An Image Based Approach to Hand Occlusions in Mixed Reality Environments

Andrea F. Abate, Fabio Narducci, and Stefano Ricciardi

VRLab, University of Salerno
{abate,fnarducci,sricciardi}@unisa.it

Abstract. The illusion of the co-existence of virtual objects in the physical world, which is the essence of MR paradigm, is typically made possible by superimposing virtual contents onto the surrounding environment captured through a camera. This works well until the order of the planes to be composited is coherent to their distance from the observer. But, whenever an object of the real world is expected to occlude the virtual contents, the illusion vanishes. What should be seen behind a real object could be visualized over it instead, generating a "cognitive dissonance" that may compromise scene comprehension and, ultimately, the interaction capabilities during the MR experience. This paper describes an approach to handle hand occlusions in MR/AR interaction contexts by means of an optimized stereo matching technique based on the belief propagation algorithm.

Keywords: mixed reality, hand occlusion, disparity map.

1 Introduction

As the number of augmented and mixed reality applications available on a variety of platforms increases, so does the level of interaction required, possibly leading to the emergence of challenging visualization issues. To this regard, it is worth to note that the illusion of the co-existence of virtual objects in the physical world (the essence of MR paradigm) is typically made possible by so called video-based[1] see-through approach in which the rendering of virtual contents is superimposed onto the surrounding environment captured in real time by means of a proper transformation. This trick works well until the order of the planes to be composited is coherent to their distance from the observer (see Fig. 1_Left). But, whenever an object of the real world is expected to occlude the virtual contents, the illusion vanishes since the order of rendered planes does not lead to a correct visualization (see Fig. 1_Right). As a result, what should be seen behind a real object could be visualized over it instead, generating a "cognitive dissonance" due to the loss of spatial coherence along the axis normal to camera plane that may compromise scene comprehension and, ultimately, the interaction capabilities during the MR experience.

[1] Optical see-through is the other well known option for MR/AR, but besides being less diffused it is inherently less suited to support processing of environment visualization.

R. Shumaker and S. Lackey (Eds.): VAMR 2014, Part I, LNCS 8525, pp. 319–328, 2014.

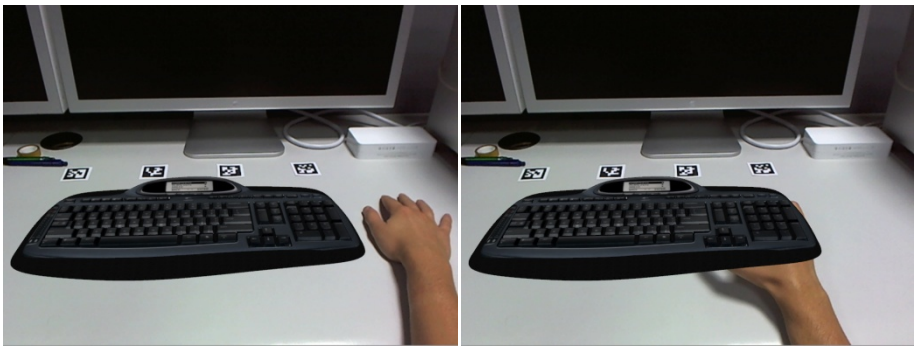

Fig. 1. *Left:* A virtual model of a keyboard rendered onto a captured frame of real environment to augment it. The hand positioned along the right side of the keyboard does not ruins the Mixed Reality illusion. *Right:* The same MR scene, but as the hand is positioned over the keyboard, it is occluded by the virtual content.

This paper describes an image-based method aimed to address effectively hand occlusion in many MR/AR interaction contexts without any additional hardware, apart from video see-through goggles enabling stereo-vision. In brief, the proposed method composites the rendered virtual objects onto the incoming video see-through streams according to a disparity map encoding real-to-virtual visualization order at a pixel level as a gray-scale image by means of stereo matching. We optimize the performance of the algorithm by segmenting the input image between hand and not-hand regions via a skin-tone filtering in the HSV color space (less affected from lighting conditions than RGB space). The purpose of this segmentation is twofold. From the one hand it is possible to reduce the region of interest (that directly affects the computational cost of the disparity map) to a cropped region of the original frame, on the other hand the contour of the segmented hand region is used as a reference to improve the edge sharpness of the disparity map.

The rest of this paper is organized as follows. Section 2 presents previous works related to this research. Section 3 describes the overall system's architecture and each of its main components. Section 4 reports about first experiments to assess the advantages and the limitations in the proposed approach. Finally, Section 5 draws some conclusions introducing future directions of this study.

2 Related Works

Hand occlusion in augmented reality is a challenging topic and scientific literature presents diverse approaches to it. In particular, displaying occluded objects in a manner that a user intuitively understands is not always trivial. Furmanski et al. [1] in 2002 developed new concepts for developing effective visualizations of occluded information in MR/AR applications. They designed some practical approaches and guidelines aimed at evaluating user's perception and comprehension of the augmented scene and distances. Many researchers aimed at solving the incorrect occlusion problem by analyzing various tracking methods or by integrating vision-based methods

with other sensors [2]. Lee and Park proposed to address this issue in AR environment introducing the usage of an Augmented Foam [3]. A blue foam mock-up is overlaid with a 3D virtual object, which is rendered with the same CAD model used for mock-up production. By hand occlusion correction, inferred by color-based detection of the foam, virtual products and user's hand are seamlessly synthesized. The advantage of the augmented foam is that it is cheap and easy to cut allowing to realize simple and complex shapes. On the other hand, it imposes that for all augmented objects has to be present in the scene the physical counterpart made of foam. A color-based similar approach is discussed by Walairacht et al [4]. They exploited the chroma-key technique to extract only the image of the hands from a blue-screen background merging the image of the real hands and the virtual objects with correct occlusion. Although chroma-key is particularly fast and efficient, it requires the use of a colored background that represents a not feasible solution in many environments. In addition, it does not provide any information about real objects in the scene and their spatial distances. Buchmann et al [5] also handled hand occlusions in augmented reality exploiting marker-based methods to determine the approximate position/orientation of user's hands and, indirectly, their contour to fix the visualization order. The disadvantages are the inconvenience to wear specific gloves featuring fiducials on each finger and the rough level of accuracy in the segmentation of the hand from the background. In the field of medicine, Fischer et al [6] exploited a Phantom tracker and anatomic volumetric models in order to support surgical interventions resolving occlusions of surgery tools. They presented a simple and fast preprocessing pipeline for medical volume datasets which extracts the visual hull volume. The resulting is used for real-time static occlusion handling in their specific AR system, which is based on off-the-shelf medical equipment. Depth/range cameras (e.g. the Kinect by Microsoft) have also been proposed [7][8][9] to provide a real-time updated depth-image of the surrounding world that can be conveniently used to evaluate whether a pixel from the captured environment is closer to the observer than the corresponding rendered pixel of virtual content, or not. This technique can lead to a more accurate result and also enables evaluating distances of real objects in the scene and their inter-occlusions with virtual objects. However, it requires additional hardware (usually an infrared pattern emitter and a dedicated infrared camera) and it should match the field-of-view of the see-through cameras, to works effectively. The generation of a disparity map by using stereo matching techniques [10][11] represent the most suited choice to correctly segment user's hands in AR environments. Results produced by this technique are comparable to the ones form depth cameras without requiring dedicated hardware, which is a central aspect of this study. In our proposal, the disparity map is generated by a belief propagation global algorithm [12] that exploits GPU's highly parallel architecture to speed up required calculations and to provide real-time performance. Some ad-hoc improvements aimed at further reducing the computational cost of the original algorithm are discussed in the following section.

3 System Description

The overall system architecture is shown in Fig. 2. The diagram highlights the main elements in the image-processing pipeline. The user wears a HMD with two embedded cameras enabling stereo vision. Two separated video streams, from left and right

camera respectively, capture the real scene from a different perspective point. On each stream, a simple and fast skin detection technique detects the user's hands in the scene. The binary image is used to apply a vertical crop to the original frame that preserves the region, including the foreground and the background, where the hands appear. On that crop two disparity maps, the one for the real scene captured and the other for the rendered content, are generated by exploiting a stereo-matching with belief propagation technique. The disparity maps are used to estimate the position of the hands in the field of view with regards to the virtual scene. The occlusion correction is achieved by comparing them and combining the result with a skin-based segmentation of the hand. An edge blurring pass is applied to the segmentation in order to smooth the edges of the hand region. The combination of disparity map with blurred color-based segmentation of hands produces a cleaner approximation of the occlusions that can be applied as top-level layer of the augmented streams sent to the HMD displays.

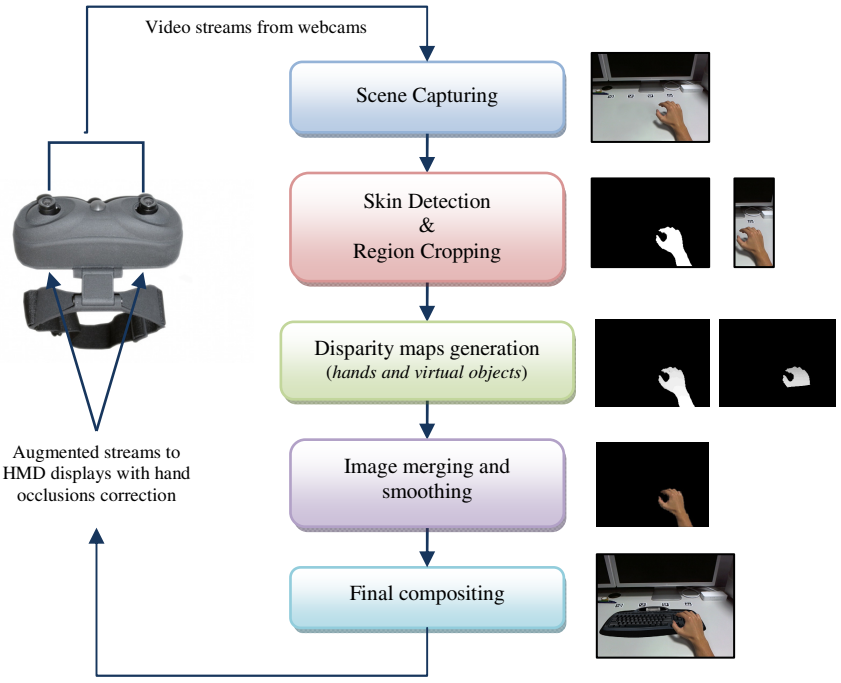

Fig. 2. The overall architecture of the approach proposed

More in detail, the first step consists in capturing the scene observed by the user through the webcams mounted on the HMD. Since the HMD is intended for a stereoscopic vision, the streams from left and right camera capture the scene from a slight different point of view. Each of the two streams is therefore separately augmented by

rendered virtual contents throughout the pipeline. Even though this implies a greater computational cost of the augmenting algorithm, it preserves the binocular vision of human eyes leading to a more reliable augmentation of the scene and the occlusion correction. Fig. 3 shows one frame captured by one of the cameras mounted on the HMD while the user wears it.

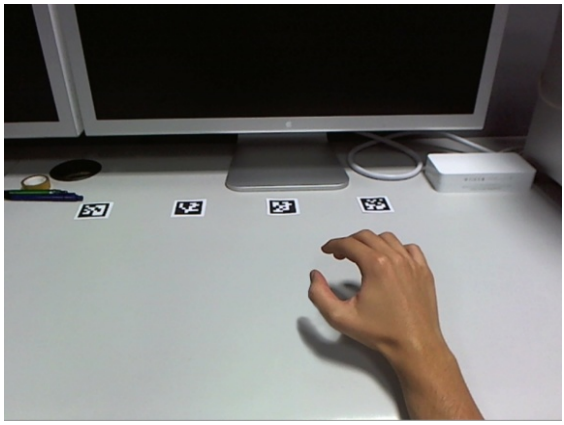

Fig. 3. The scene captured by one of the cameras mounted on the HMD

To keep computational cost of the following steps slow, the frame is properly cropped so that the algorithm can focus the execution only on the relevant portion of the entire frame. The cropping is performed by a simple and fast skin color-based technique. It converts the video frame from the RGB to the HSV color space in order to have a simply way of filtering the color of naked hands by proper ranges of hue and saturation, thus leading to a gray-scale mask. Fast closure operators enable removing little irrelevant blobs in this mask and filling holes (if any) in main closed regions (the hands in our context). Every pixel inside the region boundaries (the hands' contour) is therefore set to full white (see Fig 4a). The intersection of this first mask with the rendered content's alpha channel (see Fig 4b) results in a new mask which limits the region on which the disparity maps of rendered content has to be computed (see Fig 4c). To this aim, stereo matching with belief propagation [12] is therefore performed on these cropped regions.

By processing only a limited region of the whole scene, we manage to reduce the computational costs of this step, which is the most time consuming in the processing pipeline. Firstly the matching costs for each pixel at each disparity level in a certain range (disparity range) are calculated. The matching costs determine the probability of a correct match. Afterwards, the matching costs for all disparity levels can be aggregated within a cross-shape neighborhood window. Basically the loopy belief propagation algorithm first gathers information from a pixel's neighbors and incorporate the information to update the smoothness term between the current pixel and its neighboring pixels, and to iteratively optimize the smoothness term thus resulting in global energy minimization. Each node is assigned to a disparity level and holds its matching costs. The belief (probability) that this disparity is the optimum arises from

the matching costs and the belief values from the neighboring pixels. For each iteration, each node sends its belief value to all four connected nodes. The belief value is the sum of the matching costs and the received belief values. The new belief value is the sum of the actual and the received value and is saved for each direction separately. This is done for each disparity level. Finally, the best match is the one with the lowest belief values defined by a sum over all four directions [13] resulting in the final hand(s) disparity map. The main factor that affects every stereo-matching technique is the number of disparity ranges considered during the matching cost function. The more values are considered the more the disparity map is reliable but, the more the cost increases. Considering our main goal of performing a fast hands occlusion correction, we reduce the number of disparity ranges. We refine the both rough disparity maps obtained by composing it with the corresponding crop of the binary image from skin detection acting as alpha layer (one pass of edge blur allows to smooth the edges of the color-based segmentation).

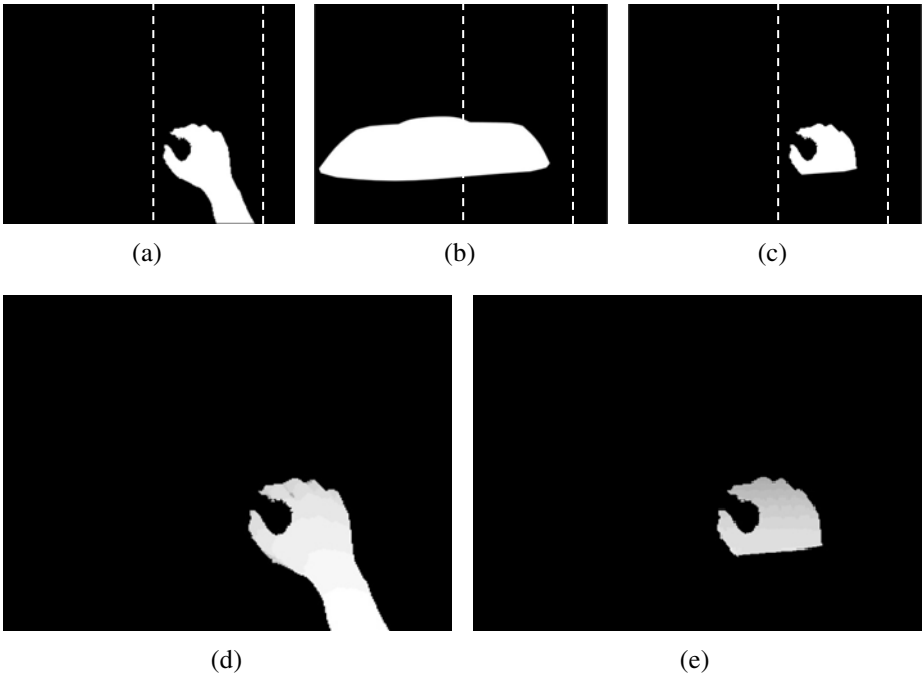

Fig. 4. Skin detection with closure functions refinement (a). Alpha channel of augmented objects rendered onto the video stream (b). Disparity map of user's hand segmented from the scene by the skin color-based detection (d). Disparity map of the crop of the region where augmented virtual contents overlap the hand (e).

Fig. 5. Compositing and final result. The original real background shown in Fig. 3 is composited according to the disparity map of the scene enabling a correct visualization and a meaningful interaction (note that hand's casted shadow is not currently handled by the proposed method).

The result is a smoother segmentation of user's hands (see Fig. 4d) that can be used for final compositing. For what concerns the rendered content, it would be simpler and faster to exploit the accurate depth info contained in the Z-buffer, but matching it coherently to the depth levels encoded in the hand(s) disparity map would be a not trivial task. The final composited frame is obtained by comparing pixel-wise the gray level of the two disparity maps. The pixel whose gray level is lower than its homologous is considered not-visible from the observer's point of view and is discarded. Fig. 5 shows an example of the final result in which the hand of a user interacting in a MR environment is properly composited onto the augmented content.

4 First Experiments

We performed a preliminary experimental trial of the proposed technique in a MR environment on a test-bed featuring an i7 Intel quad_core processor and an Nvidia GTX760 graphic board equipped with 1152 cores and 2 GB of VRAM. The user worn a Trivisio HMD that features stereo capturing by two embedded webcams (752x480 resolution, 6oFPS) and stereo vision by two 800x600 LCD displays.

Even though the technique proposed in this paper exploits time consuming algorithms, it meets the requirements of real-time application because it works only on a fraction of the whole captured scene. In addition, the improvement provided by utilizing graphics hardware acceleration makes possible to combine the time demands of stereo matching with typical marker-based tracking of the user on a stereo video stream.

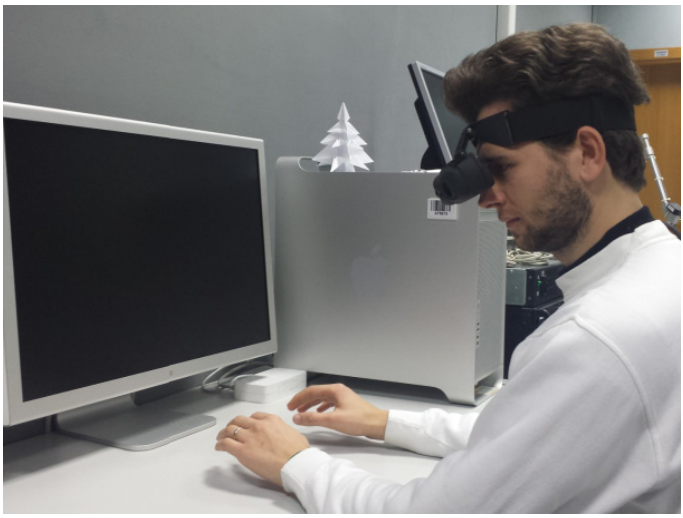

Fig. 6. A user wearing the Trivio HMD during the experimental trial

In Table 1 have been summarized the performance during the experimental session. In particular, the table shows the frame per second achieved by the solution proposed when the disparity maps are generated for 16 and 32 ranges of disparity values. During the experimental trial the user is free to move his/her hands thus implying a size of the crop of the scene that varies over time. We observed that, in normal condition of interaction, the number of pixels of user's hand covers about 1/8 to 1/6 of the whole scene for over 60% of the experimental session. When the distance between user's hand and the point of view results shorter, e.g., the user brings his/her hands closer to the cameras, the stereo matching works on a wide crop of the scene leading to a drop in performances to the limit of a smooth real-time rendering. Future improvement of our method will take into account such issue providing an adaptive amount of disparity levels to consider during the matching cost function.

Table 1. Frame per second recorded during the experimental trial at different size of the cropping region of the scene

Crop size *(fraction of the whole scene, which consists of 360960 pixels (752x480))*	**# disparity levels**	
	16	**32**
	FPS	**FPS**
< 1/8 (~ 45120 pixels)	56	48
< 1/6 (~ 60160 pixels)	42	33
< 1/4 (~ 90240 pixels)	31	22
< 1/2 (~ 180480pixels)	25	12

5 Conclusions

We presented a technique to address hand occlusion in real time when interacting in a Mixed Reality environment. The approach, designed around the stereo matching belief propagation algorithm and ad-hoc enhancements, meet the main design requirements of finer segmentation of user's hands, distance dependant occlusions, and natural interaction. Binocular scene capture and stereo rendering of virtual contents improve depth perception of real environment while stereo matching allows to estimate the distance from the observer and real/virtual objects in the scene.

The subjective system evaluation, performed by testers in an experimental environment, highlights the potential of the proposed approach, even though issues related to the hardware used (the reduced HMD's resolution/field-of-view, rough hands segmentation under rapid user's movements) have to be more carefully addressed to achieve a robust system behavior. In particular, the generation of the disparity maps for the hands when they occupy the most of the framed scene. Even though these enhancements are inherently effective only on naked hand region, even not-naked arms can be reasonably handled by the disparity info alone. As a further development of this technique, besides improving the quality of the disparity map, we are currently trying to address the incorrect visualization of the shadows casted by the hands when they should be projected onto a virtual object.

According to first users evaluations, the combination of augmentation and the detection of occlusions worked well, providing an intuitive interaction paradigm suited to a wide range of application contexts. For these reasons we expect to improve the solution proposed in this paper to resolve, by stereo-matching techniques, occlusion issues in wide environment where people are free to move around and occlude big one to one scale virtual augmenting objects.

References

1. Furmanski, C., Azuma, R., Daily, M.: Augmented-reality visualizations guided by cognition: Perceptual heuristics for combining visible and obscured information. In: Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR 2002), pp. 215–320. IEEE (2002)
2. Shah, M.M., Arshad, H., Sulaiman, R.: Occlusion in augmented reality. In: Proceedings of the 8th International Conference on Information Science and Digital Content Technology (ICIDT 2012), pp. 372–378. IEEE (2012)
3. Lee, W., Park, J.: Augmented foam: a tangible augmented reality for product design. In: Proceedings of the Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality, pp. 106–109. IEEE (2005)
4. Walairacht, S., Yamada, K., Hasegawa, S., Koike, Y., Sato, M.: 4+ 4 fingers manipulating virtual objects in mixed-reality environment. Presence: Teleoperators and Virtual Environments. MIT Press Journal, 134–143 (2002)
5. Buchmann, V., Violich, S., Billinghurst, M., Cockburn, A.: FingARtips: Gesture Based Direct Manipulation in Augmented Reality. In: Proceedings of the 2nd International Conference on Computer Graphics and Interactive Techniques (GRAPHITE 2004), pp. 212–221. ACM (2004)

6. Fischer, J., Bartz, D., Straßer, W.: Occlusion handling for medical augmented reality using a volumetric phantom model. In: Proceedings of the ACM symposium on Virtual reality software and technology (2004), pp. 174–177. ACM (2004)
7. Corbett-Davies, S., Dunser, A., Green, R., Clark, A.: An Advanced Interaction Framework for Augmented Reality Based Exposure Treatment. In: IEEE Virtual Reality (VR 2013), pp. 19–22. IEEE (2013)
8. Gordon, G., Billinghurst, M., Bell, M., Woodfill, J., Kowalik, B., Erendi, A., Tilander, J.: The use of dense stereo range data in augmented reality. In: Proceedings of the 1st International Symposium on Mixed and Augmented Reality (2002), p. 14–23. IEEE Computer Society (2002)
9. Seo, D.W., Lee, J.Y.: Direct hand touchable interactions in augmented reality environments for natural and intuitive user experiences. Expert Systems with Applications 40(9), 3784–3793 (2013)
10. Kanade, T., Okutomi, M.: A stereo matching algorithm with an adaptive window: Theory and experiment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 920–932 (1994)
11. Medioni, G., Nevatia, R.: Segment-based stereo matching. In: Computer Vision, Graphics, and Image Processing, pp. 2–18 (1985)
12. Yang, Q., Wang, L., Yang, R., Wang, S., Liao, M.: NisterD.: Real-time global stereo matching using hierarchical belief propagation, in: The British Machine Vision Conference, pp. 989–998 (2006)
13. Humenberger, M., Zinner, C., Weber, M., Kubinger, W., Vincze, M.: A fast stereo matching algorithm suitable for embedded real-time systems. Computer Vision and Image Understanding 114(11), 1180–1202 (2010)