Multimodal Fusion: Gesture and Speech Input in Augmented Reality Environment

Ajune Wanis Ismail and Mohd Shahrizal Sunar

UTM-IRDA Digital Media Centre
MaGIC-X (Media and Games Innovation Centre of Excellence)
Universiti Teknologi Malaysia, 81310 Skudai Johor, Malaysia
{Ajune,shahrizal}@utm.my

Abstract. Augmented Reality (AR) has the capability to interact with the virtual objects and physical objects simultaneously since it combines the real world with virtual world seamlessly. However, most AR interface applies conventional Virtual Reality (VR) interaction techniques without modification. In this paper we explore the multimodal fusion for AR with speech and hand gesture input. Multimodal fusion enables users to interact with computers through various input modalities like speech, gesture, and eye gaze. At the first stage to propose the multimodal interaction, the input modalities are decided to be selected before be integrated in an interface. The paper presents several related works about to recap the multimodal approaches until it recently has been one of the research trends in AR. It presents the assorted existing works in multimodal for VR and AR. In AR, multimodal considers as the solution to improve the interaction between the virtual and physical entities. It is an ideal interaction technique for AR applications since AR supports interactions in real and virtual worlds in the real-time. This paper describes the recent studies in AR developments that appeal gesture and speech inputs. It looks into multimodal fusion and its developments, followed by the conclusion.This paper will give a guideline on multimodal fusion on how to integrate the gesture and speech inputs in AR environment.

Keywords: Augmented Reality, Multimodal Fusion, Gesture and Speech Input, User Interaction.

1 Introduction

Augmented reality (AR) environment is when the real world and virtual world objects are presented together on a single display [1]. Recently, the AR applications have shown that AR interfaces can enable a person to interact with the real world in ways never before possible [2]. Recently, interaction is a crucial key inVirtual Reality (VR) and AR research area. Traditionally, keyboards and mice are common intermediary between human and machine, in most of interfaces. However, the bottleneck occurs rely on user interaction due to the unnaturalness of the interaction [3]. Many interaction methods and technologies have been proposed towards attempting to

S. Phon-Amnuaisuk and T.-W. Au (eds.), *Computational Intelligence in Information Systems,*
Advances in Intelligent Systems and Computing 331, DOI: 10.1007/978-3-319-13153-5_24

eliminate this bottleneck. By improving the ways of interacting with computers naturally and intuitively, people started to explore the human forms such speech, and gesture recognition [4]. Human gestures come in many forms, such as, hand gestures, general body gestures and facial expressions [5].The human factors need to be addressed before moving to integrate the modalities [6]. That is motivating people to study and explore multimodal interaction [7]. When it comes to unimodal, however, we usually use only one interface device at a time like typing, clicking the mouse button, speaking, or pointing with a magnetic wand. The ease with which this unimodal interaction allows us to convey our intent to the computer is far from satisfactory. The practical reason can lead to consider the use of multimodal interaction [8]. The task can be more practical and convenient with multimodal inputs. The interaction techniques that combine hand gestures provide a separate complementary modality to speech [9]. Successful embodiment of these modalities into an interface noticeable with the advances in computing and communication has the potential of easing the bottleneck in either VR or AR interfaces [8]. It has also become increasingly evident that the difficulties encountered in the analysis and interpretation of individual modalities may be overcome by integrating them into a multimodal interface. Modalities such as speech, vision-based gesture recognition, eye and facial recognition.

Another drawback of current advanced unimodal is that it lacks robustness and accuracy. Whether they use a stylus or a glove or are vision based, they are still constrained to the recognition of few predefined hand movements and are burdened by cables or strict requirements on background and camera placement [10]. However, concurrent use of two or more interaction modalities may loosen the strict restrictions needed for accurate and robust interaction with the individual modes. For instance, spoken words can affirm gestural commands, and gestures can disambiguate noisy speech. Gestures that complement speech, on the other hand, carry a complete communicational message only if they are interpreted together with speech and, possibly, gaze. The use of such multimodal messages can help reduce the complexity and increase the naturalness of the multimodal interface [11].

In the wide studies in AR area, at the early stage people however pay less attention on porting these modalities into AR. One of the most important research areas in AR is creating appropriate interaction techniques for AR applications to allow users to seamlessly interact with virtual content [3]. Many different interaction methods have been explored including natural gestures [8] and they started to look thoroughly into multimodal fusion. In multimodal interaction, users invite the hand gesture and speech input to imitate manipulation tasks in the real world either direct or indirect ways [10]. Thus, in recent years, there has been a tremendous interest in introducing various gesture and speech input into AR that will potentially resolve the user interaction limitation in AR environment. In AR, multimodal considers as the solution to improve the interaction between the virtual and physical entities [9]. It is an ideal interaction technique for AR applications since AR supports interactions in physical and virtual worlds in the real-time. Therefore, it has recently given rise to a number of novel interaction modalities. The multimodal fusion relies on unobtrusive input modalities and natural user interactions. It focuses on providing an intuitive environment, which supports natural interaction with virtual objects while sustaining accessible real tasks and interaction mechanisms. Therefore this paper will discuss the progresses in multimodal fusion in AR involves with gesture and speech input for interaction. The paper presents a few sections to detail out the related works about to recap the multimodal approaches until it recently has been one of the research trends in AR. It describes the recent studies in AR developments that appeal gesture and speech inputs for multimodal.

2 Multimodal: VR vs. AR

One of the first multimodal HCI systems can be accredited to Bolt [11], the fusion spoken input and magnetically tracked 3D hand gestures during the integration architecture. The system was used for simple management of a limited set of virtual objects such as selection of objects, modification of object properties, and object relocation. Even though the naturalness of the interaction was hindered by the limitations of the technology at that time, "*Put-That-There*" has remained the inspiration of all modern multimodal interfaces. The rest of this section focuses on some of its descendants. *QuickSet* [12] is a multimodal interface for control of military simulations using handheld PDA's. It incorporates voice and pen gestures as the modes of interaction. This interface belongs to the class of decision level fusers. It follows the [13] with recognition of pen gestures sensed through the PDA is conducted by the gesture agent.

In the past, multimodal interaction has been used not only for 2D user interfaces but also for interacting with 3D virtual contents. Chu et al. [14] showed how multimodal input can be used in VR applications to interact with virtual objects while Krum et al. [15] used it to navigate a virtual world. Laviola et al. [16] developed a prototype multimodal tool for scientific visualization in an immersive virtual environment; a user could not only interact with virtual objects but also navigate through the VR scene by using gesture input from the pinch gloves and triggering corresponding speech input. Wang [17] proposed a multimodal interface with gaze, 3D controller, voicekey and keyboard to select and manipulate the virtual object in the desktop VR environment. As shown in Fig. 2, the Fröhlich [18] meant to chain the multimodal interaction and immersive CAD systems to produce a generic demo for virtual prototyping based on VR technology. Multimodal interaction is concerned with the gesture hand recognition and speech input to drive the modifications of a 3Dvisualization scene. Meanwhile, Immersive CAD is more concerned with the design, exploration and assessment of virtual prototypes using VR simulations. The virtual prototypes are displayed in realistic size like a CAVE with multiple projection system that simplifies on both manipulative gestures for interaction with close and distant virtual parts.

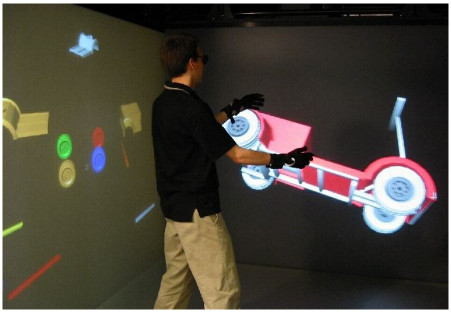

Fig. 1. Multimodal in VR prototyping using gesture and speech [18]

Interactive systems featuring multimodal interfaces are becoming widespread. They now cover many different application domains and support a wide variety of users in the performance of their tasks including in AR. Previously AR interface uses general VR interaction techniques, for example a data glove, without modifications. Adopting VR interaction techniques yield gaps between the virtual environment and real world because they only consider interaction techniques useful in VR environments. The functions the hand recognition interface provides limited and users have to wear a marker or to have a fixed hand posture [19]. The rise of AR interfaces development with these issues lead the experts to explore multimodal interaction in AR. Using speech to provide an additional input modality to the gesture in AR interface overcomes the limitations of gesture input alone.There has been some earlier work in applying multimodal in AR applications. Kaiser et al. created *SenseShape*[20] as shown in Fig.3, a multimodal AR interface in which hand to provide visual information about interaction with augmented or virtual objects and speech to provide information where the user wanted to move the object, by using words such as "*this*" or "*that*". However Kaiser et al. [20] also did not conduct user studies to measure the effectiveness of their system. A user must wear a data glove to detect the hand gestures for interaction with objects.

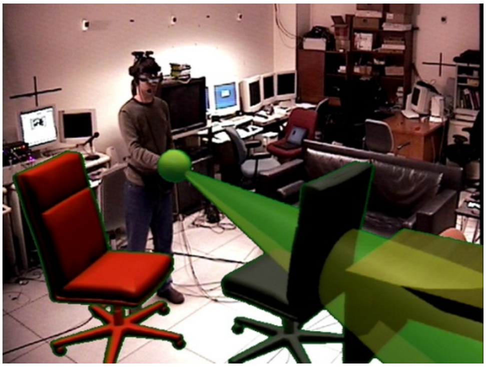

Fig. 2. SenseShape a multimodal AR interface in which hand to provide visual information [20]

Heidemann et al. [21] developed a multimodal interface for information retrieval in AR. Hand gestures with speech are adopted to move between menu options. However, the speech input was used to select the menu item that the user wanted to choose, in the same way a mouse did; thus, the system did not use multimodal input fully. Irawati et al., [22] have extended the VOMAR [23] project into multimodal interaction and they verified that combined multimodal speech and paddle gesture input was more accurate than using one modality alone. However, the system could not provide a natural gesture interface for users, and required the use of a paddle with computer vision tracking patterns on it.Lee et al. [24] developed AR multimodal interface with 3D hand vision-based recognition to precise the hand gesture recognition for interaction in AR. They develop an AR multimodal system that allows us to combine gesture and speech input with a multimodal fusion architecture that merges the two different input modalities in a natural way [22]. As presented in Neumann [25] has developed the multimodal AR interface able to remove, manipulate or add communicatively relevant multimodal information in real-time. By using an AR interception proposed by [26] and methodology explored by [27]. Pitsch et al. [28] developed a tool for linguistic studies. They built on the psycholinguistic tradition of experimenting with communicational parameters. Dierker et al. [26] proposed a prototype AR as a novel methodology to investigate human to human interaction within collaborative tasks as shown in Fig. 4. Their goal is to facilitate the recording and analysis of multimodal interaction.

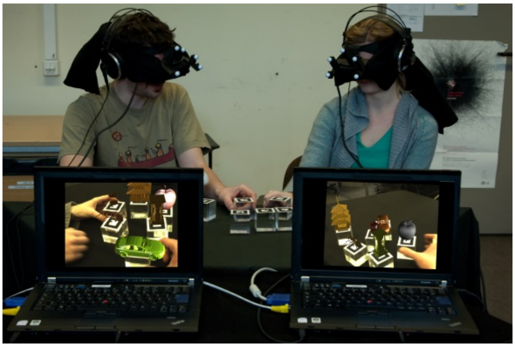

Fig. 3. AR Prototype developed to analysis the multimodal interaction in AR [26]

3 Multimodal Fusion Levels

The section before we have discussed on multimodal in VR interfaces against the AR interfaces. This section will explain multimodal fusion levels in AR. Generally, data fusion methods are divided in three main categories: first fusion which happens at features levels;second fusion which concerns the intermediate decisions fusion and lastly is thehybrid fusion which is a mixture of the two modalities.During the multimodal fusion, the question mainly comes forward is *why* to integrate or combine these modalities input. *What* are they, the appropriate modalities that are going to integrate respectively? Next, once the desired modalities are selected, need to be addressed on *when* and *how* to combine them.

3.1 Decision Levels: *Why* and *What*

Multimodal interactive systems enable users to interact with computers through various input modalities like speech, gesture, and eye gaze. Meanwhile the output channels such as text, graphics, sound, avatars, and probably the synthesized speech. At the first stage to propose the multimodal interaction. The various input modalities like speech, gesture, and eye gaze are decided to be selected before be integrated in an interface. As far as what the modalities are concerned to be selected, the work in this phase is identifying the issues raised by unimodal interfaces and its' limitations. The reasons why multimodal considered as a greater option to improve the burdensome and limitation, remains at a very high level of abstraction more focusing on the identification of problems rather than proposing solutions.

In AR, multimodal considers as the solution to improve the interaction between the virtual and physical entities. It is an ideal interaction technique for AR applications since AR supports interactions in real and virtual worlds at the same time. The most critical concerns associated with multimodal at this decision level are on cases or combination of events that lead to clearly defining what the appropriate modalities are. Machine learning has been already applied to multimodal interfaces, mainly modality recognition like speech and gesture recognition. Multimodal interface is type of user interface which does not only beneficial for enhanced accessibility, but also its usability for greater convenience. For instance, the natural input mode recognition as well as flexibility when the adaptation to context of use, to tasks or to users' preferred interaction modalities. The goal would be to define the interfaces and its fusion that are reliable and usable. Multimodal fusion is commonly known as integration stages for multiple modalities, sometimes also referred to as the fusion engine. It soon will be detailed out in the next section.

3.2 Integration Levels: *When* and *How*

The fusion is the key technical challenge for multimodal interaction systems. In general, the meanings of input streams can vary according to context, task, user, and time. Modalities with very different characteristics for instance, speech and eye gaze, facial expression and haptics input, touch-based gesture, they may not have obvious points of similarity and easy ways to combine. Perhaps the most challenging aspect is the temporal dimension. Different modalities may have different temporal constraints and different signal and semantic endurance. Some modalities such as gestures provide information at sparse, discrete points in time while others generate continuous but less time-specific output like the affect. Some modal combinations are intended to be interpreted in parallel, which others may typically be offered sequentially.

When to integrate the modalities inputs is decided on how will computer learning the interaction techniquesaffect its fusion or how will the fusion may affect the interaction system. These questions should be properly addressed by practitioners in the field in order to characterize better the applicationsand problems that multimodal fusion able to improve the conventional unimodal interfaces.

In multimodal interactive systems, multimodal fusion is a crucial step in combining and interpreting the various input modalities. Once the desired modalities are selected, an important question to be addressed is how to combine them. To address this problem, it is helpful to know how the integrating modalities relate in AR environment. Some modalities, like speech and lip movements, are more closely tied than others, such as speech and hand gestures. It is also plausible to assume that integration of such different combinations of modalities should be explored at different levels of integration. Depending on the chosen level of integration, the actual fusion can then be performed using numerous methods, ranging from simple feature concatenation to complex interaction of interface agents.

Unlike unimodal interfaces, multimodal requires having multi-signal fusion architecture to merge two or more input commands in a natural and efficient way. We should have a history of each mode of signal. With the analysis of each signal, statistical characteristics will be obtained. Then, multi-channel signal fusion is available with the provided statistical characteristics. Additionally, environmental context and task context should be considered to provide better recognition result. The main difference between a unimodal interface and a multimodal interface is that the multimodal interface requires multimodal fusion architecture to merge two or more modality input in an efficient and effective way. As presented in Table 1, multimodal fusion systems can be classified in two groups: (1) feature level fusion and (2) semantic level fusion.

Table 1. Classification of multimodal fusion on *how* to integrate modalities

Feature	Semantic
Fusion is finished before the input signals are sent to their respective recognizers	Fusion is finished after the signals are interpreted from their respective recognizers
Input signals are complex to model	Interpret the input signals independently
Difficult to train required high data training	Easy to train with existing unimodal data

Feature level fusion is done before the input signals are sent to their respective recognizers. Feature level fusion is considered as a good strategy for integrating the closely coupled and synchronized input signals, for example, lip movement and speech input whose signals correspond to each other. Typical drawbacks of the feature level fusion are that it is complex to model, intensive to compute, and difficult to train. Mostly, feature level fusion requires a large amount of training data.

Semantic level fusion is done after the signals are interpreted from their respective recognizers. Semantic level fusion is appropriate for integrating two or more signals which provide complementary information, such as, speech and pen input. Individual recognizers are used to interpret the input signals independently. Those recognizers can be trained with existing unimodal training data. Therefore, input channels needed to have complementary information to each other and time-stamp played an important role to match two different modalities for integration. Semantic level fusion is that semantic representation of the recognized input was essential for multimodal fusion and that mutual disambiguation was necessary to improve error handling and resolution.

4 Conclusion

There are numerous potential benefits in integrating multiple modalities. The reasons range from the fact that natural human interaction itself. The interaction of humans with their environment including with other humans, is naturally multimodal. The human factors need to be addressed before moving to integrate the modalities. That is motivating people to study and explore multimodal interaction. When it comes to unimodal, however, we usually use only one interface device at a time like typing, clicking the mouse button, speaking, or pointing with a magnetic wand. The ease with which this unimodal interaction allows us to convey our intent to the computer is far from satisfactory. The practical reason can lead to consider the use of multimodal interaction. The task can be more practical and convenient with multimodal inputs.

Drawback of current advanced unimodal is that it lacks robustness and accuracy. Whether they use a stylus or a glove or are vision based, they are still constrained to the recognition of few predefined hand movements and are burdened by cables or strict requirements on background and camera placement. Gestures that complement speech, on the other hand, carry a complete communicational message only if they are interpreted together with speech and, possibly, gaze. The use of such multimodal messages can help reduce the complexity and increase the naturalness of the multimodal interface. In this studies we have explored on multimodal fusion in AR. Multimodal has been a topic of research in AR since decades. Though studies have been conducted to establish the feasibility of these novel modalities using appropriate sensing and interpretation techniques, their role is still being explored to compare the partial-immersive AR system against the fully-immersive VR systems. On the first section of this paper we have described about multimodal in general. Second section later has explained the multimodal in VR against the multimodal in AR environments. We have detailed out the previous works and researches have been done in multimodal that invites modalities gesture and speech as inputs. In Next section we identified the multimodal fusion in AR. We found fusion levels in dealing with multimodal in AR.

Acknowledgments. We would to express our appreciation to Universiti Teknologi Malaysia (UTM). We also thank the Human Interface Technology Laboratory New Zealand (HITLabNZ) at University of Canterbury. This work was funded by Ministry of Science, Technology and Innovation through eScieneFund Grant 01-01-06-SF1188

References

1. Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., Blair, M.: Recent advances in augmented reality. IEEE Computer Graphics and Applications, 20–38 (2001)
2. Zhou, F., Duh, H.B.-L., Billinghurst, M.: Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In: Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality. IEEE Computer Society (2008)
3. Kaiser, E., Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., Cohen, P., Feiner, S.: Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. In: International Conference on Multimodal Interfaces, ICMI 2003, pp. 12–19 (August 2003)
4. Corradini, A., Cohen, P.: On the Relationships among Speech, Gestures, and Object Manipulation in Virtual Environments: Initial Evidence. In: Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, pp. 52–61 (2002)
5. Mitra, S., Acharya, T.: Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 37(3), 311–324 (2007)
6. Lim, C.J., Pan, Y., Lee, J.: Human Factors and Design Issues in Multimodal (Speech/Gesture) Interface. JDCTA 2(1), 67–77 (2008)
7. Jewitt, C.: Technology, literacy and learning: A multimodal approach. Psychology Press (2006)
8. Haller, M., Billinghurst, M., Thomas, B.H. (eds.): Emerging technologies of augmented reality: interfaces and design. IGI Global (2007)
9. Irawati, S., Green, S., Billinghurst, M., Duenser, A., Ko, H.: An evaluation of an augmented reality multimodal interface using speech and paddle gestures. In: Pan, Z., Cheok, D.A.D., Haller, M., Lau, R., Saito, H., Liang, R. (eds.) ICAT 2006. LNCS, vol. 4282, pp. 272–283. Springer, Heidelberg (2006)
10. Sharma, R., Pavlovic, V.I., Huang, T.S.: Toward multimodal human-computer interface. Proceedings of the IEEE 86(5), 853–869 (1998)
11. Richard, A.: Bolt: Put-That-There: Voice and Gesture at the Graphics Interface. In: Proceedings of the International Conference on Computer Graphics and Interactive Techniques, vol. 14, pp. 262–270 (1980)
12. Cohen, P.R., Johnston, M., McGee, D.R., Oviatt, S.L., Pittman, J.A., Smith, I., Chen, L., Clow, J.: Quickset: Multimodal Interaction for Distributed Applications. In: Proceedings of the Fifth Annual International Multimodal Conference, pp. 31–40 (1997)
13. Nicholson, M., Vickers, P.: Pen-Based gestures: An approach to reducing screen clutter in mobile computing. In: Brewster, S., Dunlop, M.D. (eds.) Mobile HCI 2004. LNCS, vol. 3160, pp. 320–324. Springer, Heidelberg (2004)
14. Chu, C., Dani, T., Gadh, R.: Multimodal Interface for a virtualreality based computer aided design system. In: Proceedings of 1997 IEEE International Conference on Robotics and Automation, vol. 2, pp. 1329–1334 (1997)
15. Krum, D.M., Omotesto, O., Ribarsky, W., Starner, T., Hodges, L.F.: Speech and gesture control of a whole earth 3D visualization environment. In: Proceedings of Joint Eurographics-IEEE TCVG Symposium on Visualization, pp. 195–200 (2002)
16. LaViola, J.: MSVT: A virtual reality-based multimodal scientific visualization tool. In: Proceedings of the Third IASTED International Conference on Computer Graphics and Imaging, pp. 221–225 (2000)

17. Wang, J.: Integration of eye-gaze, voice and manual response in multimodal user. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetic, vol. 5, pp. 3938–3942 (1995)
18. Fröhlich, C., Biermann, P., Latoschik, M.E., Wachsmuth, I.: Processing Iconic Gestures in a Multimodal Virtual Construction Environment. In: Sales Dias, M., Gibet, S., Wanderley, M.M., Bastos, R. (eds.) GW 2007. LNCS (LNAI), vol. 5085, pp. 187–192. Springer, Heidelberg (2009)
19. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review, 1–54 (2012)
20. Olwal, A., Benko, H., Feiner, S.: SenseShapes: Using Statistical Geometry for Object Selection in a Multimodal Augmented Reality System. In: Proceedings of the Second IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2003), Tokyo, Japan, October 7-10, pp. 300–301 (2003)
21. Heidemann, G., Bax, I., Bekel, H.: Multimodal Interaction in an Augmented Reality Scenario. In: Proceedings of International Conference on Multimodal Interfaces, ICMI 2004, pp. 53–60 (2004)
22. Irawati, S., Green, S., Billinghurst, M., Duenser, A., Ko, H.: An evaluation of an augmented reality multimodal interface using speech and paddle gestures. In: Pan, Z., Cheok, D.A.D., Haller, M., Lau, R., Saito, H., Liang, R. (eds.) ICAT 2006. LNCS, vol. 4282, pp. 272–283. Springer, Heidelberg (2006)
23. Kato, H., Billinghurst, M., Poupyrev, I., Imamoto, K., Tachibana, K.: Virtual Object Manipulation on a Table-Top AR Environment. In: Proceedings of the International Symposium on Augmented Reality, pp. 111–119 (2000)
24. Lee, M.: Multimodal Speech-Gesture Interaction with 3D Objects in Augmented Reality Environments (2010)
25. Neumann, A., Schnier, C., Hermann, T.: &Pitsch, K. Interaction Analysis and Joint Attention Tracking In Augmented Reality. In: Proceedings of the 15th ACM International Conference on Multimodal Interaction, pp. 165–172 (2013)
26. Dierker, A., et al.: Mediated attention with multimodal augmented reality. In: Proceedings of the 2009 International Conference on Multimodal Interfaces. ACM (2009)
27. Neumann, A.: Design and implementation of multi-modal AR-based interaction for cooperative planning tasks. Bielefeld University, Bielefeld (2011)
28. Pitsch, K., Neumann, A., Schnier, C., Hermann, T.: Augmented reality as a tool for linguistic research: Intercepting and manipulating multimodal interaction. In: Multimodal Corpora: Beyond Audio and Video (IVA 2013 Workshop), pp. 23–29 (2013)