Elektrotehniški vestnik 66(4–5): 260–265, 1999
Electrotechnical Review, Ljubljana, Slovenija

Synthesis of the Sign Language of the Deaf from the Sign Video Clips

Slavko Krapež, Franc Solina

University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, SI-1000 Ljubljana, Slovenia
slavkok@hermes.si, franc.solina@fri.uni-lj.si

Abstract. The aim of the article is to present a system for automatic synthesis and visualization of the Slovene sign language of the deaf (SSL) sentences. A synthesis is made by assembling individual previously filmed video clips of sign demonstrations (1800 most frequent signs). Video clips consistency and optimization of passages between individual video clips enable smooth joining of video clips on a personal computer in real time offering a high quality video. Existing systems for the sign language synthesis are based on an artificial sign language demonstrator and are merely prototypes. Our system uses a live person for signing and has a large set of sign video clips. Therefore it enables a user to perform a translation of different texts into the sign language.

Key words: multimedia, digital video, sign language of the deaf, computer vision

Sinteza znakovnega jezika gluhih z združevanjem videoposnetkov kretenj

Povzetek. Članek predstavlja sistem za avtomatsko sintezo in vizualizacijo povedi slovenskega znakovnega jezika gluhih. Sinteza poteka z združevanjem posameznih predhodno narejenih videoposnetkov kretenj (1800 najpogostejših kretenj). Konsistentnost videoposnetkov in optimizacija prehodov med njimi omogočata prikaz zveznih videoposnetkov povedi znakovnega jezika v realnem času na povprečnem osebnem računalniku ob hkratni visoki kakovosti videa. Obstoječi sistemi za sintezo znakovnega jezika temeljijo na animaciji sintetične lutke in so večinoma le prototipi. Naš sistem uporablja živega kretalca in vsebuje obsežno množico videoposnetkov kretenj. Slednje omogoča prevod najrazličnejših besedil v znakovni jezik.

Ključne besede: multimediji, digitalni video, znakovni jezik gluhih, računalniški vid

1 Introduction

Deaf people, as a marginal community, have big troubles in communicating with hearing people. Usually they have a lot of problems even with such simple tasks as understanding the written language. However, they are very skilled in using the sign language, which is their native language. A sign language is a set of signs (the Slovene sign language (SSL) contains about 4000 different signs). The sign in a sign language equals to a word in a written language. Similarly, a sentence in a written language equals to the sign sentence in a sign language. Our long-term goal is to build a system able to translate the written language (books, newspapers, e-mails, letters, HTML documents, ...) and in connection with speech recognition also speech (conversations, radio and TV programs, phone calls, ...) to the sign language. So far we have achieved this goal just partially since the translator from a text to a sequence of sign names (Figure 1) has not been implemented yet. It would be quite straightforward to implement such translator in case of the English language, because English and all sing languages are grammatically very similar. In most cases, it consists of only the word to sign translation. For languages with more difficult grammar one possible solution would be to build such a translator using a translator from that specific language to the English language and then to the sign language.

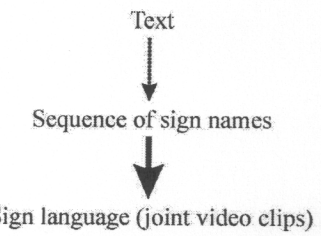


Figure 1. Translation from a text to the sign language

This article introduces a new method for joining sign

Received 15 August 1999
Accepted 29 October 1999

video clips in video clips of complete sign sentences. The joining can be processed on a personal computer in real time. We achieved smooth transitions between sign video clips using information about demonstrator's palm positions while signing. For that purpose a special program for palm position extraction from video clips was written based on methods from computer vision.

The described system contains a large set of individual sign video clips (1800), which makes possible the translation of various texts to the sign language. Words which are not in the system can be shown using video clips containing the finger alphabet. In principle any text can be shown in the sign language.

The prevalent idea in the development of a system for the sign language synthesis is to use a synthetic person [5,6]. Such systems are usually taught using datagloves which contain sensors for defining the palm and finger position and orientation.

Since realistic animation of persons is still a time consuming task, we decided to use video clips of a real person as a sign language demonstrator. A synthetic demonstrator is no match for a real one in sense of appearance and consequently also of acceptance as a teaching tool.

2 Slovene sign language dictionary for the deaf

A concept of the *Slovene* multimedia *sign language dictionary* for the deaf on CD-ROM was first presented in 1995 [1,2]. A pilot application was made in 1996 [7] and the final application in 1999 [8]. The final version (Figure 2) of the dictionary is composed of 2504 most frequent words that are used by the deaf people in every day conversations.

As the sign language can be best learned in topic groups, a simple but effective database navigation tool is provided. The user can select a topic area and then enter a word into the search entry field. As the word is entered, a list of suggested words is displayed on the screen to give a better overview over the database content and to help the deaf people with spelling. A word can be selected at any time directly from the list of suggested words. After the word is selected, the user is presented with a video clip of corresponding sign and if available also with a picture illustrating the concept. The video can be played in slow motion for easier visual tracking of the movements or even examined frame by frame by moving the slider along the time axis. The sound is played only at the normal speed playback. The sound volume can be readjusted by the user.

Sign video clips from the dictionary are consistent (since all clips are filmed in one session using just one sign demonstrator, the same start and end positions of the arms on all video clips are used, see Figure 3). Sign video clips from the CD-ROM can be assembled into sign language sentences using a special algorithm which enables

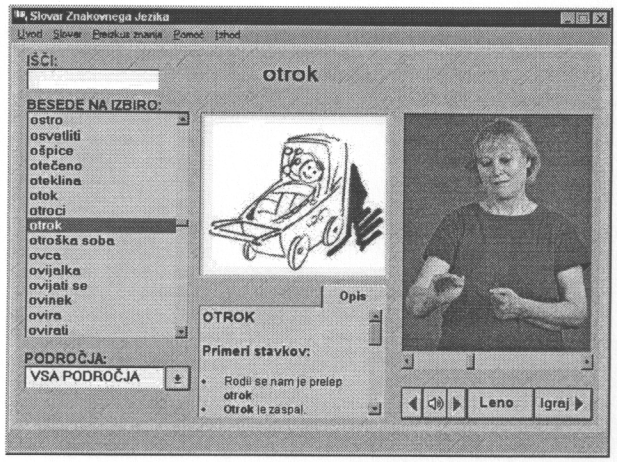


Figure 2. User interface of the *Slovene sign language dictionary* on CD-ROM

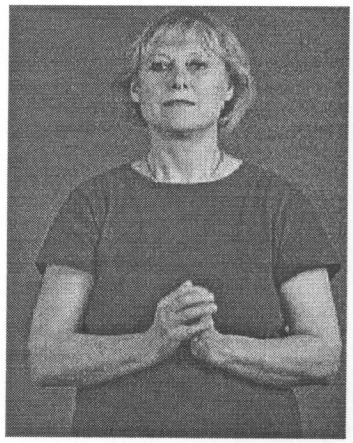


Figure 3. Example of one frame of a sign video clip. The demonstrator is in the start position

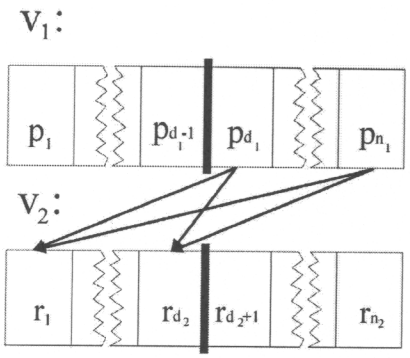
smooth transitions between individual sign video clips. The dictionary therefore represents a basis for the sign language synthesizer.

3 Joining of Video clips

Joining two digital video clips (each contained in its own file) should be done in such a way that the viewer perceives them as a single video sequence with smooth motion.

3.1 Definition of the video clip joining problem

The problem of video clip joining will be discussed in the case of two video clips. This can be easily generalized to any number of video clips. Let us represent the first video clip that we would like to show as a n_1 dimensional vector v_1



$$argmin(\text{dif}(t_1, t_2))$$

$$t_1 \in [d_1 \ldots n_1]$$

$$t_2 \in [1 \ldots d_2]$$

Figure 4. Determination of the t_1 and t_2 values

$$v_1 = \langle p_1, p_2, \ldots, p_{n_1} \rangle$$

where the components of the vector correspond to the pictures that form the video clip. n_1 is the number of pictures in the video clip. The second video clip is represented as:

$$v_2 = \langle r_1, r_2, \ldots, r_{n_2} \rangle$$

The joining can occur on any pair of pictures from v_1 and v_2. Let us indicate the point in which joining should occur with t_1 and t_2. They represent the indexes of pictures, which are, in the case of the first video clip, shown as the last and as the first in the case of the second video clip.

$$t_1 \in [1 \ldots n_1]$$

$$t_2 \in [1 \ldots n_2]$$

Let us introduce another two variables d_1 and d_2. Let d_1 be the index of the picture in the first video clip, from which on, the joining is allowed, d_2 is the index of the picture in the second video clip till which the joining is allowed. With this constraints t_1 and t_2 can occupy the following values.

$$t_1 \in [d_1 \ldots n_1]$$

$$t_2 \in [1 \ldots d_2]$$

The joining of video clips should be performed in such a way that the transition from one video clip to another wouldn't be noticed. Let us introduce a criteria function $\text{dif}(i, j)$. This function calculates the difference between pictures p_i and r_j depending on a given criterion. The smaller is the function value, the more both pictures resemble each other. The calculation of t_1 and t_2 is performed by finding the values of arguments for which the function dif has a minimum value (Figure 4).

4 Automatic sign language video clips joining

In the case of sign video clips joining we would like to achieve the following two goals:

1. as smooth as possible transitions between video clips and
2. minimal additional motion of the arms from the start position of the arms into the demonstration of the sign.

Sign video clips start and end with the demonstrator's arms in the start position (Figure 3). In the start position, the sign demonstrator joins the hands in the belt height.

In order to perform optimal joining of sign video clips some extra data is needed. Each video clip has a related file containing data about the position of demonstrator's palms performing a particular sign. Positions of palms are computed by a program *Tracking palm movements in video clips*, written in C++ using standard computer vision algorithms [4,3]. The input to that program are the sign video clips. The output are files containing data about palm positions for every frame forming the video clip.

Figure 5 shows the arm vectors computed from one frame of a sign video clip. The palm position is at the end point of the arm vector.

4.1 Suggested joining criteria

Conditions that enable smooth transition between video clips in the sense of similar palm positions are refered to as joining criteria. We propose four possible criteria to determine the transition point between two sign video clips:

1. palms in start position,
2. palms outside the start position,
3. palms over the chest,
4. palms close to each other.

Figure 5. Arm vectors

4.1.1 Start position criterion

Using this criterion, transition should occur at points where arms are in the start position. Function dif(i, j) should therefore have a minimum value for $i = n_1, j = 1$.

4.1.2 Palms outside the start position criterion

Palms are outside the start position when they are near the start position but not joined or when their distance from the start position exceeds certain value. Function dif(i, j) has its minimum in the case of a picture pair (p_i, r_j), where palms on picture p_i are outside the start position for the last time and for the first time outside the start position on picture r_j.

4.1.3 Palms over the chest criterion

According to this criterion the transition between video clips is performed when palms are over the chest. Function dif(i, j) has its minimum for a picture pair (p_i, r_j) where the palms on picture p_i are for the last time over the chest and palms on picture r_j are for the first time over the chest region.

4.1.4 Palms close to each other criterion

Using this criterion, the transition occurs at points where the palms from the first and the second video clip are close to each other. Function dif(i, j) is defined as: dif$(i, j) = d(i, j) + w(i, j)$ where $d(i, j)$ represents the distance between palms from pictures p_i and r_j. $w(i, j) = ((i - d_1) + (d_2 - j)) * K$, K is an empirical constant value (in our case $K = 5$).

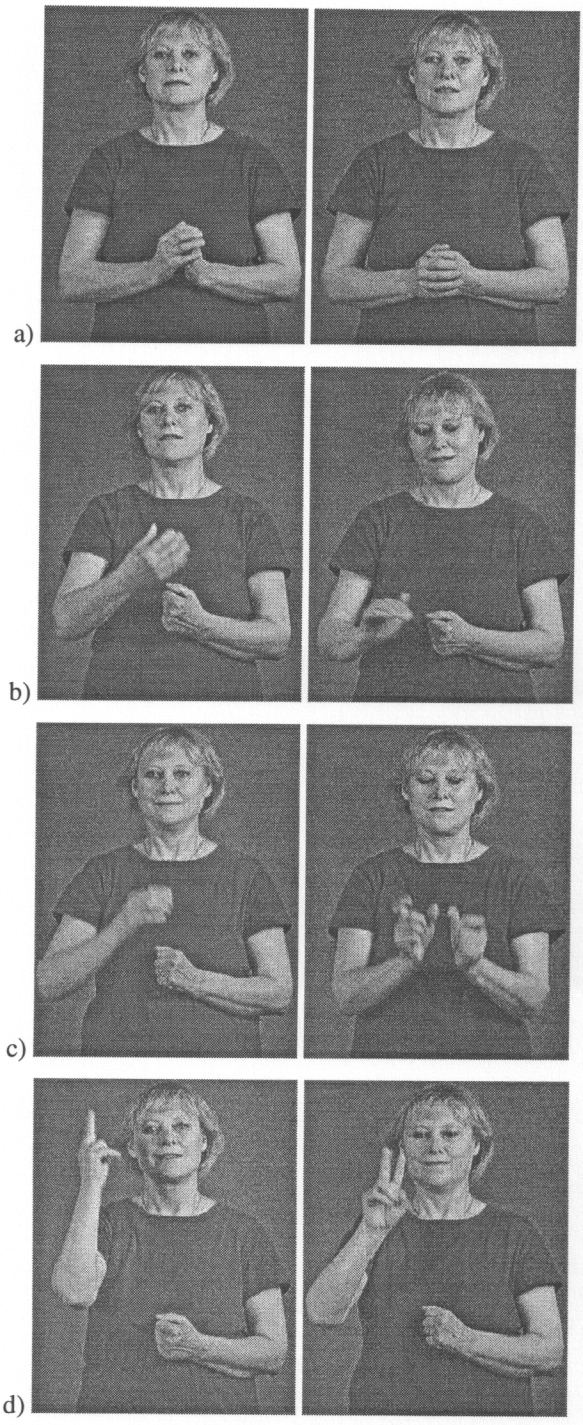


Figure 6. Shown are the last picture of the first and the first picture of the second video clip according to the a) *start position criterion*, b) *palms outside the start position criterion*, c) *palms over the chest criterion* and d) *palms close to each other criterion*

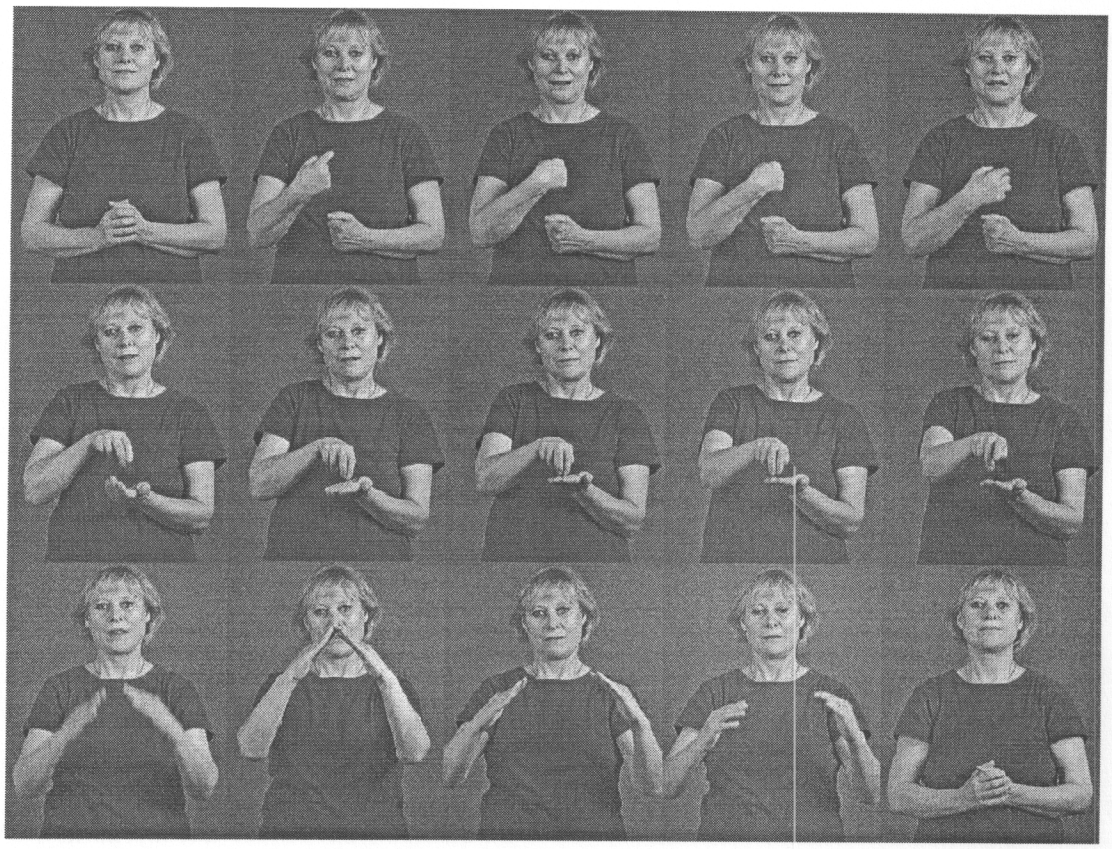


Figure 7. Example of the synthesized sentence "I am walking home"

4.2 Implementation of automatic joining of sign video clips

At this point we have to say something about the use of arms while signing. Most signs are performed in the chest and head height. Signs performed in the belt height or lower are rare but still to be considered. Approximately two thirds of signs are performed with one hand. The hand can either be the right or the left, that doesn't affect the meaning of the sign. In our case they are all performed with the right hand. We will call that signs *one-hand signs*. Approximately one third of signs is performed with both hands. We will call them *two-hand signs*.

According to that, we propose automatic joining of video clips using all four criteria for joining. The use of a specific criterion depends on the type of the two signs we would like to join. The criterion for joining two sign video clips is chosen according to the following four possibilities. If the conditions enable us to use more criteria, we use the first that matches in order it is listed:

- a delay is needed ⇒ *start position criterion*
- one of the signs is performed in the belt height ⇒ *palms outside the start position criterion*
- exactly one of the signs is a two-hand sign ⇒ *palms over the chest criterion*
- both signs are one-hand signs or two-hand signs ⇒ *palms close to each other criterion*

5 Results of automatic joining of sign video clips

Let us have a look at some results of the transitions between two video clips obtained with the method of automatic video clip joining.

Figure 6a shows the last shown picture of the current and the first shown picture of the next sign video clip using the *start position criterion*. That kind of transition is useful when we want to emphasize the end of the sentence or the end of the paragraph. It is useful also in the case where the values of the palm positions are not available. Figure 6b shows the transition between two signs one of which is performed in the belt height. *Palms outside the start position criterion* is used in this case to prevent arms from being joined in the start position. The results obtained on one-hand and two-hand signs using the *palms over the chest criterion* are shown in Figure 6c. Figure 6d shows the suggested transition between one-hand signs using the *palms close to each other criterion*.

The results of automatic sign video clip joining are very encouraging. This system allows us to join different kinds of sign video clips into sign sentences with smooth transitions between signs. The results of video clip joining are very good thanks also to the following two reasons:

- there are no visible delays in video clip playing between the transitions from one video clip to another and
- the human ability of making an impression of continuous transition between two pictures that are similar enough and shown quickly one after another.

The system for the sign synthesis can show video clips with normal speed or with lower or faster speed than normal. In this way it can be adapted to the user's ability of sign language recognition.

Figure 7 shows an example of the synthesized sentence "I am walking home", which is shown as three joined video clips. The video clips have a frame rate of 15 s^{-1} while in the figure just every fifth frame is shown.

6 Conclusion

A system for the Slovene sign language synthesis from high quality, individual sign video clips, running on a PC in real time is presented. The proposed synthesizer would enable deaf people to access different kinds of information up to now inaccessible for them. Important is that the information can be presented in the sign language, which is the most natural way of communication for deaf people. Big advantages of the suggested synthesizer is that it can be run on an ordinary personal computer and that it contains a huge set of sign video clips that enables the translation of almost any text into the sign language.

The deficiency of the system is that it expects as input a sequence of sign names, which are then joined to one or more sign sentences. Therefore, to achieve a full text to the sign language translation, we would need a translator from *text* to *sequence of sign names*. For the English language this is a relatively easy goal to achieve, since English and the sign language have a very similar syntax. However, there would be more problems with languages such as Slovene or German because they have a more complicated syntax (in that case a possible solution of the problem would be the translation of the text first in the English language and then into the sequence of sign names).

For now, no appropriate translator from the Slovene to the English language or from *text* to *sequence of sign names* exists. In these circumstances the use of a sign language synthesizer is limited to tasks such as:

- replacement of sign language demonstrators in TV programs or
- translations of previously known sentences used in education of deaf people through multimedia.

The presented synthesizer is suited for the Slovene language. Of course it would be possible to build similar synthesizers for other languages, too.

7 Acknowledgment

This work was supported by the Ministry of Science and Technology of Republic of Slovenia (Project no. L2-0521) and Zoom Promotion, Ljubljana. We would like to thank also Vito Komac, Ljubica Podboršek and Dora Vodopivec for their enormous help in the developing of the Slovene sign language dictionary on CD-ROM.

8 References

[1] Aleš Jaklič, Dora Vodopivec, Vito Komac, "Learning Sign Language through Multimedia", *Proceedings of International Conference on Multimedia Computing and Systems*, Washington, pp. 282-285, 1995.

[2] Aleš Jaklič, Dora Vodopivec, Vito Komac, M. Gašperič, "Multimedia Learning Tools for the Hearing Impaired", *Proceedings of the World Conference on Educational Multimedia and Hypermedia*, Graz, ED-MEDIA 95, pp. 354-359, 1995.

[3] Reinhard Klette, Piero Zamperoni, *"Handbook of Image Processing Operators"*, John Wiley & Sons, Wiesbaden, 1994.

[4] A. D. Marshal, R. R. Martin, *"Computer vision, Models and inspection"*, World Scientific, University of Wales, 1993.

[5] *"Recognition and Synthesis of American Sign Language"*, http://www.asel.udel.edu/gesture/ASL_synth.html.

[6] *"Sign Language Recognition and Synthesis System"*, http://www.crl.go.jp./st/st821/research/sl-e.htm.

[7] Slavko Krapež, *"Slovar znakovnega jezika za gluhe"*, Diploma Thesis, University of Ljubljana, Faculty for Computer and Information Science, Ljubljana, 1996.

[8] Slavko Krapež, *"Sinteza in razpoznavanje povedi sestavljenih iz kretenj slovenskega znakovnega jezika gluhih"*, Master Thesis, University of Ljubljana, Faculty for Computer and Information Science, Ljubljana, 1999.

Slavko Krapež received his Dipl.Ing (1996) and M.Sc. degrees (1999) in computer and information science from the University of Ljubljana. His research interests are multimedia, especially multimedia support for learning of sign language and Internet based applications. Slavko Krapež is now with Hermes SoftLab in Nova Gorica, Slovenia.

Franc Solina received Dipl.Ing (1979) and M.Sc. degrees (1982) in electrical engineering from the University of Ljubljana and Ph.D. degree (1987) in computer science from the University of Pennsylvania. He became an assistant professor of computer and information science at the University of Ljubljana in 1988, was promoted to associate professor in 1993 and to full professor in 1998. His research interests are segmentation and part-level object representation in computer vision, as well as Internet based applications of computer vision.