IDIAP RESEARCH REPORT

DATA-DRIVEN MOVEMENT SUBUNIT EXTRACTION FROM SKELETON INFORMATION FOR MODELING SIGNS AND GESTURES

Sandrine Tornay Marzieh Razavi
Mathew Magimai.-Doss

Idiap-RR-02-2019

FEBRUARY 2019

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Data-Driven Movement Subunit Extraction from Skeleton Information for Modeling Signs and Gestures

Sandrine Tornay, Marzieh Razavi and Mathew Magimai.-Doss

Abstract

Sequence modeling for signs and gestures is an open research problem. In that direction, there is a sustained effort towards modeling signs and gestures as a sequence of subunits. In this paper, we develop a novel approach to infer movement subunits in a data-driven manner to model signs and gestures in the framework of hidden Markov models (HMM) given the skeleton information. This approach involves: (a) representation of position and movement information with measurement of hand positions relative to body parts (head, shoulders, hips); (b) modeling these features to infer a sign-specific left-to-right HMM; and (c) clustering the HMM states to infer states or subunits that are shared across signs and updating the HMM topology of signs. We investigate the application of the proposed approach on sign and gesture recognition tasks, specifically on Turkish signs HospiSign database and Italian gestures Chalearn 2014 task. On both databases, our studies show that, while yielding competitive systems, the proposed approach leads to a shared movement subunit representation that maintains discrimination across signs and gestures.

1 Introduction

Sign language (SL) is a visual mode of communication for the Deaf community. To convey information, SL uses multiple visual channels such as hand gestures (hand shape, location and movement), facial expression, body posture, lip movement. In order to develop efficient sign language processing systems, it is desirable to model signs as a sequence of subunits, akin to phoneme- or phone-based speech processing [5]. As subunits allow robust parameter estimation as well as can remove the constraint that all signs in the lexicon needs to be observed during training. Furthermore, subunits can allow data sharing across languages [22]. The multistream nature of SL implies that the development of such a subunit set is a highly challenging task.

The main language structure of the SL lies in the hand gesture supplemented with non-manual components (facial expression, lip movements, body posture). Thus, most of the studies focus on the manual components i.e. hand gesture. There exist two linguistic-oriented approaches to define hand gesture: the Stokoe system [29] and the Movement-Hold model [24]. In the Stokoe system, a sign is described as a simultaneous series of three major formational units: hand shape, locations and movements, while the Movement-Hold model fragments the signs in two types of sequentially ordered segments: movement and hold (location) segments. In both approaches, the movement is a relevant structure in the sign production that needs to be modeled. In

the literature, it is well understood that the hand shape information can be modeled as a sequence of subunits based on HamNoSys [7, 17, 23]. However, the continuous aspect of the movement makes modeling of movement information as subunits difficult. The focus of this paper lies in automatic derivation of movement subunits for sign language and gesture processing. In the literature, there are two strands of research in that direction.

The first strand of research makes the assumption that some annotation of signs is available. Pitsikalis et al. [26] incorporated phonetic transcription into data-driven subunits. They first converted HamNoSys symbols into Posture-Detention-Transition-Steady Shift (PDTS) model. Then they combined these structured sequences of labels with visual tracking features for timing information via an HMM-based system to obtain the phonetic subunits. Cooper et al. [10] used hand labeled data and compared three types of subunits: appearance-based, 2D tracking-based and 3D-tracking based. Two sign-level classifiers were tested: an HMM-based approach and the sequential pattern (SP) boosting. Koller et al. [21] used gloss annotations and gloss time boundaries to generate sequences of subunits using HMM-based modeling and expectation-maximization algorithm. Elakkiya and Selvamani [11] extracted manual and non-manual features by using Parallel HMMs and introduced a novel Bayesian Parallel HMM (BPaHMM) to combine the visual and linguistic transcriptions of the sign lexicon to form a subunit gesture base.

The second strand of research involves extraction of subunits without using annotation information. In this case, subunits extraction typically involves unsupervised segmentation and clustering. There exist two lines of thoughts based on the order in which segmentation and clustering steps are carried out, i.e.,

1. clustering followed by segmentation: Bauer and Kraiss [3] used k-means algorithm to cluster the data where each cluster is then represented as a fenonic baseform. Temporal structure is then achieved with the HMM-based structure defined based on this fenonoic model [18]. Han et al. [2,15,16] used hand motion speed and trajectory to locate subunit boundaries and then temporal clustering by Dynamic Time Warping (DTW) is adopted to merge similar subunits.

2. segmentation followed by clustering: Sako and Kitamura [28] extracted depend subunits by training a multi-stream isolated sign HMM for each word where the feature vector of each frame is splited into three phonetic stream, and by clustering each state of the multistream using an inter-state distance with a tree based algorithm in order to tie the states. Fang et al. [13] segmented signs using HMMs in which each state represents one segment. Then they used a temporal clustering algorithm based on modified k-means algorithm where DTW is employed as the distance computation criterion. In that study, CyberGloves and Pohelmus 3SPACE-position trackers were used. Based only on simple position measurements obtained from video, Theodorakis et al. [30] used, as an initial segmentation step, the model based segmentation proposed by [13], and then employed a hierarchical clustering of whole dynamic models (HMMs) to find the shared segments.

In this paper, we propose a novel HMM-based approach to extract movement data-driven subunits from skeleton information for sign and gesture modeling. In this approach, no prior knowledge of the number of subunits or segmentation or linguistic annotation is used. The approach involves: (a) extraction of position and movement features from 3D skeleton information that also incorporate information related to head,

shoulders and hips; (b) inferring a left-to-right HMM for each sign by modeling the position and movement features; and finally (c) clustering the states of the HMMs across the signs through a measure of discrimination to infer subunits and representing each sign in terms of those subunits. We demonstrate the potential of the approach through sign language/gesture recognition studies on two databases: (i) HospiSign database, which contains Turkish phrases produced by native or early signers in a controlled environment and (ii) Chalearn14 database, which contains Italian gestures that do not have well defined linguistic structures akin to sign language and are produced in an uncontrolled environment.

The remainder of the paper is organized as follow: Section 2 provides a brief overview on HMM-based approach for recognizing signs and gestures. Section 3 presents the proposed approach for subunit extraction. Section 4 presents the experimental setup and Section 5 presents the results and analysis. Finally, Section 6 concludes with discussion and directions for future research.

2 HMM-Based Sign Language/Gesture Recognition

In this section we provide a short background on the HMM-based framework for sign language/gesture recognition, which also forms the basis for the subunits extraction. We present the framework in the context of sign language, while keeping the model general enough to be applicable to the gesture recognition task as well.

In the statistical sign language recognition (SLR) approach, given an input video as a sequence of images/features $X = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$, the goal is to obtain the most likely sign (in the case of isolated SLR) or sign sequence (in the case of continuous SLR) S^*,

$$S^* = \arg\max_{S \in \mathscr{S}} P(S|X, \Theta) \ , \tag{1}$$

where $\mathscr{S}$ denotes the set of all possible signs or sign sequences, S represents a sign or sign sequence and Θ denotes the set of parameters of the system. For simplicity, in the remainder of this section Θ is dropped. As direct estimation of $P(S|X)$ is a non-trivial task[1], typically Bayes' rule is applied, leading to,

$$S^* = \arg\max_{S \in \mathscr{S}} \frac{p(X|S)P(S)}{p(X)} \ , \tag{2}$$

$$= \arg\max_{S \in \mathscr{S}} p(X|S)P(S) \ . \tag{3}$$

Equation (3) is obtained as a result of the assumption that $p(X)$ does not affect the optimization. $P(S)$ is referred to as the language model, and can be estimated based on the relative frequency of the signs on the training data. A common way to model $p(X|S)$ in the literature is to use HMMs [25]. HMM is a well-known method to handle temporal pattern recognition. Furthermore its ability to compensate time and amplitude variations is valuable in sign recognition.

More precisely $p(X|S)$ in an HMM-based framework can be estimated by summing

[1] It is worth mentioning that recently there are approaches emerging which directly model $P(S|X)$ [7, 14].

over all possible state sequences $\mathcal{Q}$, i.e.,

$$p(X|S) = \sum_{Q \in \mathcal{Q}} p(X, Q|S) \ , \tag{4}$$

$$= \sum_{Q \in \mathcal{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t) P(q_t|q_{t-1}) \ , \tag{5}$$

$$\approx \max_{Q \in \mathcal{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t) P(q_t|q_{t-1}) \ , \tag{6}$$

where $Q = (q_1, \cdots, q_t, \cdots, q_T)$ denotes a sequence of HMM states. Equation (5) is obtained by making i.i.d. and first order Markov assumptions. Equation (6) is obtained by applying the Viterbi approximation. The so-called local emission score $p(\mathbf{x}_t|q_t)$ can be estimated using different techniques. In this paper, we exploit using Gaussian mixture models (GMMs) and artificial neural networks (ANNs) to estimate the emission score. The approach using GMMs is referred to as HMM/GMM approach [27], and the approach using ANNs is referred to as hybrid HMM/ANN approach [6, 22, 23, 31].

In the HMM/GMM approach,

$$p(\mathbf{x}_t|q_t) = \sum_{n=1}^{N} c_n \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \ , \tag{7}$$

where N denotes the number of Gaussian components per mixture for each state; c_n, $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ denote respectively the mixture weight, mean and covariance for the n^{th} Gaussian modeling the state.

In the hybrid HMM/ANN approach, an ANN is used to estimate the posterior probability $P(q_t|\mathbf{x}_t)$. The posterior probabilities are then converted to scaled-likelihoods (sl) of HMM states and are used as local emission score, i.e.,

$$p_{sl}(\mathbf{x}_t|q_t) = \frac{P(\mathbf{x}_t|q_t)}{p(\mathbf{x}_t)} = \frac{P(q_t|\mathbf{x}_t)}{P(q_t)} \ . \tag{8}$$

3 Proposed Approach

The proposed approach consists of three steps: (1) extraction of features based on skeleton information (Section 3.1), (2) inference of a sign-specific hidden Markov model (Section 3.2), and (3) inference of subunits by clustering HMM states across signs (Section 3.3).

3.1 Feature Extraction

In this paper, we focus on two parameters characterizing a sign: the hand location and the hand motion. To represent these manual features, inspired from [1], we decided to use continuous position and velocity features. Position features are given by the 3D coordinate of a human skeleton and velocity features are delta features computed on them. Other skeleton joints such as head, neck, shoulders and hips are used for scale normalization and also in order to have the relative position of the hands with respect to the signer's body, as described in more detail below.

For each frame t, we first normalize position features of the left and right hand, $\mathbf{p}_t$, by the width of the head. Then three types of 3D coordinate of the hands are

recalculated depending on three coordinate systems. The first one takes the head as the center; the second one uses the right shoulder as the center for the right hand, and uses the left shoulder as the center for the left hand; and the third one takes the right hip as the center for the right hand, and takes the left hip as the center for the left hand. Therefore, depending on the center **C**, the position feature $\mathbf{p}_t$ would be:

$$\mathbf{p}_t^{\mathbf{C}} := \frac{\mathbf{hand} - \mathbf{C}}{|head_y - neck_y|/4} , \tag{9}$$

where $\mathbf{C} \in \{\mathbf{head}, \mathbf{shoulder}, \mathbf{hip}\}$; **hand**, **shoulder**, **hip** are vectors of x, y, z coordinates of respectively left and right hand, shoulder and hip; and **head** contains x, y, z coordinate of the head twice. $head_y$, $neck_y$ are y coordinate of the head and neck respectively.

The velocity features, $\mathbf{v}_t^{\mathbf{C}}$, are estimated for each coordinate system by computing the difference between the position features at time t and time $t-2$.

$$\mathbf{v}_t^{\mathbf{C}} := \mathbf{p}_t^{\mathbf{C}} - \mathbf{p}_{t-2}^{\mathbf{C}} . \tag{10}$$

The resulting features are of size 36: 18 positions features—(3 left + 3 right hand position features) $\times$ 3 coordinate systems—and 18 velocity features.

3.2 Inference of Sign-Specific HMM

In this step, given the training data and cross validation data, we obtain a sign-specific HMM for each sign. This step is akin to getting a segmentation model for each sign. We investigate two methods to obtain sign specific HMMs, namely, CV-based method and train-based method.

CV-based method: In CV-based approach, first a range of possible number of states is setted: $[N_{min}, N_{max}]$. Then an HMM, $\mathscr{M}_n^{S^m}$, with n states is modeled for each sign $S^m \in \{S^1 \cdots S^M\}$, $\forall n \in [N_{min}, N_{max}]$. The emission distribution for each state is modeled by a single Gaussian distribution with diagonal covariance. Finally, from the set of HMMs for each sign, the HMM that yields the best performance for a sign on the cross validation data set is selected as the sign specific HMM for that sign. Fig. 1 illustrates this process. Given the inferred HMM with N^{S^m} states for each sign S^m, the resulting sign model is a sequence of Gaussian distributions, $\mathscr{N}_g^{S^m}$, $\forall g \in [1, N^{S^m}]$.

Train-based method: An alternative method is to obtain a segmentation model only based on the training data. We refer to this approach as the train-based approach. In this approach, the number of states, N, is common across all the signs and is defined

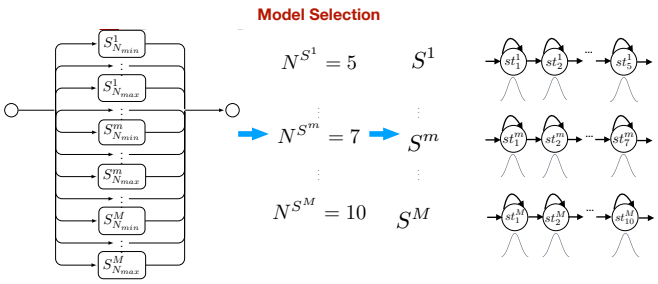

Figure 1: Illustration of the CV-based sign-specific HMM inference

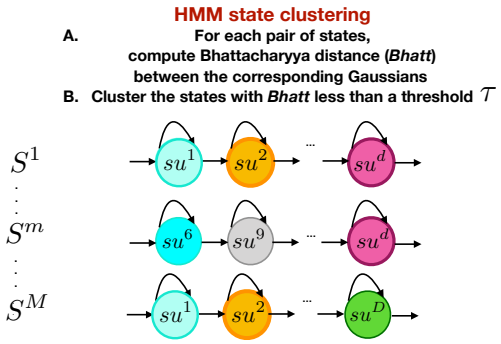

Figure 2: Clustering the HMM states based on the Bhattacharyya distance between the corresponding Gaussians. The clustered states are shown with the same color

such that the recognition accuracy saturates on the training data. This choice ensures that the states represent minimum discriminative segments. Again, similar to CV-based approach, the emission distribution of each state is modeled by a single Gaussian distribution with diagonal covariance.

3.3 Subunit Inference and Lexicon Development

Given the sign-specific HMMs and their parameters, in this step the HMM states are clustered through a measure of discrimination. More precisely, this is done by pooling all the single Gaussians of all the HMM states of all the signs and computing Bhattacharyya distance [4, 20] between each pair of Gaussian distributions:

$$\begin{aligned} Bhatt(\mathcal{N}_1,\mathcal{N}_2) = &\frac{1}{8}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\ &+\frac{1}{2}\ln\left(\frac{\det\boldsymbol{\Sigma}}{\sqrt{\det\boldsymbol{\Sigma}_1\det\boldsymbol{\Sigma}_2}}\right), \end{aligned} \tag{11}$$

where $\mathcal{N}_1 := \mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2 := \mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ are two Gaussian distributions and $\boldsymbol{\Sigma} := \frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}$. Two states are considered to be the same if the Bhattacharyya distance between their Gaussian distributions is lower than a threshold τ, see Fig. 2. The intuitive explanation is that two segments or HMM states are modeling similar movement information if the probability density functions (pdfs) of those states are similar. Since the exploited features are based on position and velocity of both hands, we can expect the clusters to describe the set of movement subunits. Furthermore, our approach of extracting subunits using skeleton information is theoretically consistent at the measurement level as well as the model level. More precisely, the measurement and the feature space are in the 3D coordinate system and the Bhattacharyya distance used to cluster the HMM states has a geometric similarity measure interpretation [4].

Given the clustered states as subunits, a lexicon can be generated in which each sign is presented as a sequence of subunits, as illustrated in Fig. 3.

The subunit based lexicon can then be used to train an HMM, $\mathcal{M}^{S^m}_{subunit}$, for each sign S^m, as illustrated in Fig. 4 for the case where the emission distributions are modeled by GMM.

In this step, the only hyper-parameter is the threshold τ. We show that this can be determined in a cross-validation manner by: (i) getting a subunit based lexicon for

different values of $\tau \in (0, 2.7]$; (ii) training a HMM/GMM SLR system for each of those subunit based lexicons; and finally (iii) selecting the lexicon that yields the best SLR system on the development data. As a by-product, this ensures that the subunit inference process maintains the discrimination between signs.

4 Experimental Setup

In this paper, we apply the proposed approach on signer-independent SLR and gesture recognition tasks to investigate its potential. In this section, we first describe the databases used for evaluating our proposed approach. We then present the setup for the recognition systems used in our studies.

4.1 Databases

We evaluated the proposed approach on two databases: (1) HospiSign database, which contains phrases produced by native signers, and (2) Chalearn14 database, which contains various gestures produced by non-experts.

4.1.1 HospiSign Database

The HospiSign database is a subset of 33 phrase classes of the continuous Bosphorus-Sign database [9]. The content is Turkish Sign Language (TSL) related to the health domain. The HospiSign subset includes 6 signers, with each sign being repeated approximately 6 times by each signer. The database is publicly available by request from the authors (`https://www.cmpe.boun.edu.tr/pilab/BosphorusSign/home_en.html`). The database has been recorded with a Kinect camera. We have used the skeletal joint coordinates that are provided in the database as the basis for our feature extraction.

In order to conduct a signer-independent experiment, we have used a leave-one-signer out cross-validation study. Furthermore, as we need a development set for tuning the hyper-parameters, we have left another signer out for this purpose. Therefore, as can be seen from Table 1, we have conducted six experiments where in each experiment, one signer is used for testing, one signer is used as the development set, and the rest of the signers are used for training. For each experiment, we have presented the average performance over the signers as the final result. Table 2 presents the average number of samples in the train, development and test sets over the six experiments.

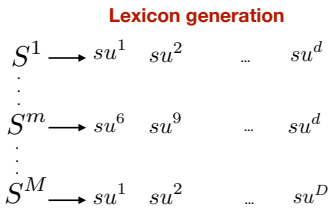

Figure 3: The generated subunit-based lexicon based on the clustered states obtained in Fig. 2

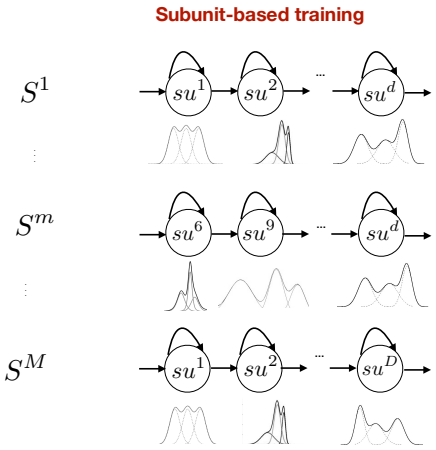

Figure 4: Subunit-based sequence modeling

Table 1: The HospiSign database segmentation of training, development and testing data according to signers. The numbers in the table refer to the signers's number

	Exp 1					Exp 2					Exp 3					Exp 4					Exp 5					Exp 6			
Train	3,4, 5,6	2,4, 5,6	2,3, 5,6	2,3, 4,6	2,3, 4,5	3,4, 5,6	1,4, 5,6	1,3, 5,6	1,3, 4,6	1,3, 4,5	2,4, 5,6	1,4, 5,6	1,2, 5,6	1,2, 4,6	1,2, 4,5	2,3, 5,6	1,3, 5,6	1,2, 5,6	1,2 ,3,6	1,2, 3,5	2,3, 4,6	1,3, 4,6	1,2, 4,6	1,2, 3,6	1,2, 3,4	2,3, 4,5	1,3, 4,5	1,2, 4,5
Dev	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4	5	5	5	5	5	6	6	6
Test	2	3	4	5	6	1	3	4	5	6	1	2	4	5	6	1	2	3	5	6	1	2	3	4	6	1	2	3

Table 2: Description of the HospiSign and Chalearn14 databases in terms of average number of samples

	Train	Dev	Test
# samples, HospiSign	874	210	210
# samples, Charlearn14	6800	2506	3579

4.1.2 Chalearn14 Database

The Chalearn14 database consists of isolated gestures drawn from 20 Italian gestures and performed by several different users. The data was recorded in the context of the Task 3 of the Chalearn challenge of 2014. The database is publicly available by request from the authors (`http://gesture.chalearn.org/2014-looking-at-people-challenge`).

For efficient comparison to the existing results we used the train/development/test setups given by the competition for the Chalearn14 database. Table 2 presents the number of samples in the train, development and test sets. It is worth mentioning that in the Chalearn 2014 competition, the segmentation of videos in the test set was not provided. Therefore the task contained two parts: (1) segmentation of videos into gestures/non-gestures, and (2) classifying the gestures. As our focus in this paper is on classifying the gestures, we have used the ground truth segmentations on the test set.

4.2 Systems

We built HMM/GMM and hybrid HMM/ANN systems using sign-specific sequence modeling approach and the subunit-based sequence modeling approach. The HMMs were trained and tested with the HTK toolkit [32] adapted for sign language. For a

better segmentation, i.e. to avoid sign irrelevant movement being taken into account in the sign, we added a transition model at the beginning and end of each HMM. For preserving the continuity of the entire model, we modeled the transition model as a three-states left-to-right HMM. The three-states transition HMM structure with one state skip is presented in Fig. 5.

4.2.1 HMM/GMM Systems

In the case of sign-specific sequence modeling, we modeled the signs with left-to-right HMMs using Gaussian state-output distributions. The number of Gaussian components per mixture varies between 1 and 56. We used the two segmentation approaches explained in Section 3.2 to determine the number of HMM states per sign. In the CV-based framework, the average number of HMM states per sign was 8 for the HospiSign database and 3 for the Chalearn14 database. In the train-based framework, the derived number of HMM states per sign was 9 for both databases.

In the subunit-based model, we trained HMM/GMM systems where each subunit was modeled with a single HMM state. The number of Gaussian components per mixture also varied between 1 to 56, and was set based on the recognition accuracy on the development set.

4.2.2 Hybrid HMM/ANN Systems

For building the hybrid HMM/ANN systems, we first obtained the alignments in terms of the HMM states using the trained HMM/GMM systems. We then trained ANNs, more precisely multilayer perceptrons (MLPs) classifying HMM states with output non-linearity of softmax and minimum cross-entropy error criterion, using Quicknet software [19]. We used 36-dimensional position and velocity features with four frames preceding context and four frames following context as the MLP input. In our experiments we trained MLPs with different number of hidden units (600, 800, 1000) and hidden layers (0, 1, 2, 3). The number of hidden units and hidden layers as well as other hyper-parameters such as learning rate and the batch size were chosen according to the frame-level accuracy on the development set.

We estimated the scaled likelihoods in the hybrid HMM/ANN systems by dividing the posterior probabilities derived from MLPs with the prior probabilities of the classes estimated from relative frequencies in the training data. These scaled likelihoods were then used as emission probabilities for HMM states.

The performance of the developed systems are evaluated in terms of recognition accuracy (RA):

$$RA = \frac{\#\ of\ correctly\ predicted\ signs/gestures}{total\ \#\ of\ signs/gestures\ in\ the\ reference}\ . \tag{12}$$

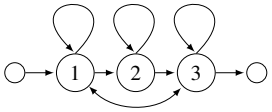

Figure 5: Structure of the three-states transition model

5 Results and Analysis

In this section, we first present the recognition results on the HospiSign and Chalearn14 databases. We then contrast the performance of the proposed approach with the existing approaches in the literature.

5.1 SLR on HospiSign

Table 3 presents the HMM/GMM SLR system results in terms of *RA* on the HospiSign database. It can be observed that the subunit-based modeling approach leads to development of a comparable SLR system to the sign-specific sequence modeling approach. This is interesting as the number of states per sign in the CV-based sign-specific system is optimized according to the segmentation approach explained in Section 3.2, and the train-based subunit-based system is able to perform comparable to this system despite considerably reducing the total number of states by in average 22%.

Table 3: HMM/GMM results on the HospiSign database depending on the segmentation approach explained in Section 3.2 (train-based and CV-based)

Experiment	sign-specific seq. modeling		subunit-based seq. modeling	
	Train-based	CV-based	Train-based	CV-based
Exp 1	92.56	91.95	91.95	90.75
Exp 2	89.48	89.73	87.9	87.96
Exp 3	91.71	92.89	92.13	90.85
Exp 4	88.11	89.47	89.49	89.72
Exp 5	91.62	90	91.6	91.17
Exp 6	92.97	91.98	90.6	88.3
Average *RA* $\pm$ std	91.08 $\pm$ 1.88	91 $\pm$ 1.44	90.61 $\pm$ 1.65	89.8 $\pm$ 1.38

Table 4 presents the hybrid HMM/ANN results on the HospiSign database. It can be observed that the use of neural networks instead of GMMs for estimating the local emission scores leads to significant improvements in the performance of all the systems.

Comparing the results on the six experimental setups explained in Section 4.1.1 shows that irrespective of the development set chosen, the systems perform similar to one another in most of the cases.

5.2 Gesture Recognition on Chalearn14

Table 5 presents the HMM/GMM and hybrid HMM/ANN results on Chalearn14 database. In the HMM/GMM systems based on the sign-specific sequence modeling, the average number of Gaussian mixtures used is 40. Indeed, the number of Gaussian mixtures plays an important role in the performance of the systems as increasing the number of mixtures from 1 to 40 leads to around 25% absolute improvement in the gestures recognition accuracy. This improvement can be explained by the wild setup and the gesture framework which implies significant signer variation. So the balance between the number of states and the number of mixtures is more difficult to set compared to HospiSign framework where increasing the number of mixtures does not change the recognition accuracy. In the Chalearn14 case, the subunit-based HMM/GMM model

Table 4: Hybrid HMM/ANN results on HospiSign database depending on the segmentation approach explained in Section 3.2 (train-based and CV-based)

Experiment	sign-specific seq. modeling		subunit-based seq. modeling	
	Train-based	CV-based	Train-based	CV-based
Exp 1	96.58	94.76	96.87	94.46
Exp 2	95.97	95.11	95.12	94.07
Exp 3	96.46	96.97	95.86	96.17
Exp 4	95.47	95.75	95.13	94.45
Exp 5	94.29	94.02	94.97	96.48
Exp 6	95.13	94.83	95.4	94.52
Average $RA \pm$ std	95.65 ± 0.86	95.24 ± 1.01	95.56 ± 0.71	95.03 ± 1.02

Table 5: HMM/GMM and hybrid HMM/ANN results on Chalearn14 database depending on the segmentation approach explained in Section 3.2 (train-based and CV-based)

System	sign-specific seq. modeling		subunit-based seq. modeling	
	Train-based	CV-based	Train-based	CV-based
HMM/GMM	80.83	83.46	86.09	78.6
Hybrid HMM/ANN	81.31	82.98	83.77	78.51

seems to better handle this balance since we can notice a significant improvement compared to the sign-specific model.

Finally, when comparing CV-based method and train-based method in obtaining the sign specific HMM, it can be seen that subunit-based sign models resulting from sign specific HMMs obtained by the train-based method yields better systems for both databases. The performance difference is more pronounced in the case of Chalearn14 database, when compared to the HospiSign database. This can be due to the fact that in HospiSign the signs were produced in a controlled scenario by native or L1 signers, where as in Chalearn14 the signs were produced in a wild scenario not necessarily by native signers. As a consequence, CV-based method may need more development data. This needs further investigation and is part of our future work.

5.3 Comparison to Existing Studies

In this section, in order to ascertain that our approach is leading to useful systems, we contrast our results with the performance of systems reported on HospiSign and Chalearn14 databases using only the skeleton information.

5.3.1 Comparison on HospiSign Database

In [9], various manual features such as hand shape, hand position and hand movement were extracted and temporal modeling using either dynamic time warping (DTW) or temporal templates was performed. In the case of using DTW, the signs were classified using k-Nearest Neighbors (k-NN). In the case of using temporal templates, random decision forest (RDF) was used for classifying the signs.

For a fair comparison, as in [9] the average performance was calculated over the signers, we also first computed the average performance for each signer over our six experiments (explained in Section 4.1.1), and then we computed the average performance

over the signers as the final accuracy. Table 6 provides the comparison of our approach using the hybrid HMM/ANN framework with the proposed approach in [9] when using DTW along with k-NNs using hand joint distances and hand movement distances as features. Furthermore, we have presented the results in the case of using temporal templates with the random decision forests as it yielded one of the best results in [9]. It can be observed from Table 6 that both sign-specific and subunit-based sequence modeling approaches yield comparable systems to the systems developed in [9]. Furthermore, the lower standard deviation w.r.t DTW & k-NN based systems indicates that the proposed approach is yielding a more consistent system across different signers.

5.3.2 Comparison on Chalearn14 Database

In the Chalearn 2014 competition, various approaches for feature extraction, temporal segmentation and classification of gestures were investigated. In order to evaluate the proposed approaches, Jaccard index was used as the evaluation metric. Jaccard index is a commonly used metric for evaluating the gesture spotting. The Jaccard index is defined as:

$$J_{s,g} = \frac{A_{s,g} \cap B_{s,g}}{A_{s,g} \cup B_{s,g}} \ , \tag{13}$$

where $A_{s,g}$ is the ground truth for gesture g at sequence s, and $B_{s,g}$ is the prediction for this gesture at sequence s [12].

Table 7 contrasts with the performance of the systems in the competition that used only the skeleton information like the proposed approach. As discussed earlier, we have evaluated using the ground truth information on the test set. In order to get an idea on how the systems resulting from the proposed approach perform when the ground truth information is not available, we evaluated our systems based on the segmentations used in the system reported in [8].[2] When considering segmentation and classification, we can observe that the systems based on the proposed approach are neither the best nor the worst. Thus, indicating that the proposed approach is worth pursuing.

6 Discussion, Conclusions and Future Directions

This paper proposed a data-driven approach for movement subunit extraction from the skeleton information for modeling signs and gestures without any annotation informa-

[2] In [8] a random forest was used to recognize the gestures and non-gestures. We would like to thank Necati Cihan Camgöz for sharing the test set segmentations with us.

Table 6: Comparison of performance on Hos piSign database w.r.t the study reported in [9].

	Approach	Features	Accuracy
Our approach	HMM/ANN with train-based sign-specific seq. modeling	position and movement	95.65 ± 0.86
Our approach	HMM/ANN with train-based subunit-based seq. modeling	position and movement	95.56 ± 0.71
Approach in [9]	DTW & k-NN using hand movement distance	position and movement	93.81 ± 6.36
Approach in [9]	Temporal templates & RDF	position, movement and hand shape	96.67 ± 1.80

tion. The approach involves three steps: (i) extracting position and movement information given measurement of relative position of the hands with respect to the signer's body; (ii) inferring a sign-specific HMM using the training or cross-validation data; and (iii) clustering the HMM states across all the signs into subunits and developing a subunit based lexicon. Validation of the derived subunits through SL and gesture recognition studies showed that the subunit based system achieves performance comparable to or better than the best whole sign/gesture HMM based system. This indicates that the subunits based sign/gesture representations obtained by our approach maintains discrimination like whole sign/gesture HMM. Furthermore, the performances obtained on the two databases are in-line with the results reported in the literature.

Table 8 compares a few closely related subunit extraction studies. As it can be seen that the previous approaches have focused on processing images or movement information captured via gloves, while our approach focuses on modeling skeleton information, which can be easily and reliably obtained nowadays. Also, not all of these works have focused on investigating signer independence and generalization of the approach on multiple databases, as done in the present paper. Although HMMs have been previously used for subunit extraction, they have not been used the way our approach does by considering discrimination at all levels, including the Sako and Kitamura approach [28] which clusters HMM states similar to our approach. More precisely, in our approach sign/gesture level discrimination is ascertained at both segmentation and clustering steps. At the segmentation step, a sign specific HMM is obtained through sign/gesture recognition. At the clustering step, the HMM states are clustered by comparing their pdfs through Bhattacharyya distance, which is a discriminative measure [20], and thresholding it by evaluating sign/gesture level discrimination on a cross-validation data. As a result, whole sign HMMs in our approach automatically become as a reference point, like whole word HMMs in speech recognition. This can not be said about the other approaches.

Our future work will focus along the following directions with extensions to continuous sign language modeling:

1. For subunits based approach to be effectively used, it is desirable that the derived subunits are database and language independent. In the case of modeling hand shape subunits based on HamNoSys, this aspect has been well understood and exploited [7, 22]. The proposed approach of movement subunit derivation generalizes across different databases. However, it is yet to be ascertained how much

Table 7: Comparing the performance of our approach with the performance of related approaches reported in the Chalearn 2014 competition in terms of Jaccard index

Team/Approach	Accuracy	Features	Classifier
Train-based subunit-based seq. modeling (ground truth seg.)	0.8655	Skeleton	HMM/GMM
CV-based sign-specific seq. modeling (ground truth seg.)	0.8422	Skeleton	HMM/GMM
Ismar [8]	0.7466	Skeleton	Random forest
Train-based subunit-based seq. modeling (seg. from [8])	0.6868	Skeleton	HMM/GMM
CV-based sign-specific seq. modeling (seg. from [8])	0.6825	Skeleton	HMM/ANN
Terrier	0.5390	Skeleton	Random forest
YNL	0.2706	Skeleton	HMM, SVM

Table 8: Comparison of our approach to existing studies

Ref.	Features based	Segment.	Clustering algorithm	Recognition study	Signer indep. study
Sako and Kitamura [28]	images processing	multi-stream HMM	tree based algorithm	✓	✓
Bauer and Kraiss [3]	gloves	HMM	*k*-means	✓	✗
Han et al. [15]	images processing	discontinuity detector	DTW	✓	✗
Fang et al. [13]	gloves	HMM	modified *k*-means	✗	✗
Theodorakis et al. [30]	images processing	HMM	HMM hierarchical clustering	✗	✗
Our approach	skeleton	HMM	pair-wise clustering with Bhatt. dist.	✓	✓

are the derived movement subunits database or language independent. We will investigate this direction by exploring methods to model the relationship between the symbolic representation of movement information in HamNoSys representation of signs and the derived subunits, similar to modeling grapheme-to-phoneme (G2P) relationship for G2P conversion in speech processing [5].

2. Hand gesture or manual components in SL consists of hand shape, location and movement information. In this work, the focus was on modeling hand location and movement information by deriving subunits. A question that remains to be answered is: how to model jointly the hand movement subunits and hand shape subunits? We will investigate this direction along with the incorporation of RGB information.

Acknowledgment

This work was conducted in the framework of the SNSF funded Sinergia project SMILE (Scalable Multimodal sign language Technology for sIgn language Learning and assessmEnt), grant agreement CRSII2_160811. This is a consortium project that involves the Idiap Research Institute, the Hochschule für Heilpädagogik (HfH, Zürich) and the Centre for Vision, Speech and Signal Processing of the University of Surrey (UK). We thank all the collaborators of the project for their valuable work and feedback.

References

[1] O. Aran. *Vision Based Sign Language Recognition: Modeling and Recognizing Isolated Signs With Manual and Non-manual Components*. PhD thesis, Bogazici University, Istanbul, Turkey, 2008.

[2] G. Awad, J. Han, and A. Sutherland. Novel boosting framework for subunit-based sign language recognition. In *Procs. on the 16th IEEE International Conference on Image Processing (ICIP)*, pages 2729–2732. IEEE, 2009.

[3] B. Bauer and K. Karl-Friedrich. Towards an automatic sign language recognition system using subunits. In *Procs. on the Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop, GW 2001 London, UK, April 18–20, 2001 Revised Papers*, 2002.

[4] A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhy: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406, 1946.

[5] M. Bisani and H. Ney. Joint-sequence Models for Grapheme-to-phoneme Conversion. *Speech Communication*, 50(5):434–451, 2008.

[6] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, Norwell, MA, USA, 1993.

[7] N. C. Camgöz, S. Hadfield, O. Koller, and R. Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *Procs. in the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[8] N. C. Camgöz, A. A. Kindiroglu, and L. Akarun. Gesture recognition using template based random forest classifiers. In *Procs. on the European Conference on Computer Vision (ECCV) Workshops*, pages 579–594, 2014.

[9] N. C. Camgöz, A. A. Kındıroğlu, and L. Akarun. Sign language recognition for assisting the deaf in hospitals. In *Procs. of the Human Behavior Understanding: 7th International Workshop, HBU 2016, Amsterdam, The Netherlands*, 2016.

[10] H. Cooper, E. Ong, N. Pugeault, and R. Bowden. Sign language recognition using sub-units. *Journal of Machine Learning Research 13*, pages 2205–2231, 2012.

[11] R. Elakkiya and K. Selvamani. Extricating manual and non-manual features for subunit level medical sign modelling in automatic sign language classification and recognition. *Journal of Medical Systems*, 41(11):175, Sep 2017.

[12] S. Escalera, X. Baró, J. Gonzalez, M. Á. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. In *Procs. of the ECCV Workshops*, pages 459–473, 2014.

[13] G. Fang, X. Gao, W. Gao, and Y. Chen. A novel approach to automatically extracting basic units from chinese sign language. In *Procs. of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 454–457 Vol.4, Aug 2004.

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Procs. of the 23rd International Conference on Machine Learning*, pages 369–376. ACM, 2006.

[15] J. Han, G. Awad, and A. Sutherland. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Pattern Recognition Letters*, 30(6):623 – 633, 2009.

[16] J. Han, G. Awad, and A. Sutherland. Boosted subunits: a framework for recognising sign language from videos. *IET Image Processing*, 7(1):70–80, February 2013.

[17] T. Hanke. HamNoSys - representing sign language data in language resources and language processing contexts. *Workshop proceedings : Representation and processing of sign languages*, pages 1–6., 2004.

[18] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press 1998.

[19] D. Johnson et al. ICSI Quicknet Software Package. http://www.icsi.berkeley.edu/Speech/qn.html, 2004.

[20] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, 15(1):52–60, February 1967.

[21] O. Koller, H. Ney, and R. Bowden. May the force be with you: Force-aligned signwriting for automatic subunit annotation of corpora. In *Procs. of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, April 2013.

[22] O. Koller, H. Ney, and R. Bowden. Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled. In *Procs. on the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3793–3802, 2016.

[23] O. Koller, O. Zargaran, H. Ney, and R. Bowden. Deep sign: Hybrid CNN-HMM for continuous sign language recognition. In *Procs. of the British Machine Vision Conference (BMVC)*, 2016.

[24] S. K. Liddell and R. E. Johnson. American Sign Language: The Phonological Base . *Sign Language Studies*, 64:195–277, 1989.

[25] V. Pashaloudi and K. Margaritis. Hidden markov model for sign language recognition: A review. In *Procs. of the 2nd Hellenic Conference AI, SETN-2002*, pages 11–12, 2002.

[26] V. Pitsikalis, S. Theodorakis, C. Vogler, and P. Maragos. Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In *Procs. of the CVPR 2011 Workshops*, pages 1–6, June 2011.

[27] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[28] S. Sako and T. Kitamura. Subunit modeling for Japanese sign language recognition based on phonetically depend multi-stream hidden Markov models. In *Procs. of the Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion*, pages 548–555, Berlin, Heidelberg, 2013.

[29] W. Stokoe. An outline of the visual communication systems of the American deaf. *Studies in linguistics: Occasional papers*, 86, 1960.

[30] S. Theodorakis, V. Pitsikalis, and P. Maragos. Model-level data-driven sub-units for signs in videos of continuous sign language. In *Procs. in the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2262–2265, March 2010.

[31] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1583–1597, 2016.

[32] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge University Engineering Department, 2002.