IEEE CVPR Workshop on Gesture Recognition, Colorado Springs, USA, June 20, 2011

Advances in Phonetics-based Sub-Unit Modeling for Transcription Alignment and Sign Language Recognition

Vassilis Pitsikalis and Stavros Theodorakis
School of Electrical and Computer Engineering
National Technical University of Athens
{vpitsik,sth}@cs.ntua.gr

Christian Vogler
Institute for Language and Speech Processing
Athena R.C.
cvogler@ilsp.athena-innovation.gr

Petros Maragos
School of Electrical and Computer Engineering
National Technical University of Athens
maragos@cs.ntua.gr

Abstract

We explore novel directions for incorporating phonetic transcriptions into sub-unit based statistical models for sign language recognition. First, we employ a new symbolic processing approach for converting sign language annotations, based on HamNoSys symbols, into structured sequences of labels according to the Posture-Detention-Transition-Steady Shift phonetic model. Next, we exploit these labels, and their correspondence with visual features to construct phonetics-based statistical sub-unit models. We also align these sequences, via the statistical sub-unit construction and decoding, to the visual data to extract time boundary information that they would lack otherwise. The resulting phonetic sub-units offer new perspectives for sign language analysis, phonetic modeling, and automatic recognition. We evaluate this approach via sign language recognition experiments on an extended Lemmas Corpus of Greek Sign Language, which results not only in improved performance compared to pure data-driven approaches, but also in meaningful phonetic sub-unit models that can be further exploited in interdisciplinary sign language analysis.

1. Introduction

Phonetic transcriptions are crucial for the performance of sign language (SL) and speech recognition systems. For the recognition of SL, which is the primary means of communication for many deaf people, this has not been practical, due to the huge level of effort required for creating detailed phonetic annotations, unlike the case of speech recognition. Another problem is the lack of appropriate phonetic models in the area of SL linguistics (although this is changing now). Thus, data-driven methods have prevailed in recent years.

We propose a novel approach to address these issues. It is based on two aspects: (1) converting SL annotations into structured sequential phonetic labels, and (2) incorporating these labels into a sub-unit-based statistical framework for training, alignment, and recognition. This framework can be applied similarly to arbitrary gesture data.

Recent successful data-driven methods include [1, 4, 2, 5, 3, 12, 8]. One employs a linguistic feature vector based on measured visual features, such as relative hand movements [2]. Another one clusters independent frames via K-means, and produces “phenones” [1]. Instead of single frames, [4, 5, 12] cluster sequences of frames on the feature level, such that they exploit the dynamics inherent to sign language. Recently, separate features and modeling for dynamic vs. static segments have been proposed [8].

These data-driven approaches allow adapting recognition systems to the concrete feature space, and work even in the face of insufficient detailed transcriptions. As mentioned before, creating such transcriptions requires an impractical amount of effort, unlike phoneme-level transcriptions for speech recognition. Yet, their value is clear: they simplify adding new words to the lexicon, and allow capturing commonalities across signs. They can also be used to create meaningful representations of intra-sign segments, for further linguistic or interdisciplinary processing.

Our approach is based on having annotations in HamNoSys [9], the creation of which requires less effort than full phonetic descriptions, and incorporating them into a statistical recognition system. This is conceptually similar to taking a written word and converting it into its pronunciation in speech recognition, and has hitherto not been possible for SL recognition. Our first contribution is that we have developed a parsing system for converting HamNoSys into structured phonetic sequences of labels, according to the Posture-Detention-Transition-Steady Shift (PDTS) system [6]. However, they do not provide any timing information, which leads us to the second contribution: We employ simple visual tracking features extracted from sign language videos. Using them in conjunction with the phonetic labels, we construct sub-units via a statistical hidden Markov model (HMM)-based system, which allows us to align the PDTS sequences with the visual data segments. The resulting output consists of sub-units that are no longer purely data-driven, in contrast to previous work. Rather, they are phonetic sub-units, each of which corresponds to a meaningful PDTS label, along with the timing information on where they occur in the data.

Once the segments have been mapped to their PDTS labels, the output of the recognition system produces phonetic labels during decoding. Such labels are invaluable in interdisciplinary research tasks, such as linguistic analysis and synthesis. We evaluate the proposed approach by performing recognition experiments on a new corpus of 1000 Greek Sign Language lemmata, with promising results.

2. Data, Visual Processing and Overview

Data: The *Greek Sign Language (GSL) Lemmas Corpus* consists of 1046 isolated signs, 5 repetitions each, from two native signers (male and female). The videos have a uniform background and a resolution of 1440x1080 pixels, recorded at 25 frames per second interlaced.

Visual Processing: For the segmentation and detection of the signer's hands and head in the Greek Sign Language (GSL) Lemmas Corpus, we employed a skin color model utilizing a Gaussian Markov Model (GMM), accompanied by morphological processing to enhance skin detection. Moreover, for tracking we employed forward-backward linear prediction, and template matching, in order to disambiguate occlusions. The adopted approach is described in [10]. The extracted feature vector has five components, and consists of the planar coordinates of the dominant hand, the instantaneous direction, and the velocity.

Overview: In the following, we adopt the Greek signs for PILE, IMMEDIATELY, and EUROPE as examples from the corpus. Figure 1 shows the initial and end frames of each sign superimposed. The arrows illustrate the movements of the hands between the frames. In the next sections we present details on the articulation of these signs via representative examples alongside the contributions.

3. Data-Driven Sub-Units without Phonetic Evidence for Recognition

Our data-driven approach is based on the work in [8]. Other previous approaches include [1, 4, 5]. We segment signs automatically and construct data-driven sub-units, which are the primitive segments that are used to construct all signs that share similar articulation parameters. Based on simple movement-related measurements for the dominant hand, the first step for sub-unit construction involves the unsupervised partitioning of the segments into two groups with respect to their movement dynamics — for each sign unit, a model-based process finds the segmentation points and assigns them the label "static" or "dynamic."

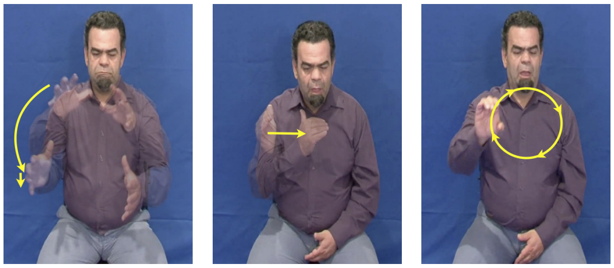

(a) PILE (b) IMMEDIATELY (c) EUROPE

Figure 1. Overview of articulation for three selected GSL signs.

For the second step, the sub-unit construction (i.e., the statistical modeling and the features employed for the static or dynamic segments) depends on the assigned label: For static segments, we employ K-means for clustering based on their position. For dynamic segments, we employ hierarchical clustering based on their DTW distances wrt. the instantaneous direction. Thus, after clustering we end up with a lexicon, where each sign consists of a sequence of dynamic and static sub-units. The characteristics of the approach above imply a sequential structure of dynamic and static segments that are explicitly accounted for by the proposed sub-unit construction and statistical modeling.

4. Conversion of Annotations to Phonetic Transcriptions

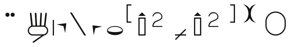
There has been little progress in the area of phonetic modeling for the purposes of SL recognition since the work of Vogler and Metaxas [11]. It is possible that the lack of widely available phonetic transcriptions in sign language corpora has contributed to this state of affairs. Because of the level of detail required, such transcriptions are time-consuming to produce and involve a steep learning curve.

In this paper, we propose a different approach that consists of generating annotations that are merely detailed enough to reproduce the sign, and having the computer convert these to the full phonetic structure. This approach has the advantage that it takes far less time and human training to produce the annotations. A disadvantage, however, is that such annotations make assumptions that require complex inferences by the conversion code. Describing such inferences in detail is beyond the scope of this paper; in the following we give a general overview of the method.

Like in the work by Vogler and Metaxas, the basic phonetic structure of a sign is a sequence of segments, which we model according to Johnson's and Liddell's recent work on the Posture-Detention-Transition-Steady Shift (PDTS) system [6]. It supersedes the older Movement-Hold model [7] used in earlier work, and fixes many of its shortcomings[1].

In this system, each sign can be considered as a sequence of key points in the form of postures (P), with associated hand configuration and location information. Transitions (T) correspond to hand movements between the key points, with attached trajectory information. Detentions (D) are like P, but the hand is held stationary; steady shifts are like T, but with a slow, deliberate movement; in this paper we distinguish only among P, D and T. In addition, we consider epenthesis movements (E) [7] to be distinct from T; the former are transitions between two locations without an explicit path, and primarily occur when the hands move into position between signs, and during repeated movements. An example of the basic structure of the sign for PILE — E P T P T P E — is shown in Fig. 2, and Table 1.

The annotations of the signs are coded in HamNoSys [9], a symbolic annotation system that can describe a sign in sufficient detail to display it in an animated avatar. It models signs as clusters of handshape, orientation, location, and movement, without explicit segmentation information, which makes it unsuitable for direct application to recognition systems. HamNoSys's philosophy is minimalist, in the sense that it avoids redundancy and strives to describe a sign in detail with as few symbols as possible. To this end, it provides symmetry and repetition operators, and describes only how a sign's configuration changes over time. As an example consider the first part of the sign for PILE:



This annotation says that the hands move symmetrically, so it needs to provide only the hand configuration and location for the right hand, and the fact that the fingers of both hands touch each other. In contrast, the left hand's information (mirrored along the x axis) is implied.

In order to model signs properly in the recognition system, we require that all information, according to the PDTS system, is made explicit for every segment; that is, Ps and Ds contain the full information on hand configuration and location, and Ts contain the full information on movement trajectories, for each hand respectively. Our conversion method from HamNoSys to the PDTS structure resolves the implied parts, and splits the signs into its constituent segments. The key step consists of accumulating deltas, which describe how a posture or transition has changed with respect to a prototype. These are then applied in a specific order. Note that this process also works for independent channels of information, such as hand configuration versus location, dominant hand versus nondominant hand, and so on, and provides relative timings of segments across channels; however, the details are beyond the scope of this paper.

Table 1. Phonetic PDTS labels of the corresponding sub-units for the sign "PILE" (location and trajectories only).

Frames	Type	PDTS label
1:12	E	rest-position — location-head
13:13	P	location-head
14:25	T	directedmotion, curve-r, direction-o, second-direction-do, tense-true
26:27	P	location-torso, side=right_beside
28:50	T	directedmotion, direction-dr, small
51:51	P	location-torso, side=right_beside_down
52:66	E	location-torso, side=right_beside_down — rest-position

Further examples of PDTS sequences can be found in Tables 1, 2. The details of the conversion are beyond the scope of this paper, due to space limitations, and will be published separately.

5. Phonetic Based Sub-units, Training, Alignment and Recognition

In the previous section we have covered our first main contribution. Our second main contribution consists of incorporating the phonetic labels into a statistical recognition system. The data-driven-only sub-units from Section 3, without any phonetic information, adapt well to specific feature spaces. However, they produce meaningless sub-unit labels, which cannot be exploited for interdisciplinary sign language processing (e.g., synthesis, linguistics).

We call the process of incorporating the phonetic information "Phonetic Sub-unit Construction for Recognition." This is the first time that the following are taken into account in an automatic, statistical, and systematic way: (1) *phonetic transcriptions* of SL, provided as described in the previous section by the PDTS system, and (2) the corresponding underlying *visual data and features* from processing the video data and the feature extraction. The procedures involved in this process involve: (1) phonetic sub-unit construction and training, (2) phonetic label alignment and segmentation, (3) lexicon construction, and (4) recognition.

5.1. Phonetic Sub-Unit Model Training

For each phonetic label provided by the PDTS system, and the features from the visual processing, we train one sub-unit HMM. These sub-units have both phonetic labels from the PDTS structure, and statistical parameters stem-

[1] Specifically, movements no longer have attached location information, which previously had prevented a direct adaptation to recognition systems. In addition, there is a strict alternation of P/D with T/S, whereas the older model could have sequences of movements without intervening holds.

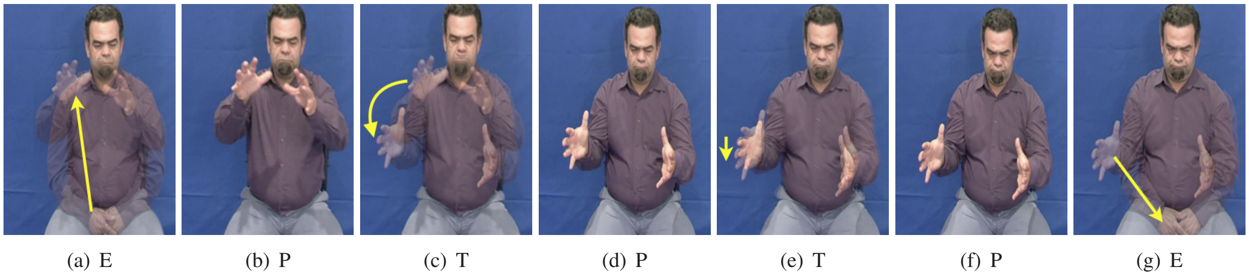


Figure 2. Sign for PILE: Segments after incorporation of PDTS phonetic labels into Phonetic Sub-unit Construction, Training and Alignment. Superimposed start and end frames of each sequence of segments, accompanied with an arrow for transitions and epenthesis. Each segment corresponds to a single phonetic label. PDTS segments labels are of type Epenthesis (E), Posture (P), Transition (T)

ming from the data-driven models, as a result of the training step. An example is illustrated in Table 1, which lists the sequence of phonetic labels for sign for "PILE".

We use different HMM parameters for each type of subunit. Distinguishing between movements (T/E) and postures/detentions (P/D) corresponds to making a distinction between dynamic and static segments, as described in Section 3. This also is consistent with the concepts in the old Movement-Hold model [7]. For T and E, we employ a 6-state and 3-state Bakis HMM topology, respectively. For P and D, we use a 1-state HMM, and a 2-state left-right HMM, respectively. One mixture and a diagonal covariance matrix was employed for each HMM. We initialize the phonetic sub-unit models in a uniform way with a flat-start procedure using the global mean and covariance of the feature space, and employ embedded training on strings of concatenated sub-unit models with unsegmented data.

5.2. Alignment and Time Segmentation

We concatenate the trained HMMs into a recognition network and decode each feature sequence via the Viterbi algorithm. This results in a sequence of phonetic PDTS labels, together with their respective starting and ending frames. Doing this for all sequences results in a lexicon with segmentation boundaries for each PDTS label.

We recognize signs by decoding unseen test data in the HMM network on the PDTS label level. We evaluate the accuracy on the sign level, based on the lexicon above.

Fig. 2 shows an example of the segmentation acquired during the decoding, which illustrates the sequence of phonetic sub-units for the above-mentioned sign for "PILE". Each image corresponds to a phonetic PDTS segment produced by the decoding. For visualization, we adopt the following conventions: (1) For T and E segments, we superimpose their respective initial and final frames. We also highlight specific movement trajectories with an arrow from the initial to the final hand position in the respective segment. (2) For P and D segments, we show only the first frame of the segment, as the hand does not move within them. In addition, the labels corresponding to this sign, along with the segmentation boundaries, are listed in Table 1.

5.3. Phonetic Sub-Units Results

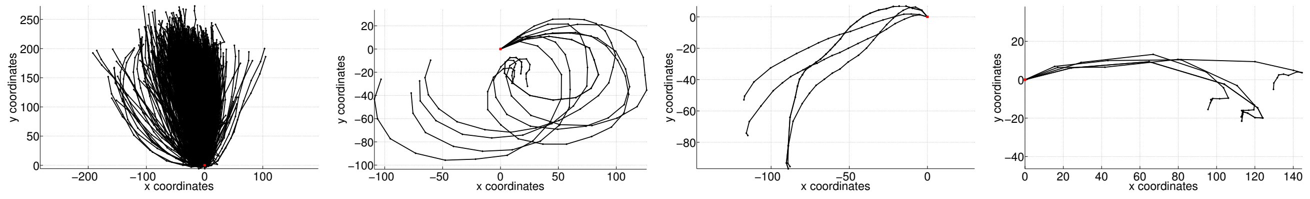
Fig. 3 and 4 show examples of movement-based subunits (T and E), using x and y coordinates mapped from the signing space. For the corresponding phonetic labels see Table 2. Fig. 3(a) shows a common epenthesis sub-unit (E-to-head). It models the movement from the rest position to the head, a common starting posture. Fig. 3(b) corresponds to a circular transition sub-unit (T-circular). An indicative sign that contains this sub-unit is "EUROPE" (see Fig. 1(c)). Fig. 3(c) and 3(d) depict directed transition subunits (T-down-right, T-in-left) with right-down and left directions respectively. Representative signs are "PILE" and "IMMEDIATELY," respectively (see Fig. 1(a), 1(b)).

In Fig. 4 we show results for the P and D sub-units, with the actual coordinates for four different postures superimposed in different colors. P-head, P-stomach, P-shoulder and P-head-top correspond to locations at the signer's head, stomach, shoulder and top of head, respectively.

In all these figures, there are cases of compact phonetic sub-units with less variance, of sparsely populated ones (i.e., few available data), and some that contain outliers. For instance, the sub-unit P-head-top is compact, but has few data. In contrast, P-head has more data and increased variance. The sub-unit for the initial transition from the rest posture to the starting position occurs in many signs, whereas other sub-units may occur in only a single sign. Outliers and high variances seem to be caused by visual processing inaccuracies (we perform 2D, rather than 3D, processing), tracking or parameter estimation errors, or human annotator errors, or actual data exhibiting such properties.

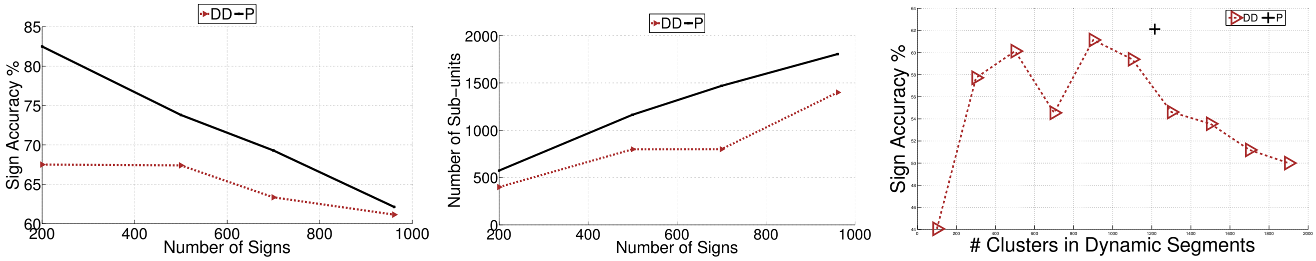
6. Sign Language Recognition Experiments

The recognition task in this paper was conducted on one signer and 961 out of the 1046 signs. Approximately half of the missing 85 signs share the same pronunciation with another sign, and thus are the same for recognition purposes, while the other half were eliminated due to unacceptably poor tracking or poor segmentation of the five repetitions



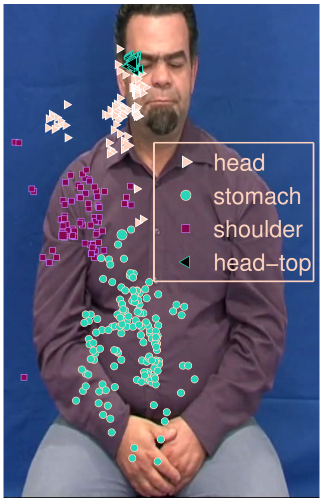
(a) E-to-head (b) T-circular (c) T-down-right (d) T-in-left

Figure 3. Sub-units after Phonetic Sub-unit Construction, Training and Alignment. (a) corresponds to an epenthesis sub-unit (E-to-head) and (b-d) to transition sub-units (T-circular, T-down-right, T-in-left). Trajectories are illustrated in the real signing space normalized wrt. their initial position (x,y) = (0,0). Red marker indicates trajectories' start position. See Table 2 for the corresponding phonetic labels.



(a) (b) (c)

Figure 5. Comparison of Data-Driven (DD) Sub-units without phonetic evidence vs. Phonetic based approach (P). (a) Sign Accuracy, (b) Number of sub-units. In (a,b) x-axis corresponds to the variation on the number of signs. (c) For the maximum number of signs, sign accuracy as affected by the number of sub-units (x-axis) in the DD case; in the Phonetic approach number of sub-units is predefined.



(a)

Figure 4. Sub-units after Phonetic Sub-unit Construction, Training and Alignment. Data for multiple posture phonetic sub-units superimposed in the signing space indicating their relative position to the signer. Sub-units with multiple colored pixels are: P-forehead, P-stomach, P-shoulder, P-head-top. Legend shows the primary locations of the corresponding phonetic labels (see also Table 2).

Table 2. Examples of phonetic subunit (PSU) and the sign where they occur. '*' correspond to multiple signs.

PSU	Sign	Type	PDTS Label
E-to-head	*	E	rest-position — location-head
T-circular	EUROPE	T	circularmotion, axis=i
T-down-right	PILE	T	directedmotion, direction=dr, small
T-in-left	IMMEDIATELY	T	directedmotion, direction=il, fast=true, halt=true
P-forehead	*	P	location=forehead
P-stomach	*	P	location=stomach
P-shoulder	*	P	location=shouldertop, side=right_beside
P-head-top	*	P	location=head-top

into individual signs. The data were split randomly into four training examples and one testing example per sign, which was the same across all experiments. Future work should expand these experiments to both signers and the full set, as more tracking results come in and improve. The visual processing and feature extraction was conducted as described in Section 2. The modeling and recognition proceeded, as described in the previous section. Our evaluation criterion was the number of correctly recognized signs, via matching sequences of phonetic labels to the lexicon.

We first compare the two approaches for sub-unit construction, as follows: (1) *Data-Driven (DD):* Data-driven sub-unit construction, which does not make use of any phonetic transcription labels. (2) *Phonetic (P):* Phonetics-based approach which makes use of the PDTS phonetic labels, via the statistically trained sub-unit models.

Second, we evaluate the relationship between lexicon size and recognition accuracy. In addition, for the DD approach, we evaluate the number of sub-units employed for static and dynamic models against accuracy. The number of phonetic sub-units is determined by the PDTS labels.

In Fig. 5(a) we compare the DD sub-unit (employing the number of SUs which corresponds to the best performance) with the P sub-unit approach, wrt. the lexicon size. By increasing the number of signs, the recognition performance for both approaches decreases; this is expected as the recognition task becomes harder. The phonetics-based approach outperforms the data-driven one across all experiments. Nevertheless, we observe that with an increasing number of signs, the recognition performance of the phonetic approach deteriorates more than the one of the data-driven approach. This is also expected, because more signs mean more PDTS labels — and a high number of labels makes the task harder (Fig. 5(b)). In contrast, with the data-driven approach without any phonetic labels, as the number of signs increases, more data are accumulated in the feature space, which is partitioned via the clustering methods afterward. Another aspect is presented in Figure 5(c), which shows the results of the phonetic approach with a varying number of data-driven sub-units. The number of static units was held constant at 500, while the number of dynamic ones varied. This experiment focuses on the details of the experiment presented in Fig. 5(a) with the 961 signs.

7. Conclusions

We have presented work on novel directions in gesture analysis and recognition, with applications in sign language. It explores new ways for the incorporation of linguistic evidence, in the form of sequences of phonetic labels, which are extracted from sign language annotations. This incorporation is based on the visual evidence from simple features, such as the tracking of the dominant hand. We construct phonetic sub-units that carry phonetic information, unlike previous data-driven approaches. The phonetic modeling also gives us the time alignment information that the phonetic labels initially lack. Finally, the decoded sequence during recognition consists of meaningful phonetic labels. Results on a GSL Lemmas corpus are promising, leading to at least 2% improvements compared to the data-driven approach on a set of close to 1000 lemmas, and 7% on average across all experiments.

Thus, phonetic modeling of signs is beneficial for computational modeling and sub-unit construction, and for capturing the relationships across different cues and modalities, irrespective of how much better it performs than data-driven approaches. The latter, in particular, cannot link phonetic transcriptions, and the corresponding statistical and time boundaries, which is important for linguistic work, as well as adding unknown signs to the lexicon, similar to the way speech recognition systems can add unknown words. Further exploitation of these results will also provide assistance to annotators. So far, we have addressed only movements and postures; however the concepts can be extended to other cues, such as the handshape. Finally, we expect that other disciplines, such as linguistics, can greatly benefit from our results for the analysis of sign languages.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135. Theodore Goulas contributed the HamNoSys annotations for GSL.

References

[1] B. Bauer and K. F. Kraiss. Towards an automatic sign language recognition system using subunits. In *Proc. of Int'l Gesture Workshop*, volume 2298, pages 64–75, 2001. 1, 2

[2] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A linguistic feature vector for the visual interpretation of sign language. In *ECCV*, 2004. 1

[3] L. Ding and A. M. Martinez. Modelling and recognition of the linguistic components in american sign language. *Im. and Vis. Comp.*, 27(12):1826 – 1844, 2009. 1

[4] G. Fang, X. Gao, W. Gao, and Y. Chen. A novel approach to automatically extracting basic units from chinese sign language. In *icpr*, 2004. 1, 2

[5] J. Han, G. Awad, and A. Sutherland. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Pat. Rec. Lett.*, 30(6):623–633, 2009. 1, 2

[6] R. E. Johnson and S. K. Liddell. A segmental framework for representing signs phonetically. sign language studies. 11(3), 2011. 2, 3

[7] S. K. Liddell and R. E. Johnson. American sign language: The phonological base. *Sign Language Studies*, 64:195 – 277, 1989. 3, 4

[8] V. Pitsikalis, S. Theodorakis, and P. Maragos. Data-driven sub-units and modeling structure for continuous sign language recognition with multiple cues. In *LREC Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 2010. 1, 2

[9] S. Prillwitz, R. Leven, H. Zienert, R. Zienert, T. Hanke, and J. Henning. HamNoSys. Version 2.0. *Int'l Studies on Sign Language and Communication of the Deaf*, 1989. 1, 3

[10] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos. Hand tracking and affine shape-appearance handshape subunits in continuous sign language recognition. In *Workshop on Sign, Gesture and Activity (SGA), ECCV*, Sep. 2010. 2

[11] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *CVIU*, 81(3):358–384, 2001. 2

[12] P. Yin, T. Starner, H. Hamilton, I. Essa, and J. Rehg. Learning the basic units in american sign language using discriminative segmental feature selection. In *Proc. ICASSP*, pages 4757–4760, 2009. 1