Semi-Automatic Sign Language Corpora Annotation using Lexical Representations of Signs

Matilde Gonzalez[1], Michael Filhol[2], Christophe Collet[3]

[1,3]IRIT (UPS - CNRS UMR 5505) Université Paul Sabatier, [2]LIMSI-CNRS
[1,3]118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9 , [2]Campus d'Orsay, bât. 508, F-91403 ORSAY CEDEX - France
[1]gonzalez@irit.fr, [2]michael.filhol@limsi.fr,[3]collet@irit.fr

Abstract

Nowadays many researches focus on the automatic recognition of sign language. High recognition rates are achieved using lot of training data. This data is, generally, collected by manual annotating SL video corpus. However this is time consuming and the results depend on the annotators knowledge. In this work we intend to assist the annotation in terms of glosses which consist on writing down the sign meaning sign for sign thanks to automatic video processing techniques. In this case using learning data is not suitable since at the first step it will be needed to manually annotate the corpus. Also the context dependency of signs and the co-articulation effect in continuous SL make the collection of learning data very difficult. Here we present a novel approach which uses lexical representations of sign to overcome these problems and image processing techniques to match sign performances to sign representations. Signs are described using Zeebede (ZBD) which is a descriptor of signs that considers the high variability of signs. A ZBD database is used to stock signs and can be queried using several characteristics. From a video corpus sequence features are extracted using a robust body part tracking approach and a semi-automatic sign segmentation algorithm. Evaluation has shown the performances and limitation of the proposed approach.

Keywords: Sign Language, Annotation, Zeebede

1. Introduction

Sign Languages (SL) are visual-gestural languages used by deaf communities. They use the (whole) upper-body to produce gestures instead of the vocal apparatus to produce sound, like in oral languages. This difference in the channel carrying the meaning, i.e. visual-gestural and not audio-vocal, leads to two main differences. The first concerns the amount of information that is carried simultaneously, body gestures are slower than vocal sounds but more information can be carried at once. The second is that the visual-gestural channel allows sign languages to make a strong use of iconicity (Cuxac, 2000). Parts of what is signed depends on, and adapts to, its semantics, usually geometrically. This makes it impossible to describe lexical units with preset phonemic values. In addition SL is strongly influenced by the context and the same sign can be performed in different ways.

For this reason automatic SL recognition systems would require huge amounts of data to be trained. Also the recognition results depend on the quality and the representativeness of the data which is, in general, manually annotated. Manual annotation is time-consuming, error prone and unreproducible. Moreover, the quality of the annotation depends on the experience and the knowledge of the annotator.

There already exist some works attempting automatic annotation. In (Dreuw and Ney, 2008) is proposed to import the results of a statistical machine translation to generate annotations. However this approach do not address the problem of collecting data since the statistical machine translation might use manually annotated training data in a basic step. In (Yang et al., 2006) is proposed to annotate video corpora but only considers low level features such as hand position and hand segmentation. In (Nayak et al., 2009) is introduced a method to automatically segment signs by extracting parts of the signs present in most occurrences. However the context-dependency of signs is not considered, e.g. object placement in the signing space. A motion-based approach is presented in (Lefebvre-Albaret and Dalle, 2010) to semi-automatically segment signs. However only the beginning and the end of the sign can be annotated.

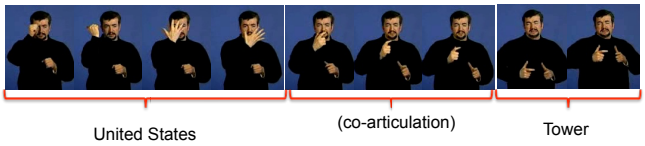


Figure 1: Example of a video sequence with the associated glosses

Unlike other approaches our method semi-automatically annotates SL corpora in terms of glosses. "Glossing" consists on writing down one language in another. It is not about translating the language but transcribing it sign for sign. Various notations can be included for the facial and body grammar that goes with the signs. Figure 1 shows an example of a video sequence of continuous SL with the associated gloss [UNITED STATES] and [TOWER], the sequence in the middle of both signs is called *co-articulation* and corresponds to the meaningless gesture linking the end of a sign and the beginning of the following sign.

We propose a descending approach using image processing techniques to extract sign features, e.g. number of hands, kind of movement, etc. A lexical model of signs is used to determine glosses whose lexical description potentially fits the performed sign. The main contributions of our work is that (i) our approach proposes a list of potential glosses to the annotator speeding up the annotation procedure; (ii)

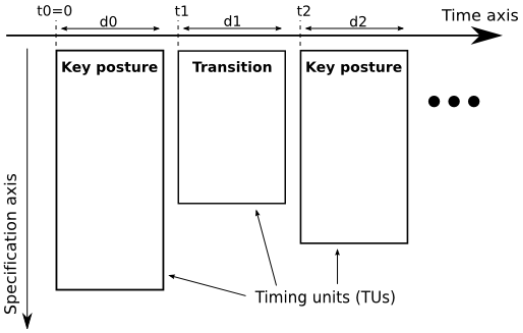


Figure 2: The two Zebedee description axes

Figure 3: Sign [BOX] in French Sign Language. "Source : IVT"

uses a lexical description of signs which takes into account sign variability making the approach context independent; and (iii) no annotated training data is needed since only low level features are extracted to match the sign description.

The remaining of this paper is organized as follows. First it is presented the formal model used in this work, Section 2.. Then it is described the manner of linking image features to a lexical representation, Section 3. Our main experiments and results are detailed in Section 4. Finally our conclusions and further work is discussed in section 5.

2. Zebedee a lexical representation of Sign Language

The formal model chosen for this work is Zebedee (Filhol, 2009) since it deals with body articulator simultaneity and integration of iconic dependencies at the lowest level of description. Unlike parametric notations like HamNoSys (Prillwitz et al., 1989), Zebedee allows grouping all possible performances of one sign under a single parametrised description entry. In Zebedee, a sign is considered as a set of dynamic geometric constraints applied to a skeleton. A description is an alternating sequence of key postures K and transitions T on the horizontal axis in fig. 2, each of which describes a global behaviour of the skeleton over its duration on (time) the vertical axis. A behaviour is a set of necessary and sufficient constraints applied to the skeleton to narrow its set of possible postures over time and make it acceptable for the given sign. In particular, key postures use primitive constraints to geometrically place and orient articulators of the body simultaneously, and transitions use various options to specify the shift from a key posture to the next, otherwise the shift is free. Every stated constraint accounts for a lexically relevant intention, not for an observation of a signed result.

Designed for sign synthesis in the first place, descriptions

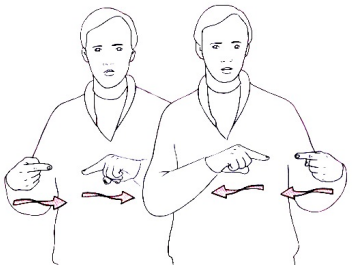


Figure 4: Sign [DEBATE] in French Sign Language

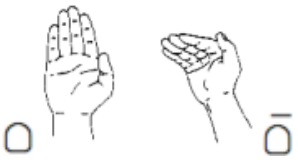


Figure 5: HamNoSys "flat" and "bent" hand configurations

take the articulatory point of view of the signer and not of the reader. It encourages systematic description of intention rather than production, even if it is invisible. For instance, the centre point of the sign [BOX], fig. 3, will be at the heart of the sign description as the hands are placed and moved around it, even if nothing actually happens at that point. Similarly, if a hand is oriented relatively to the other, its direction appears in those terms in the description, and no actual space vector explicitly appears even if it is phonetically invariable. For instance, the left and right index fingers in [DEBATE], fig. 4, respectively point right and left in most occurrences, but these phonetic values do not appear in the description, where the dependency between the fingers is preferred. Moreover, every object or value may refer to other objects or values, which accounts for dependencies between elements of the descriptions. Also, it is possible to refer to named contextual elements, which makes descriptions (hence, signs) reusable in different contexts. In particular, contextual dependencies allow descriptions to adapt to grammatical iconic transformations in context. The example of sign [BOX] is resizeable and relocatable, and according to which is the more comfortable, both  and HamNoSys hand configurations can be used (see fig. 5). It is therefore variable in a lot of ways, but all instances will fit the same zebedescription.

Using such a model for annotation purposes brings a solution to the sign variation problem: by picking up features that can be found in the zebedescriptions instead of phonetic locations or directions, all different instances of a same sign will be found without any further specification. In the case of our example, all signed variations of [BOX] will match the same single description for [BOX], which will be proposed to the user as a potential gloss.

3. Semi-Automatic Annotation of Glosses

Sign features are extracted using image processing techniques to query a zebedescription bank and find the glosses whose description match the performed sign. As result a list of potential glosses is proposed to the annotator.

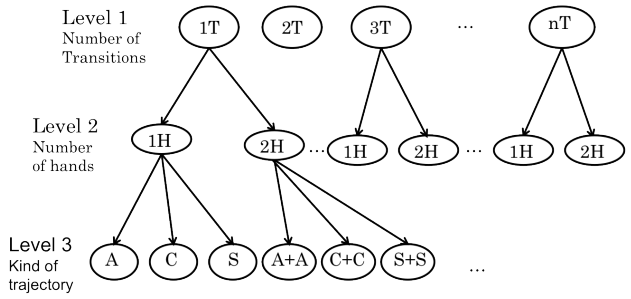


Figure 6: Gloss classification tree

We propose a descending classification method composed of three levels where each level corresponds to a feature extracted from a video sequence and explicitly zebedescribed. To filter the descriptions stored in a PostGres database, we use a dedicated command-based interface to more complex SQL queries, developed at LIMSI. Its 'FILTER' command allows to narrow down the list of descriptions, given a predicate that accepts or rejects description entries, which we give examples for below.

In Zebedee signs are described as an alternating sequence of key postures K and transitions T called *Time Structure*. The number of transitions in a zebedescription is obtained using the *Time Structure* of each sign. This is a discriminant feature for signs classification. For instance over 1600 signs in the sign zebedescription bank from LIMSI, 50% of signs correspond to one transition $(1T)$, followed by 30% and 10% for $3T$ and $2T$ respectively. This is the first feature used in Level 1, fig. 6. The command to obtain signs corresponding to n number of transitions is for example $FILTER\, transcount \sim$ "n".

Even though the number of hands, one hand $(1H)$ or two hand $(2H)$, performing the sign is not explicitly described in Zebedee, it is possible to extract it using other features in the description such as the *Movement Structure*. It corresponds to the kind of trajectory for each hand between two key postures. Trajectories are of three kinds: Arc A, Straight S and Circle C. For two hands movement trajectory is for example $S+S$ where each the first S corresponds to a straight movement for the strong hand and the second S for the weak hand. Image processing techniques are able to determine the number of hands and the kind of trajectory thanks to the velocity and the position of hands for each frame in the sequence. These features correspond to Level 2 and Level 3 in our classification tree, fig. 6, and are used to filter sign from the description bank, for example the command $FILTER\, mvtstruct \sim$ "$S + S$" which filters all signs performed by both hands with a straight movement for each hand.

The image processing techniques developed allow to extract several sign characteristics to query the zebedescription bank. First of all a body part tracking algorithm (Gonzalez and Collet, 2011a) is used to find the position of the head and the hands for each frame of the sequence. It uses the particle filtering principle to track hands and head. Occlusions are handled using the exclusion principle which penalizes other objects that the one associated to the filter. This tracking approach has been specially designed for sign language applications and is robust to hand over face occlusion. Motion features, velocity and acceleration, are extracted from the tracking results.

The proposed approach is not limited to isolated signs but can be used in continuous sign language video sequences. For this we use a semi-automatic sign segmentation approach (Gonzalez and Collet, 2011b). It uses the results of the body part tracking algorithm to extract sign features and detect limits. After motion features, hand shape features are extracted to correct the first segmentation step. Once annotator has labelled signs we are able to propose the list of potential glosses.

Although the number of transition can be determined using the changing limit between single trajectories the velocity and the acceleration of right and left hand, in this work we only address the problems of number of hands and kind of trajectory.

The number of moving hands is determined using the ratio r between the difference of average velocities of right $\bar{v}_1$ and left $\bar{v}_2$ hand and the maximal average velocity, see Eq. 1. If this rate is low that means that one hand moves much faster than the other, otherwise both hands have a similar velocity and might perform the sign. The main problem arises when we process continuous sign language. In this case signs are influenced by the previous sign and itself influences the following sign. For example when a two-hand sign follows a one-hand sign, signers tend to prepare the following sign by moving the weak hand to the beginning location of the two-hand sign. Thus one-hand sign are detected as a two-hand one.

$$r(v_1, v_2) = \frac{\|\bar{v}_1(t) - \bar{v}_2(t)\|}{max\{\bar{v}_1(t), \bar{v}_2(t)\}} \quad (1)$$

Once we have determined the number of hands performing the sign we detect the kind of trajectory which is detected using the position of hands during the whole transition. A circular trajectory is detected using the distance d_n between the first and the last point of the trajectory normalized by the total length of the curve. Thus for a circle C d_n is a low value and for an arc A or a straight S movement is close to 1. This allows to distinguish the signs with a circular trajectory but not arc or straight trajectories can be classified from this measurement.

Straight S and Arc A trajectory have to be differentiated in another way. For this we perform a linear regression and compute the ratio r^2 which give some information about the quality of the fitting. Good quality leads to r^2 close to 1 and means that the fitting has been well performed otherwise the trajectory corresponds to an arc.

Using the features extracted from a video sequence we are able to classify a sign according to our classification tree. Then a list of potential glosses can be proposed to the annotator. Decreasing the number of proposed signs leads to improve the classification tree which depends on the descriptions of signs. For example image processing techniques are able to classify hand shape, however a hand shape Zebedee filter is difficult to implement because the same hand configuration can be described in several ways. The same problem is faced for signs described in terms of a

Table 1: Movement structure statistics (%)

		Strong Hand		
		S	A	C
Weak Hand	S	35.7	0	0
	A	0	60.8	0
	C	0	0	3.46

Table 2: Feature classification results

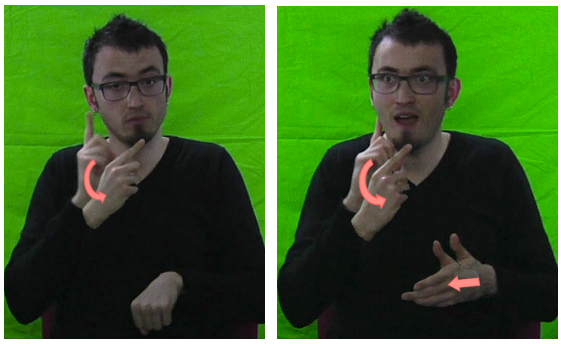
	Ground truth	
Gloss	Nb. H	Traj.
Shoulder bag	1	A
Deaf	1	A
We	1	C
Give	2	C+C

relative position. For instance placing a finger close to the face could be described using the front or the nose position. Classification improvement uses some statistics performed in our sign zebedescription bank, Table 1. For instance for $1T$ no sign performed by two hands have different kind of trajectory for right and left hand, e.g. the movement structure A+S, where A corresponds to an arc for right hand and S to a straight movement for left hand, is not inside our bank. Indeed it is hardly performed by a person. Using this little study we can correct any preparation movement done during continuous SL.

4. Experiments and results

Experiments have been performed on the French Dicta-Sign corpora where vocabulary remains completely free. Glosses have been manually segmented and annotated, table 2 shows some glosses with the number of hands and the kind of trajectory for $1T$. A selection of 95 signs with different number of transitions, number of hands and kind of trajectory is used to perform the experiment. Because of the novelty of our approach it is difficult to perform a comparison to any related work. However we show in this section some encouraging results.

Our experiment considers signs belonging to the $1T$ class which corresponds to 50% of signs in the selection. Table 3 shows the features extracted for some tested signs, number of hands *column: Nb. H* and kind of trajectory *Column: Traj.* with and without statistics improvement. Notice that the performance of signs [SHOULDER BAG] and [DEAF] in different context do not lead to the same extracted features result. Indeed without considering statistics, possible trajectories combination between strong and weak hand shown in table 1, the results are influenced by the context and do not correspond to the ground truth, see Table 2. Figure 7(a) shows the sign [DEAF] in French Sign Language (LSF). It corresponds to $1H$ and an A movement. Figure 7(b) shows the performance of the same sign in a different context, this time left hand moves straight. In this context signer prepares the following sign which corresponds to a sign performed with two hands. This is improved using statistics over the movement structure. In fact



(a) (b)

Figure 7: Sign *Deaf* in French Sign Language in different context

Table 3: Feature classification results

	Without statistics		With statistics	
Gloss	Nb. H	Traj.	Nb. H	Traj.
Shoulder bag	1	A	1	A
Shoulder bag	2	A+S	1	A
Deaf	1	A	1	A
Deaf	2	A+S	1	A
We	1	C	1	C
Give	2	C+C	2	C+C

a movement $A + S$ is hardly performed by a human and since one hand is moving to prepare the following sign the faster way of going from one point to another is through a straight S movement. Therefore the S is deleted.

Using the extracted features to query the zebedescription bank we propose the potential glosses to the annotator. The number of proposed glosses for some signs is shown in table 4. Figure 8 shows the sign [WE/US] in LSF with the potential glosses sorted alphabetically.

This results are promising and show that the selected features are discriminant though other features will be added in the future to improve our annotation approach.

5. Conclusion and Further work

We have presented an approach to assist the annotation using a lexical description of signs. This approach extracts image features from video corpora to query a sign description bank and propose the potential glosses that could correspond to the performed sign. Experiments have shown promising results. This approach can be used to annotate any kind of gestures or SL described with the formal model used in this work. Further work focus on the introduction of hand configuration in our classification tree as well as other motion features.

6. References

C. Cuxac. 2000. *Langue des signes française, les voies de l'iconicité*, volume 15–16. Ophrys.

P. Dreuw and H. Ney. 2008. Towards automatic sign language annotation for the elan tool. In *LREC Workshop on the Representation and Processing of SL: Construction and Exploitation of SL Corpora*.

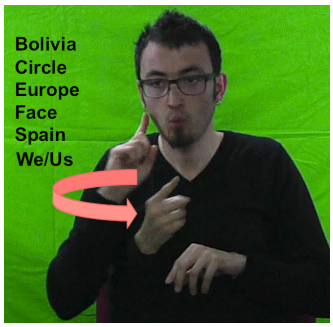


Figure 8: Sign *we/us* in FSL showing the potential glosses

Table 4: Number of potential glosses

Gloss	Nb. of proposed glosses
Shoulder bag	20
Deaf	20
We/Us	6
Give	8

M. Filhol. 2009. Internal report on zebedee. Technical Report 2009-08, LIMSI-CNRS.

M. Gonzalez and C. Collet. 2011a. Robust body parts tracking using particle filter and dynamic template. In *18th IEEE ICIP*, pages 537–540.

M. Gonzalez and C. Collet. 2011b. Signs segmentation using dynamics and hand configuration for semi-automatic annotation of sign language corpora. In 9th International Gesture Workshop, editor, *Gesture in Embodied Communication and Humain-Computer Interaction*, pages 100–103, May.

F. Lefebvre-Albaret and P. Dalle. 2010. Body posture estimation in sign language videos. *Gesture in Embodied Communication and HCI*, pages 289–300.

S. Nayak, S. Sarkar, and B. Loeding. 2009. Automated extraction of signs from continuous sign language sentences using iterated conditional modes. *CVPR*, pages 2583–2590.

R. Prillwitz, S.and Leven, H. Zienert, T. Hanke, and J. Henning. 1989. Hamnosys version 2.0, hamburg notation system for sign languages, an introductory guide. *International studies on Sign Language communication of the Deaf*, 5.

R. Yang, S. Sarkar, B. Loeding, and A. Karshmer. 2006. Efficient generation of large amounts of training data for sign language recognition: A semi-automatic tool. *Computers Helping People with Special Needs*, pages 635–642.