SLRTP 2020

23 August 2020

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

SAMUEL: Thank you very much for that wonderful talk. We will start off by going through a few of the submitted questions. So, the first question is for multi‑cue fusion, the performance appears to be getting very good, up to 95% top accuracy on some of the numbers you showed. Do you have a sense of the kinds of errors that still remain and how they might be fixable?

LALE: We actually looked into this. We tried to divide the dataset into several parts. Some signs are longer than others and some signs include repetitive movements. Some signs have holds, some signs hand shape is more important and some signs, there's no hand shape, but rather it is a hand shape but there's no hold. So, it seems that the technique is better able to represent longer and repetitive signs, as opposed to signs that are based on hand shape. So, it seems that there is still room for better hand shape representation.

SAMUEL: Another question: Is for frame level semi‑supervised learning, were there any particular methodological learnings or findings from incorporating a human into the loop when you were performing this sign cluster annotation?

LALE: This clustering idea was based on the following observation. Signs were quite different, hand shapes looked quite different from different camera angles, for example. Suppose you are doing a sign, but you have the cameras slightly off, a human may be able to still recognise the sign because it knows there's movement and so on and so forth, but apparently there are some modes in the clusters, several clusters represent a hand shape. So our main idea was to capture one different camera clusters, multiple clusters, based on different camera angles, and second, or the hand, with respect to the camera, and second, variability within the phonological variability within the signs, so the human annotator actually takes several, signs several clusters to the same hand shapes, so the hand shape is represented by several modes.   
SAMUEL: I think this is for the submitted questions. Thank you very much for the talk and those inputs. I think we will next move on to the second session of accepted papers. If you could share the screen for that, that would be fantastic.   
We will ask the authors of the four papers, which are illustrated here, if they could raise their hands so we can promote them to panellists, that will allow us to do the Q and A. I will start with the first question for paper number 11. Towards continuous recognition of illustrative and spatial structures in sign language. It's a nice move to investigate architectures for non‑lexical recognition. As the model becomes more robust and can label further data, do you think there could be scope for learning computational relationships between the lexical signs and integrating them back into learning? Something akin to word to investigate, that is for Valentin. Perhaps we can also stop the screen sharing. I don't think Valentin is here. We will move on to the next question. The next submitted question is for paper number 12. Attention is all you sign, sign language translation with transformers, the question is, as we are working on either predicted or grand truth glosses, did you conduct a study where you artificially lower the gloss accuracy to see its effect on translation in more detail?

KAYO: Thank you for the question, that is a very good idea. I did not conduct a study yet, but it is something I would like to investigate in the future research. I looked at by hand the translation between from predicted glosses and from the ground truth to see if there was any difference between the two, but I didn't see any pattern in particular.   
SAMUEL: The next question is from Lorna and it's for paper 13 how to sign large scale dataset. The question is simply, is the how to sign dataset publicly available? This is a fantastic resource for the community?

AMANDA: Unfortunately, it's not finished yet. We are looking for publish in the convention, so I will let the community know when the dataset is available. I am very excited to make this dataset available so we can move forward with it.

SAMUEL: The next question is for paper number 14. Can everybody sign now exploring sign language video generation from 2D poses? This is a very cool piece of work, a key challenge here appears to be that the synthesis model does not currently bias its capacity towards improving the quality of hand generation, what do you think are the most promising directions for improving the generated hands?

LUCAS: Hello. Connection is a bit unstable. Could you repeat the question?   
SAMUEL: The question was, it's very cool, a key challenge here appears to be that the synthesis does not currently bias its capacity towards improving specifically the quality of hand generation. What do you think are the most promising directions for addressing this to improve the generated hands?

LUCAS: We have a few connection challenges now, it may be possible to write your answer into the chat, we can return later if there is connection.

SAMUEL: Question from Oscar.   
OSCAR: I was wanting to ask, with paper 13, how to sign, in the paper you mentioned that you have these 80 hours collected. And that the annotation is ongoing, I was wondering how far are you, is that going to be done any time soon or is it a several year project?

AMANDA: We are collecting glosses from the signed videos as well as sign times for all these videos and right now it's about 30% collected, but due to the situation, it was a little bit delayed. We were expecting to have this already done, but the it was delayed with the situation, so we guess that, I don't know, it's going to take some more months to finish, but we are planning to release the dataset, even before that maybe. If we can have some experiments done with the dataset that proves this is useful for some of the tasks that doesn't need the gloss annotations yet.   
SAMUEL: I think Xavier has joined to answer the question from Lucas.   
XAVIER: Yes, so actually the question was about how to improve the results of the hands, because that's where we spot we have struggled the most in generating and producing sign language, and actually what we are presenting here it is based on a well‑known work called everybody dance now and in that work they already have a model, the way it runs, they have a full frame and then you have known generator specific for the face. Right now, what we are trying to do is train, adapt that code, to have a generator specific for the hand. This should improve, we are not really sure if that's going to be enough, because there are more challenges than just high quality with the hands, so we are also looking in other directions on construction models of the hands to see if we can model that better. But I agree for now we are focussing our efforts there, because that's where we are struggling the most.   
SAMUEL: I think we can transition to the next set of papers. Slight modification of numerical order, will be paper number 7, multi‑modal machine learning approach and toolkit to automate recognition of early stages of dementia amongst British Sign Language users. Paper number 8, score level multi‑cue fusion for sign language recognition. Paper number 9, unsupervised discovery of sign terms by K nearest neighbours approach. And paper 10, improving keyword search performance in sign language by hand shape features. Again, we will individually promote the authors of papers to panellists. So, the first question is for paper number 7. This seems like a potentially very promising application of computer vision for sign language, because it has significant implications. I was curious to know what your next steps might be with the research.   
>> Thank you for the question. We have presented the next stage is to combine with the distribution model to combine three of the models for further prediction.   
SAMUEL: Next question is for paper number 8. Score level multi‑cue fusion for sign language recognition. The question is: Your analysis of different cues very interesting, you mentioned a few mechanisms that could be used to improve performance, given the critical role played by hands, do you have any thoughts on what kind of model might ultimately be best suited to exploiting this cue? For example, estimating a full 3D hand mesh, or perhaps an appearance‑based CNN classifier.

AHMET: We believe both ideas have merit. Currently from our experiments it seems that 3D convolution-based approach seems to be the way to go, especially for real time applications such as the translation app etc. Those methods seem to give the higher performance, but if real time can be achieved with reconstruction or other similar more complete representation methods, those would be useful as well.   
SAMUEL: The next question submitted for paper number 9, unsupervised discovery of sign terms by K nearest neighbour approach. Very nice work, as you note in your paper, a critical aspect of this kind of K nearest neighbours approach is the quality of the features. What kind of unsupervised learning objectives do you think are best suited to learning good embeddings for this task?

KORHAN: My connection is unstable, so I will not be here by video, I hope that's OK. I think the best way to improve the features is to use correspondence auto includers or something like Siamese networks, which might be able to learn a better embedding of sequence to sequence payers, so I think the direction might yield good results for learning other features.   
SAMUEL: Thank you. I think we can then transition to the next set of slides. Apologies, I have missed one me here. Which is for paper number 10. Improving key word search performance in sign language with hand shape features. Here the results are promising. You mentioned that one aspect where there is room for improvement is in the cue fusion, do you have any thoughts on what could be most effective in this setting, if you were to move to an earlier cue fusion strategy?

NAZIF: We think applying fusion, for example, at the level before merging the, before the keyword search, we think like first we can apply some encoders and then transition them into another domain using another network, but since this is going to be more like deeper architecture and we are going to need more data, we just couldn't do it right now. But the earlier the better we think, yes.   
SAMUEL: I think we can now move to the next set of papers. If we could promote to panellists the authors of the paper number 15, 3D hands, face and body extraction for sign language recognition. Number 16, effect of ranking and precision of results on users' satisfaction with search by sign line dictionaries. 17, finger spelling recognition in the wild with iterative visual attention. 18, accessibility for deaf and hard of hearing users, users sign language conversational user interfaces. I think we are missing one paper, paper 15. I am not sure if anyone is present. We will continue and return to them if we manage to locate them. So, for the meantime, we can pause the screen sharing. Thank you. The first question is for paper 16. Effect of ranking and precision of results on users’ satisfaction with search by video on sign language dictionaries. This seems really useful for deciding what kinds of metrics we should optimise for in achievable accuracy. Do you expect user preferences for other variables of the retrievable system, such as the amount of time it takes to retrieve results to exhibit similar characteristics to those found for written search engines? This could be great for informing the kind of latency we should be targeting when picking a computer vision model for the query engine. Is that something you have looked into from a user preference perspective?

SAAD: So that's a very good point and we did not look into it, so we just look at two factors, which was a presentation of the search result and also the overall events of the rest of the results, so the time you can find the results, we can definitely look into that, that is a good suggestion. I do expect a trend to be there for that. We looked at the video we collected on how the users scrolled through some of the search results, but we don't have anything, any quantitative results for that. It is something we can look at in the future. Thank you for the suggestion.   
SAMUEL: So next we have a question for paper number 17. Finger spelling recognition in the world with iterative visual attention. The question is: You mentioned in the paper that the newly collected dataset may have noisier annotations because it came from crowd sourcing. Do you have a quantitative sense in terms of how much noise there is and do you think your approach would benefit further from explicitly handling this noise, either through a robust loss formulation, or other strategies for data cleaning?

SHI: Thank you for the question. When we posted the video we add two annotators and we had a measure of it, I had a number but it was in the paper, in our paper, but when our previous experiment was just using part of the data, it was carefully annotated, so we compared the two and just you have a large amount of data it is noisier compared to without any noise, but it turns out, with a larger amount of data we get a better performance. It's hard to come up with a measure to, how to say, model the noise, but we just find the explicit role we have here is having a larger amount of data.   
SAMUEL: So, then we have a few questions for paper number 18. The first one was: This is very interesting, from the initial surveys that you conducted, were there any kinds of early takeaways or findings in terms of high priority scenarios for interacting with personal assistance systems?   
>> Yes from our initial surveys, we had people in the results from our interviews that seemed more interested in what you would expect in everyday use to come out close, like weather or the news etc. We saw some interesting applications that were deaf-specific, meaning only deaf people using that would be interested in that. VRS, for example, so they can make phone calls or video chatting with other deaf people etc. So, we noticed that deaf people were interested in those specific applications as well. And there's also applications of sound awareness in the environment, what noises are happening in one's general environment that they were interested in that too, maybe for a doorbell or fire alarm etc.   
SAMUEL: Thank you. Next question also for paper number 18 is from Xavier. Many users express concerns about audio only personal assistance due to privacy concerns. In the case of sign language, have users raised concerns about this and, if so, do you have any ideas about how to address them?   
>> Yes, so mostly in the interviews, we had open‑ended feedback. Participants said they were concerned about certain things, like what if the camera is looking at me all day long while I go about my business around the house and I am working. That was one concern. People mentioned if maybe there was another deaf person in their environment and they were assigned to that person, the person assistant might pick up on that. When the intention was not to actually sign to the personal assistance. In our survey we did ask people, how important is it to have physical cover for your webcam and I guess switch, you know like you have a switch on a microphone, maybe a switch on the cover for the webcam and everybody said yes that's very important. If people see there's a physical cover on the webcam, they know that camera cannot pick up on their movements and what they are conveying in sign language and they seemed more comfortable with that.   
SAMUEL: Oscar, did you have a hand raised earlier?   
OSCAR: Yes, but that question was already answered, thanks.   
SAMUEL: The next question is for paper 11, maybe we can re-promote the author of paper 11 back to being a panellist to enable them to answer it. This was also from Valentin. I cannot see Valentin as a panellist. I will read it out. You talked about signer independent training, I am afraid, I don't know that term, could you please explain a bit more about it please. That is from Jessica. Valentin is not here. So, we don't see any further questions in the queue, then I think we can move on to the closing stage of the workshop. So, Cihan, if we could progress to the next slide. So just to wrap things up, I would like to say thank you again very much for the invited talks, to the people who put the effort into preparing them. It is fantastic to get these different perspectives where we are, to Oscar, Lale, Christian and Bencie and if we can go to the next slide. A quick note to say that all the accepted papers are available at this website. And the videos will be made available after the conference. Lastly, just to summarise the organisers, in particularly Cihan who has put a monumental amount of effort into organising this workshop. Many thanks to him. Special thanks to Bencie, Oscar for their help and advice. Many, many thanks to the translators and interpreters and captioners for helping us to organise this workshop. Dr Robert Adam for his ASL and BSL translation. Our interpreting, Anna Michaels and Brett Best and Akbar and Esther Rose Bevan and thanks to MyClearText Katie Ryder and Tara Meyer for helping us with the captions. And finally, we would like to thank the sponsors Microsoft and Google for helping to provide the resources that allowed us to put on this workshop. Thank you very much to the audience for attending and we hope that you will enjoy the rest of ECCV. Thank you.   
That about concludes our programme, normally we have a small open discussion session, so if you have any questions you would like to ask to the keynotes, or just to discuss, please raise your hand. I can assign you as panellist, or then you can ask questions to the person you would like to ask. Otherwise, if nothing happens, I will say in five minutes, have a nice Sunday and thanks for joining. Oscar has a question.   
OSCAR: I wanted to ask Christian, I was wondering, do you think one of the papers brought it up here, the detection of sign language activity in a set‑up like Zoom or Teams and then silently voicing out silent voice or something, such that the programme actually picks it up and shows that a signer is signing, is that a useful application from your point of view?   
CHRISTIAN: I am happy you have asked that question, that's great and the short answer is yes. You have to be a little bit careful if you are working with an interpreter. Because if there's many interpreters on the screen, it's just, it won't, you won't know how to focus on one person. There are so many people there. In a situation where there's a lot of people like that, it can be a bit tricky. But in general, the answer is yes.   
SAMEUL: Thank you. Lale has a question.

LALE: Lots of advances were achieved in activity recognition action recognition by challenges. Do you think it is possible to collect new data and define new challenges and that would advance the field? Is anyone thinking of organising something like this?   
>> I would like to start answering this. The whole starting point of this workshop was organising a challenge to begin with! And we wanted to start by building blocks of the signs, so you can have linguistic annotated hand shapes and build up layers of having this kind of challenge. We didn't end up doing it due to organisations limitations and couldn't decide which data and annotation, but it is on our mind, if anyone else has an answer to this question, I am happy to give the floor to them, if you raise your hand for this.   
>> I would like to add something else, prior to that, we have to discuss what are the problems. It seems that segmentation is a very important problem from the perspective that it also helps build tools, because it seems that annotation seems to be a bottleneck for this kind of thing, so maybe we should also discuss the kind of problems and give priority or order them.   
>> Another thing of complimenting what Lale said, apart from having the problems, also the type of the data that the community wants or the communities thinks that is important, I think this is a great challenge now, because when I was creating the dataset, I was looking for a set of phrases or a set of things of real‑life that the community would want to have this type of data available and it was very hard to find it, it was very hard also to come up with this phrase, or come up with real‑life problems or real‑life situations that the community thinks that is the most challenging right now. And I think address this type of data and address the type of problems that we need to work on, on this type of data I think it would be great and we would accelerate this process for the community. If we can have a discussion with the community on what type of data or what type of real‑life situations is the best to work with, I think this would be a great help for future work in developing applications for deaf or even for collecting data.   
SAMUEL: Thank you for that, we also have Richard who raised his hand.   
RICHARD: I did raise my hand. First of all, on the challenge question, we did discuss a challenge and how much work it would involve. I said, I warned Cihan how much work a workshop was going to involve. But what I wanted to do was just make some observations from all of the talks that I have seen. Excellent today, it's been really good. And then maybe open it back to the keynotes and they can comment. So we've got Christian and Bencie who picked up on the same issues, and that is that there's a lot of work that goes on in our field, that goes on without really understanding what the problem is in the first place. And that's wrong. Christian said, well, let's aim it more towards these applications and that had some overlap with what Matt was talking about, even though we have a holy grail which is AI that understands everything, that actually there's something we can do with the technology now and so maybe we should be chipping off smaller parts of the problem, actually tackle real world issues that could be of use now. And then we've also got this engagement with the deaf community, and I think we have a bad reputation with the deaf community. So much so, because of all this press that goes on, when you do approach a deaf collaborator, they are already on the back foot. So really, we have to do the hard research, the longer-term stuff, and we should be maybe chipping off smaller parts of the problem now. And making sure that we are answering a real problem. But in terms of solving the issues, what should we be doing? If we were going to organise another workshop like this, then what should we do? Is it about educating the community? Is it about trying to bring these collaborators together? What would be the best thing that we could do with the next workshop? I will open that back up to the speakers.   
>> Always happy to chip in with something or other. I certainly think all those points are absolutely the case and one often feels one is running to undo some of the damage that's been done in the past. I think there's an important decision to be made about whether one has to cope with a small dataset of naturalistic language that really is the target, or go for the simplistic material, single sign material, very constrained sentence structure, it's kind of OK to do that, I mean, the constrained system, but only if you feel that what you are getting at is going to lead you on to being able to tackle the really big issues. And you can only understand if that's the case, if either you yourselves have good skills in sign language, so that you know whether you are asking the right question, or if you are working closely in collaboration with people who can answer that question for you. I mean, not just as informants or providers of data, but really close collaboration with deaf colleagues and I would also say with linguists, who are sign language linguists who can help you ask the right question. But in the end, I just think people need a different way of work can. We cannot continue to just regard ourselves at experts at choosing the topics of research. That's really critical.   
SAMUEL: There's a comment in the chat.   
>> Amplify that hands from get go, only proceed with projects and workshop planning whilst there's meaningful deaf leadership.   
SAMUEL: Are there any more questions anyone would like to ask to anyone? Jessica asking a question. I'm going to promote her to panellist. Jessica. Can you open your video?

JESSICA: Hello, so the question that I have is kind of one of etiquette, because I am working on a project that is getting external funding, I thought I was being polite and correct in not approaching the deaf community until I had funding, so that I wasn't asking them to do unpaid labour for me. The message I'm getting from the folks here is I should have reached out to the deaf community much earlier. How do I reconcile what to me seemed like two both important competing principles? I'm not sure who that question is to.

BENCIE: I cannot speak on behalf of deaf communities, so please don't take it that I am answering on their behalf, but I would have thought if you actually consider deaf people to be potential parts of your research team, then they are just as committed to getting involved in designing research projects that they would be involved in and will benefit from. So, it shouldn't be the case that you would see yourself as a researcher and there are some people out there you consult. You have to begin to think that you are involving academics or other collaborators in helping design your research questions. And if Christian could actually add to that, it would be really helpful.   
CHRISTIAN: My perspective might be slightly different. Involving deaf researchers in the process is essential, it's important. It might be slightly different to mention that, but Lorna did mention it and I agreed with her, deaf people need to be the leaders of the work. It needs to be deaf‑led. You need to be the appearance of the work, what you see, it's not just enough to involve them. Because deaf people own their sign language, it's theirs. It's the community's and own the applications, we have to realise it's not enough just to have them working in the process of training, but we have to be, we have to have that relationship, but really that's the most important thing.

CIHAN: Thank you for the answer. We are at 6pm now. We should be mindful of our interpreters who have been working hard, which I would like to thank again for everything, for being part of and accepting part of the workshop. Thanks again for joining. I hope you enjoyed the workshop. I learned a lot from the workshop, I hope you did as well. Hopefully, you will enjoy the rest of the ECCV and talk to you soon. Have a nice rest of Sunday.