Hand Gesture Recognition within a Linguistics-Based Framework

Konstantinos G. Derpanis, Richard P. Wildes, and John K. Tsotsos

York University, Department of Computer Science and
Centre for Vision Research (CVR)
Toronto Ont. M3J 1P3, Canada
{kosta,wildes,tsostos}@cs.yorku.ca
http://www.cs.yorku.ca/~{kosta,wildes,tsotsos}

Abstract. An approach to recognizing hand gestures from a monocular temporal sequence of images is presented. Of particular concern is the representation and recognition of hand movements that are used in single handed American Sign Language (ASL). The approach exploits previous linguistic analysis of manual languages that decompose dynamic gestures into their static and dynamic components. The first level of decomposition is in terms of three sets of primitives, hand shape, location and movement. Further levels of decomposition involve the lexical and sentence levels and are part of our plan for future work. We propose and demonstrate that given a monocular gesture sequence, kinematic features can be recovered from the apparent motion that provide distinctive signatures for 14 primitive movements of ASL. The approach has been implemented in software and evaluated on a database of 592 gesture sequences with an overall recognition rate of 86.00% for fully automated processing and 97.13% for manually initialized processing.

1 Introduction

Interest in automated gesture recognition has the potential to create powerful human computer interfaces. Computer vision provides methods to acquire and interpret gesture information while being minimally obtrusive to the participant. To be useful, methods must be accurate in recognition with rapid execution to support natural interaction. Further, scalability to encompass the large range of human gestures is important. The current paper presents an approach to recognizing human gestures that leverages both linguistic theory and computer vision methods. Following a path taken in the speech recognition community for the interpretation of speech [22], we appeal to linguistics to define a finite set of contrastive primitives, termed phonemes, that can be combined to represent an arbitrary number of gestures. This ensures that the developed approach is scalable. Currently, we are focused on the representation and recovery of the movement primitives derived from American Sign Language (ASL). This same linguistics analysis has also been applied to other hand gesture languages (e.g. French Sign Language). To affect the recovery of these primitives, we make use

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 282–296, 2004.

of robust, parametric motion estimation techniques to extract signatures that uniquely identify each movement from a monocular input video sequence. Here, it is interesting to note that human observers are capable of recovering the primitive movements of ASL based on motion information alone [21]. For our case, empirical evaluation suggests that algorithmic instantiation of these ideas has sufficient accuracy to distinguish the target set of ASL movement primitives, with modest processing power.

1.1 Related Research

Significant effort in computer vision has been marshalled in the investigation of human gesture recognition (see [1,20] for general reviews); some examples follow. State-space models have been used to capture the sequential nature of gestures by requiring that a series of states estimated from visual data must match in sequence, to a learned model of ordered states [7]. This general approach also has been used in conjunction with parametric curvilinear models of motion trajectories [6]. An alternative approach has used statistical factored sampling in conjunction with a model of parameterized gestures for recognition [5]; this approach can be seen as an application and extension of the CONDENSATION approach to visual tracking [14]. Further, several approaches have used Hidden Markov Models (HMMs) [17,24,26], neural networks [10] or time-delay neural networks [31] to learn from training examples (e.g., based on 2D or 3D features extracted from raw data) and subsequently recognize gestures in novel input.

A number of the cited approaches have achieved interesting recognition rates, albeit often with limited vocabularies. Interestingly, many of these approaches analyze gestures without breaking them into their constituent primitives, which could be used as in our approach, to represent a large vocabulary from a small set of generative elements. Instead, gestures are dealt with as wholes, with parameters learned from training sets. This tack may limit the ability of such approaches to generalize to large vocabularies as the training task becomes inordinately difficult. Additionally, several of these approaches make use of special purpose devices (e.g., coloured markers, data gloves) to assist in data acquisition.

In [2,28], two of the earliest efforts of using linguistic concepts for the description and recognition of both general and domain specific motion are presented. Recently, at least two lines of investigations have appealed to linguistic theory as an attack on issues in scaling gesture recognition to sizable vocabularies [18, 30]. In [18] the authors use data glove output as the input to their system. Each phoneme, from the parameters shape, location, orientation and movement, is modelled by an HMM based on features extracted from the input stream, with an 80.4% sentence accuracy rate. In [30] to affect recovery, 3D motion is extracted from the scene by fitting a 3D model of an arm with the aid of three cameras in an orthogonal configuration (or a magnetic tracking system). The motion is then fed into parallel HMMs representing the individual phonemes. The authors report that by modelling gestures by phonemes, the word recognition rate was not severely diminished, 91.19% word accuracy with phonemes

versus 91.82% word accuracy using word-level modelling. The results thus lend credence to modelling words by phonemes in vision-based gesture recognition.

1.2 Contributions

The main contributions of the present research are as follows. First, our approach models gestures in terms of their phonemic elements to yield an algorithm that recognizes gesture movement primitives given data captured with a single video camera. Second, our approach uses the apparent motion of an unmarked hand as input as opposed to fitting a model of a hand (arm) or using a mechanical device (e.g. data glove, magnetic tracker). Third, our recognition scheme is based on a nearest neighbour match to prototype signatures, where each of 14 movement primitives of ASL is found to have a distinctive prototype signature in a kinematic feature space. We have evaluated our approach empirically with 592 video sequences and find an 86.00% phoneme accuracy rate for fully automated processing and 97.13% for manually initialized processing even as other aspects of the gesture (hand shape and location) vary.

1.3 Outline of Paper

This paper is subdivided into four main sections. This first section has provided motivation for modelling gestures at the phoneme level. Section 2 describes the linguistic-basis of our representation as well as the algorithmic aspects of the approach. Section 3 documents empirical evaluation of our algorithm instantiation. Finally, Section 4 provides a summary.

2 Technical Approach

Our approach to gesture recognition centres around two main ideas. First, linguistic theory can be used to define a representational substrate that systematically decomposes complex gestures into primitive components. Second, it is desirable to recover the primitives from data that is acquired with a standard video camera and minimal constraints on the user. Currently, we are focused on the recovery of the linguistically defined rigid single handed movement primitives of American Sign Language (ASL). The input is a temporal sequence of images that depicts a single movement phoneme. The output of our system is a classification of the depicted gesture as arising from one of the primitive movements, irrespective of other considerations (e.g., irrespective of hand location and shape). The location of the hand in the initial frame is obtained through an automated localization process utilizing the conjunction of temporal change and skin colour. We assume that the hand is the dominant moving object in the imaged scene as an aid to localization. To affect the recognition, a robust, affine motion estimator is applied to regions of interest defined by skin colour and temporal change on a frame-to-frame basis. The resulting time series of affine parameters are individually accumulated across the sequence to yield a signature that is used for classification of the depicted gesture. Details of the movement gesture vocabulary and the processing stages are presented next.

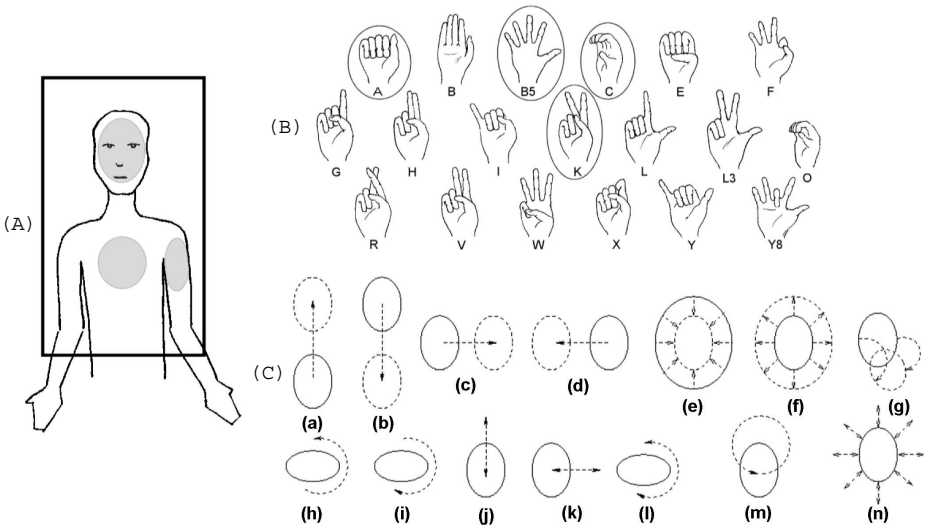


Fig. 1. Stokoe's phonemic analysis of ASL. The left panel (A) depicts the signing space where the locations reside. Shaded regions indicate locations used in our experiments. The upper right panel (B) depicts possible hand shapes. Circled shapes indicate shapes used in our experiments. The lower right panel (C) depicts possible single handed movements (a) upward (b) downward (c) rightward (d) leftward (e) toward signer (f) away signer (g) nod (h) supinate (i) pronate (j) up and down (k) side to side (l) twist wrist (m) circular (n) to and fro. The solid ellipse, dashed ellipse and dashed arrow represent the initial hand location, the final location and the path taken respectively. We investigate the recognition of movement independent of location and shape.

2.1 Linguistics Basis

Prior to William Stokoe's seminal work in ASL [27], it was assumed by linguists that the sign was the basic unit of ASL. Stokoe redefined the basic unit of a sign to units analogous to speech phonemes: minimally contrastive patterns that distinguish the symbolic vocabulary of a language. Stokoe's system consists of three parameters that are executed simultaneously to define a gesture, see Fig. 1. The three parameters capture location, hand shape and movement. There are 12 elemental locations defined by Stokoe residing in a volume in front of the signer termed the "signing space". The signing space is defined as extending from just above the head to the hip area in the vertical axis and extending close to the extents of the signer's body in the horizontal axis (see Fig. 1A). There are 19 possible hand shapes (see Fig. 1B). While Stokoe's complete vocabulary of movements consists of 24 primitives (i.e. single and two-handed movements), as a starting point, we restrict consideration to the 14 rigid *single handed* movements, shown in Fig. 1C. Current ASL theories still recognize the Stokoe system's basic parameters but differ in their definition of the constituent elements of the parameters [29]. We use Stokoe's definition of the parameters since they are gen-

erally agreed to represent an important approximation to the somewhat wider and finer grained space that might be required to capture all the subtleties of hand gesture languages.

2.2 Motion Estimation

Let $I(\mathbf{x}, t)$ represent the image brightness at position $\mathbf{x} = (x, y)^\top$ and time t. Using the brightness constancy constraint [12], we define the inter-frame motion, $\mathbf{u}(\mathbf{x}) = (u(\mathbf{x}), v(\mathbf{x}))^\top$, as,

$$I(\mathbf{x}, t+1) = I(\mathbf{x} - \mathbf{u}(\mathbf{x}), t) \tag{1}$$

We employ an affine model to describe the motion,

$$u(x, y) = a_0 + a_1 x + a_2 y, \qquad v(x, y) = a_3 + a_4 x + a_5 y \tag{2}$$

We make use of the affine model for two main reasons. First, through an analytic derivation we found that there exists a unique mapping between Stokoe's qualitative description of the movement of the hand in the world and the first-order kinematic decomposition of the corresponding visual motion fields. The first-order kinematic description includes the following measures, (differential) translation, rotation, isotropic expansion/contraction and shear: Cases (shown in Fig. 1C) a-d, j, k and m are characterized by translation, for m horizontal and vertical translation oscillate out of phase (see Fig. 2); cases h, i and l involve rotation; cases e, f and n are characterized by expansion/contraction; case g involves shear and contraction. Due to space considerations the derivation has been omitted, for details see [8]. Second, over the small angular extent that encompasses the hand at comfortable signing distances from a camera, small movements can be approximated with an affine model. To affect the recovery of the affine parameters we make use of a robust, hierarchical, gradient-based motion estimator [4] operating over a Gaussian pyramid [15]. The hierarchical nature of the estimator allows us to handle significant magnitude image displacements with computational efficiency even while avoiding local minima. This estimator is applied to skin colour defined regions of interest in a pair of images under consideration. We use skin colour to restrict consideration to image data that arises from the hand; such regions are extracted using a Bayesian maximum-likelihood classifier [32]. As a further level of robustness we restrict consideration to points that experience a significant change in intensity (i.e. dI/dt). For robustness in motion estimation, we make use of an M-estimator [13] (e.g., as opposed to a more standard least-squares approach, c.f., [3]) to allow for operation in the presence of outlying data in the form of non-hand pixels due to skin-colour oversegmentation, pixels that grossly violate the affine approximation as well as points that violate brightness constancy. The particular error norm we choose is the Geman-McClure [13].

The motion estimator is applied to adjacent frames across an image sequence. As an initial seed, the hand region in the first frame of the sequence is outlined by an automated process that consists of: utilizing the conjunction of skin colour

detection and change detection (i.e. dI/dt) to define a map of likely regions where the hand may reside, followed by a morphologically-based shape analysis [15] for the hand itself that seeks the region within the skin/change map containing the maximum circular area. No manual intervention is present. Upon recovering the motion between the first pair of frames, the analysis window is moved based on the affine parameters found (initialized identically to zero at the first frame), the affine parameters are used as the initial parameters for the motion estimation of the next pair of images and the motion estimation process is repeated. When the motion estimator reaches the end of the image sequence, six time series, each representing an affine parameter over the length of the sequence, are realized.

2.3 Kinematic Features

Owing to their descriptive power in the current context, it is advantageous to rewrite the affine parameters in terms of kinematic quantities corresponding to horizontal and vertical translation, divergence, curl and deformation (see, e.g., [16]). In particular, from the coefficients in the affine transformation (2) we calculate the following time series,

$$\begin{aligned} hor(t) &= a_0(t) \\ ver(t) &= a_3(t) \\ div(t) &= a_1(t) + a_5(t) \\ curl(t) &= -a_2(t) + a_4(t) \\ def(t) &= \sqrt{(a_1(t) - a_5(t))^2 + (a_2(t) + a_4(t))^2} \end{aligned} \tag{3}$$

Each of the kinematic time series (3) has an associated unit of measurement (e.g. horizontal/vertical motion are in pixel units) that may differ amongst each other. To facilitate comparisons across the time series for the purposes of recognition, a rescaling of responses is appropriate. We make use of min-max rescaling [11], defined as,

$$\hat{z} = \left(\frac{z - min_1}{max_1 - min_1} \right) \times (max_2 - min_2) + min_2 \tag{4}$$

with min_1 and max_1 the minimum and maximum values (resp.) in the input data z, while min_2 and max_2 specifying the range of the rescaled data taken over the entire population sample. For scaling ranges, we select $[-1, 1]$ for elements of (3) that range symmetrically about the origin and $[0, 1]$ for those with one sided responses, i.e., def.

To complete the definition of our kinematic feature set, we accumulate parameter values across each of the five rescaled kinematic time series, $\hat{hor}(t)$, $\hat{ver}(t)$, $\hat{div}(t)$, $\hat{curl}(t)$, $\hat{def}(t)$ and express each resulting value as a proportion. The accumulation procedure is motivated by the observation that there are two fundamentally different kinds of movements in the vocabulary defined in Fig. 1: those that entail constant sign movements, i.e., movements (a-i), which are unidirectional; those that entail periodic motions, i.e., movements (j-n), which move

"back and forth". To distinguish these differences, we accumulate our parameter values in two fashions.

First, to distinguish constant sign movements, we compute a *summed response*, SR_i,

$$SR_i = \sum_{t=1}^{T} p_{i,t}$$

where $i \in \{\hat{hor}, \hat{ver}, \hat{div}, \hat{curl}, \hat{def}\}$ indexes a time series, T represents the number of frames a gesture spans and $p_{i,t}$ represents the value of (rescaled) time series i at time t. Constant sign movements should yield non-zero magnitude SR_i, for some i; whereas, periodic movements will not as their changing sign responses will tend to cancel across time.

Second, to distinguish periodic movements, we compute a *summed absolute response*, SAR_i,

$$SAR_i = \sum_{t=1}^{T} |\overline{p_{i,t}}|; \text{ where } \overline{p_{i,t}} = p_{i,t} - mean_i$$

where $mean_i$ represents the mean value of (rescaled) time series i. Now, constant sign movements will have relatively small SAR_i, for all i (given removal of the mean, assuming a relatively constant velocity); whereas, periodic movements will have significantly non-zero responses as the subtracted mean should be near zero (assuming approximate symmetry in the underlying periodic pattern) and the absolute responses now sum to a positive quantity.

Due to the min-max rescaling (4), the SR_i and SAR_i calculated for any given gesture sequence are expressed in comparable ranges on an absolute scale established from consideration of all available data (i.e., min_1 and max_1 are set based on scanning across the entire sample set). For the evaluation of any given gesture sequence, we need to represent the amount of each kinematic quantity observed relative to the others in that particular sequence. For example, a (e.g., very slow) vertical motion in the absence of any other motion should be taken as significant irrespective of the speed. To capture this notion, we convert the accumulated SR_i and SAR_i values to proportions by dividing each computed value by the sum of its consort, formally,

$$SRP_i = SR_i / (\sum_k |SR_k|), \qquad SARP_i = SAR_i / (\sum_k SAR_k) \tag{5}$$

with k ranging over $\hat{hor}, \hat{ver}, \hat{div}, \hat{curl}, \hat{def}$. Here, SRP_i represents the *summed response proportion* of SR parameter i and $SARP_i$ represents the *summed absolute response proportion* of SAR parameter i. Notice that the min-max rescaling accomplished through (4) and the conversion to proportions via (5) accomplish different goals, both of which are necessary: the former brings all the kinematic variables into generally comparable units; the latter adapts the quantities to a given gesture sequence. In the end, we have a 10 component feature set SRP_i and $SARP_i$, $i \in \{\hat{hor}, \hat{ver}, \hat{div}, \hat{curl}, \hat{def}\}$ that encapsulates the kinematics of the imaged gesture.

Table 1. Gesture signatures. Each movement phoneme has a distinctive prototype signature defined in terms of our kinematic feature set. Kinematic features and movement phonemes are plotted along vertical and horizontal axes, resp. The SRP and SARP values are defined with respect to formula (5).

	SRP									**SARP**				
	upward	*downward*	*rightward*	*leftward*	*toward signer*	*away signer*	*supinate*	*pronate*	*nod*	*up and down*	*side to side*	*to and fro*	*twist wrist*	*circular*
hor	0	0	+1	-1	0	0	0	0	0	0	1	0	0	.5
ver	-1	+1	0	0	0	0	0	0	0	1	0	0	0	.5
div	0	0	0	0	+1	-1	0	0	-.5	0	0	1	0	0
curl	0	0	0	0	0	0	+1	-1	0	0	0	0	1	0
def	0	0	0	0	0	0	0	0	+.5	0	0	0	0	0

2.4 Prototype Gesture Signatures

Given our kinematic feature set, each of the primitive movements for ASL, shown in Fig. 1C has a distinctive idealized signature based on (separate) consideration of the SRP_i and $SARP_i$ values (see Table 1). Analytical relationships between the 2D kinematic signatures and the 3D hand movements are presented in [8].

Distinctive signatures for the constant sign movements (i.e., movements a-i in Fig. 1C) are defined with reference to the SRP_i values. Upward/downward movements result in responses to $ver(t)$ alone; hence, of all the SR_i, only $SR_{v\hat{e}r}$ should have a nonzero value in (5), leading to a signature of $|SRP_{v\hat{e}r}| = 1$ while $|SRP_i| = 0, i \neq v\hat{e}r$. In order to disambiguate between upward and downward movements, the sign of $SRP_{v\hat{e}r}$ is taken into account, positive sign for downward and negative for upward. Similarly, rightward/leftward movements result in significant response to $hor(t)$ alone, with the resulting signature of $|SRP_{\hat{hor}}| = 1$ while $|SRP_i| = 0, i \neq \hat{hor}$ and positive and negative signed $SRP_{\hat{hor}}$ corresponding to rightward and leftward movements, resp. The toward/away signer movements are manifest as significant responses in $div(t)$ alone. Correspondingly, $|SRP_{\hat{div}}| = 1$ while other values are zero. For this case, positive sign on $SRP_{\hat{div}}$ is indicative of toward, while negative sign indicates away. The supinate/pronate gestures map to significant responses in $curl(t)$ alone. Here, $|SRP_{\hat{curl}}| = 1$ while other values are zero with positively and negatively signed $SRP_{\hat{curl}}$ indicating supinate and pronate, resp. Unlike the other movements described so far, nod has two significant kinematic quantities which have constant signed responses throughout the gesture, namely $def(t)$ and $div(t)$. The sign of $def(t)$ should be positive, while the sign of $div(t)$ should be negative, i.e., contraction. Further, the magnitudes of these two nonzero quantities should be equal. Therefore, we have $|SRP_{\hat{div}}| = |SRP_{\hat{def}}| = 0.5$ with all other responses zero.

For periodic movements (i.e., movements j-n in Fig. 1C) distinctive signatures are defined with reference to the $SARP_i$ values. The definitions unfold

analogously to those for the constant sign movements, albeit sign now plays no role as the $SARP_i$ are all positive by construction. An up and down movement maps directly to $ver(t)$, resulting in a value of $SARP_{v\hat{e}r}$ equal to 1 with other summed absolute response proportions zero. The side to side movement directly maps to $hor(t)$, resulting in a value of $SARP_{h\hat{o}r}$ equal to 1 while other values are zero. The to and fro movement maps directly to $div(t)$, resulting in a value of $SARP_{d\hat{i}v}$ equal to 1 with other summed absolute response proportions zero. The twist wrist movement directly maps to $curl(t)$, resulting in a value of $SARP_{c\hat{u}rl}$ equal to 1 with other values zero. The circular movement has two prominent kinematic quantities, $hor(t)$ and $ver(t)$. As the hand traces a circular trajectory, these two quantities will oscillate out of phase with each other (see Fig. 2). Across a complete gesture the two summed absolute responses are equal. The overall signature is thus $SARP_{h\hat{o}r} = SARP_{v\hat{e}r} = 0.5$, with all other values zero.

For classification, we first calculate the Euclidean distance between our input signatures (i.e. SRP_i and $SARP_i$) and their respective stored prototypical signatures. The result is a set of distances d_j (14 in total). Taking the smallest distance as the classified gestures is not sufficient, since it presupposes that we know whether the classification is to be done with respect to the SRP_i (constant sign cases) or the $SARP_i$ (periodic cases). This ambiguity can be resolved through re-weighting the distances by the reciprocal norm of their respective feature vectors, formally,

$$\tilde{d_j} = (1/|\mathbf{SR}|) \times d_j; \text{ where } j \in \{\text{constant sign distance}\}$$
$$\tilde{d_j} = (1/|\mathbf{SAR}|) \times d_j; \text{ where } j \in \{\text{periodic distances}\}$$

with

$$\mathbf{SR} = (SR_{h\hat{o}r}, SR_{v\hat{e}r}, SR_{d\hat{i}v}, SR_{c\hat{u}rl}, SR_{d\hat{e}f})$$
$$\mathbf{SAR} = (SAR_{h\hat{o}r}, SAR_{v\hat{e}r}, SAR_{d\hat{i}v}, SAR_{c\hat{u}rl}, SAR_{d\hat{e}f})$$

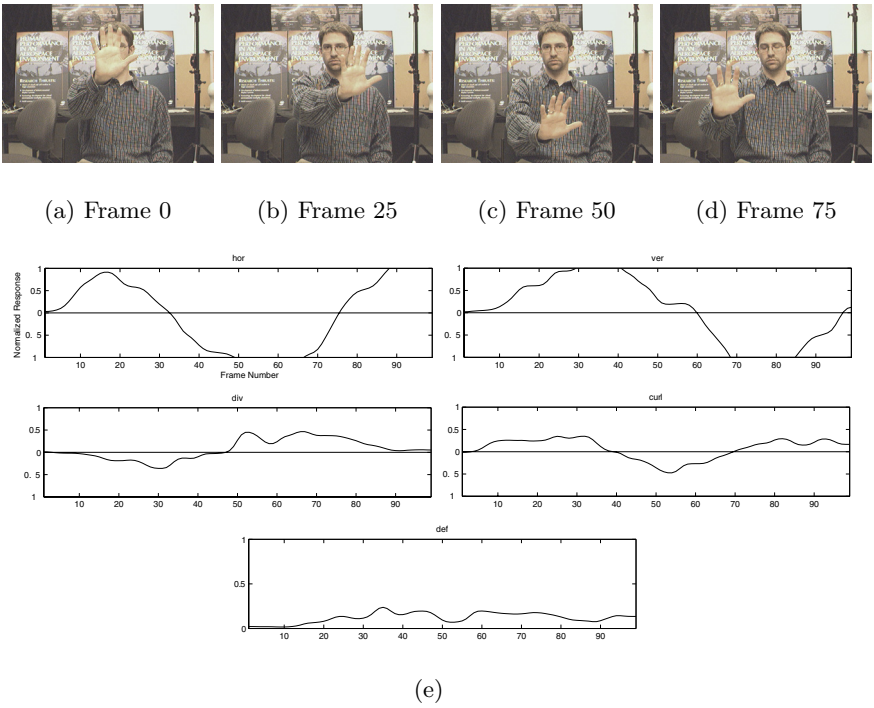
Intuitively, if the norm of **SR** is greater than that of **SAR**, then the movement is more likely to be a constant sign; if the relative magnitudes are reversed then the movement is more likely to be a periodic. Following the re-weighting, the movement with the smallest $\tilde{d_j}$ value is returned as the classification. Finally, for movements classified by distance as nod, we explicitly check to make sure $|SRP_{d\hat{i}v}| \approx |SRP_{d\hat{e}f}|$, if not we take the next closest movement. Similarly, for circular we enforce that $SARP_{h\hat{o}r} \approx SARP_{v\hat{e}r}$. These explicit checks serve to reject misclassifications when noise happens to artificially push estimated feature value patterns toward the nod and circular signatures.

3 Empirical Evaluation

To test the viability of our approach, we have tested a software realization of our algorithm on a set of video sequences each of which depicts a human volunteer executing a single movement phoneme. Here, our goal was to test the ability of our algorithm to correctly recognize movement, irrespective of the volunteer, hand location and shape of the complete gesture. Owing to the descriptive power of the phonemic decomposition of gestures into movement, location and shape

primitives, consideration of all possible combinations would lead to an experiment that is not feasible.[1] Instead, we have chosen to subsample the hand shape and location dimensions by exploiting similarities in their respective configurations. For location we have selected whole head, torso and upper arm, see Fig. 1A. These choices allow a range of locations to be considered and also introduce interesting constraints on how movements are executed. For instance, when the hand begins at the upper arm location, the natural tendency is to have the wrist rotated such that the hand is at a slight angle away from the body; as the hand moves towards the opposite side of the body, a slight rotation is introduced to bring the hand roughly parallel with the camera. For hand shape, we have selected A, B5, K and C, see Fig. 1B. The rationale for selecting hand shapes A, B5 and K is as follows: A (i.e. fist) and B5 (i.e. open flat hand) represent the two extremes of the hand shape space, whereas K (i.e. victory sign) represents an approximate midpoint of the space. Hand shape C has been included since it is a clear example of a hand shape being non-planar. This sampling leaves us with a total possible number of test cases equal to 14 (movements) $\times$ 3 (locations) $\times$ 4 (shapes) = 168. However, several of these possibilities are difficult to realize (e.g., pronating movement at the upper arm location); so, dropping these leaves us with a total of 148 cases. Three volunteers each executed all 148 movements while their actions were recorded with a video camera to yield an experimental test set of $3 \times 148 = 444$. In addition, 12 volunteers executed an approximate equal subset of the gesture space (approximately 14 gestures each). In total our experimental test set consisted of 592 gestures. It should be noted that the volunteers were fully aware of the camera and their expected position with respect to it, this allowed precise control of the experimental variables for a systematic empirical test. With an eye toward applications such control is not unrealistic: A natural signing conversation consists of directing one's signing towards the other signer (in this case a camera). During acquisition, standard indoor, overhead fluorescent lighting, was used and the normal (somewhat cluttered) background in our lab was present as volunteers signed in the foreground. Each gesture sequence was captured at a resolution of 640×480 pixels at 30 frames per second; for processing, the gesture sequences were subsampled temporally by a factor of two resulting in a frame rate of 15 frames per second. Typically, the hand region encompasses a region in a frame with dimensions approximately 100 pixels in both width and height. On average the gesture sequences spanned 40 frames for constant sign movements and 80 for periodic movements. Prior to conducting the gesture each volunteer was verbally described the gesture. This was done in order to ensure the capture of naturally occurring extraneous motions which can appear when an unbiased person performs the movements. See Fig. 2 for an example sequence.

[1] Using Stokoe's parameter definitions there would be 14 (movements) $\times$ 19 (shapes) $\times$ 12 (locations) = 3192 combinations for each volunteer.



(a) Frame 0 (b) Frame 25 (c) Frame 50 (d) Frame 75

(e)

Fig. 2. Circular movement example. A circular movement image sequence with its accompanying kinematic time series plotted. The frame numbers marked on the graphs correspond to the frame numbers of the image sequence.

3.1 Results

To assess the joint performance of the tracker and classification stages, we conducted two trials. The first trial consisted of the hand region being manually outlined in the initial frame and the second trial consisted of the automated initial localization scheme outlined in this paper. In the manually segmented trials 97.13% of the 592 test cases were correctly identified, when considering the top two candidate movements classification performance improved to 99.49%. While for the automated localization trial an accuracy rate of 86.00% was achieved and 91.00% when considering the top two candidates. Further inspection of the results found that approximately 14% of the test cases in the automated localization trial failed to isolate a sufficient region of the hand (i.e. approximately 50% of the hand). The majority of these cases consisted of the automated localization process homing in on the volunteer's head since the head was the dominant moving structure. This is contrary to our assumption that the hand is the dominant moving structure in the scene. Treating these cases as failure to acquire and omitting them from further analysis resulted in an accuracy rate of 91.55% and an accuracy of 95.09% when considering the top two candidates,

Table 2. Gesture movement recognition results. The axes of the table represent the actual input gesture (vertical) versus the classification result (horizontal). Each cell (i,j) in the table holds the percentage of test cases that were actually i but classified as j for both manually initialized localized trials (left) and automated initialized localized trials (right) (i.e. manual/automated). The diagonal (i,j) (highlighted in bold) represents the percentage of the correctly classified gestures.

	up	*down*	*up and down*	*rightward*	*leftward*	*side to side*	*toward signer*	*away signer*	*to and fro*	*supinate*	*pronate*	*twist wrist*	*nod*	*circular*
up	**100 / 92**	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 3	0 / 0	0 / 0	0 / 5	0 / 0	0 / 0
down	0 / 0	**100 / 91**	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 7	0 / 2	0 / 0	0 / 0	0 / 0	0 / 0
up and down	0 / 0	0 / 0	**100 / 95**	0 / 0	0 / 0	0 / 3	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 3	0 / 0	0 / 0
rightward	0 / 0	0 / 0	0 / 0	**100 / 92**	0 / 0	0 / 4	0 / 0	0 / 0	0 / 4	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
leftward	0 / 0	0 / 0	0 / 0	0 / 0	**97 / 85**	0 / 0	0 / 0	0 / 0	0 / 10	0 / 0	0 / 0	0 / 0	3 / 3	0 / 3
side to side	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	**100 / 86**	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 11	0 / 3	0 / 0
toward signer	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	**96 / 93**	0 / 0	0 / 0	0 / 3	0 / 0	0 / 3	4 / 0	0 / 0
away signer	0 / 0	2 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	**98 / 97**	0 / 3	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
to and fro	0 / 0	0 / 3	0 / 0	0 / 0	0 / 0	0 / 0	0 / 3	0 / 0	**92 / 84**	0 / 0	0 / 0	0 / 6	0 / 3	8 / 0
supinate	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	**97 / 95**	0 / 0	0 / 3	3 / 3	0 / 0
pronate	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	**100 / 98**	0 / 0	0 / 2	0 / 0
twist wrist	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 2	0 / 2	0 / 0	**100 / 90**	0 / 5	0 / 0
nod	0 / 0	6 / 0	0 / 0	0 / 0	0 / 0	0 / 0	3 / 0	0 / 0	6 / 3	0 / 0	0 / 0	0 / 3	**84 / 93**	0 / 0
circular	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0	0 / 2	0 / 0	0 / 0	0 / 7	**100 / 91**

see Table 2. In terms of execution speed, the tracking speed using a Pentium 4 2.1 GHz processor and unoptimized C code was 8 frames/second; the time consumed by all other components was negligible.

3.2 Discussion

A current limitation is the automated initial localization process. The majority of the failed localization cases were attributed to gross head movements, the remaining localization problems occurred with users gesturing with bare arms (although most bare arm cases were localized properly) and users wearing skin toned clothing. A review of the literature finds that most other related work has simplified the initial localization problem through manual segmentation [19,25, 30], restricting the colours in the scene [17,24,26], restricting the type of clothing worn (i.e. long sleeved shirts) [17,24,26], having users hold markers [5], using a priori knowledge of initial gesture pose [9,14], and using multiple, specially configured cameras [30] or magnetic trackers [6,10,18,30]. In our study, we make no assumptions along these lines; nevertheless, our results are competitive with those reported elsewhere. Beyond initialization, four failed tracking cases occurred related to frame-to-frame displacement beyond the capture range of our motion estimator. Drift has not been a significant factor in tracking during our experiments. This is due to the use of skin colour and change detection masks to define the region of support as well as a robust motion estimator to reject outliers. Possible solutions to tracking failure include: the use of a higher frame

rate camera to decrease interframe motion and/or the use of a motion estimator with a larger capture range (e.g., correlation-based, rather than gradient-based method).

Given acceptable tracking, problems in the classification per se arose from non-intentional but significant movements accompanying the intended movement. For instance, when conducting the "away signer" movement, some of the subjects, would rotate the palm of their hand about the camera axis as they were moving their hand forward. Systematic analysis of such cases may make it possible to improve our feature signatures to encompass such variations.

It should be noted that to realize the above results we assumed that the gestures were temporally segmented. To relax these assumptions future work may appeal to detecting discontinuities in the kinematic feature time series to temporally segment the gestures (e.g. [23]).

4 Summary

We have presented a novel approach to vision-based gesture recognition, based on two key concepts. First, we appeal to linguistic theory to represent complex gestures in terms of their primitive components. By working with a finite set of primitives, which can be combined in a wide variety of ways, our approach has the potential to deal with a large vocabulary of gestures. Second, we define distinctive signatures for the primitive components that can be recovered from monocular image sequences. By working with signatures that can be recovered without special purpose equipment, our approach has the potential for use in a wide range of human computer interfaces. Using American Sign Language (ASL) as a test bed application, we have developed an algorithm for the recognition of the primitive contrastive movements (movement phonemes) from which ASL symbols are built. The algorithm recovers kinematic features from an input video sequence, based on an affine decomposition of the apparent motion(s) across the sequence. The recovered feature values affect movement signatures that are used in a nearest neighbour recognition system. Empirical evaluation of the algorithm suggests its applicability to the analysis of complex gesture videos.

Acknowledgements. The authors thank Antonia Vezos for the illustrations in Fig. 1. Research was funded by the Institute of Robotics and Intelligent Systems one of the government of Canada's Networks of Centres of Excellence. K. G. Derpanis holds a National Sciences and Engineering Research Council of Canada PGS B fellowship. J. K. Tsotsos holds the Canada Research Chair in Computational Vision.

References

1. J.K. Aggarwal and Q. Cai. Human motion analysis: A review. *CVIU*, 73(3):428–440, 1999.
2. N. Badler. Temporal scene analysis: Conceptual descriptions of object movements. In *Dept. of Comp. Sc., Univ. of Toronto, Rep. TR-80*, 1975.
3. J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *ECCV*, pages I:5–10, 1992.
4. M.J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *ICCV*, pages 231–236, 1993.
5. M.J. Black and A.D. Jepson. A probabilistic framework for matching temporal trajectories. In *ECCV*, pages II:909–924, 1998.
6. A.F. Bobick and A.D. Wilson. A state-based approach to the representation and recognition of gesture. *PAMI*, 19(12):1325–1337, Dec 1997.
7. T. Darrell and A. Pentland. Space-time gestures. In *CVPR*, pages 335–340, 1993.
8. K.G. Derpanis. Vision based gesture recognition within a linguistics framework. Master's thesis, York University, Toronto, Canada, 2003.
9. A. Elgammal, V. Shet, Y. Yacoob, and L.S. Davis. Learning dynamics for exemplar-based gesture recognition. In *CVPR*, pages I: 571–578, 2003.
10. S.S. Fels and G.E. Hinton. Glove-talk II. *Trans. on NN*, 9(1):205–212, 1997.
11. J. Han and M. Kamber. *Data Mining*. Morgan Kaufmann, San Francisco, CA, 2001.
12. B.K.P. Horn. *Robot Vision*. MIT Press, Cambridge, MA, 1986.
13. P.J. Huber. *Robust Statistical Procedures*. SIAM Press, Philadelphia, PA, 1977.
14. M. Isard and A. Blake. CONDENSATION - conditional density propagation for visual tracking. *IJCV*, 29(1):5–28, 1998.
15. B. Jahne. *Digital Image Processing*. Springer, Berlin, 1991.
16. J.J. Koenderink and A.J. van Doorn. Local structure of movement parallax of the plane. *JOSA-A*, 66(7):717–723, 1976.
17. H.K. Lee and J.H. Kim. An HMM-based threshold model approach for gesture recognition. *PAMI*, 21(10):961–973, Oct 1999.
18. R.H. Liang and M. Ouhyoung. A real-time continuous gesture recognition system for sign language. In *AFGR*, pages 558–567, 1998.
19. S. Lu, D. Metaxas, D. Samaras, and J. Oliensis. Using multiple cues for hand tracking and model refinement. In *CVPR*, pages II: 443–450, 2003.
20. V.I. Pavlovic, R. Sharma, and T.S. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *PAMI*, 19(7):677–695, July 1997.
21. H. Poizner, U. Bellugi, and V. Lutes-Driscoll. Perception of American Sign Language in dynamic point-light displays. *J. of Exp. Psych.*, 7(2):430–440, 1981.
22. L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.
23. Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *CVPR*, pages I: 111–118, 2000.
24. J. Schlenzig, E. Hunter, and R. Jain. Vision based gesture interpretation using recursive estimation. In *Asilomar Conf. on Signals, Systems and Computers*, 1994.
25. C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *CVPR*, pages I: 69–76, 2003.
26. T. Starner, J. Weaver, and A.P. Pentland. Real-time American Sign Language recognition using desk and wearablecomputer based video. *PAMI*, 20(12):1371–1375, December 1998.

27. W.C. Stokoe, D. Casterline, and C. Croneberg. *A Dictionary of American Sign Language*. Linstok Press, Washington, DC, 1965.
28. J.K. Tsotsos, J. Mylopoulos, H.D. Covvey, and S.W. Zucker. A framework for visual motion understanding. *PAMI*, 2(6):563–573, November 1980.
29. C. Valli and C. Lucas. *Linguistics of American Sign Language: An Introduction*. Gallaudet University Press, Washington, D.C., 2000.
30. C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of American Sign Language. *CVIU*, 81(3):358–384, 2001.
31. M.H. Yang, N. Ahuja, and M. Tabb. Extraction of 2d motion trajectories and its application to hand gesture recognition. *PAMI*, 24(8):1061–1074, August 2002.
32. B. Zarit, B.J. Super, and F. Quek. Comparison of five color models in skin pixel classification. In *RATFG*, pages 58–63, 1999.