

Providing the best Third-Person Perspective to a Video-Through HMD

Patrick Salamin, Daniel Thalmann
EPFL VRLab

Lausanne, Switzerland
Patrick.Salamin@epfl.ch, Daniel.Thalmann@epfl.ch

Abstract— This paper describes a system with several fixed cameras that provide Third-Person Perspective (3PP) to the video-through HMD of the user. Once done, our system uses an algorithm to make intelligent switches between the different cameras in order to provide the best view to the user. User comfort does not suffer from the changes of perspective; some of the users even play at forcing the perspective change during the experiment. Working with an augmented environment larger than a desktop seems to be very promising for future researches in this domain.

Keywords-component; Artificial, augmented, and virtual realities

I. INTRODUCTION

During our augmented reality experiments with a video-through Head-Mounted Display (HMD), we always try to provide the best view to the user. As shown in [1], the best perspective depends on the performed action: first-person perspective (1PP) for manipulations and third-person perspective (3PP) for moving actions.

In order to propose a 3PP to the user, we need at least a second camera that follows him/her when he/she moves in the environment. Moreover, within the framework of augmented reality, it has been proven that fixed cameras avoid lots of registration issues [2][3]. Finally, if there are multiple cameras, occlusion problems can also be reduced [4].

Based on the previous researches, we propose in this paper a system with several fixed cameras combined with a mobile one on the user to provide the different perspectives to the user who wears a video-through HMD. We such a kind of system, we should have better results and provide a better comfort in almost every simulation with augmented reality.

Moreover, working with several cameras allows us to enlarge the work area to a building (or at least two rooms in this paper). In order to manage this system, we implemented an “intelligent switch” that chooses the “best view” depending on the user context (location, movement, performed action, and occlusions).

We will first have a brief overview of the related works that lead us to this system. We then present the experimentation with the equipment, participants, and procedure. Finally, we conclude with the results and their discussion.

Frédéric Vexo
EPFL VRLab

Lausanne, Switzerland
Frederic.Vexo@epfl.ch

II. RELATED WORK

Within the framework of augmented reality, there exist two main approaches for tracking and positioning [5]. One of them uses magnetic sensor such as the MotionStar. Unfortunately, this approach considerably reduces the work area to few square meters and may lead to distortions while using magnetic equipment. The other approach uses vision-based techniques. The main issues of this method, are the markers occlusion and the registration [6][5]

A well-known way to reduce the marker occlusions consists in working with multiple cameras [7][8]. Indeed, even if the marker is hidden for one camera, the other cameras (due to their strategic position) should still be able to detect the marker. The second issue, the registration, is one of the main issues in AR software and may be a source of motion sickness [9]. Notice that it has been proven in [3] that the use of a fixed camera considerably reduces this lack.

The researches cited above propose a system providing for augmented reality experiments into a static user working on a very small area like a desktop in indoor, or for a user carrying a camera outdoor [10][11]. Unfortunately, in this last case, users are carrying a camera, which is then not at a fixed location. And as shown above, this induces registration issues.

We will then use several fixed cameras to provide 3PP when the user is moving, and 1PP for the fine manipulation with a camera coupled on the users’ HMD. We change of perspective between moving and static actions for the user comfort [1]. We need then to be aware of the user context: localization, displacements, and the direction of his/her gaze. Concerning the geo-localization, we use the application developed by Hopmann et al. that uses Wi-Fi to know the current room in which the user is [12]. This gives us the set of camera available in the current room. Once done, we use our “best view” algorithm described in the following chapter to know which camera to activate.

III. DESCRIPTION OF THE SYSTEM

The goal of our system is to provide the “best view” to a user who can move in several rooms and manipulate objects in augmented reality. Based on previous studies of Salamin et al. [1][13], we know that these two actions require different perspective: third-person perspective (3PP) for navigation tasks, respectively first-person perspective (1PP) while manipulating an object with the hands. In our case,

instead of having a camera that follows the user for the 3PP, we decided to have multiple fixed cameras. Consequently, the user will not need to matter about collisions of a cumbersome backpack with wall, ceiling, doors, etc.

On the other hand, as there are multiple cameras, we need a system that will automatically detect which camera needs to be activated for the user best view. For this simulation, we will work in an area of two adjacent rooms in which we already put three cameras at strategic positions (see Figure 1).

Our system is composed of two networks (one per room). There are three computers (one server and two clients) linked to a static camera in each network and a last one connected to the notebook carried by the user. Each computer gets the video flow of “its” camera to perform two actions: sending the video flow onto its network via RTP with Java Media Framework (JMF); detecting with the help of ARToolKit if the markers visibility to inform the server of the network. The server chooses then to which video flow the user client must read and send it to him/her when he/she connects to the network.

Our system then considers that there is two networks of three cameras (one network per room). We then first have to localize the user, i.e. in which room he/she is currently. Once done, depending on the visibility of the markers and on the displacement of the user, our system will choose which video stream to provide to the user.

Here are some described cases that may happen during the simulation (the decision schema is presented in the Figure 2).

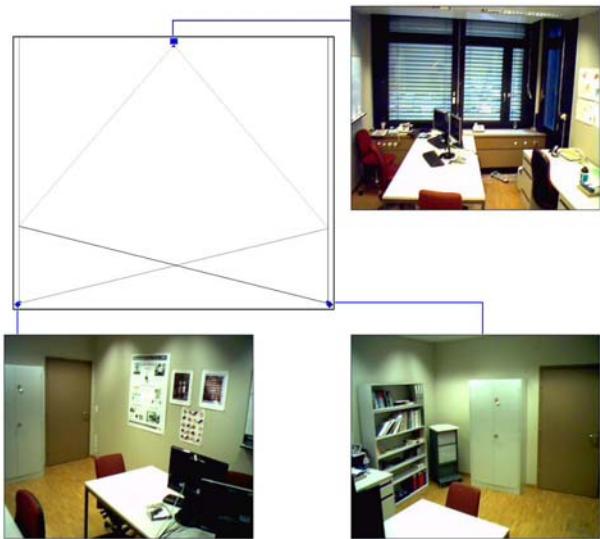


Figure 1. Schema of the cameras (in blue) disposition in the room in order to cover the whole space (their angle of view is represented by the line in different colors). Each camera is coupled with a picture representing a snapshot of its video flow sent onto the network.

Server state machine

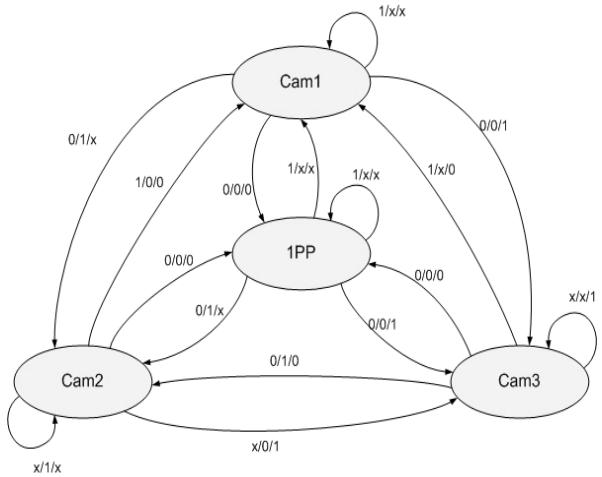


Figure 2. Presentation of the schema defining the switch of perspective in a room. The states correspond to the active camera (providing the video flow to the user) and the values on the link between them represent if they “can see” the ARToolKit marker (set to ‘1’) or not (value set to ‘0’). In this schema we can see that we privileged 3PP, and especially the first camera. (Notice that the value ‘X’ means that any value can fit).

If the user is static (no movement detected by the inertial tracker on the user directly leads to choose 1PP) with the augmented object in the hands, our system will propose to him/her the 1PP from the spy-camera video flow coupled with the HMD. In another situation, the user could be moving (movement detected leads to propose 3PP) towards an augmented object. In this case, our system will check which fixed camera can see the object and the user. Notice that there is a preference for a camera located behind the user in order to avoid “mirror effect” that may introduce biases like the right-left inversions.

Obviously, our system must avoid changing too many times the chosen perspective. In-deed we hypothesize that each camera change would perturb the user, because he/she would then need a few seconds to relocate him-/herself in the environment. Frequent perspective would then make any action impossible for the user. This is why we try to keep a view as long as it is possible, i.e. augmented object and user are visible in the video flow during the experimentation described in the next chapter.

IV. EXPERIMENTATION

We tested our system with 12 naive and voluntary participants (10 males and 2 females). They were all between 20 and 35 years old. The equipment used during the experimentation and its protocol are described in the following sections.

A. 4.1 Equipment

During the experimentation, we use static and mobile equipment. We first need a Wi-Fi antenna on the user to detect in which room he/she is. Once our system knows the

location of the user, it automatically connects to the network of the room (one Wi-Fi network per room) to access the most interesting video flow. Indeed, all the cameras in a room are connected to computer in the room network and send the video flows to the router.

As written above, we use three static cameras per room (Figure 1) and a last mobile one on the user, which means seven cameras. As we need to cover the full room with only three cameras, we used the Trust Wide Angle Live WB-6200p that provide an angle of view 45% bigger than a common webcam with a resolution of 1280 per 1024 pixels at 30 fps and a focal length at 50mm. We can then distinguish ARToolKit markers amongst opposite wall with a wide field of view from each of the cameras. The number of three cameras per room was defined in an empiric approach. Indeed, we need at least three cameras to cover the whole space in the room.



Figure 3. Presentation of the user fully equipped for the simulation with an HMD coupled with a webcam for the 1PP, a Wi-Fi USB adapter and an inertial tracker at the belt. They are all connected to the notebook in the backpack.

Once we completely cover the room space, we treat the video flow of each camera with ARToolKit in order to detect if there are markers in the field; if it is the case, we get their current distance from the camera (in order to detect the user movements). If the user is immobile, the system detects it via the XSens inertial tracker MTx (connected to the notebook

on the user) and proposes the 1PP view (camera coupled with the HMD) to the user 3. We use a HMD Sony Glasstron PLM-S700E with a resolution of 800 per 600 at 60Hz (Figure 3). Once the user is moving, our system directly switches to another video flow to provide to the user (Figure 1).

B. 4.2 Protocol

There are four main steps in the experimentation described in this paper. As written above, our working space is extended to two rooms. This means that one step will be to move from a room into the other one. This is a critical step, because lots of operations need to be performed in a little of time. First, the system must detect that the user is leaving the room (geo-tracking with Wi-Fi signal variation detected by the Wi-Fi adapter). Indeed, as the user is leaving a room, the notebook he/she wears must be disconnected from the current network to immediately reconnect to the Wi-Fi network of the other room. Once done, the system has to detect if the user is still moving. If it is not the case, it proposes the 1PP to the user. In the other case, it must detect which is the most interesting camera for the user.

The second step of this experiment is to stay in a room with no displacement and to turn on oneself. This action will help to discover the room with a more common perspective: the 1PP.

The third step consists in walking in the room. The walk direction can change several times and can be randomly chosen, or it can be to reach an augmented object. In any of these situations, our system has to provide the video flow of one of the fixed cameras of the room to the user. This means first that 3PP will be proposed to the user. But the choice of camera is neither random nor trivial. Indeed, our system must detect the direction of the user's walk to find a camera (if possible) behind him and able to "see" the user and the augmented object.

The last step is the manipulation of an augmented object. This action can be performed while walking or staying, which means a change of perspective (3PP, respectively 1PP).

All the four steps cited above are performed several times in different orders during the experimentation. Indeed, the user must first get more familiar with the environment (the two rooms), and the system. Then he/she must find the augmented objects and can manipulate them. The experimentation usually lasts around twenty minutes.

V. QUESTIONNAIRE

Once the experiment performed, we proposed a SUMI-like (Software Usability Measurement Inventory) questionnaire [14] to the users. Indeed, several types of validity studies [15][16][17] have already been conducted with SUMI, whose one of them concerns laboratory-based studies (carried out in the Human Factors Research Group). This questionnaire is composed of two parts that we describe here.

The first part is composed of questions about the user profile like the age, gender, but also if the user is used to work with computer, within VR environment and VR Stuff.

This part concludes with questions about the training for using the system (availability and length) and if the time to use the system was also adequate.

The second part of the questionnaire is composed of fifty statements. The user must answer to all of them by marking one of the three proposed boxes labeled: “Disagree”, “Undecided”, and “Agree”. It is also firstly noticed that marking the “Undecided” boxes means that the user cannot make up his/her mind, or that the statement has no relevance to the software or the situation. Secondly, it is added that marking the “Disagree” or “agree” boxes does not necessarily indicate a strong disagreement (respectively agreement) but only a general feeling most of the time.

The questions of this second part concern various topics: responsiveness of the software, quality of proposed instructions, global satisfaction about the software, possible improvements, intuitiveness, and attractiveness of the software. In the next section, we will analyze the users' answers and their behavior during the experiment.

A. Results

Globally, most of users enjoyed the system. Every step was performed by every user, even if some of them needed more time to adapt to the system. They walk a lot in the rooms looking for augmented objects and trying the perspectives. We will now first present the users' behavior during the presentation and then the questionnaire proposed to them after the experiment.

B. Behavioral Analysis

Once they move from a room to another one, the system disconnects the user notebook from the current network to connect it to the second one. This operation requires few seconds. In order not to bring the user in the blackout during this time, the system switches to 1PP until new network connection is fully performed (even if the user is moving). After the connection, the system proposes at once the best 3PP to the user.

The 1PP is only provided to the user within the case cited above and while he/she is staying immobile. In both cases, the switch from 3PP to 1PP never seemed to affect the user (e.g. fall down, loss of stability, collision with objects).

While walking in a room, it may happen that the system decides to change of 3PP provided to the user. Based on our start hypothesis, the system tries to keep as long as possible the same perspective to avoid possible issue due such a kind of switch like the loss of reference points. Anyway, after few changes of perspective, the user seems not to be too much perturbed by these switches.

Finally, the last step concerned the augmented object manipulation. This step can be divided into two cases, depending on the user movements. If the user is static (no displacements), the 1PP is provided to the user during the whole manipulation. On the other hand, if the user is moving while manipulating the augmented object, the system proposes to him one of the 3PP. In this case, depending on his orientation and direction movements, the system may switch from a 3PP by a room camera to another one in order

to keep the ARToolKit marker visible. But once again, this artifact (the switch of perspective) does not seem to disturb too much.

C. Questionnaire trends

Our adapted SUMI questionnaire was filled by every participant. Its first part, concerning the users' profile, reveal us that twenty minutes for training was widely enough all the participants but one. Indeed, during the experiment with this person, due to connection to Wi-Fi network issues that were maybe due to the overlap of the several Wi-Fi networks in our building. Globally, after twenty minutes, the users were well-trained and fifteen of them wanted to continue the experiment. Finally, length of the experiment was also considered as adequate.

The questions of the second part reveal us that our software is very accurate and fast to leave the perspective current when the augmented object disappears. But it also informs us that the reconnection to another video flow (couple of seconds) can be very perturbing at the beginning. People recognized that they always could see the augmented object -except while they were immobile (which means using 1PP) and looking at somewhere else. They seem a bit afraid at the beginning with the perspectives changing but after few minutes (around five), they already were dealing with it, and even playing/testing it by voluntarily occluding the augmented object to induce a perspective switch.

Our system was then considered as very attractive and intuitive enough, even if improvements can be done for a future version.

VI. CONCLUSION AND FURTHER WORKS

The obtained results confirm our hypotheses. User comfort does not suffer from the changes of perspective; some of the users even play at forcing the perspective change during the experiment. Working with an augmented environment larger than a desktop seems to be very promising for future researches in this domain. And globally, all the users interested in trying our system were satisfied.

Three participants, who already took part to previous 3PP experiments with a camera coupled to a backpack on their body, especially appreciate the change of perspective that avoids occlusions. They also appreciate the system preference for 3PP views when they were moving. It is also interesting to notice that none of them remarked that we also favored one of the three cameras in each room. Anyway, we think the camera should have the same importance in the room and we will change our decision graph for future experiments.

Another improvement would be to reduce the time needed for the perspectives switch and improve the image quality. Indeed, even if the video flows are compressed, sending three streams onto an access point during the whole experiment can saturate it. A solution to this problem would be to send only the video flow chosen by the system to the access point. This would allow us to send video flows with a higher resolution onto the network. In counterpart, more time would be needed for a perspective switch because three operation (for the system: stop sending current stream, start

sending new stream; and for the user: connect to the new video stream) would be required instead of only one: connect to the new video flow (already streamed over the network).

ACKNOWLEDGMENTS

This research has been partially supported by the European Coordination Action: FOCUS K3D (<http://www.focusk3d.eu>).

REFERENCES

- [1] P. Salamin, F. Vexo, and D. Thalmann. The benefits of third-person perspective in virtual and augmented reality? In ACM Symposium on Virtual Reality Software and Technology (VRST '06), pages 27–30, 2006.
- [2] Ismail Haritaoglu, David Harwood, and Larry S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:809–830, 2000.
- [3] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multi-sensor surveillance. In Proceedings of the IEEE, volume 89, pages 1456–1477, October 2001.
- [4] Anurag Mittal, Larry, and S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. In *International Journal of Computer Vision*, pages 189–203, 2003.
- [5] Ronald Azuma. Overview of augmented reality. In SIGGRAPH '04: ACM SIGGRAPH 2004 Course Notes, page 26, New York, NY, USA, 2004. ACM.
- [6] Ronald Azuma, Yohan Baillot, Reinhold Behringer, Steven Feiner, Simon Julier, and Blair MacIntyre. Recent advances in augmented reality. *IEEE Comput. Graph. Appl.*, 21(6):34–47, 2001.
- [7] Shiloh L. Dockstader and A. Murat Tekalp. Multiple camera tracking of interacting and occluded human motion. In Proceedings of the IEEE, pages 1441–1455, 2001
- [8] Kang J., Cohen I., and Medioni G. Continuous tracking within and across camera streams. In Proc. IEEE Conf. Computer Vision and Pattern Recognition, pages 267– 272, 2003.
- [9] Randy Pausch, Thomas Crea, and Matthew Conway. A literature survey for virtual environments: military flight simulator visual systems and simulator sickness. *Presence: Teleoper. Virtual Environ.*, 1(3):344–363, 1992.
- [10] Tobias Hllerer, Steven Feiner, Tachio Terauchi, and Gus Rashid. Exploring mars: developing indoor and outdoor user interfaces to a mobile augmented reality system. *Computers and Graphics*, 23:779–785, 1999.
- [11] Adrian David Cheok, Kok Hwee Goh, Wei Liu, Farzam Farbiz, Siew Wan Fong, Sze Lee Teo, Yu Li, and Xubo Yang. Human pacman: a mobile, wide-area entertainment system based on physical, social, and ubiquitous computing. *Personal Ubiquitous Comput.*, 8(2):71–81, 2004.
- [12] Mathieu Hopmann, Daniel Thalmann, and Frédéric Vexo. ed 'Thanks to geolocalized remote control : the sound will follow. In Cyberwords 2008, pages 371–376, 2008.
- [13] Patrick Salamin, Daniel Thalmann, and Frédéric Vexo. Improved third-person perspective: a solution reducing occlusion of the 3pp? In VRCAI 2008, the 7th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry, 2008.
- [14] R. McSweeney. Sumi – a psychometric approach to software evaluation. Unpublished MA (Qual) thesis in Applied Psychology, University College Cork, Ireland, 1992.
- [15] Susannah Ravden and Graham Johnson. Evaluating usability of human-computer interfaces: a practical method. Halsted Press, New York, NY, USA, 1989.
- [16] Carol Stoak Saunders and Jack William Jones. Measuring performance of the information systems function. *J. Manage. Inf. Syst.*, 8(4):63–82, 1992.
- [17] G. Wong and R. Rengger. The validity of questionnaires designed to measure user-satisfaction of computer systems. Technical report, National Physical Laboratory report DITC 169/90, Teddington, Middx., UK, 1990.