2011-05-15

Exploiting phonological constraints for handshape inference in ASL video

Boston University, Computer Science, Technical Report BUCS-TR-2011-013

Exploiting Phonological Constraints for Handshape Inference in ASL Video

Ashwin Thangali[†], Joan P. Nash[‡], Stan Sclaroff[†], Carol Neidle[‡]
Computer Science Department[†], and Linguistics Program[‡] at Boston University, Boston, MA
tvashwin@cs.bu.edu, joanpnash@gmail.com, sclaroff@cs.bu.edu, carol@bu.edu

Abstract

Handshape is a key linguistic component of signs, and thus, handshape recognition is essential to algorithms for sign language recognition and retrieval. In this work, linguistic constraints on the relationship between start and end handshapes are leveraged to improve handshape recognition accuracy. A Bayesian network formulation is proposed for learning and exploiting these constraints, while taking into consideration inter-signer variations in the production of particular handshapes. A Variational Bayes formulation is employed for supervised learning of the model parameters. A non-rigid image alignment algorithm, which yields improved robustness to variability in handshape appearance, is proposed for computing image observation likelihoods in the model. The resulting handshape inference algorithm is evaluated using a dataset of 1500 lexical signs in American Sign Language (ASL), where each lexical sign is produced by three native ASL signers.

1. Introduction

Computer models that exploit the linguistic structure of the target language are essential for development of sign recognition algorithms that are scalable to large vocabulary sizes and have robustness to inter and intra-signer variation. Computer vision approaches [1, 9, 26] for sign language recognition, however, lag significantly behind state-of-the-art speech recognition approaches [16] in this regard. Towards bridging this gap, we propose a Bayesian network formulation for exploiting linguistic constraints to improve handshape recognition in monomorphemic lexical signs.

Signs in American Sign Language (ASL) can be categorized into several morphological classes with different principles and constraints governing the composition of signs. We limit our attention here to the most prevalent class of signs in ASL and other signed languages: the class of lexical signs, and further restrict our attention to monomorphemic signs (i.e., excluding compounds). Lexical signs are made up of discriminative components for articulation (phonemes) that consist of hand shapes, orientations, and locations within the signing space – which can change in ways that are linguistically constrained between the start and end point of a given sign – as well as movement type and, in rare instances, non-manual expressions (of the face or upper body).

This work was supported in part through US National Science Foundation grants 0705749 and 0855065.

We specifically exploit the phonological constraints [5, 22] that govern the relationships between the allowable start and end handshapes[2] for handshape recognition. The transition between the start and end handshapes generally involves either closing or opening of the hand (Fig. 1). With the exception of a small number of signs that include explicit finger movements (e.g., wiggling of fingers), the intermediate handshapes are not linguistically informative.

Furthermore, as with spoken languages, there is a certain amount of variation in the production of phonemes articulated by same or different signers [4]. Different realizations of a phoneme are called *allophones*. The occurrence of allophonic variations in handshape is general across the language (i.e., these variations are not specific to a particular sign), and, hence is amenable to a probabilistic formulation. In this paper, we focus on incorporating variations that do not involve contextual dependencies (the latter are observed at morpheme boundaries in compound signs, and, at sign boundaries in continuous signing). Examples of handshape variation are shown in Fig. 2.

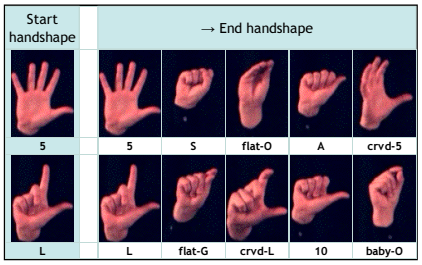

Figure 1. Example start → end handshape transitions for lexical signs in ASL. Each row shows common end handshapes for a particular start handshape ordered using probabilities for handshape transitions estimated in the proposed model.

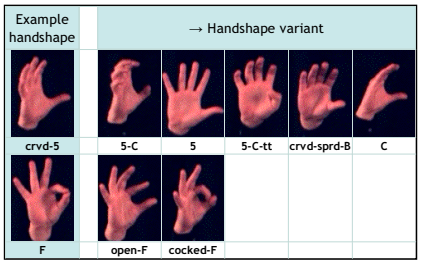

Figure 2. Common variations for two example handshapes ordered using estimated probabilities for handshape variation.

Our contributions: We propose a Bayesian network for-

[2]There is, however, no general agreement regarding the exact number of handshape phonemes in ASL [5]; for this work, we employ the ≈ 80 handshapes identified for annotations by the ASLLRP [20] project.

mulation for handshape inference in lexical signs which:

(i) exploits phonological constraints concerning {start, end} handshape co-occurrence, and,

(ii) models handshape variations using the property that a subset of similar handshapes may occur as allophonic variants of a handshape phoneme.

We also propose a non-rigid image alignment algorithm for computing image observation likelihoods in the model, which yields improved robustness to variability in handshape appearance. In experiments using a large vocabulary of ASL signs, we demonstrate that utilizing linguistic constraints improves handshape recognition accuracy.

2. Related work

Tracking hand pose in general hand gestures. Several approaches have been proposed to track finger articulations in a video sequence [13]. However, these approaches impose strong constraints on hand articulation: hands are typically assumed to be stationary (little global motion), to occupy a large portion of the video frame, and/or to be viewed from certain canonical orientations (the palm of the hand is oriented parallel or perpendicular to the camera). Approaches that use a $3D$ computer graphics hand model [8, 11] need good initialization and sufficiently well-resolved hand images in addition to the orientation constraints.

Handshape recognition in sign language. An Active Appearance Model (AAM) for sign language handshape recognition from static images is proposed in Fillbrandt et al. [15] and uses a PCA based method to capture shape and appearance variations. The learnt modes of variation, however, are tuned to the exemplars in the training set. Athitsos et al. [2] propose a fast nearest neighbor method to retrieve images from a large dictionary of ASL handshapes with similar configurations to a query hand image. The database is composed of renderings from a $3D$ graphics model for the human hand. The synthetic nature of these images does not yield a robust similarity score to real hand images.

Handshape appearance features are used along with hand location and movement descriptors in a sign spotting framework by [12, 1, 26]. Farhadi et al. [14] propose a transfer learning approach, where sign models learnt in a training domain are transferred to a test domain utilizing a subset of labelled signs in the test domain that overlap with those of the training domain (for instance, sign models learnt from one viewpoint can be transferred to a different viewpoint). These approaches do not explicitly distinguish between different handshapes and as a result do not leverage linguistic constraints on handshape transitions.

Buehler et al. [9] describe an approach to automatically extract a video template corresponding to a specified sign gloss (e.g., 'GOLF') from TV broadcast continuous signing video with weakly aligned English subtitles. A similarity score for a pair of windowed video sequences is defined based on image features for shape, orientation and location of the hands. This framework, however, treats the sign recognition problem as an instance of a general temporal sequence matching problem and does not exploit phonological constraints on signing parameters. Inter-signer variations are not addressed and the image alignment between hand image pairs is restricted to $2D$ rotations.

HMM models. Vogler and Metaxas [25] propose the 'Parallel HMM' approach assuming independent sequential processes for hand location and movement employing $3D$ tracks for arms and hands obtained using multiple cameras and physical sensors mounted on the body. A Markov model utilizing multiple articulation parameters was also proposed in [7], however only a small number of handshape classes (6) were considered. A HMM was proposed for fingerspelled word recognition in [19] using a lexicon consisting of proper nouns (names of people). Legal state transitions in the model correspond to letter sequences for words in the lexicon. In this paper, we model linguistic constraints on handshape transitions in lexical signs (handshape transitions for signs in this class follow certain general rules) and further incorporate variations across different signers.

In summary, while there has been work that has looked at handshapes, none has modelled the linguistic constraints on the start and end handshapes in lexical signs.

3. Approach

An overview of our approach is shown in Fig. 3. For a given video of a lexical sign (in this example for the gloss

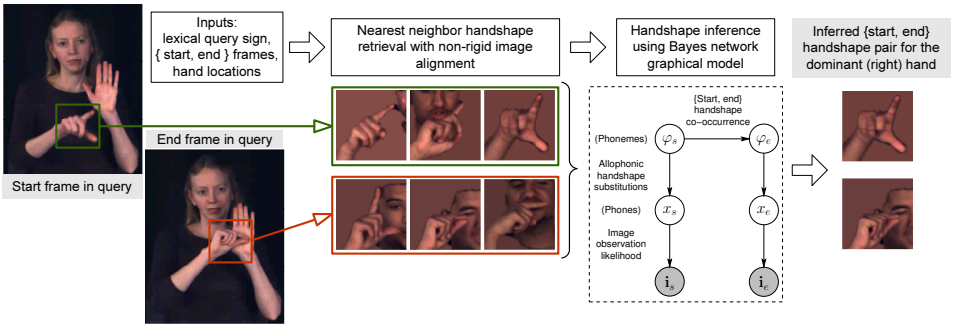

Figure 3. The proposed approach for handshape inference in lexical signs is illustrated here for handshapes on the dominant hand.

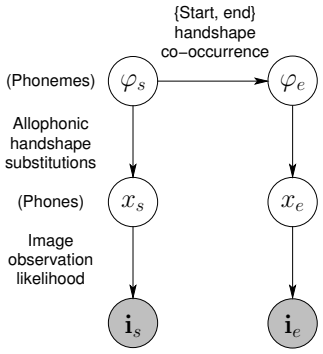

Figure 4. Graphical model to exploit {start, end} handshape co-occurrence and handshape variations in lexical signs for the dominant hand. Here, (x_s, x_e) are handshape labels we wish to infer given observed hand images $(\mathbf{i}_s, \mathbf{i}_e)$.

APPOINT), the handshape recognition algorithm takes as input a pair of images $(\mathbf{i}_s, \mathbf{i}_e)$ corresponding to the {start, end} handshapes in the video. For the purpose of illustrating our approach, we restrict our attention here to handshapes on the dominant hand. Candidate handshapes for the input {start, end} hand images are independently retrieved from a database of handshape images using a nearest neighbor method. The retrieved results (i.e., a ranked list of handshapes) are used to compute observation likelihoods in a Handshape Bayesian network (HSBN) designed to exploit {start, end} handshape co-occurrence and certain allophonic handshape variations. Computing posterior distributions in the HSBN enables inference for the {start, end} handshapes to satisfy phonological constraints.

3.1. Handshapes Bayesian Network (HSBN)

The proposed Handshapes Bayesian network (HSBN) model is shown in Fig. 4. The phoneme layer with variables (φ_s, φ_e) captures the {start, end} handshape co-occurrence probabilities. We model sets of handshapes that occur as allophonic variations of other handshapes; we introduce the phone layer with variables (x_s, x_e) to account for these variations. Determination of the appropriate linguistic analysis of the essential distinctive (phonemic) handshapes, orientations, locations, and movement trajectories, and of allowable (phonetic) variants of each of those is an active area of research in sign language linguistics. In this context, we develop here an algorithm to infer the posterior distributions and evaluate handshape recognition performance in the phone layer where it is easier to annotate the ground-truth. The HSBN in Fig. 4 yields a decomposition over the handshape labels (phones):

$$P(x_s, x_e) \;=\; \sum_{\varphi_s, \varphi_e} \boldsymbol{\pi}_{\varphi_s} \;\; \mathbf{a}_{\varphi_s, \varphi_e} \;\; \mathbf{b}^s_{\varphi_s}(x_s) \;\; \mathbf{b}^e_{\varphi_e}(x_e) \;. \quad (1)$$

The parameters $\boldsymbol{\lambda} = \{\boldsymbol{\pi}, \mathbf{a}, \mathbf{b}^s, \mathbf{b}^e\}$ above correspond to the following multinomial probability distributions:

$$\boldsymbol{\pi}_{\varphi_s} = P(\varphi_s); \qquad \mathbf{a}_{\varphi_s, \varphi_e} = P(\varphi_e | \varphi_s);$$

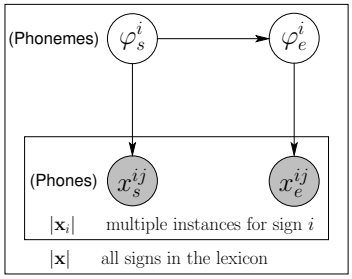

Figure 5. Plate representation of the training data used in learning the parameters for the hidden layers of the HSBN.

$$\mathbf{b}^s_{\varphi_s}(x_s) = P(x_s | \varphi_s); \qquad \mathbf{b}^e_{\varphi_e}(x_e) = P(x_e | \varphi_e). \quad (2)$$

We depart here from a conventional kernel density based observation likelihood model due to the small available dataset of handshape instances sampled from a large space of possible handposes. We use the k-nearest neighbor handshape instances retrieved from a database to postulate an expression for the posterior form of the image observation likelihood,

$$P(x_s | \mathbf{i}_s) \;\overset{\text{define}}{\propto}\; \sum_{i=1}^{k} e^{-\beta i} \;\; \delta(\; x^i_{\text{DB}},\; x_s \;) \;. \quad (3)$$

Where, k is the number of retrieved examples, δ the indicator function, and, β specifies a decaying weight. This yields the following posterior joint distribution for the {start, end} handshape labels given an input handshape image pair,

$$P(x_s, x_e | \mathbf{i}_s, \mathbf{i}_e) \propto P(x_s | \mathbf{i}_s) \;\; P(x_e | \mathbf{i}_e) \;\; \frac{P(x_s, x_e)}{P(x_s) P(x_e)} \;. \quad (4)$$

$P(x_s), P(x_e)$ can be computed as marginals of Eqn. 1.

3.2. Variational Bayes learning of HSBN

We adopt the variational Bayes (VB) [6] method to learn the parameters (Eqn. 2) for the proposed HSBN. The VB approach has been demonstrated in [6] (and references therein) to be robust to the exact choice for the parameter prior (i.e., the hyper-parameters) and also to incorporate an intrinsic penalty for model complexity. The latter property biases the VB method towards favoring sparse distributions, an essential feature for learning with small datasets.

A plate representation for learning in the proposed HSBN is shown in Fig. 5. The training set provided to the learning algorithm comprises {start, end} handshape labels annotated by linguists for monomorphemic lexical signs. Each sign in the dataset is produced by multiple signers. During learning in the HSBN, the phonemes $(\varphi^i_s, \varphi^i_e)$ constitute a hidden layer while phones (i.e., handshape labels) (x^{ij}_s, x^{ij}_e) correspond to the observed variables. We assume here that the label-set for the phonemes is a subset of the phone labels ($\approx$ 80 handshapes).

The proposed HSBN accounts for one-to-many associations between the hidden and observed random variables; whereas, in HMMs a one-to-one relationship between these

Inputs: Parameters for Dirichlet priors $\{\boldsymbol{\nu}^{\circ}, \boldsymbol{\alpha}^{\circ}, \boldsymbol{\beta}^{s\circ}, \boldsymbol{\beta}^{e\circ}\}$ and handshape label pairs $\mathbf{x}$ for signs in a training set. The latter can be decomposed as follows,

$$\mathbf{x} \;=\; \{\mathbf{x}_1, \cdots, \mathbf{x}_N\} = \big\{\{\boldsymbol{x}_{11}, \cdots, \boldsymbol{x}_{1|\mathbf{x}_1|}\}, \cdots, \{\boldsymbol{x}_{N1}, \cdots, \boldsymbol{x}_{N|\mathbf{x}_N|}\}\big\}\,; \quad \text{with, } \boldsymbol{x}_{ij} = (x_s^{ij}, x_e^{ij})\,. \tag{5}$$

Outputs: Posterior distributions for model parameters; these again belong to the Dirichlet family with parameters $\{\boldsymbol{\nu}^{\star}, \boldsymbol{\alpha}^{\star}, \boldsymbol{\beta}^{s\star}, \boldsymbol{\beta}^{e\star}\}$.

Variational Bayes lower bound: Introduce variational distributions $\{Q_{\boldsymbol{\lambda}}, Q_{\boldsymbol{\varphi}_i}\}$ to derive a lower bound $\mathcal{F}$ for the posterior distribution $P(\mathbf{x})$,

$$\begin{aligned}\ln P(\mathbf{x}) \;&=\; \ln \int d\boldsymbol{\lambda}\; P(\mathbf{x}|\boldsymbol{\lambda}) P(\boldsymbol{\lambda}) \;=\; \ln \int d\boldsymbol{\lambda}\; Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) P(\mathbf{x}|\boldsymbol{\lambda}) \frac{P(\boldsymbol{\lambda})}{Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})} \;\geq\; \int d\boldsymbol{\lambda}\; Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) \ln P(\mathbf{x}|\boldsymbol{\lambda}) \frac{P(\boldsymbol{\lambda})}{Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})} \;=\; \int d\boldsymbol{\lambda}\; Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) \left[\sum_{i=1}^{N} \ln P(\mathbf{x}_i|\boldsymbol{\lambda}) + \ln \frac{P(\boldsymbol{\lambda})}{Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})}\right] \\ &=\; \int d\boldsymbol{\lambda}\; Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) \left[\sum_{i} \ln \sum_{\boldsymbol{\varphi}_i} P(\mathbf{x}_i, \boldsymbol{\varphi}_i|\boldsymbol{\lambda}) + \ln \frac{P(\boldsymbol{\lambda})}{Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})}\right] \;\geq\; \int d\boldsymbol{\lambda}\; Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) \left[\sum_{i} \sum_{\boldsymbol{\varphi}_i} Q_{\boldsymbol{\varphi}_i}(\boldsymbol{\varphi}_i) \ln \frac{P(\mathbf{x}_i, \boldsymbol{\varphi}_i|\boldsymbol{\lambda})}{Q_{\boldsymbol{\varphi}_i}(\boldsymbol{\varphi}_i)} \;+\; \ln \frac{P(\boldsymbol{\lambda})}{Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})}\right] \;=\; \mathcal{F}(Q_{\boldsymbol{\lambda}}, Q_{\boldsymbol{\varphi}_i})\,. \end{aligned} \tag{6}$$

VB-M step: Maximize the lower bound $\mathcal{F}$ w.r.t. $Q_{\boldsymbol{\lambda}}$ to obtain an update for the latter distributions; $Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$ here approximates the desired posteriors over parameters $P(\boldsymbol{\lambda}|\mathbf{x})$,

$$\ln Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) \;=\; \ln \mathrm{Dir}\,(\,\boldsymbol{\pi}\,|\,\boldsymbol{\nu}^{\star}\,) \;+\; \sum_{\varphi_s} \ln \mathrm{Dir}\,(\,\mathbf{a}_{\varphi_s}\,|\,\boldsymbol{\alpha}^{\star}_{\varphi_s}\,) \;+\; \sum_{\varphi_s} \ln \mathrm{Dir}\,(\,\mathbf{b}^{s}_{\varphi_s}\,|\,\boldsymbol{\beta}^{s\star}_{\varphi_s}\,) \;+\; \sum_{\varphi_e} \ln \mathrm{Dir}\,(\,\mathbf{b}^{e}_{\varphi_e}\,|\,\boldsymbol{\beta}^{e\star}_{\varphi_e}\,) \quad \text{, where,} \tag{7}$$

$$\boldsymbol{\nu}^{\star}_{\varphi_s} = \boldsymbol{\nu}^{\circ}_{\varphi_s} + \sum_i Q_{\varphi_s^i}(\varphi_s)\;\;;\;\; \boldsymbol{\alpha}^{\star}_{\varphi_s,\varphi_e} = \boldsymbol{\alpha}^{\circ}_{\varphi_s,\varphi_e} + \sum_i Q_{\varphi_s^i,\varphi_e^i}(\varphi_s,\varphi_e)\;\;;\;\; \boldsymbol{\beta}^{s\star}_{\varphi_s}(x) = \boldsymbol{\beta}^{s\circ}_{\varphi_s}(x) + \sum_i \sum_{j=1}^{|\mathbf{x}_i|} \delta(x, x_s^{ij})\, Q_{\varphi_s^i}(\varphi_s)\;\;;\;\; \boldsymbol{\beta}^{e\star}_{\varphi_s}(x) = \boldsymbol{\beta}^{e\circ}_{\varphi_s}(x) + \sum_i \sum_{j=1}^{|\mathbf{x}_i|} \delta(x, x_e^{ij})\, Q_{\varphi_e^i}(\varphi_e)\,.$$

VB-E step: Maximizing $\mathcal{F}$ w.r.t. $Q_{\boldsymbol{\varphi}_i}$ yields an update for the statistics,

$$\ln Q_{\boldsymbol{\varphi}_i}(\varphi_s^i, \varphi_e^i) \;=\; -C_{Q_{\boldsymbol{\varphi}_i}} \;+\; \psi\left(\boldsymbol{\nu}^{\star}_{\varphi_s^i}\right) - \psi\left(\sum_k \boldsymbol{\nu}^{\star}_k\right) \;+\; \psi\left(\boldsymbol{\alpha}^{\star}_{\varphi_s^i,\varphi_e^i}\right) - \psi\left(\sum_k \boldsymbol{\alpha}^{\star}_{\varphi_s^i,k}\right) \;+\; \sum_{j=1}^{|\mathbf{x}_i|}\left[\psi\left(\boldsymbol{\beta}^{s\star}_{\varphi_s^i}(x_s^{ij})\right) - \psi\left(\sum_k \boldsymbol{\beta}^{s\star}_{\varphi_s^i}(k)\right) \;+\; \psi\left(\boldsymbol{\beta}^{e\star}_{\varphi_e^i}(x_e^{ij})\right) - \psi\left(\sum_k \boldsymbol{\beta}^{e\star}_{\varphi_e^i}(k)\right)\right],$$

ψ here is the *digamma* function and $C_{Q_{\boldsymbol{\varphi}_i}}$ are normalizing constants for the variational distributions $Q_{\boldsymbol{\varphi}_i}$ (sum-to-one constraints). (8)

Expansion for the lower bound $\mathcal{F}$ The expansion below is guaranteed to increase monotonically through the EM steps,

$$\mathcal{F}_{\text{current}} \;=\; \sum_i C_{Q_{\boldsymbol{\varphi}_i}} \;-\; \mathrm{KL}(\boldsymbol{\nu}^{\star} \,\|\, \boldsymbol{\nu}^{\circ}) - \sum_{\varphi_s} \mathrm{KL}(\boldsymbol{\alpha}^{\star}_{\varphi_s} \,\|\, \boldsymbol{\alpha}^{\circ}_{\varphi_s}) - \sum_{\varphi_s} \mathrm{KL}(\boldsymbol{\beta}^{s\star}_{\varphi_s} \,\|\, \boldsymbol{\beta}^{s\circ}_{\varphi_s}) - \sum_{\varphi_e} \mathrm{KL}(\boldsymbol{\beta}^{e\star}_{\varphi_e} \,\|\, \boldsymbol{\beta}^{e\circ}_{\varphi_e})\,. \tag{9}$$

$\mathrm{KL}(\boldsymbol{\nu}^{\star} \,\|\, \boldsymbol{\nu}^{\circ})$ is the divergence between Dirichlet distributions with parameter vectors $\boldsymbol{\nu}^{\star}, \boldsymbol{\nu}^{\circ}$ (expansion in appendix for [6]).

Figure 6. VB-EM algorithm to estimate posterior distributions over parameters $\boldsymbol{\lambda} = \{\boldsymbol{\pi}, \mathbf{a}, \mathbf{b}^s, \mathbf{b}^e\}$ in the proposed HSBN.

two sets of variables is typically assumed. This hence necessitates an adaptation of the VB approach for HMMs presented in [6] as described below.

VB algorithm for learning in HSBN:

The VB approach employs a lower bound to the posterior likelihood $P(\mathbf{x})$ given training data $\mathbf{x}$; this is needed since the complete data-likelihood is intractable to compute directly (the hidden parameters introduce dependencies between latent variables associated with different training samples). Through the process of maximizing this lower bound, the VB approach yields an approximation to the desired posterior distribution over model parameters $P(\boldsymbol{\lambda}|\mathbf{x})$. Choosing Dirichlet priors with parameters $\{\boldsymbol{\nu}^{\circ}, \boldsymbol{\alpha}^{\circ}, \boldsymbol{\beta}^{s\circ}, \boldsymbol{\beta}^{e\circ}\}$ for the multinomial distributions in the model (Eqn. 2) yields posterior distributions from the same family (denoted here with parameters $\{\boldsymbol{\nu}^{\star}, \boldsymbol{\alpha}^{\star}, \boldsymbol{\beta}^{s\star}, \boldsymbol{\beta}^{e\star}\}$).

The sequence of steps in the VB approach are outlined here (with details in Fig. 6):

1. Inputs: prior distributions and handshape labels for signs in the training set, Eqn. 5.
2. Introduce variational distributions $Q_{\boldsymbol{\lambda}}, Q_{\boldsymbol{\varphi}_i}$ to derive a lower bounding function $\mathcal{F}$ for the posterior likelihood $P(\mathbf{x})$, Eqn. 6.
3. Maximize $\mathcal{F}$ independently with respect to each of the two variational distributions employing Lagrange multipliers to derive updates for the respective distributions; these two updates constitute the E and M steps in the VB-EM algorithm, Eqns. 7, 8. These two key equations differ from those of the VB formulation for HMMs by including the one-to-many associations between hidden and observed variables.
4. The variational distributions $Q_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})$ obtained as a result of maximizing the lower bound in the iterative VB-EM algorithm is an approximation to the desired posterior distributions over model parameters $P(\boldsymbol{\lambda}|\mathbf{x})$.
5. The mean for the estimated posterior given by, $\widehat{\boldsymbol{\lambda}} = \mathrm{E}_{Q_{\boldsymbol{\lambda}}}[\boldsymbol{\lambda}]$ yields a point estimate for the model parameters and is commonly employed for prediction with new inputs.

During handshape inference for a query image pair $(\mathbf{i}_s, \mathbf{i}_e)$, we use the estimated model parameters $\widehat{\boldsymbol{\lambda}}$ in Eqn. 4.

3.3. Handshape observation likelihood

Given a {start, end} handshape image pair, we need to compute the handshape observation likelihoods for use in the HSBN. For this purpose, we employ a nearest neighbor (NN) method: each observed handshape image is matched to a database of labelled handshape images, and database images with the best appearance-based similarity scores are used in computing the observation likelihoods (Eqn. 3). We propose a non-rigid image alignment method for handshape image pairs to accommodate some of the variations in handshape appearance.

A sparse feature representation (e.g., edges or corners) is difficult to extract in a repeatable fashion for handshapes

due to the more gradual changes in image gradient within the hand region; we instead choose to locate feature points on a regular grid. In computing an appearance based similarity score for a hand image pair $(\mathbf{i}, \mathbf{j})$, we compute vectors $\mathbf{a}^{\mathbf{i}\to\mathbf{j}}$ that map feature locations in image $\mathbf{i}$ to pixel locations in image $\mathbf{j}$ by minimizing an alignment cost,

$$\mathbf{a}^{\mathbf{i}\to\mathbf{j}} = \min_{\mathbf{a}} \left[\mathrm{E}_{\text{data association}}(\mathbf{a}) + \mathrm{E}_{\text{spatial smoothness}}(\mathbf{a}) \right]. \quad (10)$$

For a general class of smoothness priors, the max-product LBP algorithm within a MRF representation yields state-of-the-art results, e.g., [17], and SIFTflow [18]. LBP approaches are based on message passing and typically assume a discrete label set for the alignment vectors. A quantization performed using a locally sampled grid within a window $\mathcal{W}$ for each feature yields a set of candidate alignment vectors. The message passing cost for general smoothness priors scales quadratically in the label set size, $|\mathcal{W}|$. Hence, this precludes using large densely sampled local search regions.

Choosing a smoothness prior from the Free Form Deformation (FFD) family, given by $\mathrm{E}_{\text{spatial smoothness}}(\mathbf{a}) = \mathbf{a}^T\ \mathbf{K}\ \mathbf{a}$, admits an efficient solution via gradient descent. This involves solving a sequence of sparse linear systems of equations (LSEs). Gradient descent, however, is susceptible to local minima. Motivated by the RANSAC algorithm, we include a randomization step in our LSE minimization that tends to perform well in practice. We will now describe this formulation in greater detail.

Handshape alignment algorithm:

We present the LSE formulation below which suggests an iterative approach to minimize the alignment cost.

$$-\nabla_{\mathbf{a}}\, \mathrm{E}_{\text{data assoc.}}(\mathbf{a}) \;=\; \mathbf{K}\,\mathbf{a} \quad \Big\} \text{ Local minima condition}$$

$$\text{Let, } \mathbf{f}_{\mathbf{a}}^{n} \;=\; -\nabla_{\mathbf{a}^n}\, \mathrm{E}_{\text{data assoc.}}(\mathbf{a}) \quad \Big\} \begin{array}{l}\text{Local displacements}\\ \text{to decrease } \mathrm{E}_{\text{data assoc.}}\end{array}$$

$$\mathbf{f}_{\mathbf{a}} \;=\; \mathbf{K}\,\mathbf{a} \quad \Big\} \begin{array}{l}\text{Solve LHS and RHS}\\ \text{in alternation}\end{array}$$

An outline for the proposed algorithm that adapts the above formulation to compute an alignment $\mathbf{i} \to \mathbf{j}$ for an input hand image pair is presented in Fig. 7. A global linear transformation is incorporated via an affine alignment (Eqn. 11). In each iteration of the non-rigid alignment procedure, we use local-search (employing a feature matching cost) within window $\mathcal{W}$ to predict a local alignment vector $\mathbf{a}_n^{\mathbf{u}}$ for a feature location n. To incorporate robustness to local minima, we use either the weighted average, or, a randomly chosen vector among the top-U locations in $\mathcal{W}$. The weights and ranked ordering are computed using the feature matching scores.

Because of the articulated nature of the human hand we found it beneficial to employ a non-uniform spatial smoothness prior. We propose a spring-mesh system where the

Inputs:	Image pair $\mathbf{i}, \mathbf{j}$; *Output:* Image alignment $\mathbf{a}^{\mathbf{s}:\mathbf{i}\to\mathbf{j}}$
Initialization:	Compute an affine alignment using $\mathbf{a}^{\mathbf{u}:\mathbf{i}\to\mathbf{j}}$ described below (11)
Iterations:	Update feature locations, the local search windows $\mathcal{W}$, repeat
Local alignment $\mathbf{a}^{\mathbf{u}:\mathbf{i}\to\mathbf{j}}$	In alternate iterations, choose between { random among top-U, weighted avg. of top-U } local alignments in $\mathcal{W}$ for each feature location. (12)
Stiffness matrix $\mathbf{K}$	Adapt spring stiffness κ_l using predicted local alignments, $\kappa_l = \frac{\kappa_{base}}{\text{avg}(\lvert\mathbf{a}_n^{\mathbf{u}}\rvert + \lvert\mathbf{a}_m^{\mathbf{u}}\rvert)}$. (13)
Define forces $\mathbf{f}$	Use normalized local displacements, $\mathbf{f}_n = \frac{\mathbf{a}_n^{\mathbf{u}}}{\lvert\mathbf{a}_n^{\mathbf{u}}\rvert}$. (14)
Candidate alignment $\mathbf{a}$	Solve linear system for $\mathbf{a}$, $\mathbf{f} = \mathbf{K}\,\mathbf{a}$. (15)
Smooth alignment $\mathbf{a}^{\mathbf{s}:\mathbf{i}\to\mathbf{j}}$	Line-search to determine the scaling parameter α, $\mathbf{a}^{\mathbf{s}} = \alpha^*\mathbf{a}$, $\alpha^* = \underset{\alpha\,\in\,[0,\,\alpha_{\max}]}{\text{argmin}}\ \mathrm{E}_{\text{data assoc.}}(\alpha\,\mathbf{a})$. (16)

Figure 7. Proposed algorithm for hand image alignment.

spring stiffness values are adapted to provide more flexibility in image regions with larger predicted deformation. We specify the stiffness values for each spring l using the magnitudes of predicted local alignments at the end nodes, Eqn. 13. Normalizing the local alignments yields force vectors Eqn. 14. Solving the LSE in Eqn. 15 and refinement using line search in Eqn. 16 yields one iteration of the alignment algorithm. Summing the data association costs corresponding to the independently computed alignments $\mathbf{a}^{\mathbf{s}:\mathbf{i}\to\mathbf{j}}$ and $\mathbf{a}^{\mathbf{s}:\mathbf{j}\to\mathbf{i}}$ yields a similarity score for the image pair.

We show alignment results for an example hand image pair in Fig 8. The first column visualizes the inferred spring stiffness values in the final iteration of the alignment algorithm. We observe that the ring structure with two of the fingers is essentially rigid and hence higher stiffness values (darker link colors) are inferred within it and conversely, lower stiffness values are inferred in regions surrounding the extended fingers. Results for the MRF-LBP approach minimizing the same alignment cost (but with a spatially uniform spring-mesh smoothness prior) is shown in the last column. In practice, while both approaches yield comparable alignment results, the proposed approach is an order of magnitude faster (2.4s vs. 58s) which allows a larger fraction of the database to be scanned during filter+refine NN search. We demonstrate in our experiments that the proposed stiffness adaptation with deep-NN search improves handshape retrieval accuracy over MRF-LBP.

4. Implementation details

This section gives some details about parameters for our implementation. The VB learning algorithm (Fig. 6) takes as input the training set of handshape labels. We use frequency counts computed in the training set for each of the model parameters to specify the initial posterior parameters. We also use thresholded frequency counts to specify the prior parameters (counts $<$ threshold are set to zero,

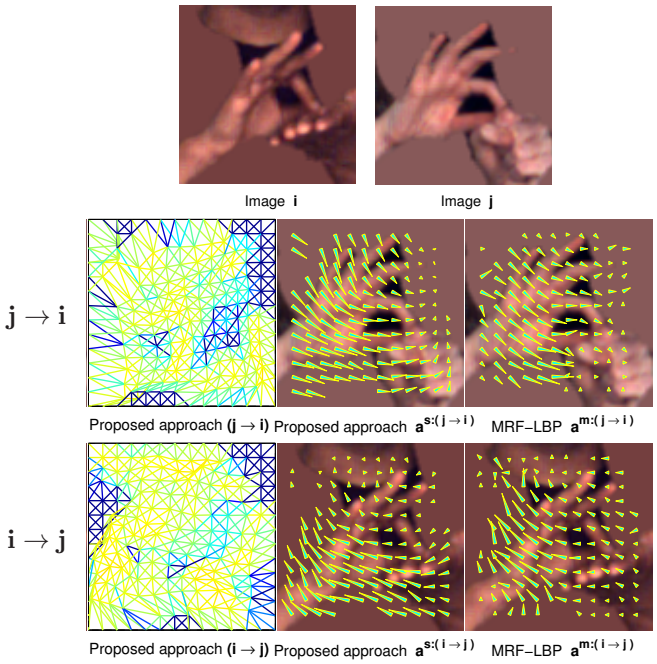

Figure 8. Bi-directional alignment. Top: Example handshape image pair $(\mathbf{i}, \mathbf{j})$. Middle: spring-mesh system for $\mathbf{j} \rightarrow \mathbf{i}$ adapts its stiffness to provide higher rigidity in areas where less deformation is expected (darker colors indicate higher stiffness); displacement field computed by the proposed approach vs. MRF-LBP. Bottom: Results for alignment $\mathbf{i} \rightarrow \mathbf{j}$.

and, a constant value otherwise). We investigated different strengths for the Dirichlet parameters; and used the same setting across all experiments.

The inference step in HSBN uses Eqn. 3 for computing the observation likelihood. The parameters here were chosen empirically as $k = 100$ and $\beta = 10^{-2}$.

In our implementation of the alignment algorithm, image descriptors are defined on a 12×12 grid. The descriptor (common to all approaches below) comprises HOG features [10] extracted for 9 local orientations of the image patch at each feature location and also at its predicted pixel location for alignment. We define the appearance matching cost in Eqn. 10 for each feature as the minimum HOG distance over these local orientations. The capture setup and image processing applied to the video sequences are described in a prior work [23].

We select a single value over the whole dataset for the base stiffness parameter κ_{base} (Eqn. 13). The other parameters specified are, local-search window size $\mathcal{W} = 17 \times 17$ grid with 2 pixels spacing, and, $\mathtt{U} = 3$ in Eqn. 12.

5. Experiments

5.1. Dataset for evaluation

We utilize the ASL Lexicon Video Dataset (ASLLVD) [3] comprising $\approx$ 1500 individual lexical signs in citation form in our experiments. Each sign here was produced by three native signers (i.e., signers raised in Deaf families who learned ASL as a first language). The signers were presented with video of the signs from the Gallaudet dictionary [24] and asked to perform each sign as they would normally produce it. Linguistic annotations, including {start, end} frames of each sign, {start, end} handshapes and gloss labels were carried out using SignStream®[21][3]. Since the focus of this work is handshape recognition which on its own is a challenging problem, we include annotations for {start, end} hand location bounding boxes in our experiments.

The dataset contains $\{1473, 1208, 1220\}$ lexical signs with handshape annotations for the three signers {M1,F1,F2} (one male and two female participants). {Start, end} hand locations were annotated for 419 signs from M1 and in a total of 1016 (start and end) frames for F1. The hand image regions are $\approx 90 \times 90$ pixels. In the experiments reported here, we use handshape images from M1 as the query set. We employ images from F1 as the database for the nearest neighbor (NN) retrieval layer in the HSBN. The different anthropometric properties of the query and database signers make handshape recognition in this dataset a challenging problem. We utilize handshape annotations from the three signers - excluding handshape labels corresponding to the query signs from M1 - to learn parameters in the HSBN (Sec. 3.2).

5.2. Experimental evaluation

Using the above dataset, we have conducted an experimental evaluation of our system. Fig. 9 shows handshape retrieval results for five query signs from the test set. The first column in the figure shows the {start,end} hand images from each query video for signer M1. The subsequent images in each row shows the top matches for the {start, end} handshapes, which were obtained via our HSBN inference method. The correct matches for the query sign are highlighted in green. Ideally, the correct match for the start and end query handshape should appear in the first position. In four of the examples shown, the correct matches appear within the top five. In the fifth example (shown at the bottom of Fig. 9) the correct match does not appear in the top five. However, close inspection of the retrieved handshape image chips shows that many of the retrieved handshapes have similar appearance.

We conducted quantitative experiments on the full test set to compare simple nearest-neighbor retrieval (NN), vs. handshape inference using the proposed HSBN. We further compared performance of our proposed alignment method vs. three other approaches for measuring appearance sim-

[3] We used a beta (pre-release) version of SignStream3, a Java re-implementation of SignStream2 (www.bu.edu/asllrp/signstream/index.html), which includes new features for annotating phonological properties of signs in ASL.

ilarity: simple HOG score (without nonrigid alignment), affine alignment based on HOG score, and MRF-LBP alignment based on HOG score. In each case, the experimental setup for computing the HOG score was the same as the one used in the implementation of our approach. In computing an affine alignment, we employ the least squares method utilizing the local displacements followed by a line-search Eqn. 11. For the proposed and MRF-LBP methods we use a spring-mesh system connecting the feature nodes (Fig. 8) as the spatial smoothness prior.

For quantitative evaluation of the recognition performance, we extract unique handshape labels from the retrieved list retaining the highest ranked match for each handshape label and removing duplicates. This yields a ranked order for the handshapes (with max-rank $= 82$ the number of handshape labels).

The table in Fig. 10 summarizes the results of our quantitative experiments. For each alignment method, results are reported for the HSBN vs. retrieval using alignment only (i.e., without HSBN). The results obtained for each alignment method without HSBN are shown in parentheses, beneath the corresponding results obtained with the HSBN. For instance, the proposed approach for non-rigid alignment with HSBN ranked the correct handshape in the first position for 32.1% of the test cases, whereas NN retrieval using alignment-only yielded the correct handshape in the first position for 26% of the test cases. A similar trend is observed as we increase the threshold on correct retrieved rank, with the proposed approach consistently giving the best results. Furthermore, HSBN inference consistently improves the retrieval accuracy vs. simple NN for all alignment approaches. We observed that the additional computation needed for HSBN inference was negligible compared to computing the alignment cost.

The graph in Fig. 10 shows a plot of the same experiments. The solid curves in the graph show the accuracy of the corresponding alignment methods with HSBN inference. These curves show performance that is consistently better than retrieval without HSBN (shown as dashed curves in the graph).

6. Conclusions and future work

We have demonstrated how the HSBN model, which models linguistic constraints on start/end handshapes in ASL, can improve the handshape recognition accuracy on a challenging dataset. Furthermore, we have proposed a handshape image alignment algorithm that yields results on-par with an MRF/LBP formulation, yet is an order of magnitude faster. However, there still remains significant room for improvement in future work.

The VB method lends itself to an approach for minimizing the state space for the hidden variables, i.e., the number of phoneme labels. This is an important aspect that we plan to investigate further in future work. There are also dialectical and ideolectical variations (i.e., phonological variations produced by groups of signers or by individuals) which are not depicted in the present model to simplify factorization of the likelihood distribution. Incorporating these properties is one further direction for future investigation.

The proposed approach can be extended to incorporate

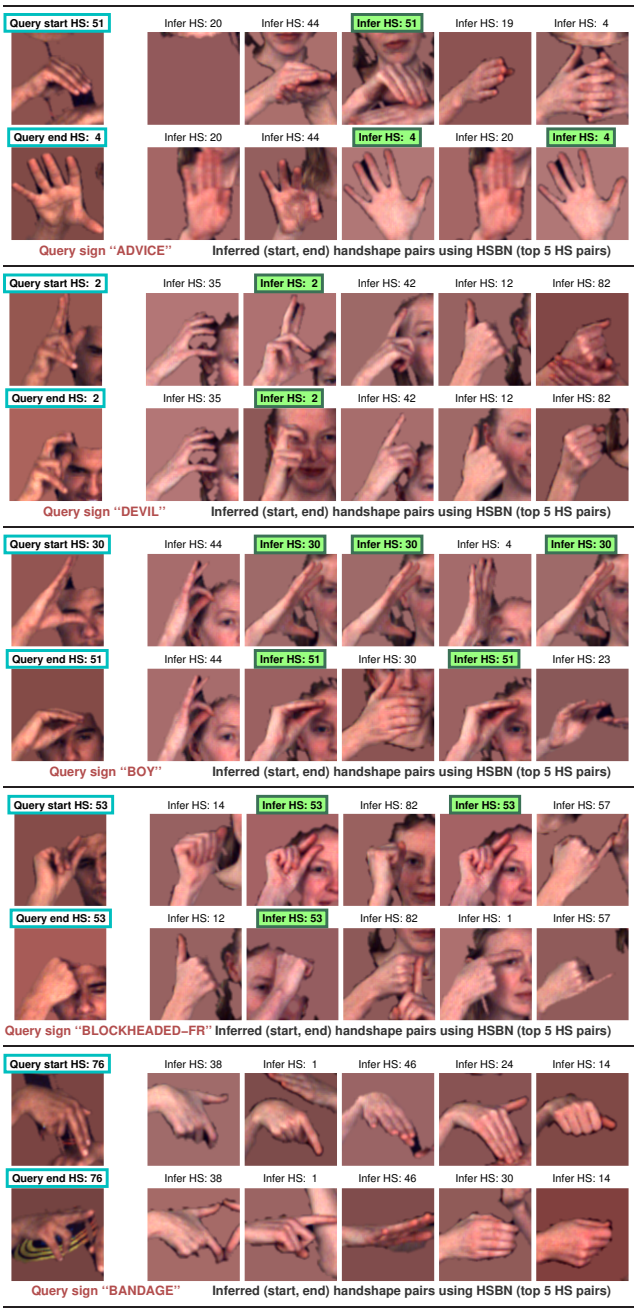

Figure 9. The first column shows query {start, end} hand images (from M1). The remaining columns show {start, end} handshape pairs inferred by HSBN (top-5 pairs) using the proposed image alignment for NN retrieval. Correct matches are marked in green.

(a). Rank of first correct retrieved handshape (max rank = #handshape labels = 82) → % of queries ↓ (419 query handshape pairs)	1	5	10	15	20	25
No spatial alignment (0.00s avg.)	25.9 (18.1)	53.3 (47.7)	66.1 (60.6)	74.8 (72.8)	81.5 (80.7)	86.4 (85.0)
Affine alignment (0.57s avg.)	27.3 (22.7)	58.7 (51.7)	71.1 (66.2)	77.8 (75.1)	83.7 (81.9)	88.4 (87.0)
Proposed approach for non-rigid (2.43s avg.)	**32.1** (26.0)	**61.3** (55.1)	**75.1** (71.4)	**81.0** (80.2)	**85.9** (84.5)	**89.6** (88.7)
MRF-LBP solver for non-rigid (58.33s avg.)	26.4 (24.5)	59.7 (52.9)	72.1 (68.3)	76.6 (76.1)	82.6 (82.1)	87.5 (86.6)
Rows (with, without) parentheses := (independent retrieval, handshape inference using the HSBN).						

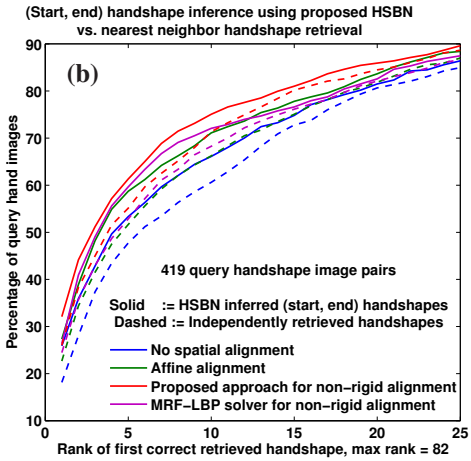

Figure 10. **(a,b).** Evaluation of handshape recognition approaches: presents nearest neighbor (NN) handshape retrieval performance (numbers in parenthesis, dashed curves in plot) for four image alignment approaches and corresponding results for handshape inference using the HSBN (no-parenthesis, solid curves). For example, (first, second) columns give % query images in which correct handshape is (at rank 1, within top-5) for NN retrieval and HSBN inference.

handshapes on the non-dominant hand. In signs where the handshapes are the same on the two hands, observations from the two handshapes can be combined to improve the accuracy of handshape recognition. When the two hands assume different handshapes, the non-dominant hand is limited to a small set of basic handshapes.

Finally, we envision handshape recognition as part of a larger system for sign recognition and retrieval. The handshape phonemes inferred using the HSBN can be used in conjunction with other articulation parameters (which include hand location, trajectory, and orientation) to facilitate progress towards person-independent large vocabulary sign recognition/sign retrieval systems.

References

[1] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *PAMI*, 31(9):1685–1699, 2009.

[2] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. BoostMap: An embedding method for efficient nearest neighbor retrieval. *PAMI*, 30(1):89–104, 2008.

[3] V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali. The American Sign Language lexicon video dataset. In *CVPR4HB*, 2008.

[4] R. Battison. *Analyzing variation in language, papers from the Colloquium on New Ways of Analzing Variation*, chapter A Good Rule of Thumb: Variable Phonology in American Sign Language, pages 291–301. Georgetown University, 1973.

[5] R. Battison. *Linguistics of American Sign Language: An introduction*, chapter Analyzing Signs, pages 193–212. Gallaudet University Press, 2000.

[6] M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[7] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A linguistic feature vector for the visual interpretation of sign language. In *ECCV*, 2004.

[8] M. Bray, E. Koller-Meier, and L. Van Gool. Smart particle filtering for high-dimensional tracking. *CVIU*, 106(1):116–129, 2007.

[9] P. Buehler, M. Everingham, and A. Zisserman. Learning sign language by watching TV (using weakly aligned subtitles). In *CVPR*, 2009.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[11] M. de La Gorce, N. Paragios, and D. J. Fleet. Model-based hand tracking with texture, shading and self-occlusions. In *CVPR*, 2008.

[12] P. Dreuw and H. Ney. Visual modeling and feature adaptation in sign language recognition. In *ITG Conference on Speech Communication*, 2008.

[13] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *CVIU*, 108:52–73, 2007.

[14] A. Farhadi, D. Forsyth, and R. White. Transfer learning in sign language. In *CVPR*, 2007.

[15] H. Fillbrandt, S. Akyol, and K. F. Kraiss. Extraction of 3D hand shape and posture from image sequences for sign language recognition. In *Face and Gesture*, 2003.

[16] F. Jelinek. *Statistical methods for speech recognition*. The MIT Press, 1997.

[17] D. Kwon, K. J. Lee, I. D. Yun, and S. U. Lee. Nonrigid image registration using dynamic higher-order MRF model. In *ECCV*, 2008.

[18] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. SIFT flow: Dense correspondence across different scenes. In *ECCV*, 2008.

[19] S. Liwicki and M. Everingham. Automatic recognition of fingerspelled words in british sign language. In *CVPR4HB*, 2009.

[20] C. Neidle. SignStream annotation: Conventions used for the American Sign Language Linguistic Research Project. Technical report, Boston University, Reports No. 11 (2002) and 13 (addendum, 2007).

[21] C. Neidle, S. Sclaroff, and V. Athitsos. SignStream: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, and Computers*, 33(3):311–320, 2001.

[22] R. Tennant and G. Brown. *The American Sign Language Handshape Dictionary*. Gallaudet University Press, 2004.

[23] A. Thangali and S. Sclaroff. An alignment based similarity measure for hand detection in cluttered sign language video. In *CVPR4HB*, 2009.

[24] C. Valli, editor. *The Gallaudet Dictionary of American Sign Language*. Gallaudet University Press, 2005.

[25] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of American Sign Language. *CVIU*, 81:358–384, 2001.

[26] R. Yang, S. Sarkar, and B. Loeding. Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. *PAMI*, 32, no.3:462–477, 2010.