An HMM Approach with Inherent Model Selection for Sign Language and Gesture Recognition

Sandrine Tornay, Oya Aran, Mathew Magimai.-Doss
Idiap Research Institute, Ecole Polytechnique Fédérale de Lausanne, De La Salle Univerity
Martigny CH-1920 Switzerland, Lausanne CH-1015 Switzerland, Malate Manila Philippines
{sandrine.tornay, mathew}@idiap.ch

Abstract

HMMs have been the one of the first models to be applied for sign recognition and have become the baseline models due to their success in modeling sequential and multivariate data. Despite the extensive use of HMMs for sign recognition, determining the HMM structure has still remained as a challenge, especially when the number of signs to be modeled is high. In this work, we present a continuous HMM framework for modeling and recognizing isolated signs, which inherently performs model selection to optimize the number of states for each sign separately during recognition. Our experiments on three different datasets, namely, German sign language DGS dataset, Turkish sign language HospiSign dataset and Chalearn14 dataset show that the proposed approach achieves better sign language or gesture recognition systems in comparison to the approach of selecting or presetting the number of HMM states based on k-means, and yields systems that perform competitive to the case where the number of states are determined based on the test set performance.

Keywords: Sign Language Recognition, Hidden Markov Models, Model Selection

1. Introduction

Following the recent developments and advancements on automatic speech recognition, commercial systems are becoming available in our daily life, where users can use their everyday communication means, i.e. speech, to interact with machines. However these systems rarely address the deaf members of our society, who predominantly use sign languages as their primary means of communication. For more accessible technology, research on automatic Sign Language Recognition (SLR) should also advance, which is currently in its infancy.

Sign languages are the natural communication media of deaf people. In spoken languages, the words are produced through the vocal tract and are perceived as sounds, whereas in sign languages, the signs are produced alone or simultaneously, by use of hand shape, hand motion, hand location, as well as facial expression, head motion, and body posture, and they are perceived visually. The production of signs follow both sequential and parallel nature: signs come one after the other showing a sequential behavior; at the same time each sign may contain parallel actions of hands, face, head or body (Stokoe, 2005). Due to its visual nature, the multimodal properties, and parallel use of different channels, sign languages present other challenges for automatic recognition, from feature extraction to modeling, in comparison to spoken languages.

HMMs offer a natural solution for SLR with their power in handling sequential and multimodal data. They are extensively used and have proven successful in the sign language recognition domain (Ong and Ranganath, 2005; Cooper et al., 2011). One of the challenges of using HMMs in sign language processing is that sign languages are inherently under-resourced i.e. few well developed resources with several signers are available, and HMMs require a certain amount of training data for robust parameter estimation. Another challenge is to select the structure of the HMM, i.e. the number of states, which directly can affect the performance of sign language recognition system. Unlike speech processing, where the spoken words are represented as a sequence of subword units (e.g. phones) and the subword units are modeled through an HMM with minimum duration constraint (Bourlard and Morgan, 1994), there is no such prior knowledge for sign language. As discussed in detail in Section 2, in many studies the number of states in the HMM is fixed for all the signs in the dataset. This may not be optimal, as the temporal structure of signs can differ, akin to temporal differences in spoken words.

The present paper focuses on addressing the challenge related to defining or determining the HMM structure in sign language processing. Specifically, we develop a HMM-based approach where, during the training phase, each sign is modeled by a set of HMMs with different number of states. During the recognition phase, the sign language recognition system determines the number of states for each sign independently such that the joint likelihood of the HMM state sequence and the feature observation is maximized. In other words, the approach selects the best matching HMM during testing time. The motivation being that, as there is no prior knowledge to determine the HMM structure, treat the number of states for each sign as an hidden information. Further, for a single sign (or lexical entity) there can be signer variations. For instance, signers can sign at different speeds (fast or slow) while varying the hand movement. Having multiple HMMs per sign could also potentially handle signer variation. To draw an analogy to spoken language processing, speech recognition systems typically handle pronunciation variation (introduced by speakers) by having multiple pronunciations as well as by changing minimum duration constraints (Strik and Cucchiarini, 1999). Besides that, we also propose incorporation of a transition model, similar to silence modeling in speech recognition (Young et al., 2002), to model portions of visual signal before and after production of signs.

We investigate the proposed approach by modeling hand movement information based on skeleton information on three different datasets, namely, (a) German sign language DGS corpus consisting of isolated signs collected as part of EU project DictaSign, (b) Turkish sign language dataset HospiSign consisting of phrase classes and Chalearn 14 dataset consisting of Italian hand gestures and compare with the entropy-based k-means approach presented in (Li et al., 2016) to set the number of HMM states. Our studies show that the proposed approach consistently outperforms the k-mean approach and performs comparable to the case where the fixed number of HMM states are determined based on the performance on the *test set*. Furthermore, systems incorporating the transition model yield improved performances.

The paper is organized as follows. In Section 2., we present the related work on SLR and HMM modeling. Our proposed approach is explained in Section 3.. Section 4. and Section 5. presents the experimental setup and the results and analysis of the experiments, respectively. Conclusions and future work is given in Section 6..

2. Related Work

In this section, we present several key works on sign language recognition focusing on the works that primarily use HMMs for sign modeling. Interested readers can refer to more exhaustive surveys covering the sign language recognition literature (Ong and Ranganath, 2005; Cooper et al., 2011).

Signs include multiple dynamic elements that are performed in parallel. For recognizing signs, and recognizing sign language in general, one must use methods that are capable of modeling the temporal and multimodal characteristics of the signs. Among several methods that have been used in the literature, HMMs have proven successful in several kinds of SLR systems and have been the baseline model for modeling signs due to their success in modeling sequential data and their flexibility to handle multiple parallel streams of data.

The initial works in SLR have used HMMs for modeling and classifying signs (Starner and Pentland, 1995; Vogler and Metaxas, 1997). In (Starner and Pentland, 1995), the authors propose a vision based system and used a four-state left to right HMM with one skip transition for recognition. In (Vogler and Metaxas, 1997), the authors address the co-articulation effects in continuous signing and model the transition movements between signs and the signs themselves with different HMM topologies. The experiments in these initial works present a signer dependent setup as the data used are from a single user.

Not only the initial works but a large number of recent studies also use HMMs for sign modeling and classification (Ong and Ranganath, 2005; Cooper et al., 2011). A large range of HMM variants have also been used in the literature. In (Kumar et al., 2017), the authors use coupled HMM and present a multi sensor fusion framework for isolated sign recognition. In (Keskin and Akarun, 2009), the authors used an Input-Output HMM (IOHMM) for modeling signs where the hand shape features are provided as input to the IOHMM. Parallel HMMs have been used in (Vogler and Metaxas, 1999; Zaki and Shaheen, 2011) where a separate HMM is used for modeling right and left hands or different feature sets such as hand configuration, shape and motion.

Methods other than HMMs have also been used in the literature, such as Dynamic Time Warping (DTW) (Lichtenauer et al., 2008), Conditional Random Fields (CRF) (Kong and Ranganath, 2014), Gaussian Process Dynamical Models (Gamage et al., 2011), and continuous iterated conditional modes (Nayak et al., 2012). In (Caridakis et al., 2012), Self Organizing Maps (SOM) have been used to model the spatial aspect of the signs and Markov models for its temporal counterpart.

In most of the works on SLR that use HMMs, a left-to-right HMM structure, with or without skip states has been used. The number of states of the HMM has generally been fixed for all the signs/subunits in the dataset. In (Liu et al., 2004), the authors compare three different HMM topologies with different number of states, without presenting any model selection approach: fully connected, left-to-right with skip states and left-to-right without skip states and conclude that the left-to-right HMM provides the best performance, confirming the popularity of the left-to-right HMMs for SLR. Only a couple of works in the literature have investigated a model selection approach for HMMs for SLR. In (Siddiqi et al., 2007), the authors present a state splitting algorithm for HMMs. In their experiments on a dataset of signs from Australian sign language, their proposed approach is faster and achieves better performance than the conventional HMM Baum Welch training. In (Matsuo et al., 2008), the HMM topology is automatically constructed from an initial topology by modifying it using segments, which are formed based on the segmentation of hand motion. In (Wang et al., 2015), low rank approximation is used to determine the key frames of a sign which guides the selection of the number of states of HMM independently for each sign. In (Li et al., 2016), an entropy-based k-means algorithm is used to determine the number of states in an HMM, where each sign is modeled by one HMM. With this approach, each sign is represented by an HMM with different number of states. Additionally, an artificial bee colony algorithm is used together with the Baum Welch algorithm to determine the HMM structure. Their experiments show that the proposed approach achieves better performance than a left-to-right HMM structure with fixed number of states. However, no results have been reported to understand how much of the performance increase comes from the selection of number of states and from the determination of the HMM structure through the swarm optimization algorithm.

When sign based modeling is used, the scalability problem arises: the number of HMMs that needs to be trained increase with the increased vocabulary. This problem is more evident if bi-gram or tri-gram models is to be used for continuous SLR. As a possible solution to the scalability problem, one can identify basic subunits of signs, analogous to the phonemes in speech, which would then be used to constitute all the signs in the vocabulary. Identifying subunits decreases the total number of models that needs to be trained, as the number of subunits is expected to be less than the number of signs in the vocabulary. In (Vogler and Metaxas, 1998), an approach based on modeling the subunits of the signs rather than the whole sign has been presented and applied to continuous SLR. In (Wang et al., 2002), the authors present their work on large vocabulary continuous SLR based on subunit modeling of signs. Parallel HMMs have also been used to model subunits of signs instead of the whole sign words (Theodorakis et al., 2014; Vogler and Metaxas, 2001).

3. Proposed Approach

The proposed approach consists of two steps: (1) extracting the features based on the skeleton information and (2) inferring a sign-based hidden Markov model.

3.1. Feature Extraction

Signing takes place in 3D and around the upper body region. The components of a sign contain manual signals such as hand shape, hand motion, hand position, and non-manual signals, such as facial expressions, head motion and body posture. While the manual signals are the basic components that form the signs, several other key body parts such as the face, shoulders and arms are also important in the analysis of manual signs in order to understand the relative position of the hands with respect to the body (Aran, 2008).

For extracting the features, we rely on the tracked 3D coordinates of a human skeleton. The 3D trajectories of the two hands as well as the other skeleton joints such as head, neck, shoulders and hips form the basis for our continuous features of hand motion information, in particular hand position and velocity. While using discretized features is an option, in the presence of enough training examples, continuous features outperform discretized features as discretization results in data loss in most cases (Aran, 2008). We use three coordinate centers to normalize for the translation: the head, the shoulders and the hips. The distance between the neck and the head is used to normalize for the scale. After normalization, the stack of the continuous hand motion and position values related to the three coordinate centers give us the necessary information on the hand trajectory and position with respect to the signer's body.

For each frame t, we first calculate the 3D normalized position features for the left and right hands according to each of the three coordinate centers. First, the x, y, z coordinates of the hands are recalculated with respect to the coordinate center. Next, the coordinates are normalized by an estimate of the head size, which has been calculated as the quarter of the absolute distance between the y coordinates of neck and head:

$$\mathbf{p}_t^{lhnd} = \frac{\mathbf{lhnd} - \mathbf{center}}{|neck_y - head_y|\,/4} \quad (1)$$

$$\mathbf{p}_t^{rhnd} = \frac{\mathbf{rhnd} - \mathbf{center}}{|neck_y - head_y|\,/4} \quad (2)$$

where $\mathbf{lhnd}$, $\mathbf{rhnd}$, $\mathbf{center}$ are vectors containing the x, y, z coordinates of related joints: left hand, right hand, respectively and head, right/left shoulder, right/left hip for the $\mathbf{center}$ where the right shoulder/hip center is used to compute the right hand position vector and the left shoulder/hip for the left one; $neck_y$, $head_y$ are the y coordinate of the neck and the head joint at time frame t.

The velocity features, $\mathbf{v}_t^{lhnd}, \mathbf{v}_t^{rhnd}$, are then computed by the difference of the hand position vectors between time t and $t-2$:

$$\mathbf{v}_t^{lhnd} = \mathbf{p}_t^{lhnd} - \mathbf{p}_{t-2}^{lhnd} \quad (3)$$

$$\mathbf{v}_t^{rhnd} = \mathbf{p}_t^{rhnd} - \mathbf{p}_{t-2}^{rhnd} \quad (4)$$

The resulting feature vector is the stack of $\mathbf{p}_t^{lhnd}$, $\mathbf{p}_t^{rhnd}$, $\mathbf{v}_t^{lhnd}$ and $\mathbf{v}_t^{rhnd}$ according to the three coordinate centers, leading to a sequence of F 36 dimensional feature vectors, where F is the total number of frames and the 36 dimensional feature vector consists of 18 position features (= 3 coordinates × 3 coordinate centers × 2 hands) and 18 velocity features.

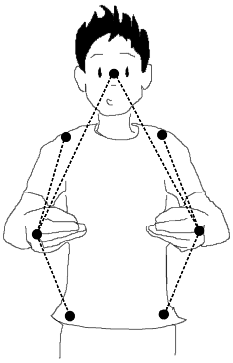

Figure 1: The features are extracted using x, y, z coordinates of the skeleton joints of left hand, right hand according to three coordinate centers (head, shoulder and hip joint).

3.2. Sign-based HMM Inference

Given the feature vectors, a sign-based left-to-right HMM is trained for each sign. The choice of the number of states used to model each sign is an important issue, which directly affects the performance of the model. To handle this problem, we propose the following model selection approach. The quality of the data segmentation can also affect the sign-based model. To outperform this issue, we propose a transition model explained in this section.

The proposed model selection approach assumes that each sign could have different number of stationary states. Intuitively an HMM state can represent a specific position, orientation or shape of the hands depending on the feeding observation features. Therefore we made the assumption that the complexity of a sign influences the appropriate number of states used to model it. To the authors' knowledge, no method in the literature today allows to set this number beforehand. An exhaustive search using cross validation is not feasible in the sign language domain as the number of signs in the datasets is typically high. Thus, instead of setting this number beforehand, the proposed model selection approach selects the appropriate one in an interval of possibilities at the recognition stage. More precisely, an interval of possible number of states is first chosen, let's say N_{min} to N_{max}. Then, for all n in the defined range, a left-to-right HMM with n states is trained for each sign. Then at the recognition stage, the model leading to the maximum likelihood is chosen as the appropriate one (see Figure 2). Thus $S \cdot (N_{max} - N_{min} + 1)$ models are tested, in comparison to S in the first approach, where S is the number of signs in the dataset.

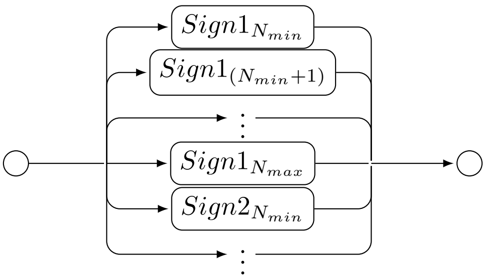

Figure 2: Recognition Network of the proposed model selection approach.

Moreover, the exact start and end of a sign is not perfectly defined, especially in a continuous context where each sign is being followed by other signs. In the isolated context, the segmentation is not necessarily optimized leading to the same problem. Thus in both cases, there is some transition phase at the beginning and at the end of the performed sign. This period can represent the absence of movement or even some slight insignificant movement. To prevent this being taken into account in modeling the sign, we propose to add a transition model, common to each sign, before and after each sign-based HMM. For preserving the continuity of the entire model, we modeled it as a three-state left-to-right HMM with one-state-skip (see Fig.3 for the structure).

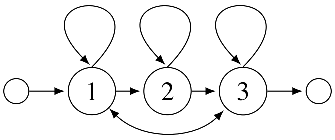

Figure 3: Structure of the transition model.

4. Experimental Setup

We validated the proposed approach on the isolated sign language recognition and gesture recognition tasks. To do so, three datasets were used: the Chalearn14, the DGS and the HospiSign datasets. In this section, the description of the datasets used for evaluating our proposed approach is given as well as the setup of the presented systems.

4.1. Chalearn14 Dataset

The Chalearn14 dataset contains the data used in the Chalearn 2014 challenge, which includes a vocabulary of 20 Italian cultural/anthropological gestures (Escalera et al., 2013). The users are recorded in front of a Kinect, performing natural communicative gestures and speaking in fluent Italian. Each sign has been repeated several times by each user. We have used the skeletal joint coordinates provided in the dataset. The dataset is publicly available[1].

4.2. DGS Dataset

The DGS dataset contains 40 signs from German Sign Language (DGS). The dataset includes data from 14 non-native right-handed signers, where each sign is repeated approximately 5 times by each person. There are a total of 3186 signs in the dataset. The DGS is a challenging dataset as the signs performed by the non-native signers contain a large variety. The dataset has been recorded with a Kinect camera and the 3D coordinates of a human skeleton has been tracked using the OpenNI framework. The resulting skeletal joint coordinates has been shared with us by the authors of (Ong et al., 2012), which we used as the basis for feature extraction. More information about the DGS dataset can be found in (Ong et al., 2012).

4.3. HospiSign Dataset

The HospiSign dataset is a subset of the BosphorusSign dataset (Camgöz et al., 2016a), containing 33 phrase classes from Turkish Sign Language (TSL) related to the health domain. The HospiSign subset includes 6 signers, with each sign being repeated approximately 6 times (Camgöz et al., 2016b). The dataset is publicly available by request from the authors[2]. The dataset has been recorded with a Kinect camera. We have used the skeletal joint coordinates that are provided in the dataset as the basis for our feature extraction.

4.4. Systems

In this section, we present three systems, namely, (a) proposed model selection based system, denoted as *msHMM*, (b) a system where the number of states are varied, denoted as *sdHMM* and best performing system is selected on the test set and (c) a system where the number of states is set using k-means, denoted as *kmHMM* systems. We also present corresponding *tr-msHMM*, *tr-sdHMM* and *tr-kmHMM* systems that incorporate transition models at the beginning and end of the sign.

For all presented systems, we used the position and velocity features of both hands as input features. The resulting feature vector is of size 36, see Section 3.1. for details. Moreover we used the leave-one-signer-out cross validation to report signer independent accuracy in the DGS and HospiSign studies. For the Chalearn14 case, we kept the given data segmentation of the challenge. The performance accuracy used in this paper is the ratio of the number of correct recognized sign on the total number of signs. All the HMMs have been trained with the HTK toolkit (Young et al., 2002). In all cases, each HMM state emission distribution is modeled with a single multivariate Gaussian with diagonal covariance matrix. To set the appropriate range of states, we assume that the minimum duration constraint of a sign is half a second. Any HMM architecture used should be able to model a sign that is at least 0.5 seconds. This assumption allows us to determine how much stationary states we need: with a 25fps frame rate (common to all the datasets in this study), half a second corresponds to 12.5 frames, rounded of to a maximum of 13 states since we used a left-to-right HMM structure.

[1] http://gesture.chalearn.org/mmdata#Track3

[2] https://www.cmpe.boun.edu.tr/pilab/BosphorusSign/home_en.html

(A) The *msHMM* system stands for our proposed model selection approach (see Section 3.), where N is different for each sign and is not fixed beforehand. Only an interval of possibilities from N_{min} to N_{max} has to be defined. The range tried was from 3 to 13 (according to the minimum duration constraint explained above).

(B) In most of the HMM existing studies (see Section 2.), the number of states N is common to all the sign. For the sake of comparison, the *sdHMM* system represents this standard approach, i.e. each sign was modeled with a N states left-to-right HMM, for all N between 3 and 13 (according to the duration constraint presented above). The system corresponding to the number of states that yield the best performance on the test serves as a baseline system. We refer to this system as pseudo-oracle system.

(C) To validate the proposed model selection model, we implemented the entropy-based k-means algorithm presented in (Li et al., 2016). This system is referred as the *kmHMM* system. For fair comparison, k was taken into the same range as the *msHMM*, i.e. between 3 to 13.

The *tr-msHMM*, *tr-sdHMM*, *tr-kmHMM* systems are a combination of the transition model and the *msHMM*, *sdHMM*, *kmHMM* systems respectively. More precisely, we added the transition model, tr_N, before and after each sign-based model. Since by adding the three states transition model we increase the number of state of the sign-based model by at least four states (two states before and after each model, see Fig. 3), we decided to adapt the range of possibilities, leading to 3 to 9. Figure 4 depicts the recognition network of the *sdHMM* and *kmHMM* systems, while Figure 5 the *msHMM* system.

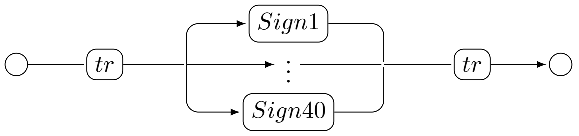

Figure 4: Recognition Network of the *tr-sdHMM* and the *tr-kmHMM* system.

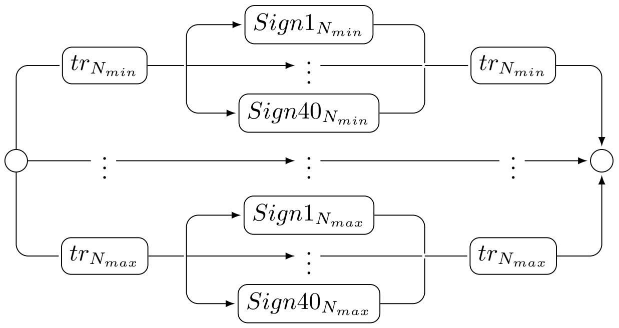

Figure 5: Recognition network of the *tr-msHMM* system.

5. Results and Analysis

First the recognition results of the systems (see Section 4.4.) for the Chalearn14, DGS and HospiSign datasets are presented. Next, we contrast the performances obtained by our approach with the existing studies reported on those datasets to demonstrate that the results obtained by our systems are competitive.

5.1. Comparison of Systems

Figures 6 presents the recognition accuracy of all the signer-independent systems without transition model. We can observe that the proposed approach (*msHMM*) consistently outperforms the approach of setting number of states based on k-means (*kmHMM*). Furthermore, we can also observe that *msHMM* system yields performance comparable to pseudo-oracle system, i.e. *sdHMM* system with fixed number of states yielding the best performance on the test data.

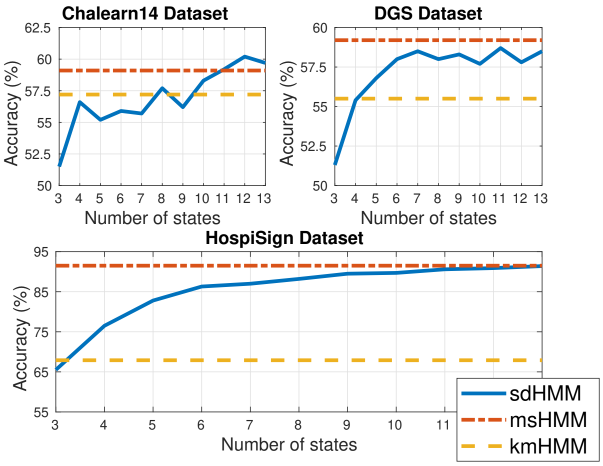

Figure 6: Recognition accuracy of the *sdHMM*, the *msHMM* and the *kmHMM* systems.

Figures 7 presents the recognition accuracy of all the systems containing the transition model. We can observe that the recognition performance of all the systems considerably improve. As the transition model is common to all signs, the improvement can be attributed to the modeling of sign-independent irrelevant information at the beginning and end of the visual signal. When comparing *tr-msHMM*, *tr-kmHMM* and *tr-sdHMM* systems, the trend remains the same, i.e. *tr-msHMM* system is better than *tr-kmHMM* system and is comparable to the pseudo-oracle *tr-sdHMM* system.

For the sake of completeness, Table 1 summarizes the recognition accuracy with standard deviation for all the systems.

	Chalearn14	**DGS**	**HospiSign**
msHMM	59.1	59.2 ± 9.6	91.5 ± 7.0
sdHMM (# state)	60.2 *(12)*	58.7 ± 11.5 *(11)*	91.4 ± 6.0 *(13)*
kmHMM	57.2	55.5 ± 10.5	67.9 ± 4.4
tr-msHMM	60.2	63.1 ± 10.3	91.2 ± 6.5
tr-sdHMM (# state)	61.5 *(6)*	62.3 ± 9.8 *(5)*	92.2 ± 4.9 *(8)*
tr-kmHMM	57.9	60.9 ± 9.5	86.1 ± 6.3

Table 1: Recognition accuracy of the systems on the Chalearn14, DGS and HospiSign dataset. *sdHMM (# state)* and *tr-sdHMM (# state)* denote pseudo-oracle systems.

Figure 8 shows the histogram of number of states of the HMMs selected during recognition phase of *tr-msHMM*

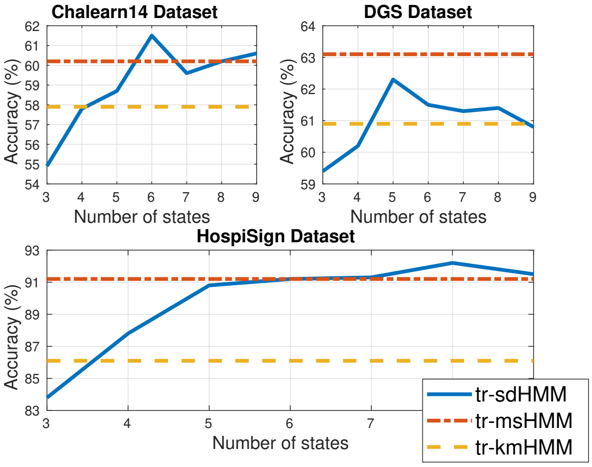

Figure 7: Recognition accuracy of the *tr-sdHMM*, the *tr-msHMM* and the *tr-kmHMM* systems.

system for Chalearn14, DGS and HospiSign. As expected, it can be observed that HMMs with different number of states are selected at run time. In the case of DGS and HospiSign models, the histogram is skewed towards higher number of states. However, it is not the case for Chalearn14. One possible reason for that could be that Chalearn14 has simple gestures (hand up and down movement) in a "wild" (or uncontrolled) environment.

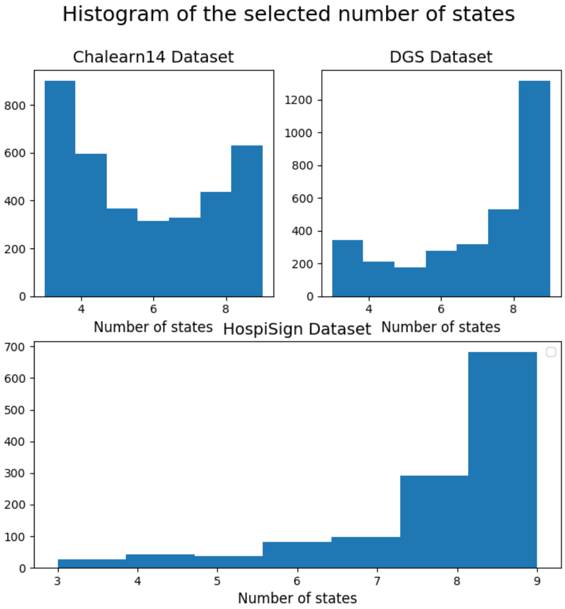

Figure 8: Histogram of the selected number of states during the recognition process using the *tr-msHMM* systems.

5.2. Comparison to Existing Studies

In this section, we contrast the performances obtained on DGS dataset and HospiSign dataset to existing studies reported on these datasets. These studies have used the same protocols as we have. In the case of Chalearn14, the evaluation is based on Jaccard index that involves joint evaluation of segmentation and recognition of gestures (Escalera et al., 2015). A fair comparison is not feasible, as it is difficult to separate the contribution of segmentation errors and recognition errors of the systems reported in (Escalera et al., 2015). Thus, we do not contrast for Chalearn14.

5.2.1. DGS Dataset

Table 2 compares performance with our models with two other works on the DGS dataset. In (Ong et al., 2012), the authors use a multi-class sequential pattern tree with boosting for classifying signs, using binary features based on the hand motion and location information. In (Cooper et al., 2012), the authors propose a subunit based approach using a Sequential Pattern Boosting classifier, where the subunits are extracted based on the different modalities that make up a sign, i.e. shape, location, motion, and hand arrangement. It is important to note that the dataset used in (Cooper et al., 2012) contains signs from one extra signer, which we do not have access to in our dataset. Based on the reported signer independent performance of 49.4% in (Cooper et al., 2012), we calculated the accuracy range for the remaining 14 users (assuming that the accuracy on the 15th user could take a value between 0% and 100%). This calculation gives us a range of $[45.7, 52.9]$, which is still lower than the performance achieved the proposed *tr-msHMM* system.

Method	Signer Indep. (%)
Sequential Pattern Trees (Ong et al., 2012)	55.4
Boosted Subunits (Cooper et al., 2012)	49.4
tr-msHMM system	**63.1**

Table 2: Comparison of our systems with existing studies for the DGS dataset

5.2.2. HospiSign Dataset

Table 3 compares the performance of our system with the performance reported in (Camgöz et al., 2016a). Briefly, in (Camgöz et al., 2016a), 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 Neighbours (k-NN). We contrast to the system where only hand movement information is modeled. We also trained a *tr-msHMM* system that uses the same hand movement and hand joint feature as in (Camgöz et al., 2016a). In both cases, we can observe that the proposed *tr-msHMM* system yields performance close to the best reported system.

6. Conclusion

This paper presented a model selection approach where, each sign is modeled by a set of HMMs with different number of states during training and the best matching model is automatically selected during recognition based on maximum likelihood criteria. We also investigated the use of a transition model taking inspiration from silence modeling in speech processing. Our investigations on sign language recognition and gesture recognition tasks on three different

Method	Signer Indep. (%)
Hand Joint and Movement Distances (Camgöz et al., 2016a)	**93.8 ± 6.36**
tr-msHMM system	91.2 ± 6.5
tr-msHMM system using the same "Hand Joint and Movement Distances" features as (Camgöz et al., 2016a)	91.6 ± 6.07

Table 3: Comparison of our systems with existing studies for the HospiSign dataset

datasets show that the proposed model selection approach yields better systems than the approach of presetting the number of HMM states and yields systems competitive to the best performing systems with fixed number of HMM states determined on the test set. Furthermore, incorporation of a transition model to model portion of visual signal before and after the production of each sign helps in improving the performance of systems.

It is worth mentioning that, although the investigations were carried out on isolated signs, gestures and phrases, the approach can be extended to continuous sign language processing. As (elucidated in Section 1), the different HMMs for each sign serve a similar role as multiple pronunciations for each word in speech recognition systems. The decoder can handle that. The present work focused on modeling hand movement information. It is worth mentioning that the model selection approach can be adopted when modeling both hand movement and hand shape information (Tornay et al., 2019). Also, the model selection approach could potentially be exploited for hand movement subunits extraction (Tornay and Magimai.-Doss, 2019). Our future work will pursue investigation of the proposed model selection for modeling hand movement and hand shape information and for subunits extraction in the context of continuous sign language processing.

7. Acknowledgments

This work was funded by the SNSF through the Sinergia project SMILE *(Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment)*, grant agreement CRSII2_160811. We thank all the collaborators in the project for their valuable work. We thank our collaborators from the University of Surrey for giving access to the DGS corpus.

8. Bibliographical References

Aran, O. (2008). *Vision Based Sign Language Recognition: Modeling and Recognizing Isolated Signs With Manual and Non-manual Components*. Ph.D. thesis, Bogazici University, Istanbul, Turkey.

Bourlard, H. and Morgan, N. (1994). *Connectionist Speech Recognition - A Hybrid Approach*. Kluwer Academic Publishers.

Caridakis, G., Karpouzis, K., Drosopoulos, A., and Kollias, S. (2012). Non parametric, self organizing, scalable modeling of spatiotemporal inputs: The sign language paradigm. *Neural Networks*, 36:157–166.

Cooper, H., Holt, B., and Bowden, R. (2011). Sign language recognition. In *Visual Analysis of Humans*, pages 539–562. Springer London.

Cooper, H., Ong, E.-J., Pugeault, N., and Bowden, R. (2012). Sign Language Recognition Using Sub-units. *J. Mach. Learn. Res.*, 13(1):2205–2231, July.

Escalera, S., Baró, X., Gonzàlez, J., Bautista, M. A., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H. J., Shotton, J., and Guyon, I. (2015). Chalearn looking at people challenge 2014: Dataset and results. In Lourdes Agapito, et al., editors, *Computer Vision - ECCV 2014 Workshops*, pages 459–473, Cham. Springer International Publishing.

Gamage, N., Kuang, Y. C., Akmeliawati, R., and Demidenko, S. (2011). Gaussian Process Dynamical Models for hand gesture interpretation in Sign Language. *Pattern Recognition Letters*, 32(15):2009–2014.

Keskin, C. and Akarun, L. (2009). STARS: Sign tracking and recognition system using input–output HMMs. *Pattern Recognition Letters*, 30(12):1086–1095. Image/video-based Pattern Analysis and {HCI} Applications.

Kong, W. and Ranganath, S. (2014). Towards subject independent continuous sign language recognition: A segment and merge approach. *Pattern Recognition*, 47(3):1294–1308. Handwriting Recognition and other {PR} Applications.

Kumar, P., Gauba, H., Roy, P. P., and Dogra, D. P. (2017). Coupled HMM-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, 86:1–8.

Li, T. H. S., Kao, M. C., and Kuo, P. H. (2016). Recognition System for Home-Service-Related Sign Language Using Entropy-Based K -Means Algorithm and ABC-Based HMM. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(1):150–162, Jan.

Lichtenauer, J. F., Hendriks, E. A., and Reinders, M. J. T. (2008). Sign Language Recognition by Combining Statistical DTW and Independent Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):2040–2046, Nov.

Liu, N., Lovell, B. C., Kootsookos, P. J., and Davis, R. I. A. (2004). Model structure selection training algorithms for an HMM gesture recognition system. In *Ninth International Workshop on Frontiers in Handwriting Recognition*, pages 100–105, Oct.

Matsuo, T., Shirai, Y., and Shimada, N. (2008). Automatic generation of HMM topology for sign language recognition. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Dec. hmm model selection/formation.

Nayak, S., Duncan, K., Sarkar, S., and Loeding, B. (2012). Finding Recurrent Patterns from Continuous Sign Language Sentences for Automated Extraction of Signs. *J. Mach. Learn. Res.*, 13(1):2589–2615, September.

Ong, S. C. W. and Ranganath, S. (2005). Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):873–891.

Siddiqi, S., Gordon, G., and Moore, A. (2007). Fast State Discovery for HMM Model Selection and Learning. In

Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AI-STATS).

Starner, T. and Pentland, A. (1995). Real-Time American Sign Language Recognition From Video Using Hidden Markov Models. In *SCV95*.

Stokoe, J. W. C. (2005). Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf. *The Journal of Deaf Studies and Deaf Education*, 10(1):3.

Strik, H. and Cucchiarini, C. (1999). Modeling Pronunciation Variation for ASR: A Survey of the Literature. *Speech Communication*, 29(2-4):225–246.

Theodorakis, S., Pitsikalis, V., and Maragos, P. (2014). Dynamic–Static Unsupervised Sequentiality, Statistical Subunits and Lexicon for Sign Language Recognition. *Image and Vision Computing*, 32(8):533–549.

Tornay, S. and Magimai.-Doss, M. (2019). Subunits inference and lexicon development based on pairwise comparison of utterances and signs. *Information*, 10:298.

Tornay, S., Razavi, M., Camgoz, N. C., Bowden, R., and Magimai.-Doss, M. (2019). HMM-based Approaches to Model multichannel Information in Sign Language inspired from Articulatory Features-based Speech Processing. In *Proceedings of ICASSP*.

Vogler, C. and Metaxas, D. (1997). Adapting Hidden Markov models for ASL recognition by using three-dimensional computer vision methods. In *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 156–161.

Vogler, C. and Metaxas, D. (1998). ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis. In *Sixth International Conference on Computer Vision (ICCV '98)*, page 363, Washington, DC, USA. IEEE Computer Society.

Vogler, C. and Metaxas, D. (1999). Parallel Hidden Markov Models for American Sign Language Recognition. In *International Conference on Computer Vision, Kerkyra, Greece*, volume 1, pages 116–122.

Vogler, C. and Metaxas, D. (2001). A framework for recognizing the simultaneous aspects of American sign language. *Computer Vision and Image Understanding*, 81(3):358–384.

Wang, C., Shan, S., and Gao, W. (2002). An Approach Based on Phonemes to Large Vocabulary Chinese Sign Language Recognition. In *Fifth International Conference on Automatic Face and Gesture Recognition*, volume 00, page 0411, Los Alamitos, CA, USA. IEEE Computer Society.

Wang, H., Chai, X., Zhou, Y., and Chen, X. (2015). Fast sign language recognition benefited from low rank approximation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6, May. hmm model selection.

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

Zaki, M. M. and Shaheen, S. I. (2011). Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters*, 32(4):572–577.

9. Language Resource References

Camgöz, N. C., Kindiroglu, A. A., Karabüklü, S., Kelepir, M., Özsoy, A. S., and Akarun, L. (2016a). BosphorusSign: A Turkish Sign Language Recognition Corpus in Health and Finance Domains. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Camgöz, N. C., Kındıroğlu, A. A., and Akarun, L. (2016b). Sign Language Recognition for Assisting the Deaf in Hospitals. In *International Workshop on Human Behavior Understanding*, pages 89–101. Springer International Publishing.

Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., Athitsos, V., and Escalante, H. (2013). Multimodal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 445–452. ACM.

Ong, E.-J., Cooper, H., Pugeault, N., and Bowden, R. (2012). Sign language recognition using sequential pattern trees. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2200–2207. IEEE.