# [20210412] INFOMMMI - Multimodal Interaction - 3 - thuis

**Course: BETA-INFOMMMI Multimodal interaction (INFOMMMI)**

## Comments

Some comments on possible solutions. Notice that these are incomplete and for some questions other answers exist that might give full credit, too.

Note: this file only contains the questions covering lectures 5-7 (by W. Huerst)

| | |
|---|---|
| **Duration:** | 2 hours |
| **Number of questions:** | 10 |
| **Generated on:** | Apr 11, 2021 |

**Contents:** Pages:

# [20210412] INFOMMMI - Multimodal Interaction - 3 - thuis

## Course: Multimodal interaction (INFOMMMI)

**EXAM CONTENT AND DURATION**

Question1-6 cover lectures 1-4 by Peter Werkhoven (max. 50 points).
Questions 7-10 cover lectures 5-7 by Wolfgang Hürst (max. 50 points).

***Important:*** Questions 7-10 contain 32 sub-questions, so plan your time accordingly. The maximum time to answer all questions is two hours. It is up to you how much time you spend on which questions. If you spend too much time looking up things in external sources, you will likely run out of time.

**By partaking in the exam you agree to the following CODE OF CONDUCT**

This test takes place under special circumstances in which we, even more than usual, rely on your professionalism and integrity. By partaking in this digital exam, you agree to the following code of conduct:

- You are logged in with your own account and take this exam in your own name.
- You will take this exam yourself, without contact or help from others.
- You will not copy, “screen dump”, or otherwise record or distribute questions or answers during or after the exam.
- You will only use permitted tools and resources. In this case, since it is an open book exam, these are notes, books, printouts, and online resources.

By partaking, you also confirm that you are aware of the following things:

- Violation of the aforementioned agreements is regarded as Fraud (see OER art 5.14).
- Answers can be checked for plagiarism.
- The results of this exam are conditional: if deemed necessary, the examiners can invite you for an additional oral exam at a later stage.

Good luck with the exam!

**Number of questions:** 10

**You can score a total of 100 points for this exam, you need 50 points to pass the exam.**

## 7 AR systems (different handheld AR systems)

Augmented reality (AR) can be realised in various ways. In this group of questions we want to look into different handheld AR systems, that is, AR systems created using your mobile phone.

*Because it was asked for "debatable" cases, obviously there are many different answers that can be considered as correct in addition to the ones mentioned below. In fact, using the right arguments, the AR UFO game could have been used as an example for all three. The most important thing here was that your answer demonstrated a good understanding of the respective criteria. Many used the same example for the first and last one, which makes sense, but the reason / explanation is slightly different.*

2 pt. **a.** (max. 2 pts) Shortly explain in your own words how AR can be created on a mobile phone.

(*No details needed. About two sentences describing the basic idea could be enough to get full credits.)*

*The answer is basically on slide 7 of lecture 6. The perfect solution must contain:*
*1) Video for real, graphics integrated for augmented*
*2) Sensors used for tracking / to place graphics correctly (e.g., IMU to track phone and CV/camera image to get info about environment)*

Azuma's definition of AR provides a framework that allows us to compare different systems against each other via three characteristics of an AR system.

2 pt. **b.** (max. 2 pts) The first characteristic is that it "combines real and virtual".

Give one example of a mobile app that people commonly refer to as AR, although it technically violates this rule, does not completely fulfil it, or where it is at least debatable if it is really fulfilled. Shorty explain why.

(*A high-level description of the example is sufficient. One or two sentences such as "An app that …, because it …" could be sufficient to get full credits.*)

*Good answers are basically everything where the real and virtual part are not strictly combined or either of them is not really needed. For example:*
*A navigation app that does not directly place arrows indicating directions on locations in the real world but just shows them on the screen could also be used without the real-life video stream, thus not really combining real and virtual.*
*Pokemon Go, because it can also be played with the live video stream turned off. If the "reality part" of the game is not needed, it is debatable if one should call it an AR game.*
*Some mentioned examples using Google Streetview or Google Earth, which are actually very good solutions, too (if the related explanation was correct). Another good example people came up with is HUDs.*

2 pt. **c.** (max. 2 pts) The second characteristic of an AR system is that it "is interactive in real-time".

Give one example of a mobile app that people commonly refer to as AR, although it technically violates this rule, does not completely fulfil it, or where it is at least debatable if it is really fulfilled. Shorty explain why.

(*A high-level description of the example is sufficient. One sentence such as "An app that …, because it …" could be sufficient to get full credits.*)

*A correct answer would be anything where the interaction with the virtual elements is non-existent, very limited, or only indirectly done (remember the example from the Terminator movie shown in lecture 5). A good answer in relation to handheld AR would be the information browsers that we discussed, so something like:*

*"A mobile information browser, because users don't interact directly with the virtual elements directly, but they are only indirectly changed when moving the phone."*

2 pt. **d.** (max. 2 pts) The third characteristic of an AR system is that it "is registered in three dimensions".

Give one example of a mobile app that people commonly refer to as AR, although it technically violates this rule, does not completely fulfil it, or where it is at least debatable if it is really fulfilled. Shorty explain why.

(*A high-level description of the example is sufficient. One sentence such as "An app that …, because it …" could be sufficient to get full credits.*)

*A correct answer would be anything where the virtual parts are not correctly integrated into the live video stream, but only superimposed. We saw several of them in the lecture (e.g., the UFO attack game, the Beatles crossing Abbey Road).*

## 8 AR systems (different displays, pros and cons)

Now let's look into other AR systems that are created with different displays. In particular, we want to compare AR systems created via optical-see-through head-mounted displays (**OST HMDs**), via projections (**projected spatial AR**), and via spatial see-through displays (**see-through spatial AR**). The latter refers to fixed installations of see-through displays (in contrast to head-worn ones). We did not discuss them in detail, but we saw an example in the fifth lecture (i.e., the SpaceTop system shown in video 6 of lecture 5).

We want to compare these systems with a concrete **use case: an AR ride in an amusement park**. In the lecture, we saw an example of the Mario Kart AR ride in the Super Nintendo World at Universal Studios. Here, AR elements are embedded via projections or large screens in the environment (projected spatial AR) and glasses handed out to the riders (OST HMDs). If the cars had a windshield that serves as see-through display, we could also integrate AR elements there (see-through spatial AR).

Assume you are an engineer at an amusement park that wants to build an AR ride as well. You are now discussing with your colleagues which of these three technologies would be the best to use.

(*Short answers are sufficient in the following. In most cases, one sentence or even a short phrase could be enough to get full credits.*)

*Many different but correct answers exist for the following six sub-questions. It basically comes down to the pros and cons of each approach and if and how the related characteristics play a role in this concrete scenario.*

*Correct answers that people came up with ranged from rather simple aspects (e.g.: HMDs must be worn personally whereas spatial AR solutions are integrated in the environment. Resulting disadvantages include discomfort, higher risk of device damage, longer boarding times), rather complex ones, or very sophisticated technical details.*

1 pt. **a.** (max. 1 pt) Give one advantage that using projected spatial AR could have in this context compared to OST HMDs.

1 pt. **b.** (max. 1 pt) Give one advantage that using OST HMDs could have in this context compared to projected spatial AR.

1 pt. **c.** (max. 1 pt) Give one advantage that using see-through spatial AR could have in this context compared to OST HMDs.

1 pt. **d.** (max. 1 pt) Give one advantage that using OST HMDs could have in this context compared to see-through spatial AR.

1 pt. **e.** (max. 1 pt) Give one advantage that using projected spatial AR could have in this context compared to using see-through spatial AR.

1 pt. **f.** (max. 1 pt) Give one advantage that using see-through spatial AR could have in this context compared to using projected spatial AR.

Now assume that the ride is supposed to be a group experience, that is, you can do it together with a group of friends in the same car or two cars next to each other and there are some interactive elements that you do as a group. Assume, for example, that each rider has a gun and can shoot virtual AR balloons that burst when they are hit (which implies that a balloon cannot be shot again by you or members of your group after it got hit once; riders who are not in your group might still see it though and can still shot it). At the end of your ride, you get a group score based on how many balloons your group burst.

2 pt. **g.** (max. 2 pts) Considering this multi-user experience, which of the three AR solutions from above would you recommend to use? Give a short explanation why.

(*A short explanation is sufficient as long as it convincingly justifies your choice*.)

*If you read the description very carefully, there is actually only one correct answer: HMDs, because technically, group members can be distributed over different cars and members from different groups could be in one car. Thus, all spatial AR versions would not allow a view that is unique for each group.*

*Some people did not consider the "people from different groups could end up in one car" (which admittedly was also not explicitly stated in the question text), and gave good arguments for see-through spatial AR. If they were convincing, they still got full credits.*

In the Mario Kart AR ride mentioned above, the engineers decided not to use one form of AR but two (namely projected spatial AR and OST HMDs).

1 pt. **h.** (max. 1 pt) Give one good reason why this is a good decision.

*Various correct answers exist. Most argued that this way you can have a shared experience for everyone (via projected spatial AR) combined with a more personalized one (via HMDs), e.g., to show a personal score or to have different visuals for different target groups (kid-friendly vs. more scary stuff for adults), all of which were nice ideas.*

*You could also argue from a more technical side: HMDs are newer / more advanced / cooler, but more prune to failure. Adding spatial AR makes the ride interesting and still functional even when the HMDs cannot be used. I actually found this statement in one of the articles that I read about the ride and some students used it (or a variation of it) in the exam as well.*

Although the term "display" is commonly associated with a visual display, we sometimes also talk of displays in relation to other modalities; for example, "auditory displays" and "haptic displays". Likewise, we can also create AR systems that have no visual component but are focused on another modality. Let's look at "**auditory AR systems**", i.e., AR systems that combine real and virtual sounds. Again, we will do this in relation to an example from an amusement park.

Assume an amusement park has an animatronic placed at a fixed location on a sidewalk. (Animatronics are robot-like puppets that look and move like real people, animals, or creatures.) The animatronic is placed at a fixed location, but can turn, move its hands, arms, head, and other body parts in a realistic way and interactively talks to people approaching it. There are hidden cameras that give a live feed to the system in real-time. It shows what is happening in the field of view of the animatronic.

Because the animatronic is purely physical (and thus "real"), we should not call this "mechanical part" of the installation an AR system. But what about the audio part? Use Azuma's definition to discuss shortly if or to what degree this setup could qualify as an auditory AR system.

*The purpose of this question was to test if and how well you understood the three criteria and also their characteristic, i.e., not being a binary "yes/no" criteria, but a characteristic against which to compare different systems. E.g., "interactive in real-time" can mean a lot of things and for most systems, you can argue that they are interactive in real-time in some ways but not fully in others. Likewise, the description above does not give all information that is needed (which was intended, since the answer is speculative anyhow). Thus, for the grading, it was not that much relevant what answer was given concretely, but how it was justified.*

2 pt. **i.** (max. 2 pts) To what degree does the auditory part of this system fulfil Azuma's first criteria, i.e., combines real and virtual?

*One correct answer giving full credits could be:*

It is fully fulfilled, because virtual sounds (speech of the animatronic) and real ones (surrounding environmental sound and people talking) are mixed seamlessly.

*Some interpreted the sounds and the setup differently, others interpreted the criteria in a multimodal way (e.g., stating that the animatronic is real, but the sound is virtual, thus combining real and virtual), which technically is not in line with the question (which clearly asked to look at it from a pure auditory AR system perspective), but demonstrated that you have a good and correct understanding of the matter and thus also gave full credits.*

2 pt. **j.** (max. 2 pts) To what degree does the auditory part of this system fulfil Azuma's second criteria, i.e., is interactive in real-time?

*The text in the question clearly suggests that it is interactive in real-time to some degree ("The animatronic … interactively talks to people approaching it"). Yet, it is not clear how detailed and sophisticated this interaction is. The existence of a camera suggests that the system can react to people individually (e.g., even addressing them individually with things such as "Hey, you in the red shirt."), but it all depends on various things. Also, does the system have an AI component and text-synthesis that automatically creates sentences related to a current situation or does it just use pre-recorded speech snippets? Are there microphones and speech recognition to have a "real" conversation or just cameras? What if a person approaches the animatronic from the side and starts talking to it when being outside of the camera's FOV? There are tons of things that come into place here, but no detailed analysis was required, so basically stating that it is interactive due to the quote given above gave one point in the grading. Stating that the level of "how interactive" it really is with, e.g., one example gave another point.*

2 pt. **k.** (max. 2 pts) To what degree does the auditory part of this system fulfil Azuma's third criteria,i.e., is registered in three dimensions?

*One correct answer giving full credits could be:* Assuming the sound from the animatronic is realized in a way that is perceived as coming from its mouth, the virtual part seems perfectly registered in 3D.

*Many other different phrasings exist that could give full credits. It is important though that this is about the integration of the virtual parts into the real world, so not about "is the sound 3D" but is the location of origin a real space in 3D. E.g., a 2D visual can also be perfectly registered in 3D, even if it is not a 3D graphic itself. Again, was more important that your answer demonstrated a good understanding of this criteria rather than being the same as the one given above (because you could interpret the situation also differently).*

## 9 VR and AR comparison

Virtual reality (VR) and AR are related and share many techniques, hardware, and algorithms. Yet, there are also differences between them and the concrete techniques that are applied.

2 pt. **a.** (max. 2 pts) Give one tracking-related problem in AR that does not exist in VR or is commonly easier to deal with in VR.
(A *short answer with, e.g., one or two sentences could be sufficient to get full* credits)

2 pt. **b.** (max. 2 pts) Give one tracking-related problem in VR that does not exist in AR or is commonly easier to deal with in AR.
(A *short answer with, e.g., one or two sentences could be sufficient to get full* credits)

*Many brought concrete examples in (a) for AR that require you to have detailed information about the environment in order to have full 3D registration. Likewise, several students used examples in (b) that require body-tracking in VR, which is commonly less needed in AR. (It is needed for interaction, but not to create a first-person avatar, since, well, you see your own body in AR and thus don't have to model it).*
*Others referred to the problem of jitter and lags leading to "swimming objects" for AR, and lags leading to cyber sickness due to problems with proprioception for VR. Also, good and valid answers for (a) and (b), respectively.*

In their paper "Touch the wall: Comparison of virtual and augmented reality with conventional 2d screen eye-hand coordination training systems," Batmaz et al. mention that "highlighting (of) objects" is known to increase selection time and throughput. They also state that other visual cues could be used "to obtain a stronger spatial comprehension of the VE, which may also increase selection performance" (quote from the paper). They list shadows, motion parallax, and texture as examples, but there are may others, too.
We also discussed visual cues in the lecture (but I used the term "depth cues" instead of visual cues). One sub-group of these depth cues are pictorial depth cues. They can be used to improve depth perception in AR and VR. Yet, this is often more difficult in AR than in VR.

1 pt. **c.** (max. 1 pt) Name one pictorial depth cue that is clearly more difficult to use in AR than in VR and shortly explain why.
(*One sentence could be sufficient to get full credits.*)

**1** pt. **d.** (max. 1 pt) Name one pictorial depth cue that is generally as easy to use in AR as in VR and shortly explain why.
(*One sentence could be sufficient to get full credits.*)

*Key here is integration in real world. E.g., size can be used for both (thus, good example for second question), but shadows need knowledge about the real environment in AR, but not in VR (thus, good example for first question).*

Assume you liked the paper by Batmaz et al. so much that you want to do a follow up study as a master thesis. Your idea is to test another visual cue in a similar setup and analyse its effect on the measurements done by Batmaz et al.

**2** pt. **e.** (max. 2 pts) Tell me which visual depth cue you would choose and give me a good reason why. Your reason should be strong enough to convince me to be your supervisor, that is, you should not just mention any visual cue, but have a good reason why this particular cue would be worth studying.
(*Any reason that reflects that you have a good understanding of the context gives full credits, even if I do not fully agree that this visual cue would be the best one to study*.)
*No concrete answer here; many possible ones exist.*

The paper by Batmaz et al. also addresses active and passive haptics and introduce "visuo-haptic augmented reality (VHAR)", which "enhances reality through haptic interaction and enables users to interact more precisely" (quote from paper).

2 pt. **f.** (max. 2 pts) Give one reason why or an aspect, characteristic, situation where realising active haptics is generally more difficult to achieve convincingly in AR than in VR.

(*A short answer is sufficient, as long as it clearly shows that you know what active haptics means in this context and your statement is convincing.)*

*Active haptics (e.g., force feedback on your fingers when touching a virtual object) usually need some additional equipment (e.g., gloves) which you you always see in AR.*

2 pt. **g.** (max. 2 pts) Give one reason why or an aspect, characteristic, situation where realising passive haptics is more difficult to achieve convincingly in VR than in AR.

(*A short answer is sufficient, as long as it clearly shows that you know what passive haptics means in this context and your statement is convincing.*)

*Passive haptics in projected AR. Since we project onto a surface, passive haptics exist per default (although there can be a mismatch in texture, material, etc.).*

In their discussion, Batmaz et al. state that their "result also confirms that there is (still) a difference between real-world 2D screen systems and VR and AR application, however our results reveal that especially VR systems have a strong potential for eye-hand coordination training systems" (quote from the paper).

1 pt. **h.** (max. 1 pt) Give one technical improvement that current VR systems need to not just make them a potential but a true replacement for 2D screen systems.

(*A short sentence or phrase could be sufficent to get full credits.*)

*Apart from the results and general issues listed in the paper, basically all shortcomings of current VR systems that have a potential impact for this particular task are a good answer here.*

1 pt. **i.** (max. 1 pt) Shortly explain why current AR systems are not mentioned as potential replacement for 2D screens in the sentence from Batmaz et al. quoted above.

(*You do not have to refer to the results in the paper but some general statements with respect to AR and VR could be sufficient to get full credits.)*

*Some did not understand this question correctly (but plenty did, so it looks like the problem was not the phrasing). It comes down to either what the authors say on page 190: "… we believe that the difference between AR and VR conditions was caused by the drawbacks and limitations of current AR headsets." Or to what I said in the lecture, i.e., that AR technology is kind of behind VR in terms of quality and sophistication due to the challenges being much higher. A general statement like this, as well as some convincing concrete examples were sufficient to get full credits for this question. It was important though that the reason why VR is mentioned as potential alternative with some improvement, but AR is not becomes clear (so, simply stating that the results were worse is not sufficient).*

## 10 AR interaction

In the following group of questions, we want to address **interaction in AR** in more detail. Most approaches for manipulation in AR can be categorised as either isomorphic or non-isomorphic.

1 pt. **a.** (max. 1 pt) Give one example for an isomorphic approach and explain why it is isomorphic.

(*No lengthy explanation of the approach and the reason is needed; any short text that shows that you know why it is isomorphic and did not just guess or write down the name of a random technique can be sufficient to get full credits.*)

1 pt. **b.** (max. 1 pt) Give one example for a non-isomorphic approach and explain why it is non-isomorphic.

(*No lengthy explanation of the approach and the reason is needed; any short text that shows that you know why it is non-isomorphic and did not just guess or write down the name of a random technique can be sufficient to get full credits.*)

**1** pt. **c.** (max. 1 pt) Give one reason, use case, or situation where using an isomorphic approach would intuitively be better and shortly explain why.

(*Again, no lengthy explanation is needed, but just a few comments illustrating that you understand the context and are not just guessing can be sufficient to get full credits.*)

*See slide 38 from the last lecture for the definition of isomorphic / non-isomorphic and the following ones for different approaches. Other approaches exist and gave full credits, too, if explained correctly. Likewise, various correct answers exist for sub-question (c); basically everything that is likely to be easier to understand or use if it is "more natural / realistic" or benefits from "imitating physical reality" (quotes from slide 38).*

In their paper "A comparative analysis of 3d user interaction: How to move virtual objects in mixed reality," Kang et al. compare the three interaction approaches with each other: gaze and pinch interaction, direct touch and grab interaction, and worlds-in-miniature interaction. They also list two additional ones (go-go interaction and ray-casting) and we discussed even more in the lecture (e.g., cone casting, image plane, 3d bubble cursor, and more).

**2** pt. **d.** (max. 2 pts) Pick one technique that was not tested by Kang et al. (it can be one of the two they listed or one of the others that we discussed in the lecture) and explain it shortly in your own words.

(*A short, general explanation that illustrates that you understood the technique is sufficient to get full credits.*)

*See various techniques listed on slides 39 to 43 in the last lecture.*

In the lecture, we said that the "effectiveness of any technique is task dependent" (quote from a slide in the last lecture). In their paper, Kang et al. found evidence that "the object's visual appearance" can play an important role, too, when "designing natural hand interaction" (quotes from abstract). For example, they quote the following comment from one participant: "Even if it's not real, you think it's furniture. It feels like it should be heavy, and I should be using both hands."

2 pt. **e.** (max. 2 pts) Give one other example for a visual appearance of an object where users might try to interact with the object differently if this characteristic is changed.

(*One or two sentences could be sufficient to get full credits.*)

*Several correct solutions exist here, including "shape" (e.g., intuitively, you would grab the handle of a teacup differently than a teaspoon or sugar box) and "material" (e.g., you might grab a fragile looking object more carefully and thus use a different grip).*
*Many gave just one concrete example. I noticed that the question can easily be misinterpreted (my mistake) and thus also gave full credits for such answers.*

The authors note that many participants indicated in the interviews "that they want to combine all three interactions rather than picking one" (quote from the paper).

2 pt. **f.** (max. 2 pts) Explain why it might indeed be a good idea to not just implement one option but offer different ones when creating such an AR system.

(*No detailed explanation is needed. A general example or intuitive reason is sufficient to get full credits.*)

*The answer is basically in the sub-question above. Different objects might suggest a different interaction mode (e.g., it seems both more intuitive to grab a teacup handle with two fingers but a sugar cup with your whole hand), so why not just offer both options?*

While your reason given in the previous sub-question might sound convincing, there are also reasons that speak against such a combination of approaches.

2 pt. **g.** (max. 2 pts) Give one convincing reason or example that would speak against combining different approaches into one system but suggests it is better to use just one.

(*Again, no detailed explanation is needed, but a rather short intuitive reason is sufficient to get full credits.*)

*There are also various intuitive reasons that speak against offering multiple options. E.g., above example are "obvious" cases, but what about grabbing objects where the "obvious" way to grab them is not clear? Other aspects include inconsistency (esp. in relation to performance), potential confusion of users, implementation (do you implement all options for all objects or only the "obvious" one for each), etc.*

In their study Kang et al. use an AR setup created with a video-see-through head-mounted display (VST HMD). Now let's assume you want to recreate their experiment but instead of a VST HMD, you are using an optical-see-through head-mounted display (OST HMD). It seems fair to assume that some of the results might change, while others may stay the same. We don't know for sure until we do the experiment, but we can certainly make some educated guesses.

2 pt. **h.** (max. 2 pts) Pick one difference between VST HMDs and OST HMDs, state one aspect of the evaluation, make a prediction about how it will change (or not, if the prediction is that it will not have an impact) and justify your prediction. It is easiest to phrase this in the form of a hypothesis (but don't forget the justification).

(*Something like this could be sufficient to get full credits: "Compared to VST HMDs, OST HMDs have/are/… <difference>. Because … <justification>, I expect that … <aspect of the evaluation> will be … <prediction>." Other phrasings are okay, too. Because it is a prediction, there is no single correct answer. Any convincing reason for your hypothesis will give you full credits.*)

*Example solution:*

*Latency can lead to jitter and "swimming artefacts" for OSTs.*
*Because these do usually not happen with VSTs, I expect that the accuracy will be lower.*

Thank you for participating in the course. We hope you enjoyed it.