Data-Driven Sub-Units and Modeling Structure for Continuous Sign Language Recognition with Multiple Cues

Vassilis Pitsikalis, Stavros Theodorakis and Petros Maragos

National Technical University of Athens, School of ECE, Athens 15773, Greece.
{vpitsik,sth,maragos}@cs.ntua.gr.

Abstract

We investigate the automatic phonetic modeling of sign language based on phonetic sub-units, which are data driven and without any prior phonetic information. Visual processing is based on a probabilistic skin color model and a framewise geodesic active contour segmentation; occlusions are handled by a forward-backward prediction component leading finally to simple and effective region-based visual features. For sign-language modeling we propose a modeling structure for data-driven sub-unit construction. This utilizes the cue that is considered crucial to *segment* the signal into parts; at the same time we also *classify* the segments by implicitly assigning labels of Dynamic or Static type. This segmentation and classification step disentangles *Dynamic* from *Static* parts and allows us to employ for each type of segment the *appropriate* cue, modeling and clustering approach. The constructed Dynamic segments are exploited at the model level via hidden Markov models (HMMs). The Static segments are exploited via k-means clustering. Each Dynamic or Static part, exploits the appropriate cue related to the movement. We propose that the movement cues are normalized in order to be translation and scale *invariant*. We apply the proposed modeling for further combination of the movement trajectory individual cues. The proposed approaches are evaluated in recognition experiments conducted on the continuous sign language corpus of Boston University (BU-400) showing promising preliminary results.

1. Introduction

Sign languages, i.e., languages that essentially convey information via visual patterns, commonly serve as an alternative or complementary mode of human communication. Visual patterns, as opposed to the audio ones used in the oral languages, are formed by hand shapes and manual or general body motion, lip movements and facial expressions. Their expressiveness facilitates human interaction and exchange of information not only in the existence of hearing-impaired people but also in situations where speech is impractical, e.g., in loud workspaces. However, efficient communication by these means is only feasible between specially trained interacting parties. In this context, automatic sign-to-text and text-to-sign translation can be viewed as the intermediate technological modules that can partially lift this restriction. Moreover automatic sign language recognition may have contributions across other are as as linguistics for the study of sign languages or for the semi-automated processing of corpora.

Early attempts on automatic Sign Language Recognition (SLR) were restricted to simple recognition tasks [Ong and Ranganath2005] similarly to cases of speech recognition a few decades ago. An informal correspondence of the word in spoken language is a sign unit, given that sign languages tend to be monosyllabic [Emmorey2002]. There are several metaphors between sign and speech recognition that allow for the exchange of methods between the two areas. However, there exist points of difference too. A diversity that also has practical effects concerns phonetic sub-units: There is not yet a well-defined unit equivalent to the phoneme in speech. Another difference concerns the multiple parallel cues that are articulated during sign language generation. In this paper, as far as the segmentation, modeling and recognition are concerned, we focus on automatic data-driven modeling of sub-units without any phonological or linguistic information.

The field of SLR is certainly in the focus of quite intense research lately [Ong and Ranganath2005]. It is considered to be a multilevel problem and it poses significant challenges regarding visual processing and information stream modeling for recognition. [Vogler and Metaxas2003] broke down signs into their constituent sub-units using the basic ideas of the Movement-Hold model [Liddell and Johnson1989] and applied successfully the so-called Parallel HMMs. [Bauer and Kraiss2001], on the other hand worked also at the sub-unit level exploring a data-driven approach for modeling the intra-sign units. They cluster independent frames utilizing K-means. [Fang et al.2004] and [Han et al.2009] have also proposed approaches for data-driven sub-unit modeling. They employed clustering by considering segments and not only independent frames as [Bauer and Kraiss2001] at the feature level, taking advantage of the dynamics that are essential in sign language. Modeling at the sub-unit level provides a powerful method in order to increase the vocabulary size and deal with more realistic data conditions.

The main objective of the proposed modeling approach is the automatic segmentation and construction of data-driven sub-units: these sub-units are the intra-sign primitive segments that shall be reused to reconstruct signs that share similar articulation parameters. We are inspired by both perceptual and linguistic evidence [Emmorey2002, Liddell and Johnson1989] on the functionality of the multiple cues. Among all cues the ones that we heavily exploit next are based on the planar (2D) coordinates of the dominant hand's centroid, and some of its products. We shall refer to these features from now on as the *movement-position* cues. Besides, movement and position are among the main characteristics that describe a sign [Emmorey2002].

Based on simple movement, position measurements, we proceed on the automatic sub-unit modeling of sign lan-

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

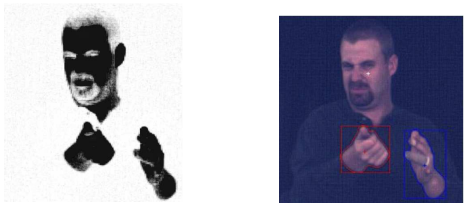
guage at the model level, that refers to the modeling of intra-sign segments. This modeling involves the synergy of the multiple cues and the modeling structure that these cues are incorporated: 1) the partitioning of segments into dynamic or static with respect to their dynamics; we employ for each sign unit, a model based segmentation at the state level, that apart from the segmentation assigns also labels to the segments. 2) The modeling of the static or dynamic segments depending on the label that they were assigned in the previous modeling step. Each type of segment shall be modeled by the cues and the model that are more appropriate for each case. Given the segmented sign we are equipped with a prosperous initialization step to face appropriately the modeling the dynamic vs. static intra-sign segments. For the case of dynamic segments, our goal is to cluster not the independent frames as if they were in a common pool [Bauer and Kraiss2001], neither the feature frames sequences as segments themselves at the feature level [Fang et al.2004,Han et al.2009]. Instead, we propose to hierarchically cluster whole dynamic models (HMMs) [Smyth1997] based on a similarity measure among models via their parameters. Another point to stress is that the models are first normalized wrt. 1) the initial segment's position, for each segment, and 2) the maximum scale of the movement's trajectory. These normalization steps are crucial, since by incorporating them we end up modeling the actual movement data independently to the existing mixed scales or initial positions: this makes the models more compact, increases the training data per model, and reduces the total number of models required. For the case of static segments, the main measurement that characterizes them is the one of position, corresponding to more clear postures on which the velocity has been on average close to zero. We evaluate the proposed methods on real data from the Boston-University continuous American Sign Language corpus (BU400) [Dreuw et al.2008]. In the experiments we explore a variety of feature streams and evaluate the efficacy of the proposed modeling scheme in preliminary automatic recognition experiments showing promising results. These experiments investigate the efficacy of the employed features, as well the integration of the multiple movement-position cues.

2. Visual Processing of Sign Language

2.1. Segmentation and Tracking

For the segmentation of the video frames we are based on the Geodesic Active Regions (GAR) approach [Paragios and Deriche2002], as this has been adapted on previous work [Diamanti and Maragos2008] for sign language processing. The GAR are deformable 2D contours, which evolve to minimize an energy functional, designed to meet the needs of the segmentation process. The intensity image is partitioned into two separable regions, one being the union of the skin-colored regions, and the other consisting of the rest of the image pixels, referred to as background. We adapt the GAR model to introduce a new force for skin segmentation.

$$F_{color} = \log\left((P_s(\vec{x}))/(P_b(\vec{x}))\right) + cg(I) \qquad (1)$$



(a) (b)

Figure 1: (a) Likelihood ratio per pixel belonging to skin or not, shown as a grayscale image. (b) Segmentation after employing GAR on the likelihood ratio map.

where I is the image, P_s, P_b denote the probability of a certain pixel $\vec{x}$ belonging to the skin or background regions, respectively, and $g(I)$ is the edge-detection stopping function. To estimate the probabilities P_s and P_b we employ two probabilistic models to account for the skin and background color, respectively. After the estimation of P_s and P_b by taking their ratio we result with a measure of a pixel belonging to skin. The above likelihood ratio map is then used as a force in the GAR model enforcing the curve to converge eventually to the edges that separate the skin region from the background. The result of the hand detection that we use is shown in Fig. 1. Due to the dynamic nature of sign language articulation, the skin color regions of interest may occlude each other. For these cases we employ techniques in order to disambiguate occlusions such as linear forward-backward prediction and template matching.

2.2. Feature Summary

Employing the segmentation and tracking process, we extract features related to the position and the movement. More specifically using the fitted ellipses on each hand we extract the features related to these ellipses such as x, y centroid coordinates. In addition, we construct features which are products from the x,y coordinates of the hands' centroids, such as the velocity $vel(t) = [\dot{x}; \dot{y}]$, the acceleration $acc(t) = [\ddot{x}; \ddot{y}]$ and the instantaneous direction $dir(t) = [\dot{x}; \dot{y}]/(\dot{x}^2 + \dot{y}^2)^{1/2}$. For the scope of our current modeling and recognition we are using only the x,y coordinates of the dominant hand centroid using as reference point the centroid of the signer's head and its aforementioned products.

3. Continuous Sign Language Recognition

We tackle the issue of sub-unit probabilistic modeling in order to deal with continuous sign language recognition. We propose 1) the organization of the modeling in a tree-like modeling structure that employs on each modeling level the *appropriate feature(s)* with the appropriate modeling depending on the functionality of the features; 2) the normalization of the features that are modeled: We focus in this way on the actual underlying phenomena we wish to tackle and avoid from getting our modeling consumed on mixed factors; 3) the incorporation of the dynamics *at the model level* – and not at the feature level of separate frames or sequences of frames' level. We consider that it is both 1) the *modeling structure* and 2) the modeling with *normalization*, that are important as it is discussed next.

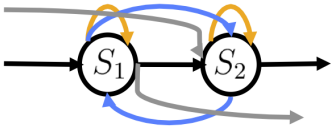


Figure 2: The 2-state HMM topology that is employed for segmentation and implicit classification of the segments.

3.1. Model-based Segmentation and Classification

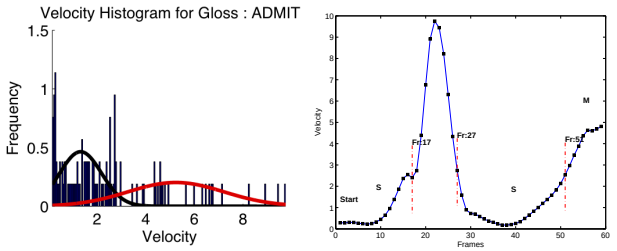
Modeling the Velocity Cue Our goal is to separate the so called from now on, "dynamic" from the "static" parts w.r.t. movement. This is the level of segmentation and classification of the segmented parts of the signal: dynamic parts shall correspond to movements and static parts to non-movements. This approach is inspired by linguistic modeling [Liddell and Johnson1989] of "movements and holds". We assume that movements correspond to high on average velocity, and non-movements to low relative velocity. Although the fuzziness of the 'high' and 'low' terms we appropriately incorporate them by adopting a suitable model-based approach. The feature that shall be utilized for this characterization is the *velocity*, whereas the *acceleration* could add further detail. The velocity feature vector consists of the dominant hand's centroid velocity that is constructed as the norm of the coordinate derivatives. Our goal is met if we consider the HMM structure of two states, as shown in Fig. 2. This allows the entrance and the exit from both states and the full transition from each state to the other, since the dynamic or static parts may alternate one another and do not obey any grammar rule.

Gloss Specific Modeling Next, we create one model for each gloss that is trained using all realizations of the specific gloss. Each HMM gloss model models the velocity profile of the corresponding gloss. Each one of the HMM states results in modeling a single velocity level. Given the population of data from large portions of the training set, the two state levels correspond to a low- and a high-level of the corresponding feature, i.e. velocity. This is further understood if we observe the velocity distribution over the different realizations for a specific gloss in Fig. 4(a). After training each HMM we perform a Viterbi alignment for each realization given the gloss resulting to the most probable *segmentation points* at the state level *together* with the labels of the velocity profiles. An example of segmentation obtained for one instance of the sign "ADMIT" is depicted in Fig. 4(b) for the feature level, whereas Fig. 3 shows the actual frames of the segments (subsampled).

Automatic vs. Manual Segmentation One way to evaluate the proposed segmentation approach is by comparing its output with the corresponding manual annotation by experts. At this point we show the results of a preliminary such effort. Figure 5 presents both the automatic and manual annotation[1] for a realization of the sign "ADMIT". For the automatic production of both segmentation points and the classification of the segments we make use of the velocity modeling providing two different labels. By comparing the results it seems that the automatic segmentation via the proposed approach is on average close to the manual segmentation points.

[1]The manual annotation has been provided by Annelies Braffort at CNRS-LIMSI.

Figure 3: Segmentation using the velocity cue for one instance of the sign "ADMIT". Each row corresponds to a different segment.



(a) (b)

Figure 4: (a) Velocity distribution (histogram) superimposed with the fitted (b) Segmentation shown superimposed on the velocity profile for an instance of the sign ADMIT.

The proposed model-based approach provides various advantages: 1) we get not only the segmentation but also the result of a classification since we have encapsulated implicitly the dynamic and static characteristics into the states of the same model. 2) Another asset is that we don't need to define any threshold manually (as other approaches for segmentation at the feature level), since these are handled inherently after setting the model parameters.

3.2. Modeling Dynamic Segments

We tackle next the issue of intra-sign sub-unit modeling at the HMM model level instead of the feature level. In this way we take advantage of the explicit dynamic modeling at the state level that the HMMs yield. Dynamic modeling is crucial for the modeling of movement. After all, HMMs have been employed successfully in other applications of sign language modeling too [Vogler and Metaxas2003]. Afterwards, a model level approach adds up a probabilistic viewpoint that can be further exploited, and finally fits well with the automatic recognition framework.

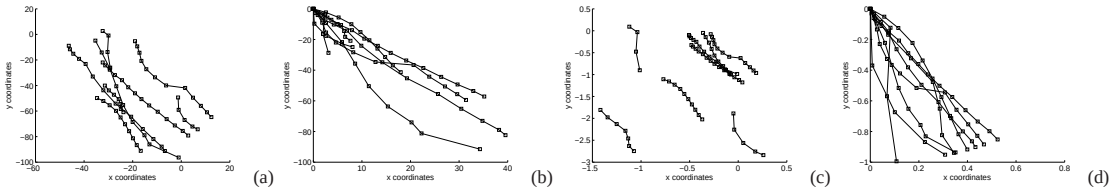


Figure 6: Trajectories of dynamic movements mapped onto the 2D signing space: (a) Without any normalization. (b) After normalization to the initial position. (c) After normalization to scale. (d) After normalization to both the initial position and scale.

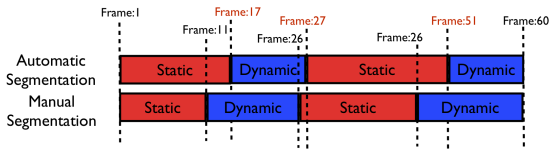


Figure 5: Automatic vs. Manual Segmentation and segments' classification for a realization of the sign "ADMIT".

3.2.1. Feature Normalization

Initial Position Our goal in this task is to model the dynamics of movement during the signs. The main feature for each dynamic segment is the *movement trajectory*. Each position sequence is initiated from the previous actual position that is arbitrary. The modeling of such features, leads to the consumption of the modeling effort due to the increased variance that the arbitrary initial positions of the movement trajectories introduce, so as to account for all different initial positions. This is encountered by normalizing the feature segments, each one with its corresponding initial position. This step results on the translation invariant movement modeling, i.e. independently to the initial position. An example of this normalization is presented in Fig. 6(a,b): we present the movement trajectories as they are mapped onto the initial 2D signing space before they are employed in the sub-unit construction process; we demonstrate the *same* trajectories with and without normalization. Moreover, normalization methods are well-known in the ASR community [Rabiner1989]. Another advantage of the normalization is the increase of the data requirements per model and at the same time we decrease the total number of models required.

Scale Similarly to the above, scale also affects the modeling of movement trajectories. Scale normalization of each movement results in scale invariant modeling, increase of data examples per model, end more efficient modeling with less models. At the same time, we do keep the scale parameter itself for further incorporation and modeling as a separate feature. An example of this normalization is shown in Fig. 6(a,c): the figure shows the same segments before they are employed in the sub-unit construction procedure with and without normalization. Finally Fig. 6(d) shows the same trajectories after both scale and initial position normalization. It shall be next more efficient to incorporate these normalized segments in the corresponding HMM models instead of the non-normalized, since they shall capture the actual dynamics independently to both the initial position (compare with Fig. 6(a,c)), and the maximum scale (compare with Fig. 6(b,c)).

3.2.2. HMM Clustering

We initialize the segments by first applying the segmentation procedures, as it has been described in the previous Section 3.1.. Given that the segments contain movement our goal is to cluster whole dynamic models (HMMs) [Smyth1997] that correspond to these movement trajectories. Clustering states at the model level has been employed succesfully in ASR applications. Herein we cluster not just the states, but *whole* HMMs. Thus, we fit N 3-state HMMs, one for each individual sequence or segment S_i, $i = 1 \ldots N$. Afterwards we use a similarity measure between pairs of HMM models $H_k, k = 1, 2$, by adopting among proposed approaches in the literature [Juang and Rabiner1985] that are based on the Kullback-Leibler divergence. Similarly we employ

$$D(H_1, H_2) = \sum_{O_i^{H_1}} \frac{1}{T_i} \log \frac{P(O_i^{H_1}|H_1, S_i^{H_1})}{P(O_i^{H_1}|H_2, S_i^{H_2})}$$

where $O_i^{H_k}$ corresponds to the observation sequences that have been generated from each of the H_k model, of length T_i and $\log P(O_i^{H_k}|H_k, S_i^{H_k})$ is the log probability of the observation given the HMM model and the optimum state sequence $S_i^{H_k}$, for $k = 1, 2$. The sequences used to compute the log probabilities are generatively constructed by each H_k model employing 20 sequences. The distance similarity matrix among all models is exploited via an agglomerative hierarchical clustering algorithm. We end up with the total likelihood of the specific clustering, given the number of clusters employed.

3.3. Dynamic Sub-Units for Each Feature

Next, we explore the modeling of features that are appropriate for dynamic segments modeling. The output of the clustering on the HMM level corresponds to a partition on the feature space. Each cluster in this partition is defined as a distinct sub-unit, presented next for different cases of features.

Movement Trajectories After the normalization steps each segment is modeling the plain normalized trajectory in the 2D planar signing space. We show in Fig. 7(b) indicative sub-units: these are clusters that have been constructed by the HMM hierarchical clustering at the model level, and are then mapped onto the 2D signing space. This mapping retains the sub-unit identity that is encoded by means

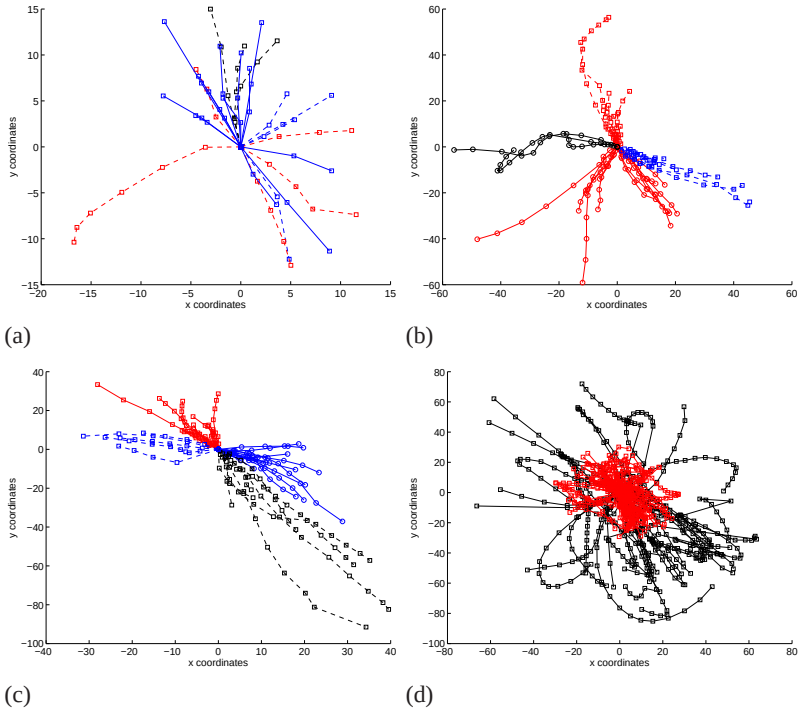


Figure 7: The trajectories for different sub-units as they have been mapped on the 2D signing space. With different color we represent different sub-units that correspond to the different clusters. (a)Trajectories of sub-units obtained using as feature the movement trajectories (P) *without* any normalization. (b)Trajectories of sub-units that incorporate both scale and initial position normalization. (c) Trajectories of sub-units that incorporate the Direction cue after normalization of the trajectories to the initial position. (d) Trajectories for two different sub-units that correspond to different scales.

of color in the presented figures. In Fig. 7(a) we show a case of sub-units as a result of clustering, but without the normalization steps. It is evident by comparing with the previous case (Fig. 7(b)) that the modeling is much looser since the models are consumed totally on the explanation of the different initial positions or scales. The non-normalized constructed sub-units as shown mapped on the original 2D signing space make it hard to understand what exactly each cluster represents. The clusters after normalization actually implicitly incorporate direction information. This is something expected as the modeling contains the direction information encapsulated with the geometry of the whole trajectory. As a matter of fact, each model's state from the first to the last explains points in the trajectory that have on average increasing distance from the segments initial position.

Scale We may have normalized with the scale of each trajectory, being the maximum distance of all points in a trajectory, but this information shall not be disregarded. It is modeled on its own in order to investigate how it affects the modeling. We show in Fig. 7(d) indicative sub-units: these are clusters that have been constructed by the clustering at the model level, and are afterwards mapped on the 2D signing space. This mapping retains the sub-unit identity or equivalently the cluster index that is encoded by means of color in the presented figures. The presented sub-units are presented to model trajectories entirely based on their scale independently to their direction.

Direction The sub-units constructed by the direction feature show similar results as the ones that model the normalized trajectories. As expected each sub-unit consists of movements with similar on average direction over time. Figure 7(c) shows indicative examples of movements over the same or different clusters having similar on average or different directions respectively.

3.4. Dynamic Sub-Units for Multiple Features

In the previous section we presented the sub-unit construction for the dynamic segments using a single cue at each time for each sub-unit type. Thus we constructed sub-units that account for single different characteristics of a movement such as the direction, the scale or the movement trajectory. Next, we explore sub-unit construction for the dynamic segments by using for each sub-unit multiple cues. This extension is seamlessly incorporated given the discussed framework. As mentioned in Section 3.2.2. the sub-unit clustering is based on HMMs. In order to account for multiple features during sub-unit construction we employ a multi-stream HMM instead of one simple single-stream HMM. More specifically by incorporating both direction and scale into a multi-stream HMM we create multiple-cue sub-units that model movements based jointly on their direction and their scale. This sub-unit construction is shown via the corresponding trajectories that correspond to the distinct sub-units of Fig.8. In these, instead of the different directions (as seen in Fig. 7(c)) we have created sub-units that explain at the same time the direction for each one of the different scales.

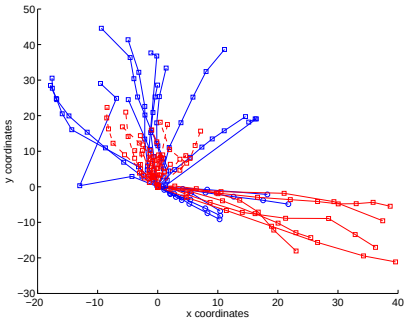


Figure 8: The trajectories for four different sub-units mapped on the 2D signing space represented with different color/marker. Sub-units account for the multiple-cues of both direction and scale of the dynamic segments.

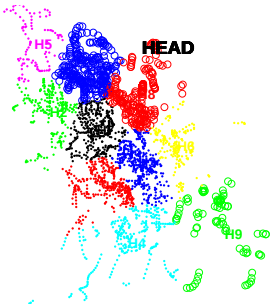


Figure 9: Partitioning of the 2D signing space by K-means. Different colors discriminate the sub-unit.

3.5. Modeling Static Segments

Given the discrimination and separate modeling of the dynamic segments, the remaining segments correspond to the low velocity profiles. We model *only* these static positions and not all positions as they lay across time within movement segments. After applying k-means clustering to the non-normalized positions we get a partitioning of the static positions relative to the head of the signer. Figure 9 shows the constructed sub-units as they are mapped on the 2D signing space.

4. Lexicon: Recombining the Dynamic and Static Segments

4.1. Lexicon Construction

After decomposing the dynamic and static segments for separate modeling, we re-compose them via the lexicon so as to form the complete signs via a concatenation of the sub-units at a symbolic level. Each sub-unit is in this case a 'symbol' that is uniquely identified by the feature that has been employed for its construction and the index that has been assigned during the clustering procedure. This lexicon is completely data-driven and does not employ any linguistic information. The lexicon re-composes the two levels of 1) the Dynamic Movement Segments (D) and the 2) Static Position Segments (S). An example of three different lexicons that have been obtained using Position (P) for the static segments and Direction (D) or Movement Trajectories (SPn) or Scale (S) for the dynamic segments respectively is shown in Fig. 10.

Figure 10: Lexicon sample for two different type of features (from left to right) SPn, S for the dynamic and P for static segments.

4.2. Multiple Pronunciations

The realization of signs during continuous natural signing introduces factors that increase the articulation variability. Among the reasons responsible for these multiple pronunciations is the existence of features that do not remain constant during each gloss articulation. For instance there might be cases of the same gloss that is represented by the same sequence of movements but in multiple realizations that involve different initial positions. An example of varying pronunciation for a specific gloss is illustrated by the sample lexicons shown in Fig. 10. Each line in a lexicon sample consists of 1) a gloss identifier concatenated by 2) an index that corresponds to the pronunciation realization case. Figure 10 includes two cases of features for the Dynamic segments combined in all cases with the Position feature for the Static segments. In the shown example, gloss "BECAUSE" is being mapped to three different sub-unit sequences. These specific sub-unit sequences share the first sub-unit of static modeling, while the second one adds at a movement sub-unit, e.g. MSPn1, and the third one adds another static sub-unit.

4.3. Sub-Unit sequences to Multiple Glosses Mapping

Among the reasons responsible for these multiple pronunciations is the non-sufficient during this stage of modeling w.r.t. the features employed. For instance there might be cases of glosses that are represented by the same sequence of movements-positions but they involve different handshape configurations that are not accounted yet. Such a case are signs "WITH" and "FOOTBALL" which share common sequence of movements-positions but different handshape configuration. Another factor is the model order we employ, or in other words how loose is the sub-unit clustering we apply. For example if we use a small number of clusters in order to represent all space of movements, although we might have used sufficient features, multiple different movements shall be mapped to the same sub-unit creating looser models.

5. Recognition Experiments

Experimental configuration

In the experiments described we use only the front camera video stream. Among the whole corpus, we restrict our processing on six videos that contain stories narrated from a single signer[2]. We utilize 50 randomly selected signs

[2]Videos are identified namely as: `accident`,

among the most frequent ones. We employ 60% of the data for training and 40% for testing. This partitioning samples data from all videos, and among all realizations per sign in order to equalize gloss occurrence. For the evaluation of the recognition results we employ the standard measure of accuracy in the sub-unit level and the gloss level.

Experiments: Next we describe recognition experiments that evaluate the main aspects discussed. 1) We examine the incorporation of the segmentation and classification component referred to as Static vs. Dynamic Classification; this step affects also the adapted modeling w.r.t the employed multiple cues and clustering. 2) Another contribution discussed is the feature normalization for the Dynamic parts that on its turn affects both the modeling and the recognition results. 3) Finally, we further evaluate the incorporation of multiple cues in the Dynamic parts modeling. The employed cues are encoded as Direction (D), Movement Trajectory after scale and initial-position normalization (SPn), Scale (S) and non-normalized Position (P). The results contain both gloss-level and sub-unit level accuracies.

Number of Sub-Units: The number of sub-units we use in each case is depended on the existing experimental dataset and on prior linguistic information. The dynamic segments employ 24 sub-units given motivation on the different type of movements (8 for each of straight or curved or other more complex movements). We use four sub-units for scale modeling and 22 sub-units for the static segments' sub-units which corresponds to different but arbitrary places of articulation. These numbers imply the total number of sub-units employed on each recognition experiment described next and are shown on Table 1. Note that for tasks that are to be compared we employ equal number of sub-units. More sub-units imply a more complex task. Another point to stress, (see also the discussion in Section 4.), is that the gloss level results should be viewed given the "single sub-unit sequence mapping to multiple glosses" due to the missing cues (e.g. handshape). The above gloss accuracy considers a gloss as correct if it exists in the set of targets of the specific sub-unit sequence. This is the case *even* if other glosses are present in the same set. That is, the recognition performance evaluation functions towards our favor even if there are multiple glosses mapped from a specific sub-unit sequence.

Single-Stream Continuous SL Recognition: Here, we evaluate the efficacy of the various movement-position cues employed in single stream recognition experiments and at the same time without incorporating the Dynamic-Static Classification. Figure 11(c,d) shows the results for the four single cue cases: P, D, S and SPn. These results should be seen under the following point of view. The sub-unit accuracy is dependent each time on the complexity of the task: For the case of S the employed number of sub-units is much lower compared to the other single cue cases thus the high performance is for a much easier task (see Table 1).

Dynamic-Static Segmentation and Classification: In this case we compare two variants. The first variant evaluates the modeling that exploits the Dynamic-Static Classification (DSC) obtained during segmentation. The second one, corresponds to the case in which we employ only the segmentation *without* the Dynamic-Static Classification (no-DSC) of the segments. For the first case above (DSC) we employ for the Dynamic segments the cues of D, SPn and S. On the contrary for the static segments we employ only the P cue. For the second case of no-DSC all segments share the same cue. For this case among all multiple-cue combinations we show the one that performs best (SPn-S-P). The incorporation of the DSC is encoded in the Fig. 11 by the "+" symbol, e.g. A+B shall correspond to the A cue for the dynamic modeling and the B cue for the static. Where two cues are concatenated by "-" as in A-B, this corresponds to the plain concatenation via multiple streams.

First, we should note that by comparing the single cue experiments with the DSC multiple cue case the latter show improved performance. The overall recognition performance for the DSC case Fig. 11(a,b). outperforms the no-DSC case Fig. 11(c,d). More specifically, using the Position (P) cue naively combined with other features (S, SPn, D) implies increased model variance. On the contrary, see Fig.11(a,b), when the cues (SPn, D, S) are modeled plainly in the dynamic parts and the Position cue (P) is only incorporated on the static modeling the results are improved significantly.

Feature Normalization: The importance of normalization is observed for the no-DSC case since the SPn cue outperforms the non-normalized P cue. For the multiple-cue DSC case on which the P is better incorporated, the SPn+P performs much higher than the non-reported accuracy of P+P (i.e. non-normalized cue in the Dynamic modeling resulting on 38% gloss accuracy).

Multiple Cues in Dynamic Modeling: By incorporating multiple cues in the Dynamic modeling as shown in the DSC case, see for instance D-S+P and SPn-S+P compared to S+P, SPn+P, D+P in Fig.11(a,b), there are slight improvements, that should be considered given the number of sub-units reported in Table 1.

6. Conclusions

We propose a modeling structure that incorporates movement-position cues in an unsupervised manner. Each cue is adopted with the appropriate modeling given its functionality during sign language articulation. The modeling is based on the discrimination between Dynamic and Static cases of the movement-position cues, which provides a segmentation and classification of the segments. Secondly, for each type of modeling we incorporated the appropriate cues after normalization. The dynamic sub-units are constructed by clustering at the *model level*. The evaluation of the proposed multiple-cue modeling approach in recognition experiments on the BU400 continuous sign language corpus shows promising results. However, in order to be able to reach more mature conclusions, we shall 1) incorporate phonological and linguistic information, 2) as well as handshape information, that is currently explored via a model based approach and shall be integrated in a common framework.

`biker_buddy`, `boston_la`, `football`, `lapd_story` and `siblings`.

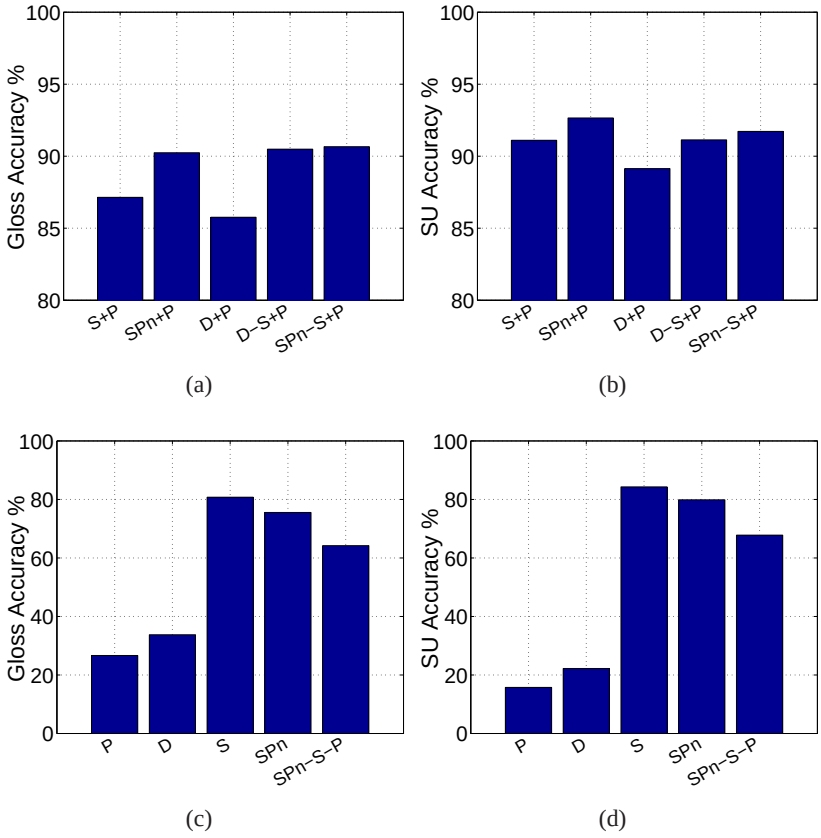


Figure 11: Recognition performance:(a,b) Gloss and Sub-unit accuracy of multiple cues while incorporating Dynamic-Static Classification (DSC), (c,d) Gloss and Sub-unit accuracy of single and one multiple cue without incorporating DSC.

Table 1: Feature identifier corresponding to the recognition experiments and number of sub-units employed.

Feature	S	D	SPn	P	S+P	SPn-S-P	SPn+P	D+P	D-S+P	SPn-S+P
# SUs	4	46	46	46	4+22(46)	24x4+22(118)	24+22(46)	24+22(46)	24x4+22(118)	24x4+22(118)

7. References

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

O. Diamanti and P. Maragos. 2008. Geodesic active regions for segmentation and tracking of human gestures in sign language videos. In *icip*.

P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. 2008. Benchmark databases for video-based automatic sign language recognition. In *Proc. Int'l Conf. on Language Resources and Evaluation (LREC)*, May.

K. Emmorey. 2002. *Language, cognition, and the brain: insights from sign language research*. Erlbaum.

G. Fang, X. Gao, W. Gao, and Y. Chen. 2004. A novel approach to automatically extracting basic units from chinese sign language.

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

B. H. Juang and L. R. Rabiner. 1985. A probabilistic distance for hidden markov models. *AT & T Technical Journal*.

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

S. Ong and S. Ranganath. 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. *ieeetpami*, 27(6):873–891.

N. Paragios and R. Deriche. 2002. Geodesic Active Regions: A New Framework to Deal with Frame Partition Problems in Computer Vision. *Journ. of Vis. Commun. and Image Repres.*, 13(1/2):249–268.

L. R. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.

P. Smyth. 1997. Clustering sequences with hidden markov models. In *In Advances in Neural Information Processing Systems*, volume 9, pages 648–654.

C. Vogler and D. Metaxas. 2003. Handshapes and movements: Multiple-channel american sign language recognition. In *Gesture Workshop*, pages 247–258.