Video retrieval in sign language videos: how to model and compare signs ?

F. Lefebvre-Albaret, P. Dalle

Institut de Recherche Informatique de Toulouse
Université Paul Sabatier, 118 Route de Narbonne
F-31062 TOULOUSE CEDEX 9
lefebvre@irit.fr, dalle@irit.fr

Abstract

This paper deals with the problem of finding sign occurrences in a sign language (SL) video. It begins with an analysis of sign models and the way they can take into account the sign variability. Then, we review the most popular technics dedicated to automatic sign language processing and we focus on their adaptation to model sign variability. We present a new method to provide a parametric description of the sign as a set of continuous and discrete parameters. Signs are classified according to there categories (ballistic movements, circles ...), the symmetry between the hand movements, hand absolute and relative locations. Membership grades to sign categories and continuous parameter comparisons can be combined to estimate the similarity between two signs. We set out our system and we evaluate how much time can be saved when looking for a sign in a french sign language video. By now, our formalism only uses hand 2D locations, we finally discuss about the way of integrating other parameters as hand shape or facial expression in our framework.

1. Introduction

The automatic Sign Language (SL) processing is a specific problem related to both speech and gesture processing. Various parameters are involved in a sign production: hand shape, placement, movement, facial expression, and gaze. These sign features are heterogeneous and vary deeply between two realizations of the same sign.

This paper addresses the problem of facilitating the step of looking for a specific sign in a video containing French Sign Language (LSF) utterances. The target videos (in which we want to locate the sign query) are free expressions or translations of pieces of news. Their lexicon and grammatical structure have not been constrained. The signers do not wear any additional visual markers to make the automatic tracking easier.

The article first deals with the problem of sign models in French Sign Language. We then explain how we modified some methods dedicated to SL processing to take into account the sign variability. We finally evaluate our parametric model in order to quantify the time gain for a sign retrieval.

2. Linguistic model

The first step is to be able to look for a sign in video is to define a sign model. Several approaches have been proposed by the specialists of linguistic and computer sciences to select the relevant features in the video in order to correctly identify the signs.

Stokoe (W.C. Stokoe, D. Casterline and C. Croneberg, 1978) proposed a parametric model based on the sign decomposition into parameters. He identifies three relevant parameters: the hand shape (linked with the hand orientation), the movement, and the global sign placement.

Lidell and Johnson (Liddell and Johnson, 1990) proposed a multi-segmental sign model based on a temporal segmentation of the sign. In this formalism, a gesture is described as a sequence of hold and movement timing units. This model takes into account the synchronization of the parameters during the sign production. In each timing unit, the sign parameters are described separately. This sign representation highlights a lot of regularities in sign temporal structures like the repetitions or the sign dynamics. The relationships between the movements of right and left hands (symmetries, translations, alternation) are also easily visible. Any gesture can be modeled with this method, even if its structure is not compatible with the French Sign Language phonology.

The mono-segmental model proposed by (Channon, 2002) can be opposed to the former multi-segmental model. The author first underlines the impressive amount of the American Sign Language gestures that involve repetitions compared to the part of English words with a repetitive structure (respectively 50% vs. 1%) to justify the necessity of considering the signs as single-segments, even if they can be further decomposed into smaller units. The sign is then characterized by a set of features like the repetition, the relationship between the right and the left hand, the hand shape, the orientation, the motion direction and the location.

Nowadays, most of the algorithms dedicated to automatic SL processing are based either on Hidden Markov

Models or Dynamic Time Warping and implicitly rely on multi-segmental models. On the contrary, the models used to generate sign videos (Filhol, 2008) (Losson, 2000) include more and more parameters as repetitions or symmetries that refers to mono-segmental models.

Our goal is to be able to locate a sign in signed utterance videos. The tracking in the video query produces hand trajectory estimations that may include some estimation errors caused by the noise in the input video. For this reason, we choose to base our algorithm on a mono-segmental sign model that gives restrictions about sign structures and makes out method more robust. However, we verified that the mono-segmental sign representation is able to deal with French Sign Language signs. Among 4027 signs of the dictionary (Moody, 1997) (we excluded all the compound signs), 87% of projected 2D movements on the video plane can be classified into a small set of movement categories described in §5.1.. The other 13% involve either too elaborated movements or complex relationships between the two hands. Like Channon (Channon, 2002), we noticed an important frequency of sign repetitions (35% of the signs) and of the symmetries between the right and left hands (32%).

3. Existing processing methods

A lot of contributions have already been proposed in the field of automatic SL recognition. The best outcomes come from the processing of corpora acquired by motion capture (Gao et al., 2004)(Vogler and Metaxas, 2003) where up to 5000 signs can be correctly segmented with a 90% correct identification rate. When the only available inputs are uncalibrated videos (like videos that can be found on internet), such a recognition rates (90%) can only be obtained from a sign corpus of at most hundreds of signs.

The majority of the studies try to adapt the methods dedicated to automatic speech processing like in (Zahedi et al., 2006). The most frequently used methods are Hidden Markov Chains (Zahedi et al., 2006)(Pei et al., 2009)(Bauer and Kraiss, 2002)(Vogler and Metaxas, 1999)(Brand et al., 1997)(Deng and Tsui, 2002) and Dynamic Time Warping based algorithms (Han et al., 2007) (Alon, 2006) that make use of dynamic programming.

It is interesting to notice that those traditional methods take more and more into account the parametric nature of signs:

- A first adaptation consists in dissociating the different sign features (movement and hand shape of each hand) that are processed separately. This approach is explained in (Vogler and Metaxas, 1999)(Brand et al., 1997) and (Deng and Tsui, 2002). The models used for each sign feature can be used to model several signs. As a consequence, the total size of the recognition models decreases.

- Other improvements proposed by (Kim et al., 2001) make use of Markov Chains to model recurrent primitives of the Korean Sign Language. The modification of Markov Chain structures allows the model to explicitly take into account the repetitions. The grouping of similar sign patterns is also an approach that we used in (Lefebvre-Albaret and Dalle, 2008) to segment the video into signs.

- The last approach uses parametric DTW or HMM. It is the case in (Wilson and Bobick, 2001) where the states of the Markov Model are modified according to the sign orientation and the sign amplitude. (Alon, 2006) takes also into account the sign translation in the signing space in parametric DTW.

4. Our method

Those evolutions are justified at the linguistic level (cf. §2.) and show better recognition rates than the traditional approaches. Our method combines the enhancements that were mentioned in the §3.: the dissociation and synchronization of the sign feature detection, the signs categorization and the modeling of their variability.

The algorithm is only based upon the projected movement of hand and head centroids. Our method processes three-dimensional data (two dimensions in time and one in space). The head and hand trajectories have been evaluated by means of the tracking algorithm described in (Lefebvre-Albaret and Dalle, 2009). We do not integrate yet the configuration and orientation information although those features are obviously important for a sign identification. The determination of these two parameters from a monocular video is so far an open problem. We want to evaluate the results that can be obtained in using only the hand and head trajectories. As a lot of signs only differ in their hand orientations and hand shapes, our aim is not to achieve a complete sign recognition. We rather want to define similarity criterion based on sign movements, in order to reduce the searching space where the sign query might be in the video like in (Alon, 2006).

The signs are characterized by means of a parametric approach. Each sign is represented as a collection of the following parameters: relative hand locations, place of the hands besides the head, hand contacts, hand trajectories, movement dynamics, relationship between the two hand movements, orientation and amplitude of the movement. The detection filters are applied to each time interval $[t_1, t_2]$ of the video but the filters do not allow any time warping unlike the HMM and DTW methods. This constraint of linear time deformation is intentional because we estimate that it better models the regular structure observed in the repetitive signs that are very frequent in French Sign Language.

5. Sign Characterization

Our system takes as an input 2D coordinates of the right hand $(X_r(t), Y_r(t))$, of the left hand $(X_l(t), Y_l(t))$, and of the head $(X_h(t), Y_h(t))$ at each time step t. Those measures come from a tracking algorithm (Lefebvre-Albaret and Dalle, 2009) on 2D videos of the signer. The movement characterization is based upon the instantaneous speeds and locations of the signer's hands.

In each time interval where one filter has to be applied, the measures are first resampled so that each interval can be characterized by vectors of N instantaneous speeds (we choose $N = 16$ in our implementation). Those speeds will be named $(Vx_r(i), Vy_r(i))$ and $(Vx_l(i)\ , Vy_l(i))$ in the following pages.

Two types of processing are applied to the tracking measures:

- The **categorization filters** provide grade of membership of a video time segment in a sign category. For the moment, we only take into account six motion categories: ballistic motions, back and forth, repeated ballistic, circular, repeated circular and angular because they are highly represented in the French Sign Language lexicon. We also distinguish the signs where only one hand moves from the signs where the hand movements are symmetrical. As a consequence, we consider 6 x 2 = 12 sign categories that gather 87% of the French Sign Language lexicon.
- The **comparison operators** indicate the similarity of two signs (of the same category) based on one of their features. Those operators deal with continuous parameters. The comparison operators that are used in our algorithm take as an input the gesture amplitude, orientation, relative hand position and sign location.

5.1. Categorization filers

The categorization filters are based on the geometry of the sign trajectory, the motion dynamic as well as the relationship between the two hand movements. Each filter processes resampled speed of right and left hands and provides a membership grade in the sign to the category ranging from 0 to 1. The filters have been set manually according to sign models described in (Losson, 2000). We plan to replace it by sign models learned from motion capture data in order to make it closer to the real sign utterances.

The **geometrical filter** qualifies the shape of hand trajectories. It is made of an array of N angles $[\alpha_1, \alpha_2 \ldots \alpha_N]$ that represent the instantaneous orientation of hand speeds. The membership grade of the sign in the geometrical categories are computed in the orthonormal frame $(\vec{u}, \vec{v})$:

$$\overrightarrow{Lin} = \sum_{i=1}^{N} \left(V_x(i)cos(\alpha_i) + V_y(i)sin(\alpha_i)\right) \vec{u}$$
$$. \quad + \sum_{i=1}^{N} \left(V_y(i)cos(\alpha_i) - V_x(i)sin(\alpha_i)\right) \vec{v}$$

A rotation is applied to each instantaneous speed.

$$\textbf{GS} = \frac{\| \overrightarrow{Lin} \|}{\sum_{i=1}^{N} \sqrt{V_x^2(i) + V_y^2(i)}}$$

It is important to notice that this operator is not affected by the sign translation, scaling or rotation in the 2D plan of the video.
It is then possible to determine the orientation Theta of the dominating hand[1].

$$\theta = (\vec{u}, \overrightarrow{Lin})$$

It is possible to project the instantaneous speeds on the theoretical sign trajectory. Those projected speed $V'(i)$ will be used to compute the sign dynamic.

$$V'(i) = V_x(i).cos(\alpha_i - \theta) + V_y(i).sin(\alpha_i - \theta)$$

The **dynamical filter** characterizes the speed profile of the sign. The filter is made of a speed profile template of N speeds $[V_p(1), V_p(2) \ldots V_p(N)]$. The membership grade of the sign in the dynamic category is computed by means of a normalized scalar product between the template and the real profile.

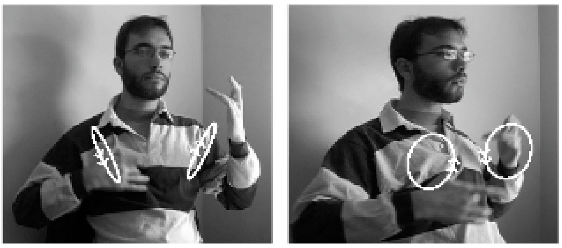
$$\textbf{DS} = \frac{\sum_{i=1}^{N} V'(i).V_p(i)}{\sqrt{\sum_{i=1}^{N} V'^2(i)}\sqrt{\sum_{i=1}^{N} V_p^2(i)}}$$

The result is smaller than 1 and is not affected by he sign translation, scaling or rotation in the 2D plan of the video.
Other measures quantify the dependency between the two hand movements. As shown in (Filhol, 2008), most signs are effectuated either with one hand, or with two hands having a symmetrical movement (footnote: symmetry means here simple dependency between the two hand speeds. The meaning is here extended to relationships like translation or alternating movements). It is important to be careful because the kind of symmetry might change according to the camera point of view. The fig.1 presents a sign seen with two different points of view :

- When the camera faces the signer, the movement seems to have an alternated symmetry $(Vr_x(i) \approx Vl_x(i)\ and\ Vr_y(i) \approx -Vl_y(i))$
- When the camera films the signer on its side, the two hand motions seem to be linked by a central symmetry $(Vr_x(i) \approx -Vl_x(i)\ and\ Vr_y(i) \approx -Vl_y(i))$

The **static hand filter** MS detects whether only one hand is moving during the sign realization.

[1] right hand for a right handed person



[Theatre] front view
$\sigma = -125°$
Alternated symmetry

[Theatre] side view
$\sigma = -71°$
Central symmetry

Figure 1: Projection bias

$$\textbf{MS} = \frac{\sum_{i=1}^{N}(\| \overrightarrow{V_r(i)} \| - \| \overrightarrow{V_l(i)} \|)}{\sum_{i=1}^{N}(\| \overrightarrow{V_r(i)} \| + \| \overrightarrow{V_l(i)} \|)}$$

The **symmetry filter** SYM detects a symmetry between the two hand movements $(Vr_x(i) \approx \pm Vl_x(i) \;\; and \;\; Vr_y(i) \approx \pm Vl_y(i))$.

$$C_x = \frac{sgn(\sum_{i=1}^{N} Vr_x(i).Vl_x(i))\sqrt{\left|\sum_{i=1}^{N} Vr_x(i).Vl_x(i)\right|}}{\sqrt{\sum_{i=1}^{N} max(V_r^2(i), V_l^2(i))}}$$

The measure C_y is computed like C_x.

$$\textbf{SYM} = \sqrt{C_x^2 + C_y 2}$$

As it can be seen on the fig. 2, the signs of C_x and C_y can be analyzed to provide the kind of symmetry involved in the sign. However, we decided to use a continuous angle measure σ that is less affected by the projection bias.

It is possible to combine all those confidence measures to compute a membership grade of the video time segment in one of the sign categories. The filter results are combined in a probabilistic way (with a naive Bayesian fusion). The confidence measures of the right and left hand are respectively denoted (DS_r, GS_r) and (DS_l, GS_l).

The membership grades of a video segment in a category are computed in the following way:
For a sign effectuated with the right hand
$SCORE = GS_r.DS_r.MS$
For a sign effectuated with the left hand
$SCORE = -GS_l.DS_l.MS$
For a sign effectuated with both hands
$SCORE = SYM.\sqrt{GS_r.DS_r.GS_l.DS_l}$

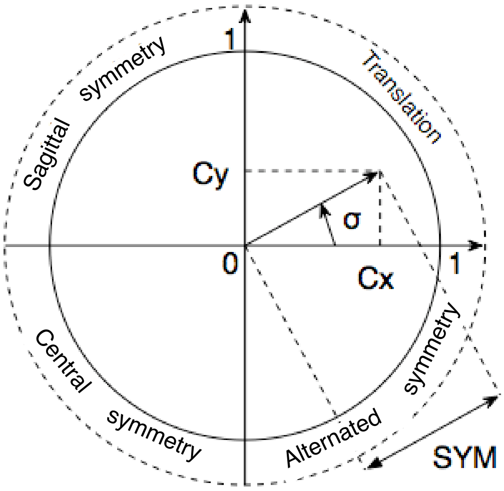


Figure 2: symmetry classes

5.2. Comparison operators

Other measures that we will explain in the present chapter evaluate the similarity between two signs. Each of them compares one of the parameters of the query sign to the corresponding parameter in the target sign and provides a similarity measure between 0[2] and 1[3].

Several sign parameters are compared:
AMP: Movement amplitude
CONT: Contact between hands
TR: Relative hand location
PL: Hand location (relatively to the head location)
OR: Sign orientation (Measured by θ)
TS: Kind of symmetry (measured by σ)

5.3. Similarity measure between two signs

The similarity grade of the query sign S_s with a time segment of the target video will be computed in combining the results of the categorization filters and the comparison operators. The fusion method is chosen according to the sign category. There are four different kinds of fusion. The following example corresponds to a sign where both hands are effectuating a ballistic movement and are linked by a translation. In the following similarity grade computation, the confidence measures GS_r, DS_r, GS_l, DS_l, SYM refer to the video segment to characterize with the cat_s category.

$SCORESIM = (SYM * GS_r * DS_r * GS_r * DS_l) * (CONT * AMP * TR * PL * OS * TS)$

The final similarity grade between the query sign S_s

[2] 0 : The parameters of the query and target signs have nothing in common.

[3] 1 : The parameters of the query and target signs are equal.

and the time segment of the target video is computed by means of a product of the categorization filter and comparison operator results.

6. Video query processing

We seek to solve the problem of finding in a video V_t all the occurrences $S_t(j)$ of a sign. The query sign S_s is contained in the video V_s. Rather than directly solving this problem like in (Alon, 2006), we choose to split it into two steps:

A - Sign query characterization, determination of the cat_s category.

1. The video query including the sign query is recorded. An other algorithm described in [12] tracks the hand and head positions.
2. Membership grades of the query sign to each sign category cats are computed for each time segment [t1,t2] of Ss as explained in §5.1..
3. The propositions are sorted by membership grade and presented in the form of sign pictures (like in the figure 1). The user can then select the right category Cat_s by using the arrow shape indicating the sign pattern as well as the relationship between the two hand movements.

B - Sign search in the target video

1. Hand and head trajectories are tracked in the target video Vt.
2. The categorization filters corresponding to Cat_s and the comparison operators are applied to each time interval of Vt which duration is less than Tmax (2 s in our implementation).
3. The video segments are sorted by similarity grade with the sign query. Each segment is then presented to the user.

7. Evaluation

The goal of our work is to make the sign search in a video easier. As a consequence, it is natural to check whether our algorithm generates time saving in the task of looking for a specific sign in a video. To test our algorithm on a representative set of French Sign Language utterances, the test videos are chosen from free narrations and piece of news translations provided by Websourd[4] society. The evaluation consists in processing 103 different requests. Some of the signs are present several times in the target videos. As a consequence, there are 178 good answers to the requests. We intentionally excluded all signs of less than 3 frames (corresponding to 0.1s) that are hard to identify out of context, even by a native signer. Although our method is based on a comparison of sign movements, we decided not to exclude all the signs that involve only a small (or no) projected 2D movement. Those sign represent 10% of the French Sign Language gestures and involve other parameters like the hand shape or hand orientation changes. However, those signs can be characterized by the relative hand location or the overall sign placement.

For each sign query, our algorithm processes each time segment of the target video and classifies them according to their similarity grade with the sign query. The relative rank[5] of the correct[6] answer(s) corresponding to the request are reported in the diagram of the fig. 3.

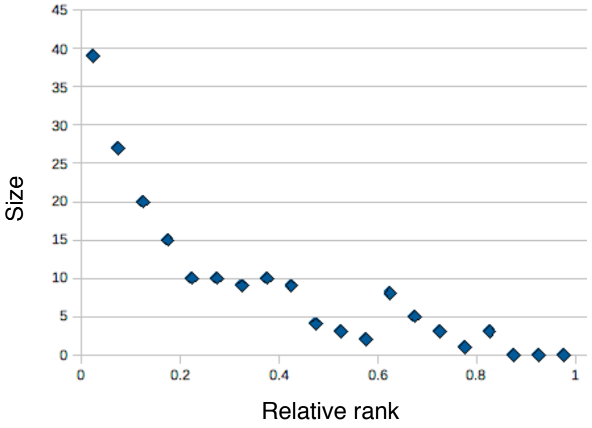


Figure 3: relative ranks of the good answer to the video queries

We can draw several quantitative conclusions of the above-mentioned results:

- More than 50% of the correct answers are located in the first time segments of the video presented to the user. Those propositions represent only 15% of the total amount of segments.
- It is possible to deduce from the curve (fig. 1) that the sign search will be about two times faster in visualizing the segments proposed by our algorithm ranked by similarity grades.
- Sign involving large movements like the sign “building” can be located in average 15 times faster when using our method.

[4] available at www.websourd.org

[5] The relative rank of the good answer is the ratio between the rank of the correct answer and the overall amount of possible answers

[6] A time segment is said to be correct when it contains more than the half of a sign corresponding to the video request

Other qualitative tendencies can be observed from the results:

- The signs involving small movements are much harder to locate in the video. We assume that other parameters like hand orientation and hand shape should be necessary to achieve a sufficient characterization.

- In our evaluation, we observed systematic changes of parameters between the sign query and the corresponding signs in the target videos. For instance, it is frequent that repeated sign loose their repetitions when they are used in signed utterances. However, using a lot of other sign features allows our algorithm to give satisfying results, even in those cases.

8. Conclusion and perspectives

Our results are very promising and some applications could be developped on the short run. One of them consists in optimizing the navigation in long French Sign Language documents by means of keysigns.

The presented framework could easily embed other sign features like hand shape, hand orientation and facial expression in order to use them in sign queries. It would be interesting to use learning algorithm to optimize the filters used to detect the sign geometry and dynamic.

As shown in studies like (Channon, 2008), some constraints of symmetry and repetitions can also be found in gestures involved in co-verbal communication. Then, our algorithm should be used in this domain. In the same way, the parametric gesture modeling could lead to significant improvement in the field of Human Computer Interaction (Wilson and Bobick, 2001). An adaptation of our algorithm to those two domains could help to take into account the variability in the production of gesture units.

9. References

J. Alon. 2006. *Spatiotemporal gesture segmentation*. Ph.d. diss., University of Boston, USA.

B. Bauer and K. Kraiss. 2002. Video-based sign recognition using self-organizing subunits. In *ICPVR*, Quebec.

M. Brand, N. Oliver, and A. Pentland. 1997. Coupled hidden markov models for complex action recognition. In *ICCVPR*, Porto Rico.

R.E. Channon. 2002. *Signs are single segments: phonological representations and temporal sequencing in American Sign Language and other Sign Languages*. Ph.d. diss., University of Maryland, USA.

R.E. Channon. 2008. he symmetry and dominance conditions reconsidered. In Chicago Linguistic Society, editor, *Journal Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 40, pages 45–47, Chicago.

J. Deng and H.T. Tsui. 2002. A two-step approach based on pahmm for the recognition of amercian sign language. In *ACCV*, Melbourne.

M. Filhol. 2008. *Descriptive model for an automatic Sign Language processing (in French)*. Ph.d. diss., University Paris-1, France.

W. Gao, G. Fang, D. Zhao, and Y. Chen. 2004. Transition movement models for large vocabulary continuous sign language recognition. In *Face and Gesture*, pages 553–558, Seoul.

U. Han, G. Awad, and A. Sutherland. 2007. Subunit boundary detection for sign language recognition using spatio-temporal modelling. In *ICVS*, Bielefeld.

J.B. Kim, K.H. Park, W.C. Bang, J.S. Kim, and Z. Bien. 2001. Continuous korean sign language recognition using automata based gesture segmentation and hidden markov model. In *ICCAS*, Jeju, Corea.

F. Lefebvre-Albaret and P. Dalle. 2008. An approach of french sign language segmentation (in french). In *TALN*, Avignon.

F. Lefebvre-Albaret and P. Dalle. 2009. Analysis of sign language videos, methods and strategies (in french). In *ORASIS*, Trégastel.

S.K. Liddell and R.E. Johnson. 1990. American sign language: the phonological base. *Sign Language Studies*, 64.

O. Losson. 2000. *Synthesis of communicative gestures, application to French Sign Language (in French)*. Ph.d. diss., University of Lille, France.

B. Moody. 1997. *Sign Language (in French), volume 2 and 3*. IVT.

T. Pei, Starner, H. Hamilton, I. Essa, and J.M. Rehg. 2009. Learning basic units in american sign language using discriminative segmental feature selection. In *ICASSP*, Taipei.

C. Vogler and D. Metaxas. 1999. Parallel hidden markov models for american sign language recognition. In *ICCV*, Kerkyra.

C. Vogler and D. Metaxas. 2003. Handshapes and movements: Multiple-channel american sign language recognition. In *Gesture Workshop*, pages 247–258, Genova.

W.C. Stokoe, D. Casterline and C. Croneberg. 1978. *A dictionary of American Sign Language on Linguistic principles*. Linstok Press.

A.D. Wilson and A. Bobick. 2001. Hidden markov models for modeling and recognizing gesture under variation. In *International Journal on Pattern Recognition and Artificial Intelligence*, volume 15, pages 123–160.

M. Zahedi, P. Dreuw, D. Rybach, T. Deselaers, J. Bungeroth, and H. Ney. 2006. Continuous sign language recognition – approaches from speech recognition and available data resources. In *LREC Workshop on the Representation and Processing of Sign Languages*, pages 21–24, Genova.