Head Re-enactment for Sign Language synthesis

A project report by Savvas Apostolidis for the academic year 2020-2021.

Abstract— In this project, we evaluate the potential of head re-enactment methods to be used for Sign Language synthesis, more specifically the method of Head2Head++[DC21]. Our goal is to create and test scenarios involving Sign Language videos to discover what artifacts appear in the process of the re-enactment, what are the limitations of the current method proposed and which possible improvements can be suggested to further support Sign Language synthesis using head re-enactment methods similar to Head2Head++.

1. INTRODUCTION

Reenactment is the method of transferring the movements of a source actor to a target actor through video. State of the art facial reenactment methods have achieved very realistic expressions on real human faces using 3D face models and renderers, compared to the usual appearance of realistic expressions in digital avatars such as video game characters. What these methods lack is the ability to express motion on the head itself such as rotations and orientation, which problem is being tackled by head reenactment methods that support both facial reenactment and are able to copy head movements for more realistic and expressive results.

One of the many possible applications of such methods is Sign Language synthesis, a required step in the process of automatically translating Spoken Language to Sign. Currently, 3D avatars are being used as the interpreters which can introduce a low level of realism which, similarly to robotic synthetic voices, creates problems in interpretation and the engagement of Deaf users with these methods, for a characteristic example see here. Another issue is that these methods focus too much on the use of manual signs, which is the use of hand gestures to articulate messages, instead of focusing on other important ways of signing such as the non manual signs, which involve head and face motions. However non manual signs play an important role in the lexical meaning of a sign, the syntax and the prosody which is the patterns of stress and intonation in a language, for examples of non manual signs you can see here and here.

Another reason for studying the use of head reenactment for Sign Language synthesis is that Sign language involves sets of movements that create a very challenging scenario for head reenactment methods. Some examples are extreme head rotations, angles and speed of transitioning from one to another, occlusion from hands while signing and facial expressions that are not used in normal everyday speaking scenarios. Studying them is very useful for the development of these methods, alongside the important benefits of potentially creating technologies to assist deaf people.

The goal of our project is to test Sign Language videos on the target actors of the dataset of politicians that the Head2Head++ method provides, create our own target actors that better match the movements and expressions of the source actors to compare, gather information about what kind of artifacts can happen and possible ways to improve the results and study what are the limitations of this method.

Three stages of experiments were conducted. The first set of experiments tested the application of Sign Language videos on the original actors, the second set of experiments involved creating target actors custom to the Sign Language scenario in order to improve upon the re-enactment

results and finally a user study were users were evaluating the realism of the results and also asked to recognise reenacted signs in combination with self reenactment tests to compare speech to signing scenarios.

The results of the experiments helped us understand that some artifacts derive from the lack of coverage in the training footage, some derive from the camera setup while recording the footage and which are some of the limitations of this method. On another hand, the results also showed promise that head reenactment methods like Head2Head++ with some improvements can certainly be used to produce realistic and interpetable videos paving the way to the development of reliable Sign Language synthesis systems.

2. RELATED WORK

2.1 3DMM's and Facial reconstruction

In the field of face reconstruction one of the first researches was that of 3D morphable models by Blanz and Vetter [BV99], which has been adapted and used into many newer researches ever since for example [RA18,BJ18]. 3DMM's are parametric models that can generate 3D representations of human faces. Using anatomical correspondences of 3D facial scans, they can recreate never seen before faces as a linear combination of the training set.

2.2 Facial and full head reenactment methods

A pioneering facial reenactment method is Face2Face [TJ16] , which uses monocular 3D face reconstruction for both source and target videos and can also operate in real time. Some reenactment methods lack the ability of dealing with large variations of the head pose, controlling the target gaze and providing a mouth appearance similar to the target actor, such as the method of Averbuch-Elor et al. [AH17] which uses a target portrait photo instead of a video. Other methods do not adapt for the difference of the extracted landmarks of the source actor and the geometry of the target, creating identity mismatches in the generated video, such as Zakharov et al. [ZE19] which uses a few shot adversarial learning approach on a network conditioned on landmarks to synthesise the frames. More recently, Hyeongwoo et al. proposed the method of Deep Video Portraits [KH18], which uses an image-to-image translation neural network based on 3D facial information, hence an image-based head reenactment method. Due to the independence of the synthesised frames from one another, temporal incoherence can appear between generated frames and especially on the mouth region, which is a downside of the method and a common pattern for reenactment methods to have problems in the mouth region.

2.3 Image and video synthesis

For photo-realistic image and video synthesis, generative adversarial networks, referred to as GANs, have been used. Oftenly used data types for synthesis are class labels, such as in [MM14] or images, such as [IP17]. A well known method that utilises GANs for video synthesis tasks is vid2vid [WT18]. Similarly, both Head2Head [KR20] and in Head2Head++ [DC21] methods a GAN-based approach is used for rendering highly realistic video frames, in order to provide temporal stability. Our project is based on the latter method and we provide further details in the next section.

3. FULL HEAD REENACTMENT AND Head2Head++

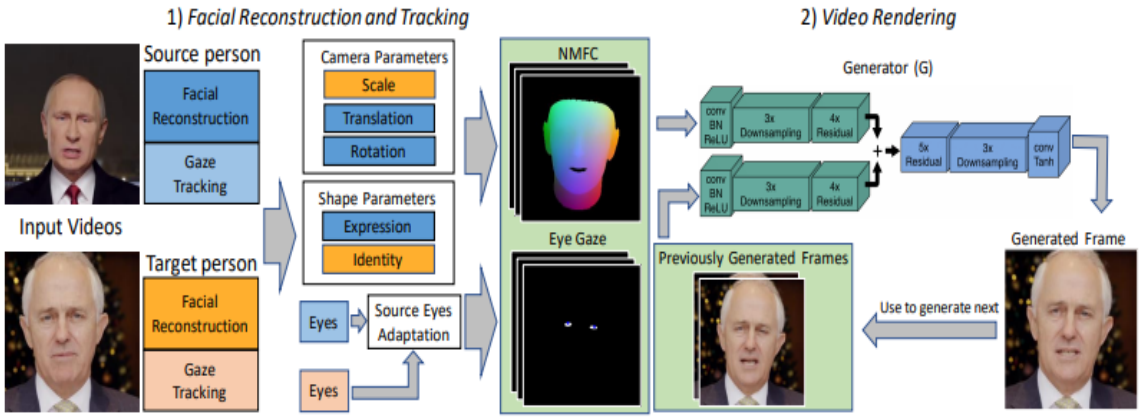


Fig. 1: The Head2Head pipeline[KR20]. 1)First the facial information of both source and target actors is extracted and reconstructed and the eye gaze of the source actor is tracked. 2) Then the NMFC and eye gaze sequences are used to drive the video synthesis. Figure taken from [KR20]

Both Head2Head [KR20] and Head2Head++ [DC21] methods improve upon state of the art head reenactment methods in the ability to track the eye gaze and the visual quality of the mouth area and by providing a more coherent and consistent representation of the identity of the actors.

The pipeline has 2 main stages, illustrated in Fig 1. Facial reconstruction and tracking stage and a learning based video rendering stage. The 3D reconstruction stage follows a batch based 3DMM approach to take advantage of the information contained in facial videos, extracting information about 3 sets of parameters, the identity, the expression and the camera parameters. For a set frame, given the estimated shape and camera parameters of the source and target actors, the identity coefficients and scale parameters of the source actor replace those of the target actor creating the ‘’hybrid’’ shape and camera parameters. After a more meaningful representation of the data in image space is being created, the normalised mean face coordinates of images or the so-called ‘’NMFC”, which is a rasterized version of the reconstructed face of the source after modifying it to take the identity of the target. Given a sequence of reconstructed NMFC images and the eye gaze, the video rendering network learns to translate the conditional input video into a highly realistic and temporally coherent output video showing the target actor performing exactly the expressions and motions of the actor in the source video. In order to train the models self reenactment is being used.

Limitations: One of the limitations is in the ability of Head2Head++ to track the head movements when the actor moves far away from the center of capture, illustrated in Fig 2, resulting in failure to detect the face and the method to produce a failure in the preprocessing stage. This error derives from the maximum distance of bounding boxes that limits head movements within a certain distance to the center of the camera , even if the head is visible in the footage.

Fig. 2: Example of the limitations of Head2Head++ method during the preprocessing stage. In this instance when the actor tilts her core during a sign, even if her body is still visible in the footage, in the detection phase the head goes out of bounds, resulting in inability to detect a face.

Another issue is in the preprocessing stage of the algorithm which imposes a time limitation on the duration of the training footage, see Sec. 4.3 for details.

4. Head2head++ REENACTMENT FOR SIGN LANGUAGE

In this section we evaluate the ability of the method to perform head reenactment in Sign Language scenarios. In order to set up for the testing we focused on the first step of the pipeline, which is the footage itself, both source and target. We conducted two stages of experiments and observed the resulting videos. Goal of the first stage was to evaluate how well the head reenactment performs when using source footage from Sign Language videos on already trained models of politicians, meaning on the original models that the method provided. Goal of the second stage was to improve upon these results by creating our own models with respect to the source footage and record any artifacts that persisted and which improvements were achieved.

4.1 Re-enactment using already-trained models on politicians

In the first stage of testing, we only control the source footage and test it on pre-trained models, in this case models of politicians, whose scenario was a speech. Two sets of videos were used for this stage of testing.

Fig. 3: Example of original footage between the source and target actor where it can be observed there is a vertical camera angle difference between the actors.

First set: The first set of source footage derived from the HealthSign dataset [KD20] containing numerous model actors and Sign Language movesets. During testing, it was discovered that due to a vertical difference in the camera angles that captured the footage of the actors in the database and the politicians, specifically the source actors having their chins in a higher position constantly as shown in Fig 3, it was impossible to produce stable results because every consecutive frame

had sudden changes in the estimated pose. Most commonly occuring artifact was a constant twitching of the head and the disability to reproduce the expressions on the faces of the resulting footage, illustrated in Fig 4.

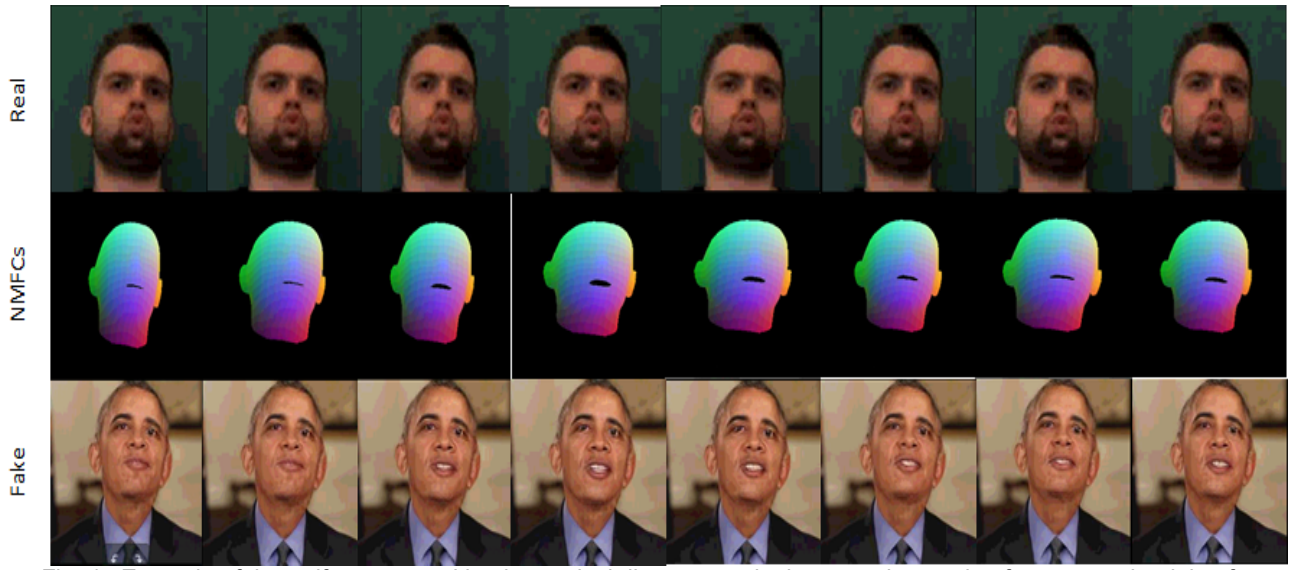


Fig. 4: Example of the artifacts created by the vertical discrepancy in the capturing angle of source and training footage. On the first row we see the source actor remaining stationary doing a sign involving frowning of the lips, but the NMFCs constantly change frame by frame producing the twitching resulting in the target actor's mouth and nose to move up and down.

Worst result was a combination of extreme contrast in the vertical rotation of the chin and the horizontal rotation of the head of the source actor which in combination with the target politician's opposite positioning of the head during the speech created a dysmorphic and constantly twitching result, illustrated in Fig 5. Head tracking seemed to accurately follow the source footage head positioning nonetheless in all cases.

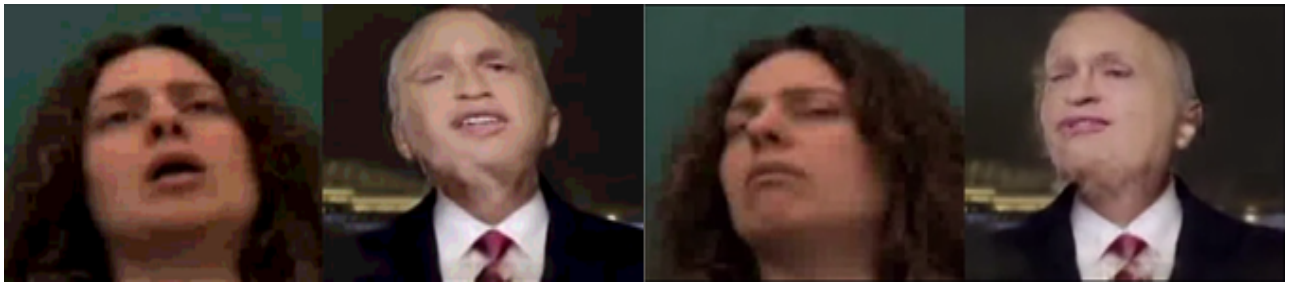


Fig. 5: Example of the worst result so far. Source footage mostly had the actor facing the upper left but the training footage never faced in that direction combined with the difference in the angle of capture creating a very challenging case.

Second set: The next set of source footage was videos of Sign Language selected from the platform of youtube containing a plethora of expressions and movements ensuring that the angle of capture would not create the same problem and any artifacts occuring would not be related to the previously stated issue.

Good results included the ability of the Head2Head++ method to greatly track head rotations and movements even in fast scenarios, resulting in seamless head movements when it was covered by the training material , keeping up in speed with the original footage. Fig 6 shows that lip tracking also seemed to work pretty well, which was evident by the ability of the politicians to mimic word mouthing such as the ''pow'' word and generally easy lip movements that even speech

could cover for. When the facial expressions of the politicians closely matched the source footage the results were good, further indicating that training footage is very important for the results. Small hand occlusions that did not cover important areas of the face such as the eyes and the mouth, did not create significant artifacts if any at all.

Fig. 6: Examples of promising results of head reenactment for Sign Language by the Head2Head++ method.

For bad results, as shown on Fig 7 below, it was observed that various expressions cannot be generalised from common speech and if not provided in the training material produce faulty results in the reenactment. Such expressions are frowning the lips, eyebrows or the forehead, the so -called ‘’duck-face’’ and puffing the cheeks. Another common artifact derived from head rotations not appearing in the training material. Sign Language often needs head rotations to convey the meaning and normal speech scenarios rarely move the head. This issue resulted in distorted reconstructions of the models face when the source actor rotated their head to sign and the politician face did not have material for these different angles. As to be expected, any expressions and mouth movements, such as grinning , dysphoria or biting the lips with the upper teeth, that are not provided in the training material can not be reenacted without artifacts which makes politician speeches not the most favourable training footage candidates. Closely related to head rotations is the movement of the core, such as downwards movements, which produced a challenging scenario for the politician footage that lacked such movements, resulting in further artifacts.

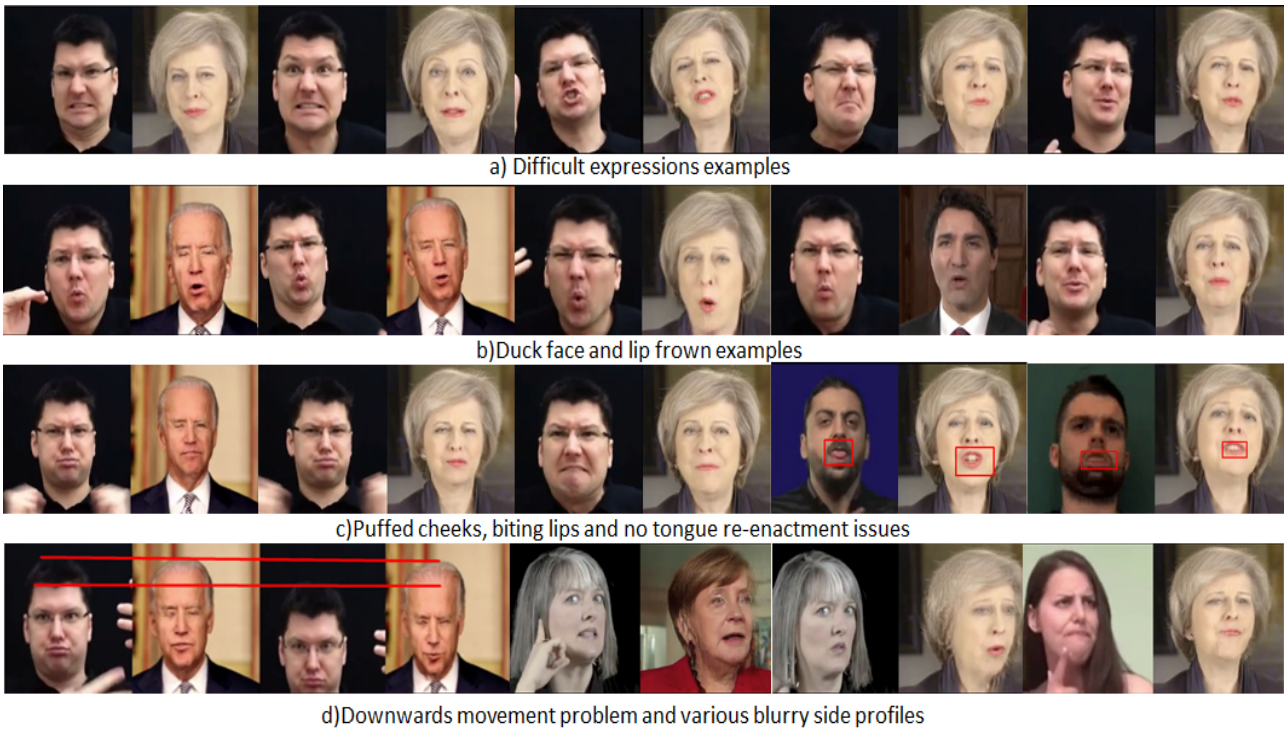


Fig. 7: Examples of artifacts produced with head reenactment on Sign Language scenarios by the Head2Head++ method. In each row we see examples of artifacts that occurred when using Sign Language videos as source and reenactment models trained to politicians as target actors. Depending on the training footage we can see that the reenacted face might be in an expressionless state or create a blurry side profile for head poses not included in the training footage.

Lastly, signs that involved hand occlusions that covered the entire face resulted in NMFCs changing predictions every frame similarly to the twitching artifact and the smoothening that the renderer applied created blurry results as a combination of the previous and current estimated pose, illustrated in Fig 8.

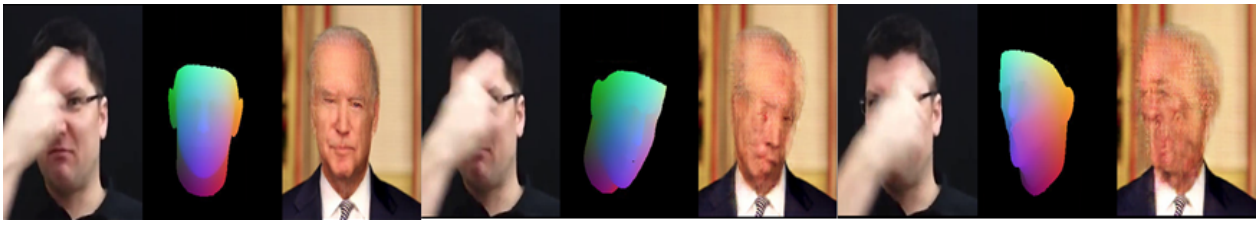


Fig 8: Frame by frame example of hand occlusions resulting in jumps of the predicted NMFCs and producing blurry faces.

4.2 Re-enactment training on novel footage

We conducted a set of trials of training novel models. Based on the observations from these trials, we conducted the training of the final models that we used in the Experimental Evaluation (see Section 5). For these trials, we captured new videos to create new training footage for head reenactment models. The purpose of creating new training footage was to target the artifacts that were observed in the first stage of testing by providing expressions, head angles and rotations and core movements that resulted in artifacts. The training footage can be divided in three main sets of trials: Freestyle general movements in an attempt to cover for most of the basic Sign Language movesets, targeted movements that copy Sign Language footage and using some of the source actors from the youtube videos as trained models.

Fig. 9: Example poses from the training footage for the first set of models.

Fig. 10: Examples of results of head reenactment for Sign Language on trained models created as part of this project specifically for this scenario. In this figure we can see that not every expression can be flawless with basic training footage of frowns and head movements. Tongue is not able to be trained and remains a problem. Side profiles are less blurry and core movements can follow the actor but introduce some blurring unless the sign is copied completely.

First set of trials of training novel models: In the first set of trials, two models were created by studying the source material and providing various possible head rotations and movements, as many of the expressions that were observed and copying the core movements that created ‘’extreme’’ cases to test if the artifacts can be reduced. Example poses can be observed in Fig 9.

The resulting footage showed that a lot of the expressions that were problematic in the first stage , like the ‘’duckface’’ or various frowns of the eyes, eyebrows and mouth improved. Side profiles no longer produced extreme artifacts when it was provided in the training material, same observation with core movement moving downwards, as illustrated in Fig 10.

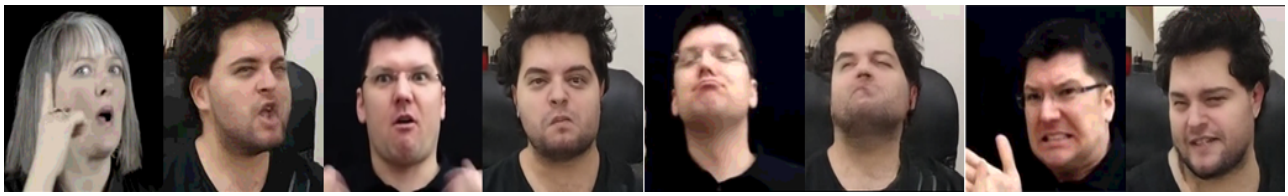


Fig. 11: Examples of bad results from untrained expressions and head poses.

In a similar way, as shown in Fig 11, artifacts still occurred with challenging head poses and rotations and unique expressions that did not fit in the training footage recorded style of basic lip and eyebrow frowns. Some basic speaking in the training footage covered for most of the basic lip movements that Sign Language needs, but some extreme word mouthing phrases cannot be covered without copying the movement.

Fig. 12: Examples of the blurred low quality model.

The second model, as shown in Fig 12, although it followed in the steps of the good results like the first one, was recorded in a bad lighting scenario and the video capture quality of the camera was not adequate, which introduced a lot of blurring in the results. Both models were captured by 1080p commercial phone cameras which had an impact in the quality of the reconstructed models, which can be seen in the examples shown above.

Fig. 13: Results from first set of clips on the first models. It can be observed that some expressions were not detected correctly.

Second set of trials of training novel models: The second set of trials of models included mimicking source footage from a Sign Language expert. Originally 7 clips were provided, and the movements involved in these were used as the target actor tried to copy them as close as possible to create the training footage. The results had much fewer artifacts, as shown in Fig 13, compared to the models of the first set that used general movements in an attempt to create a model for general purpose.

The results were not flawless, due to a minor discrepancy in the angle of capture of the first model and the Sign Language expert, in which case a second model was created to match the angle of capture, reducing even more the artifacts that were caused from this difference, illustrated on Fig 14.

Fig. 14: Results from the second model that was recorded with a matching camera angle to the expert, improvements can be observed.

As a next step, 7 new clips were provided and tested on these models, where the source actor was instructed to capture the footage from an angle closer to the eye level to help increase the accuracy of detected moves and to test the ability of these models to generalise.

Fig. 15: Results from the second sets of clips, it can be observed that the models possess some generalizing ability and show promise for the future of reenactment for Sign Language synthesis.

With the instructed camera angle results improved, as shown in Fig 15 and although not perfect, because artifacts still were produced when an expression was not provided in the material, were very promising.

Fig. 16: The ‘’fish-face’’ pose. It can be seen that the reenacted pose has defaulted to a basic lip frown which is the closest pose to cover the lip movements but the cheeks are wrong in both cases.

Lastly, there was an extreme case of the ‘’fish-face’’ pose which posed an extremely difficult scenario for the model, as illustrated in Fig 16.

In general the new 7 clips produced very good results and a possible reason is that all 14 clips were recorded with a simple use of facial expressions in Sign Language in mind, compared to the first models that were tested in various difficult head poses, expressions and angles or rotations. The new models were recorded with a more controlled environment in mind and were able to produce great results with only 1 minute and 20 seconds of training footage compared to the 2 minutes average duration of training footage of the first models which introduced the idea that for simple Sign Language scenarios it requires less footage to create a sufficiently good model. These new models were captured using a Gopro Hero 7 camera at 720p.

Fig. 17: Examples from the NMS actor reenactment clips. Due to the heavily specialised expressions in the training footage it can be observed that most expressions in the reenactment footage are influenced from the expressiveness of the actor instead of completely copying the source footage which indicates that even when the training footage is high level use of Sign Language, a good model that generalises to lots of expressions is not an easy task.

Third set of trials of training novel models: The third set of trials of models involved two of the source actors used as trained models by selecting two minutes worth of footage from their source material. The purpose of this test was to simulate the process of head reenactment in a speech scenario, where different speeches of politicians were used to train models and mixed the source footage between them to drive each one of them.

The first model ,which is called ‘’NMS’’, performed similarly to the first iteration of models. When the material of the source video appeared in the training footage the results were good, but this model did not generalise very well in general Sign Language, often producing artifacts due to the very specialised tone in the youtube video portraying very certain expressions and emotions, illustrated in Fig 17.

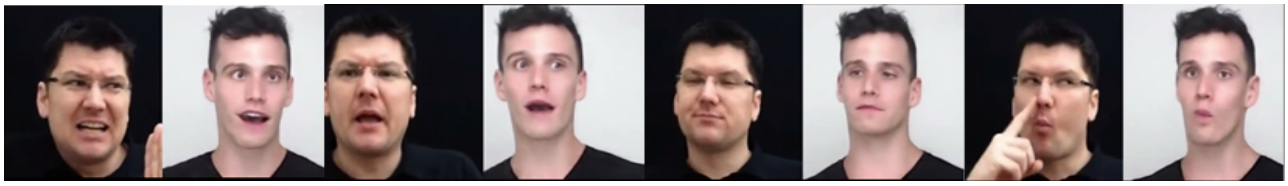


Fig. 18: Examples of the problematic second model of the third set of trials. It can be seen that although generally the poses were descently reenacted, the eyes did not follow a normal path and produced artifacts where the target actor would look into different directions.

The second model named ‘’ConC’’ was originally chosen because it provided a wide range of head rotations and had a lot of action in the lip movements, but the resulting model had failed to reconstruct the eyes properly and every output footage had eye artifacts that could not be ignored, thus a failed model, which can be seen in Fig 18.

4.3 General observations

Positive aspects: In general, throughout the testing it was observed that the method of Head2Head++ has the potential to produce very realistic reenactment expressions in this challenging scenario of Sign Language. Most of the artifacts observed during the first stage of testing were improved by targeting the training footage to better match the task at hand. Even the most challenging expressions could reach a high level of realism when included in the training material which indicates that the method is able to greatly reenact this scenario if the training stage is improved to allow a wider range of available facial expressions and head or core movements.

Limitations: The Head2Head++ method has some limitations in regard to Sign Language that cannot be covered for in the training footage. One of them is the tongue which is used in Sign Language but cannot be reenacted. There is also a limit to the extension of the head rotations

before reconstruction problems are introduced such as an extreme variation of the ‘’chin up’’ pose , if the head is raised too much for extended periods of time the method fails to reconstruct the NMFC images properly and twitching appears.

Training challenges: During the creation of training footage for new models, eyebrows are easier to train compared to the lips, because eyebrows have a smaller range of possible movements that can be predicted and trained for. To be specific, training for the lips can be difficult because there are a lot of possible words to be ‘’mouthed’’ that are unique from everyday speech and in order to reenact them successfully they have to be included specifically in the training footage, meaning there is a limit to how many signs can be covered during the available time. The same problem appears with expressions, certain expressions are very unique and cannot be generalised from basic movements that cover for the basics of Sign Language. Head rotations and positions can be covered effectively but the problem is the range limit that the method imposes before the actor goes out of frame. When recording the training footage, it is best if the movements of the actor are constant and fluent to help ensure the maximum amount of head rotations and positions and to help interpolate between facial expressions because it was observed how easily artifacts can appear due to the difficulty in predicting the possible movements a sign can require.

Recording of training footage: For recording footage settings, good lighting is important for the creation of the trained model, in order to capture the face from variable rotations and to prevent hiding eyebrow and lip movements during expressions. A good camera angle is at the height of the eyes or at least at the same height as the shoulders. For camera quality, 720p is sufficient with a good camera or 1080p with a more conventional camera such as a phone. Training duration differs with what is the goal of the model. For very specific coverage of certain Sign Language motions, the training footage can be under 1 minute to cover only the basic necessities for the movements in mind, but in order to maximise the generalising ability of the model, at least 2 minutes of training footage are needed, in order to have sufficient time to provide various head rotations and expressions in combination and also have enough time to produce lip movements for speech to cover for basic word mouthings. The machine used for training and testing was a 64gb ram setup with NVIDIA TITAN V gpu and specifically on the detection stage, during facial reconstruction, often ran out of memory when video duration approached the 2 minute and 20 second mark in 1080p which is the reason why only 2 minutes were used for the training footage at most. In general, challenging scenarios such as in Sign Language would benefit from longer training footage duration, in order to provide a wider range of expressions and head poses and rotations to create more complete models.

Challenging facial expressions of Sign Language: The ability of Head2Head++ to reenact expressions of Sign Language from source to target has some problems. One of them is a gap between an expression being provided in the training material but not being chosen as a good candidate expression that matches the source footage due to minor differences between the two expressions. This problem became apparent when creating models that directly copied the moves shown in the source footage but in the reenactment videos a lot of the times a pose that was trained for did not activate properly. Correlated with this problem is the inability of the method to ‘’stitch’’ different parts of expressions together. For example, if the source footage has a pose which requires squinted eyes and a duck face, if in the training material the exact same combination was not provided in the same pose the method could not pick two different expressions, one with squinted eyes and another one with the duck face and combine them to create the required pose, which would help a lot in the ability to reenact difficult poses that are required in Sign Language and also make training footage less demanding on the actor. The same problem also occurs with head rotations, an expression that does not appear in a certain head rotation will not be reenacted even if it was captured in a different head pose. These problems create a difficult scenario for the target actor when creating the footage to be trained. A person that does not use Sign Language in everyday life can not perform on a level required to activate

expressions that match source footage of Sign Language users and therefore its best if this gap was bridged, which will be discussed in better detail in the section of conclusions and future work.

Generalization ability of trained models: Studying the reenactment results of Sign Language videos shows that the generalization ability of trained models is not sufficient for this scenario with the current implementation of the method. It is best to contain the range of expressions and signs that a model can provide in order to have good results and based on all these observations from the intermediate testing we designed the protocol for the final experiments, which will be described in the next section.

5. EXPERIMENTS

The final experiments we conducted included two user studies and a self reenactment test where a model reenacts part of the footage that was not included in the training.

In the first user study we asked users to recognise Sign Language signs that were shown in the videos and pick the best fitting label to measure the interpretability of our created models. In the second user study we asked users to rate the realism of the reenactment footage, based on a set perfect score example using real footage from a Sign Language Expert.

The self reenactment experiment included 2 tests. The first test was a scenario of speech, where the target actor was asked to create a video talking to the camera for 2 minutes from which 66% of the footage was used to train the model, the same ratio of the Head2Head++ self reenact experiments. The second test was a scenario of Sign Language usage, where the target actor was asked to continuously sign for 2 minutes and the same ratio was used for training and testing.

5.1 Evaluation of Recognition of Facial Actions during Signing

This stage of the user study included three actors. One of the actors was a Sign Language Expert whose footage shown was real and not reenacted. The other two actors were models created for the experiments whose footage shown was strictly reenacted. Both models were trained based on a set of videos that were placed also in the set of test videos to recognise as part of the total videos shown, because as stated on the previous section containing the range of total signs greatly helps the results. The rest of the videos were from a new set of never seen before footage to measure the generalisability of these models. All three actors used the same eight clips, from which five, named Set A, were included in the training material in terms of mimicking the expressions shown in them and three, named Set B, were from the new set. The clips were shuffled but a static order of the actors was used in order to reduce bias in the selections as much as possible. For each video, each action was repeated three times and it was allowed to repeat the video as many times as needed. Average duration for each clip was 10 seconds but the length was variable since signs in Sign Language need a different amount of time between them to be complete and interpretable. For more details of how this stage was conducted please refer to APPENDIX A.

5.1.1 Results of the first stage of the user studies

Table 1: Recognition rates (%) over all annotators. Mean and standard deviation values are reported.

	Real Signer	Target 1	Target 2	**Target 1 & 2 (re-enactment)**
Set A	93.3 ± 10.3	73.3 ± 16.3	60.0 ± 0.0	66.6 ±8.1
Set B	50.0 ± 18.2	38.8 ± 25.0	22.2 ± 17.2	30.5 ± 19.4
All clips	77.0 ± 9.4	60.4 ± 14.6	45.8 ± 6.4	53.1 ± 10.2

By measuring the mean and standard deviation of the recognition rates over all annotators, as shown in Table 1, we understood that reenacted videos reduce the interpretability of Sign Language videos but the results were within a margine that shows promise for the future. The standard deviation showed that the results were not only based on the videos but also on the ability of the annotators to recognise actions in general due to the differences between the recognition rates on the real Signer footage, which pattern reappeared for the target actors as well.

There is a difference between the recognition rates of Target 1 and Target 2 and the reason for this difference was the training footage provided by both actors. Although they were instructed in the same way and provided the same sets of signs, the second actor lacked expressiveness in some range of movements, such as frowning the lips, which had an impact on the ability of the model to reenact the signs in enough detail to be completely interpretable.

In the second set of videos the recognition rate reduced a lot, which is a sign that shows that never seen before footage has an impact in the ability of a model to produce interpretable results.

Table 2: Recognition rates (%) per annotator.

	Real Signer Set A | Set B | All	Target 1 Set A | Set B | All	Target 2 Set A | Set B | All	**Target 1 & 2 (re-enactment)** Set A | Set B | All
Annotator 1	100 | 33 | 75	60 | 66 | 62	60 | 33 | 50	60 | 50 | 56
Annotator 2	80 | 66 | 75	80 | 66 | 75	60 | 33 | 50	70 | 50 | 62
Annotator 3	100 | 66 | 87	100 | 33 | 75	60 | 33 | 50	80 | 33 | 62
Annotator 4	100 | 66 | 87	60 | 33 | 50	60 | 0 | 37	60 | 16 | 43
Annotator 5	100 | 33 | 75	80 | 33 | 62	60 | 33 | 50	70 | 33 | 56
Annotator 6	80 | 33 | 62	60 | 0 | 37	60 | 0 | 37	60 | 0 | 37
min	80 | 33 | 62	60 | 0 | 37	60 | 0 | 37	60 | 0 | 37
median	100 | 50 | 75	70 | 33 | 62	60 | 33 | 50	65 | 33 | 56
max	100 | 66 | 87	100 | 66 | 75	60 | 33 | 50	80 | 50 | 62

In order to ensure that all results recorded were indeed reliable and did not have outliers we measured the performance of each annotator separately. As shown in Table 2, across all annotators there is not a clear outlier and all results were significant. There were occasions of exceptional results, as it can be seen in the maximum recognition rates, which further supports that the recognition rate also gets affected by the individual's ability to interpret Sign Language in general.

It can be observed that Set B performed poorly for both Real Signer and the Target actors which implied that there was another reason why the results of Set B were much lower than those of Set A. As shown in Fig 19 below, the third clip set B was not recognised by any annotators even for the Real Signer footage, which reduced the recognition rate for that set. A possible reason for this problem was a misinterpretation of an action based on the available options, whose labels we provide in Table 3, which could have introduced a level of difficulty when trying to discern between two different actions. For the rest of the clips, the number of annotators that picked the correct action for each clip usually had the most correct answers for the real target but both targets performed relatively well and on some occasions even better than the real actor.

Table 3: Labels of the Sign Language actions provided to the annotators as choices in the first stage of the user studies.

Set and number of label	Real Label provided	Equivalent label in the plots	Shortened equivalent
A, 1	ΜΑΚΡΙΝΟ ΠΑΡΕΛΘΟΝ: βλέμμα προς τα πάνω, κεφάλι προς τα πάνω	A.1.Μακρινο_παρελθον	A.1
A, 2	ΜΕΓΑΛΗ ΒΑΡΕΤΗ ΔΙΑΡΚΕΙΑ ΧΡΟΝΟΥ: κεφάλι μπρος-πίσω και στο πλάι	A.2.Μεγαλη_διαρκεια	A.2
A, 3	ΕΠΙΘΕΤΟ ΠΟΛΥ ΜΕΓΑΛΟ-ΧΟΝΔΡΟ: ύψωμα φρυδιών, άνοιγμα ματιών, φούσκωμα μάγουλων	A.3.Επιθετο_χονδρο	A.3
A, 4	ΑΡΝΗΣΗ 1: κεφάλι δεξιά-αριστερά	A.4.Αρνηση_1	A.4
A, 5	ΑΡΝΗΣΗ 2: κεφάλι προς τα πίσω, άνοιγμα ματιών	A.5.Αρνηση_2	A.5
B, 1	ΕΡΩΤΗΣΗ ΜΕΡΙΚΗΣ ΑΓΝΟΙΑΣ: σμίξιμο φρυδιών, σαγόνι μπροστά, χείλη προς τα έξω	B.1.Μερικη_αγνοια	B.1
B, 2	ΕΡΩΤΗΣΗ ΟΛΙΚΗΣ ΑΓΝΟΙΑΣ 1: ύψωμα φρυδιών, άνοιγμα ματιών, κίνηση κατάφασης κεφαλιού	B.2.Ολικη_αγνοια1	B.2
B, 3	ΕΡΩΤΗΣΗ ΟΛΙΚΗΣ ΑΓΝΟΙΑΣ 2: ύψωμα φρυδιών, σαγόνι μπροστά	B.3.Ολικη_αγνοια2	B.3

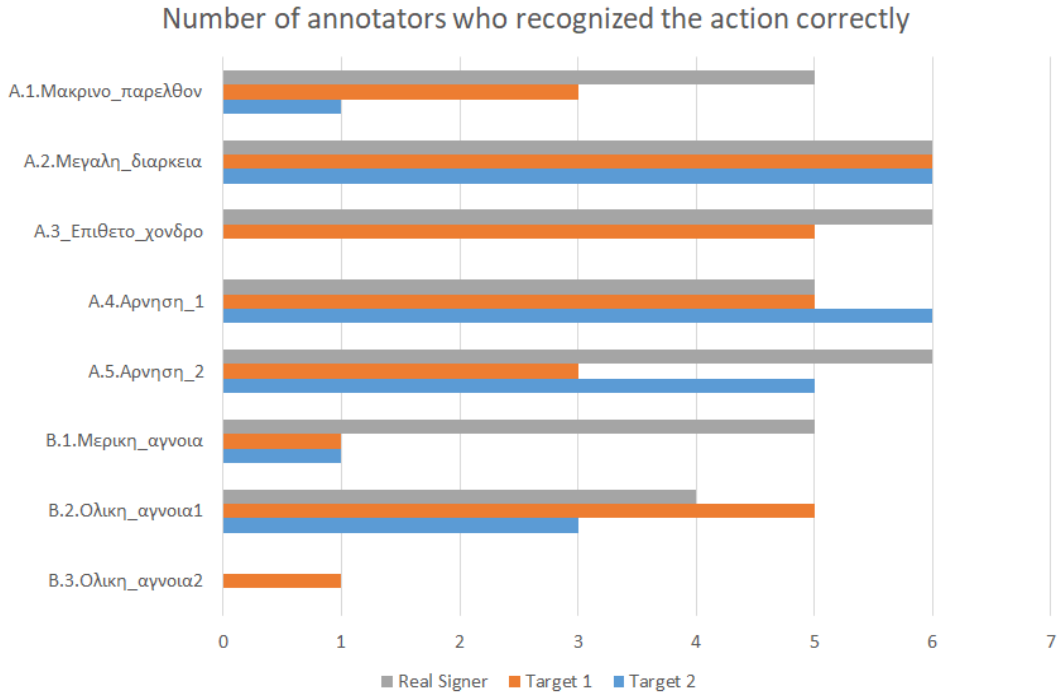


Fig. 19: Plot for the first stage of the user studies. For each row in the plot the number of annotators that picked the correct label for each clip is measured, up to 6 annotators.

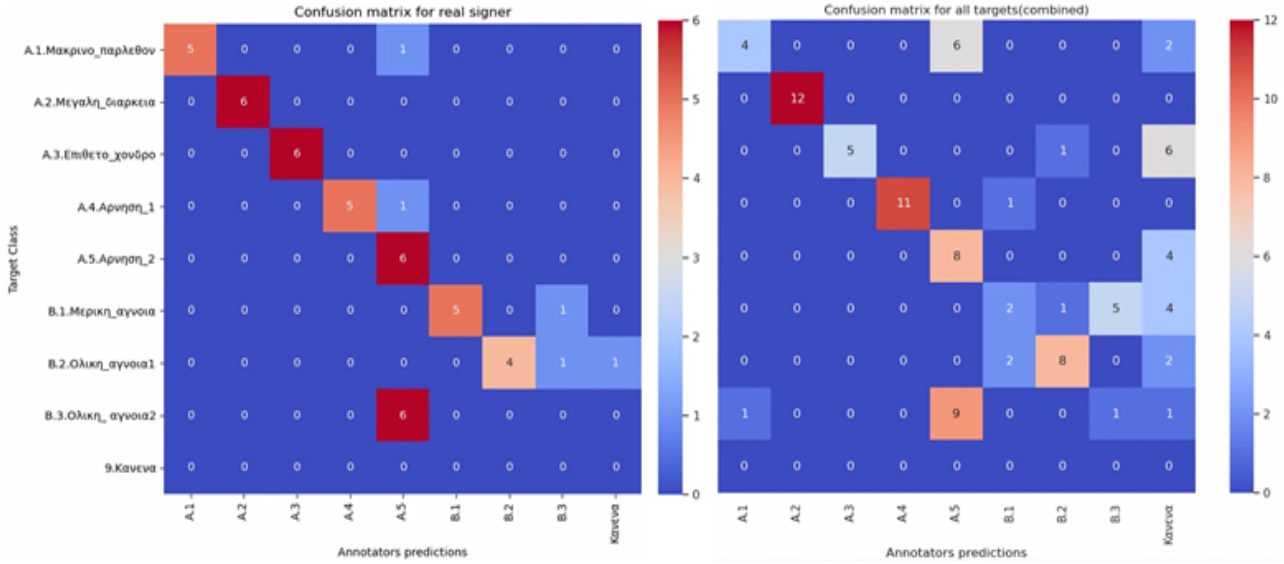


Fig. 20: Confusion matrices for the real signer (on the left) and for both targets (on the right) for the first stage of the user study.

In order to get a better understanding of the results, we also created confusion matrices for the real signer and both actors to get an overall view of the performance of the annotators.

As shown in Fig 20, on the Real Signer matrice (left one) across the diameter we can observe that for most clips, with one exception, the majority of the annotators were able to correctly recognise the sign, providing the goal in mind for the target actors. On the matrice for both targets (right one) we can see that for the first set of clips (A.1-A.5) the majority was able to recognise the sign correctly, but there were occasions were an action was misinterpreted for a different one and there were occasions were an action was not recognised (option 9, “none” or “none of the above” as was provided in the questionnaire).

On the second set of clips (B.1-B.3) we can observe that the number of correct choices reduced a lot, with occasions of misinterpreting clips within the set or not interpreting the clips at all. One important result was that of clip B.3, which we can see followed the exact pattern of the answers in the Real Signer, where the majority confused the action for the option 5 meaning that although it was the wrong answer it produced the same result as the real footage, implying that this was a problem of labeling rather than recognising an action correctly which is in fact a positive result that was not obvious with just numeric measurements of recognition rates.

In general, the second set results showed that never seen before footage produces reenactment videos that less often are recognised as a specific action, but rather the artifacts introduced with the lack of training footage creates results that often can be confusing and difficult to discern between them and interpret them correctly. For example in clip B.1 when the annotators were wrong the answers were spread across the clips more evenly than other occasions, but not to a degree that does not show promise as the results were better than expected, for example for clip B.2 8/12 answers were correct across the 2 targets for the 6 annotators and 9/12 answers were the same answer as the real signer, even if the answers were all misinterpreted due to labelling.

Finally, we investigated the confusion matrices per target actor in order to figure out how much the quality difference in the expressive ability between the two targets impacted the results and thus the overall view. As shown in Fig 21 below, Target 1 performed relatively better on the first set of clips, results that were close to the real signer’s. Both targets had problems in the second set of clips with Target 1 performing a little better on the second clip(B2). Although the general performance between the targets showed that Target 1, whose training footage resembled better the footage included in the clips of set A, performed relatively better on the second set of clips that involved never seen before footage, both models faced the same issues which showed that

performing better in a contained environment did not mean that the generalization ability of interpretation would also increase.

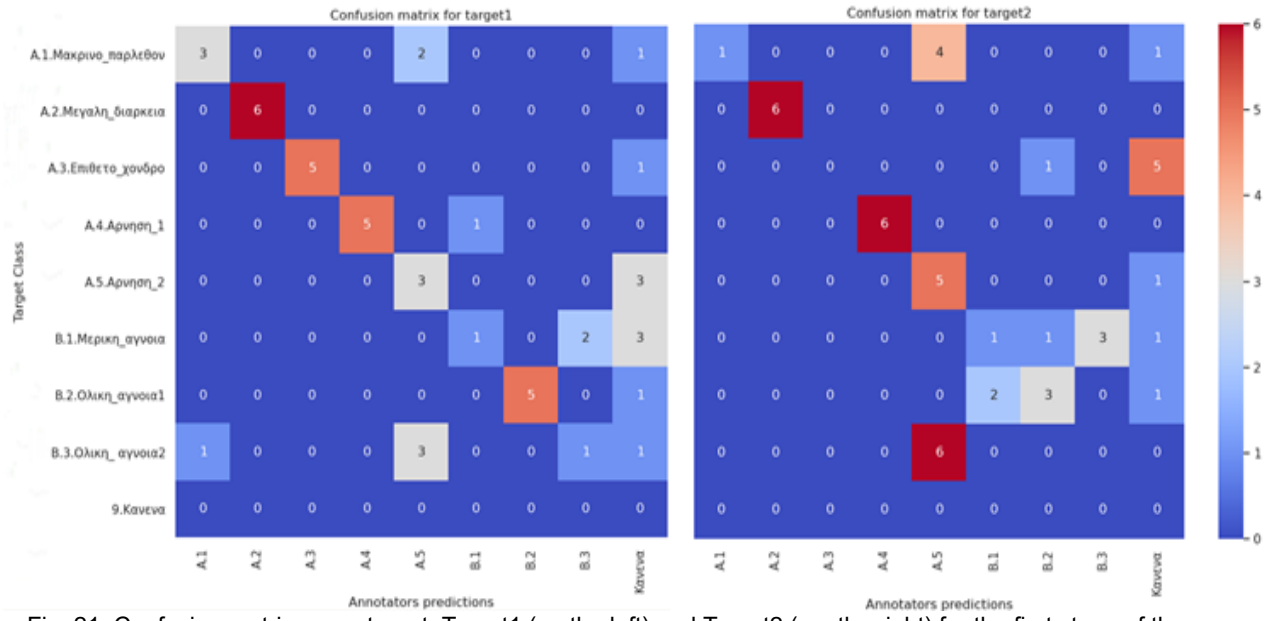


Fig. 21: Confusion matrices per target. Target1 (on the left) and Target2 (on the right) for the first stage of the user studies.

5.2 Evaluation of Realism of Synthetic SL videos

The second stage used the same models and randomly selected videos from both seen and never seen footage and asked different users from the first stage to evaluate the realism of the reenactment videos. In total the number of participants was 13. The scale used was 1 to 5, 1 meaning they considered the footage absolutely fake and 5 meaning that the footage was considered absolutely real. In total two videos of reenactment per target were shown, one for each set of clips and each video included two clips from that set. For each video, the first clip was common for both targets and each video in order had a different target from before, meaning they swapped in order of appearance, to reduce bias as much as possible. Before asking the users to rate the videos, they were shown some real footage of the Sign Language Expert to set the bar for what was considered to be the perfect mark of 5 for a realistic video. For more details on this stage please refer to APPENDIX B.

5.2.1 Results of the second stage of the user studies

The average realism scores for both targets, illustrated in Table 4 below, showed that on average the videos were neither considered absolutely fake or absolutely real. The standard deviation of 1 in a scale of 1 to 5 showed that the opinions of the annotators varied a lot and could possibly be up to personal standards and preference as to what they considered to be realistic, other than the models themselves lacking in realism while signing.

Table 4: Average score of Realism over two videos per target. Mean and standard deviation. Scale from 1 to 5

	Target 1	Target 2
First video shown	2.3 ±1.1	2 ± 1.0
Second video shown	3.1 ± 1.3	2.5 ± 0.9
Average score over Both videos	2.7 ± 1.3	2.2 ± 1.0

Table 5: Median (min, max) realism score over two videos per target.

	Target 1	Target 2
First video shown	2 (1,4)	2 (1,4)
Second video shown	3 (1,5)	3 (1,4)
Statistics over Both videos	3 (1,5)	2 (1,4)

As shown Table 5, every video for both targets involved answers from both sides of the spectrum in the scale of scores meaning that for some people the resulting reenactment videos were realistic enough to be considered real videos and for some other people the artifacts in the videos could not be ignored thus resulting in a score of absolutely fake. Although the second video for each target was from the set B, meaning the footage was never seen before, it scored better for both targets on average possibly due to the randomised selection of the clips from both sets that removed any bias in expressions from seen before footage that produced greater results within that set. Even if all signs were trained for in set A, not every trained pose was flawless providing further proof that even if an expression is known and trained for it can still impose a challenging scenario for reenactment.

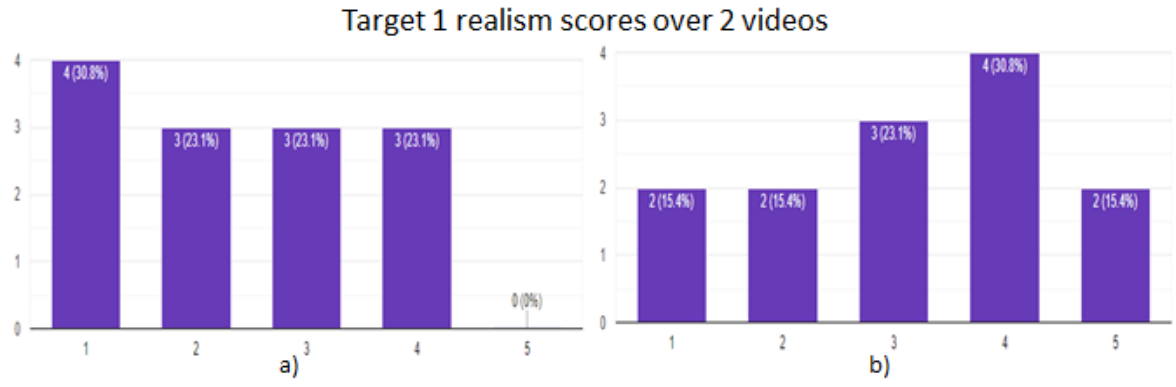


Fig. 22: Realism scores for target 1 on the second stage of experiments. A) is the first video shown belonging in set A and B) is the second video shown derived from clips of set B.

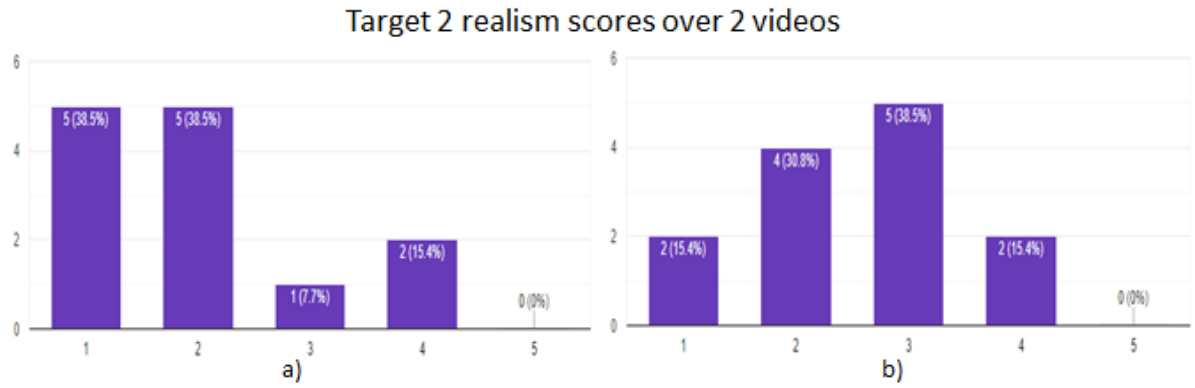


Fig. 23: Realism scores for target 2 on the second stage of experiments. A) is the first video shown belonging in set A and B) is the second video shown derived from clips of set B.

In figures 22 and 23 above it is shown that between the two models the first one performed much better on a realism scale compared to the second. In addition to interpretability, the level of realism is also greatly affected by the performance level of the target actor when recording the training footage, meaning that an actor who can perform expressions and signs better than another actor will produce more realistic results in his model even for never seen before footage. In order to provide an equal opportunity for all target actors to create realistic results, there is a need to detach the performance abilities of the target actor from the resulting model produced after training upon that footage, something that will be discussed in better detail on the section of conclusions and future work.

5.3 Evaluation of Self-reenactment Quality

Goal of the self-reenactment testing was to quantitatively evaluate the difference between the speech and signing scenario. Both resulting reenactment videos were compared pixel-by-pixel & frame-by-frame with their corresponding real test videos. We calculated three metrics to quantify the differences, of them two were included also in [DC21]:

Average Pixel Distance (APD): is computed as the average L2-distance of RGB values across all spatial locations and frames, between the ground truth and generated data.

Masked Average Pixel Distance (MAPD): similar to APD, it tests the reconstructive performance. A mask computed from NMFC frames is used to constrain the metric on the **facial area**, where conditional information is available.

Mouth region - Masked Average Pixel Distance (Mouth-MAPD): as MAPD, but based on the mask of the inner-mouth region (instead of a mask of the facial area, which excludes the inner mouth). This is to test specifically the realism of the inner mouth synthesis, which is challenging.

Table 6: Metrics of pixel differences during self-reenactment, averaged over all pixels, all frames and all videos. In all cases lower numbers mean better results

	APD	MAPD(face)	Mouth-MAPD
Continuous speaking	11.88	17.08	19.36
Continuous signing	12.57	21.16	24.61

As shown in Table 6, the scenario of signing resulted in higher measurements for all metrics confirming that signing is a more difficult scenario than speaking. Although the APD measurement was similar for both cases, it takes into account the background which is static for both scenarios meaning that the reason for similarly good results is the common factor of an unchanging and completely identical portion of each frame of the video. Through the Face-MAPD and Mouth-MAPD metrics it became more apparent that there is a definite difference between the two scenarios. On average signing introduced a 23 to 26% increase in the error measures.

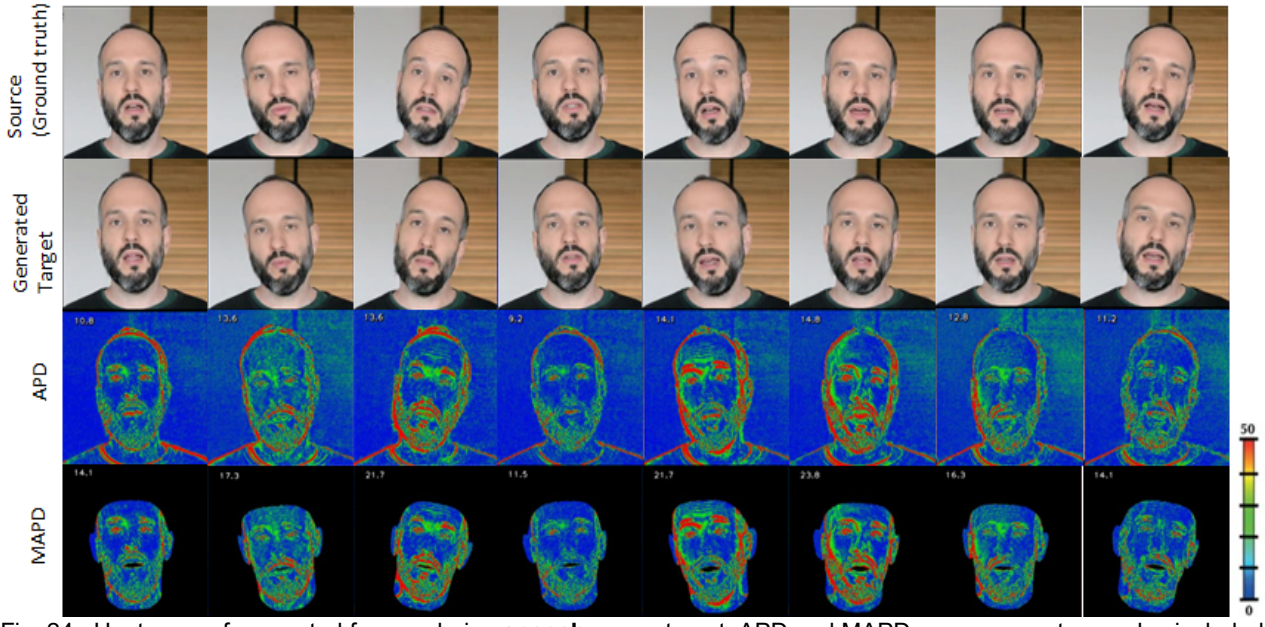


Fig. 24: Heatmaps of reenacted frames during **speech** reenactment. APD and MAPD measurements are also included.

During speech reenactment measurements were all within a margin of error around the average measurements, without any significant differences between each frame. The reason is that during a speech the movements that a person performs do not include a wide range of expressions or head movements which creates a fairly easy challenge for head reenactment techniques. As shown in Fig 24, there was a variation in the measurements during certain head rotations and poses, implying that in general rotations of the head always introduce some slight errors as it is difficult to train for every possible movement, but there was not an extreme error measurement that created significant artifacts. As a measure of reference, in [DC21] the measurements of APD and MAPD for the speech scenario produced similar results with our measurements within a small margin of error that is based on multiple factors such as camera quality, length of footage, lighting during capture and the fact that a different amount of models were used to measure the results.

As shown in Fig 25 below, during Sign Language reenactment there is a wider range of measurements. Some expressions and head positions resulted in low measurements implying that the reenactment was successful but there were occasions where extreme measurements were recorded during a significant artifact. Usually frontal positioning resulted in smaller measurements,

where only artifacts in the expression itself resulted in an increase of error measurements. Side-ways positioning resulted in bigger measurements, when there were artifacts, due to the combination of errors in both the expression of the face but also in the reconstructed shape of the head from either blurring or missing parts of the head, as shown in the fourth column of Fig 25 which also resulted in one of the highest measured errors.

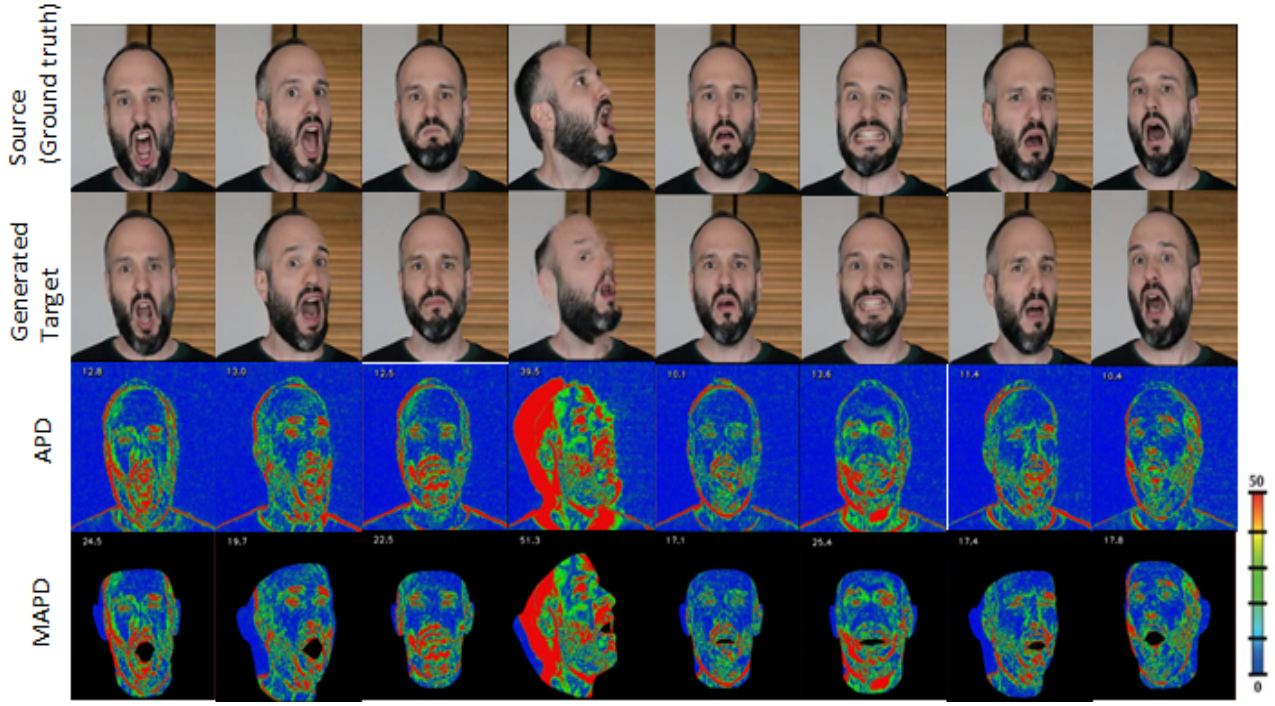


Fig. 25: Heatmaps of reenacted frames during **Sign Language** reenactment. APD and MAPD measurements are also included.

6. CONCLUSIONS AND FUTURE WORK

To conclude, the results of head reenactment for Sign Language scenarios using the method of Head2Head++ showed that the method can perform exceptionally when expressions and movements have been trained for but it is very difficult to train for a wide range of them. In order to improve the performance of the method, two important areas to target should be the ability to better match expressions in the source footage with selected expressions available by the models and to make training new models more accessible to the average user who does not know Sign Language.

As discussed in section 4.3, challenging poses that require combinations of previously seen expressions currently are not supported and force the target actor to provide more footage than needed. A solution to this problem would be to segment the head of the models into sectors such as the eyes, the eyebrows, the mouth , cheeks and the forehead and stitch different expressions from each sector together from provided material to accommodate the needed outcome, improving the ability of the method to more accurately reenact challenging expressions.

Improving the training stage could be done with two different possible solutions, one of them being better than the other.

The first solution entails lengthening the available training time from the current limitation of around 2 minutes, due to implementation issues, which would allow the target actor to provide more material and enforce the generalizability of the model. But, as discussed in section 5.2.1, a remaining issue of this solution is the performance abilities of the target actor that have an impact on both interpretability and the realism of the results.

The second solution targets both problems and entails detaching the performance of the target actor from the quality of the model created. In order to achieve that, the training method should follow a few-shot learning approach for the creation of the models, which is stated as a planned course of action in the method of Head2Head++, that allows creating models using only a few frames of the actor. The remaining problem of this solution would be how to provide all the complex expressions that Sign Language is compromised from.

In order to provide material, in the same way the current proposed method of Head2Head++ works, a database of Sign Language data should be created and used in a similar manner that the database of Face Forensics++ [RA19] provides material to improve the quality of models in the currently proposed work.

Lastly, remaining problems include the limitation on the range of core movements the actors can do before going out of frame, which problem can be solved by implementing a wider threshold of what is considered out of frame for the method, and hand occlusions creating artifacts in the reconstructed frames. Hand movements are vital for interpreting signs and should be supported in reenactment videos. In order to combine future implementations of hand reenacting with the proposed method of Head2Head++ there is a need to improve the robustness of the method against occlusions.

For future work it is important, as mentioned above, to provide support for hand reenactment to create models more accustomed to the scenario of signing. Another aspect of signing that needs support is in-mouth movements, such as the tongue. Finally, there is a need for much larger scale user studies, both in number of tested videos and in number of participants, to be conducted that will help gather more concrete information about the performance of the method in this challenging scenario of Sign Language reenactment.

References

[BV99] Blanz, V. and Vetter, T., 1999, July. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (pp. 187-194).

[RA18] Koujan, Mohammad Rami, and Anastasios Roussos. "Combining dense nonrigid structure from motion and 3D morphable models for monocular 4d face reconstruction." *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*. 2018.

[BJ18] Booth, James, et al. "3D reconstruction of “in-the-wild” faces in images and videos." *IEEE transactions on pattern analysis and machine intelligence* 40.11 (2018): 2638-2652.

[TJ16] Thies, Justus, et al. "Face2face: Real-time face capture and reenactment of rgb videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[AH17] Averbuch-Elor, Hadar, et al. "Bringing portraits to life." *ACM Transactions on Graphics (TOG)* 36.6 (2017): 1-13.

[ZE19] Zakharov, Egor, et al. "Few-shot adversarial learning of realistic neural talking head models." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

[KH18] Kim, Hyeongwoo, et al. "Deep video portraits." *ACM Transactions on Graphics (TOG)* 37.4 (2018): 1-14.

[MM14] Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." *arXiv preprint arXiv:1411.1784* (2014).

[IP17] Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[WT18] Wang, Ting-Chun, et al. "Video-to-video synthesis." *arXiv preprint arXiv:1808.06601* (2018).

[KR20] Koujan, Mohammad Rami, et al. "Head2head: Video-based neural head synthesis." *arXiv preprint arXiv:2005.10954* (2020).

[DC21] Doukas, Michail Christos, et al. "Head2Head++: Deep Facial Attributes Re-Targeting." *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2021).
[KD20] Kosmopoulos, D., et al. "Towards a visual Sign Language dataset for home care services." *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020.
[RA19] Rossler, Andreas, et al. "Faceforensics++: Learning to detect manipulated facial images." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

APPENDIX A

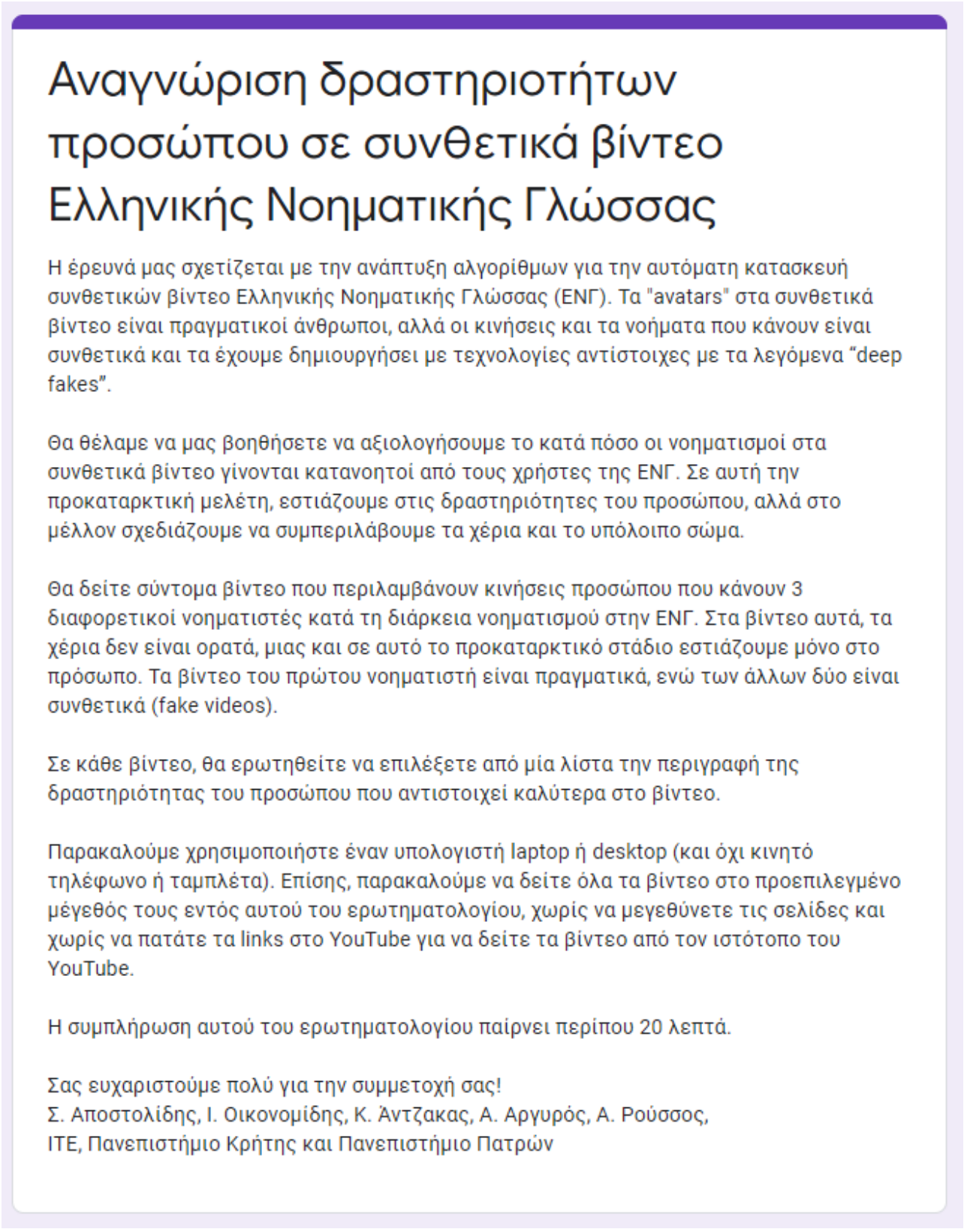
Evaluation of Recognition of Facial Actions during Signing (user study)

Introduction

During this stage of the user study, each user was asked to recognize facial action for a total of 24 clips, 3 targets in total 8 clips each. All 8 clips had the same source footage for each actor and one actor presented real footage while the other two actors presented reenacted footage. Five of the clips were used to help create the models for this stage and three of the clips were never seen before footage, namely set A and set B but the users were not informed of the set that each clip belonged to.

At first, the participants were shown an introductory page that explained the purpose of the study and the process of the evaluation, illustrated in Fig 26. For each page of the questionnaire, a singular video of the total pool was shown, in shuffle per clip but in order per target in turns of Real signer - Target 1 - Target 2 to prevent any bias. For every video they were asked to first watch the video, which repeated each clip 3 times, and after to choose which of the 8 provided labels in a multiple choice setup better fitted the action shown. In case no action was recognised they were also given the option to vote that no action from the displayed list of options fit to the video. An example of the pages is illustrated in Fig 27. The total number of participants was 6.

The link to the user study can be found here.



Αναγνώριση δραστηριοτήτων προσώπου σε συνθετικά βίντεο Ελληνικής Νοηματικής Γλώσσας

Η έρευνά μας σχετίζεται με την ανάπτυξη αλγορίθμων για την αυτόματη κατασκευή συνθετικών βίντεο Ελληνικής Νοηματικής Γλώσσας (ΕΝΓ). Τα "avatars" στα συνθετικά βίντεο είναι πραγματικοί άνθρωποι, αλλά οι κινήσεις και τα νοήματα που κάνουν είναι συνθετικά και τα έχουμε δημιουργήσει με τεχνολογίες αντίστοιχες με τα λεγόμενα “deep fakes”.

Θα θέλαμε να μας βοηθήσετε να αξιολογήσουμε το κατά πόσο οι νοηματισμοί στα συνθετικά βίντεο γίνονται κατανοητοί από τους χρήστες της ΕΝΓ. Σε αυτή την προκαταρκτική μελέτη, εστιάζουμε στις δραστηριότητες του προσώπου, αλλά στο μέλλον σχεδιάζουμε να συμπεριλάβουμε τα χέρια και το υπόλοιπο σώμα.

Θα δείτε σύντομα βίντεο που περιλαμβάνουν κινήσεις προσώπου που κάνουν 3 διαφορετικοί νοηματιστές κατά τη διάρκεια νοηματισμού στην ΕΝΓ. Στα βίντεο αυτά, τα χέρια δεν είναι ορατά, μιας και σε αυτό το προκαταρκτικό στάδιο εστιάζουμε μόνο στο πρόσωπο. Τα βίντεο του πρώτου νοηματιστή είναι πραγματικά, ενώ των άλλων δύο είναι συνθετικά (fake videos).

Σε κάθε βίντεο, θα ερωτηθείτε να επιλέξετε από μία λίστα την περιγραφή της δραστηριότητας του προσώπου που αντιστοιχεί καλύτερα στο βίντεο.

Παρακαλούμε χρησιμοποιήστε έναν υπολογιστή laptop ή desktop (και όχι κινητό τηλέφωνο ή ταμπλέτα). Επίσης, παρακαλούμε να δείτε όλα τα βίντεο στο προεπιλεγμένο μέγεθός τους εντός αυτού του ερωτηματολογίου, χωρίς να μεγεθύνετε τις σελίδες και χωρίς να πατάτε τα links στο YouTube για να δείτε τα βίντεο από τον ιστότοπο του YouTube.

Η συμπλήρωση αυτού του ερωτηματολογίου παίρνει περίπου 20 λεπτά.

Σας ευχαριστούμε πολύ για την συμμετοχή σας!
Σ. Αποστολίδης, Ι. Οικονομίδης, Κ. Άντζακας, Α. Αργυρός, Α. Ρούσσος,
ΙΤΕ, Πανεπιστήμιο Κρήτης και Πανεπιστήμιο Πατρών

Fig. 26: Intro page of the questionnaire for the first stage of the user study.

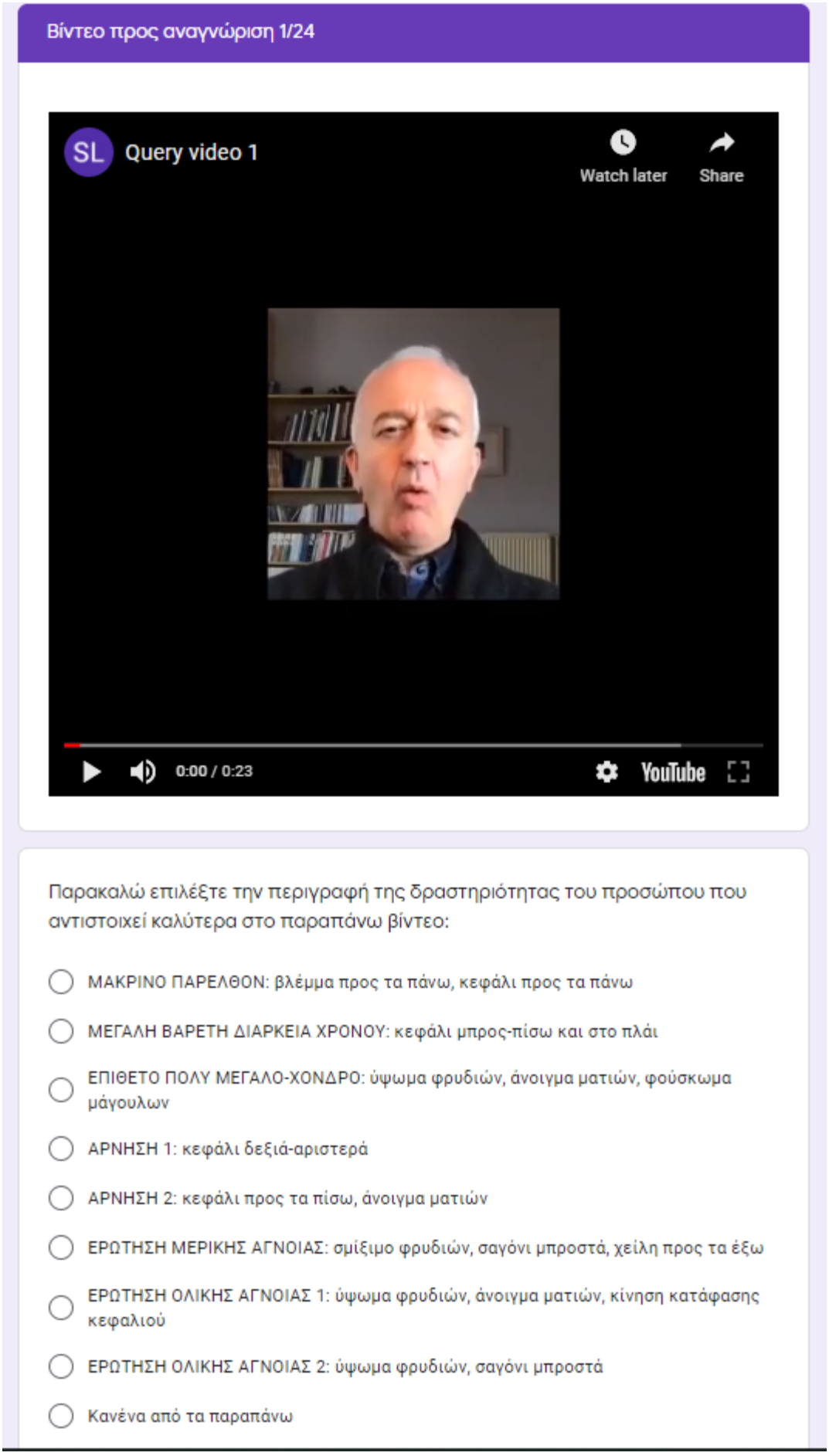


Fig. 27: Example page of the multiple choice selection in the first stage of the user stage.

APPENDIX B

Evaluation of Realism of Synthetic SL videos (User study)

Introduction

For the second stage of the user study, each participant was tasked with rating 4 videos based on a realism scale of 1 to 5, where 1 meant absolutely fake and 5 meant absolutely real. Two target actors in total were used for this stage of the experiment, the same models as in the first stage. For each actor two videos were created, compromised from two clips per set for each video. The two sets used for this stage were the same sets used in the first stage. Each video began with the same source footage on the first of the two clips in order to preserve bias in the results. The targets followed a set order of Target 1- Target 2 in order to prevent bias in the scores.

In a similar manner with the first stage, the questionnaire began with an intro page that explained the purpose of the user study and the process, illustrated in Fig 28 . After the intro page, an example video of real footage was provided to set the perfect grade of 5 to which the participants need to compare when scoring the videos on a realism scale, illustrated in Fig 29. Each page after included one video which repeated the two clips twice and then asked users to rate the video based on how realistic it was, illustrated in Fig 30.

The link to the user study can be found here.

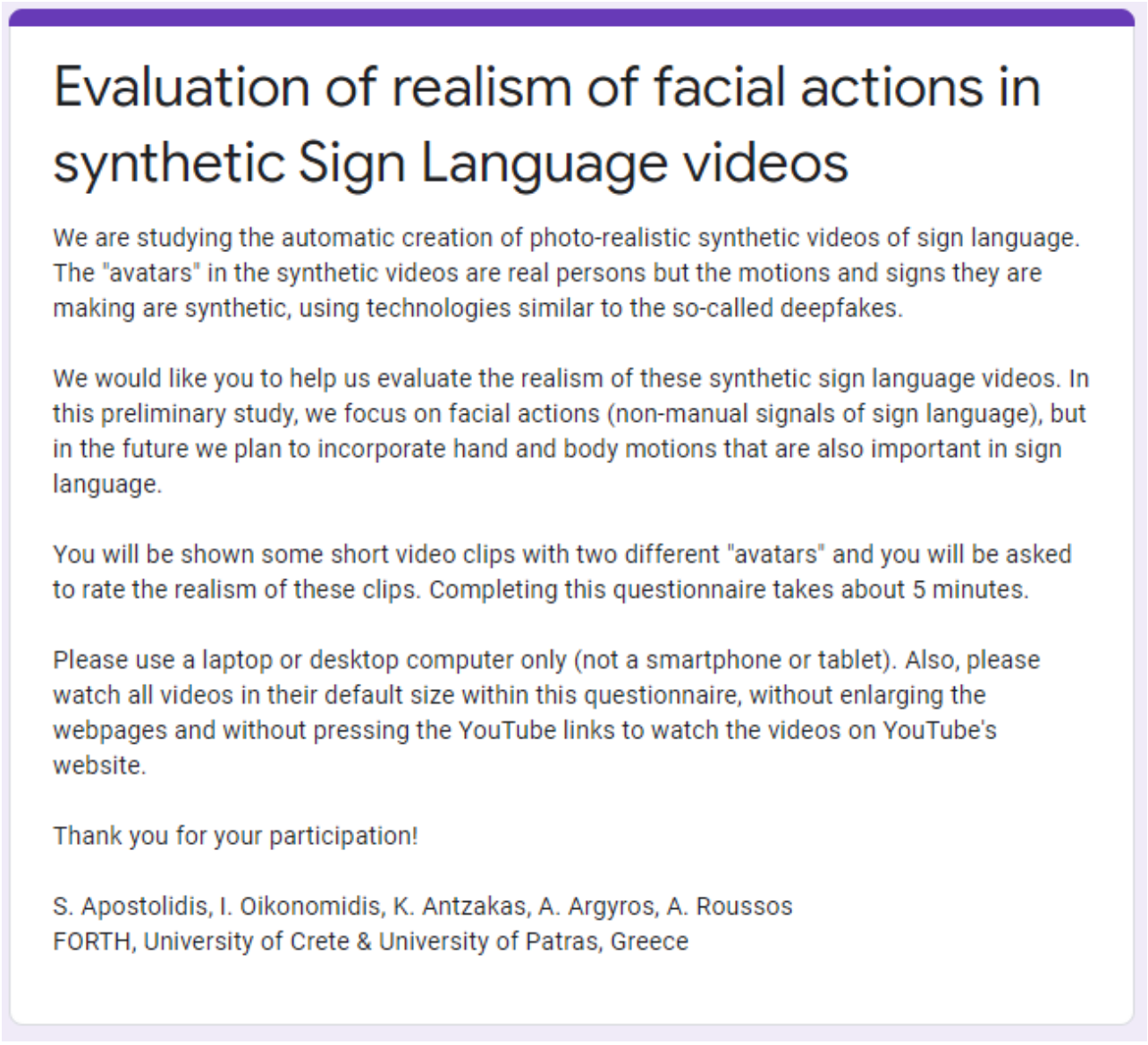


Fig. 28: Intro page for the second stage of the user study

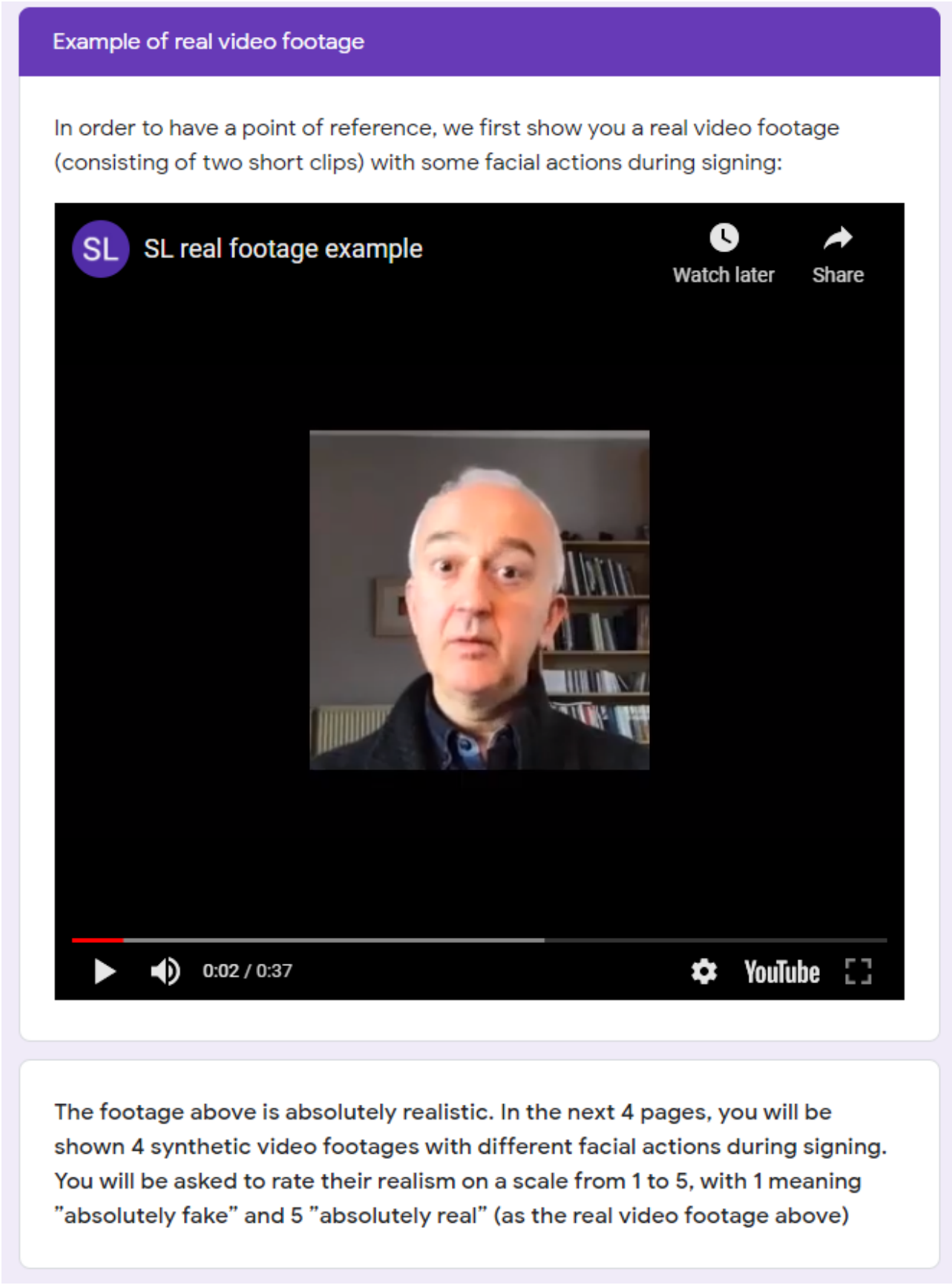


Fig. 29: Example video of real footage for the second stage of the user study.

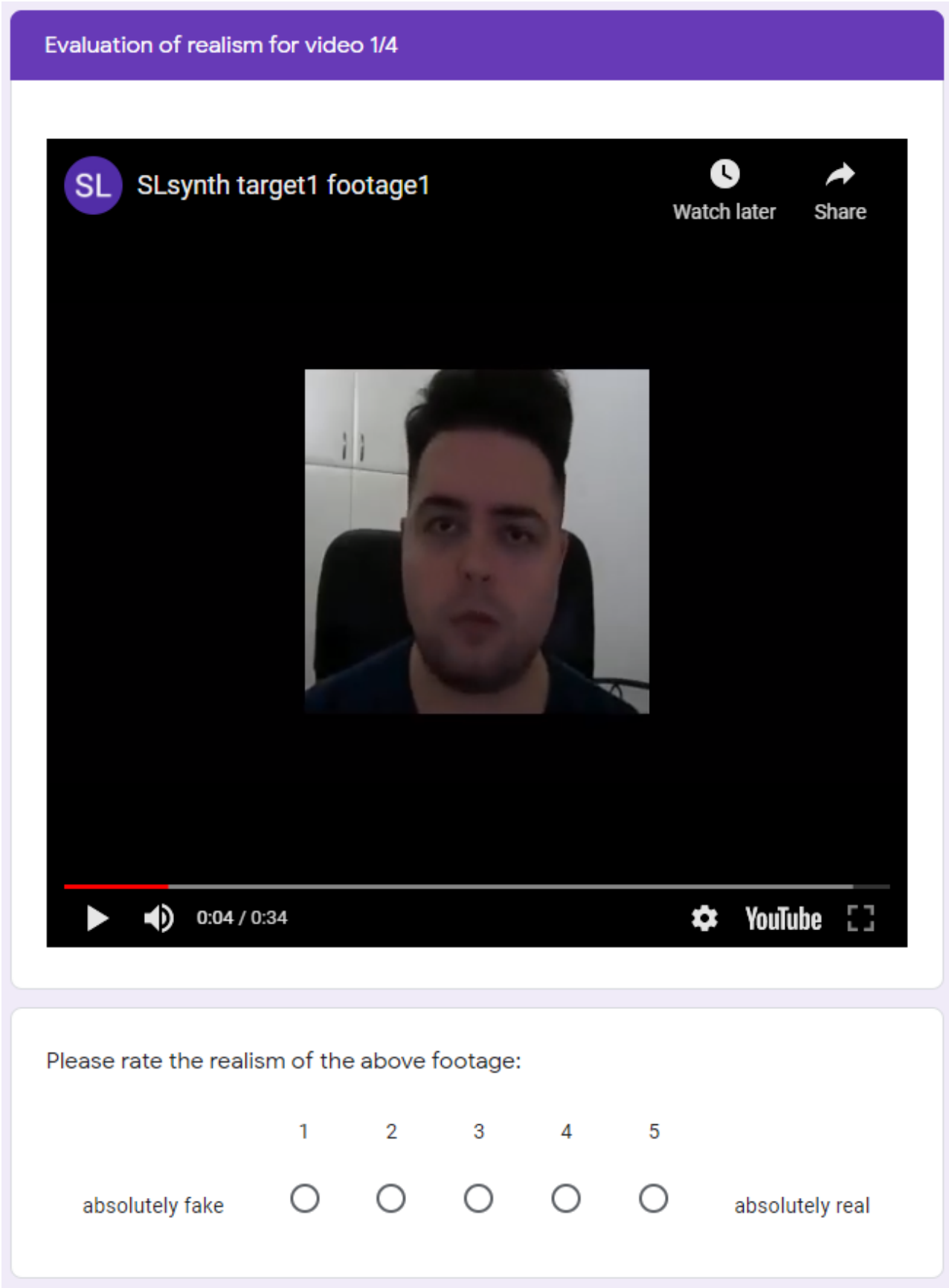


Fig. 30: Example page of the realism score questions for the second stage of the user study.