Image and Vision Computing 32 (2014) 533–549

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition ☆

Stavros Theodorakis *, Vassilis Pitsikalis, Petros Maragos

School of Electrical and Computer Engineering, National Technical University of Athens, Greece

ARTICLE INFO

Article history:
Received 22 May 2013
Received in revised form 27 March 2014
Accepted 30 April 2014
Available online 9 May 2014

Keywords:
Automatic sign language recognition
Data-driven subunits
Sub-sign phonetic modeling
Unsupervised
Segmentation
HMM

ABSTRACT

We introduce a new computational phonetic modeling framework for sign language (SL) recognition. This is based on dynamic–static statistical subunits and provides sequentiality in an unsupervised manner, without prior linguistic information. Subunit "sequentiality" refers to the decomposition of signs into two types of parts, varying and non-varying, that are sequentially stacked across time. Our approach is inspired by the Movement–Hold SL linguistic model that refers to such sequences. First, we segment signs into intra-sign primitives, and classify each segment as dynamic or static, i.e., movements and non-movements. These segments are then clustered appropriately to construct a set of dynamic and static subunits. The dynamic/static discrimination allows us employing different visual features for clustering the dynamic or static segments. Sequences of the generated subunits are used as sign pronunciations in a data-driven lexicon. Based on this lexicon and the corresponding segmentation, each subunit is statistically represented and trained on multimodal sign data as a hidden Markov model. In the proposed approach, dynamic/static sequentiality is incorporated in an unsupervised manner. Further, handshape information is integrated in a parallel hidden Markov modeling scheme. The novel sign language modeling scheme is evaluated in recognition experiments on data from three corpora and two sign languages: Boston University American SL which is employed pre-segmented at the sign-level, Greek SL Lemmas, and American SL Large Vocabulary Dictionary, including both signer dependent and unseen signers' testing. Results show consistent improvements when compared with other approaches, demonstrating the importance of dynamic/static structure in sub-sign phonetic modeling.

1. Introduction

Sign languages are natural languages that manifest themselves via the visual modality in the 3D space. They convey information via visual patterns and serve for communication in parts of Deaf communities [2]. Visual patterns are formed by manual and non-manual cues. The automatic processing of such visual patterns for the Automatic Sign Language Recognition (ASLR) can bridge the communication gap between the deaf and the hearing. Since the early work of [3], there has been progress in visual processing, sign language phonetic modeling, and automatic recognition [1,4,5]. Moreover ASLR may contribute to other disciplines such as linguistics for the study of Sign Languages (SLs), via automated processing of corpora, whereas it is broadly related to human computer interaction.

Herein we focus on sign language articulation produced by manual cues. The term "manual cues" refers to the movements and handshapes of both hands, one of which is considered as dominant. The dominant hand articulates the main phonetic parts. The other hand is referred to as non-dominant (ND). The ND hand contributes to symmetric/anti-symmetric movements or as a Place-of-Articulation (PoA). By PoA we refer to the location of the dominant hand in relation to either the body or the non-dominant hand. When the ND hand contributes in sign articulation, it is called active. Handshape, the form of the hand, equally plays a central role.

A coarse correspondence of a "word" in spoken language is a "sign" in SL. The phonemes constituting a spoken word are concatenated sequentially across time as the English word "admit" is phonetically transcribed as [ədm'ɪt]. As discussed next, signs make use of both simultaneous [2] and sequential phonetic structure [6]. Signs tend to be monosyllabic [7]. Due to the larger articulators, for instance the hands versus the tongue, this sequential compositionality is transformed into simultaneity via multiple cues accommodating similar amounts of information in the spoken or signed propositions respectively [8]. Take for instance the signs in Fig. 1: articulation parameters such as type of movement, handshape, as well as facial cues may vary in parallel. Yet there are studies on the sequential structure of SL [9], as the seminal work of Liddell and Johnson (L&J) [6]. Varying and non-varying

☆ This paper has been recommended for acceptance by Vassilis Athitsos.

* Corresponding author at: Zografou Campus, Athens 15773, Greece.
E-mail addresses: sth@cs.ntua.gr (S. Theodorakis), vpitsik@cs.ntua.gr (V. Pitsikalis), maragos@cs.ntua.gr (P. Maragos).

http://dx.doi.org/10.1016/j.imavis.2014.04.012
0262-8856/

Fig. 1. ASL signs from the (a, b) BU400 and (c, d) ASLLVD. (e–h) GSL signs from the GSL-Lem. Signs are formed by movements, non-movements (postures), handshapes and non-manual cues. A dominant hand constructs the main phonetic parts (b, c, e, f). The non-dominant (ND) hand contributes in symmetric/anti-symmetric movements (a, d, g, h), or as a Place-of-Articulation (PoA). By PoA we refer to the place the dominant hand is located in relation to either the body or the non-dominant hand: e.g. neutral space (a, c, e, f), eye (e), mouth (f). Handshape, the form of the hand, equally plays a central role. For further information see [1] and references therein.

phonetic parts are sequentially stacked across time. We link the terms "varying/non-varying" to the cases of movements/non-movements respectively; for a familiar example refer to the corresponding, in a broad sense, vowel-consonant case in speech. Take for instance the Greek Sign Language (GSL) sign "SAY" in Fig. 1f. This sign is articulated employing the dominant hand. It consists of two different positions, at the mouth, and in the neutral space, before and after the downward movement. These three movements/non-movements parts are stacked *in sequence*, as "position-mouth", "downward movement", and "position-neutral space." Thus, we conclude that the concept of the "phoneme" in SL is not to be taken for granted as in speech. There is still work in this direction both in the linguistic community [2,6,10], as well as from practical viewpoints such as computational recognition [11–13,1].

In this context, the phonetic modeling for automatic recognition is challenging. First, as other authors mention too [14,13], there is a lack of formal dictionaries with sub-sign phonetic transcriptions, based on well-defined phone inventories and on a standard notation system. In automatic speech recognition (ASR), such resources are easily accessible, standard for several spoken languages, and reusable among research teams. For sign languages, the cases that employ sub-sign phonetic level dictionaries are as follows: On the one hand, data-driven approaches define a set of basic units computationally without the need of manual annotation; indicative examples include [11,14, 13]. On the other hand, formally defined dictionaries are based on linguistic models such as the Movement–Hold [6], and sign notation systems such as the Stokoe system [2], the Hamburg notation system (HamNoSys) [15], or SignWriting [16]. These dictionaries are constructed by manual phonetic annotation which is time-consuming as in [12], by linguistic dictionary compilation [17], or recently via automatic processing as in [18,19]. In between, one finds approaches [20–23] that incorporate linguistic–phonetic concepts, ranging from Stokoe-driven decomposition to syllable phonetics. However, they do not lead to broadly reusable sub-sign transcriptions according to some known notation system or linguistic model [2,15,16,6]. Finally, there is a lack of phonetically transcribed data since annotation at the phonetic level is highly time consuming. Meanwhile, new SL corpora are being built [24–26], increasing the need for automatic processing. All the above render research in phonetic modeling for ASLR challenging.

This paper introduces a novel SL phonetic modeling approach for unsupervised dynamic–static sequentiality with statistical subunits (2-S-U). By "dynamic–static subunit sequentiality" we refer to the sequential stacking of dynamic and static subunits across time. This approach provides by construction both sequential and simultaneous phonetic structure. This is accomplished without any linguistic prior. A valuable result of the above is the construction of an unsupervised data-driven subunit-level lexicon that shares the aforementioned properties. The 2-S-U approach includes first the unsupervised model-based sign segmentation and classification into dynamic and static segments, i.e., movements and non-movements, and then the construction of data-driven statistical dynamic and static subunits (SUs). The latter is implemented in a state synchronous multistream Hidden Markov Model (HMM) framework that encapsulates movement's dynamics. Moreover, it integrates movement and position cues as multiple stream observations. This scheme lets us employ different features and models for the dynamic and the static cases. The HMM-based SUs are the intra-sign primitives that are reused to reconstruct the signs in the lexicon. Although we do not incorporate any linguistic information, our approach is inspired by L&J's work on Movement–Hold [6]. As L&J suggested that signs are formed by movements and non-movements (postures), we explicitly model movements and non-movements. In this way we actually *generate* a sequential structure of sub-sign models. This sequential structure is considered partially "phonetically meaningful"; this holds in the above explained terms of movements and non-movements. An example of this decomposition into Movements (M) and Holds (H) for sign ADMIT – H M H – is illustrated in Fig. 2. We represent movements and non-movements by different feature cues in each case; these correspond to the above movements and holds. We call these cues "*movement–position* cues" (M–P), and they are used for the explicit training of the corresponding Dynamic and Static models. Finally, handshape is also incorporated as a parallel information cue.

The overall framework is evaluated on data from three corpora and two SLs: Boston University SL corpus (BU400) [27], GSL Lemmas corpus (GSL-Lem) [26] and American Sign Language (ASL) Large Vocabulary Dictionary (ASLLVD) [24]. The experiments address multiple aspects such as exploitation of the M–P cues, integration of handshape information, employment of a single training example per sign, testing on unseen signers, and compensating for unseen pronunciations by employing a few development data. Finally, we present comparisons with three SU-level approaches [14,11,23], one sign-level approach [28] from the state of the art, and one similar approach to 2-S-U, without D/S discrimination (see Section 10). 2-S-U leads to improvements that show the importance of D/S sequentiality in sub-sign phonetic modeling.

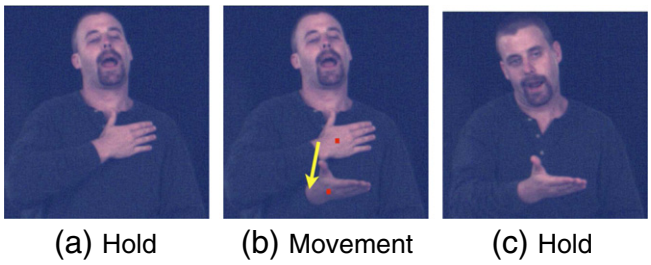

Fig. 2. Movements and Holds decomposition for ASL sign ADMIT (BU400).

2. Related literature and differences

2.1. Overview

Automatic sign language recognition is a multilevel problem posing significant challenges on feature extraction and information stream modeling –for a review refer to [5,1,4]. Most recent works are based on visual processing, instead of color gloves [11], data gloves [31,14,38], motion capture [12,36,13], and others [33]. We extract features after visual processing based on our earlier work [39]. In the following paragraphs we summarize several important aspects of ASLR, as related to our work. At the same time in Table 1 we present a list of indicative works as grouped w.r.t. some of these aspects. The issues discussed next include: 1) Learning and modeling techniques. 2) Other related tasks, such as sign spotting. 3) ASLR approaches inspired by Stokoe's phonetic decomposition. 4) Employment of model-based subunits and of a subunit-level lexicon. 5) Sequentiality and related works. 6) Unsupervised segmentation tasks. 7) Experiments with respect to the training data and the signers. and 8) Our earlier related work.

2.2. Modeling

ASLR involves multiple dynamically varying streams. It requires handling cues of variable duration and as discussed above, it involves an unknown phone inventory. Approaches addressing these aspects can be of parametric type, e.g. based on hidden Markov models (HMMs), conditional random fields (CRFs), or not, e.g. based on dynamic time warping (DTW). Hidden Markov model constitutes a popular approach because of its ability to account for dynamics [40]. Early attempts employed HMMs to build sign-level models [3,34] whereas various later works accounted for subunits, either explicitly [41,11,14] or implicitly [35]. Another important contribution concerns the parallel HMMs (PaHMMs) [12] that accommodate multiple cues simultaneously. In addition, other hybrid approaches appeared too, combining HMMs and recurrent networks [31,30], or the known tandem combination from the ASR community of multi-layer perceptrons with GMMs [35]. Markov chains are employed by authors in [20], and DTW can be found in exemplar-based cases [28]. Others stress discriminative aspects as in statistical DTW with discriminative features [42], HMMs with discriminative segmental features [33], multi-class Fischer kernels [32], and sequential pattern boosting with weak classifiers [23]. In 2-S-U we employ HMMs for explicit subunit models.

2.3. Other related tasks

Apart from sign recognition, other tasks have dragged attention and are worth mentioning, such as the detection of sign coarticulation points with CRFs [43], sign spotting [44] to distinguish non-sign patterns with threshold CRFs, and the modeling of epenthesis movements [34,38]. Authors in [45] explore sign extraction in subtitled videos, employing multiple instance learning in weak supervision, whereas in [46] they find the common patterns of signs, via iterative conditional modes on multiple sequences.

2.4. Stokoe's work and ASLR

A seminal work that has inspired many researchers is the one of Stokoe [2] who among other contributions proposes a parallel decomposition of signs into multiple components: tab (sign location), dez (handshape), and sig (motion). Several works have invested in the modeling of related components. Kadir et al. [20] employ a description based upon Stokoe's components for sign classification. Authors in [37] model the three basic components of signs by specific algorithms that recover in detail their 3D structure and recognize separately each component; finally, they combine the components in a tree-like structure. Derpanis et al. [47] recognize isolated movement phonemes by deriving mappings between the phonemic movements and the kinematic description of visual motions. The authors in [36] study sign inflections by modeling the systematic variations as parallel cues with independent feature sets employing a dynamic Bayesian network. Others combine these cues by forming subunits with regard to the basic components of signs. Cooper et al. learn weak classifiers, and combine them in sign-level classifiers via Markov chains or sequential pattern boosting [23]. The former scheme is employed to encode temporal changes and

Table 1
Indicative list of related works.[a]

Works		Sensor/FE	SU-Segm	Modeling	M–H seq.	Unseen signer
[28]	Sign-level	Vis.	✗	Exemplar based (DTW)	✗	✓
[29]		Vis.	✗	HMMs	✗	✓
[30]		Vis.	✗	HMMs/RNN	✗	n.a.
[3]		Vis.	✗	HMMs	✗	n.a.
[31]		d-Gloves	✗	SRN/HMMs	✗	✓
[32]		Vis.	✗	Multi-Class Fisher Score	✗	✓
[33]	SU-implicit	d-Gloves	DIST-SBHMMs	HMMs	✗	n.a.
[34]		MoCap	✗	CD-HMM + Epenthesis	✗	n.a.
[35]		Vis.	✗	MLP/HMM	✗	✗
[36]		MoCap	✗	DBN/MH-HMM	✗	✗
[22]		Vis.	Motion disk.	WC/Adaboost	✗	✗
[37]		Vis.	✗	Tree-based	✗	✓
[23]		Vis.	Rule-based	WC/SP,MC	✗	✓
[20]		Vis.	Rule-based	WC/MC	✗	✗
[11]	SU-explicit	Vis. + c-gloves	K-means	HMM	✗	n.a.
[14]		d-Gloves	LR-HMM	MKM-DTW, HMM	✗	n.a.
[38]		d-Gloves	LR-HMM	MKM-DTW, HMM + Epenthesis	✗	n.a.
[12][b]		MoCap	✗	PaHMM + Epenthesis	✗	n.a.
[13]		MoCap	rule-based	HMM	✗	✗
2-S-U		Vis.	2S-ERG HMM	MS-HMM, PaHMM	✓	✓

[a] FE refers to feature extraction, Segm. to segmentation and M–H Seq. to Movement–Hold sequentiality. Vis. refers to visual processing, d-gloves to datagloves, MoCap to various motion capture devices and c-gloves to color gloves. LR-HMM refers to left-right HMM, motion disk. to motion discontinuities and 2S-ERG HMM to a two-state ergodic HMM. DIST refers to discriminative state-space tying, MKM-DTW to modified k-means employing DTW and Hier. to hierarchical clustering. SBHMMs refers to Segmentally Boosted HMMs, RNN to recurrent neural network, SRN to simple recurrent network, DBN to dynamic Bayesian network, MH-HMM to multichannel hierarchical HMM, WC to weak classifiers, SP to sequential pattern boosting, MC to markov chains, MS-HMM to multistream HMM, and CD-HMM to context-dependent HMM. Finally, n.a. refers to the case of non-availability of the specific information in the corresponding publication.

[b] Employs manual SU-level annotation.

the latter to apply discriminative feature selection and to encode temporal information. Han et al. explicitly perform sub-sign segmentation based on motion discontinuities into motion subunits, inspired by syllable phonetics [22]. Next, they combine weak classifiers with boosting into sign-level classifiers. As far as the segmentation is concerned this work shares similarities with our velocity based segmentation; however all subunits are of a single type in contrast to our case. In [48] they report increased performance by "sharing features across classes." Data-driven units (called "fenemes") are computed in [33] after discriminative segmental feature selection. All the above, – unlike works that employ global image features, as [35] – model articulatory components inspired by Stokoe, and finally combine them in sign-level models. 2-S-U similarly exploits as features, local cues, inspired by Stokoe and L&J. Nevertheless, we employ explicit statistical sub-sign units, referred to as subunits (SUs) instead of whole sign models.

2.5. Advantages of model-based SUs

Explicit sub-sign models have attracted interest because of several advantages when compared with sign-level models. First, they scale well with increasing vocabulary size requiring smaller amount of training data, since subunits are shared across signs. Another point concerns the SU-level lexicon; the SU-level lexicon allows the incorporation of new signs without requiring model retraining. Apart from linguistic based SUs this also holds for data-driven SU approaches given that: 1) the training phonetic data, account for the new sign's phonetic data, 2) there is at least one iteration for the new sign to construct the SU pronunciation after SU-level decoding. This pronunciation is then inserted in the dictionary as a new sign entry. Finally, model-based SUs allow the adaptation to different conditions or signers, to decrease the mismatch with test data.

2.6. Explicit model-based SU approaches

Indicative works for statistical SUs are the following: Bauer and Kraiss introduced a data-driven approach for SU-level segmentation and modeling [11]. They cluster independent frames via K-means to construct a data-driven SU-level lexicon and employ HMMs to model SUs. Fang et al. [14] employ a 3-state left-right HMM for SU-level segmentation and modified k-means with DTW to cluster segments, exploiting the dynamics that are essential in ASLR. Kong and Ranganath [13] segment motion trajectories via rule-based segmentation. They extract features based on principal component analysis (PCA) and cluster them by K-means. However, all the above do not account for any concept similar to dynamic–static that implies the sequential phonemic contrast: all subunits are of a single type.

2.7. D/S sequentiality

To linguistically account for both simultaneous and sequential phonemic contrast Liddell and Johnson proposed the Movement–Hold model [6]. They introduce two classes of segments *Movements* and *Holds*: "Movements" correspond to segments during which some aspect of the sign's configuration changes, such as a movement or a change in handshape. In contrast, "Holds" correspond to segments during which no aspect of the sign's configuration change. As a result, signs are made up of movements' and holds' sequences. 2-S-U introduces an unsupervised statistical phonetic modeling framework inspired by the above work. To our knowledge it is the first time that a computational unsupervised data-driven model is introduced based on these concepts for ASLR. The first works in computational sub-sign statistical modeling were in [41,12]. This presented an ASLR framework, by breaking down the signs into subunits, employing manual phonetic transcriptions based on the Movement–Hold model and then statistically modeling them with parallel HMMs [41,12]. As [11] noted, although the sequential model of L&J "seems to be more appropriate for the recognition of SL, as it is partitioned in a sequential way" it requires time consuming manual transcriptions and is thus in practice not feasible. We alleviate this problem and via 2-S-U we provide an unsupervised data-driven perspective for which no manual phonetic transcriptions are employed. 2-S-U introduces sequential phonemic contrast in an unsupervised computational manner, via the discrimination between dynamic and static SUs in contrast to [11,14,13].

2.8. Unsupervised segmentation

Unsupervised segmentation into D/S segments is implemented by an ergodic HMM; see Sections 3 and 5. On its own this specific modeling approach is implemented in other domains as well, such as unsupervised speaker segmentation [49], segmentation of emotions with regard to facial expressions [50], and gesture spotting [51]. Nevertheless, the way it serves our purposes to gain sequentiality in an unsupervised way, in the overall HMM framework is different. The above is partially related to methods explicitly employing hierarchical techniques, as hierarchical-HMMs [52] for unsupervised video segmentation or segmentation of meeting data [53]. Herein, we do not employ a hierarchical model, but implicitly build two layers of models via unsupervised segmentation (Section 5) and SU construction with statistical training (Section 8).

2.9. Experiments, training data and signers

Important aspects for the experimental evaluation of an approach are the size of the employed training data and whether the test signer is unseen. Wang et al. present a sign lookup dictionary tool proposing an exemplar-based approach based on dynamic time warping that deals with small quantities of training data [28]: they report results for 10-best sign recognition accuracy 78% in a tough recognition task, with 1113 signs, two train instances per sign and testing on an unseen signer. Kadir et al. [20] employ a single training example per sign, and report 76.2% accuracy in 164 signs, on signer dependent experiments. Since [31] several authors apply signer independent testing [32,29]. Cooper et al. [23] present results of 76% and 49.4%, for 20 and 40 signs respectively. Further, Fang et al. [31] show results up to 92% for 208 signs employing data gloves. Overall, unseen signer testing deteriorates performance significantly, when compared with the signer dependent case, as for instance: 55 percentage points (pp) for 232 signs in [29], and 16 pp or 10.4 pp for 20 or 40 signs respectively [23]. We evaluate 2-S-U in signer dependent testing and in unseen signer experiments with a single training signer and a single sign instance for training.

2.10. Our related work

A brief presentation of our visual tracking system can be found in [39]. In [18] we introduced an approach based on linguistic information via phonetic transcriptions, in contrast to this work, which does not employ any intra-sign phonetic transcriptions; then, [54] extends [18]. Among earlier works the more relevant ones are [55], and mainly [56], being exploratory and preliminary respectively. The differential to the more related, second one, is significant and includes: 1) The dynamic–static framework, that is integrated via a statistical SU HMM based scheme. 2) Intermediate results highlight the unsupervised lexicon and differences on signers' pronunciations. 3) Handshape integration and SUs. 4) Incorporation of non-dominant hand. 5) Experiments in data from multiple corpora and comparisons.

3. System overview and contributions

An overview of the proposed framework is presented in Fig. 3, consisting of: 1) Unsupervised D/S Sequentiality, SUs and Lexicon, 2) Statistical SU Training, and 3) Recognition.

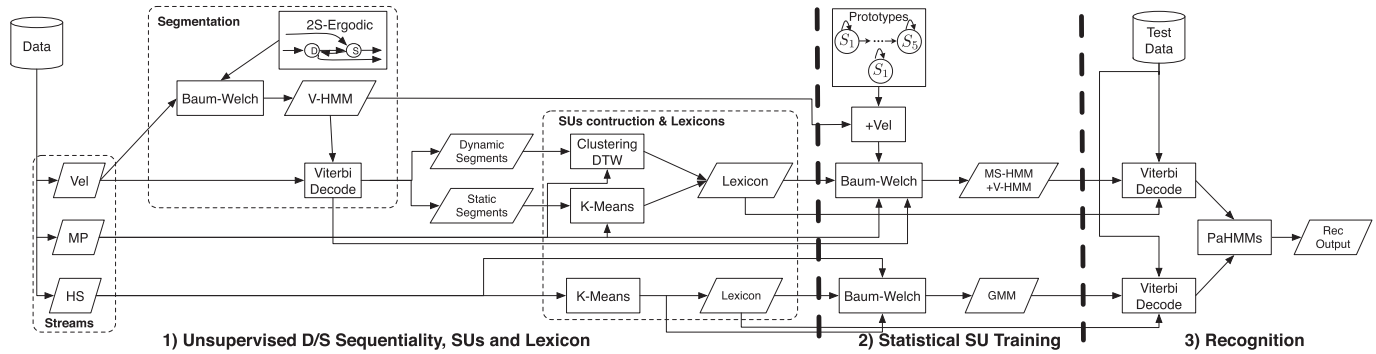

Fig. 3. Overall 2-S-U HMM-based framework and components for automatic sign language recognition. Rectangles represent procedures; parallelograms represent input and output data. 1) Unsupervised D/S sequentiality and subunit construction: exploits the velocity cue, segments the signs into sub-sign segments and clusters separately the dynamic and static ones. 2) Statistical HMM SUs: incorporates the D/S statistics, integrates the D/S SUs into multistream HMMs for SU training. 3) Recognition: decoding and late integration of handshape and M–P cues. In all cases, "data" corresponds to already extracted features by the visual front–end, i.e., velocity (Vel), movement–position (M–P), and handshape (HS) feature vectors. V-HMM corresponds to the trained D/S Gaussian models employing velocity. "+Vel" is the encapsulation of the D/S pdfs in the multi-stream HMM.

3.1. Unsupervised D/S sequentiality, SUs and lexicon

The first part of our contribution includes the SU-level lexicon and the incorporation of D/S phonetic sequentiality. This is realized via segmentation into dynamic and static intra-sign segments for the movement–position (M–P) cue. For this segmentation, we employ a two-state ergodic HMM (2S-Ergodic) to model the movement dynamics via the velocity (Vel) feature (Section 5). In SU construction, for each segment type we employ the appropriate cues, and clustering (Section 6): This is hierarchical clustering for the dynamic segments with a dynamic time warping (DTW) metric and K-means for the static ones. K-means is similarly applied on handshape (HS). Finally, we recombine the segmentation and cluster information to construct two SU-level lexica: for M–P and HS (Section 7). Outputs such as the dynamics' distributions, the sequential D/S labels and the segments' clusters hold a major role next.

3.2. Statistical SU training

Another part of our contribution, concerns D/S statistical SUs training (Section 8.2): we employ a state synchronous multi-stream HMM scheme (MS-HMM) to integrate the M–P cues, and to incorporate the D/S sequential structure. In this scheme we encapsulate ("+Vel") the trained velocity probability distributions (V-HMM). Furthermore, we employ stream weights to use only the features that contribute in the dynamic or static case (Section 8.1). Handshape SUs are modeled by a Gaussian model.

3.3. Recognition

Herein we employ the trained SU models and the SU-level lexica separately for M–P and HS. Recognition finds both the D/S sign segmentation, and the most probable SU per segment, among the dynamic or static SUs. This results to the most probable D/S SU sequence (Section 8.2). The sign recognition output (Rec. Output) is obtained via a late fusion scheme, after the integration with the HS via Parallel HMMs (PaHMMs) [12], combining the introduced D/S sequentiality with multi-cue parallelism.

4. Visual processing of sign language videos

Next, we summarize the main parts of our visual processing front–end and the produced features. These features are the input for the system presented in Section 3.

For image segmentation and tracking we employ our previous work [39], the main components of which include: Estimation of the hands and head locations based on color cue, by a skin color model; we handle occlusions via ellipses in each body-part, and then employ a forward–backward linear prediction for the estimation of ellipse's parameters. For signer dependent parameters, such as body size and scale, we apply a simple calibration for the body of the signer w.r.t. a reference signer. This is based on foreground detection, registration of the user's binary mask, and finally estimation of the rotation and scale parameters.

After image segmentation and tracking, we extract features that represent position (as PoA) and movement. Specifically, we extract the (x,y) centroid coordinates using as reference point the centroid of the signer's head. This, although a convention, is due to the head's importance as a PoA. Moreover, we construct features that are products from the (x,y) coordinates of the hands' centroids. These are the velocity $V(t) = (\dot{x},\dot{y})$, and the instantaneous direction $D(t) = (\dot{x},\dot{y})/\left(\dot{x}^2+\dot{y}^2\right)^{1/2}$.

For handshape feature extraction we employ the concept of spatial pyramids [57]. For the hand segmentation we employ the aforementioned tracking system. Next, we extract dense Scale Invariant Feature Transform features [58], and apply K-means clustering of a random set of patches from the training set to form a visual vocabulary. The size of this vocabulary in the experiments is set to 10. Afterwards, we compute the histograms of the visual vocabulary in 3-level pyramids similar to [57]. After concatenating the histograms of each pyramid level we embed this feature vector in a feature space in which the inner product between two vectors is equal to the histogram intersection distance before the embedding [59]. Finally, we employ PCA for dimensionality reduction keeping the first 100 eigenvectors out of 630. All the above parameters are set experimentally.

5. Unsupervised segmentation and D/S labeling

The first component concerns sign segmentation into intuitive sequential sub-sign segments and classification into Dynamic (D) and Static (S), i.e., movements and non-movements respectively. The output of this section implies the D/S sequential structure.

5.1. Dynamic and static modeling

The classification into dynamic and static segments is based on the movement dynamics. For this we exploit the velocity cue. We assume that dynamic and static segments share on average relatively high and low velocity respectively. Since the segments are of two types (D/S) we employ two single Gaussian models for the segmentation procedure. Then we employ a 2-state ergodic HMM scheme to combine these single Gaussian models. In this HMM, the first state corresponds to the Static and the second to the Dynamic Gaussian model. We train the ergodic HMM by the Baum–Welch algorithm, employing all sign

realizations in the training dataset. Thus we end up with two trained Gaussian models one for the static and one for dynamic segments. In Fig. 4a we illustrate the velocity distribution superimposed with the probability density functions (pdfs) corresponding to the trained D/S Gaussian models. In this way, we implicitly estimate the threshold to separate movements from non-movements.

After training the ergodic HMM we find via Viterbi decoding the most probable state sequence, i.e., the segmentation into D/S segments for each sign instance in the training dataset. Fig. 4b shows a segmentation example for an instance of the sign ADMIT. The D/S structure of the sign ADMIT is "S D S." This result should be seen in comparison with Fig. 2 where the manual decomposition based on the Movement–Hold Model is "H M H."

5.2. Summary and outputs

This segmentation model-based approach offers various advantages. First, we obtain both the segmentation and the D/S labels since we encapsulate implicitly the dynamic and static notions in the states of the same model. Second, we do not need to optimize any parameter or to manually set any threshold. Then, the whole D/S segmentation approach, including re-training of Dynamic/Static models and decoding, is applicable to other datasets too. Finally, the model-based nature fits with the probabilistic framework. The outputs are next exploited as follows: The D/S segmentation is applied to different cues, such as the direction, resulting on the actual segmented signals per cue employed in clustering (Section 6). Then, the lexicon construction employs the D/S sequence and the mapping of segments to their assigned clusters (Section 7). The D/S pdfs are exploited to encapsulate discriminative dynamics information in the statistical SU HMMs (Section 8.1). Finally, the D/S segmentation is employed during the SU models training (Section 8.2).

6. Dynamic and static subunits

We present the clustering procedure for D/S SUs. We take as input the aforementioned segmentation and employ the appropriate features according to the D/S classification. At the end, the SUs consist of clustered segments; in Section 8 we model the statistically within the HMM framework.

6.1. Construction of dynamic subunits

For the dynamic SUs we take advantage of dynamic information, in sequences of frames, which is considered important for the modeling of movements. For the modeling of the movements we next present the employed feature representations. Then, we describe the clustering of the segments based on the underlying features.

The employed feature representation is either the instantaneous direction feature, or the actual positions across time normalized w.r.t. scale and initial position. The direction feature vector has been defined in Section 4. Next, we describe the normalizations applied in the position feature vector. The modeling of the movement trajectories by employing the position feature without any normalization increases the model's variance. This increase is because of the translation of the movements to various places in the signing space. Segment normalization by its corresponding initial position leads to a translation-invariant modeling. In Fig. 5a and b, we illustrate the movement trajectories with and without normalization. Scale, which corresponds to the amplitude of movements, also affects their modeling. Scale normalization yields scale-invariance. At the same time, we do keep the scale parameter for further use. An example of this normalization is presented in Fig. 5a and c. Finally, Fig. 5d shows the same trajectories after both scale and initial position normalization (SPn). It is more effective to incorporate these normalized segments for clustering instead of the non-normalized ones.

6.2. Clustering dynamic segments

We start with the segments produced in Section 5. Next, we cluster sequences of features, by employing DTW to compute a similarity matrix among the segments. Take for instance two arbitrary segments $X = (X_1, X_2, \ldots, X_{T_x})$ and $Y = (Y_1, Y_2, \ldots, Y_{T_y})$ where T_x, T_y are the number of frames of each one. We define the warping path $W = ((x_1,y_1),\ldots,(x_N,y_N))$ where $1 \le x_i \le T_x, 1 \le y_i \le T_y, N$ is the length of the warping path and the notation of the pair (x_i,y_i) signifies that frame x_i of X corresponds to frame y_i of Y. The measure $d(X_{x_i}, Y_{y_i})$ is the Euclidean distance. DTW aims to search the minimal accumulating distance and the associated warping path: $D(X,Y) = \min_W \sum_{n=1}^{N} d(X_{x_i}, Y_{y_i})$. Finally, the distance similarity matrix among all segments is exploited via hierarchical agglomerative clustering employing as end criterion the number of clusters. Technical details are omitted due to space limitations [60]. As follows, we construct clusters of segments accounting for the dynamics. Each cluster defines a dynamic SU, that is to be modeled later on via an HMM. The number of employed clusters is set experimentally based on recognition performance on a development set, discussed in the experiments (Sections 10–12).

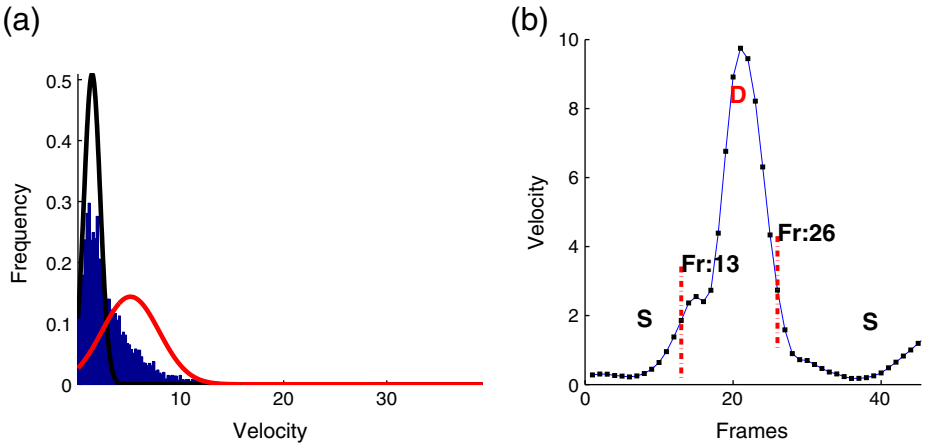

Fig. 4. (a) Velocity distribution (histogram) superimposed with the pdfs (red and black curves) corresponding to the two states of the ergodic HMM. Black corresponds to the static Gaussian distribution, and red to the dynamic one. The unit for the x axis is pixels per frame, and for the y axis is the normalized frequency. (b) Segmentation points shown on the velocity profile for sign ADMIT with the D/S labels per segment.

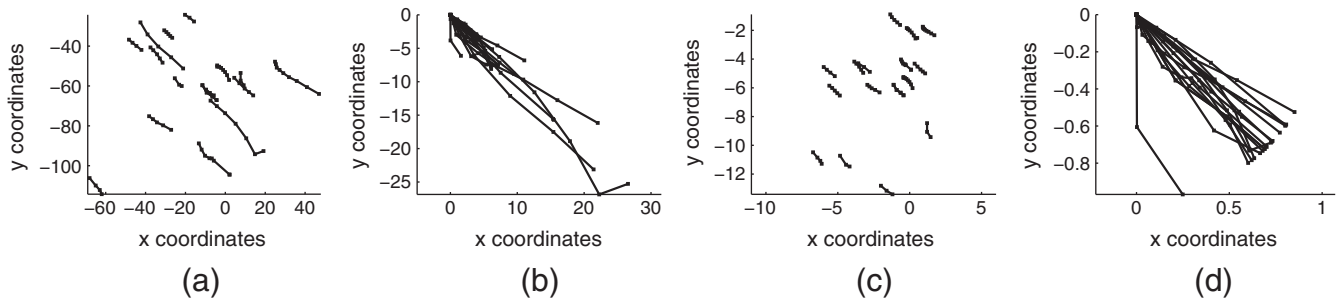

Fig. 5. Trajectories of dynamic movements mapped onto the 2D signing space: (a) Without any normalization. (b) After initial position normalization. (c) After scale normalization. (d) After initial position and scale normalization.

6.3. Dynamic subunits for different or multiple cues

Next, we explore the features that are employed for the dynamic segments. The output of the clustering partitions each feature space separately. Each cluster in this partition is a distinct subunit; this is identified by the feature employed and the assigned cluster id.

After normalization steps, each segment corresponds to a normalized trajectory. We show in Fig. 6a indicative SUs: these clusters are constructed after hierarchical clustering, and are then mapped onto the 2D signing space. This mapping retains the SU identity, encoded by a distinct color. For instance SU "SPn1" corresponds to curved movements with direction down-left. Characterizations as "down-left" concern our interpretation, and does not correspond to any transcription. An example in which "SPn1" SU appears is sign "END" in Fig. 1b. In Fig. 6b we show indicative cases of SUs employing as feature the non-normalized positions. It is evident by comparing with the previous Fig. 6a, that the SUs are less intuitive since the models are consumed on the explanation of different initial positions or scales. In addition, the clusters produced by the normalized trajectories implicitly incorporate direction information; this is since the direction is dependent on the geometry of the trajectory.

SUs constructed with the direction feature, show similar results as the ones for the normalized movement trajectories. Each SU consists of movements with similar direction on average. Fig. 6c shows indicative examples of movements over different clusters having on average different directions. For instance subunit "D10" models curved movements with direction down-right. An example in which the SU "D10" appears is sign "HERE" in Fig. 1a. Concerning the scale of each trajectory we show in Fig. 6d two indicative scale SUs. These model trajectories according to their scale. Note that the subunits with labels "S9" and "S3" appear in the ASL signs "END" (Fig. 1b) and "HERE" (Fig. 1a) respectively.

6.4. Multiple movement cues

Herein, we employ multiple cues by concatenating the multiple features. By incorporating both direction and scale we create multiple-cue SUs that model movements based jointly on direction and scale. Such SUs appear in Fig. 7a, via the corresponding trajectories in the signing space. Each SU refers to both direction and scale. Finally, we also show examples of joint direction-scale SUs for two ASL signs, "DEAF", "DECIDE", by superimposing their initial and final frames with an arrow depicting the trajectory. The movement in Fig. 7b corresponds to the direction-scale SU D2S2: This is a straight movement with direction D2 (up-left) and scale S2 (small). The movement in Fig. 7c corresponds to the direction-scale SU D1S4: This is a straight movement with direction D1 (down-right) and scale S4 (medium). The above labels in parentheses come from our interpretation; they have been added for a qualitative description of the involved cues, to assist their presentation.

In the experiments (Section 10) we have explored all the above features, these include both single-cue feature vectors (i.e., movement trajectories, direction, and scale) and combinations of them (multi-cue feature vectors). However, after experimentation as discussed in Section 10 of the experiments, we concluded on employing for the dynamic SUs the single-cue direction.

6.5. Static subunits

Static segments correspond to the low velocity profile of the ergodic HMM (Section 5). For the static SU construction we cluster *only* the static segments. Specifically, we apply K-means on the position feature vector. In this way we get a partitioning relative to the signer's head. Fig. 7d shows in different color the constructed SUs together with the centroids for each cluster as mapped on the 2D space. These are dependent on the

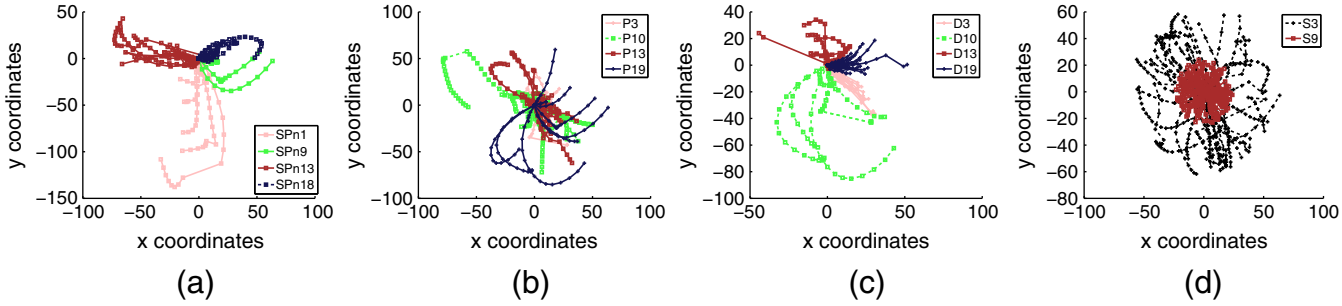

Fig. 6. The trajectories for different SUs mapped on the 2D signing space after normalization w.r.t. initial position. With different color we represent different SUs corresponding to different clusters. (a) Trajectories of SUs that incorporate both scale and initial position normalization (SPn). (b) Trajectories of SUs obtained using as feature the movement trajectories (P) without any normalization. (c) Trajectories of SUs that incorporate Direction (D). (d) Trajectories for two different SUs that correspond to different Scales (S).

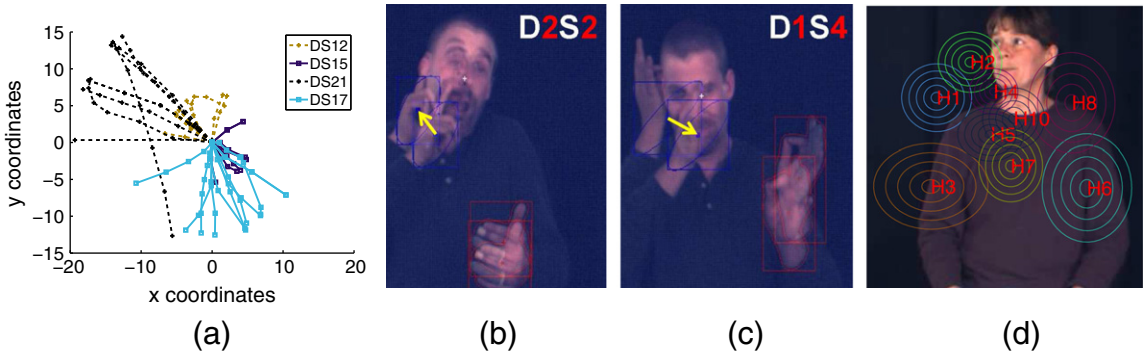

Fig. 7. (a) Trajectories for multi-cue SUs mapped on the 2D signing space with different color/marker. SUs account for both direction and scale. (b,c) Examples of multi-cue SUs for direction and scale that correspond to the movement for the ASL signs "DEAF", "DECIDE" (BU400). (d) Partitioning of the 2D signing space by K-means for the static subunit construction, superimposed on a frame for signer Lana (ASLLVD).

employed space of the signer, as they appear in the dataset. Finally, the number of clusters is set experimentally based on recognition performance on a development set, as discussed in the experiments.

6.6. Handshape subunits

For the handshape SUs we do not employ the D/S component. Handshape SUs are constructed in a data-driven way similar to [11]. All frames are considered in a feature pool in which we apply K-means; for this we employ the Euclidean distance. Each cluster corresponds to a different SU. In Fig. 8 we show samples from different handshape SUs as they appear in GSL-Lem data. For each SU, we show the corresponding original data samples. As expected, this correspondence involves similar handshapes.

7. Lexicon and segmentation results

Given the lack of phonetic transcriptions, we construct data-driven phonetic lexica for the M–P and handshape cues. These are based on outputs of the D/S segmentation component (Section 5), and the clustering of the D/S segments which leads to the D/S SU construction (Section 6). The M–P lexicon inherits the D/S sequential structure. This is in contrast to the handshape lexicon, for which the D/S segmentation is not employed. Afterwards, the lexica are used in training and in sign accuracy evaluation (Section 8.2).

7.1. Lexicon for the movement–position cue

After decomposing and clustering the D/S segments we recompose the labels, hereafter referred as symbols, producing the lexicon. In this way the lexicon consists of an entry for each sign instance as it appears in the dataset. Each SU label is a symbol identified by a concatenation of the assigned Dynamic (D) or Static (P) SU label, and the cluster id assigned after clustering. For the D/S SUs we employ the direction and the position features respectively. The non-dominant (ND) hand is taken into account during postures when both hands are non-moving (static), and during transitions when both hands are moving (dynamic). In all other cases only the dominant hand is processed. The differentiation between static and dynamic parts is done by the trained D/S Gaussian models; these employ the velocity feature as in Section 5. Thus if the ND hand is active the SU identifier accounts for both hands: for instance SU "D6–D8" (Fig. 9) corresponds to a dynamic SU (D) and id 6 for the dominant hand, and a dynamic SU with id 8 for the ND.

7.1.1. Other approaches and results

By employing different approaches, we construct the lexica and segmentations that we next compare. These correspond to Fang et al., 2004 approach [14] (SU-Segm), Bauer and Kraiss, 2001 approach [11] (SU-Frame), and SU-noDSC. In brief, none of them discriminates between D/S SUs; especially SU-noDSC closely resembles 2-S-U, by sharing common segmentation results. For details on these approaches see also Section 10. The notation for each SU consists of a constant string "SU" with the cluster identifier (id) assigned after clustering. In Fig. 9 we illustrate in the horizontal, time axis, each image frame with both the SU symbols and segmentations for all approaches, for the ASL sign "ACCIDENT." Fig. 12 (bottom) shows an additional example for the decomposition of sign "ANY", with the corresponding HMM SUs (Section 8).

7.1.2. 2-S-U results

As shown in Figs. 9 and 12, 2-S-U decomposes each sign into a D/S SU sequence, following an estimate of the actual articulated movements and postures. Movements are explicitly modeled by Dynamic SUs (D) and postures by Static SUs (S). For instance, sign "ACCIDENT" consists of a posture modeled by S1–S8, a simultaneous movement of both

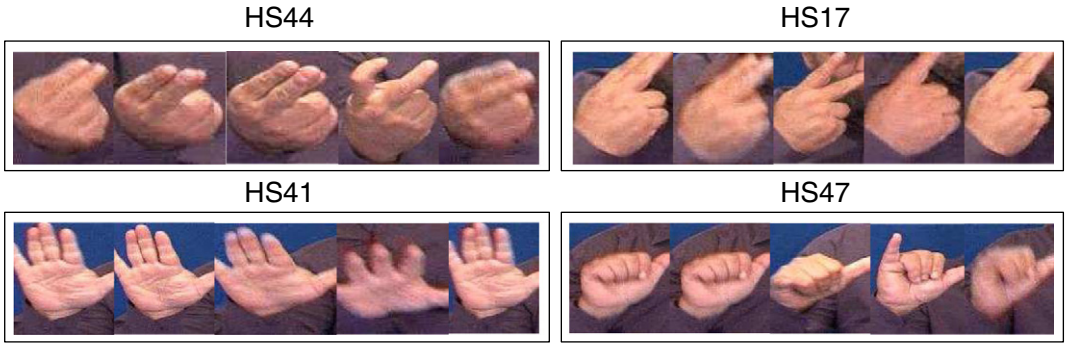

Fig. 8. Samples from different handshape SU clusters (GSL-Lem).

Fig. 9. Lexicon and segmentation results for ASL sign "ACCIDENT" (ASLLVD). SU sequence of symbols and segmentation after dynamic–static decomposition for 2-S-U. Comparison of segmentation and SU results for multiple approaches.

hands represented by D6 and D8 respectively and finally a posture (S4–S10). Similarly, sign "ANY" is decomposed into a posture (S5), followed by two consecutive movements of the dominant hand D6 (up-right) and D8 (up-left), and finally a posture (S1). Moreover, SUs are shared across multiple signs. For instance S1 and D6 appear in both signs. In addition, same SUs are shared across both hands: D8 (up-left movement) appears in both signs, "ANY" for the dominant hand, and "ACCIDENT" for the ND hand.

7.1.3. Comparison of results

Both the SU-noDSC and 2-S-U result in the same segmentation, since they employ the same velocity-based segmentation algorithm. The substantial difference is that SU-noDSC does not discriminate between Dynamic and Static segments. As a consequence SU-noDSC concatenates both movement and position cues, and then constructs subunits by clustering all segments independently of their Dynamic or Static label. As a result the SU partitioning is done in this multi-cue feature space. The SU-Segm and SU-Frame approaches lead to different segmentations and SU decompositions. These segmentations are not characterized by the D/S concept, and do not contain distinct moving and non-moving parts. In addition, a single movement or posture may be segmented into multiple segments.

7.2. Lexicon for the handshape cue

After handshape SU construction (Sec.sec-wp2-su-HS) we recombine the produced symbols in a corresponding lexicon. This consists of a lexical entry per sign pronunciation. SU notation consists of the identifier of each handshape SU (HS) and the cluster id after clustering. Fig. 10 shows the handshape SU segmentation and clustering for two signs as articulated in the GSL-Lem corpus. Sign "SEE" consists of the SU HS17 followed by HS44. Refer to Fig. 8 where we show handshape samples for these two handshape SUs. Although these SUs correspond to the same handshape they differ in their 3D pose. Finally, GSL sign "ABROAD" consists of two subunits (HS47 and HS41) that correspond to different handshapes.

8. HMM dynamic/static sequentiality and statistical SUs

According to the lexicon that is generated as described in Sections 6 and 7, each sign is composed by a sequence of dynamic/static subunits. Further, the D/S segmentation provides the temporal boundaries of each subunit. Here, we aim to employ a probabilistic HMM scheme for training and recognition. This should account for D/S sequential structure, but also allow for multi-cue parallelism.

8.1. Overview

For training we wish to impose the D/S sequential structure implied by the existing segmentation. Further we wish to employ the D/S segments' clusters in the training of the statistical SUs. In recognition we aim to find both the most probable D/S segmentation, given the sequence of features (observations), and the statistical SU models that best match each type of dynamic/static segment. The above goals are fulfilled by a HMM that encapsulates the D/S velocity discriminative pdfs, and integrates the movement–position cues as multistream observations. For this we employ a state-synchronous multistream HMM. The multi-stream models for the D/S cases include the velocity (Vel), the position (Pos) and movement (Mov) cues, all for the dominant hand (D). Similarly, we integrate the non-dominant (ND) hand (Section 8.3). Next, we present how to employ these cues to serve our goals.

8.2. Dynamics' encapsulation and D/S sequentiality

First, we describe the encapsulation of the velocity pdfs, then the employment of the multistream HMM, and the role of stream weights. Finally, we compare our view with the typical multistream HMM case.

The velocity pdfs correspond to the states of the ergodic HMM (Section 5) are used for initialization of the velocity streams for the SU HMMs. Specifically, for all static and dynamic SU models we employ the velocity pdf that models low-velocity and high-velocity segments respectively. Further, the stream weight employed to the velocity stream affects the resulting likelihoods. In this way we implicitly deal with the different feature magnitudes. This stream weight is set

Fig. 10. Handshape SU decomposition for GSL signs "SEE" and "ABROAD" (GSL-Lem). "SEE" contains a single handshape with varying appearance due to the 2D data; "ABROAD" contains a varying handshape.

experimentally based on the recognition performance in the development data set.

8.2.1. Multistream HMM

Before proceeding on the training of the statistical SUs, it is essential to place constraints on the features of each SU model: the SUs corresponding to movements and the ones corresponding to the non-movements should depend *only* on the movement and position cues respectively. Movements can be seen as stacked in sequence and between them there are "gaps"; these gaps are actually non-movements. Then, we wish to employ different features and models in each segment. Thus, the employment of the multistream HMM paradigm in a typical way does not match our requirements. This is since both types of features, movement and position, would be taken into account. Here comes our view on how to employ the multiple streams, to serve our requirement. We transform the otherwise non-linear sequential stacking of models with different cues (see Fig. 11a), into a multistream scheme. In this scheme we view the movement and position cues as "parallel" streams across time. See Fig. 11b against the previous one. Nevertheless, there is still one element missing.

8.2.2. Sequentiality and stream weights

Then here comes the role of stream weights. A stream weight (SW) is a weighting factor that multiplies the log-probability of each emitting state, generating the corresponding observation of the specific stream. Specifically we employ one weight per stream and per SU model. In other words each SU multi-stream HMM model has its own stream weights, one for each stream. Here, we wish to employ SW, to implicitly constrain that dynamic SU models depend *only* on movement cue, and that static SU models depend *only* on position cue. This is accomplished by construction as follows: For static HMM models the SW for the movement streams are set equal to zero and for the position streams are set to one. Vice versa, for dynamic models the SW for the position streams are set to zero and for the movement to one. Thus, we account for different feature streams for the dynamic and static SUs, as if they are interlaced across time (Fig. 11c).

8.2.3. Our interpretation on streams and stream weights. The multistream paradigm is employed to model independent streams or different temporal resolutions; as for instance in audio–visual and multiband ASR [61]. These compensate for the relative reliability or the importance of each stream by equalizing the likelihoods of the different information streams. Herein we take advantage of the multistream scheme, and exploit an extreme case of stream weight compensation. "Extreme" refers to the canceling of the corresponding likelihood in the following way. We consider for the dynamic models, the position features as *inappropriate*, instead of more or less reliable. We thus assign on this stream and for the duration of this specific dynamic model and segment, the extreme weight of *zero*. This implies also a zero likelihood. The opposite holds for the static case.

8.3. D/S SU training and recognition

8.3.1. Training

For the training of the subunits we employ the described multistream scheme: this makes use of a 5-state HMM with Bakis topology [40] for the dynamic SUs and a 1-state HMM with one Gaussian per stream for the static SUs; stream-weights are as discussed above (Section 8.1). The time boundaries for each D/S segment have been extracted during segmentation (Section 5). These segments together with the clustering information are used to map the training examples and the corresponding SU models. We initialize the multistream HMM models employing an iterative scheme. The Viterbi algorithm is used to find the most likely state sequence for each subunit instance. This is repeated for each training example. Then we estimate the HMM parameters. As a by-product of the Viterbi state alignment, we get the log-likelihood of all training data. The whole estimation process is repeated until we obtain no further increase in likelihood. After this initialization, we apply Baum–Welch re-estimation [40]. For each training example we consult the SU-level lexicon to convert each sign into the D/S SU sequence, and construct a composite D/S network employing the corresponding multistream SU models (HMMs). This network is employed to collect the necessary statistics for the re-estimation. When all the training examples are processed, the total set of accumulated statistics is used to re-estimate the parameters of all of the dynamic and static HMMs. The training of the handshape SUs is done separately, by employing the handshape lexicon and the corresponding segmentation boundaries after the handshape frame-level clustering (see Fig. 3). Since for the handshape we do not consider the D/S segmentation, we employ a single-stream Gaussian model.

Fig. 12 illustrates an example highlighting some of the above. On top, the D/S HMM, outputs the segmentation and the D/S symbol sequence, not explicitly depicted here. It also feeds the appropriate velocity pdf, D or S, to each HMM SU, of Gaussian distributions (shown at the second layer). The encapsulated Dynamic and Static pdfs are presented in different colors. Movement and position cues are incorporated in separate streams. Note the shaded boxes that correspond to zero stream weights, constraining the D/S sequential structure. These statistical models are linked in a network as prescribed by the D/S sequences in the lexicon, to construct a composite D/S network. Finally, the HMMs output the observation symbols per stream: these are the visual observations corresponding to the image sequence for the ASL sign "ANY". This network consists of one static HMM SU (S5), followed by two dynamic HMM SUs (D6 and D8), and finally one static HMM SU (S1). We also show the D/S segmentation output at the bottom of the frames, comparing with multiple SU-level methods (discussed in Section 7.1).

8.3.2. Recognition

Recognition is conducted employing the trained D/S subunit models and the recognition network. First, we construct the aforementioned composite D/S networks for each pronunciation. These networks employ the trained HMMs as they appear in the SU-level lexicon. Then we construct the recognition network by combining the composite D/S networks. In this way, we end up with a recognition network that consists of nodes, these are the HMM subunits connected by arcs. Every path in the recognition network that passes through exactly T emitting HMM states is a potential recognition hypothesis for a test example with T frames. Each such path has a log-probability that is computed by summing the log-probability of each individual transition in the path, and the log-probability of each emitting state generating the corresponding observation. At each time instance, we find the path maximizing the above log-probability, i.e., the most probable D/S SU

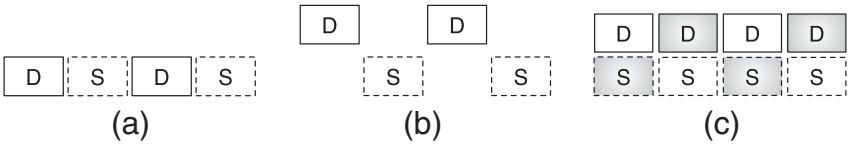

Fig. 11. Each box, corresponds to either dynamic (continuous line) or static (dotted line) model and features. (a) Intended D/S sequential structure. (b) Splitting into separate streams the D from the S models and features. (c) Actual implementation of D/S sequentiality, via multiple streams; gray boxes are inappropriate, and correspond to zero stream weight.

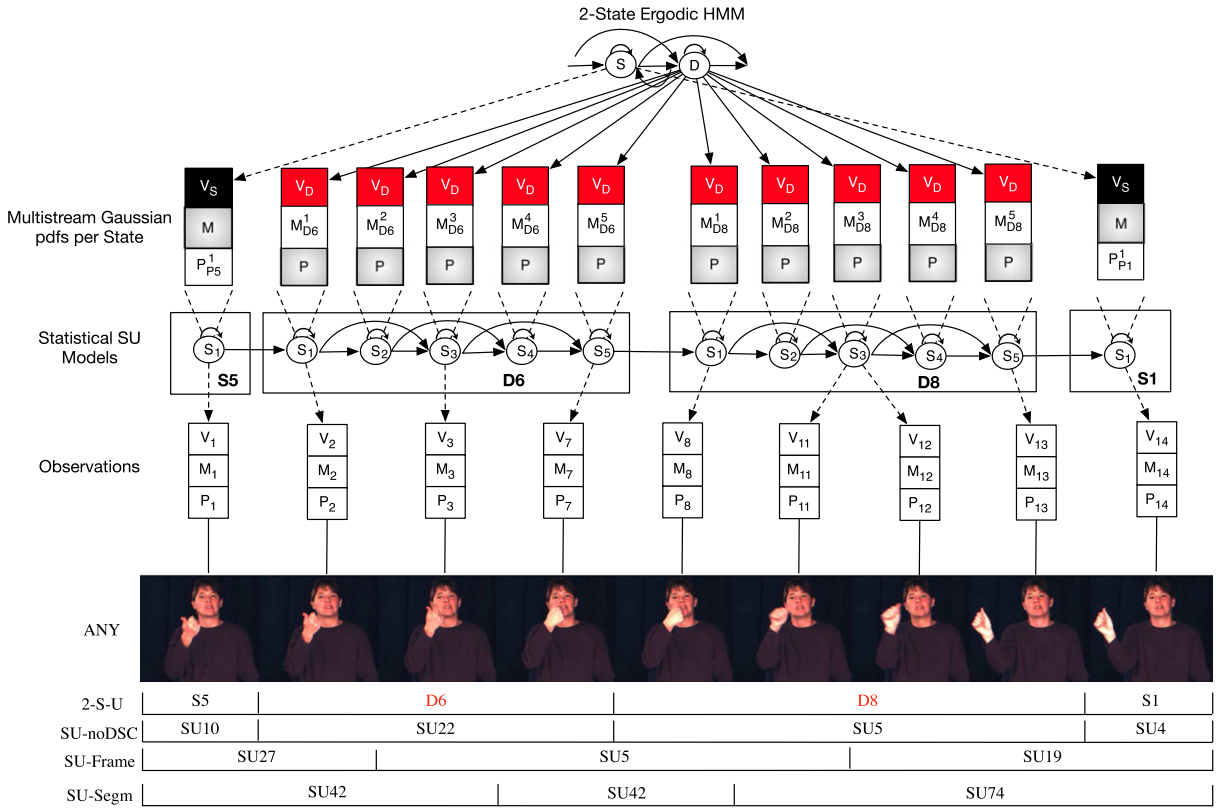

Fig. 12. D/S sequentiality and statistical subunits, for ASL sign "ANY" (ASLLVD). D/S HMM: (top) feeds the appropriate D/S pdf per HMM SU. Multistream Gaussian pdfs: the encapsulated Dynamic (V_D) and Static (V_S) ones (velocity) in different colors; shaded boxes correspond to zero stream weights. Altogether, they prescribe the D/S sequential structure. These pdfs correspond to each state of the next layer's HMMs: e.g. M^1_{D6} corresponds to the pdf for the first state of the *D6* dynamic HMM. Statistical HMM: linked in a network as prescribed by the D/S sequence in the lexicon; they construct a composite D/S network. Observations: the HMMs output the observations (features) per stream (V_i,M_i,P_i, i is the frame number), corresponding to the sequence of images for sign "ANY." Frames and segmentations (bottom): comparison of methods. See also Section 8.

sequence. The decoding time for the GSL Lemmas database is on average 0.69 × RT (RT refers to real-time)[1].

8.4. Incorporation of the non-dominant (ND) hand

The incorporation of the non-dominant hand fits the described HMM scheme. Specifically, we add in each multi-stream HMM the three extra streams: these are the velocity, the position and the movement cues of the non-dominant hand. In this way, as described next, each multi-stream HMM models both hands.

Recall as discussed in Section 5, the training of the two velocity pdfs as incorporated in the states of the ergodic HMM. The one pdf models low-velocity segments (V-L) and the other high-velocity segments (V-H). The V-L and V-H pdfs are used for initialization of the velocity streams for the SU HMMs, as follows: 1) For the *static SU models* we initialize the velocity streams of both hands employing the V-L pdf. Then we set to zero the stream weights for the movement cues and to one the stream weights for the position cues. 2) For the *dynamic SUs* that model the movements only by the dominant hand, we initialize the dominant's hand velocity stream employing the V-H pdf. In contrast, the non-dominant's hand velocity stream is initialized employing the V-L pdf. Then, we set to zero the stream weights of the position cues for both hands and the movement cue for the non-dominant hand. In addition, we set equal to one the stream weight of the movement cue for the dominant hand. 3) For the *dynamic SUs* that model movements of both hands, we initialize the velocity streams for both hands, by employing the V-H pdf. Then we set to zero the stream weights for the position cues and to one the stream weights for the movement cues. Finally, we tie the corresponding streams (movement, position) if the same SU is performed either by either hand (D or ND). By tying we refer to the sharing of the statistical parameters of the underlying pdfs; each time all models are updated.

In Fig. 13 we show an example diagram of this scheme: three dynamic SUs (D6, D8 and D6–D8) and three static SUs (S5, S1 and S5–S1) appear in the signs "ANY" and "ACCIDENT" (Figs. 9 and 12). The D6–D8 SU share distributions with D6 and D8 SUs in the dominant and non-dominant movement streams respectively, shown as Mov-D and Mov-ND. Similarly, the S5-S1 SU shares the pdf with S5 and S1 SUs in the dominant and non-dominant position streams (Pos-D and Pos-ND).

9. Lexicon: multiple signers' results & data-driven compensation of unseen pronunciations

9.1. Articulation variability

The articulation of signs is dependent on the signer, and results in variability when we consider different signers. This variability is observed for instance as follows: In signs that consist of multiple movement iterations, the number of which may vary; in signs pronounced in a compound variant; or in signs with different movement pronunciation, to list but a few. See for example the articulation of sign "QUIET" by two signers in Fig. 14. This shows an example in which Signer-A articulates it differently compared with Signer-B, by articulating an additional component. In both cases however, the sign is perceived the same. One way to address such issues is to compensate for them at the lexicon, preventing consequent recognition errors. Given the

[1] We used an AMD Opteron(tm) Processor 6386 at 2.80 GHz.

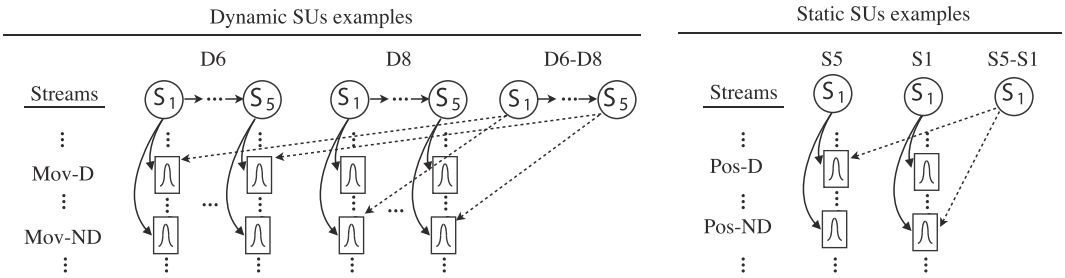

Fig. 13. Dynamic/Static SUs tying example for D and ND hand.

data-driven lexicon, we can easily face such cases, within the same framework, by generating new data-driven pronunciations. For this we employ a few development data of the unseen signer that is to be tested.

9.2. Compensating for unseen signer pronunciations

In the training phase we build SU models employing data only from the signer-A, referred to, as "training" signer. We also construct a lexicon, that contains only pronunciations from Signer-A. The average number of pronunciations per sign and signer, based on the decoded SU sequences, are 3.5. Herein our goal is to compensate the unseen pronunciations for the unseen Signer-B, referred as "test" signer. This compensation is conducted by generating new data-driven pronunciations; for these we employ a development dataset from Signer-B. To sum up, with the trained SU models (of Signer-A) and for each sign articulation in the development dataset, we find the most probable SU sequence, i.e., sign pronunciation, given the sequence of features. These new SU sequences construct a new lexicon. This lexicon fits best the way that the new test signer articulates each sign. In this way, we also highlight the differences between the pronunciations of the signers, since data are decoded with the same models. Moreover, as discussed next, by comparing the new lexical entries with the previous ones, these differ in SU substitutions, insertions and deletions in interpretable ways.

9.3. Examples of pronunciation differences between signers

In Table 2 we illustrate the sign pronunciations for three instances of GSL signs: "QUIET", "RECEPTION", and "SOMETIMES." These correspond to the lexical entries after employing the training data (Signer-A) versus the development data (Signer-B); in both cases the employed models are the ones trained only on Signer-A. By comparing the SU sequences, i.e., pronunciations, of the signers we observe the following: First, we show their difference by highlighting the corresponding SU subsequences, after applying typical pairwise sequence alignment [62], adapted for the case of SUs. These differences can be seen as implicit mappings on the alignments; Table 2 presents a few examples. Such mappings are of various types, indicating candidates responsible for variability. For instance the variation for the sign "QUIET" is represented by a substitution: {D21 S1 D16 S4 D21} → {D29} (Table 2, Fig. 14). In addition, the articulation variation for sign "SOMETIMES" in Table 2 is manifested via a difference in the number of iterative movements, which may vary. Finally, in sign "RECEPTION" the articulation of the movement is different for the two signers; compare also Fig. 1g with h.

10. Recognition experiments on BU400

Next, we employ the BU400 continuous ASL corpus [27]. We process the following six videos: Accident, Biker-Buddy, Boston-La, Football, Lapd-story, and Siblings; these contain stories narrated from a single signer, and thus the experiments of this section are signer dependent. In addition, as we do not account for inter-sign transition we use sign-level transcriptions to pre-segment the stories into separate signs. The vocabulary size is 94 signs, and the running glosses are 1202. We employ 60% of the data for training, 30% for testing, and 10% for development; all experiments employ 3-fold random selection for the train and test set, and we show finally average results. For more details on the train and test data partitioning refer to [63]. In addition, in the following experiments we take into account both the movement–position cues for both hands and the handshape cue for the dominant hand.

10.1. Other approaches

We compare 2-S-U with the following approaches: 1) The SU-noDSC is similar to 2-S-U, employing the same segmentation via a 2-state ergodic HMM. Nevertheless, it does not discriminate between dynamic and static segments. Consequently the same features are employed in each segment; then we cluster all segments employing DTW as a similarity measure. 2) The SU-Segm [14] employs a 3-state left-right HMM model for segmentation. It still accounts for whole segments as SU-noDSC and 2-S-U. In addition, for the SUs and lexicon construction we employ DTW as a similarity metric among segments to cluster them. 3) The SU-Frame [11] for SU and lexicon construction is based on frame-level clustering and segmentation without considering segments, but by applying K-means on frames. In both (2) and (3), each SU is statistically trained via HMMs whereas, there is no discrimination between D/S segments. In all competitive approaches we employed the same movement–position (M–P) cues for the dominant and non-dominant hand, and implement the modeling as in each publication. For the HS cue we use the same modeling in all SU-level approaches.

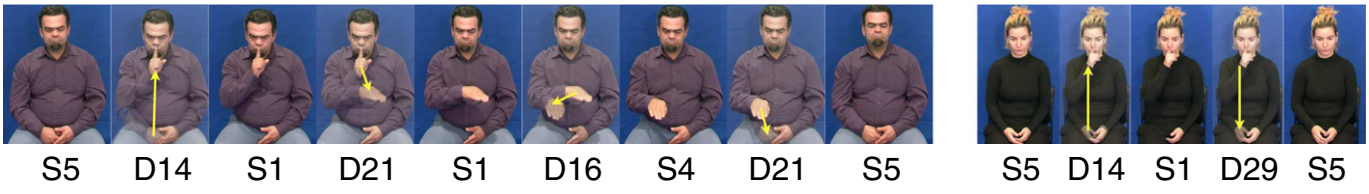

Fig. 14. Sign "QUIET" by two signers (GSL-Lem) with the SU-level decomposition. Note the pronunciation difference corresponding to the SU sequence "D21 S1 D16 S4 D21" of signer Kostas (left), with "D29" of signer Olga (right). The former articulates a supplementary movement component. See also Table 2.

Table 2
Correspondence of subunits after data-driven compensation of pronunciations between Signers-A and -B. GSL signs are "QUIET", "RECEPTION", and "SOMETIMES." After each pair of SU sequences, we show the mappings (see Map.) between the sign sub-sequences responsible for the pronunciation differences. Rightmost column contains a description (Descr.) of these differences.

Signer	Sign	Sign pronunciations	Descr. of difference
A	QUIET	S5 D14 S1 D21 S1 D16 S4 D21 S5	Extra compound sign
B	QUIET	S5 D14 S1 D29 S5	
	Map:	{D21 S1 D16 S4 D21} ⟶ {D29}	
A	RECEPTION	S5-S3 D19 D26-D26 S3-S4 D21-D27 D27-D27 D29-D29 S5-S5 D16-D16 S5-S3	Different pronunciation of movements
B	RECEPTION	S5-S3 D20-D20 S4-S2 D27-D27 D22-D22 S5-S5 D16-D16 S5-S3	
	Map:	{D19 D26-D26 S3-S4 D21-D27} ⟶ {D20-D20 S4-S2}, {D29-D29} ⟶ {D22-D22}	
A	SOMETIMES	S5-S3 D14 D8 D20 S3-S3 D28 S3-S3 D28 S3-S3 D29 S5-S3	Different number of iterations & movements' pronunciation
B	SOMETIMES	S5-S3 D14 D8 D20 D2 S4-S3 D2 S3-S3 D22 D29 S5-S3	
	Map:	⟶ {D2 S4-S3 D2}, {D28 S3-S3 D28 S3-S3} ⟶ {D22}	

The integration of M–P and HS cues for both SU-Segm and SU-Frame is done by early feature concatenation as described in [14,11]. However this leads to lower performance compared with late integration. Thus, for fair comparison we integrate them via PaHMM.

10.2. Feature notation

The features and their notation of the movement–position cues for the dominant and non-dominant hands are as follows: Direction is denoted as "D", Movement Trajectory after scale and initial-position normalization as "SPn", Scale as "S", and non-normalized Position as "P". Incorporation of multiple features (Fig. 15), is encoded by "−" (e.g., A–B). For the 2-S-U approach, A–B indicate that A cue corresponds to the dynamic segments and B to the static ones. In contrast, for all other approaches A–B cues are concatenated, and do not employ the D/S discrimination. Finally, handshape (HS) cue in all approaches is incorporated via PaHMMs and is indicated by "+HS."

10.3. Subunits' number

In the following experiments we set the number of SUs based on recognition performance on the development set that has no overlap with the test data. For 2-S-U we use 20, 30, and 110 SUs for the position, movement, and handshape cues respectively. For the SU-noDSC, SU-Segm and SU-Frame we employ 150, 100, 100 SUs for the movement–position cue respectively, and 110 SUs for the handshape cue. As we observe, for the movement–position cue the number of the SUs employed for the 2-S-U is smaller compared to the other approaches. This is due to the discrimination between dynamic and static SUs, which allows us to employ a smaller number of SUs to model the deconvolved feature space.

10.4. Comparisons with other approaches

Average results appear in the first row (with label Feat.) of Table 3. In addition, Fig. 15a shows more detailed results. The 2-S-U approach outperforms SU-noDSC while employing as features SPn-P + HS or D–P + HS. This indicates that the D/S discrimination is crucial. This concerns the employment of different, but appropriate features in the sequential segments, instead of naively combining the features. We employed the SPn or the D feature vector for the dynamic segments, and the P feature vector for the static segments. Finally, by averaging over the experiments that employ different features (see Table 3) the 2-S-U approach results on 2% increase compared with SU-Frame, 5.6% with SU-noDSC, and 8.2% with SU-Segm.

10.5. Features and combinations

Herein, we evaluate the efficacy of multiple features and their combinations. First, the importance of normalization w.r.t. to the initial position and to the movement's scale (Section 6.1) is also reflected in the recognition results. Fig. 15a shows that by employing the SPn-P +

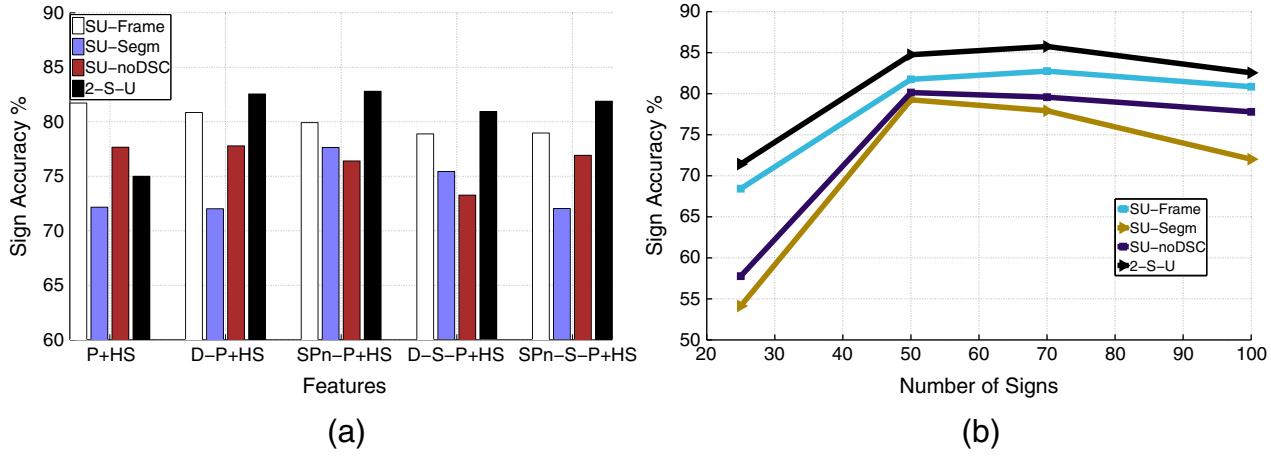

Fig. 15. Recognition experiments in BU400: a) Comparison with other approaches and feature combinations b) Variation of the number of signs.

Table 3
Overview of recognition experiments on BU400.[a]

Exp.	Method	Segm	D/S Incorp.	#G	Avg. sign acc. %
Feat.	2-S-U	2S-ERG	✓	94	82.04
	SU-noDSC	2S-ERG	✗		76.4
	SU-Segm	3S-LR	✗		73.8
	SU-Frame	✗	✗		80.06
#G	2-S-U	2S-ERG	✓	{25,50,70,94}	81.1
	SU-noDSC	2S-ERG	✗		73.8
	SU-Segm	3S-LR	✗		70.8
	SU-Frame	✗	✗		78.4

[a] Segm. refers to the HMM used in the segmentation. "2S-ERG" refers to 2-state ergodic HMM and "3S-LF" refers to 3-state left-right HMM. "Exp." corresponds to experiments that account for variation of the feature (Feat.), or of the number of glosses (#G).

HS or D–P + HS feature cue in 2-S-U we achieve higher performance than using the P + HS feature. In contrast, SU-Frame achieves the best recognition performance with the P feature vector. This cross-validates our intuition since the proposed approach is not designed to incorporate the non-normalized position. Another observation is that by employing the SPn-P + HS or D–P + HS features in 2-S-U the performance is similar (Fig. 15a). This is expected as SPn contains information of each movement's direction (Fig. 6a). Finally, by incorporating multiple cues in the dynamic modeling as shown in the 2-S-U case the accuracy is of the same order; see the cases of D-S-P + HS and SPn-S-P + HS compared to SPn-P + HS and D–P + HS respectively in Fig. 15a. Thus, in all the following experiments in Sections 11 and 12, the features employed for the M–P cues are the direction and position respectively, i.e., D–P.

10.6. Variation of the vocabulary size

An overview with average results is shown in the second row (label #G) of Table 3. Next, we compare 2-S-U with the above methods, while varying the vocabulary size. These experiments show results for the D–P + HS feature (Fig. 15b). By increasing the number of signs from 25 to 50 the recognition performance increases in all approaches. This is because more data are employed during the SU construction. Thus, the resulting SUs describe better the articulation variability of the signs. By averaging over recognition experiments for different number of signs (group of rows with label #G., Table 3) the 2-S-U results on an average absolute increase of 2.7% compared with SU-Frame, 7.3% with SU-noDSC, and 10.3% with SU-Segm.

11. Recognition experiments on GSL lemmas

Herein we present experiments taking into account both the movement–position cues for both hands and the handshape cue for the dominant hand. The evaluations contain the following scenarios: 1) Signer dependent experiments, that is, training and testing on the same signer. 2) Test on an unseen signer, that is, training on Signer A, and testing on data from a different Signer B; Signer's B data have not been employed in any way. 3) Experiments that fall in between (1) and (2). Similar to (2), we make use of models that are still unseen in terms of the trained data to the test signer. However, we allow a few development data to be employed to compensate for the unseen pronunciations of the test signer (see Section. 9).

11.1. Data, Subunits' number and feature notation

The database employed in the following experiments is the GSL Lemmas Corpus (GSL-Lem) [64]. This consists of 1046 different signs with 5 repetitions each, conducted by two native signers (referred to as "Kostas" and "Olga").

In these experiments we set the number of SUs based on maximizing recognition performance on a randomly selected development set; this contains the 20% of the data and has no overlap with the test data. For the 2-S-U we use 10 SUs for the position cue, 30 SUs for the movement cue and 500 SUs for the handshape cue. For the handshape SUs, each one models a different hand configuration together with the 3D hand orientation, since we process 2D data. For the SU-noDSC, SU-Segm, and SU-Frame we employ 150, 300, 150 SUs respectively for the movement-position cues and 500 SUs for the handshape cue.

The information cues include Movement (M), Position (P) and Handshape (HS). The M, P combination is noted with a "—" (Section 10): "M–P" indicates employment of both. HS incorporation is indicated by a "+." Thus, "M–P + HS" indicates that all cues are employed. In detail, the features employed are the non-normalized position for the position cue, the direction for movement cue, and the features of Section 4 for HS.

11.2. Other approaches

We compare 2-S-U with three SU-level approaches: SU-Segm [14], SU-Frame [11], and SU-noDSC (see Section 10.2). For the M–P case we implement the modeling as in each publication. For the HS we employ the same modeling in all SU-level approaches, as in 2-S-U. We also compare with the sign-level approach of Wang et al. 2010 (Sign-DTW) [28]: this is an exemplar-based method that constructs for each sign multiple templates. Recognition is based on similarity via DTW. All the above approaches employ the same visual features. Finally, we compare with approaches presented by Cooper et al. in [23]: these are based on *Markov Chains* (MC) and *Sequential Patterns* (SPs). For these, we report the exact recognition results presented. The authors therein employed the same vocabulary, dataset, and visual tracking output (see Section 4), and are thus directly comparable for this signer dependent experiment.

11.3. Signer dependent scenario

Herein we present signer dependent experiments on a single signer (Kostas). The vocabulary consists of 984 signs. This reduction on the number of signs is due to tracking errors in 62 of the signs, which were removed. The data are split randomly into four training examples and one test example per sign; these are kept the same for all experiments. For more details on the train/test partitioning refer to [63].

Table 4 presents the recognition results employing all information cues. The 2-S-U, SU-Segm, and SU-Frame approaches result to similar recognition performance. Furthermore, the proposed approach 2-S-U outperforms the *MC* and *SPs* [23] methods leading to 25.5% and 22.8% absolute improvements respectively. Moreover, the Sign-DTW performs 2% better than 2-S-U. Note that this is a signer dependent task. The employment of multiple signers increases articulation variation, and evaluates the generalization on unseen signers. For this, next follows a task where the test signer is unseen.

11.4. Unseen signer scenario

Herein we present results by testing on an unseen signer, that is, no data from the test signer are employed in the training. We train the SU models with all repetitions per sign from a single signer, and then test on the unseen signer. The vocabulary consists of 300 signs out of the 984 signs. This reduction on the number of signs is because of the unavailability of the hand tracking for the second signer (Olga). In Table 5 we show the sign recognition accuracy for the different cues and methods.

11.4.1. Movement–position cues

Table 5 shows that 2-S-U outperforms the SU-noDSC approach leading to an absolute improvement of 18% on average for both signers. This focused comparison, indicates that the exploitation of the D/S concept together with its multistream integration, increases sign discrimination. In addition, by comparing with the SU-Segm and SU-Frame approaches

Table 4
Signer dependent sign recognition accuracy % for multiple approaches on 984 signs from GSL-Lem.

2-S-U	SU-Frame	MC	SPs	SU-Segm	Sign-DTW
96.98	96.2	71.4	74.1	96.2	99

Table 5
Unseen signer experiments. Sign recognition accuracy % on 300 signs from GSL-Lem.

Signer	Cue	2-S-U	SU-noDSC	SU-Segm	SU-Frame	Sign-DTW
Olga	M–P	30.1	11.3	14.23	11.4	25.8
	HS	38.8	38.8	38.8	38.8	42.2
	M–P + HS	61.2	46.6	54.4	40.53	57.9
Kostas	M–P	29	11.8	9.1	11.9	24.4
	HS	28.8	28.8	28.8	28.8	32.7
	M–P + HS	50.1	33.2	32.6	35.53	46.3

the 2-S-U leads to absolute improvements of 17.8% and 17.9 respectively on average for both signers. Finally, when comparing with the sign-level approach (Sign-DTW), 2-S-U increases recognition performance by 4.5% on average for both signers.

11.4.2. Other cues

Table 5 shows that 2-S-U, SU-noDSC, SU-Segm, and SU-Frame approaches lead to the same result in the HS case. This is because of the employment of exactly the same type of modeling, since for the handshape cue we do not discriminate between D/S cases. Finally, for the case of the M–P + HS cues, 2-S-U outperforms the other approaches in the experiments of both signers. Specifically, the recognition performance increases on average for both signers as follows: 15.8% over SU-noDSC, 12.1% over SU-Segm, 17.6% over SU-Frame, and 3.5% over Sign-DTW.

11.4.3. Confusability and errors

We discuss indicative cases of confusability of some GSL signs from the above experiment. First, we focus on signs recognized correctly by 2-S-U, but incorrectly by other SU-based approaches (see also graph in Fig. 16). Methods lacking the D/S concept lead to errors as follows: 1) Signs that differ in an extra posture after a movement for instance the signs "RICE", "SAY", and "SEE" (see Figs. 1e, 10, 1f): all contain a posture in the neutral space, and are incorrectly recognized as "SWEET" that does not contain this posture. This is since there is no subunit representing explicitly the specific postures as in 2-S-U. 2) Signs that differ in an extra movement. Sign "SOUND" contains a small movement, for which there is no explicit SU for SU-noDSC, SU-Segm, and SU-Frame, and it is recognized incorrectly to the signs "TASTY", "SHINE", and "WHY", respectively. 3) D/S SUs affect also two-handed signs; in the D/S absence they can be confused to a single-handed sign that produced higher likelihood: e.g. SU-noDSC confused "AUDIENCE" with "SHINE", sharing the same handshape. Second, we also examine 2-S-U's errors. 1) Small movements, i.e., wrist rotations and finger-play are not detected. Thus signs that differ only in these are not discriminated. Take for instance the compound sign "SPORTS". In this, the first component contains a wrist rotation that is not represented in the SUs resulting on a confusion to sign "THINGS." However, the latter corresponds to the second component of the compound "SPORTS." Sign "SIXTY" appears the same, but contains fingers' movement in contrast to "SIX." 2) Same movement and similar appearing handshapes as in "STORE" vs. "WHOLE." 3) 3D information is not available, and movements are mapped in 2D: Sign "SALT", consisting of a 3D circular movement is confused to "WALKER." 4) Signs "WHAT" and "WHY" are homonyms. Fig. 16 shows other cases too.

11.5. Compensating for unseen pronunciations

As observed in the unseen signer scenario the performance decreases significantly compared with the signer dependent case. This holds for all approaches. The unseen signer scenario complements the overall evaluation by quantifying the generalization of each approach. To achieve high recognition performance the lexicon has to account for the pronunciation variation of the unseen signer, so we act as follows. Herein we evaluate the 2-S-U by compensating for unseen pronunciations of the test signer as described in Section 9. We train the SU models employing all repetitions for each sign from a single signer (Signer-A). Then we employ a development dataset from the unseen test signer (Signer-B) to generate new pronunciations. Finally, we evaluate on the rest of the still unseen test signer's data.

Table 6 shows the results in sign accuracy: in these we employ 300 signs while varying the percentage of the development set of the new signer. By employing 20% of the new signer's data, that is only one repetition per sign, the recognition performance increases 30% at least for both signers, leading to 91.1% and 86.4% for Olga and Kostas respectively. As the percentage of the development dataset increases, the performance increases too, since more pronunciations are generated, and the lexicon is implicitly adapted to the articulation variation of the test signer. This indicates that the generation of new pronunciations from the test signer can be proved beneficial when dealing with a new signer: even with a single example per sign, performance is increased significantly.

12. Recognition experiments on ASLLVD

Herein we present recognition experiments on a subset of the ASL Large Vocabulary Dictionary corpus [24]. The vocabulary consists of 97 signs with one repetition each, from two native signers (Dana, Lana). For the training of the SU models we employ a single repetition per sign from one signer. For the testing we employ the data from the other signer that is kept unseen during the training of the models.

The number of the SUs employed in each cue was set to maximize recognition performance in a randomly selected development set. This set constitutes 20% of the data, and does not overlap with the test set. We employ a 5-fold cross-validation selection for the development and test sets, and present average results. The median number of SUs employed for each cue is as follows: 10 SUs for the position cue: each one models a different Place-of-Articulation; 10 SUs for the movement

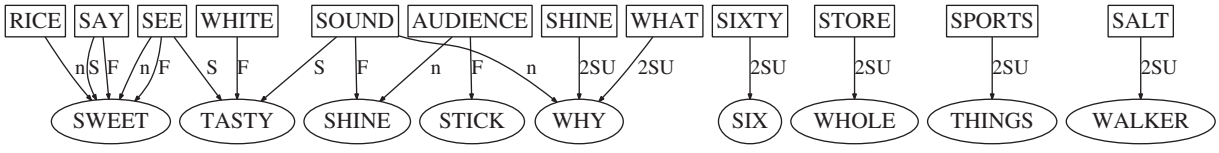

Fig. 16. Sign confusability graph. Nodes: rectangulars correspond to tested signs (transcriptions); ellipses to recognized signs. Archs: link a sign, e.g. "SEE", to the sign that is confused to, e.g. "TASTY" because of an error by one method among: SU-noDSC (n), SU-Frame (F), SU-Segm (S); these signs were recognized correctly by 2-S-U. Archs with a 2SU label show 2-S-U errors. GSL sign samples can be viewed in [26].

Table 6
Compensating for unseen pronunciations via a development set of 0–4 instances (Inst.) per sign, of the unseen signer. Sign recognition accuracy % on 300 signs from GSL-Lem.

Test signer	Olga					Kostas				
inst.	0	1	2	3	4	0	1	2	3	4
Cue										
M–P	30.1	69.2	75.3	75.8	76.2	29	64.3	72	76.2	80.5
HS	38.8	85	91.4	94	93.6	28.8	76	86	90.6	91
M–P + HS	61.2	91.1	95.5	96.1	96.6	50.1	86.4	92.6	94.33	95.66

cue: each one models a different movement; 200 SUs for the handshape cue. The features employed are the non-normalized position for the position cue, the direction for the movement cue, and the features of Sec.sec:feat for the handshape cue.

The recognition results appear in Table 7. By employing the movement–position (M–P) and handshape (HS) cues separately 2-S-U leads to an average absolute increase over both signers, of 9.3% and 4.8% respectively. After employing all cues (M–P + HS) the average absolute increase over both signers is 7.5%.

13. Conclusions and discussion

We introduce a novel computational SL phonetic modeling framework (2-S-U) of *dynamic–static* segmentation, classification and modeling for subunit construction in ASLR. Our main contribution lies on the introduction of data-driven unsupervised D/S *sequentiality* without any linguistic prior information. At the same time we preserve the parallelism of multiple cues. This is implemented via 1) the segmentation and classification into dynamic and static segments, 2) the employment of the appropriate model and *different features* in each SU type, and 3) the integration of D/S statistical SUs in a HMM framework. An important output is the intuitive data-driven lexicon: this lexicon inherits the D/S sequential structure inspired by the L&J's work. In this way the constructed lexicon is not only data-driven, but it has the phonetic property that each sign consists of sequentially stacked movement (Dynamic) and non-movement (Static) parts.

The 2-S-U approach is evaluated in ASLR experiments, on data from three different corpora and two SLs: Boston University SL corpus (BU400) with a vocabulary of 94 signs, GSL lemmas with 984 signs for the signer dependent experiments and 300 signs for unseen signer testing, and ASL Large Vocabulary Dictionary with 97 signs. In the experiments we incorporate the dominant and non-dominant hands as well as handshape. The experiments provide evaluations by employing a single training example per sign, and testing on an unseen signer. Note also that although we deal with isolated signs, we model and recognize sub-sign phonetic units. The final recognition output is evaluated at the sign-level, via the SU-level lexica. Extensive comparisons are conducted with three different SU-level approaches [14,11,23], and one sign-level approach [28]. The average over the multiple signers relative improvements w.r.t. other approaches for the GSL-Lem with unseen signer testing are as follows (300 signs): 23% for [14], 31.4% for [11] and 28.8% for SU-noDSC; the latter adds a supplementary focused comparison, and is as 2-S-U, but lacks the D/S component. The average relative improvements from [28] over multiple signers with unseen signer tests and for both GSL-Lem (300 signs) and ASLLVD (97 signs) is 9.3%.

Table 7
Unseen signer experiments. Sign recognition accuracy % with a single training example per sign on 97 signs from ALSLVD.

	Test signer	M–P	HS	M–P + HS
2-S-U	Dana	40.31	44.21	63.15
2-S-U	Lana	38.2	40.1	61.3
Sign-DTW	Dana	26.3	41	55.78
Sign-DTW	Lana	33.6	35.7	53.6

Finally, the relative improvements over Markov chains and sequential patterns of [23] for 984 signs are 26.4% and 23.6% respectively.

These results together with the intermediate qualitative discussion, validate the significance of D/S sequentiality, which increases sign recognition performance. 2-S-U's D/S sequentiality is supported by both linguistic evidence and computational phonetic modeling after the seminal works of [6] and [41,12] respectively. Moreover, the D/S results are intuitive [6,65]: movements are thought to correspond to the most sonorous parts of the signs, as the nuclei of syllables, like the vowels in speech. On the other hand, the places of articulation (positions) are of consonantal type. Thus, the incorporation in an unsupervised way of this D/S sequential structure with appropriate features in each case, and in accordance with the above concepts, as the vowel–consonant one, in ASLR is considered rather important.

The main aspects of 2-S-U can be extended. Its data-driven nature is useful in the absence of phonetic level annotations. However, future research should also incorporate linguistic–phonetic information where available; ongoing work in this direction shows promising results [18]. Other aspects, include inter-sign transitions: these are related to continuous recognition for which the statistical SUs have a great potential. Other directions concern first the application to SL cases by exploring fusion schemes in relation to the phonological structure of the involved cues, following the research on linguistic models; second, the application to more general cases concerning gesture, face, or articulators during speech production. Finally, generalization of the approach is also of interest, by means of feature selection [66]. This would allow the employment of the appropriate cues for different cases automatically, in a different scenario from the one presented. Concluding, the overall 2-S-U framework, shows the importance of accounting for unsupervised D/S sequentiality in sub-sign phonetic modeling, and is expected to affect fields such as automatic corpora processing and the study of SLs.

Acknowledgments

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

References

[1] U. Agris, J. Zieren, U. Canzler, B. Bauer, K.F. Kraiss, Recent developments in visual sign language recognition, Univ. Access Inf. Soc. 6 (2008) 323–362.
[2] W.C. Stokoe, Sign language structure, Annu. Rev. Anthropol. 9 (1980) 365–390.
[3] T. Starner, A. Pentland, Real-time American sign language recognition from video using hidden Markov models, Motion-Based Recognition, Springer, 1997, pp. 227–243.
[4] H. Cooper, B. Holt, R. Bowden, Sign language recognition, Visual Analysis of Humans, Springer, 2011, pp. 539–562.
[5] S. Ong, S. Ranganath, Automatic sign language analysis: a survey and the future beyond lexical meaning, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 873–891.
[6] S.K. Liddell, R.E. Johnson, American sign language: the phonological base, Sign Lang. Stud. 64 (1989) 195–277.
[7] G. Coulter, On the nature of asl as a monosyllabic language, Annual Meeting of the Linguistic Society of America, San Diego, CA, 1982.
[8] E. Klima, U. Bellugi, The signs of language, Harvard Univ. Press, 1979.
[9] D. Corina, W. Sandler, On the nature of phonological structure in sign language, Phonology 10 (2008) 165–207.
[10] W. Sandler, Sequentiality and simultaneity in American Sign Language phonology, (Ph.D. thesis) Univ. of Texas, Austin, 1987.
[11] B. Bauer, K.F. Kraiss, Towards an automatic sign language recognition system using subunits, Proc. of Int'l Gesture Workshop, vol. 2298, 2001, pp. 64–75.
[12] C. Vogler, D. Metaxas, A framework for recognizing the simultaneous aspects of american sign language, Comput. Vis. Image Underst. 81 (2001) 358.
[13] W. Kong, S. Ranganath, Sign language phoneme transcription with rule-based hand trajectory segmentation, J. Signal Process. Syst. 59 (2010) 211–222.
[14] G. Fang, X. Gao, W. Gao, Y. Chen, A novel approach to automatically extracting basic units from chinese sign language, Proc. Int'l Conf. on Pattern Recognition, vol. 4, 2004, pp. 454–457.
[15] S. Prillwitz, R. Leven, H. Zienert, R. Zienert, T. Hanke, J. Henning, HamNoSys. Version 2.0, Int'l Studies on SL and Communication of the Deaf, 7, 1989, pp. 225–231.
[16] V. Sutton, Sign writing, Deaf Action Committee (DAC), 2000.
[17] Multilingual Sign Language Dictionary, [Online] http://www.signbank.org/signpuddle2.0 (Accessed 12 Nov. 2013).
[18] V. Pitsikalis, S. Theodorakis, C. Vogler, P. Maragos, Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition, IEEE CVPR Wksp on Gesture Recognition, 2011.

[19] O. Koller, H. Ney, R. Bowden, May the force be with you: force-aligned sign writing for automatic subunit annotation of corpora, Int'l Conf. on Automatic Face & Gesture Recognition, 2013.
[20] T. Kadir, R. Bowden, E.J. Ong, A. Zisserman, Minimal training, large lexicon, unconstrained sign language recognition, Proc. British Machine Vision Conference, 2004.
[21] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, M. Brady, A linguistic feature vector for the visual interpretation of sign language, Proc. European Conf. on Computer Vision, 2004.
[22] J. Han, G. Awad, A. Sutherland, Modelling and segmenting subunits for sign language recognition based on hand motion analysis, Pattern Recogn. Lett. 30 (2009) 623–633.
[23] H. Cooper, E. Ong, N. Pugeault, R. Bowden, Sign language recognition using sub-units, J. Mach. Learn. Res. 13 (2012) 2205–2231.
[24] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, A. Thangali, The american sign language lexicon video dataset, Proc. Computer Vision and Pattern Recognition Wksp, 2008, pp. 1–8, (IEEE).
[25] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, A. Thangali, H. Wang, Q. Yuan, Large lexicon project: American sign language video corpus and sign language indexing/retrieval algorithms, Proc. of Wksp on Representation and Processing of SL: Corp. and SLT, 2010.
[26] Dicta-Sign Language Resources, Greek sign language corpus, [Online] http://www.sign-lang.uni-hamburg.de/dicta-sign/portal 2012 (Accessed 2 May. 2012).
[27] C. Neidle, C. Vogler, A new web interface to facilitate access to corpora: development of the ASLLRP Data Access Interface, Proc. of 5th Wksp on Representation and Processing of SL: Interactions between Corpus and Lexicon, 2012.
[28] H. Wang, A. Stefan, S. Moradi, V. Athitsos, C. Neidle, F. Kamangar, A system for large vocabulary sign search, Proc. ECCV Wksp on Sign, Gesture and Activity, vol. 1, 2010, (IEEE).
[29] J. Zieren, K.-F. Kraiss, Robust person-independent visual sign language recognition, Patter Recognit. Image Anal. (2005) 333–355.
[30] C. Wah Ng, S. Ranganath, Real-time gesture recognition system and application, Image Vis. Comput. 20 (2002) 993–1007.
[31] G. Fang, W. Gao, X. Chen, C. Wang, J. Ma, Signer-independent continuous sign language recognition based on SRN/HMM, Proc. Gesture and Sign Language in HCI, 2002, pp. 163–197.
[32] O. Aran, L. Akarun, A multi-class classification strategy for fisher scores: application to signer independent sign language recognition, Pattern Recog. 43 (2010) 1776–1788.
[33] P. Yin, T. Starner, H. Hamilton, I. Essa, J. Rehg, Learning the basic units in American Sign Language using discriminative segmental feature selection, Int'l Conf. on Acoustics, Speech and, Signal Processing, 2009, pp. 4757–4760.
[34] C. Vogler, D. Metaxas, Adapting hidden Markov models for asl recognition by using three-dimensional computer vision methods, Proc. Int'l Conf. on System, Man and Cybernetics, vol. 1, 1997, pp. 156–161.
[35] Y. Gweth, C. Plahl, H. Ney, Enhanced continuous sign language recognition using pca and neural network features, Computer Vision and Pattern Recognition Wksp, IEEE, 2012, pp. 55–60.
[36] S. Ong, S. Ranganath, A new probabilistic model for recognizing signs with systematic modulations, Int'l Conf. on Analysis and modeling of faces and gestures, 2007, pp. 16–30.
[37] L. Ding, A. Martinez, Modelling and recognition of the linguistic components in american sign language, Image Vision Comput. 27 (2009) 1826–1844.
[38] G. Fang, W. Gao, D. Zhao, Large-vocabulary continuous sign language recognition based on transition-movement models, IEEE Trans. Syst. Man Cybern. A 37 (2007) 1–9.
[39] A. Roussos, S. Theodorakis, V. Pitsikalis, P. Maragos, Hand tracking and affine shape–appearance handshape sub-units in continuous sign language recognition, Proc. ECCV Wksp on Sign, Gesture and Activity, 2010.
[40] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–286.
[41] C. Vogler, D. Metaxas, Toward scalability in ASL recognition: breaking down signs into phonemes, Gesture-Based Comm. in HCI, 1999. 211–224.
[42] J. Lichtenauer, E. Hendriks, M. Reinders, Sign language recognition by combining statistical dtw and independent classification, IEEE Trans. Pattern Analysis and Machine Intelligence 30 (2008) 2040.
[43] R. Yang, S. Sarkar, Detecting coarticulation in sign language using conditional random fields, Proc. Int'l Conf. on Pattern Recognition, vol. 2, 2006, pp. 108–112, (IEEE).
[44] H. Yang, S. Sclaroff, S.-W. Lee, Sign language spotting with a threshold model based on conditional randomfields, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 1264–1277.
[45] P. Buehler, M. Everingham, A. Zisserman, Learning sign language by watching TV (using weakly aligned subtitles), Proc. Conf. on Computer Vision & Pattern Recognition, 2009, pp. 2961–2968.
[46] S. Nayak, K. Duncan, S. Sarkar, B. Loeding, Finding recurrent patterns from continuous sign language sentences for automated extraction of signs, J. Mach. Learn. Res. 13 (2012) 2589–2615.
[47] K.G. Derpanis, R.P. Wildes, J.K. Tsotsos, Definition and recovery of kinematic features for recognition of American sign language movements, Image Vis. Comput. 26 (2008) 1650–1662.
[48] G. Awad, J. Han, A. Sutherland, Novel boosting framework for subunit-based sign language recognition, Proc. Int'l Conf. on Image Processing, IEEE, 2009, pp. 2729–2732.
[49] J. Ajmera, C. Wooters, A robust speaker clustering algorithm, IEEEWksp on Automatic Speech Recognition and Understanding, 2003, pp. 411–416, (IEEE).
[50] I. Cohen, A. Garg, T.S. Huang, et al., Emotion recognition from facial expressions using multilevel hmm, vol. 2, NIPS, 2000.
[51] H.-K. Lee, J.-H. Kim, An HMM-based threshold model approach for gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell. 21 (1999) 961–973.
[52] L. Xie, S.-F. Chang, A. Divakaran, H. Sun, Unsupervised discovery of multilevel statistical video structures using hierarchical hidden Markov models, Proc. Int'l Conf. on Multimedia and Expo, vol. 3, 2003, (IEEE).
[53] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, Modeling individual and group actions in meetings with layered hmms, IEEE Trans. Multimedia 8 (2006) 509–520.
[54] S. Theodorakis, V. Pitsikalis, I. Rodomagoulakis, P. Maragos, Recognition with raw canonical phonetic movement and handshape subunits on videos of continuous sign language, Proc. Int'l Conf. on Image Processing, 2012.
[55] S. Theodorakis, P. Pitsikalis, P. Maragos, Model-level data-driven sub-units for signs in videos of continuous sign language, Int'l Conf. on Acoustics, Speech and Signal Processing, 2010.
[56] V. Pitsikalis, S. Theodorakis, P. Maragos, Data-driven sub-units and modeling structure for continuous sign language recognition with multiple cues, Proc. of Wksp on Representation and Processing of SL: Corp. and SLT", 2010.
[57] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2 (2006) 2169–2178.
[58] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[59] A. Vedaldi, A. Zisserman, Efficient additive kernels via explicit feature maps, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 480–492.
[60] J. Ward, H. Joe, Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963) 236.
[61] P. Gerasimos, C. Neti, G. Gravier, A. Garg, A.W. Senior, Recent advances in the automatic recognition of audiovisual speech, Proc. IEEE 91 (2003) 1306–1326.
[62] E. Myers, An O(ND) difference algorithm and its variations, Algorithmica 1 (1986) 251–266.
[63] S. Theodorakis, V. Pitsikalis, P. Maragos, Experiments' data reference webpage, [Online] http://cvsp.cs.ntua.gr/research/sign/2su 2013 (Accessed 12 Nov. 2013).
[64] Dicta-Sign Project, Corpus annotations, [Online] http://www.dictasign.eu 2012 (Accessed 2 May 2012]).
[65] D. Brentari, Modality differences in sign language phonology and morphophonemics, Modality and structure in signed and spoken languages, 2002. 35–64.
[66] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.