(INFOMMMI) Multimodal interaction - 6 april 2020

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)

(INFOMMMI) Multimodal interaction - 6 april 2020

Course: Multimodal interaction (INFOMMMI)

EXAM CONTENT AND DURATION

Question 1-8 cover lectures 1-4 by Peter Werkhoven (max. 60 points).
Questions 9-11 cover lectures 5-7 by Wolfgang Hürst (max. 40 points).

Important: Questions 9-11 contain a total of 26 sub-questions, so plan your time accordingly. The maximum time to answer all questions is two hours. If you spend too much time looking up things in external sources, you will likely run out of time.

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: 11

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

9 In his paper "A Survey of Augmented Reality", R. Azuma uses three characteristics to define augmented reality (AR). One of them is that AR "**combines real and virtual**". In the following group of questions, we want to look into this aspect in relation to the different display technologies that are used to achieve this, along with their characteristics and potential limitations.

Because it was asked to "explain in your own words", many different correct answers exist. The ones below are just some examples illustrating one possible answer that would give full credits. Some people gave very informal descriptions, which is totally okay. As long as it became clear that you really understood the matter, you got full credits, even if your answer was informal and not very scientific or may have even missed things of minor relevance.

2 pt. **a.** [max. 2 pts] When creating a virtual character in AR with **optical see-through displays** (OST displays), we often get a so-called "ghost effect". Explain in your own words what this effect is and why it happens.

OSTs add a light source between the real world and your eyes. Yet, they only add light, but cannot block out the real light sources behind them. Thus, both virtual and real light streams hit your eyes making the virtual parts often look like "transparent ghosts".

1 pt. **b.** [max. 1 pt] Given an example for a useful AR application where such a "ghost effect" is ***not*** a problem or maybe even wanted.

Any example where we want some sort of "X-ray" view, e.g., **the luggage size check app by KLM** *mentioned in a later question or the* **"parcel size check" app** *we saw in the lecture.*

*A very simple (and in retrospective obvious) answer to this would be something like "***an AR game with ghosts***". I didn't think about that, but some of you did, so kudos to them and full credits. Some also used "***navigation apps***", since it is safer to also see the real world behind the augmentation, which is a nice example, too.*

1 pt. **c.** [max. 2 pts] Another effect that we sometimes get with OST displays are "swimming artifacts". Explain in your own words what this means. What is the major reason for this effect?

Because of latency, a virtual object's location in the real world is updated with a slight delay, making the object appear "floating" like an object that is swimming in the water.

Various different phrasings exist and were also considered correct (e.g., not stating "latency" but explicitly but describing it, using "jitter" or other words instead). Using "calibration" is technically not fully correct (meaning it is not incorrect, but doesn't cover the whole situation), but if you used it, it showed that you understood the issue very well, so you still got full credits.

1 pt. **d.** [max. 1 pt] What problem do **Light Field Displays** resolve? Name the problem and explain it in your own words.

Light field displays resolve the accommodation-vergence conflict which is a mismatch between accommodation (i.e., how your lenses focus on an object in the distance) and vergence (i.e., how both eyes rotate to focus on an object in the distance).

1 pt. **e.** [max. 1 pt] Give a convincing example for an AR system that would ***not*** benefit from Light Field Displays, but where a regular OST display would be totally sufficient. Explain why.

The accommodation-vergence conflict appears when people focus on objects at different depths. Thus, any example where all virtual objects are placed at a fixed distance from the eye would be correct here because they are always displayed at the same focal distance. A good example could be:

In a navigation app, where directions are shown at a fixed distance from your eyes (e.g., via arrows), your eyes don't need to refocus, so the accommodation-vergence conflict will not happen.

(No detailed description of any system is needed. It is sufficient to give a short description of, e.g., the characteristic of the system or application that is the reason for this.)

1 pt. **f.** [max. 1 pt] Another issue that OST displays suffer from is a low field of view (FOV). Give an example for a situation where such a limited FOV will likely result in a decrease of performance in a visual search task and explain why.

That question referred to one of the three papers, but given that it was an open book exam, I did not put this in the question, but assumed everyone who read it would immediately notice it. Many people gave a more informal answer that basically said that you see objects outside of the FOV later or that they are harder to find, which was okay, too, and gave full credits.

Because of the low FOV, there is no peripheral viewing field, which means targets placed there might get recognized later.

1 pt. **g.** [max. 1 pt] Assume we could build OST displays with a FOV that perfectly matches a human's FOV. Give an example for a situation where this would actually decrease search performance in a visual search task and explain why.

Targets in the peripheral view can also be distracting. Thus, there are also situations where eliminating it may increase search performance.

Again, this referred to the paper, but I also mentioned it in the lecture. Remember the Japanese video that I showed with the cognitive overload.

2 pt. **h.** [max. 2 pts] Explain the two terms field of view (FOV) and field of regard (FOR) and discuss them with respect to OST displays and **handheld AR displays** (e.g., when using your mobile phone to create AR).

FOR refers to the area in which virtual objects can be shown, whereas FOV refers to the area where they can be shown at the same time.

For both displays, the FOR covers all your surroundings, since virtual objects can appear everywhere when you turn your head or move the handheld device. For OSTs, the FOV only covers the narrow area on the glasses that can show the virtual images at a time. For handheld AR, the FOV covers the whole screen of the device.

Note that this is a very informal description of FOV; others exist and since I didn't ask for "describe it in your own words", more formal ones that resembled the actual text on the slides were okay, too. I used the informal one here to illustrate that such descriptions also gave full credit, since they clearly show that you understood it. Also, many people described the two examples very different, but equally correct and thus got full credits. Notice also that depending on how you argue, you could also make the case that the FOR of handhelds is just the devices display (so FOR = FOV). If you argued correctly, this gave full credits, too.

2 pt. **i.** [max. 2 pts] While we can create a certain kind of AR system on our mobile phones, we cannot create an AR system with them where we hold the phone in front of our eyes and use it similar to an OST display, i.e., in a way that the augmented, virtual elements seamlessly blend into the real environment around us. Shortly explain in your own words why.

When you hold the phone too close, you cannot focus on the screen anymore. If you move it further away, the FOV of the camera and your view of the environment don't match, and even if they do, you cannot focus on both the environment and the phone at the same time.

Some people did not address all of these three issues, but still got full credits, since their answer reflected that they understood the essential difference for these two cases very well.

3 pt. **j.** [max. 3 pts] In their paper "Breaking the Barriers to True Augmented Reality", Sandor et al. introduce an AR Turing test. They also discuss what kind of technology would be needed to achieve a "true AR" that would pass such a test. Obviously, none of today's AR systems is able to do this. Yet, in reality, most AR systems serve a concrete purpose. It might well be that this concrete purpose results in a usage where people are not be able to distinguish between real and virtual elements of the AR. Therefore, such a system might actually pass the AR Turing test (despite its limitations).

Give an example for such a system and context. Name the display technology used for it and shortly list the characteristics that it must have to pass the AR Turing test in this context (not in the general way described in the paper).

This question was difficult and required some thinking. It was put in here basically for students aiming for a top grade (9 or higher), so I didn't expect many to figure it out. Yet, some provided some good comments, which gave at least partial points. Some even got it totally right (even if they used a different example than the one below).

The AR Turing test is basically passed if a person cannot distinguish between real and virtual objects anymore. So, you just need to think of an example where the technology is already so far ahead that we can do this. I actually made a related comment (but didn't refer to the AR Turing test though) when we had the slide comparing OST with VST displays (remember the dragon and other stuff placed on a table). Thus, a correct answer could be:

For VSTs, both real and virtual world are displayed as a video. If the virtual elements are objects, they can hardly be distinguished from real ones. Thus, if you don't have to interact with them, such a system would likely pass the AR Turing test.

Side note: if you interpret "display" not only as visual but also as possibly "audio display" or "haptic display" (which are terms Peter used and are totally correct here), you could also bring an audio example, since the technology to create virtual audio sounds that are indistinguishable from real ones and register them perfectly in 3D exists. Someone actually did that and got full credits.

Many people just explained the AR Turing test and, if done correctly, got 1 pt for this. Several said it cannot be passed without dedicated technology (e.g., Light Field displays), which suggests they misunderstood the question. Thus, if their explanation was correct, they were still rewarded with up to 1 pt (depending on how well they argued and explained it).

10 Another characteristic used by R. Azuma to define AR is that it "**is registered in three dimensions**". In the following group of questions, we mostly want to look into this aspect, what it means, and what sensor technology is used to achieve it in different contexts.

2 pt. **a.** [max. 2 pts] For now, we restrict our discussion to the visual part of AR. Explain in your own words what "registered in three dimensions" means in this context.

Virtual objects are placed at an exact position in the 3d space of the real world.

Various other ways exist to explain this. The important part here is that the "3d" refers to the real world, not the virtual object (i.e., the object can also be 2d, but it is placed in and associated with a fixed 3d location of the real world).

Assume we want to create a **handheld AR** system using a mobile phone. Mobile phones generally contain so-called *inertial measurement units (IMUs)*, which in turn contain several sensors that can be used to implement the tracking of such handheld AR systems.

2 pt. **b.** [max. 2 pts] Give an example of a useful handheld AR application that cannot be implemented using solely the sensors on the IMU but needs at least one other sensor that we commonly find on today's mobile phones. Give the sensor's name and shortly explain why or for what purpose it is used.

(No lengthy explanation is needed, but it is sufficient to give a simple description of the feature or characteristic that this sensor provides and why it is important for this application.)

The IMU only gives you the orientation of the device, so every AR app that needs information about the environment would be a correct answer here. Could be something like:

An AR game that is played on a real table. Marker-tracking or natural feature tracking with the camera could be used to detect this table.

An information browser that gives you information about the buildings surrounding you. GPS could be used to get your location in the real world.

Most people stated "navigation apps" and "GPS", which is correct.

1 pt. **c.** [max. 2 pts] In the lecture, we saw a handheld AR app from the Dutch airline KLM, where you could point your phone at your hand luggage to check if it complies with the airline's size regulations. What kind of tracking approach are they using for this? Shortly explain why this is the best solution in this context.

Natural feature tracking using the phone's camera. It is sufficient to get the location of the floor, which can be done reliably with this technique and you don't need, e.g., markers, but just a camera which all phones these days have anyhow.

Other explanations also gave full credits as long as they illustrate a good understanding of the matter. Some students proposed other types of tracking (or ways of tracking, e.g., the luggage instead of the floor), which, if explained correctly and making sense in this context also gave partial credits (or full, e.g., if you track the luggage instead of the floor although it is technically not needed).

Give three potential registration issues that can happen when using *fiducial tracking.*

(Only list problems directly related to using fiducial markers. Mentioning general tracking problems will not give you any credits.)

1 pt. **d.** [max. 1 pt] 1st problem: **Marker may be (partially) out of the camera's view**

1 pt. **e.** [max. 1 pt] 2nd problem: **Markers only give you one location in 3d but no other environmental data (which is a problem, e.g., when AR objects are moving over the sides of a table)**

1 pt. **f.** [max. 1 pt] 3rd problem: **Registration errors due to changing lighting conditions**

The question asked for "registration issues", not general disadvantages of using markers. Yet, related comments that are indeed issues (but not registration-related) did get partial credits, too (0.5 instead of 1). Other correct examples related to registration issues include:

- **Orientation / small viewing angles from the camera**
- **Large distance from marker / marker too small**
- **Multiple markers could interfere with each other when they are too close (or accidentally overlap when they are not fixed)**

Some mentioned damaged, dirty, or worn out markers, which is a nice and correct aspect, too.

Imagine the following scenario: you are sitting in a circle with five virtual characters that are created with some AR technique. Assume your AR system has a wide FOV, no occlusion problems, and the tracking works smoothly in real-time without any noticeable delay.

1 pt. **g.** [max. 1 pt] Now you want to play a ball game with these virtual characters where you all remain seated and pass a virtual ball to each other in random order using a paddle. While the virtual characters, their paddles, and the ball are all simulated by the AR system, you are using a real table tennis paddle to pass the ball. What is the technical term for such an AR interaction approach?

TUIs / Tangible User Interfaces

Now we want to complement this visual AR system with **audio AR**. That is, we want to augment the real sound from your environment with virtual sounds that fit into the real environment in a similar way as virtual visuals blend into the real world in a 'perfect' visual AR system. Assume you are wearing headphones that do not block out the real environment, that is, you can hear the sound that they produce as well as any real sound in the environment surrounding you. The headphones can create perfect 3D audio, that is, they can simulate a sound source at any place in the real environment surrounding you.

3 pt. **h.** [max. 3 pts] Instead of playing ball, imagine now that you want to have a discussion with the five virtual characters, who are still sitting in a circle with you. Would the above-mentioned headphones be sufficient to create such an audio-visual AR system? If yes, explain why. If no, explain why not and illustrate what functionality or technology needs to be added to achieve this.

The answer is yes and no. To realize such a system you would need:

- *3D audio, which is provided via the headphones*
- *Information on where you are looking at (to simulate a "real", life-like conversation), which is provided via the head-tracking (which in turn must be there to create the visual characters and the 3D audio)*
- *A microphone, so the system gets your input to simulate the communication, which is not mentioned in the description text*

Some students misunderstood the described scenario (e.g., thought that these are "teleported characters" instead of "virtual characters") but gave a good explanation for this alternative understanding. Thus they got some credits, too. In general, this was a question that was intended to test your understanding of the matter and how well you can apply it to a context not directly discussed in the lecture (same for the following two sub-questions). Thus, as long as you reflected a good understanding of the matter, you still got credits (quite a lot in some cases), even if you did not answer the question in the way intended. I double-checked and verified the gradings for all students again at the end to make sure that the credits were awarded fairly and consistently. Your answers did indeed show a clear difference. Some students didn't provide answers at all or wrote things that clearly reflected that they didn't understand it. Many provided good answers that showed that they had a good understanding of the matter, even if they were incomplete or partially wrong. And several (quite a lot actually) demonstrated that they really developed a good understanding of it that went beyond just being able to explain knowledge that they gained from the course.

1 pt. **i.** [max. 1 pt] In the above audio-visual AR system (i.e., the five virtual characters having a conversation with you while you are all sitting in a circle), give one aspect where creating the visual part of the AR is more difficult than creating the auditory part of the AR (assuming the same characteristics for the technology as described above).

Generating realistic sound given that you have 3D audio headsets is easier than modelling 3D characters *(because even with a "perfect" display technology (e.g., no ghost views) it is still challenging to create realistic human animations)*

Many students brought up different ideas here, which were also correct. Some were not fully correct, but demonstrated a very good understanding of the matter and thus also got partial or sometimes even full credits.

1 pt. **j.** [max. 1 pt] For the same example, give one aspect where creating the auditory part of such an AR experience is more difficult to achieve than creating the visual part of it.

Providing a realistic conversation compared to providing realistic visual motions.
Modelling the motions of a realistic conversation, although not perfect, can still be done quite well and is definitely easier to achieve than a real-life conversation with virtual characters. (I admit that this was a tricky one that also goes beyond what we discussed in the lecture. But again, it was intended as one of the questions for the students aiming for a 9 or higher.) You could also argue that speech synthesis doesn't work that well, so you need pre-recorded speech segments (which would lead to a similar statement as above, i.e., that the content of the speech signal is the difficult part here). And again, several students came up with other ideas which were either equally correct or good enough to get partial or full credits.

11 The final characteristic used by R. Azuma to define AR is that it is "**interactive in real time**". In this group of questions, we address various aspects related to interaction.

1 pt. **a.** [max. 1 pt] In AR, we often use devices for interaction that support 6 DOF tracking. Yet, this is not always needed. Give an example where only three DOF are needed.

(No detailed explanation is required. A short description of a simple scenario, situation, or characteristic describing an AR system that obviously only needs these three DOF as input is sufficient. Don't forget to mention what these DOF are in your case.)

The 3 DOFs are location (x, y, z) and orientation (around 3 axes). While you can probably come up with an example for every possible triple of these six degrees, the easiest ones are likely the ones that don't need orientation but just location. The motivation of this and the next question was to see if you understood what these 6 DOFs are but also what they mean for an actual interaction. It was asked to give an example that describes such a case (and thus not necessarily a concrete application). The simplest one is probably:

An app where we interact with the AR with a 3D mouse pointer or cursor. Because we only need to know the location of the tip of the mouse pointer (x, y, z), there is no need to track orientation (i.e., the other 3 DOFs).

Note that it was not necessary to explain it in detail, if it was clear from your description that the answer is correct. E.g., above, just the first sentence would have been sufficient to give you full credits.

1 pt. **b.** [max. 1 pt] Give an example where only two DOF are needed.

(No detailed explanation is required. A short description of a simple scenario, situation, or characteristic describing an AR system that obviously only needs these two DOF as input is sufficient. Don't forget to mention what these DOF are in your case.)

Any example where the interaction is restricted to a 2D sub-space of the 3D world would only need x- and y-coordinates (i.e., any example that could be controlled in AR with a traditional mouse) or two rotation angles (e.g., if the sub-space is curved). Menu selection, heads-up displays that operate at a fixed distance from the user, etc. comes to mind, but it would also give full credits if you just stated the first sentence of this explanation.

Several students came up with the nice example of controlling a flying AR object for the previous question (where you need three coordinates to control its location in the air) and controlling, e.g., a car that drives on a flat surface such as a table for this question (where you need two coordinates to control its location on said table).

In the paper " Interacting with Distant Objects in Augmented Reality", Whitlock et al. address the problem of interacting with objects in AR that are not in your close proximity (e.g., within arm's length), but placed further away from the user.

1 pt. **c.** [max. 1 pt] The authors mention *ambiguities* as one aspect that might be more difficult when interacting with distant objects compared to ones that are in close proximity. Explain in your own words what this means in this context.

(No lengthy explanation is needed. A simple example could be sufficient to get full credits.)

See paper. Many came up with explanations not described in the paper that were also correct and gave full credits.

1 pt. **d.** [max. 1 pt] Another aspect mentioned in the paper are changes in a user's *mental model.* Explain in your own words what this means in this context.

(No lengthy explanation is needed. A simple example could be sufficient to get full credits.)

See paper. Many came up with explanations not described in the paper that were also correct and gave full credits.

One common approach for 3D interaction (both in AR as well as VR) is *ray casting.* Yet, ray casting has several issues when interacting with objects that are far away from the user, which is why several modifications of it have been introduced.

3 pt. **e.** [max. 3 pts] Give one example of such a modification that was discussed in the lecture. Mention its name, shortly explain how it works, what problem it solves (and why), and at what price.

(By "at what price" it is meant that most of the solutions we discussed solve the problem but introduce another one or make another problem even more critical. Which problem is that and why is it bigger with this approach?)

The two most obvious examples are **cone casting (aka flashlight)** *and the* **3d bubble cursor**. *Both create a bigger selection area (using a cone-shaped ray which is broader in the distance or a bubble-shaped cursor, respectively). That makes it easier to select (small) objects at a distance, but at the price of a more difficult selection of objects that are very close to each other. Other examples than these two exist and gave full credits, too.*

Ray casting and its variations all follow a so-called egocentric metaphor. Yet, there are also approaches following an *exocentric metaphor.*

1 pt. **f.** [max. 1 pt] Give one major drawback that these approaches have when we apply them to AR that does not exist with VR.

They disconnect the augmented / virtual parts from the real-world parts (or require you to visualize a partial replica of the real world).

The answer to this question was actually in the paper by Withlock et al. ("… decontextualize virtual content from the physical world"). The text in bracket (which was not needed to get full credits) refers to the image at the bottom right on the related slide from the lecture.

Some came up with other disadvantages (e.g., difficult to manipulate virtual and real objects that are next to each other at the same time).

In the paper by Withlock et al. that is mentioned above, the authors implemented translation of objects via voice-based interaction.

1 pt. **g.** [max. 1 pt] Give one disadvantage of their implementation compared to a traditional ray casting approach.

(It is not necessary to give a disadvantage that is mentioned in the actual paper. When looking at their implementation, there are also other obvious disadvantages.)

Several correct answers exist. One not mentioned in the paper is:

They impose a time restriction because objects move at a certain speed.

Some students wrote something along the lines of “speech recognition errors” or “doesn’t work well in noisy environments”, which is correct, too.

In their experiments, the authors of the above paper use virtual object sizes as depth cue in their AR system. In the future work section, they mention that one could also use other supplementary virtual depth cues.

2 pt. **h.** [max. 2 pts] Assume you want to make a follow-up project where you are using the authors' system but add another virtual depth cue to it and study its impact on the results with a comparable experiment. Which depth cue would you add and why?

(There is no "perfect" answer to this question. The purpose of it is just to verify if you got a deeper understanding of the knowledge that would allow you to make such a decision. Remember that we listed several possible depth cues in the AR lecture on displays. Pick one that you think would be most interesting to study and explain why you think such an experiment would be worth your efforts.)

Side note: The motivation behind this question was the following. Assume you are a student who followed the course, saw this paper, and is interested in doing a related thesis project. You approach me with the idea of doing one of the things they mentioned in their future work section, namely testing different depth cues. The first thing that I would likely ask you is, which depth cue and why. A student who has barely passed the course, would come up with a depth cue, but have difficulties explaining why he/she would want to study it. A good student who learned a lot from the course should be able to come up with a good argument why. Thus, students who just listed a depth cue but didn’t give a good reason why only got 1 pt, students who gave a good reason got full credits (even if I disagreed with their reasoning, as long as it reflects that they developed a good understanding of the matter).