INTELLIGENT AUTOMATION & SOFT COMPUTING, 2018
https://doi.org/10.1080/10798587.2017.1327549

Portrait Vision Fusion for Augmented Reality

Li-Hong Juang[a], Ming-Ni Wu[b] and Feng-Mao Tsou[b]

[a]School of Electrical Engineering and Automation, Xiamen University of Technology, No.600, Ligong Road, Jimei, Xiamen, 360124, P.R.China; [b]Department of Information Management, National Taichung University of Technology, Taichung, Taiwan

ABSTRACT
Video communication is a common way to communicate via interactive technology, especially using webcams for remote interaction and for each participant to see each other's characteristics from the screen display. In this paper, the main goal is to augment some dynamic interactive virtual environments. Towards this goal , a method using superimposing a segmented human portrait on a panoramic background is proposed, then the limb interactive element is added into these videos involved with a dynamic portrait segmentation method meanwhile using a Kinect (+openCV) device to extract a portrait for amendment, finally acquires a full portrait of information. Because the face is the most important identification region in a portrait, a head skeleton tracking method is also used to strengthen the remedy for its head segmentation, further uses the edge transparent processing to synthesize them into the video. The approach leds the users can verbally and physically communicate through these video interactive much more vibrantly.

KEYWORDS
Kinect(+openCV); Dynamic portrait segmentation; Skeletal tracking; Edge transparent processing; Video interactive

1. Introduction

Human interaction is an important process for building relationships, comparing with the traditional telephone communications, a fast developing paradiam in network and intelligent mobile devices, which lets people communicate face to face, and is for long distance interaction. In recent years, mobile devices are used as social sites (such as Facebook, and Google+ wretch). Pictures and text in multimedia communication have made the interactive process more pixel-rich, now it has become the major trend of communication for the youth groups, but it also indirectly led the interaction confined to the tiny device interface, therefore bow-in-response word was born. Furthermore, the video is also a common way to communicate in the interactive multimedia, using network cameras letting different people around the world seeing each other's screen characters and interact with each other in real time. Currently, intelligent network cameras can also be equipped with mobile devices, so that video is no longer restricted by a fixed location; these techniques are implemented using so called Augmented Reality (AR) based on computer generated garments, which are superimposed on real data. AR has gained more and more significant attention in different emerging applications. Yuan et al. (2013) used AR for improving the shopping experience by assisting customers to make purchase decisions or at least help to narrow down selections before physically trying them on. Chen, et al. (2008) used AR to make it possible to increase the understanding of viewed objects by combining real and synthesized information. In recent years, AR is always used for images captured by video cameras in a virtual fitting room, which are augmented with the images of garments in real-time to build 3D models by scanning the objects (Alexiadis & Zarpalas, 2013) to recognize human actions (Cai, Zhou, Wu, Luo, & Li, 2016) using depth cameras. Furthermore, spatial augmented reality (Menk & Koch, 2013), (Sheng, Yapo, & Cutler, 2010) is interesting for the design process of a car, because a lot of virtual content and corresponding real objects are used. Moreover, AR allows to seamlessly insert virtual objects in an image sequence (Marchand, et al., 2016).

For AR processing, the scene in an image can be roughly divided into the foreground and the background, the significant objects are mainly in the foreground (Ratnaweera, et al., 2004), (Julià, et al. , 2011), (Mutto, et al., 2010, 2012). The approach is quick to classify the objects' areas in the foreground and the background as well as to process the so-called dynamic image segmentation of portraits in any environment through the objects' corresponding depth position when a scene is combined with the 3D depth information. However, the special equipment has a high price to acquire 3D information (Mutto et al., 2010). Kinect (+openCV) is a motion-sensing device produced by Microsoft, not only reduces the cost, but also obtains the rich information such as color information, depth information, portrait recognition and human skeleton information at the same time. Kinect (+openCV) uses infrared projection to calculate depth. In the process of projection, infrared rays might be refracted by the smooth objects or be absorbed by the blacker objects. As a result, it is unable to receive the information and it has a great loss. This will generate irregular serration on the edge of images.

The interactive multimedia outlets from the foregoing processing can also be as members of sport interactions, and in the modern equipment, it can use the somatosensory system to do the limb simulation for an indoor interactive feature. For a 3D depth detection, because Kinect (+openCV) prices compared with other 3D detection equipment is low and has the functions to track the human skeleton and the prospects for these capturing portraits, it was chosen in this study. Kinect

CONTACT Li-Hong Juang ✉ lipuu@qq.com

(+openCV) is mostly used for an application on the skeleton tracking, however its dynamic portrait segmentation is not generally too accurate, so for this reason, we will refine it by a new method. In this paper, a method for superimposing a segmented human portrait on a panoramic background is proposed. The main goal of the research is to augment a dynamic interactive virtual environment.

Mutto et al., (2012) merged color and depth information to optimize the image segmentation by comparing with the various depth-capturing equipment. Moreover, the researchers also segmented these scenes using a depth-of-field shallow. First, they seperated the whole area into the foreground, the background and the undefined areas by using the "trimap" classification method were watershed segmentation processing. Then, the undefined areas can be processed by the "Alpha matte" method. Finally, they found that after these procedures, the captured images are more complete. From all the merits of above surveys, we refined further this method to let it be compared to uncertain areas by the color information and the depth information. Through this refined method, the portrait will be more complete by complementing the information in the segmentation process.

The proposed method is based on segmenting a human body where data is acquired from the Kinect (+openCV) sensor. The segmented human portrait is obtained by means of a relative segmentation method. This research proposes to amend the dynamic portrait segmentation from Kinect (+openCV) sensor (Lee et al., 2012), (Hsieh, et al., 2010) and acts from the relative news from network camera for their combination, and uses the background removal method (Khongkraphan & Kaewtrakulpong, 2011), (Vincent & Soille, 1991), (Luo et al., 2012) and adds the different people in different locations to enter into the same background scene, then they look like each other on the picture, therefore the Kinect (+openCV) user can follow the holder with this smart type action device who is in travel and do limb exchanges to reach a far distant interaction.

2. Methods

2.1. Somatosensory Technology

Somatosensory technology is that people within the detection range of the device can directly control the limb movements without the use of complex operations. This technology is based on somatosensory modality principles and their difference can be divided into three broad categories (http://www.86wiki.com/view/5338106.htm); the inertial sensor, the light sensor and the combination sensor, respectively. The optical sensor is the use of lasers for sending and receiving the human body images as well as the depth information in terms of strength and angle. The combination sensor consists of a three-axis gyroscope and infra-red sensors, which can accurately detect the wrist rotation movement.

The Kinect (+openCV) sensor equipment used in this study, was developed by Microsoft was primarily used for XBOX360 interactive game products for Windows and commercial applications. Kinect (+openCV) usage has the light sensing technology and supports the portrait identification, voice recognition features, and its video aspects is composed by the color camera and the infra-red launching players and the infra-red receiving players. The infra-red launching players will be a ray shot with average projection to measure space and reflect through the infra-red and the receiving players receive each point location. It further forms its depth images by the internal operation, finally go through tracking the prospects of portrait depth operation from the twenty human skeleton locations. Because Kinect (+openCV) is designed for using for body movements, it means to detect the body, which is preferring on the speed of the operation and has poor accuracy (Gonzalez-Jorge, et al., 2013).

2.2. Image Segmentation

Image segmentation is the scene of calculus and an assortment of features such as color or texture to distinguish each object in response to the users needs.

There are several classification methods in this research; a portrait is usually located in the foreground of a scene (Luo et al., 2012). Segmenting the foreground image is a complex issue, some methods solve this problem such as the prospect segmentation for a low depth field image (Liu et al., 2010), or use the triangulation method to detect the image of the human limb for segmenting (Luo et al., 2012) and so on. For segmenting video images, it is needed to operate quickly, this research uses a dynamic portrait, a simplified method based on Kinect (+openCV)'s depth image using the trimap concept for further analysis, and uses the skeleton tracking to remedy the head for furthermore segmenting. The concept is to categorize a trimap as the foreground, the background and the uncertaint areas, as long as the uncertainties region can be analyzed, it can fully separate foreground and background images (Juan & Keriven, 2005), which is a major focus of this study.

2.3. Interactive Video

The interactive video is boosted up due to the expansion of network bandwidth, with internet users increasing and more video talking since 2005. The interactive video type can be divided into the custom interactive video, and the exploring and dialogue online interactive video (http://en.wikipedia.org/wiki/Interactive). The custom interactive video means that a user can view and edit with interactive components such as buttons, depending on user preferences. The exploring interactive video allows users to see through a space from multiple angles of an object, as an art show on the network can rotate through multiple perspectives and carefully appreciates the art works. The dialogue interactive video is now commonly used for the remote real-time video communication, using network cameras and other equipment for "one to one" or "one to many" video conversations. The applications like the interactive video game for the muscles of lower limb rehabilitation training for the elderly (Chen et al., 2012) and the doctor conveys the instructions to perform the rehabilitation campaign at home through the video. However, these interactions are user interactions with virtual objects, therefore here a new method for the multi-user offsite for physical interaction is proposed for online groups and participants of blends interactive video, then let the interactive video provide more real exchanges.

3. Experiments

The objective of this system is to enable users the different physical interactions between the two parties, share each other's space environment, uses dynamic portraits segmentation processing, and reduce the fusion violation produced by portrait and video. This section will describe the overall system architecture and the approach of each device. Figure 1 shows the system process for this case, in this experiment users are “User A” and “User B”, two parties, respectively. User A, the end of part is the Kinect (+openCV) portrait sense equipment for the production of data in this research method, after the computer operation processing produces a dynamic full portrait segmentation. In User B, the end of part produces the simple network depending on the news, and synchronizes with User A. This study does not take into account these issues such as network traffic, mainly use User A is for portrait and video fusion.

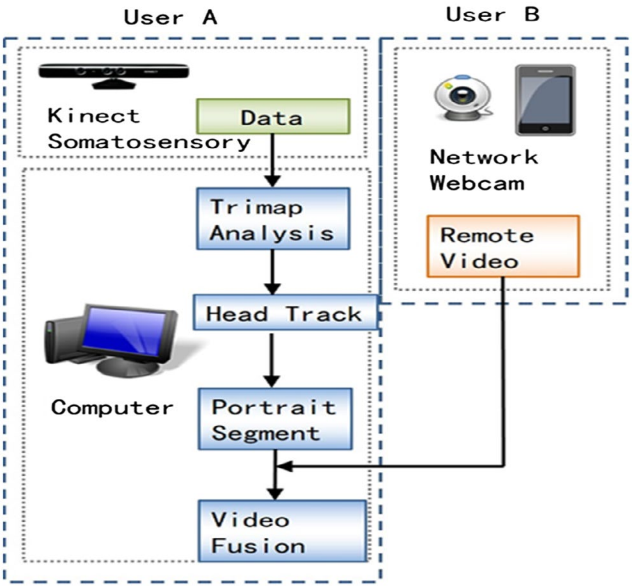


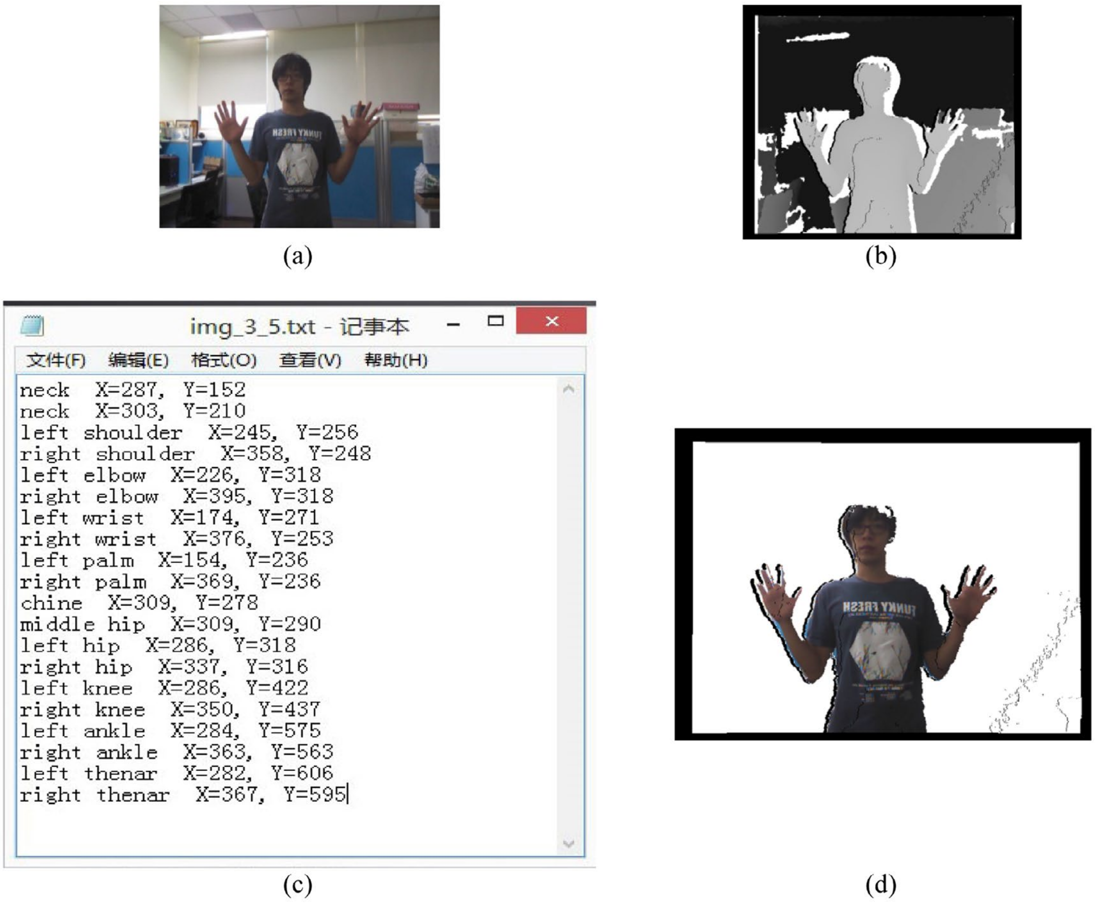
Figure 1. The System Processing Flowchart.

3.1. Acquiring Data

In this study, Microsoft Kinect (+openCV) Windows SDK v1.5 is used for acquiring somatosensory data, as Figure 2 shows, this data is for color information, depth and the twenty skeleton tracking points. In the depth information, the Kinect (+openCV) device has its own searching portrait feature information, however the depth information is detected by the infra-red 3D depth sensor. Because infra-red is easy to have some gloss of refraction property and black of things due to interference, led the depth information being is not complete. As shown in Figure 2d for explanation, the direct depth portrait segmenting will let portrait segmenting not complete, especially the head is seriously, therefore next we will use the proposed method to settle down this problem.

3.2. Image Analysis and Head Tracking

In the part of image analysis, Kinect (+openCV) is used for the face information, which will be found by the in-depth information, the trimap classification for the foreground, the background, and the uncertain areas. As shown in Figure 3, the foreground in the studied portrait is in red, the background is the outside of portrait scene indicated in green and the uncertain areas is likely not to be classified the areas of foreground or background presented in white. Here also has the depth information cannot detect or incorrect within the error range classified as the uncertain areas. For the portrait information,



(a)

(b)

```
neck  X=287, Y=152
neck  X=303, Y=210
left shoulder  X=245, Y=256
right shoulder  X=358, Y=248
left elbow  X=226, Y=318
right elbow  X=395, Y=318
left wrist  X=174, Y=271
right wrist  X=376, Y=253
left palm  X=154, Y=236
right palm  X=369, Y=236
chine  X=309, Y=278
middle hip  X=309, Y=290
left hip  X=286, Y=318
right hip  X=337, Y=316
left knee  X=286, Y=422
right knee  X=350, Y=437
left ankle  X=284, Y=575
right ankle  X=363, Y=563
left thenar  X=282, Y=606
right thenar  X=367, Y=595
```

(c)

(d)

Figure 2. Somatosensory Technology Data from Kinect(+openCV) (a) Color Information (b) Depth Information (c) Twenty Skeletal Tracking Points Data Information (d) Tracking Human Image Segmentation.

it is found the damage of the head portrait is most serious. The face is the main information identifying a person, it must be maintained for its integrity so that the special tracking ro remedy the skeleton of head and neck. To prove the accuracy of trimap image analysis, a second example test shown in Figure 4, is made as the second example for the same procedure process.

3.3. Dynamic Portrait Segmentation

This study focuses on 640X480 pixels and thirty frames for color and depth information per second and designs the image processing system using Visual C # 2010 Express, as shown in the flowchart of Figure 5. To make the face a denser depth portrait image in color images, we made adjustments based on the experimental device, for example, upward three pixels' displacement, and offset one pixel to the right. Because there are many gaps between the depth disparity resulting from the correction, we expanded the known 3X3 portrait depth image to fill up spaces between two depths.

To enhanc the skeleton tracking of the upper human body process procedure, it is necessary to pick up these key points of skeleton tracking from the upper human body as shown in Figure 6. It will be albe to strengthen the major head broken caused by the Kinect (+openCV) portrait process. Next it is to remedy the more serious damaged head, from our experience, the portrait will be included in the scope of foreground and uncertain areas, and it takes advantage of the head for its skeleton tracking. A portrait in the detection range of uncertainty will enter into the completed known portraits of the remedied head. The detection range as Equations (1) and (2) are the followings:

$$\text{Distance}^{\text{skeleton}} = \overline{\text{Head}_m, \text{Neck}_n}$$

$$\text{Distance}_i^{\text{pixel}} = \overline{\text{Head}_m, \text{Pixel}_i}$$

$$T = \text{Distance}^{\text{skeleton}} \times 0.6 \tag{1}$$

$$\text{if}\left(\left(\text{Distance}_i^{\text{pixel}} < \text{T}\right)\text{and}\left(\text{Pixel}_i \in \text{Unsure}^{\text{region}}\right)\right) \text{ then } \left(\text{Sure}_i^{\text{regio } n} = \text{true}\right) \tag{2}$$

In Equation (1), $\text{Distance}^{\text{skeleton}}$ called skeleton distance representatives for the head skeleton points Head_m to the neck skeleton points Neck_n of the shortest distance, and $\text{Distance}_i^{\text{pixel}}$ for the current pixel distance is by the head skeleton points Head_m to the current operation pixel points Pixel_i of the shortest distance. Here m, n and i are for each pixel location, and T for the detection threshold of the head area is 0.6 times the skeleton distance, here times is the multiples for better measurement results under many experiments. In Equation (2), $\text{Unsure}^{\text{region}}$ represents for the uncertain regions, $\text{Sure}_i^{\text{region}}$ represents the known portrait pixels, while the current pixel distance is less than T, and the current pixels are in the uncertain regions, here classifing the pixels as the part of head and add to a known portraits region. Finally, it is moved on the known portrait for its edge transparent processing, mainly it will let the portrait synthesis show a better fusion of video vision and reduce the cutting sawtooth on the edge. However, before the transparent processing, first it must be for edge detection, because of the series stream pixel information being of real-time processing, it cannot detect for unready pixels, therefore the left, right and top location of portrait are for edge detection. it needs to set the directional decreasing transparency matrix as Equation (3):

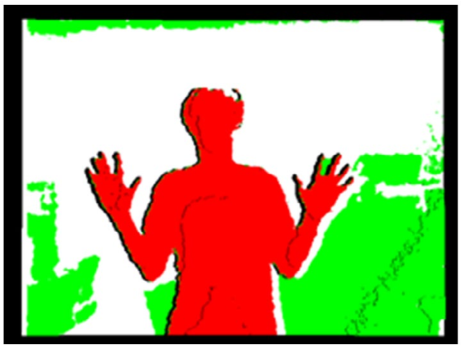


Figure 3. The Trimap Image Analysis.

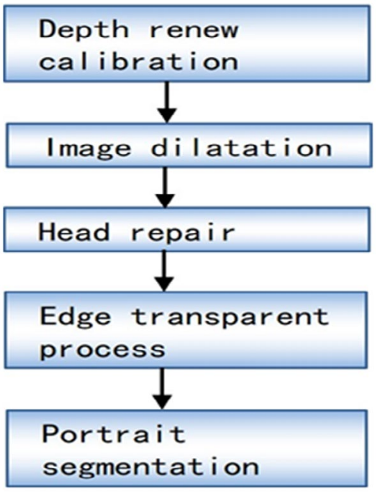


Figure 5. The Procedure for Dynamic Portrait Segmentation.

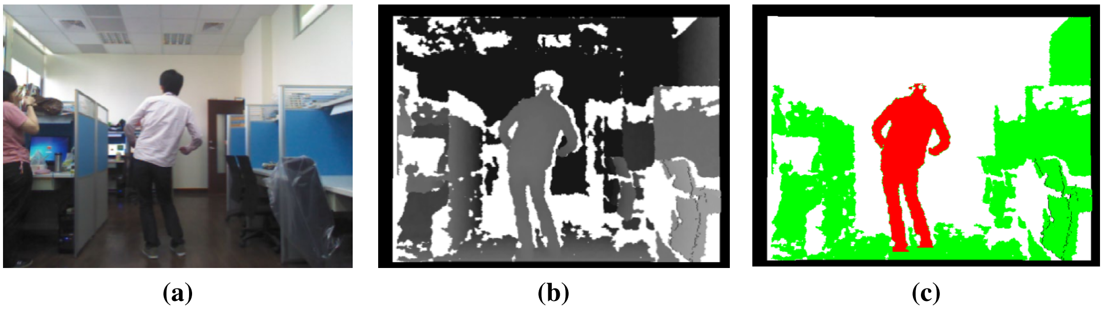


Figure 4. The Second Example for the Trimap Image Analysis (a) Color Information (b) Depth Information (c) Trimap Results.

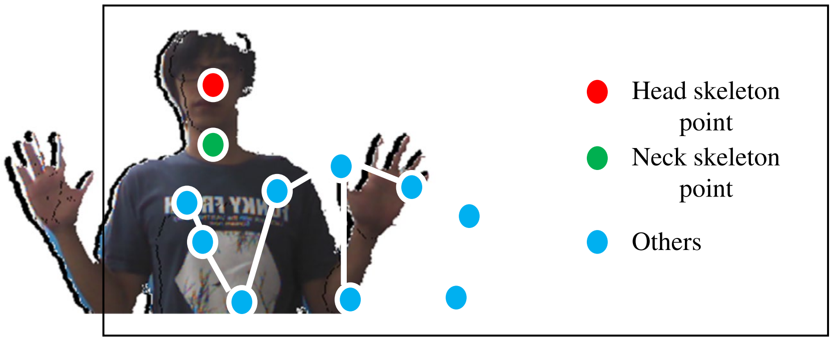
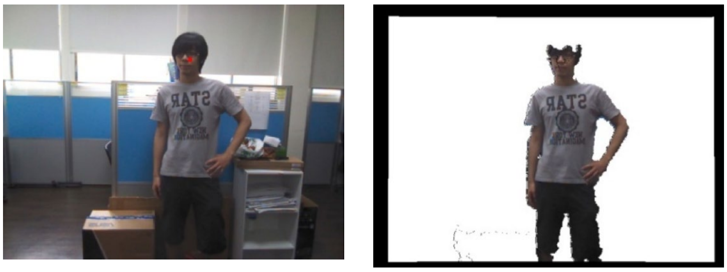


Figure 6. The Skeleton Tracking for the UpperHuman Body.



(a) (b) (c)

Figure 7. The Results of Dynamic Portrait Segmentation (a) Kinect(+openCV) Head Skeletal Tracking (b) Kinect(+openCV) Portrait Interior Detection (c) Final Portrait Segmentation.

$$\text{Alpha}^{\text{left}} = \begin{bmatrix} 90 & c0 & e0 \\ 90 & c0 & e0 \\ 90 & c0 & e0 \end{bmatrix}$$

$$\text{Alpha}^{\text{right}} = \begin{bmatrix} e0 & c0 & 90 \\ e0 & c0 & 90 \\ e0 & c0 & 90 \end{bmatrix}$$

$$\text{Alpha}^{\text{top}} = \begin{bmatrix} 90 & 90 & 90 \\ c0 & c0 & c0 \\ e0 & e0 & e0 \end{bmatrix} \tag{3}$$

where $\text{Alpha}^{\text{left}}$, $\text{Alpha}^{\text{right}}$ and $\text{Alpha}^{\text{top}}$ are for the left, right and top decreasing transparency matrix respectively. Their values will be hexadecimal values, the value of FF is for fully opaque and the value of 00 is for fully transparent. These matrices will separately apply for the respective directions of the edge, and the edge detection equations is as follows:

$$\text{if}\left(\left(\text{Pixel}_{(i-6)\sim(i-1)} \in \text{Orther}^{\text{region}}\right)\text{and}\left(\text{Pixel}_i \in \text{Sure}^{\text{region}}\right)\right)$$

$$\text{then }\left(\text{Alpha}^{\text{left}}\left(\text{Pixel}_i^{\text{alpha}}\right)\right) \tag{4}$$

$$\text{if }((\text{Pixel}_{((i-5)\sim i)} \in \text{Orther}^{\text{region}})$$

$$\text{and }(\text{Pixel}_{i-6} \in Sure^{\text{region}})$$

$$\text{then }(\text{Alpha}^{\text{right}}(\text{Pixel}_{i-6}^{\text{alpha}}) \tag{5}$$

$$\text{if }((\text{Pixel}_{\text{i(width}\times(1\sim5))} \in \text{Orther}^{\text{region}})$$

$$\text{and }(\text{Pixel}_i \in Sure^{\text{region}})$$

$$(\text{then }\ \text{Alpha}^{\text{top}}(\text{Pixel}_i^{\text{alpha}}) \tag{6}$$

where $\text{Orther}^{\text{region}}$ is for the non-portrait zone. Equation (4) is for the left edge detection, when the operation goes through five continuous non-portrait pixels and the current pixel is located in the portrait region, then it is considered as the left edge and uses its pixel transparency $\text{Pixel}_i^{\text{alpha}}$ for a center, meanwhile use $\text{Alpha}^{\text{left}}$ for the edge transparent processing. Equation (5) is for the right edge detection and the same process as in Equation (4), but use $\text{Alpha}^{\text{right}}$ for the right edge transparent processing. Equation (6) shows the top edge detection; here width is for the data width of the color image, as the same process as above, but use $\text{Alpha}^{\text{top}}$ for the top edge transparency.

4. Results and Discussion

Because of the device limitations and the instruction cycle considerations, in this paper, in order to simplify operational complexity for segmenting and video image fusion, we set a rule of thirty frames per second, using I7–3770 3.40 GHz CPU, 12 GB RAM, WIN7 64-bit operating system, webcam devices; and Logitech Pro 9000 as our computing platform.

Consider Figure 7 for the example, the red point on the head in Figure 7a is for the tracking point of the head skeleton, Figure 7b shows the Kinect (+openCV) automated ripping portrait image, where one can see some parallax correction errors and the head shows a incomplete pose. Figure 7c shows a dynamic portrait segmenting result; the proposed method

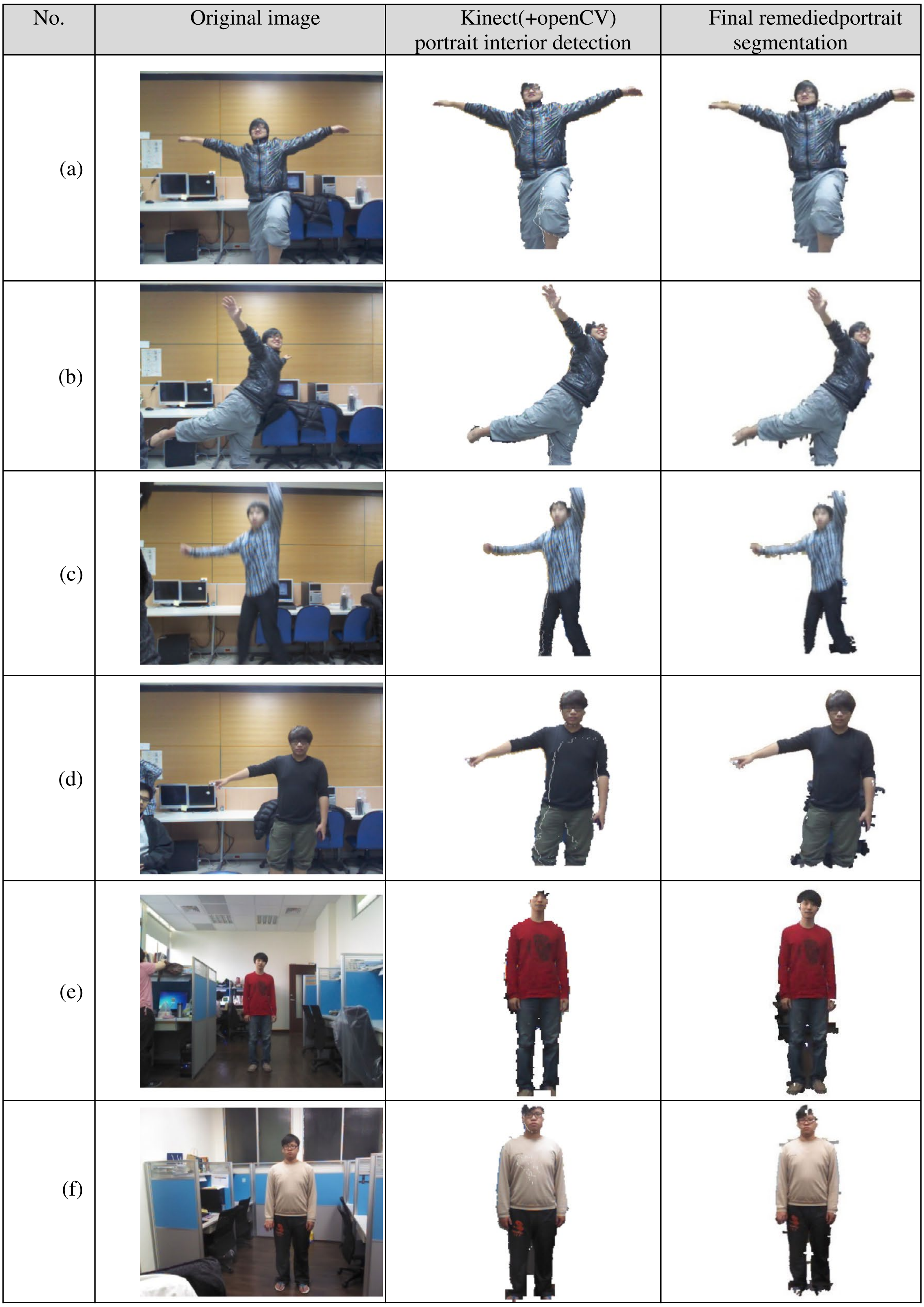

No.
Original image
Kinect(+openCV) portrait interior detection
Final remediedportrait segmentation
(a)
(b)
(c)
(d)
(e)
(f)

Figure 8. These Testing Results for More Complex Postures from the Original Image, Kinect(+openCV) Portrait Interior Detection and Final Remedied Portrait Segmentation.

obviously improves the Kinect (+openCV) portrait segmenting problem and made a full portrait image. Furthermore, more testing was done to prove the proposed method's accuracy. Figure 8 shows these testing results for more complex postures from the original image, the Kinect (+openCV) portrait interior detection and final remedied portrait segmentation, we can get the full head skeleton by using the proposed method. These results also demonstrate all images made by the Kinnect (+openCV), which are broken at their heads, so they can not further use video fusion for augmented reality, however, after our proposed remedying method, almost all heads have full detail, these are our major contributions. Figure 9 presents the flowchart for the actual video fusion processes. Figure 10 is the final fusion result. All the experiments show that the proposed method improves the Kinect (+openCV) segmentation and demonstrates that the portrait can have relative

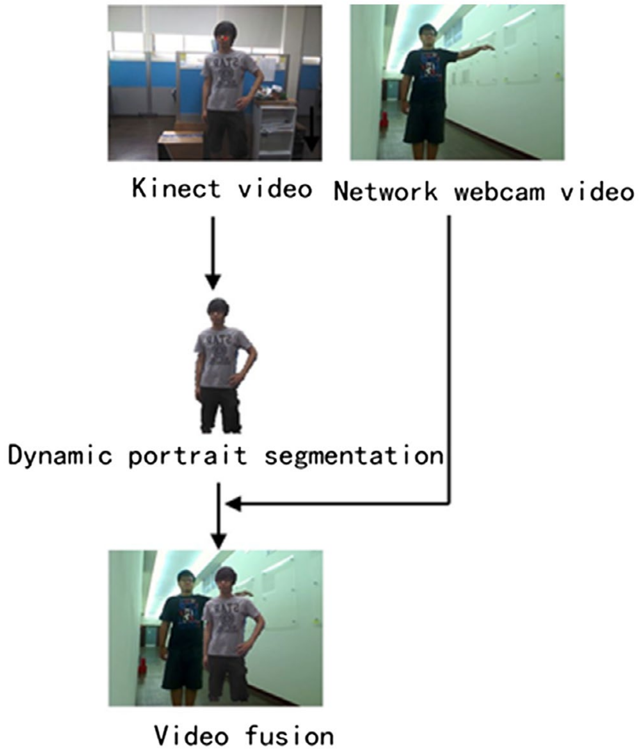


Figure 9. The System Schema for Interactive Video Fusion Processing.

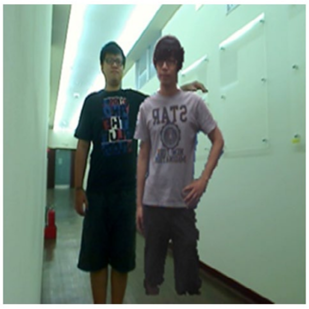


Figure 10. The Final Video Fusion Result.

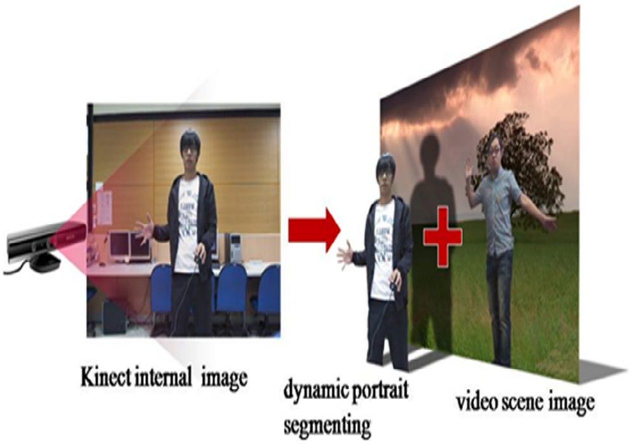


Figure 11. The Display for Video Scene Fusion.

integrity and the error rate is within 10%. To the best of our knowledge no other similar researches and experiments can make a comparison with ours. This is a remarkable result with the Kinect (+openCV).

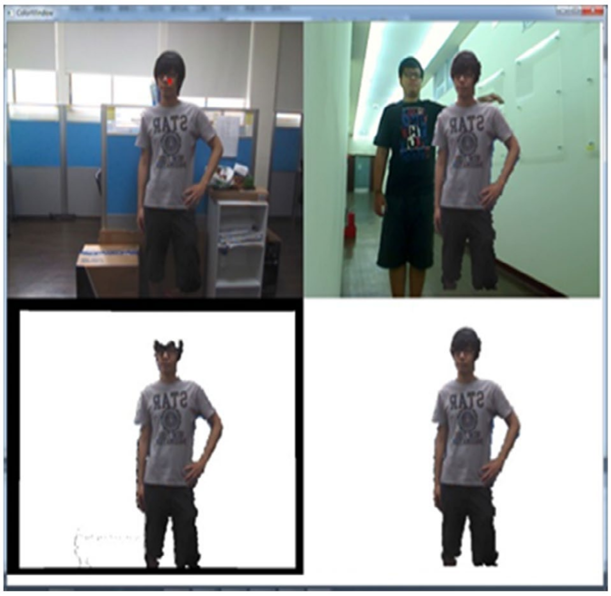


Figure 12. The Serial Scenes for Video Fusion.

It is expected that this system can render a remote tour, where the player sets up multiple network photography in the touring scene in advance by dynamic scenery cams, users can select these views, then enter the scene in the video through the capture of dynamic portraits and set up private interactive zones, furthermore invite some friends and families to join into this video sharing physical and verbal interaction and enhancing their relationship. Figure 11 shows the display for video scene fusion. Figure 12 is the system for running the dynamic portrait fusion of interactive vision in Visual C # 2010 Express screen.

5. Conclusions

This research excludes the massive computing time needed for dynamic image segmentation, the dynamic portraits segmentation can be full of expression for portraits, and particularly the head area is the most significant, segmenting a good portrait by the transparent edge treatment, then it can be better integrated into a video. The main contribution is to augment dynamic interactive virtual environments by means of a relative segmentation method. The experimental results show that the proposed method has significantly improved the Kinect (+openCV) segmenting problem with a complete portrait, people will increase further intention to use it for video fusion.

Authors'contributions

A method using superimposing a segmented human portrait on a panoramic background is proposed, then the limb interactive element is added into these videos involved with a dynamic portrait segmentation method meanwhile uses Kinect (+openCV) sensor to extract a portrait for amendment, finally acquires a full of portrait information.

Acknowledgement

The authors deeply acknowledge the financial support from Xiamen University of Technology, Fujian, P.R. China under the Xiamen University of Technology Scientific Research Foundation for Talents plan.

Disclosure statement

The authors declare that they have no competing interests.

Notes on contributors

Li-Hong Juang received a B.S. degree in Civil Engineering from the National Chiao Tung University, Taiwan, in 1990 and an M. S. degree in Applied Mechanics from the National Taiwan University, Taiwan, in 1993, and a Ph.D. degree in Control and Embedded System Group from Department of Engineering at Leicester University, UK, in 2006. Now he is a Professor at School of Electrical Engineering and Automation, Xiamen University of Technology, P. R. China. His research interests are in analysis, modeling, smart system design, image process, machine vision, robot control and medical system.

Ming-Ni Wu received a Ph.D. degree in Computer Science & Information Engineering, National Chung Cheng University, Taiwan. She is an associate professor at the Department of Information Management, National Taichung University of Technology, Taiwan. Her research interests are in image process and information management.

Feng-Mao Tsou received an M. S. degree in Department of Information Management, National Taichung University of Technology, Taiwan. His research interests are in image processing.

References

Alexiadis, D.S., & Zarpalas, D. (2013). Real-time, full 3-D reconstruction of moving foreground objects from multiple consumer depth cameras. *IEEE Transactions on Multimedia, 15*, February, pp. 339–358.

Cai, X., Zhou, W., Wu, L., Luo, J., & Li, H. (2016). Effective active skeleton representation for low latency human action recognition. *IEEE Transactions on Multimedia, 18*, February, 141–154.

Chen, Y.H., Chia, T.L., Lee, YK.., Huang, S.Y., & Wang, R.Z. (2008). A vision-based augmented-reality system for multiuser collaborative environments. *IEEE Transactions on Multimedia, 10*, 585–595, June 2008.

Chen, P.Y., Wei, S.H., Hsieh, W.L., Cheen, J.R., Chen, L.K., & Kao, C.L. (2012). Lower limb power rehabilitation (LLPR) using interactive video game for improvement of balance function in older people. *Archives of Gerontology and Geriatrics, 55*, 677–682.

Gonzalez-Jorge, H., Riveiro, B., Vazquez-Fernandez, E., Martínez-Sánchez, J., & Arias, P. (2013). Metrological evaluation of microsoft kinect and Asus Xtion. *Measurement,* 1800–1806.

Hsieh, J.W., Chuang, C.H., Chen, S.Y., Chen, C.C., & Fan, K.C. (2010). Segmentation of human body parts using deformable triangulation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 40* MAY, 596–610.

Juan, O., & Keriven, R. (2005). Trimap segmentation for fast and user-friendly alpha matting. *Variational, geometric, and level set methods in computer vision, lecture notes in computer science, 3752*, 186–197.

Julià, C., Moreno, R., Puig, D., & Garcia, M.A. (2011). Shape-based image segmentation through photometric stereo. *Computer Vision and Image Understanding, 115*, 91–104.

Khongkraphan, K., & Kaewtrakulpong, P.. (2011). A novel reconstruction and tracking of 3D-articulated human body from 2D point correspondences of a monocular image sequence. *IEICE Transactions on Information and Systems, E94–D*, 1090–1098.

Lee, C.Y., Leou, J.J., & Hsiao, H.H. (2012). Saliency-directed color image segmentation using modified particle swarm optimization. *Signal Processing, 92*, 1–18.

Liu, Z., Li, W., Shen, L., Han, Z.G., & Zhang, Z.Y. (2010). Automatic segmentation of focused objects from images with low depth of field. *Pattern Recognition Letters, 31*, 572–581.

Luo, W., Li, H.G., Liu, G.G., & NgiNgan, K. (2012). Global salient information maximization for saliency detection. *Signal Processing: Image Communication, 27*, 238–248.

Marchand, E., Uchiyama, H., & Spindler, F. (2016). Pose estimation for augmented reality: A hands-on survey. *IEEE Transactions on Vision and Computer Graphics., 22*, December, 2633–2651.

Menk, C., & Koch, R. (2013). Truthful color reproduction in spatial augmented reality applications. *IEEE Transactions on Vision and Computer Graphics., 19*, February, 236–248.

Mutto, C.D., Zanuttigh, P., & Cortelazzo, G.M. (2010). A probabilistic approach to ToF and stereo data fusion. In Proc. 3DPVT, Paris, France, 2010.

Mutto, C,D., Zanuttigh, P., & Cortelazzo, G,M. (2012). Fusion of geometry and color information for scene segmentation. *IEEE Journal of Selected Topics in Signal Processing, 6*, 505–521.

Ratnaweera, A., Halgamuge, S.K., & Watson, H.C. (2004). Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients. *IEEE Transactions on Evolutionary Computation, 8*, 240–255.

Sheng, Y., Yapo, T.C., & Cutler, B. (2010). Global illumination compensation for spatially augmented reality," *Computer Graphics Forum, 29*, May, 387–396.

Vincent, L., & Soille, P. (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*, 583–598, 1991.

Yuan, M., Khan, I.R., Farbiz, F., Yao, S., Niswar, A., & Foo, M.H.(2013). A mixed reality virtual clothes try-on system. *IEEE Transactions on Multimedia, 15*, December, 1958–1968.