarity has significant effect on perceptual duration of simple images, but this indication is implicit in complex natural images [15]. They concluded that HVS relies on highly automatic feed-forward mechanisms in perceiving highly complex images, which are very little influenced by familiarity. While, categorical representations are crucial for memory load when visual content maintained for a certain duration [16]. This mechanism is also adopted in machine recognition of multimedia [17].

The other factor is the cognitive mode, precisely, the order of the perception process. In most situations, the attentional mechanisms of the visual system tend to coexist in both top-down and bottom-up modes. Posner and Petersen's study of the attentional properties showed that HVS pays attention to visual stimuli consciously when detecting signals for focal objects, and the consciousness is prominent in cognitive accounts of attention [18]. Kastner and Ungerleider investigated the mechanisms of interaction between bottom-up and top-down modalities, and asserted that HVS has selective and active mental activity properties [19, 20]. Bar's study focused on and revealed the role of top-down perceptual mechanisms during low spatial frequency image perception [21]. In addition, the attentional mechanisms of HVS are strongly linked to emotional states to a large extent. Fan et al. confirmed the relationship between image attributes and audience’s emotions by studying subjects' feedback on 31 imagery attributes [22]. Zhang et al. also insists that emotions play important role in video cognition [23].

We consider that these mechanisms and characteristics of HVS lie in the efficiency and accuracy of audience’s perception of video information. If the temporal efficiency of video perception is analyzed quantitatively, the results will be contributive to improving the accuracy and efficiency of information dissemination in video art. This is the reason for conducting the experiment and clarifying the significance of this research work.

3. Experimental Design and Execution

In navigation perspective, there are usually two modes of visual perception: a top-down perception mode driven by subjective consciousness, in which the viewer intentionally intensify attention to certain visual information or actively search for confirmation of the expected information during the viewing process; the other is a bottom-up perceptual mode driven entirely by visual stimuli, in which the audience's perception depends on the stimulus element of the object itself. It usually occurs under conditions in which the viewer has no expectations or contextual preparation for the content and information, and perception is formed gradually from the process of visual stimuli [19, 24]. Top-down attention is mainly governed by consciousness, and visual perception actively seeks the target [25], while bottom-up attention depends primarily on the visual perceptual content. In most viewing experience, audience do not know what specific visual information they are about to access, and can only gradually form cognitive concepts by following the video narration, so it mainly belongs to the bottom-up perceptual mode. The perceptual mode set in this experiment is the bottom-up one, and the subjects participating in this experiment will not be informed of the relevant content before the test so as to maintain their ignorance of the visual information.

3.1. Testing Video and Questionnaire Design

In order to ensure the diversity of video types within the limited testing volume while restore the general viewing experience, we selected six video clips (Figure 1). The six scenes cover camera moving and still shots, camera perspectives ranging from overhead to elevation shots, close up to panoramic shots, with contents of characters, animals and vehicles moving, relatively still scenes with natural landscapes and architectural spaces, etc. (Table 1). Since each shot is prepared for various exposure duration, it is necessary to ensure that the motion of each video is continuously stable so as to avoid significant information fluctuations. These clips are largely stable both in camera motion and target content, and can satisfy the process of arbitrarily cutting part of the motion.

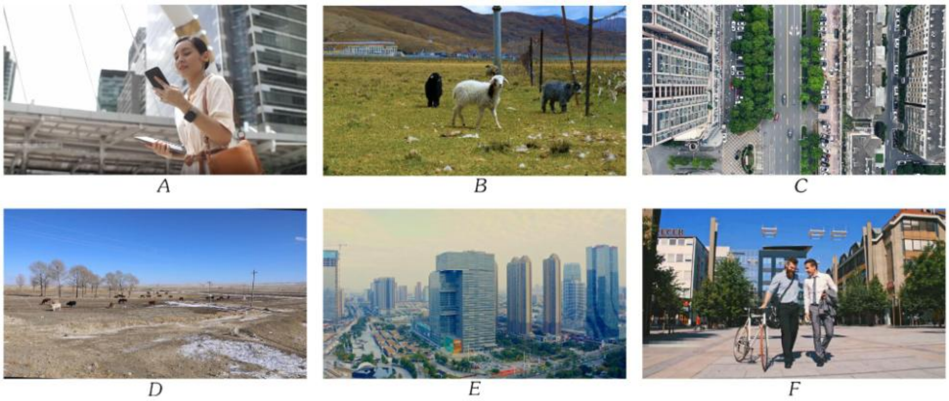

Figure 1. The six clips of testing videos. The clips are intensively chosen according to the testing objects. Clip A and F are provided by https://www.vcg.com, and B, C, D, E from https://www.vjshi.com.

Table 1. Contents and features of the six testing clips.

Clip No.	*Camera Motion*	*Character and Content*	*Action*	*Testing Information*
A	Sideways dolly	city street, Woman, cellphone, tablet, watch, bag	walk, smiling, eyes shifting, holding the tablet	environment, woman, facial expression (smile), dress, hairstyle, bag, belt, cellphone, tablet, watch
B	still	grassland, sheep	sheep moving	grassland, wired fence, mountains, the white sheep, flock of sheep
C	Pilot view	roads, cars, buildings, trees	car running	running cars, roads, trees, buildings, roof view, cars in parking, electric bicycle, pedestrians, traffic signs on the ground
D	still	sky, plain, cattle, trees	almost still	blue sky, barren plain, trees, cattle and sheep, white mark on the ground, white cow foreground, aircraft trails in the sky
E	still	skyscrapers, buildings	cars running	main building, neighboring skyscarapers, buildings in the distance, traffic flow
F	following	street, two men, bicycle	walking nearer, talking	two men, tie, jacket on arm, walking forward, bicycles, briefcase

This experiment is based on materials of the same video clips. Since normal duration of a single clip is generally not less than one second, our test videos are set from one second on, with five gradient samples up to five second long. According to the experimental duration design, each original clip was cut for five samples, and five test videos were synthesized by same durations of every clip. The contents of each test video are the identical, but the segment durations are different, i.e., the one-second testing video is stitched with the six original videos in one second clips, the two-second testing video with the original video in two-second clips, and so on to the five-second-long video. In order to facilitate testing operation, and avoid the montage effect formed by consecutive shots, which would affect the visual perception effect of the subject, three-second-long full-screen in medium-gray (128 of 256) were inserted between each two shots of the synthesized video to make intervals in between. The output size of these videos is 1920*1080 pixels with frame rate of 25p.

In order to evaluate the subjects' perception of the video content, information of the video was listed according to the detail levels, and a questionnaire was designed for the subjects to fill. The questionnaire provided three options for each listed object of the video contents: *definite*, *vague* and *null*. Subjects were asked to check one of the options based on their impressions after watching the testing videos.

3.2. Selection of Subjects

109 participants were invited for this experiment, all of them were university students of 18 to 25. Their academic majors are randomly scattered. In the original stage, 100 of the participants were divided into five groups, gender balanced, and each group watched only one test video of the specific duration, without repeating test for other videos (Figure 1). Some of the subjects did not finish all the test procedures as expected, and a few of them came to the results that were obviously not in line with common sense (possibly due to their reluctance to cooperate with the experiment). Their test data were excluded and other participants was asked to fill the gap. Finally, the number and gender ratio of subjects in each group were ensured to be comparable, with 100 questionnaires, half to half from each gender.

3.3. Experimental Environment and Equipment

The minimum difference in luminance that can be perceived by the human eye at a particular background is the luminance sensitivity level. According to Weber's law, the ratio of luminance sensitivity to background is commonly a constant. In an image of 256 gray levels, human vision tends to be highly sensitive to the areas with a gray level of about 128 [26, 27]. The six videos used in this experiment are all on luminance level of about 128, that guarantee the subjects better perceiving the video information. The brightness contrast of the video images is mainly in the medium. The entire experiment was conducted in a room without strong light interference, and the videos were played on a Dell 27-inch display (U2720Q) running on a PC with Windows OS. The subjects kept eyes about 70 cm away from the display. The video player program was Potplayer.

3.4. Testing Process

The experiment was operated and recorded by an operator. The operator first communicated with the subjects and informed them of the test procedure, but did not reveal any information about the contents of the testing videos. The operator then started playing the video when the subject is ready. The video starts with a 5-second countdown sign, and then the first test scene is presented. The operator paused the video after each shot of the six scenes is played. The subjects were then given a questionnaire about the scene and was informed that the items in the questionnaire might not be presented in the video. The questionnaire was used to confirm what the subjects really perceived, with three options for each item: definite, vague, and not perceived. After finishing the questionnaire, the subject continued to watch the next scene, then the corresponding questionnaire, and all the video scenes were tested in this order. Figure 2 shows the work flow of the experiment.

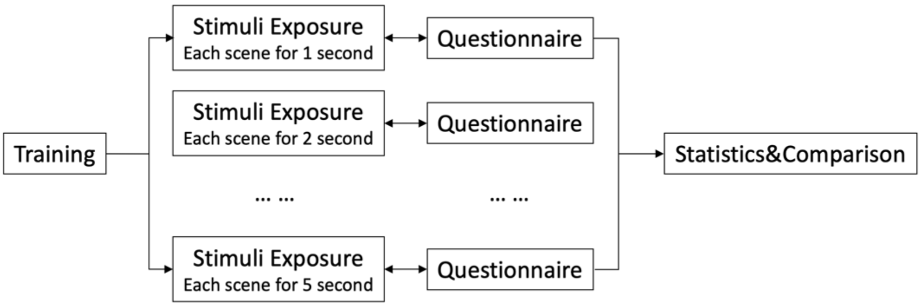

Figure 2. The work flow of the experiment. Each clip was inserted with gray background of three second between every scene. The subjects were asked to finished the questionnaire after watching each scene of the clip.

4. Measures for Data Validity

4.1. Design of the Test Videos

The themes of the six video clips were animals, human characters, buildings and nature spaces. These are common themes and typical types of general videos. In order to focus on the correlation of visual perception and exposure duration, we intentionally typified the video content when selecting the video clips and ensured that the motion trajectories of the main objects in the videos were conservatively stable, and the motion of the character objects was controlled within 1°/sec. This ensures both the typicality of the videos and the stability of the visual information when cutting shots for various durations.

HVS usually do not fully grasp the visual signals entering the human eye, there is but an Internal Generative Mechanism to interpret the input visual signals [28]. HVS will derive and predict the visual contents of the scenes to be recognized mainly based on the memory and experience, while the unintelligible and uncertain information will be discarded. From this perspective, our experiment is about the HVS efficiency of visual information derivation. The efficiency is influenced by human memory. For instance, people are familiar with scenes viewed from a normal viewpoint, as opposed to unfamiliar things, so that people perceive differently in things of familiarities [14]. The overhead perspective is not a common observation angle, so there will be little prior memory information. For this reason, we deliberately chose an aerial video of an overhead view, hoping to access whether the amount of visual experience make differences in visual perception by comparing with other items.

HVS is inevitably influenced by complex factors when perceiving content. Prior contextual information constrains anticipation and visual navigation, facilitating the recognition and search for objects in complex images [29]. To avoid such contextual cues, we inserted a 3-second null screen between each scene to discard consecutive shot combinations that might form extended interpretation of primitive contents, and also to facilitate the pause operations during the experiment.

4.2. Design of the Test Questionnaire

The test questionnaire for this experiment underwent two major versions during the design phase. The original questionnaire included open questions, asking the subjects what they saw in the footage that had just been shown. We found that this approach would result in the subjects missing lots of important information that he actually perceived, and cause uncertainty in the subjects' descriptions that could not be translated into countable data. Thus, we redesigned the questionnaire to list all the content that appear in the videos for the subjects to check off the items that they perceived. However, the results of this test also showed significant inaccuracies, as the subjects appeared irresolute on some of the content and randomly check the items on their uncertainty. It indicates that the results would be seriously affected by the fact that subjects with this level of perception when filling out the questionnaire. Finally, we decided to provide three options for each item: *definite*, *vague*, and *null*. Only those items checked for *definite* were counted as the positive results, so as to filter out some ambiguous perceptions and improve the accuracy of the information perceived with certainty (Table 2).

Table 2. Questionnaire for Scene D of the testing clips (executed in Chinese version). Subjects were asked to check the perceived items according to their impression.

	blue sky and earth	*trees*	*cattle and sheep*	*white mark on the ground*	*white cow in the foreground*	*aircraft trace in the sky*
definite	√		√		√	
vague		√		√		
null (missed)						√

4.3. Plan of the Testing Process

The major challenge of this experiment is ensuring the accuracy of the result data, since visual perception is easily affected by complex factors, such as empirical information, experimental environment, and individual differences. Theoretically, having each subject check all the test videos of various durations can eliminate bias of the individual subjects, but the results would be obviously affected by visual memory and contextual cues when the subjects were repeatedly exposed to the videos of identical contents. Therefore, we decided to have each subject watch every video only once regardless the duration. Even though the test results were still inevitably influenced by individual differences, mental state, and attitudes of the subjects when filling out the questionnaire. In order to reduce these influences, we specifically limited the test subjects to university students, who are of similar age, mental state, cognitive ability and life experience, and therefore have relatively minor individual differences. The test was conducted in their spare time for relax and attentive mood. Nevertheless, many problems were still found in the statistics of the test data. In particular, for one-second duration test, some questionnaires ticked all the listed items as *definite*. As a matter of fact, HVS is unlikely to grab all the information in such a short time. Considering that these questionnaires would weaken the significance of the statistical results, we eliminated the answer sheets with all the items checked as *definite*, and invited more subjects to participate in the questionnaire for the absence.

5. Results and Discussion

The perceptual procedure of human vision usually starts from the overall observation of the imagery, then quickly locates the important targets for in-depth examination. The visual attention actively focuses on the targets of interest, while selectively reducing attention or even ignoring other uninteresting objects. This active and selective observation behavior is also called the *visual attention* mechanism. Imagery in the vision usually contains complex visual information consisting of various objects juxtaposed or superimposed, rather than a single figure, which interact or interfere with each other causing *visual masking effect* as Macknik referred to [30]. Some information is dominant in the interrelations, showing active aggressiveness, while others are subsidiary, negative or weak. Thus, the objects in motion pictures often present a very rich hierarchy of strengths and weaknesses due to visual perception mechanisms. Video information can be divided into three categories according to the visual perceptual strength: the global information, the major objects of attention, and the subsidiary objects. Taking the perception of these three information categories into account, it is assumed that the order of human visual perception is from the global information to the major objects and then expands to the subsidiary objects, where the global and major categories of information are usually clearer and more explicit in the audience’s visual perception, while the subsidiary information is often intentionally weakened by the creator to be vague in the audience’s perception. Therefore, when providing the three questionnaire options of *definite*, *vague* and *null*, the subsidiary information was intentionally filtered out. The target of the survey was focused on the main objects of the video artists’ emphasis. The results of the questionnaires in which the subjects chose the *definite* option remarkablely showed this tendency, i.e., the main information of the scenes were mostly found in the *definite* option (Figure 3).

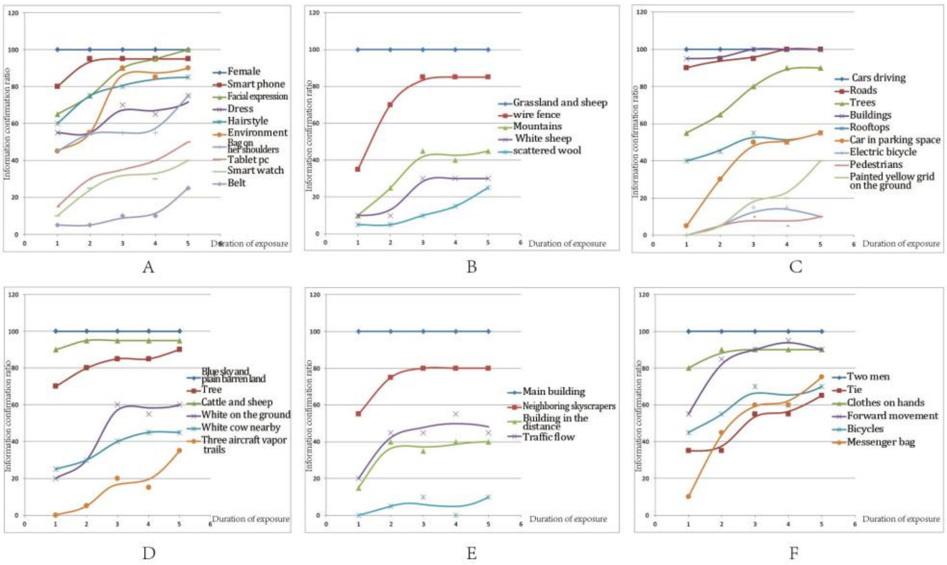

Figure 3. Statistics of the cognitive trends according to the audience questionnaire on the six-scene test. The major the content, the higher the confirmation rate shows; and the longer the exposure duration, the higher the confirmation rate.

The statistic trend shows that as the perceived time increases, more items is confirmed by the subjects (Figure 3). Among them, the global information in the uppermost layer is mostly grasped on one-second duration. The information of main objects is most sensitive to the perceived duration, with some major details in the upper level in the graph and others in the middle level. The perception curves of most major details show significant surge between two and three seconds of exposure duration, and leveling off after three seconds. Among them, the trend curves of scene *E* come to the high stage period significantly early. It is easy to notice, by inspecting the content of the video and the statistic curves, that the content of scene *E* are more concise and has less major details. It is very likely the main reason that supports the rapid achievement of higher perceptual certainty. Therefore, it is the detailed information of main objects that is most obviously affected by perceived duration. As the shot duration increases, the perceptual certainty of the characters and primary details increases significantly. The perceptual certainty of the subsidiary details also increases along the increase of perceived duration, but with a certain lag comparing to the characters and primary information, and the increase is relatively smoother. They are mainly unimportant details with weak attraction, many of them are even completely ignored by the subjects. However, most information of these items does not affect the subject's understanding of the video. In general, the statistical results of the test data seem in consonance with the visual attention mechanism, and reveal the perceptual order and attention distribution of the three levels of visual information.

In this experiment, the subjects were divided into two identical groups of both genders, in order to eliminate gender bias, and also for convenience of inspecting the gender differences. Judging from the features of the questionnaire data, the gender difference is significantly reflected in the test of scene *A* (Figure 4). The statistical results of the questionnaire on scene *A* show that female subjects pay more attention to the character's dress, bag, hair style and belt in the video than the male group, and the intensity and certainty of perception increases earlier. Whereas male subjects tend to be much acute on electronic items as the tablet, cellphone and smartwatch. There seems to be a lack of attention to the women character's belt. For scene *D*, gender differences had less impact on the perception of global information, and both male and female subjects were able to identify the grass and sheep in the picture in just one second. The barbed wire fence belongs to the second category of information, i.e., the major object, which was fully perceived with perception up to three seconds, but the subsidiary information did not significantly increase in perceptual intensity within all the testing durations (Figure 5). In general, gender differences do exist in visual perception and are likely to be influenced primarily by interest and culture.

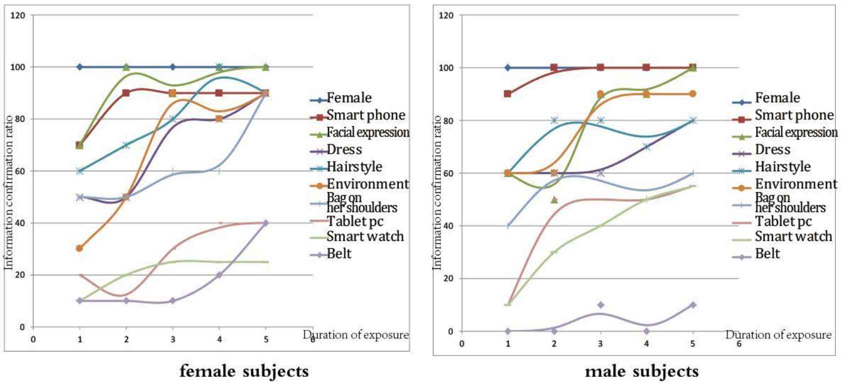

Figure 4. Gender difference trend according questionnaire on scene A.

An interesting phenomenon appeared that most male subjects showed a significant surge of awareness of the frost on the ground in scene *D* at the third second. There were also perceptual differences on the white cow, with more males than females definitely grasped it in one second, but this advantage did not increase as the exposure duration increased, while the number of females who positively checked this information increased significantly. This may imply that there are differences in the order of visual perception, not only between genders, but also between individuals of the same gender.

Besides, there shows no significant difference, merely by the results of the questionnaire, in the video duration of both normal viewpoints and unconventional perspective, such as aerial footage.

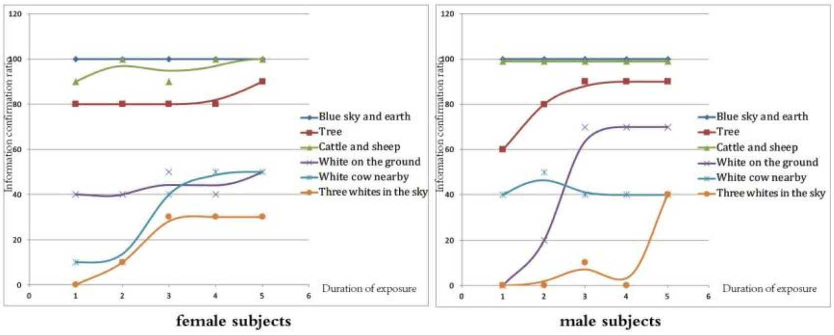

Figure 5. Gender difference trend according questionnaire on scene D.

6. Conclusion

The experimental results are basically consistent with our initial hypothesis, but the specific data fluctuate according to the video content. It generally shows that three seconds is a very important threshold. HVS can perceive the overall contextual information and the main character-objects in videos within exposure duration of three seconds. In fact, once these two types of information are mastered, the content and intention of the video can be basically understood. This indicates that audience can grasp the main information of video shot from three seconds up. However, this perceived duration fluctuates due to the information density of the footage. In the case of relatively low information density, the shot duration can be as short as two seconds, but in most cases segment duration of three seconds or longer appears to be more secure. Gender differences also affect the perception of certain types of information, such differences depend mainly on the audience's interests rather than on the visual perception mechanism per se.

Although the experimental results are generally consistent with our hypothesis, there are still shortcomings that affect the accuracy of the experimental results. First, the greatest impact on the accuracy of the experimental results might be the subjective factors of the questionnaire. Although we screened the received questionnaires, it is still hard to say the retained questionnaires were sufficiently objective. The other uncertainty is that the testing process was conducted simultaneously by two testing operators. Although we trained the operators before the test and controlled the testing process equally, the collaboration differences between the two operators and subjects were still unavoidable and could affect the test results. In addition, this experiment only selected six typical videos for testing, it might be inadequate to cover the miscellaneous and complex perceptual cases. Plenty of factors on video perception needs to be investigated further

Acknowledgment. This document is the results of research project supported by The National Social Science Fund of China [20ZD09] and China Academy of Art [ZD202205].

Data Availability. The raw data supporting the conclusions of this article will be made available by the authors by email or website.

References

1. Gkalelis, N., Goulas, A., Galanopoulos, D., Mezaris, V. Object Graphs: Using Objects and a Graph Convolutional Network for the Bottom-up Recognition and Explanation of Events in Video. Computer Vision and Pattern Recognition, IEEE. (2021)
2. Schwenzow, J., Hartmann, J., Schikowsky, A., Heitmann, M. Understanding videos at scale: How to extract insights for business research. Journal of Business Research 123:367-379. (2021)
3. Zhang, J., Yu, X., Lei, X., Wu, C. A Novel Deep LeNet-5 Convolutional Neural Network Model for Image Recognition. Computer Science and Information Systems 19(3):1463-1480. (2022)
4. Conrad, M., Cin, M.D., Marr, D. Approaches to biological information processing. Science 190:875-876. (1975)
5. Marr, D. Early processing of visual information, Philosophical Transactions of the Royal Society of London. Biological Sciences 275(942):483-519. (1976)
6. Johansson, G. Visual perception of biological motion and a model for its analysis. Atten. Percept. Psychophys 14:201-211. (1973)
7. Johansson, G. Spatio-temporal differentiation and integration in visual motion perception. Psychological Research 38(4):379-393. (1976)
8. Albright, T.D., Stoner, G.R. Visual motion perception. Proceedings of the National Academy of Sciences of the United States of America 92(7):2433-2440. (1995)
9. Chun, M.M. Contextual cueing of visual attention. Trends Cognit Sci, 4(5):0-178. (2000)
10. Watt, R.J. Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus. Journal of the Optical Society of America A-optics Image Science & Vision 4(10):2006-2021. (1987)
11. Bicanski A., Burgess, N. A Computational Model of Visual Recognition Memory via Grid Cells. Current Biology 29(6):979-990. (2019)
12. Rybak, I.A., Golovan, A.V., Gusakova, V.I. Behavioral model of visual perception and recognition. Proceedings of SPIE - The International Society for Optical Engineering 1913:548-560. (1993)
13. Thorpe, S., Fize, D., Marlot, C. Speed of processing in the human visual system. Nature 381:520-522. (1996)
14. uBülthoff, I., Newell, F.N. The role of familiarity in the recognition of static and dynamic objects. Progress in Brain Research 154:315-325. (2006)
15. Fabre-Thorpe, M., Delorme, A., Marlot, C., Thorpe, S. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. Journal of Cognitive Neuroscience 13(2):171-180. (2001)
16. Zhou, C., Lorist, M.M., Mathôt, S. Categorical bias as a crucial parameter in visual working memory: The effect of memory load and retention interval. Cortex 154:311-321. (2022)
17. Zafar, B., Ashraf, R., Ali, N., Ahmed, M., Jabbar, S., Naseer, K., Ahmad, A., Jeon, G. Intelligent Image Classification-Based on Spatial Weighted Histograms of Concentric Circles. Computer Science and Information Systerms 15(3):615-633. (2018)
18. Posner, M.I., Petersen, S.E. The attention system of the human brain. Annual Review of Neuroscience 13(1):25-42. (1990)
19. Kastner, S., Ungerleider, L.G. Mechanisms of visual attention in the human cortex. Annual review of neuroscience 23:315-341. (2000)

20. Intraub, H. The representation of visual scenes. Trends in Cognitive Sciences 1(6):0-222. (1997)
21. Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmidt, A.M., Dale, A.M., Hamalainen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., Halgren, E. Top-down facilitation of visual recognition. Proc Natl Acad Sci USA 103(2):449-54. (2006)
22. Fan, S., Koenig, B.L., Zhao, Q., Kankanhalli, M.S. A Deeper Look at Human Visual Perception of Images. SN Computer Science 1(1):58. (2020)
23. Zhang, J., Wen, X., Whang, M. Recognition of Emotion According to the Physical Elements of the Video. Sensors 20(3):648. (2020)
24. Privitera, C.M., Stark, L.W. Algorithms for defining visual region-of-interesting: comparison with eye fixations. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(9):970-982. (2000)
25. Baluch, F., Itti, L. Mechanisms of top-down attention. Trends in Neurosciences 34(4):210-224. (2011)
26. Netravali, A.N., Haskell, B.G. Digital Pictures: Representation and Compression. New York: Plenum (1988)
27. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4):600-612. (2004)
28. Peter, R.J., Iyer, A., Koch, C., Itti, L. Components of bottom-up gaze allocation in natural scenes. J. Vison 5(8):692-692. (2005)
29. Joubert, O.R., Rousselet, G.A., Fize, D., Fabre-Thorpe, M. Processing scene context: fast categorization and object interference. Vision Research 47:3286-3297. (2007)
30. Macknik, S.L., Livingstone, M.S. Neuronal correlates of visibility and invisibility in the primate visual system. Nature Neuroscience 1(2):144-149. (1998)

Jianping Song, lecturer of digital media art at Zhejiang Agriculture and Forestry University, is mainly engaged in teaching and research in the field of digital graphics and videos. His interests are in the human visual perception of mages, especially from the perspective of video creation and artificial intelligence.

Tianran Tang, PhD of Design, associate Professor, is the deputy dean of Artificial Intelligence College of Dongguan Polytechnic. His main interest is in teaching and research of digital art and animation design, including the promotion of digital art and animation brands.

Guosheng Hu, MA in art and design, PhD in computer technology, is an associate professor of China Academy of Art. His work and research interests are related to computer graphics, color design, and multimedia & interaction art.

Received: September 19, 2022; Accepted: January 16, 2023.