Advances in Dynamic-Static Integration of Movement and Handshape Cues for Sign Language Recognition

Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos*

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

Abstract. We explore the integration of movement-position (MP) and handshape (HS) cues for sign language recognition. The proposed method combines the data-driven subunit (SU) modeling exploiting the dynamic-static notion for MP and the affine shape-appearance SUs for HS configurations. These aspects lead to the new dynamic-static integration of manual cues. This data-driven scheme takes advantage of the dynamic-static sequential SU modeling. Recognition evaluation on the continuous sign language corpus BU400 demonstrates promising results.

Keywords: automatic sign language recognition, movement-position, handshape, data-driven subunits, multiple streams, fusion, integration.

1 Introduction

One of the key points of difference of sign languages when compared with their spoken counterparts is the articulation of parallel information streams. The integration of multiple streams is a known but challenging issue within many fields and is still an open issue for automatic sign language recognition[1]. From the linguistic point of view there is an ongoing evolution of concepts regarding the relations of the multiple streams [10, 6]. In this work, building on single-cue subunits [8, 9] we tackle the integration of multiple manual cues, by taking advantage of the dynamic - static concept of movement-position cues.

Existing works rarely consider explicitly multiple cues integration, but rather include bundles of features. Another issue is the recognition level of the task since most studies address isolated signs and whole sign models [3, 2]. Some works study integration issues via specific modeling architectures. Indicative cases include [11] with parallel HMMs given manual transcriptions. Handshape, motion and place of articulation are combined in [4] in a tree-like structure for isolated sign recognition. Independent feature sets are employed in [7] to analyze inflections by modeling the systematic variations as parallel channels.

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

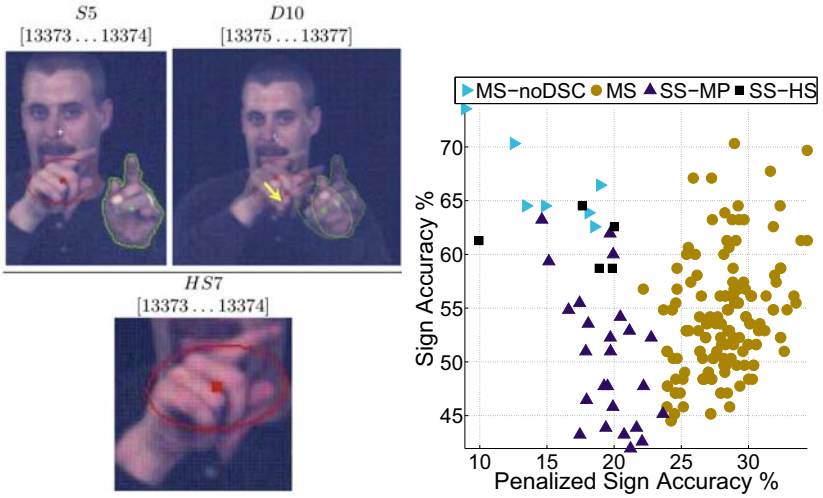
On previous work we employed subunits for the dynamic-static *movement-position* (MP) [8] and the affine shape-appearance models (Aff-SAM) for handshapes (HS) [9] single-cues separately. These aim at the automatic time segmentation and construction of data-driven subunits (SUs). For the *MP cues* the SU modeling consists of the partitioning of segments into dynamic or static with respect to their dynamics. Then, for each type of segment we employ a clustering approach that results on the construction of SU models. The latter gives birth to the lexicon construction that recomposes the dynamic-static SUs (DSSU) to form each sign. For the *HS cue* we are based on Aff-SAM and construct handshape SUs (HSSU) and the corresponding lexicon based on HSSU.

In this paper we introduce a method for the data-driven integration of manual cues that exploits the dynamic-static MP concept. Given the sequential structure of sign language [6] we explore an integration scheme of the manual cues that is driven by the static subunits (i.e. postures, detentions). Based on the single-cue subunit models, of both MP [8] and HS [9], we train fused static position and handshape models. For the statistical modeling we utilize multi-stream hidden Markov models (HMMs). The proposed framework is evaluated in recognition experiments on the continuous sign language corpus of Boston University (BU400) and provides promising results.

2 Dynamic-Static Integration of Manual Cues

Herein we exploit the *dynamic-static* nature of the movement-position subunits (SUs) by considering handshapes only during detentions and postures. These are assumed to correspond to segments classified as static. We proceed by assigning the handshape SUs separately on each static position SU. For the dynamic movement segments, we construct a handshape-during-movements SU model that collects all handshapes during movements. In this way we implicitly account for the non-dealt variation of the handshape during movements, for instance due to 2D instead of 3D processing. Given the sequential structure of dynamic-static segments this is shown to be a good compromise.

This scheme results on *fused subunit models* of Static-Positions with Handshapes (SP_i-HS_j), and Dynamic-Movements independent to the handshape (D_k-HS*) i, k and j correspond to the single-cue subunits of dynamic-movements (D), static-position (S) and handshapes (HS). After constructing the combined subunits we statistically train them in the HMM framework: employing a multi-stream GMM and HMM for the S+HS models (S_i-HS_j) and the D models (D_k-HS*) respectively. The automated data-driven analysis ends up on a lexicon that decomposes the signs wrt. to the movement-position and handshape SUs. An indicative combination of MP and HS SUs is illustrated in Fig. 1(a). Sign BUT consists of a static-position SU (S5), followed by a dynamic-movement SU (D10); the handshape cue remains static over the sign (HS7). Therefore in the fused MP+HS lexicon we end-up with a static-position+handshape SU (S5-HS7) and a dynamic-movement SU (D10-HS*).



(a) (b)

Fig. 1. (a)Decomposition of BUT into Movement-Position+Handshape SUs. (b)Single-stream (SS) and multi-stream (MS) recognition. Markers map to results depending on the number of Movement-Position (MP) and Handshape (HS) SUs MS refers to the proposed Dynamic-Static integration approach. MS-noDSC employs MP+HS cues without the Dynamic-Static concept. Axes refer to sign Generalization and Discrimination.

3 Recognition Experiments

The experiments are conducted on the Boston-University continuous American Sign Language corpus (BU400) [5][1]. The difficulty of the task is increased by not considering the test sign realizations for the lexicon construction. For the evaluation of the performance we employ the Sign Accuracy SignAcc $= \frac{N-S}{N}$ 100%: N is the number of examples and S refers to substitution errors. Because of the single subunit sequence mapping to multiple signs [8, 9] the $SignAcc$ considers a sign as correct if it is in the set of signs related to this subunit sequence; this holds *even* if other signs are present in the set. We introduce the Penalized Sign Accuracy Low Bound (PenAcc) that accounts for the above effect: for each i example classified as correct for a set of N test examples we increase the PenAcc by $\text{PenAcc}_i = \frac{1}{BF \cdot N}$ where BF is the cardinality of the set of signs that the specific subunit sequence is mapped to. With the standard measure the increase in the accuracy percentage would instead be $Acc_i = \frac{1}{N}$.

Figure 1(b) illustrates SignAcc (y-axis) and PenAcc (x-axis) for the movement-position (SS-MP) and handshape (SS-HS) single streams. SS-MP and SS-HS result on average on 50% and 60% SignAcc. However, their discriminability is

[1] We employ the stories narrated from a single signer: `accident`, `biker_buddy`, `boston_la`, `football`, `lapd_story` and `siblings`; 50 glosses are randomly selected among the most frequent; 75% of the data are employed for training.

quite low: the PenAcc is 20% and 15% on average. This is expected as the single-cue modeling cannot discriminate e.g., signs sharing a movement-position with different handshape. With the proposed dynamic-static integration scheme (MS) the accuracy measure shows an increase or not depending on the parameters and approach employed. Nevertheless, the discrimination is increased. The combined measures show that the proposed scheme both generalizes and discriminates among more signs. We also compare with a competing approach (MS-noDSC) that does not employ the dynamic-static concept while sharing the same parameters and time segmentation. MS-noDSC results on 65% on average in sign accuracy however the penalized sign accuracy is 15%. Therefore the proposed approach (MS) leads to better performance, as shown by the marker points concentrated in the upper right corner of the evaluation figure.

4 Conclusions

We explore integration schemes for the movement-position and handshape cues. These are based on data-driven subunits while exploiting the dynamic-static notion and the sequential structure of sign languages. Experiments on the BU400 show promising results, when evaluating wrt. both the generalization and the discrimination ability of the proposed approach.

References

1. Agris, U., Zieren, J., Canzler, U., Bauer, B., Kraiss, K.F.: Recent developments in visual sign language recognition. Universal Access in the Information Society 6, 323–362 (2008)
2. Bowden, R., Windridge, D., Kadir, T., Zisserman, A., Brady, M.: A linguistic feature vector for the visual interpretation of sign language. In: ECCV (2004)
3. Buehler, P., Everingham, M., Zisserman, A.: Learning sign language by watching TV (using weakly aligned subtitles). In: CVPR. pp. 2961–2968 (June 2009)
4. Ding, L., Martinez, A.M.: Modelling and recognition of the linguistic components in american sign language. Im. and Vis. Comp. 27(12), 1826 – 1844 (2009)
5. Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., Ney, H.: Benchmark databases for video-based automatic sign language recognition. In: Proc. LREC (May 2008)
6. Liddell, S.K., Johnson, R.E.: American sign language: The phonological base. Sign Language Studies 64, 195 – 277 (1989)
7. Ong, S.C.W., Ranganath, S.: A new probabilistic model for recognizing signs with systematic modulations. In: amfg. pp. 16–30 (2007)
8. Pitsikalis, V., Theodorakis, S., Maragos, P.: Data-driven sub-units and modeling structure for continuous sign language recognition with multiple cues,. In: Proc. LREC Workshop: Corpora and Sign Language Technologies (May 2010)
9. Roussos, A., Theodorakis, S., Pitsikalis, V., Maragos, P.: Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. In: Workshop on Sign, Gesture and Activity (SGA), ECCV-2010 (Sep 2010)
10. Stokoe, W.: Sign language structure. Annual Review of Anthropology 9(1), 365–390 (1980)
11. Vogler, C., Metaxas, D.: Handshapes and movements: Multiple-channel american sign language recognition. In: Gesture Workshop. pp. 247–258 (2003)