MODEL-LEVEL DATA-DRIVEN SUB-UNITS FOR SIGNS IN VIDEOS OF CONTINUOUS SIGN LANGUAGE

Stavros Theodorakis, Vassilis Pitsikalis and Petros Maragos

National Technical University of Athens, School of ECE, Athens 15773, Greece.
{sth,vpitsik,maragos}@cs.ntua.gr.

ABSTRACT

We investigate the issue of sign language automatic phonetic subunit modeling, that is completely data driven and without any prior phonetic information. A first step of visual processing leads to simple and effective region-based visual features. Prior to the sub-unit modeling we propose to employ a pronunciation clustering step with respect to each sign. Afterwards, for each sign and pronunciation group we find the time segmentation at the hidden Markov model (HMM) level. The models employed refer to movements as a sequence of dominant hand positions. The constructed segments are exploited explicitly at the model level via hierarchical clustering of HMMs and lead to the data-driven movement sub-unit construction. The constructed movement sub-units are evaluated in qualitative analysis experiments on data from the Boston University (BU)-400 American Sign Language corpus showing promising results.

Index Terms— sign language, subunit modeling, HMM

1. INTRODUCTION

Sign languages, i.e., languages that essentially convey information via visual patterns, commonly serve as an alternative or complementary mode of human communication. Visual patterns, as opposed to the audio ones used in the oral languages, are formed by hand shapes and manual or general body motion, lip movements and facial expressions. Their expressiveness facilitates human interaction and exchange of information not only in the existence of hearing-impaired people but also in situations where speech is impractical, e.g., in loud workspaces. However, efficient communication by these means is only feasible between specially trained interacting parties. In this context, automatic sign-to-text and text-to-sign translation can be viewed as the intermediate technological modules that can partially lift this restriction.

First attempts on automatic Sign Language recognition were restricted to simple recognition tasks [1] similarly to cases of speech recognition a few decades ago. An informal correspondence of the word in spoken language is a sign unit, given that sign language tend to be monosyllabic [2]. There are several metaphors between sign and speech recognition that allow for the exchange of methods between the two areas. However, there exist points of difference too [2]. A diversity that has also practical effects concerns phonological sub-units. There is not yet a well-defined unit equivalent to the phoneme in speech. In this paper, we focus on automatic data-driven modeling of sub-units without any phonetic information. This research direction is important both in order to face the phonetic modeling of intra-sign sub-units and for the practical case of automatic recognition.

This research work was supported by the EU under the research program Dictasign with grant FP7-ICT-3-231135.

The field of sign language recognition is certainly in the focus of quite intense research lately [1]. It is considered to be a multilevel problem and it poses significant challenges regarding data collection, visual processing and information stream modeling for recognition. Vogler and Metaxas [3] broke down signs into their constituent sub-units using the basic ideas of the Movement-Hold model [4] and applied successfully the so-called Parallel HMMs. Bauer and Kraiss [5], on the other hand worked also at the sub-unit level exploring a data-driven approach for modeling the intra-sign units. They cluster independent frames utilizing K-means. They produced sub-units named as phenones and further employed a 2-state HMM for their modeling. Fang et al. [6] and Han et al. [7] have also proposed approaches for data-driven sub-unit modeling. They employed clustering by considering segments and not only independent frames as [5] at the feature level taking advantage of the dynamics that are essential in sign language. Modeling at the sub-unit level provides a powerful method in order to increase the vocabulary size and deal with more realistic data conditions.

Based only on simple position measurements, we proceed on the sub-unit modeling of sign language at the model level, that refers to the modeling of intra-sign segments. Given the lack of annotation information within the sign units we assemble our approach by attempting at first an initial segmentation step. We employ for each sign, a model based segmentation at the state level, similar to [6]. Yet, this may be proved poor due to factors that introduce variation in the realization of signs. To cope with this pronunciation variation per sign we propose to precede the segmentation step by pronunciation clustering. Given the segmented sign we are equipped with a prosperous initialization step to face the intra-sign segments' modeling. Our goal is to cluster not the independent frames as if they were in a common pool [5], neither the feature frames sequences as segments themselves at the feature level [6, 7]. Instead, we propose to hierarchically cluster whole dynamic models (HMMs) [8] based on a similarity measure among models. We evaluate the proposed methods by qualitative experiments analyzing the mapping among the created models and the real movement data, showing promising results. In all experiments we employ real data from the Boston-University continuous American sign language corpus (BU400) [9].

2. DATA AND FEATURE EXTRACTION

2.1. Continuous Sign Language Corpus: BU400

The BU400 [9] is a continuous American Sign Language (ASL) database and consists of 843 utterances over a vocabulary of 406 words and four different signers. The background is uniform. The camera setup consists of three cameras, among which we used the one facing front. The transcriptions are in the sign level, consisting of English signs, with annotated start and end points.

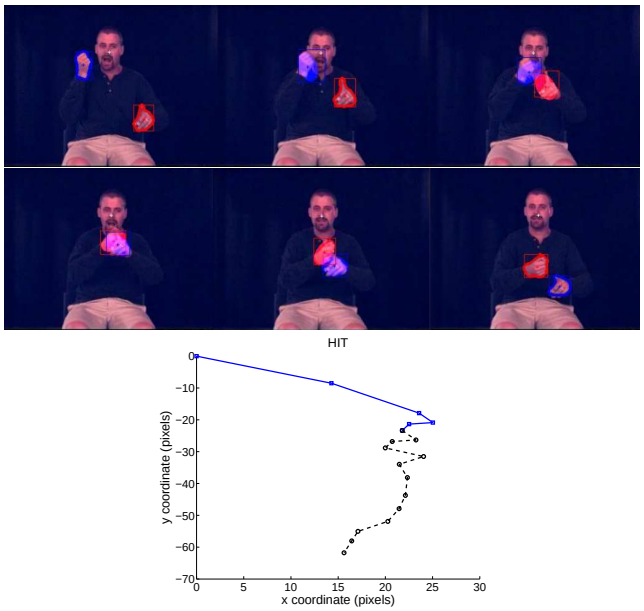

Fig. 1. Video frames (left to right, upper to bottom row): Progressive frames that correspond to the realization of the sign "HIT". Bottom row: Movements as sequences of positions in x,y-coordinates normalized to the initial position of the sign; the same realization for the sign "HIT" as shown in the video frames. 1) Continuous lines show the 1st segments, 2) dashed lines correspond to the 2nd segments as they have been produced by the segmentation step. The marker (and color) in the lines corresponds to the sub-unit cluster that each segment belongs.

2.2. Feature extraction on the BU400 Database

For the hand and head detection we employ a probabilistic skin color model that uses as initialization manually annotated skin color areas. In this way we estimate the probability of each pixel belonging to skin. This probabilistic map is then used as a force in the Geodesic Active Contour model [10, 11] enforcing the curve to converge eventually to the edges that separate the skin region from the background.

Next, we face the cases of occlusions during tracking that emerge when one hand is in front of the other or the head. We disambiguate occlusions by a linear forward-backward prediction of the centroid of each hand and looking on following or previous frames. This is combined with a template matching scheme.

Finally, we extract features related to the position, the movement and the shape of the hands. In the presentation that follows we only take advantage of movement features. Via such simple features we aim on understanding their effect on sub-unit modeling. Besides, movement and position are among the main characteristics that describe a sign [2, 1].

3. SUB-UNIT MODELING AT THE MODEL LEVEL

Data Selection: In the experiments described next we use only the front camera video stream. Among the whole corpus, we restrict our processing on six videos that contain stories narrated from a single signer; these are identified namely as: `accident`, `biker_buddy`, `boston_la`, `football`, `lapd_story` and `siblings`. We utilize 20 signs among the most frequent, sampled from all stories.

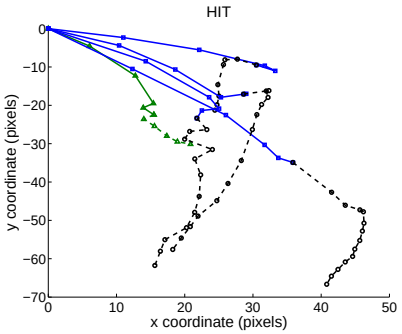

Fig. 2. Multiple realizations of the same sign "HIT", containing the realization shown in Fig. 1. We show movements as sequences of positions in x,y-coordinates normalized to the initial position of the sign. 1) Continuous lines show the 1st segments, 2) dashed lines correspond to the 2nd segments as they have been produced by the segmentation step. The marker and color in the lines corresponds to the sub-unit cluster of each segment.

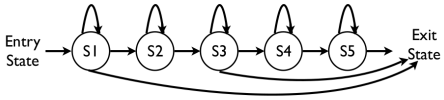

Fig. 3. HMM topology as employed for segmentation.

3.1. Sign Pronunciation Clustering

To account in a simplified way for the variation of the different realizations of the same Sign we construct clusters with respect to each Pronunciation of the sign (SP); this step is repeated for all signs. The rationale for Gloss Pronunciation (GP) clustering is that the produced clusters are more compact, affecting on their turn the segmentation step. In order to cluster the different examples we are using an hierarchical clustering algorithm. Distances among sequences are computed by employing dynamic time warping to account for segments of different lengths combined with the $L2$ norm. After experimental observations, we practically employ 3-5 clusters per gloss.

3.2. Segmentation

Using the partitioning of the GP clustering we train one HMM for each sign pronunciation. For training we consider all different realizations of each GP cluster. The HMM topology employed is a 5-state left-right HMM (Fig. 3) allowing entrance and exit transitions from its first and third state. After training each GP HMM we perform a Viterbi alignment resulting to the most probable segmentation point at the state level. The duration in the model between the first and the third state and between the third and the fifth state corresponds to the first and the second segment respectively. In this way we constrain the decomposition of signs in one or two sub-units. This fits with sign language aspects that sign consist of a single or two movements [2].

3.3. HMM based sub-units by HMM clustering

Attacking the issue of intra-sign sub-unit modeling at the HMM model level provides advantages compared to the signal level approach: for instance, we take advantage of the explicit dynamic mod-

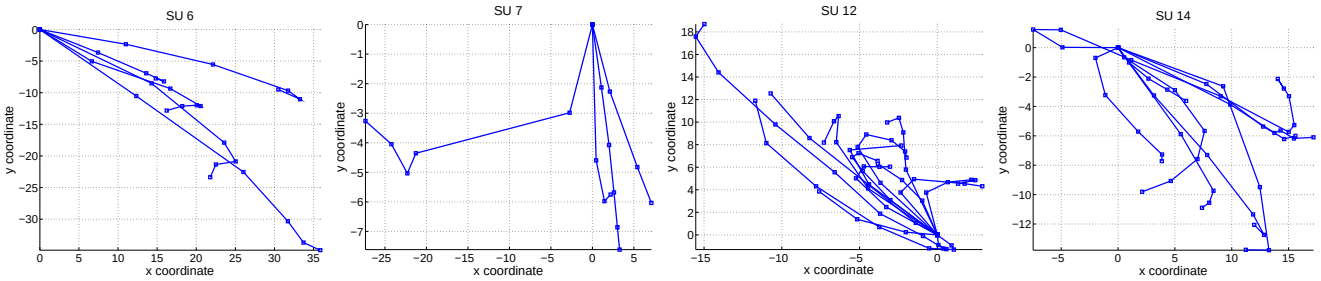

Fig. 4. Segments of position sequences normalized to the initial position of the segment. Each figure contains segments that correspond to a single sub-unit model after HMM clustering.

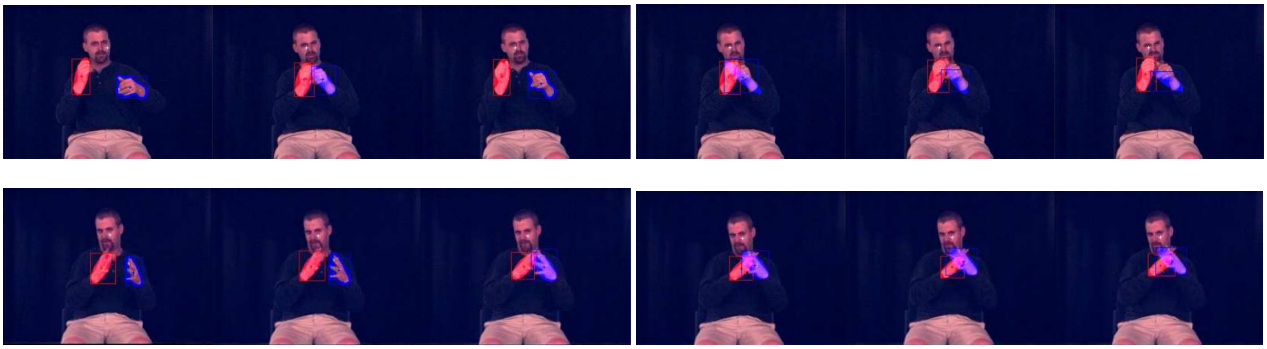

Fig. 5. Video frames (left to right), upper row: sign "WITH"; bottom row: gloss "FOOTBALL".

eling that the HMMs yield. This dynamic modeling is requisite for the modeling of movement, and has been employed successfully in numerous applications [3]. Afterwards, a model level approach gives a probabilistic viewpoint and fits well with the automatic recognition framework. We initialize the segments by first applying the pronunciation clustering and segmentation procedures, as described in the previous Sections 3.1,3.2. Since our goal is to model the dynamics of movement during the signs we employ instead of the position feature vector the position normalized by the initial position of the segment. In this way we explicitly force our models to be translation invariant. This additional characteristic, requires the application of one more normalization step, similarly, on the rest of the segments apart from all the first that have not been normalized, by subtracting their own initial position. Given the normalized to the initial position segments our goal is to cluster whole dynamic models (HMMs) [8]. Clustering states at the model level has been employed successfully in ASR applications [12]. Herein we cluster not just the states, but whole sequences of states.

Next, we do not use explicitely the GP clustering; its application is restricted to the segmentation step. Thus, we fit N 3-state HMMs, one for each individual sequence S_i, $i = 1 \dots N$. Then we use a similarity measure between pairs of HMM models $H_k, k = 1, 2$, by adopting among proposed approaches [13] that are based on the Kullback-Leibler divergence. Similarly we employ

$$D(H_1, H_2) = \sum_{O_i^{H_1}} \frac{1}{T_i} \log \frac{P(O_i^{H_1}|H_1, S_i^{H_1})}{P(O_i^{H_1}|H_2, S_i^{H_2})}$$

where $O_i^{H_k}$ corresponds to the observation sequences that have been generated from each H_k, of length T_i and $\log P(O_i^{H_k}|H_k, S_i^{H_k})$ to the log probability of the observation given the HMM model and the optimum state sequence $S_i^{H_k}$, for $k = 1, 2$. The distance similarity matrix among all models is exploited via an agglomerative hierarchical clustering algorithm. We end up with the total likelihood of the specific clustering, given the number of clusters employed. Up to now the number of clusters was assumed to be known. In order to select the number of centers we follow a Monte-Carlo cross-validation approach [8]: We partition the data into a fraction of 0.5 for testing and training and repeat 5 times the steps for clustering using only the training data for a range of numbers of centers. The 5 partitions are randomly chosen on each run. We evaluate each realization by the total log-likelihood of the models. Based on these average measurements we select the number of 15 clusters over different sign selections.

The existing automatic sub-unit modeling approaches exploit the data-driven characteristic at the feature level [5, 6, 7]. In [6] they employ segments that result from a segmentation step, instead of the individual frames. Next, they apply a hierarchical clustering algorithm by utilizing dynamic time wrapping. The employing of dynamic information in [6, 7] seems more appropriate for feature cues that evolve dynamically compared to a static clustering [5] that acts upon features in isolated time instances. In contrast to these approaches we propose the incorporation of the dynamics at the model level.

4. EVALUATION AND DISCUSSION

After the HMM hierarchical clustering we get clusters of models. By mapping back to the initial segments of features we show on Fig. 4 a few indicative clusters which contain the sequences of positions i.e., movements. We observe that the grouping of segments obeys loosely formed patterns. However there are clusters that contain only a few segments or outliers. For instance different models seem to

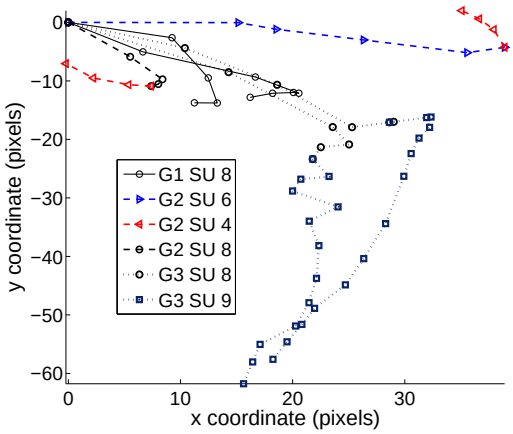

Fig. 6. Multiple realizations of three different signs, "FOOTBALL" (G1), "WITH" (G2), "HIT" (G3). Their segments share at least one sub-unit model. Line type identifies the sign. Markers and color indicate sub-units; see Fig. 5 for video frames.

map to a different type of movement pattern with respect to factors such as direction, scaling, and tracking: first figure shows straight movements from left to right and downwards, while 3rd figure shows curved movements pointing upwards and left.

In Fig. 2 we can see different realizations of the sign "HIT". The sub-unit sequences of the figure shows pairs of segments in sequence that combined create the whole signs. All instances shown, start with the 1st segment movement being from left to right; the 2nd segment movement evolves upwards to downwards. We observe that the first segments (continuous lines in the figure) for all realizations except from one which is clustered separately, are clustered in the same sub-unit; sub-units are identified by the marker type (and color). This is the case also for the second segment (dashed line) where all the segments are mapped into the same sub-unit cluster. A single instance among all instances in Fig. 2 is shown in Fig. 1. The movement pattern of the specific sign is also observed in the sequence of frames shown in the top row of the figure.

A complementary result is illustrated in Fig. 6 that is accompanied too by corresponding video frames for two instances of signs in Fig. 5. As it is shown, the signs are segmented and mapped on the sub-unit sequences as follows: the first instance of "FOOTBALL" (G1) is mapped on SU8, next, "WITH" (G2) is mapped on SU6+SU4 or SU8+SU4 and "HIT" (G3) on SU8+SU9. We see the advantage of sub-unit construction at the model level: different realizations of the same sign or of a completely different sign may share sub-units, i.e., they do not share just states of their HMMs but whole HMMs which contain sequences of states.

5. CONCLUSIONS

We investigate the issue of data-driven phonetic modeling of intra-gloss sub-units. We perform a pronunciation clustering step at the gloss level, followed by a model based segmentation that defines segments of movements. We cluster these segments by employing a clustering at the HMM level that is based on each model log-likelihood. We finally construct the model based sub-units. We evaluate qualitatively this modeling on real continuous sign language video from the BU400 corpus. The results highlight the benefits of the proposed approach. Moreover, in our ongoing research we deal with the recognition problem and the subunit approach yields promising results. Although we have applied the proposed framework by using movement only features, there is on-going work in 1) extending this approach by incorporating hand-shape and other informative cues of sign language and 2) taking advantage of the probabilistic character of the modeling in favor of other challenging issues in sign language modeling. Finally, it would be fruitful to employ phonetic information aiming at the deeper understanding of the mechanisms and phenomena involved.

Acknowledgements: We wish to thank Boston University and C. Neidle for providing the BU400 video database, A. Potamianos and C. Vogler for insightful discussions.

6. REFERENCES

[1] S. Ong and S. Ranganath, "Automatic sign language analysis: A survey and the future beyond lexical meaning," *Trans. Pattern Anal. Mach. Intellig.*, vol. 27, no. 6, pp. 873–891, 2005.

[2] K. Emmorey, *Language, cognition, and the brain: insights from sign language research*, Erlbaum, 2002.

[3] C. Vogler and D. Metaxas, "Handshapes and movements: Multiple-channel american sign language recognition," in *Gesture Workshop*, 2003, pp. 247–258.

[4] S. K. Liddell and R. E. Johnson, "American sign language: The phonological base," *Sign Language Studies*, vol. 64, pp. 195 – 277, 1989.

[5] B. Bauer and K-F. Kraiss, "Towards an automatic sign language recognition system using subunits," in *Int'l Gesture Workshop*, 2001, vol. 2298, pp. 64–75.

[6] G. Fang, X. Gao, W. Gao, and Y. Chen, "A novel approach to automatically extracting basic units from chinese sign language," in *Proc. ICPR*, USA, 2004, vol. 4, pp. 454–457.

[7] J. Han, G. Awad, and A. Sutherland, "Modelling and segmenting subunits for sign language recognition based on hand motion analysis," *Pat. Rec. Lett.*, vol. 30, no. 6, pp. 623–633, 2009.

[8] P. Smyth, "Clustering sequences with hidden markov models," in *In Advances in Neural Information Processing Systems*, 1997, vol. 9, pp. 648–654.

[9] P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and Ney H., "Benchmark databases for video-based automatic sign language recognition," in *Proc. International Conference on Language Resources and Evaluation (LREC)*, May 2008.

[10] G. Papandreou and P. Maragos, "Multigrid geometric active contour models," *IEEE Trans. on Image Process.*, vol. 16, no. 1, pp. 229–240, Jan. 2007.

[11] O. Diamanti and P. Maragos, "Geodesic active regions for segmentation and tracking of human gestures in sign language videos," in *Proc. ICIP*, 2008.

[12] V. Digalakis, P. Monaco, and H. Murveit, "Genones: generalized mixture tying in continuous hidden markovmodel-based speech recognizers," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 4, pp. 281–289, Jul 1996.

[13] B.-H. Juang and L. R. Rabiner, "A probabilistic distance for hidden markov models," *AT & T Technical Journal*, 1985.