Revisiting Concatenative Video Synthesis with Relaxed Constraints

Sangyong Gil
Computer Science Department
KAIST
Daejeon, South Korea
sygil@nlp.kaist.ac.kr

Jong C. Park
Computer Science Department
KAIST
Daejeon, South Korea
park@nlp.kaist.ac.kr

ABSTRACT
A number of sign language translation systems have recently been proposed with the shared goal of providing improved accessibility for the deaf people. However, they are still largely in the stage of achieving correctness in a very limited domain, and much further work is needed to achieve naturalness with full emotion and non-manual signals as compared to human signers and, equivalently, full video. Concatenative video synthesis has earlier been proposed to address scalability of full video, but the technique has become less popular with associated shortcomings. In this paper, we propose to improve it by relaxing certain constraints, such as on signer and background, and see how the proof-of-concept videos are perceived by the deaf people. The result of this study shows that the revised concatenative video synthesis may provide a solution with adequate naturalness and complementary to sign language translation systems.

Categories and Subject Descriptors
K.4.2 [**Social Issues**]: Assistive technologies for persons with disabilities; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems- Video

General Terms
Design, Human Factors, Experimentation

Keywords
Korean Sign Language (KSL), Sign Language Translation, Concatenative Synthesis

1. INTRODUCTION
Sign language is the primary language of the deaf people, which is well known as quite different from spoken languages, for instance from the perspective of grammar and in its extensive use of spatial dimension. Learning spoken language is difficult for the deaf people, with reasons comparable to, if not more complex than, those that second language learners have in general. Consequently, the deaf people have much difficulty in acquiring knowledge as expressed primarily in spoken and written languages. This presents a serious problem for the deaf people from achieving adequate information access.

To deal with this problem, quite a few sign language translation systems have already been proposed to turn written language expressions into sign language animations [1-5]. These systems convert expressions in source written language into scripts that correspond to the meaning or the movements as defined in target sign language. The scripts are then used as directives for creating video or computer animation, with the help of a sign language lexicon and other rules that combine individual movements. The end result of this process is computer animation where a 3D human-like character, or an avatar, performs a sign language gesture that corresponds to the source language expression. The individual movement of an avatar's motion is based on the lexicon of a sign language translation system. The lexicon contains various sign language gestures in a unit of either words or phones, where the unit size depends on how the avatar animation is synthesized [3].

The previous studies have shown that the animation generated from the motion description of these units is now considered acceptable by the deaf people, at least if we focus on the manual signals alone. Furthermore, there are more recent researches [6,7] to devise a suitable representation to mediate the information for both manual signals and non-manual signals (NMSs). Nonetheless, we have yet to see results that are fully scalable to the level in the real world with respect to lexicon construction, due to the complicated multi-tier information units of the natural sign language.

Building up the lexicon containing these units is labor-intensive because of their diversity; the tiers of information units include manual signs and NMSs, such as movements of lips, eye brows, and tongue for facial expression [8,9]. In particular, for an appropriate facial expression, the transcription becomes not only labor-intensive, but also difficult. As a small difference in the transcription may change the meaning of facial expression quite radically, it would be safe to ask a sign language expert to examine and endorse the result, especially if the domain does not tolerate inaccurate translation. Moreover, it is quite another matter to synchronize hands and facial expressions during the synthesis stage, and to maintain the resulting precision to an adequate degree.

While recording and representing natural facial expressions for animation is difficult, achieving naturalness is no less important for sign language communication. In particular, experiments with an eye-tracker have shown that the deaf person concentrates primarily on the face while he/she is engaged in a conversation with another deaf person or when he/she watches a sign language

SLTAT 2011, 23 October 2011, Dundee, UK

video [1,10,11]. If deaf people can communicate with a hearing signer who is otherwise expressive but suffers from a low level of sign language communication skill, we can assume that facial expressions with full emotional information may be more meaningful than those produced by an abstract grammar of sign language, at least for the purpose of comprehension by the deaf people.

In this paper, we propose to re-visit the idea of sign language translation through concatenative video synthesis. Concatenative video synthesis has been criticized mostly for the poor quality of transition between video fragments as well as for the contextually mismatched signals due to the static, unchangeable video information. We discuss these issues in further detail in this paper, and argue that concatenative video synthesis may work to complement the more rigorous use of an avatar technology for sign language expressions, at least until the technology matures.

In Section 2, we look into previous researches including sign language translation with a video lexicon and its limitations. In Section 3, we describe our proof-of-concept system, and address the well-known limitations of concatenative video synthesis. A user study with the deaf people focusing on comprehension is discussed in Sections 4 and 5. Section 6 concludes our paper.

2. Related Work

If we focus on the simplicity of sign language generation, the simplest method would be to use pre-recorded video. For instance, a pre-recorded video is usually presented to signers in news programs such as weather forecast. It is interesting to note that a real human signer is employed in these programs. This kind of interface has been criticized for its mostly static and contextually inappropriate motion across different sentences [4]. Another system addressed these limitations by putting together videos for each sign language word and for each sentence performed by the same human signer [12,13].

If we look for simplicity, the idea of putting together sign language clips deserves further attention. Capturing video is much simpler than representing each word by some notations or motion capture data. Furthermore, it would be much less dependent on the full theory of sign language phonetics.

The idea would also be reasonable from the perspective of ensuring acceptability. For instance, even though the video in [12] was found neither so fluid nor seamless by 20 deaf people, it was also found that the more important issue for the deaf people is naturalness of signs particularly on the facial expression [14] than simple seamlessness and fluid sign movements. It is certainly a non-trivial trade-off, especially when we are forced to make a choice between naturalness and seamlessness.

One of the problems in these systems would be, however, that the generated video represents only a signed version of spoken language, not the full natural sign language. The video is composed of the unit of word so that it cannot show the full use of a spatial dimension with manual and non-manual signals, the important characteristics of sign language.

Another problem is the limited scalability of the system. It has been assumed that a video of each word should be performed by a single and the same human signer [14], but it would be almost impossible to compose sign language video this way for the whole words by a single signer against the same background [15]. However, to the best our knowledge, the assumption has not been empirically proven yet.

3. Proof-of-concept System

3.1 Overview

We propose to improve on the old idea of concatenative video synthesis focusing on achieving scalability. Unlike other related systems, the video clips in the lexicon are not captured from the gesture of only one human signer. The sign performance by a number of human signers against a different background is usually captured and stored into the lexicon. The proposed technique allows this lexicon to scale up naturally and easily.

To increase the scalability of the system even further, the video clips of the lexicon are now composed of video clips not only of words in a multimedia dictionary of sign language, but also of the segmented words from a number of sign language video fragments performed by other sign language translators, which are recorded against various backgrounds. The proposed technique can thus deal easily with an increased number of sources to achieve relevant data for the lexicon so that it is scaled up easily.

However, the proposed technique does not respect the previously held assumption that the sign gestures of the lexicon should be performed by a single human signer against the same background. As the assumption has not been proven valid yet, we conducted a preliminary user study with 14 deaf people (4 young adults, 10 adults) to see whether or not the appearance of different signers affects comprehension. We believe that the findings are suggestive enough at this stage, despite the small number of participants. This will be discussed further in Section 5.

3.2 Lexicon construction

In this preliminary study, the lexicon of video data is constructed from two sources: a multimedia dictionary of Korean Sign Language (KSL) with subtitles [16,17], and videos where a sign language translator performs a sign language expression [18]. The video clips of the lexicon composed of the unit of words in Korean, and the videos of a sign language expression are segmented in the unit to construct the lexicon.

The video clips of words are annotated with the meaning (glossing), the gender of signer, the speed of gesture, and the background. Other features such as NMS and emotion are not annotated, but we believe that this does not present a serious problem for the present proposal, as we are here focusing primarily on the acceptability of the proposed idea: video composed by different signers on different backgrounds.

3.3 Word-based synthesis

With these annotated clips, new videos can be composed by piecing together smaller clips. In this stage, we need to address three problems: flickering between word changes, spatial dependency, and difficulty on representing figurative expression.

Videos of various sign language sentences are synthesized from the annotated video clips. The unit of synthesis is a video clip in the lexicon, which is also the unit of words in Korean. We have not used any video engineering technique during this process, and this gives rise to a video with flickering when a word changes.

As the unit of synthesis is a word, some parts of the generated video do not show fully natural sign language. In particular, the use of a spatial dimension in sign language is not fully expressed through our synthesis, because videos of words with fixed (rigid) spatial information are concatenated.

This spatial dependency raises another problem of representing figurative expression. In a natural sign language, figurative

expressions are quite common, helping to tell a long story visually in a short time period. However, because our synthesis does not make a full use of the spatial dimension, it is rather difficult to represent figurative expressions.

These problems can be a part of important factors affecting the comprehension by the deaf people. In the next section, we will discuss in detail how these problems show up in sample videos, later evaluated by the deaf people.

4. User study

4.1 Experiment design

4.1.1 Sample videos

We conducted a survey with the deaf people on the reactions by the potential users of our proof-of-concept videos. In particular, we focused on the question of whether or not the video synthesized in a way as described in Section 3.3 can be understood well by the deaf people, though these videos suffer potential problems such as flickering and spatial dependency. We constructed 5 sample videos for each representative expression which we believe poses potential problems to the comprehension.

Table 1. The sentence of 1st sample video

In Korean	어린이는 나라의 보배. elininun nalauy popay (in Yale notation)		
In Korean words	어린이 elininun	나라 nala	보배 popay
In English words	children	country	treasure
In English	Children are treasures of the country.		

Figure 1. Snapshots: Children | country | treasure

The first video shows the problem of hosting different signers for a single expression. Each word video shows a different signer, a different position, and a different background. Except for these differences, the video shows a simple sign language expression with a limited use of spatial dimension where each word is sequentially presented.

Table 2. The sentence of 2nd sample video

In Korean	내 친구는 국가 대표가 되길 바란다. nay chinkwunun kwukka tayphyoka toykil palanta				
In Korean words	나 na	친구 chinkwu	국가 kwukka	대표 tayphyo	바라다 palata
In English words	I	friend	country	representative	hope
In English	My friend hopes to become a representative of the country.				

Figure 2. Snapshots: I | friend | country | representative | hope

The second video shows the problem of employing different signers, together with signer-dependent changes of a facial expression. The word "hope" is naturally expressed together with the emotion expression for "desire" on the face of the signer. The strength of the emotion expression for "desire" alters the sentential meaning with respect to "how much he/she hopes to become someone." However, as there is no standardized way to represent the strength, the overall meaning of the expression may not be preserved well over multiple video clips that express collectively a changing strength.

Table 3. The sentence of 3rd sample video

In Korean	여행할 때 기차가 제일 편해요. yehaynghal ttay kichaka ceyil phyenhayyo				
In Korean words	여행 yehayng	때 ttay	기차 kicha	제일 ceyil	편하다 phyenhata
In English words	Travel	time	train	most	comfortable
In English	For travel, a train is the most comfortable means.				

Figure 3. Snapshots: Travel | at the time | train | most | comfortable

The third video shows the same problem with different signers, along with different signer positions regarding tilting. For each word in a sentence, a different signer is shown, tilting the body to the left or to the right, but not always to the front. These different tilts might confuse the deaf people to grasp the intended meaning of the expression, especially since tilting usually happens when a signer wishes to express a role shift in a natural sign language.

Table 4. The sentence of 4th sample video

In Korean	고양이가 나무 위로 올라갔다 koyangika namwu wilo ollakassta			
In Korean words	고양이 koangi	나무 namwu	위 wi	올라갔다 ollakassta
In English	Cat	tree	up	climbed

words				
In English	The cat climbed up the tree.			

Figure 4. Snapshots: Cat | tree | up | climb

The fourth video shows the problem of a figurative expression, which is often used in a natural sign language. The signer signals a tree in front of him/her, and then figuratively shows a cat climbing up the tree.

Unlike this natural expression, the fourth video represents it in a nonfigurative way. Signer A establishes a tree in front, and then another signer B performs another sign gesture of "climbing up", which is a common gesture that describes a person climbing up a hill.

Table 5. The sentence of 5th sample video

In Korean	서울에서 부산까지 비행기로 몇 시간 걸려요? sewuleyse pwusankkaci pihayngkilo myech sikan kellyeyo?						
In Korean words	서울 Sewul	에서 eyse	까지 kkaci	부산 pwusan	비행기 -날다 pihayngki -nalta	시간 sikan	걸려요? kellyeyo?
In English words	Seoul	from	to	Pusan	airplane-fly	time	take?
In English	How long does it take from Seoul to Pusan by airplane?						

Figure 5. Snapshots: Seoul | from A | to B | Pusan | airplane-fly | time | take?

The final video shows the problem of spatial dependency. In a natural sign language, the signer would place the location gestures "Seoul" on the left and "Pusan" on the right, both in front of the signer. Then, he/she would play the sign gesture in which an airplane flies from the left to the right in order to represent an airplane flying from "Seoul" to "Pusan".

However, this spatial dependency is hard to establish with our synthesis method, since the sign gesture for "airplane flying" cannot be modified dynamically as the spatial allocation of the departure and destination changes from video clip to video clip. In this video, however, the sign gesture for "airplane flying" is represented as if it has no relation with the spatial allocations of departure and destination.

4.1.2 Experiment structure

With the videos in Section 4.1.1, we asked potential users of our system about the level of comprehension, inconvenience, and the opinions on the problems in the video as discussed in Section 3, in particular as for the appearance of a different signer in a single video, flickering changes between words, ungrammatical uses of the spatial dimension, and the nonfigurative expression.

Table 6. The number of participants in each age group

ages	~18	~29	~39	~49	~59	~69
# of participants	4	1	2	4	2	1

For this study, 14 deaf people participated in our survey. The survey was taken twice in different places. The first survey was conducted on 4 pre-lingually deaf students (3 males, 1 female) in high school at the ages of 17~18. Their level of written language skill is the primary school level, and they learned the oralism. The present user study is conducted in the classroom of the school for the deaf. In this first survey, each student watched the video on their own 19-inch monitor individually.

Before conducting the survey, 4 sample videos not in the survey are shown to familiarize the participants with the synthesized video. The time taken to show these videos was about 5 minutes.

After watching the videos, the participants received an online questionnaire with 5 sample videos of Section 4.1.1. The questions are all in a written language, and they are also explained in a sign language by a sign language translator.

A group interview on the shown videos followed the survey. The participants were allowed to freely express their feelings, suggestions, and encountered problems in understanding the video. In addition, we showed the synthesized animations generated by previous researches [19] using SignSmith Studio [20] and gathered feedback on the preferences to video versus animation. As we focused on the preferences only, the group interview proceeded informally without a quantitative analysis. The expressions of these animations were not the target sentences in Section 4.1.1. The sentences of these animations contained one sentence that does not show any spatial dependency and three sentences that show spatial dependency.

The second survey was conducted on 10 pre-lingually deaf adults of the local deaf community at the ages of 25~64, whose level of written language skill is the primary school level at minimum. The user study was conducted at a local branch of the Center for Deaf Association in Korea. In this second survey, all the participants watched the videos on a 60-inch screen from a 5 meter distance due to the lack of personal computers. Because of the same reason, the printed version of the online questionnaire substituted the previous one. Except for these changes, the survey was conducted in a similar manner as the first survey.

A group interview followed the second survey, in a manner similar to that in the first survey. At this time, other sign animations generated by worldwide projects [7, 23] were shown in addition to our animations to elicit broader feedback and to form a realistic perspective on the state-of-the-art of the avatar technology. We clearly explained that the sign languages in the animations are not KSL, and asked the participants to imagine the KSL animation where the motions are more fluid and natural as in other animations of the worldwide projects. We waited until they decided and gathered their preferences.

4.1.3 Questionnaire

The questionnaire has 3 questions for each of the 5 videos. The first question showed a synthesized video and asked each participant how much he/she can understand the sign language expression as shown in the video. The answer is in the format of Likert scale of 5 from "very easy to understand" (1) to "impossible to understand" (5). This question is designed to record the comprehension level.

The second question showed both synthesized video and the original video of natural sign language performed by a sign translator for the same expression. The participants acknowledged that both videos convey the same expression, and are asked to answer the question about which aspects caused him/her to have difficulty for understanding the synthesized video as opposed to the original video. The answer consists of flickering, the characteristics of the signer such as gender, position, and inappropriate sign gesture. This question is designed to identify those aspects that affect comprehension.

The final question asked the participants how much he/she felt discomfort while watching the synthesized video. The answer is in the format of Likert scale of 5 from "absolutely no discomfort to watch" (1) to "great discomfort to watch" (5). This question is designed to show whether or not the video would make the deaf person feel tired when reading/watching longer synthesized video.

5. Evaluation

5.1 Results of survey

The survey focused on the comprehension and inconvenience, and its result shows that our proof-of-concept is promising for both young adults group and adults group.

5.1.1 Survey with the young adults group

On the first survey with 4 deaf students, no negative answer was collected in the comprehension and inconvenience questions about 5 videos.

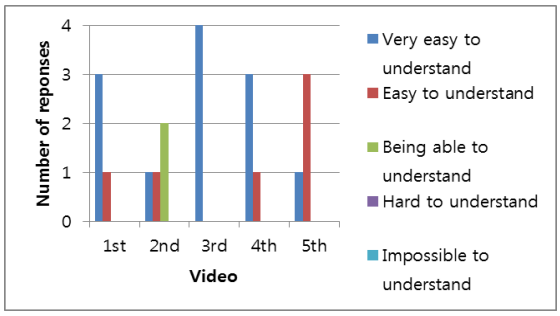

Figure 6. Responses to the comprehension question

The most negative response in this first survey, "being able to understand", was given only to the second video, and it was found that the problem was due to a different dialect of those 2 respondents. The expected problem, change of facial expression by a different signer, was not pointed out at all by any of the participants.

Except for the second video, the distribution of answers of the fifth video is remarkable. The distribution shows that the comprehension level decreased a little bit as compared to other distributions. Considering that the fifth video has a spatial dependency problem, this result might imply that an expression with spatial dependency should be carefully synthesized, though the participants didn't say anything about this issue. This problem was seriously shown in the second survey.

For the second question, asking aspects that caused difficulty on understanding, the responses included substitution of gesture for dialect usage. Only one participant responded that flickering is bothersome but acceptable. Other aspects such as different signer, tilting and position of signer, figurative expression, and spatial dependency are not pointed out.

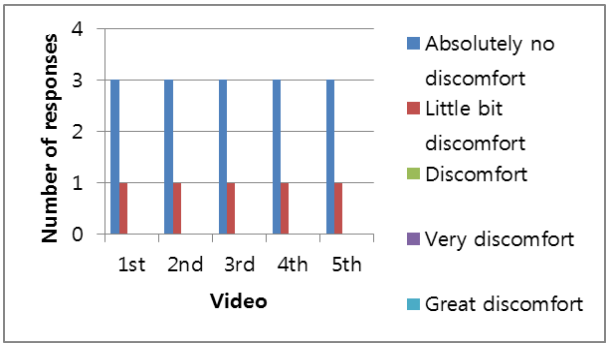

Figure 7. Responses to the inconvenience question

For the third question, or the inconvenience test, on average, three people answered "absolutely no discomfort to watch", and one person answered "a little discomfort to watch". The participant who answered "a little discomfort to watch" is the one who responded that the flickering is bothersome, but acceptable.

5.1.2 Survey with the adults group

On the second survey with 10 adults, the responses were quite different from the previous one.

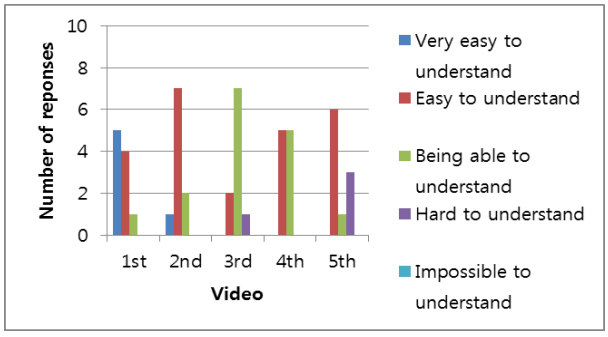

Figure 8. Responses to the comprehension question

On the average, the comprehension level is lower than that of the young adults group. From 3rd video to 5th video, the responses showed clearly more negative rates than those in the first survey.

The 3rd video was rated much lower than the first survey because of a technical problem, or that of video resolution. The last word clip of the video had a coarser resolution than other clips, and as

the video was shown on the 60-inch screen, the detailed movement of the word was lost, which contributed to making the participants confused about the meaning.

The 4th video was rated lower than the first survey because of the figurative expression problem. All the participants responded that the video is understandable, but not exactly the same as the original versions as these videos show different nuances.

The 5th video was rated low as expected in the first survey, but it is still significant that it was responded as "hard to understand" by 3 participants. Although other participants were able to understand the video, this result might suggest the limitations of concatenative synthesis. We will discuss this issue further in Section 5.3.

For the second question, that of asking those aspects that caused difficulty in understanding, the adults group responded that flickering was the most serious, followed by a different signer especially on gender and different position of the signer. Those aspects were discussed in the group interview session.

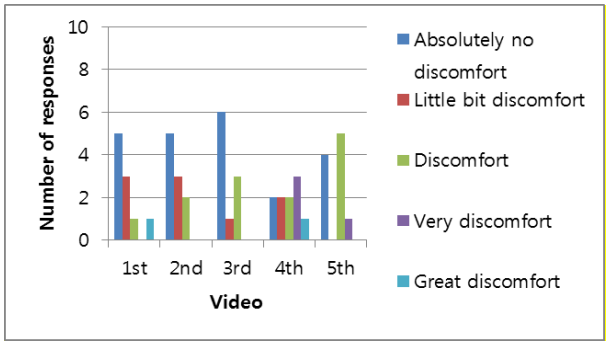

Figure 9. Responses to the inconvenience question

For the third question, the inconvenience test, the participants responded with a similar tendency as they responded to the comprehension test. Interestingly, they mostly felt discomfort on the 4th and 5th videos, but not on the 3rd video, even though these videos showed clear movements with no resolution problem. From this result, we suggest that an appropriate use of signing space gives comfort to the deaf people.

5.2 Results of the group interview

All videos of 5 expressions have the problem of flickering between words, and we received feedback on this problem first in the following group interview. The responses of the young adults are slightly different from those of the adults group as discussed earlier. At the first survey with 4 young adults, 3 deaf students responded that the flickering didn't make them bothered to watch the sign language expression in video, and one deaf student responded that it was bothersome but acceptable. Moreover, they didn't mind at all the change of background and signers including all suggested aspects: gender, age, and clothes. One student commented that the flickering video is more "cool" to watch than the original video. The only problem they discussed is the sign gesture which has various versions for each dialect usage.

As for the preference to video versus animation, all the deaf students preferred video to animation. They responded that the sign movement of the animation was "robotic" and "awkward". Unlike the animation, the video was regarded as "vivid".

At the second survey with 10 adults, however, all the participants responded that the flickering between words was bothersome, though acceptable. Some of them raised a question whether the flickering can be made smoother or slower. They pointed out that flickering occurred so unexpectedly that they cannot sometimes follow the next word. Moreover, they suggested showing translators of the same gender, not to be confused by different signers. In addition, one participant advised to fix the signer's position on the screen. The participant said that the small changes in the position caused much confusion to identifying the gesture, especially between the gestures that have the same manual signals with a different signing space.

As for the preference to video versus animation, we showed two different animation groups as described in Section 4.1.2. For the first comparison, we showed our animations, and all the adults preferred video to our animation without hesitation. In fact, they had more difficulty in understanding the animation than the video. This might be because our videos suffered from some "robotic" motions due to a technical problem of SignSmith Studio. However, considering the fact that the animations are definitely more seamless than the videos and have synchronized NMSs such as eye gaze, mouthing and head tilting, with a full use of spatial dependency, the preference of the adults group is quite surprising.

For the second comparison, we showed the animations of worldwide projects, and asked the participants to imagine more fluid and natural KSL animations as in these animations. The participants discussed for approximately 5 minutes, and responded that they still prefer the video to animation. However, they couldn't elaborate on the reason. We suspect that this informal comparison might have been biased from the first comparison as the participants watched and rated our animation first. However, we also believe that the result indicates that the video contains some critical information that those animations don't. We will conduct a more formal user study to identify the factors that affected the preference.

5.3 Discussion

The results of the first survey group and second survey group are quite different from each other, especially regarding the participants' responses to the flickering and changes of signers. The ages of the participants might be one of the crucial factors that affected the result as the statistics of Table 7 shows.

Table 7. The correlation value between age and the responses

	comprehension level	inconvenience level
correlation value	0.470	0.434
p-value	<0.001	<0.001

However, only for the adults group, the p-value of correlation value between age and comprehension level was 0.114.[1] It is not clear whether the age is the factor, and this result is not conclusive enough due to the small size of this group.

In addition to age, these two groups show many other differences, but we have insufficient background information to pin-point the factors that gave rise to this result. However, we believe that the

[1] The correlation value between age and comprehension level was 0.226, and the value between age and inconvenience level for the adults group was 0.276 with p-value 0.052.

duration of the deaf person with exposure to a natural sign language is one of the important factors.

The young adults who participated in the first survey usually spend their daily lives in the school with hearing teachers who use a signed version of Korean, not KSL, the natural sign language. However, the adults participated in the second survey spend their daily lives with other deaf people in the local deaf community, which uses KSL as the primary language, but not a signed version of Korean.

As our synthesized video is much similar to the signed version of Korean, it is reasonable to assume that, given the aforementioned difference of the respective communities, the young adults understood the videos well while the adults had some difficulties. This seems to be prominent from the results of the 4^{th} and 5^{th} videos which use truly a signed version of the Korean language. The young adults understood these videos well, but the adults showed a much lower comprehension level.

Likewise, current video exhibits no distinguishing characteristics of a natural sign language such as figurative expressions and the full use of a spatial dimension. In fact, what we have conducted in this research at this point can be categorized simply as transliteration, an interpretation from a spoken language to the signed version of that language. One may thus claim that there is no real challenge for the concatenative video synthesis of this kind. However, transliteration is widely accepted and used especially for education. In particular, Napier and colleagues [21] showed that transliteration works as good as translation to natural sign language (free interpretation) in assuring the comprehension of the deaf people during education. Some of our findings, such that nonfigurative expressions and inappropriate uses of a spatial dimension can be accepted and even understood by some deaf people, in particular by young adults, show that transliteration can be applied at least to the people in the community where active education takes place. It may thus be true that research on transliteration complements the work on translation quite well in providing accessibility for the deaf people, especially when it is augmented with techniques for a more natural sign language expression.

Nonetheless, we admit that the full use of our concept of mixing various videos might pose a serious problem to spatial dependency as the user study has clearly shown. Until now, this problem has not been addressed adequately by pure video techniques. In this regard, we are working towards bringing the avatar technology into our proposal. In particular, we are working on a hybrid approach where animation and video are combined, as animation serves well to address spatial dependency and video serves well to address other problems where emotion and naturalness are particularly important. While this idea should certainly be rigorously examined by the deaf people, we believe it is quite promising as the deaf people are found to be surprisingly tolerant of the flickering and changes of signers. In the future, we will conduct a user study on another proof-of-concept system based on this idea. It is anticipated that the resulting system can achieve both the scalability of the video clips and the flexibility of the animation [22].

6. Conclusion

In this paper, we proposed to use the concatenative video synthesis technique in a more scalable manner. The proposed idea for scalability is to use video clips by a different signer for the lexicon, and to synthesize a video by putting together these individual clips.

We should note again that, while the proposed idea may not fully account for the characteristics of a natural sign language such as figurative expressions and the extensive use of a spatial dimension, it still preserves natural facial expressions in the video as much as possible. A simple user study on comprehension showed that this idea works on various types of short expressions. Much further work is needed, such as a more rigorous user study with a larger user group, as well as a systematic approach to overcoming the problems as discussed in the paper.

7. ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2011-0004422). We would also like to express our special thanks to the participants of our user study: the students at Daejeon Won-myeong School, and the people of Yuseong Deaf Association in Korea. The comments and feedback of the anonymous reviewers are also much appreciated, though not reflected fully in the present version.

8. REFERENCES

[1] Lu, S., Igi, S., Matsuo, H., Nagashima, Y.: Towards a dialogue system based on recognition and synthesis of Japanese Sign Language. In *Wachsmuth, I., Fr¨ohlich, M., editors, International Gesture Workshop, GW'97. Lecture Notes in Artificial Intelligence*. Springer-Verlag, Berlin Heidelberg New York (1997) 259–271.

[2] Elliott, R., Glauert, J. R. W., Kennaway, J. R., and Marshall, I. 2000. The development of language processing support for the ViSiCAST project. In *ASSETS 2000 - Proc.4th International ACM Conference on Assistive Technologies*, November 2000, Arlington, Virginia, pages 101–108.

[3] Grieve-Smith, Angus B. 2001. Signsynth: A sign language synthesis application using Web3D and Perl. In *Ipke Wachsmuth and Timo Sowa, editors, Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop GW2001* (LNAI vol.2298, pages 134–145, Springer).

[4] Zwitserlood I, Verlinden M, Ros J, van der Schoot S. 2004. Synthetic signing for the deaf: Esign. In *Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment* (Granada, Spain, July 2004).

[5] Morrissey, S. 2008. Assistive translation technology for deaf people: translating into and animating Irish sign language. In *12th International Conference on Computers Helping People with Special Needs* (9-11 July 2008, Linz, Austria).

[6] Huenerfauth, M. 2006. Generating American Sign Language classifier predicates for English-to-ASL machine translation. Dissertation, Department of Computer and Information Science, University of Pennsylvania.

[7] Wolfe, R., Cook, P., McDonald, J. C., and Schnepp, J. 2011. Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in American Sign Language. In *Nonmanuals in Sign Language Special issue of Sign Language & Linguistics* 14:1 pages 179-199.

[8] Yoon, B., and Kim, B. 2004. The Study on the Linguistic Characteristics of Non-manual Signals in Korean Sign Language. In *Journal of Special Education: Theory and Practice*, volume 5(1), pages 253-277.

[9] Hanke, T. 2004. HamNoSys—representing sign language data in language resources and language processing contexts. In *Streiter,O., Vettori, C., editors, Fourth International Conference on Language Resources and Evaluation (LREC 2004). Representation and Processing of Sign Languages Workshop, pp. 1–6. European Language Resources Association,* (Paris, 2004)

[10] Muir, L. and Richardson, I. 2005. Perception of sign language and its application to visual communications for deaf people. In *Journal of Deaf Studies and Deaf Education*, volume 10, pages 390–401.

[11] Emmorey, K., Thompson, R., and Colvin, R. 2009. Eye gaze during comprehension of American Sign Language by native and beginning signers. In *Journal of Deaf Studies and Deaf Education*, volume 14(2), pages 237–243.

[12] Inclusive Learning Scotland (2007), Science Signs, http://www.ltscotland.org.uk/inclusiveeducation/findresources/bslsciencesigns/BSLinteractive/bslscience.asp

[13] Solina, F., Krapez, S., Jaklic, A., and Komac, V. 2001. Multimedia Dictionary and Synthesis of Sign Language. In *Rahman, S. M., editors, Design and Management of Multimedia Information Systems.* Idea Group Publishing pages 268-281.

[14] Naqvi, S. 2007. End-User Involvement in Assistive Technology Design for the Deaf - Are Artificial Forms of Sign Language Meeting the Needs of the Target Audience? In *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision & Hearing Impairments* (Granada, Spain).

[15] Huenerfauth, M., and Hanson, V.L. 2009. Sign language in the interface: Access for deaf signers. In *Stephanidis, C., editor, The Universal Access Handbook*. CRC Press.

[16] The Dictionary of Korean Sign Language, The National Institute of the Korean language, http://222.122.196.111/

[17] Korean Sign Language Dictionary, Korea National College of Rehabilitation and Welfare, http://support.hanrw.ac.kr/signdic/

[18] Forum of Learning Korean Sign Language, http://cafe.naver.com/ksign

[19] Chung, J., Lee, H., and Park, J. C. 2010. Sentence Type Identification in Korean: Applications to Korean-Sign Language Translation and Korean Speech Synthesis. In *Journal of the HCI Society of Korea*, volume 5(1), pages 25-35.

[20] Vcom3D. 2008. Company website. http://www.vcom3d.com

[21] Napier, J., and Barker, R. 2004. Accessing University Education: Perceptions, Preferences, and Expectations for Interpreting by Deaf Students. In *Journal of Deaf Studies and Deaf Education*, volume 9(2), pages 228-238.

[22] Chung, J., and Park, J. C. 2011. Text Parsing for Sign Language Generation with Combinatory Categorial Grammar. In *Second International Workshop on Sign Language Translation and Avatar Technology* (Accepted).

[23] Lombardo, V., Nunnari, F., and Damiano, R. 2011. A Virtual Interpreter for the Italian Sign Language. In *First International Workshop on Sign Language Translation and Avatar Technology* (Berlin, Germany)