Hand Tracking and Affine Shape-Appearance Handshape Sub-units in Continuous Sign Language Recognition

Anastasios Roussos, Stavros Theodorakis, Vassilis Pitsikalis and Petros Maragos[*]

School of E.C.E., National Technical University of Athens, Greece

Abstract. We propose and investigate a framework that utilizes novel aspects concerning probabilistic and morphological visual processing for the segmentation, tracking and handshape modeling of the hands, which is used as front-end for sign language video analysis. Our ultimate goal is to explore the automatic Handshape Sub-Unit (HSU) construction and moreover the exploitation of the overall system in automatic sign language recognition (ASLR). We employ probabilistic skin color detection followed by the proposed morphological algorithms and related shape filtering for fast and reliable segmentation of hands and head. This is then fed to our hand tracking system which emphasizes robust handling of occlusions based on forward-backward prediction and incorporation of probabilistic constraints. The tracking is exploited by an Affine-invariant Modeling of hand Shape-Appearance images, offering a compact and descriptive representation of the hand configurations. We further propose that the handshape features extracted via the fitting of this model are utilized to construct in an unsupervised way basic HSUs. We first provide intuitive results on the HSU to sign mapping and further quantitatively evaluate the integrated system and the constructed HSUs on ASLR experiments at the sub-unit and sign level. These are conducted on continuous SL data from the BU400 corpus and investigate the effect of the involved parameters. The experiments indicate the effectiveness of the overall approach and especially for the modeling of handshapes when incorporated in the HSU-based framework showing promising results.

1 Introduction

Sign languages convey information via visual patterns and serve as an alternative or complementary mode of human communication or human-computer interaction. The visual patterns of sign languages, as opposed to the audio patterns used in the oral languages, are formed mainly by handshapes and manual motion, as well as by non-manual patterns. The hand localization and tracking in a sign video as well as the derivation of features that reliably describe the pose and configuration of the signer's hand are crucial for the overall success of an automatic Sign Language Recognition (ASLR) system. Nevertheless, these

tasks still pose several challenges, which are mainly due to the great variation of the hand's 3D shape and pose.

Many approaches of hand detection and tracking have been reported in the literature, e.g. [1–4]. As far as the extraction of features of the hand configuration is concerned, several works use geometric measures related to the hand, such as shape moments [5]. Other methods use the contour that surrounds the hand in order to extract various invariant features, such as Fourier descriptors [6]. More complex hand features are related to the shape and/or the appearance of the hand [1, 3, 4]. Segmented hand images are normalized for size, in-plane orientation, and/or illumination, and Principal Component Analysis (PCA) is often applied for dimensionality reduction, [7, 8]. In addition, Active Shape and Appearance Models have been applied to the hand tracking and recognition problem [9, 10]. Apart from methods that use 2D hand images, some methods are based on a 3D hand model, in order to estimate the finger joint angles and the 3D hand pose, e.g. [11].

In the higher level, ASLR provides challenges too. In contrast with spoken languages, sign languages tend to be monosyllabic and poly-morphemic [12]. A diversity that also has practical effects concerns phonetic sub-units: A sign unit has a different nature when compared to the corresponding unit in speech, i.e. the phoneme. This concerns the multiple parallel cues that are articulated simultaneously during sign language generation. Handshape is among the important phonetic parameters that characterize the signs together with the parameters of movement and place-of-articulation. In addition, modeling at the sub-unit level [13, 14] provides a powerful method in order to increase the vocabulary size and deal with more realistic data conditions.

In this paper, we propose a new framework that incorporates skin-color based morphological segmentation, tracking and occlusion handling, hand Shape - Appearance (SA) modeling and feature extraction: these are all integrated to serve the automatic construction of handshape sub-units (HSU), on their employment in ASLR. Our contribution consists of the following: 1) In order to detect and refine the skin regions of interest, we combine a basic probabilistic skin-color model with novel shape filtering algorithms that we designed based on mathematical morphology [15]. 2) We track the hands and the head making use of forward-backward prediction and incorporating rule-based statistical prior information, 3) We employ SA hand images for the representation of the hand configurations. These images are modeled with a linear combination of affine-free eigenimages followed by an affine transformation, which effectively accounts for modest 3D hand pose variations. 4) Making use of the eigenimage weights after model fitting, which correspond to the handshape features, we construct in an unsupervised way data-driven handshape sub-units. These are incorporated in ASLR as the basic phonetic HSUs that compose the different signs. 5) We evaluate the overall framework on the BU400 corpus [16]. In the experiments we investigate the effectiveness of the SA modeling and HSU construction in the task of ASLR that refers to the modeling of *intra-sign* segments by addressing issues such as: a) the variation of involved parameters, as for instance the model order during sub-unit construction, and the employment of initialization during clustering; b) the vocabulary size. c) Finally, we provide intuition concerning the lexicon and the sub-unit to sign maps via qualitative and quantitative ex-
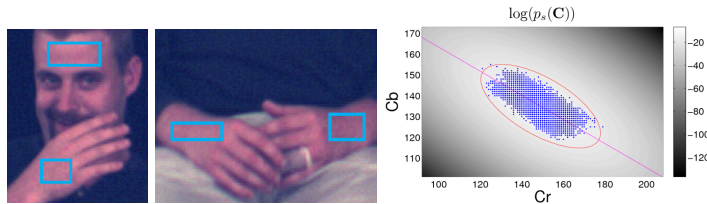
Fig. 1. Skin color modeling. (*Left, Middle*) Examples of manual annotations of skin regions (rectangles) that provide training samples of skin color. (*Right*) Training samples in the C_bC_r space and fitted pdf $p_s(C)$. The ellipse bounds the colors that are classified to skin, according to the thresholding of $p_s(C(x))$. The line determines the projection that defines the mapping g used in the SA images formation.

periments. Under these points of view the conducted experiments demonstrate promising results.

## 2 Visual Front-End Processing

2.1 Segmentation and Skin Detection

Probabilistic Skin Color Modeling First of all, a preliminary estimation of the hands and head locations is derived from the color cue, similarly to various existing methods [1–3]. For this, we assume that the signer wears long sleeved clothes and the colors in the background differ from the skin color. More precisely, we construct a simple skin color model in the YC_bC_r space and we keep the two chromaticity components C_b,C_r. In this way we obtain some degree of robustness to illumination changes [17]. We assume that the C_bC_r values $C(x)$ of skin pixels follow a bivariate gaussian distribution $p_s(C)$, which is fitted using a training set of skin color samples from manually annotated skin areas of the signer, Fig.1. A first estimation of the skin mask S_0 is thus derived by a thresholding of $p_s(C(x))$ at every pixel x, Figs.1-*right*, 2(b). The corresponding threshold constant is determined so that a percentage of the training skin color samples are classified to skin. This percentage is slightly smaller than 100%, in order to cope with training samples outliers.

Morphological Refinement of the Skin Mask The extracted skin mask S_0 may contain spurious regions as well as holes inside the head area because of the signer's eyes or potential beard. For these reasons, we propose a novel morphological algorithm to regularize the set S_0: First, we use the concept of *holes* $\mathcal{H}(S)$ of a binary image S; these are defined as the set of background components which are not connected to the border of the image frame [15, 18]. In order to fill also some background regions that are not holes in the strict sense but are connected to the image border passing from a small "canal", we apply the following *generalized* hole filling that yields a refined skin mask estimation S_1:

$$S_1 = S_0 \cup \mathcal{H}(S_0) \cup \{\mathcal{H}(S_0 \bullet B) \oplus B\} \tag{1}$$

where B is a structuring element of small size and $\oplus$ and $\bullet$ denotes dilation and closing respectively. For efficiency reasons, we chose B to be square instead

(a) Input   (b) S_0 (c) S_2 (d) $S_2 \ominus B_c$ (e) segmented S_2
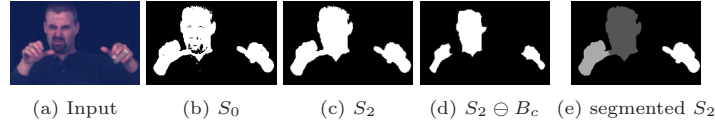
Fig. 2. Indicative results of the skin mask extraction and segmentation system.

of disk, since dilations/erosions by a square are much faster to compute while showing an almost equal effectiveness for this problem.

Afterwards, in order to remove potential spurious regions, we exploit prior knowledge: the connected components (CCs) of relevant skin regions 1) can be at most three, corresponding to the head and the hands, and 2) cannot have an area smaller than a threshold A_{min}. Therefore, we apply an area opening with a varying threshold value: we find all the CCs of S_1, compute their areas and finally discard all the components whose area is not on the top 3 or is less than A_{min}. This yields the final estimation S_2 of the skin mask, Fig. 2(c).

Morphological Segmentation of the Skin Mask Since the pixels of the binary skin mask S_2 correspond to multiple body regions, next we segment it, in order to separate these regions, whenever possible. For this, we have designed the following method. In the frames where S_2 contains 3 CCs, these yield directly an adequate segmentation. However, the skin regions of interest may occlude each other, which makes S_2 to have less than 3 CCs. In many such cases though, the occlusions between skin regions are not essential: different regions in S_2 may be connected via a thin "bridge", Fig. 2(c), e.g. when one hand touches the other hand or the head. Therefore we can reduce the set of occluded frames by further segmenting some occluded regions based on morphological operations as follows:

If S_2 contains N_{cc} connected components with $N_{cc} < 3$, find the CCs of $S_2 \ominus B_c$ (e.g. Fig. 2(d)) for a structuring element B_c of small size and discard those CCs whose area (after a dilation with B_c) is smaller than A_{min}. A number of remaining CCs not bigger than N_{cc} implies the absence of a thin connection, thus does not provide any occlusion separations. Otherwise, use each one of these CCs as the seed of a different segment and expand it in order to cover all the region of S_2. For this we propose a *competitive reconstruction opening* (see Fig. 2(e)), this is the result of an iterative algorithm, where in every step 1) each evolving segment is expanded using its conditional dilation by the 3×3 cross relative to S_2, 2) the pixels that belong to more than one segment are determined and excluded from all segments. This means that the segments are expanded inside S_2 but their expansion stops wherever they meet other segments. This procedure converges since after some steps the segments remain unchanged.

2.2   Tracking and Occlusion handling

After employing the segmentation of the skin mask S_2, we tackle the issue of hands/head tracking. This consists of 1) the assignment of one or multiple body-part labels, *head, left* and *right hand*, to all the segments of every frame and 2) the estimation of ellipses at segments with multiple labels (occluded). For that, we distinguish between two cases: the segmentation of S_2 yielded a) 3 segments in the *non-occlusion case* and b) 1 or 2 segments in the *occlusion case*.

(a)	(b) HR	(c) HR	(d) HRL	(e) HRL	(f) HR

Fig. 3. Hands & head tracking in a sequence of frames where occlusion occurs (b-f), among Head (H), Right (R) or Left (L) hand.

Non-Occlusion case: The segment with the biggest area is assigned the label *head* assuming that its area is always larger than hands. For the hands' labels, given that they have been assigned to the previous frames, we employ a linear prediction of the centroid position of each hand region taking into account the 3 preceding frames; the predictor coefficients correspond to a model of constant acceleration. Then, we assign the labels based on the minimum distances between the predicted positions and the centroids of the segments. We also fit one ellipse on each segment assuming that an ellipse can coarsely approximate the hand or head contour [2]. We plan to employ the fitted ellipses in cases of occlusions.

Occlusion case: Using the parameters of the body-part ellipses already computed from the last 3 preceding frames, we employ similarly to the previous case the linear forward prediction for all ellipses parameters of the current frame. Due to the sensitivity of this linear estimation with respect to the number of the consecutive occluded frames, non-disambiguated cases still exist. We face this issue by obtaining an auxiliary centroid estimation of each body-part via template matching of the corresponding image region between consecutive frames. Then, we repeat the prediction and template matching estimations backwards in time through the reverse frame sequence. Consequently, forward and backward prediction, are fused yielding a final estimation of the ellipses' parameters for the signer's head and hands. Fig.3 depicts the tracking result in a sequence of frames with non-occluded and occluded cases. We observe that our system yields an accurate tracking even during occlusions.

Statistical parameter setting: The aforementioned front-end processing involves various parameters. Most of them are derived *automatically* by preprocessing some frames of the video(s) of the specific signer. For this, we consider non-occluded cases of frames on which we compute the following statistics. By adopting gaussian models we train the probability density functions p_H, p_{RL} of the signer's head and hand areas respectively. We also compute the maximum displacement per frame d_{max} and the hand's minimum area A_{min}.

3   Affine Shape-Appearance Handshape Modeling

Our next goal is to extract hand configuration features from the signer's *dominant* hand, which is defined manually. For this purpose, we use the modeling of hand's 2D shape and appearance that we recently proposed in [19]. This modeling combines a modified formulation of Active Appearance Models [10] with an explicit modeling of modest pose variations via incorporation of affine image transformations.
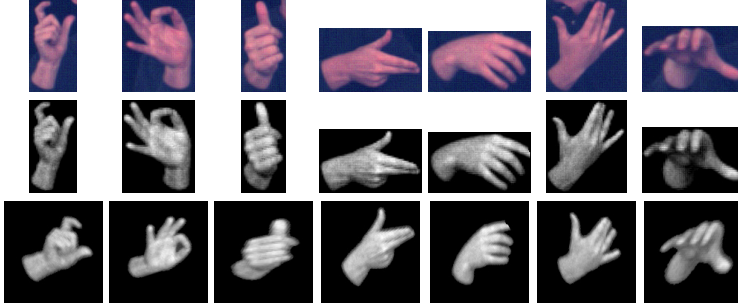
Fig. 4. (*Top row*) Cropped images $I_k(x)$ of the hand, for some frames k included in the 200 samples of the SAM training set. (*Middle row*) Corresponding SA images $f_k(x)$. (*Bottom row*) Transformed $f(W_{p_k}(x))$, after affine alignment of the training set.

First, we employ a hybrid representation of both hand shape and appearance, which does not require any landmark points: If $I(x)$ is a cropped part of the current color frame around the hand mask M, then the hand is represented by the following *Shape-Appearance (SA) image*: $f(x) = g(I(x))$, if $x \in M$ and $f(x) = -c_b$ otherwise. The function $g : \Re^3 \rightarrow \Re$ maps the color values of the skin pixels to a value that is appropriate for the hand appearance representation (e.g. we currently use the projection of the C_bC_r values on the principal direction of the skin gaussian pdf, Fig. 1). c_b is a background constant that controls the balance between shape and appearance: as c_b gets larger, the appearance variation gets relatively less weighted and more emphasis is given to the shape part. Figure 4-*middle* shows examples on the formation of hand SA images.

Further, the SA images of the hand, $f(x)$, are modeled by a linear combination of predefined variation images followed by an affine transformation:

$$f(W_p(x)) \approx A_0(x) + \sum_{i=1}^{N_c} \lambda_i A_i(x), \; x \in \Omega \tag{2}$$

$A_0(x)$ is the mean image, $A_i(x)$ are N_c eigenimages that model the linear variation; W_p is an affine transformation with parameters $p \in \Re^6$. The affine transformation models similarity transforms of the image as well as small 3D changes in pose. It has a highly nonlinear impact on the SA images and drastically reduces the variation that is to be explained by the linear combination part. The parameters of the model are p and $\lambda = (\lambda_1 \cdots \lambda_{N_c})$, which are considered as features of hand pose and shape respectively.

A specific model of hand SA images is defined from images of the linear combination of the model, $A_i(x)$, $i = 0, .., N_c$. In order to train this model, we employ a representative set of handshape images, Fig. 4-*top*. Given this selection, the training set is constructed from the corresponding SA images. In order to exclude the variation that can be explained by the affine transformation part of the model, we apply an affine alignment of the training set by using a generalization of the procrustes analysis of [10], Fig. 4-*bottom*. Afterwards, $A_i(x)$ are learned using Principal Component Analysis (PCA) on the aligned set and keeping a relatively small number ($N_c = 25$) of principal components. In the PCA results of Fig. 5, we observe that the influence of each eigenimage at the modeled hand SA image is fairly intuitive.
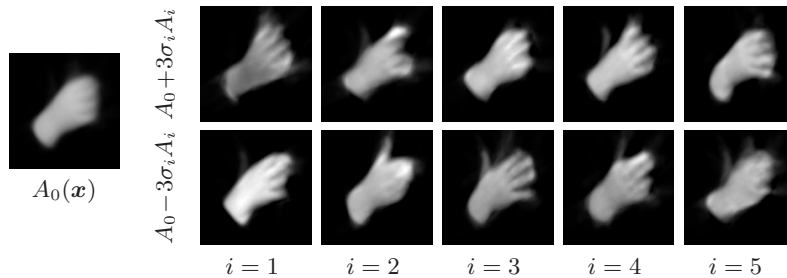
Fig. 5. PCA-based learning of the linear variation images of Eq.(2): Mean image $A_0(\boldsymbol{x})$ and variations in the directions of the first 5 eigenimages.
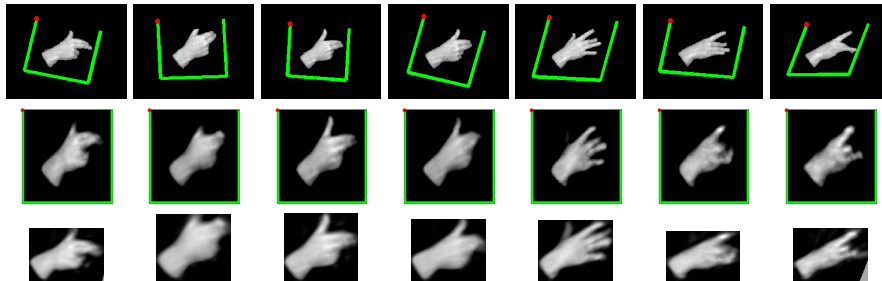

Fig. 6. SA model Fitting. (*Top*) SA images and rectangles determining the optimum affine parameters $\boldsymbol{p}$. (*Middle*) Reconstructions at the SA model domain determining the optimum weights $\boldsymbol{\lambda}$. (*Bottom*) Reconstructions at the domain of input images.

Finally, we extract hand features from the tracked region of the dominant hand at every frame via the SA model fitting. We find the optimum parameters $\boldsymbol{p}$ and $\boldsymbol{\lambda}$ that generate a model-based synthesized image that is "closest" to the corresponding hand SA image $f(\boldsymbol{x})$. Thus, we minimize the energy of the reconstruction error (evaluated at the model domain):

$$\sum_{\boldsymbol{x}} \left\{ A_0(\boldsymbol{x}) + \sum_{i=1}^{N_c} \lambda_i A_i(\boldsymbol{x}) - f(W_{\boldsymbol{p}}(\boldsymbol{x})) \right\}^2 , \tag{3}$$

simultaneously wrt $\boldsymbol{p}$ and $\boldsymbol{\lambda}$. This nonlinear optimization problem is solved using the Simultaneous Inverse Compositional (SIC) algorithm of [20]. We initialize the algorithm using the result from the previous frame. For the first frame of a sequence, we use multiple initializations, based on the hand mask's area and orientation, and finally we keep the result with the smallest error energy. Note that we consider here only cases where the hand is not occluded. In most of these cases, our method yields an effective fitting result, without any need of additional constraints or priors on the parameters. Figure 6 demonstrates fitting results. We observe that the results are plausible and the model-based reconstructions are quite accurate, despite the relatively small number $N_c = 25$ of eigenimages. Also, the optimum affine transforms effectively track the changes in the 3D pose.

Note that in works that use the *HOG descriptors*, e.g. [3,4], the components of shape and appearance are also combined. However, in contrast to these approaches, the proposed method offers a direct control on the balance between these two components. In addition, unlike to [3,8], the used handshape features

Mean Shape   Instances
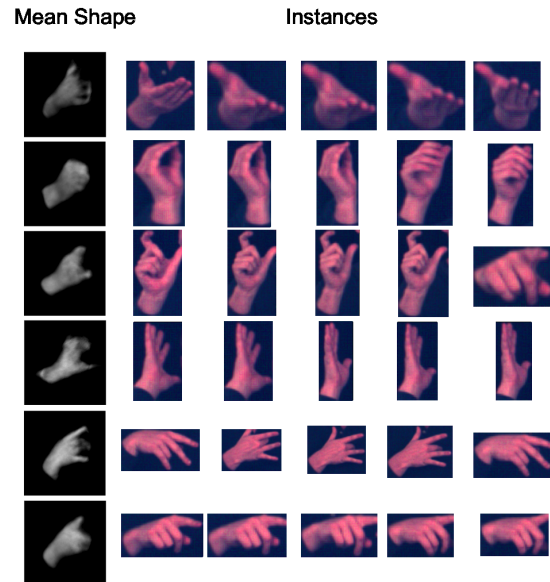
Fig. 7. Rows correspond to handshape sub-units. Left: mean shape-appearance reconstructed images of the centroids for the corresponding clusters. Next follow five indicative instances of the handshapes assigned to each centroid.

$\boldsymbol{\lambda}$ are invariant to translation, scaling and rotation within the image plane. This property holds also for the methods of [1, 7], but a difference is that in our model the features are also invariant to modest changes in the 3D hand pose. Such changes affect *only* the fitted affine transform parameters $\boldsymbol{p}$.

4 Handshape Sub-unit Based Recognition Framework

Our sign language recognition framework consists of the following: 1) First, we employ the handshape features produced by the visual front-end as presented in the Section 3. 2) Second, follows the sub-unit construction via clustering of the handshape features. 3) Then, we create the lexicon that recomposes the constructed handshape sub-units (HSU) to form each sign realization. This step provides also the labels for the intra-sign sub-units. 4) Next, the HSUs are trained by assigning one GMM to each one of them. 5) Finally, for the testing at the handshape sub-unit level we employ the sign-level transcriptions and the created labels in the lexicon.

4.1 Handshape Sub-unit Construction

We consider as input the visual front-end handshape features, the sign level boundaries and the gloss transcriptions. The HSUs are constructed in a data-driven way similar to [14]. All individual frames that compose all signs are considered in a common pool of features. In that way we take into account all frames in each sign and not just the start and the end frame. We apply next on this superset of features a clustering algorithm. The first approach explored is an

Table 1. Sample indicative part of a lexicon showing the mapping of Signs to hand-shape sub-unit sequences. HSx denotes the artificial occlusion sub-unit. Each sub-unit "HSi" consists of a sequence of frames where handshape remains fixed.

Gloss	HOSPITAL		LOOK		FEEL	
Pronunciation	P1	P2	P1	P2	P1	P2
SU-Seq	$HSx\ HS4$	$HSx\ HS4\ HS5$	$HSx\ HS4$	$HS4$	$HS7$	$HSx\ HS7$

unsupervised one. We start with a random initialization of the K centers, and get a partitioning of the handshape feature space on K clusters. K-means provides actually in this way a vector quantization of the handshape features space. The second approach we apply takes advantage of prior handshape information in the following sense. Once again we employ K-means to partition the hand-shape feature space. However, this time we employ the clustering algorithm *with Initialization* by specific handshape examples that are selected manually.

Herein, we illustrate indicative cases of sub-unit results as they have been constructed by the second method. Figure 7 presents five selected HSUs. For each one we visualize 1) the initial cropped handshape images for indicative instances in the pattern space that have been assigned to the specific sub-unit after clustering and 2) the reconstructed mean shape that corresponds to the centroid in the feature space of the specific cluster. It seems that the constructed HSUs in this way are quite intuitive. However there exist outliers too since the results depend on the employed model order as well as on the initialization.
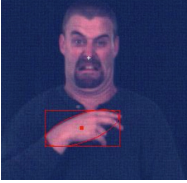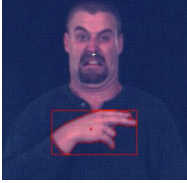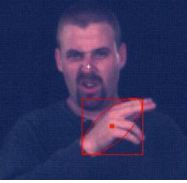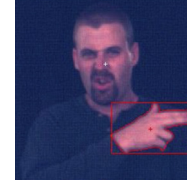
Handling Occlusions: After tracking and occlusion disambiguation there are several cases of occlusion that still remain. This is inevitable due to the nature of the data and also because of the present visual front-end that takes under consideration 2D information. During these non-resolved occluded cases we face a situation where we actually have missing features. We take advantage of the evidence that the visual front-end provides on the reliability of the features; this evidence at the present time is of binary type: occluded or non-occluded (i.e. unreliable). We explicitly distinguish our models by creating an artificial (noise-like) occlusion model. This is responsible to model all these unreliable cases. In this way we manage to keep the actual time frame synchronization information of the non-occluded cases, instead of bagging all of them in a linear pool without the actual time indices.

4.2   Handshape Sub-unit Lexicon

After the sub-unit construction via clustering we make use of the gloss labels to recompose the original signs. Next, we create a map of each sign realization to a sequence of handshape sub-units. This mapping of sub-units to signs is employed in the recognition stage for the experimental evaluation at the sign level. Each sub-unit is in this case a symbol HS that is identified by the the arbitrary index i, as HSi, assigned during clustering. The artificial sub-unit that corresponds to the occlusion cases is denoted as HSx

An example of a lexicon is shown in Table 1. This illustrates part of a sample lexicon. Each column consists of 1) a gloss string identifier e.g. LOOK, followen 2) by a pronunciation index, e.g. P1, and the corresponding sub-unit sequence.

Table 2. Two signs' realization sharing a single sub-unit sequence. Each sub-unit "HSi" consists of a sequence of frames where handshape remains fixed.

Signs:	**LOOK**			**HOSPITAL**		
SUSeq:	HSx	$+$	$HS4$	HSx	$+$	$HS4$
Frames:	[1,...,**3**]		[**4**,...,7]	[1,...,**4**]		[**5**,..,10]

The realization of signs during continuous natural signing introduces factors that increase the articulation variability. Among the reasons responsible for the multiple pronunciations as shown in the sample lexicon, is the variation by which each sign is articulated. For instance, two realizations of the sign HOSPITAL map on two different sub-unit sequences $HSx\ HS4$ and $HSx\ HS4\ HS5$ (Table 1). The extra sub-unit ($HS5$) is a result of the handshape pose variation during articulation.

Sub-Unit sequences to Multiple Glosses Map 1) Among the reasons responsible for the "single sub-unit sequence map to multiple signs" is the non-sufficient representation during modeling w.r.t. the features employed since in the presented framework we do not incorporate movement and place-of-articulation cues. 2) Another factor is the model order we employ during clustering, or in other words how loose or dense is the sub-unit construction we apply. For instance, if we make use of a small number of clusters in order to represent the space of handshapes multiple handshapes shall be assigned to the same sub-unit creating on their turn looser models. 3) Other factors involve front-end inefficiencies like the tracking errors, 3D to 2D mapping as well as the pose variation that is not explicitly treated in this approach. An example of the aforementioned mapping for signs HOSPITAL and LOOK is presented in Tables 1,2. We observe that both signs, although they consist of different hand movements map on the same sub-unit sequence $HSx\ HS4$: both consist of a segment where the right hand is occluded followed by a segment with the same HSU ($HS4$).

Sign dissimilarity: In order to take into account the mapping of sub-unit sequences to multiple signs we quantify the distance between different signs in terms of the shared sub-unit sequences. This is realized by counting for the i-th sign the number of realizations $R(i, j)$ that are represented by each sub-unit sequence j. For a set of $i = 1, 2, \cdots N_G$ signs and $j = 1, 2, \cdots N_S$ sub-unit sequences this yields $R_n(i, j) = R(i, j)/N_i$ where we also normalize with the i-th sign's number of realizations N_i. Next, we define the metric $d_s(m, n)$ between a pair of signs m, n as:

$$d_s(m, n) = 1 - \sum_{j=1}^{N_S} \min(R_n(m, j), R_n(n, j)) \tag{4}$$

When d_s between two signs equals zero, signifies that all the sub-unit sequences that map to the one sign are also shared by the second sign and with the same distribution among realizations, and vice versa. After computing d_s for all pairs
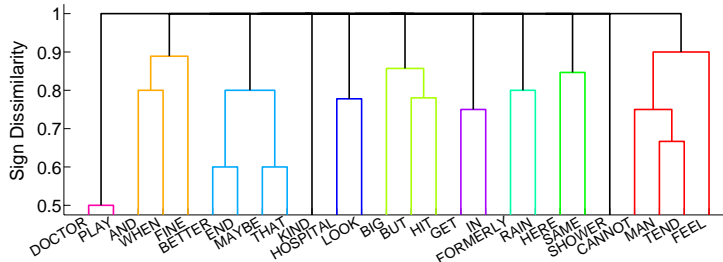
Fig. 8. Sign dissimilarity dendrogram.

of signs we hierarchically cluster the signs and construct the corresponding dendrogram. A sample dendrogram case is shown for Fig. 8. This is obtained for **26** randomly selected glosses to facilitate visualization. We observe for instance the signs "Doctor" and "Play" are quite close to each other as their distance is low. In this way we manage to find at the top level of the dendrogram the effective signs that are actually considered during recognition instead of the initial assumed greater number of sings.

4.3   Handshape Sub-units for Sign Recognition

Statistical modeling of the handshape features given the constructed HSUis implemented via GMMs by assigning one GMM to each one of the HSU. The HSU GMMs are trained on 60% of the percentage that is selected randomly as the training set. Given the unsupervised and data-driven nature of the approach there is no ground truth for the sub-unit level. In contrast for the sign level we have available the sign level transcriptions. The assignment of the sub-unit labels in the test data is accomplished by employing k-means: We compute the distance between each frame and the centroids of the sub-unit clusters that have been constructed. Eventually we assign the sub-unit label of the sub-unit whose centroid has the minimum distance error. The evaluation is realized on the rest unseen data. We apply Viterbi decoding on each test utterance, getting the most likely model fitting given the trained GMMs.

5   Sign Language Recognition Experiments

The experiments provide evaluation on the main aspects involved. These include the *Number of Subunits* and the *Vocabulary Size*. Sub-unit construction is in all cases unsupervised. However, we also evaluate the case that the clustering is initialized with manually selected handshapes.

Data and Experimental Configuration We employ *data* from the continuous American Sign Language Corpus BU400 [16]. Among the whole corpus, we select for processing 6 videos that contain stories narrated from a single signer[1].

---

[1] The original color video sequences have resolution of 648x484 pixels. Videos are identified namely as: `accident`, `biker_buddy`, `boston_la`, `football`, `lapd_story` and `siblings`. Total number of handshapes in the intra-sign segments is **4349**.

We utilize a number of *randomly selected* glosses, 26 and 40, among the most frequent ones. These are sampled from all six stories. For gloss selection we also take into account the frequency of the non-occluded right-handshape cases: we constraint gloss selection by jointly considering the most frequent ones in terms of occurrences and at the same time the ones that have the more reliable segments of non-occluded right-handshape features. We split the data at a 60-40% train and test percentages respectively. The partitioning samples data among all realizations per sign in order to equalize gloss occurrences. At the same time all experiments are conducted by employing *cross-validation*: we select three different random sets and finally show the average results. The number of realizations per sign are on average 13, with a minimum and maximum number of realizations in the range of 4 to 137. The number of non-occluded right handshapes per sign are on average 9 with a lower acceptable bound of 3. The employed features Affine Shape-Appearance Modeling are abbreviated as Aff-SAM . The results contain sub-unit level and sign-level level accuracies. At the same time we also present results on the average number of independent signs.

Number of Sub-Units and Vocabulary Size: There are two contrasting trends we take into account. On one hand, the smaller the model order, the easier the handshape measurements are classified in the correct cluster, since the models generalize successfully: this implies high recognition results. At the same time the discrimination among the different points in the handshape feature space is low. On the other hand, the greater the model order, the more the different handshapes can be discriminated. Next, we present results while varying the number of sub-units. We observe, as shown in Fig. 9, that for small number of clusters we achieve high accuracies i.e. most handshapes are recognized correctly since there is a small number of clusters. However, because of the single sub-unit sequence map to multiple signs there is no sign discrimination: as shown in Fig. 9(c), where the number of effective glosses is very low. On the other hand when we increase the number of clusters we get higher sign discrimination Fig. 9(c); at the same time our pattern space is be too fragmented, the models are overtrained and as a sequence they don't generalize well. To conclude, we trade-off between generalization and discrimination by selecting the number of clusters at the middle range of values. At the same time, although this selection is not based on explicit prior linguistic information, it refers implicitly in a quantitative way to a set of main frequent handshapes that are observed in ASL. Next, we present results of the Aff-SAM while varying the vocabulary size. We observe that for a higher number of signs Fig.9(b) the feature space is more populated and sub-unit recognition accuracy increases. At the same time although the task gets more difficult, the performance is similar in the sign recognition accuracy Fig.9(a). This is promising as sign recognition performance is not being affected from the increase of the number of signs.

Sub-unit construction with Initialization: Herein we present results when the unsupervised sub-unit construction is modified by considering prior initialization with manually selected handshapes. This initialization is conducted by selecting, after *subjective* inspection cases of handshape configurations. The handshapes are selected so that roughly 1) they span enough variance of observed handshapes and 2) they are quite frequent. This initialization is not the output of an experts' study on the more salient handshapes. We rather want to show how the employed
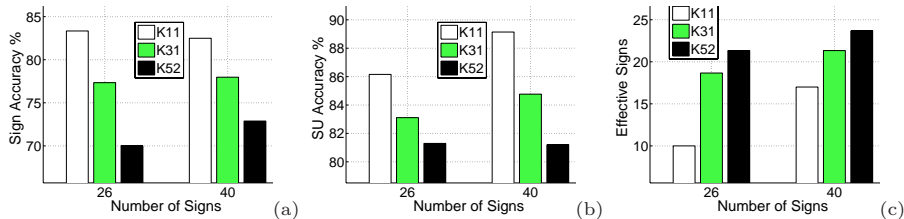
Fig. 9. ASLR Experiments on the BU400 data. Variation of the vocabulary size (26, 40 glosses) for three cases of clustering K model order parameter (11, 32 and 52). (a) Sign Accuracy, (b) Sub-unit accuracy and (c) Number of Effective Glosses.
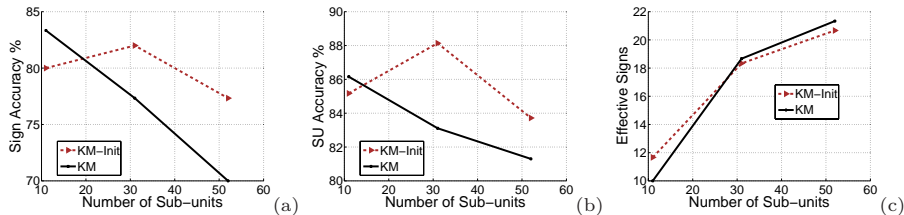
Fig. 10. ASLR Experiments on the BU400 data. Sub-unit construction with and without initialization of the clustering, while the clustering K model order parameter is increased. (a) Sign Accuracy, (b) Sub-unit accuracy and (c) Number of Effective Glosses.

framework may be employed by experts to initialize the sub-unit construction and provide more linguistically meaningful results or facilitate specific needs. We employ different cases of initialization so as to match the sub-unit number in the corresponding experiments conducted without any initialization. The larger handshape initialization sets, are constructed by adding supplementary classes. We observe in Fig. 10(a) that the handshape sub-unit construction with initialization performs on average at least 4% better. However this difference is not significant and still the accuracy of the non-initialized SU construction is for the smaller number of SUs acceptable. The average number of effective signs is similar for the two cases signify that sign discrimination is not being affected from the initialization. Concluding, via the presented framework even with the completely unsupervised data driven scheme the constructed handshape SU seem to be on average meaningful and provide promising results.

6   Conclusions

We propose an integrated framework for hand tracking and feature extraction in sign language videos and we employ it in sub-unit based ASLR. For the detection of the hands we combine a simple skin color modeling with a novel morphological filtering that results on a fast and reliable segmentation. Then, the tracking provides occlusion disambiguation so as to facilitate feature extraction. For handshape feature extraction we propose an affine modeling of hand shape-appearance images (Aff-SAM), which seems to effectively model the hand configuration and pose. The extracted features are exploited in unsupervised sub-unit construction creating in this way basic data-driven handshape phonetic units that constitute the signs. The presented framework is evaluated on a variety of recognition experiments, conducted on data from the BU400 continuous sign

language corpus, which show promising results. At the same time, we provide results on the effective number of signs among which we discriminate. To conclude with, given that handshape is among the main phonological sign language parameters, we have addressed important issues that are indispensable for automatic sign language recognition. The quantitative evaluation and the intuitive results presented show the perspective of the proposed framework for further research as well as for integration with other major sign language parameters either manual, such as movement and place-of-articulation, or facial.

References

1. Bowden, R., Windridge, D., Kadir, T., Zisserman, A., Brady, M.: A linguistic feature vector for the visual interpretation of sign language. In: ECCV. (2004)
2. Argyros, A., Lourakis, M.: Real time tracking of multiple skin-colored objects with a possibly moving camera. In: ECCV. (2004)
3. Buehler, P., Everingham, M., Zisserman, A.: Learning sign language by watching TV (using weakly aligned subtitles). In: CVPR. (2009) 2961–2968
4. Liwicki, S., Everingham, M.: Automatic recognition of fingerspelled words in British sign language. In: Proc. of CVPR4HB. (2009)
5. Hu, M.K.: Visual pattern recognition by moment invariants. Information Theory, IRE Transactions on **8** (1962) 179–187
6. Conseil, S., Bourennane, S., Martin, L.: Comparison of Fourier descriptors and Hu moments for hand posture recognition. In: EUSIPCO. (2007)
7. Birk, H., Moeslund, T., Madsen, C.: Real-time recognition of hand alphabet gestures using principal component analysis. In: Proc. SCIA. (1997)
8. Wu, Y., Huang, T.: View-independent recognition of hand postures. In: CVPR. Volume 2. (2000) 88–94
9. Huang, C.L., Jeng, S.H.: A model-based hand gesture recognition system. Machine Vision and Application **12** (2001) 243–258
10. Cootes, T., Taylor, C.: Statistical models of appearance for computer vision. Technical report, University of Manchester (2004)
11. Stenger, B., Mendonca, P., Cipolla, R.: Model-based 3D tracking of an articulated hand. In: CVPR. (2001)
12. Emmorey, K.: Language, cognition, and the brain: insights from sign language research. Erlbaum (2002)
13. Vogler, C., Metaxas, D.: Handshapes and movements: Multiple-channel american sign language recognition. In: Gesture Workshop. (2003) 247–258
14. Bauer, B., Kraiss, K.F.: Towards an automatic sign language recognition system using subunits. In: Proc. of Int'l Gesture Workshop. Volume 2298. (2001) 64–75
15. Maragos, P.: Morphological Filtering for Image Enhancement and Feature Detection. In: The Image and Video Processing Handbook. 2nd edn. Elsevier (2005)
16. Dreuw, P., Neidle, C., Athitsos, V., Sclaroff, S., Ney, H.: Benchmark databases for video-based automatic sign language recognition. In: Proc. LREC. (2008)
17. Zabulis, X., Baltzakis, H., Argyros, A.: Vision-based Hand Gesture Recognition for Human-Computer Interaction. In: The Universal Access Handbook. LEA (2009)
18. Soille, P.: Morphological Image Analysis: Principles&Applications. Springer (2004)
19. Roussos, A., Theodorakis, S., Pitsikalis, V., Maragos, P.: Affine-invariant modeling of shape-appearance images applied on sign language handshape classification. In: Proc. Int'l Conf. on Image Processing. (2010)
20. Gross, R., Matthews, I., Baker, S.: Generic vs. person specific active appearance models. Im. and Vis. Comp. **23** (2005) 1080–1093