applied sciences

MDPI

Article

Method for Viewing Real-World Scenes while Recording Video

Daehee Park * and Cheoljun Lee

Samsung Research, Seoul 06765, Korea; cheoljun.lee@samsung.com
* Correspondence: daehee0.park@samsung.com

Abstract: Because smartphones support various functions, they are carried by users everywhere. Whenever a user believes that a moment is interesting, important, or meaningful to them, they can record a video to preserve such memories. The main problem with video recording an important moment is the fact that the user needs to look at the scene through the mobile phone screen rather than seeing the actual real-world event. This occurs owing to uncertainty the user might feel when recording the video. For example, the user might not be sure if the recording is of high-quality and might worry about missing the target object. To overcome this, we developed a new camera application that utilizes two main algorithms, the minimum output sum of squared error and the histograms of oriented gradient algorithms, to track the target object and recognize the direction of the user's head. We assumed that the functions of the new camera application can solve the user's anxiety while recording a video. To test the effectiveness of the proposed application, we conducted a case study and measured the emotional responses of users and the error rates based on a comparison with the use of a regular camera application. The results indicate that the new camera application induces greater feelings of pleasure, excitement, and independence than a regular camera application. Furthermore, it effectively reduces the error rates during video recording.

Keywords: smart cameras; HCI; affective design

Citation: Park, D.; Lee, C. Method for Viewing Real-World Scenes while Recording Video. *Appl. Sci.* **2021**, *11*, 4617. https://doi.org/10.3390/app11104617

Academic Editor: João Carlos de Oliveira Matias

Received: 27 April 2021
Accepted: 16 May 2021
Published: 18 May 2021

1. Introduction

Since the first versions of Apple's iPhone and Samsung's Galaxy S series smartphones were first released globally, they have become extremely popular and important devices in daily life [1]. In 2009, smartphone penetration in the US was 25%, and 14% of the mobile phones shipped worldwide were smartphones [2]. Furthermore, the number of smartphones sold to end users worldwide has increased dramatically from 2007 to 2020, expanding from 122 million to 1.56 billion devices sold [3]. Smartphone technologies have been evolving through several device generations, and cameras have increased in importance and are currently regarded as a marketing tool to attract customers [4]. Because smartphones can achieve various functions, people use them frequently and carry them almost everywhere. One of the important functions of a smartphone is photography. People use their smartphones whenever they want to capture a moment that has an important meaning to them. Thus, relevant camera technologies have evolved to record everyday moments with high quality [2]. However, one significant problem when recording video of an important event is the fact that the user generally views the scene recorded through the mobile phone screen rather than the actual real life event itself (see Figure 1).

For example, when recording a dance performance, the user will likely focus on the mobile phone screen rather than the actual live dancing owing to concerns with the video recording. This occurs because the system cannot identify the user's intention when recording a video. As a result, when recording, the user should focus on the mobile phone screen, and the user cannot perceive the real world directly. In addition, at that time, the user may have an unpleasant or out-of-control feeling in terms of affection because the user cannot see the real-world event and cannot be confident of the quality of the recording.

Figure 1. People view events through the mobile phone screen rather than the real-world event itself.

Through a focus group discussion (FGD), we investigated the reasons why people feel compelled to concentrate on the mobile phone screen when recording a video. The FGD results indicate that there are several reasons why people focus on the scenes displayed on the mobile phone screen rather than the actual event itself. First, users are concerned regarding whether the target object will be properly recorded. Second, the users cannot be sure whether the video recording contains what they are seeing in real life. If users do not see the mobile phone screen while recording a video, the result of recording sometimes cannot involve the exact scenes the user wants and the quality of recording cannot meet the user's standard. Third, they cannot be sure if the video recording is going well. Focusing on recording the video results in a reduced level of emotion because the user cannot focus on the actual event, and there has been a lack of studies dealing with this issue.

The technical and practical problems discussed here emerged from the experience of developing and deploying video recording capabilities as a part of research conducted through Samsung's Camera Improvement Project. In this study, we propose new camera functions that support object tracking and head movement recognition. This helps record target objects while tracking them and can record scenes based on the user's head movements.

Our aim is to help the camera user focus on the real-life event instead of viewing it through the screen. In the aspect of user experience, this will create a better emotional feeling for the user during the video recording. Thus, we proposed a new method to see the real-world scenes while a video recording, which is our main contribution. In addition, we contributed that we measured emotional responses after video recording to verify the effectiveness of the proposed method in the aspect of user experience. In the following sections of this paper, we describe a method proposed to solve the above-mentioned problems. We then evaluate the new system to prove the effectiveness of the proposed application. Although the proposed approach improves the user's positive emotion by allowing the real-world event to be viewed directly while recording a video, further improvements of the application are required.

2. Related Studies

2.1. Object Tracking Technology: Minimum Output Sum of Squared Error (MOSSE)

Visual tracking has received a significant amount of attention in recent years. Visual tracking technology can be regarded as important and has many practical applications in video processing [5]. Bolme et al. insisted that if a target is located in one frame of a video, it is recommended to track that object in subsequent frames [5]. If the target is successfully tracked, it provides more information about the activity of the target. Because tracking is easier than detection, tracking algorithms require fewer computational resources than an object detector. Many types of tracking technologies have been suggested that can check changes in the target appearance and track the complex motions of the target, including

incremental visual tracking [6], robust fragment-based tracking [7], graph-based discriminative learning [8], and multiple instance learning [9]. These tracking technologies are effective [10]; however, they are not easy to apply and involve complex appearance models and optimization algorithms. In addition, it is difficult to maintain the 25–30 fps produced by many modern cameras. To address the above issues, Bolme et al. proposed a new type of correlation filter (CF) called MOSSE [5]. This technology, when initialized, produces stable correlation filters using a single frame. The research of Bolme et al. showed that simpler MOSSE CFs can solve visual tracking problems more effectively than heavy weight classifiers, complex appearance models, and stochastic search techniques [5]. MOSSE provides an effective algorithm that is accurate, easy to implement, and much faster than other technologies. The advantage of a MOSSE filter is its robustness when dealing with variations in lighting, scale, pose, and non-rigid deformations during processing at 669 fps. In addition, it supports tracker pausing when the object disappears and resumes where it left off when the object reappears [5].

2.2. Face Detection Technology: Histograms of Oriented Gradients (HOG)

Object detection is considered an important step in high-level computer vision [11]. Pang et al. insisted that object detection is an essential function for video analysis and image comprehension [11]. Among a variety of objects, human faces and bodies are regarded as the most salient objects in images and videos [11]. Extensive research has indicated that detecting humans in videos is difficult because of the diversity in appearance, illumination, and background [12–14]. In the computer vision area, HOG combined with a support vector machine (SVM) (HOG + SVM) is well-known and has been regarded as the most successful human detection algorithm [15]. However, it has the disadvantage of being time-consuming [11], which Pang et al. suggested can be solved in two ways: the first is to reuse features in blocks to build the HOG features for intersecting the detection windows [11], and the second is to utilize sub-cell-based interpolation to efficiently calculate the HOG features for each block [11]. Pang et al. indicated that the combination of the two approaches results in a significant increase in human detection [11]. The results indicate that the combination of both approaches computes five times better than the HOG + SVM [11].

3. Methods

To view a real-world event rather than the mobile phone screen while video recording, it is necessary to calculate the direction of the user's face. The camera should record the scene that the user is focused on. In addition, the camera should track the target object continuously while video recording. To fulfill the above requirements, we adopted two methods to obtain and capture information about the area of interest while recording the video. First, we used the front camera to calculate the user face's direction. Second, we tracked the object using a rear camera. For the simulation, we used a Samsung Galaxy S20 Ultra phone and built our own simulator application operated on the Android system.

3.1. Object Tracking

To track a target object during a video recording, we reviewed various algorithms in OpenCV [16]. Grabner, Grabner, and Bischof proposed an object tracking algorithm in 2006 called on-line AdaBoost [17]. Object tracking algorithms have subsequently been developed and evolved over the past several years. For example, Babenko, Yang, and Belongie suggested a new algorithm called multiple instance learning (MIL), which showed a more robust performance in real-time than on-line AdaBoost [9]. Kalal, Mikolajczyk, and Matas [18] proposed a new method for tracking a failure detection called a median flow, which provides a better performance than MIL, despite certain drawbacks [19]. Varfolomieiev and Lysenko insisted that a median flow is not a suitable algorithm in the context of embedded systems [19]. It should be noted that we often have to consider the trade-offs among reliability, operational speed, power consumption, size, mass, and/or cost of the entire system [19]. Based on extensive considerations regarding reliability,

operational speed, and power consumption, we chose MOSSE, which, although it lacks accuracy, supports high-speed processing, making it suitable for an embedded system. In addition, it makes it possible to quickly find and track a missing target object.

The operation of the MOSSE algorithm can be described as follows: first, users utilize their smartphones to record moving objects. At this time, the user records a video on a widescreen at a distance from the target object. The camera then recognizes the direction of the user's face in real time, identifies moving objects in the recognized direction, and saves the video of the selected areas around them (Figure 2).

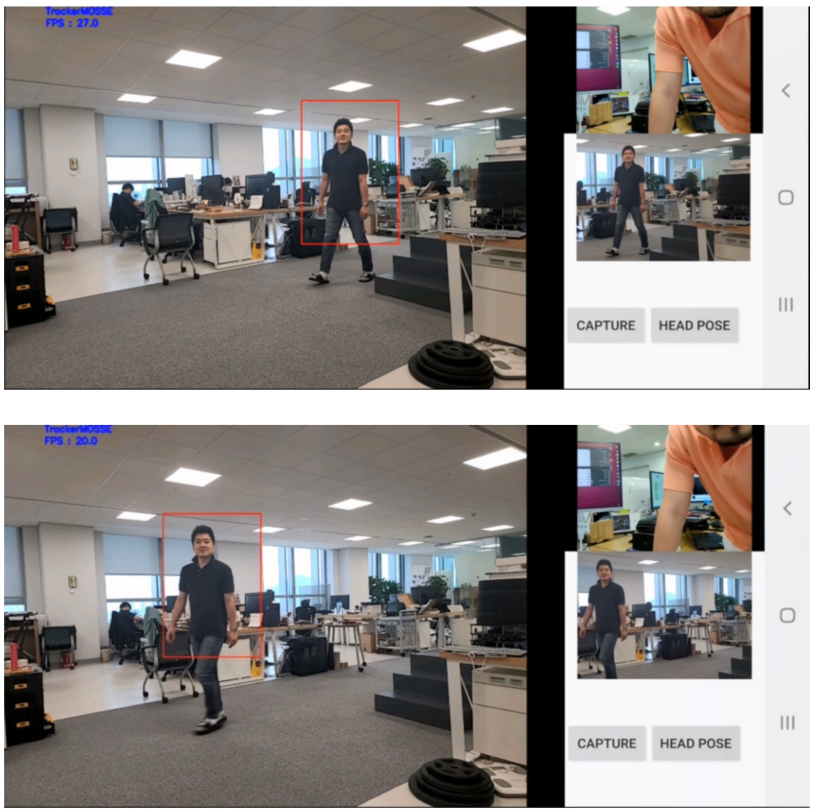
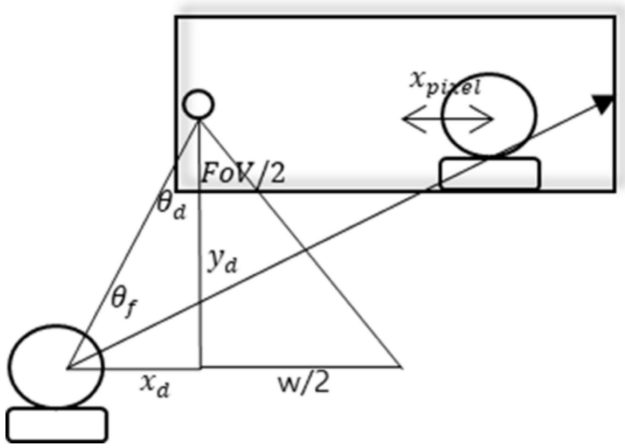


Figure 2. Working example of MOOSE algorithm.

3.2. Camera Follows the Direction of the User's Face

To calculate the direction of the user's face, we adopted the HOG, which uses a face detection algorithm in Dlib [20]. HOG provides a significantly increased performance in human detection compared with traditional methods [8]. During the camera detection of the user's face, it recognizes 68 face characteristics. In addition, it utilizes the OpenCV library to recognize the direction of the face using the front camera [16]. It then converts 2D vision data into 3D data. To account for this formula, we defined the location of the user's face and the distance from the user's face to the mobile phone (Figure 3). The calculation results of these two variables are subsequently utilized to calculate the direction of the object. To obtain the direction of the object, we can calculate θ2, as shown in Figure 4. At this stage, Holzmann and Hochgatterer indicated that we can calculate the value of y2 using single-camera stereo vision while recording a video [21].

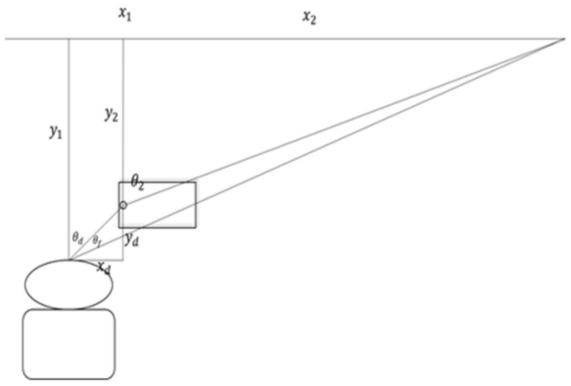


w: length of left most to right most in FoV
x_d: position of videographer from center
y_d: distance between device and videographer (usually 25cm)

$$\tan\left(\frac{FoV}{2}\right) = \frac{\frac{w}{2}}{y_d}, \mathrm{w} = 2 \times y_d \times \tan\frac{FoV}{2}$$

$$x_d = \frac{x_{pixel}}{width_{pixel}} \times w$$

Figure 3. Formula for calculating the location of the face and distance from the mobile phone to the face.



$$\theta_1 = \theta_f + \theta_d$$

$$x_1 = x_2 + x_d$$

$$y_1 = y_2 + y_d$$

$$\tan(\theta_1) = \tan(\theta_f + \theta_d) = \frac{x_2 + x_d}{y_2 + y_d}$$

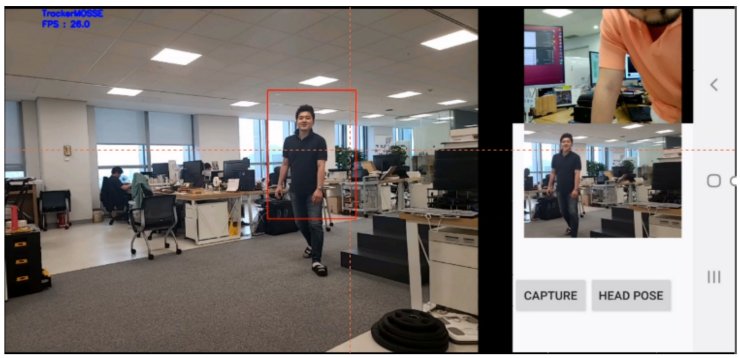
$$x_2 = \tan(\theta_f + \theta_d) \times (y_2 + y_d) - x_d$$

$$\theta_2 \approx \theta_1 = \tan^{-1}\frac{x_2}{y_2} = \tan^{-1}\left(\frac{\tan(\theta_f + \theta_d) \times (y_2 + y_d) - x_d}{y_2}\right)$$

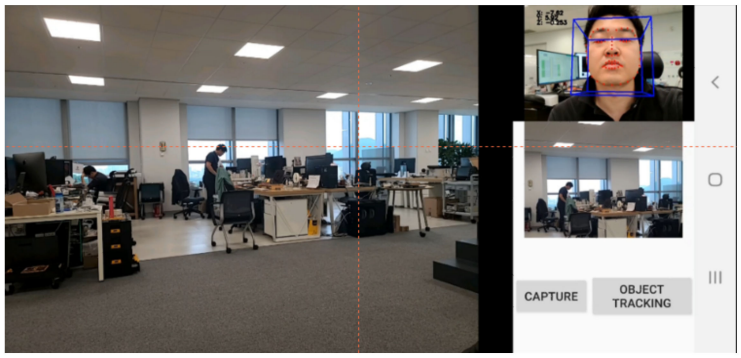
Figure 4. Formula for calculating the direction of the object.

3.3. Application Flow

We developed an application that employs object tracking and user face recognition technologies (Figures 5 and 6). The application has two different modes. First, the user checks the target object, and the camera then automatically tracks that object. This mode helps the user feel confident of not missing the target object while recording the video. Second, the user can choose the mode that identifies the direction of the head through face detection. Hence, the angle of the camera moves automatically according to the direction of the user's head. This helps the user become more immersed in the real world because the camera detects the direction of the head and records the user's view.



(**a**)



(**b**)

Figure 5. New camera application: (**a**) first application mode, object tracking, and (**b**) second application mode, face detection.

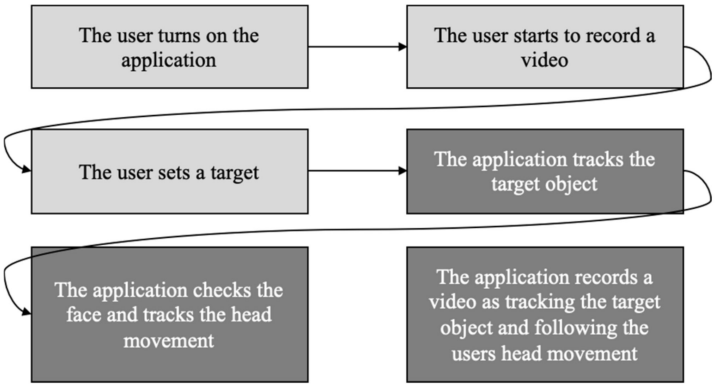


Figure 6. Application flow.

4. Case Study

4.1. Hypothesis

As mentioned previously, the users generally record a video by looking at the mobile phone screen to avoid missing the target object. The user may feel anxious about missing the target object and lack confidence that the recording is progressing well unless viewing the recording screen directly. Thus, we hypothesized that, by viewing the mobile phone screen rather than the real-world scene, the users may have a low feeling of valence, arousal, or dominance when recording a video. Thus, we compared two systems, including our application: viewing the real-world scene versus viewing the screen of a regular mobile phone camera. When comparing the two systems, we measured the affection scores and compared them. In addition, we hypothesized that seeing a mobile phone camera screen has disadvantages in object detection and tracking because the angle of the human eyes is narrow when looking at the mobile phone screen when recording.

4.2. Case Study Design

We recruited 20 participants (16 males and 4 females) from among Samsung Electronics employees who were not otherwise involved in this project (mean age = 38.5 years). Parkkola and Saariluoma asserted that 8–10 participants are sufficient to generate the majority of action types [22]. We subsequently conducted a case study to measure the affection of the user and the error rate during the object detection and tracking between two situations: recording the video while looking at the mobile phone screen and recording the video while viewing the real-world event. First, in regard to measuring the affection scores, it is not easy to express affections verbally while using a product [23]. Therefore, Desmet asserted that a tool for measuring nonverbal emotions is needed to measure the affective responses in a satisfactory manner [24]. The author outlined [24] the advantages and disadvantages of five methods that can measure emotions: self-assessment manikin (SAM), emocards, expressing experiences and emotions (3E), feedback application, and experience clip. Among these affection measurement methods, SAM is an effective method for measuring subjective emotions using three axes: pleasure, arousal, and dominance [2]. Through the SAM approach, it is possible to collect quantitative data with a simple image of the three axes. Therefore, we used SAM as a tool to collect quantitative affective data to compare viewing the camera screen with viewing the real-world event. We also measured the object detection and tracking error rates to compare the two situations.

4.3. Metrics

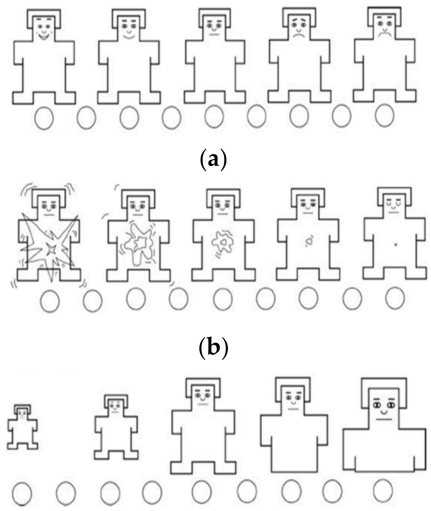
4.3.1. SAM

SAM was initially developed by Lang [25] to suggest a solution to the problems that correspond to measuring emotional responses to advertising. SAM consists of pleasure, arousal, and dominance (PAD) dimensions. It was designed as an alternative to the sometimes cumbersome verbal self-reporting measures [25]. SAM measures each PAD dimension with a graphical character on a nine-point scale [26] (Figure 7). We measured the emotional responses after recording a video because we hypothesized that there is a statistically significant difference between seeing the mobile phone screen and seeing the real-world scene.

4.3.2. Measuring Error Rates

We measured error rates during object detection and tracking using two different materials, as shown in Figure 8. We used a TV as a medium to present the object virtually. For the purpose of object tracking, a target image shaped like a child moved quickly across the TV screen for 15 s. For the material speed, we adopted and used the moving speed of Microsoft PowerPoint's animation function. The project was started under the assumption that the user might miss the target object when recording while looking at the mobile phone screen; hence, the error rate was measured during object tracking. In addition, we measured the object detection error rates. During the object tracking session, the participants were

asked to track the target object as soon as possible. For object detection, we designed a simulated fireworks environment, and the target object appeared randomly for 1.5 s and then disappeared. During object detection, the participants were asked to find the location of the target object and record it.

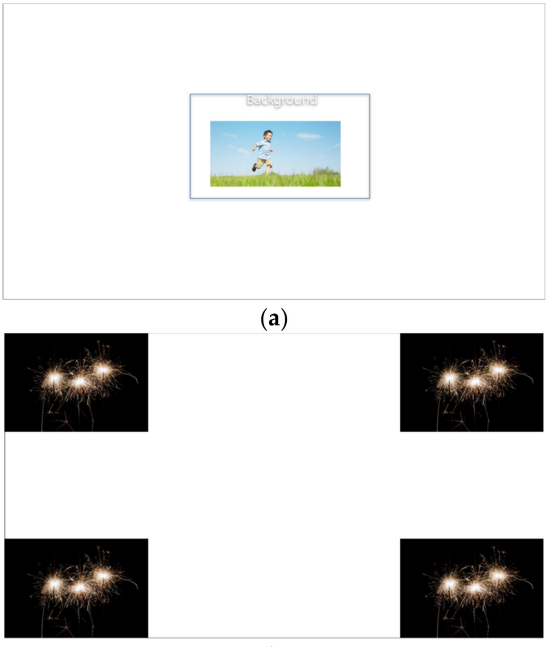


(**a**)

(**b**)

(**c**)

Figure 7. Self-assessment manikin (SAM) [26]: (**a**) valence, (**b**) arousal (**c**) dominance.



(**a**)

(**b**)

Figure 8. Materials used for measuring error rates: (**a**) object tracking, (**b**) object detection.

4.3.3. Case Study Procedure

Figure 9 shows the case study procedure. At the instruction stage, the participants were given a description of the purpose and procedures of the case study. They were then asked to follow the child's movements and record a video for 10 s using two types of mobile phone: a regular mobile phone requiring the participants to look at the phone screen during video recording; and a mobile phone equipped with the new camera application, where the participants were asked to look at the real-world event while recording the video.

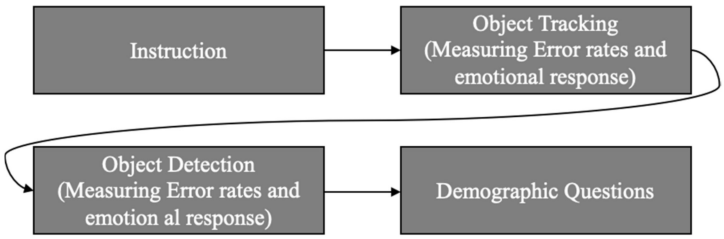


Figure 9. Case study procedure.

Subsequently, the researchers asked the participants about their emotional responses when using the mobile phone while recording the video. The participants were then asked to conduct an object detection using a regular mobile phone and a mobile phone with the new camera application. Four fireworks appeared successively for 1.5 s in the corner of the screen, and the participants were instructed to find and record them. To avoid the order having an effect on the outcome, the order of the videos was randomized for both approaches: looking at the mobile phone screen and looking at the real-world event. Finally, all participants were asked about their age and gender, and then dismissed.

5. Results

The results of the case study were analyzed from two perspectives: emotional responses and error rates (Figures 10–12).

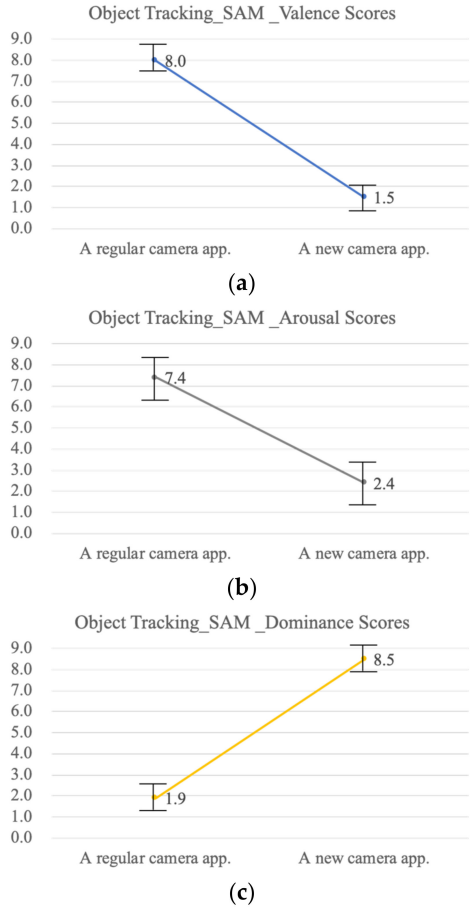


Figure 10. SAM result during object tracking: (**a**) valence score, (**b**) arousal score, (**c**) dominance score.

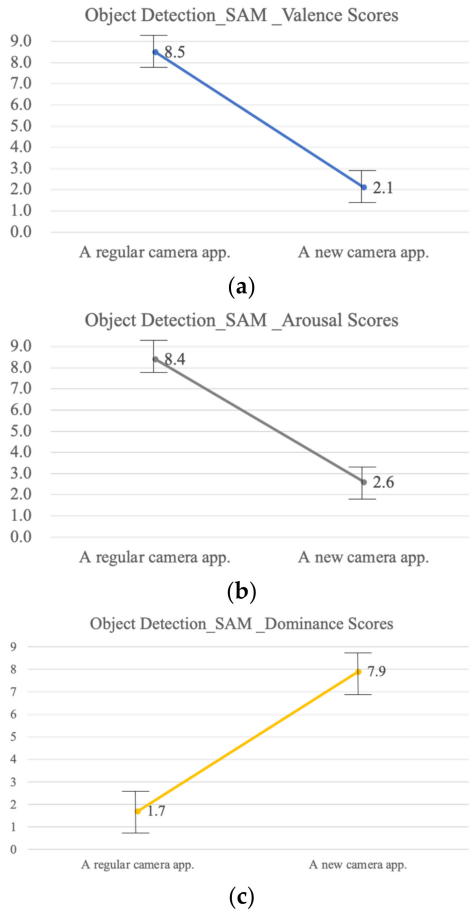


Figure 11. SAM result during object detection: (**a**) valence score, (**b**) arousal score, (**c**) dominance score.

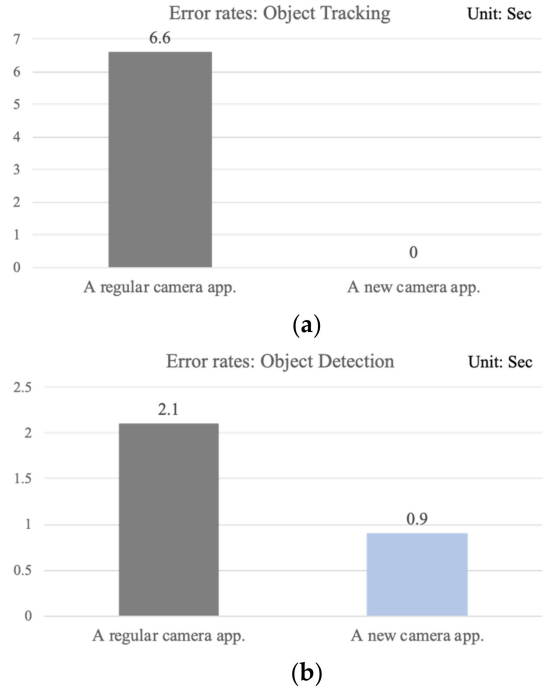


Figure 12. Error rate results during object detection: (**a**) object tracking error rates, (**b**) object detection error rates.

Emotional responses: We measured the emotional responses when conducting two different experiments (object tracking and object detection) and compared them under two conditions (using a regular camera app and using the new camera app). As described in Figure 7, each emotional dimension consists of a nine-point scale [27]. On the valence scale, the closer the participant response was to 1, the more pleasant the participant felt. In addition, the closer the participant response was to 9 on the scale, the more unpleasant the participant felt. On the arousal score scale, the closer the participant response is to 1, the more excitement the participant expressed. By contrast, the closer the participant response is to 9 on the scale, the more calm the participant felt. By contrast, on the dominance score scale, the points are reversed. The closer the participant response is to 1, the more dependency the participant expressed. In addition, the closer the participant's response is to 9, the more independent the participant felt. We statistically analyzed the SAM results. First, for the object tracking valence score, the mean score when using the new camera application was 1.5, as compared to a mean score of 8.0 for the regular camera application ($F = 304.20$; $p < 0.01$). The object tracking arousal scores were 2.4 when using the new camera application, as compared with a mean score of 7.4 when using a regular camera application ($F = 52.57$; $p < 0.01$). For the dominance score, the mean when using a new camera application was 8.5, compared to a mean of 1.9 when using a regular camera application ($F = 292.57$; $p < 0.01$). The object tracking results indicate that the participants felt a sense of pleasance, excitement, and independence when using the new camera application.

During the object detection experiment, for the arousal dimension, the mean score when using the new camera application was 2.1, as compared to a mean score of 8.5 when using a regular camera application ($F = 157.54$; $p < 0.01$). For the arousal dimension, the mean score when using the new camera application was 2.6, as compared to a mean score of 8.4 when using a regular camera application ($F = 82.27$; $p < 0.01$). For the dominance dimension, the mean score when using the new camera application was 2.6, compared to a mean score of 8.4 when using a regular camera application ($F = 82.27$; $p < 0.01$).

Error rates: For object tracking, the total duration of the child's movement was 10 s. We measured the time up to which the participants missed the target object during video recording. In addition, for object detection, four firework images appeared, and we measured the number of missed target objects during the video recording. The results indicate that the mean amount of time the target object was missed when using a regular camera application was 6.6, compared to a mean time of 0 when using the new camera application ($F = 272.25$; $p < 0.01$). It turns out that it is extremely difficult to track the movement of the target object. Obviously, the new camera application automatically tracks the target object, and hence the missed time is 0. It can be assumed that the result is associated with the SAM result. The participants did not feel pleasure, excitement, or independence when using a regular camera application because they could not track the target object well. The object detection error rate indicates that the mean number of missed target objects using a regular camera application was 2.1, compared to the 0.9 mean number of missed targets for the new camera application ($F = 4.35$; $p = 0.05$).

6. Discussion

Although many people record a video to memorize important moments, it is difficult to see the entire real-world scene because they need to focus on the recording. We hypothesized that people produce negative emotional responses when seeing the scene through the mobile phone screen and concentrate on recording rather than fully enjoying the moments. In general, the results indicate that there is a significant difference in emotional responses between using a regular camera application and using the new camera application that applied our novel technology.

Object Tracking: Most participants felt greater pleasure, more excitement, and more independence when using the new camera application. This is associated with the error rate results. In object tracking, the error rates of using a regular camera application were significantly higher than those of using the new camera application. Recording a video

using a regular camera application led the participants to focus on the video recording itself rather than on watching the actual scene, and induced them to feel negative emotions. By contrast, while video recording using the new camera application, the participants saw the entire real-world scene and were confident that the new camera application tracked the target object automatically, which induced more pleasurable, exciting, and independent emotional responses for the participants.

Object Detection: The error rate results indicate that the error rate of using the regular camera application was higher than that of using the new camera application, although the difference was not statistically significant. In object detection, the participants felt more pleasure, excitement, and independence when using the new camera application than when using the regular camera application. Using the new camera application helped the participants see the entire real-world scene. This indicates that the participants can use a wider field of view when using the new camera application, which induced emotional responses of greater pleasure, excitement, and independence. When the participants used the regular camera application, they felt more negative emotions. It was assumed that generating negative emotions from looking at the mobile phone screen might further shrink the field of view. The research of Que et al. asserted that the field of view can be shrunk by negative emotions [28]. Thus, although there was no significant effect in object detection between the two situations, there was a significant difference in the emotional response.

Limitations and Future Studies: Although our new camera application can help people see the real-world scene rather than through the mobile phone screen, there are some limitations and a need remains for additional future studies. First, it is not necessary to look at the mobile phone screen during video recording; however, the users must hold the mobile phone at the same height as when using a regular camera app. The camera requires a field-of-view expansion, especially in the up and down areas. In this research, we could not consider the method to correct the user's posture when using the new video recording application even if their posture is wrong. Future works should include the study of a user's posture whilst holding a mobile phone during a video recording. Second, it is difficult to recognize the user's intentions indoors based only on the movement of the head pose. We believe that the movement of the head pose might be acceptable in object detection in outdoor situations such as concerts or musicals. There are some limitations in this study. First, our case study was limited to an indoor experimental laboratory. The indoor experiment indicated that there was no significant difference in object detection between conditions. This is because the field of view was narrow in the indoor activity. Second, the algorithms applied in the application should be improved. The MOSSE algorithm is related to the filtering algorithm, and it cannot support a long-term scene. Hence, we believe that eye-tracking technology can be applied in our new camera application to avoid the above issues. In addition, the free-viewpoint video can be applied in future work to improve the quality of recording a video [29]. This is because eye-tracking technology and the free-viewpoint video help the system to recognize the users' intention—where they want to record. It can be assumed that it can supplement the weakness of movement of the head pose and can easily detect the user's intention. In addition, the improved algorithms for object tracking suggested by the studies of Wang et al. [30] and Voigtlaender et al. [31] can be applied in the application, and we expect those algorithms to increase the velocity and stability of the application. Third, an "egocentric video" can be applied in the system to monitor daily living through a user wearing the camera. Ortis et al. and Funari et al. suggested that applying egocentric videos can observe the scene flow from the user's perspective and improve the system, allowing the user's behaviors and intentions to be understood [32,33]. Last, regarding metrics, we used SAM to measure emotional responses by self-assessment. In order to measure the user's responses correctly, it requires a facial emotional response to detect the emotional change in real-time by using a real-time machine learning algorithm [34,35].

7. Conclusions

A smartphone supports various functions, and video recording using a camera application has become one of its most important functions. The user saves memories and moments by recording them on video. However, when the user records a video using a smartphone, the problem is that they need to look at the scene through the mobile phone screen rather than seeing the real-world scene. This is because the user feels the uncertainty of recording a video. For example, they are not sure if the recording is good, and worry about missing the target object. To overcome this, we developed a new camera application that utilizes two main algorithms, MOSSE and HOG, to track the target object and recognize the direction of the user's head. It was assumed that the functions of the new camera application will solve the anxiety of the users during video recording. In addition, we believe that the approach provides more effective and positive emotion in the aspect of a new user experience to the users while recording a video. To test the effectiveness of the application from an affection standpoint, we conducted a case study and measured the emotional responses and error rates as a comparison between using a regular and the new camera applications. The results indicate that the new camera application induced feelings of greater pleasure, excitement, and independence compared with the regular camera application. Furthermore, it effectively reduces the error rates during video recording. In a future study, it will be necessary to expand the up and down sides of the field of view and apply eye-tracking technology to the new camera application.

Author Contributions: Conceptualization, D.P. and C.L.; methodology, D.P. and C.L.; software, C.L.; validation, D.P.; formal analysis, C.L.; investigation, D.P.; resources, C.L.; data curation, D.P. and C.L.; writing—original draft preparation, D.P.; writing—review and editing, D.P.; visualization, D.P.; supervision, C.L.; project administration, C.L. funding acquisition, D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: I would like to record my appreciation to all people that involve in writing this conference paper. First of all, my appreciation goes to my team, Think Tank Team in Samsung Research. Especially, Leo Jun and Sajid Sadi supporting me until I complete this paper successfully.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, R.; Yu, C.; Yang, X.; He, W.; Shi, Y. EarTouch: Facilitating smartphone use for visually impaired people in mobile and public scenarios. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019.
2. Nguyen, D.; Marcu, G.; Hayes, G.; Truong, K.; Scott, J.; Langheinrich, M.; Roduner, C. Encountering SenseCam: Personal recording technologies in everyday life. In Proceedings of the 11th International Conference on Ubiquitous Computing, Orlando, FL, USA, 30 September–3 October 2009.
3. O'Dea, S. Global Smartphone Sales to End Users 2007–2021. Available online: https://www.statista.com/statistics/263437/global-smartphone-sales-to-end-users-since-2007 (accessed on 7 September 2020).
4. DCW Team. The Best Camera Phone in 2020: Which Is the Best Smartphone for Photography? Available online: https://www.digitalcameraworld.com/uk/buying-guides/best-camera-phone (accessed on 7 September 2020).
5. Bolme, D.; Beveridge, J.; Draper, B.; Lui, Y. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
6. Ross, D.; Lim, J.; Lin, R.; Yang, M. Incremental learning for robust visual tracking. *IJCV* **2008**, *77*, 125–141. [CrossRef]
7. Adam, A.; Rivlin, E.; Shimshoni, I. Robust fragments based tracking using the integral histogram. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006.
8. Zhang, X.; Hu, W.; Maybank, S.; Li, X. Graph based discriminative learning for robust and efficient object tracking. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007.
9. Babenko, B.; Yang, M.; Belongie, S. Visual Tracking with Online Multiple Instance Learning. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.

10. Han, K. Image object tracking based on temporal context and MOSSE. *Clust. Comput.* **2017**, *20*, 1259–1269. [CrossRef]
11. Pang, Y.; Yuan, Y.; Li, X.; Pan, J. Efficient HOG human detection. *Signal Process.* **2011**, *91*, 773–781. [CrossRef]
12. Chen, Y.; Chen, C. Fast human detection using a novel boosted cascading structure with meta stages. In *IEEE Transactions on Image Processing*; IEEE: Piscataway Township, NJ, USA, 2008; Volume 17, pp. 1452–1464.
13. Xie, S.; Shan, S.; Chen, X.; Meng, X.; Gao, W. Learned local Gabor patterns for face representation and recognition. *Signal Process.* **2009**, *89*, 2333–2344. [CrossRef]
14. Jin, Z.; Lo, Z.; Yang, J.; Sun, Q. Face detection using template matching and skin-color information. *Neurocomputing* **2007**, *70*, 794–800. [CrossRef]
15. Lian, G. Pedestrian detection using quaternion histograms of oriented gradients. In Proceedings of the 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 28–30 July 2020.
16. Bradski, G.; Kaehler, A. *Learning OpenCV: Computer Vision with the OpenCV Library*; O'Reilly Media, Inc.: Newton, MA, USA, 2008.
17. Grabner, H.; Grabner, M.; Bischof, H. Real-time tracking via on-line boosting. In Proceedings of the British Machine Vision Conference 2006, Edinburgh, UK, 4–7 September 2006; Volume 1, p. 6.
18. Kalal, Z.; Mikolajczyk, K.; Matas, J. Forward-backward error: Automatic detection of tracking failures. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2756–2759.
19. Varfolomieiev, A.; Lysenko, O. An improved algorithm of median flow for visual object tracking and its implementation on ARM platform. *J. Real-Time Image Process.* **2016**, *11*, 527–534. [CrossRef]
20. King, D. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
21. Holzmann, C.; Hochgatterer, M. Measuring Distance with Mobile Phones Using Single-Camera Stereo Vision. In Proceedings of the 32nd International Conference on Distributed Computing Systems Workshops, Macau, China, 18–21 June 2012; pp. 88–93.
22. Parkkola, H.; Saariluoma, P. Would Ten Participants Be Enough for Design of New Services? In *Quality and Impact of Qualitative Research*; Institute for Integrated and Intelligent Systems, Griffith University: Queensland, Australia, 2006; p. 86.
23. Park, H.; Lee, J.; Bae, S.; Park, D.; Lee, Y. A Proposal for an Affective Design and User-Friendly Voice Agent. In *International Conference on Human Systems Engineering and Design: Future Trends and Applications*; Springer: Cham, Switzerland, 2018; pp. 249–255.
24. Desmet, P.; Overbeeke, K.; Tax, X. Designing products with added emotional value: Development and application of an approach for research through design. *Des. J.* **2001**, *4*, 32–47. [CrossRef]
25. Lang, P. *The Cognitive Psychophysiology of Emotion: Fear and Anxiety*; Routledge: London, UK, 1985.
26. Bradley, M.M.; Lang, P.J. Measuring emotion: The self assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [CrossRef]
27. Morris, J. Observations: SAM: The Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *J. Advert. Res.* **1995**, *35*, 63–68.
28. Que, W.; Hakoda, Y.; Onuma, N.; Morikawa, S. The effect of negative emotion on eyewitness functional field of view. *Shinrigaku Kenkyu Jpn. J. Psychol.* **2001**, *72*, 361–368.
29. Meyer, B.; Lipski, C.; Scolz, B.; Magnor, M. Real-time free-viewpoint navigation from compressed multi-video recordings. In Proceedings of the 3D Data Processing, Visualization and Transmission (3DPVT), Padova, Italy, 19–21 June 2010; pp. 1–6.
30. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
31. Voigtlaender, P.; Krause, M.; Osep, A.; Luiten, J.; Sekar, B.; Geiger, A.; Leibe, B. MOTS: Multi-object tracking and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
32. Ortis, A.; Farinella, G.; D'Amico, V.; Addesso, L.; Torrisi, G.; Battiato, S. Organizing egocentric videos of daily living activities. *Pattern Recognit.* **2017**, *72*, 207–218. [CrossRef]
33. Furnari, A.; Farinella, G.; Battiato, S. Temporal segmentation of egocentric videos to highlight personal locations of interest. In *European Conference on Computer Vision*; ECCV: Amsterdam, The Netherlands, 2016.
34. Mehendale, N. Facial emotion recognition using convolutional neural networks (FERC). *SN Appl. Sci.* **2020**, *2*, 446. [CrossRef]
35. Dashtipour, K.; Gogate, M.; Cambria, E.; Hussain, A. A novel context-aware multimodal framework for persian sentiment analysis. *arXiv* **2021**, arXiv:2103.02636 2021.