Image and Vision Computing 26 (2008) 1650–1662

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Definition and recovery of kinematic features for recognition of American sign language movements

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

York University, Department of Computer Science and Engineering, 4700 Keele Street, Toronto, Ont., Canada M3J 1P3
York University, Centre for Vision Research (CVR), 4700 Keele Street, Toronto, Ont., Canada M3J 1P3

ARTICLE INFO

Article history:
Received 23 August 2006
Received in revised form 16 February 2008
Accepted 3 April 2008

Keywords:
Gesture recognition
Motion estimation
Human computer interaction (HCI)
Linguistics
American sign language (ASL)

ABSTRACT

An approach to recognizing human hand gestures from a monocular temporal sequence of images is presented. Of 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 beyond the scope of the present paper. We propose and subsequently 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% for fully automated processing and 97% for manually initialized processing.

1. Introduction

1.1. Motivation

Interest in automated gesture recognition stems from the potentially powerful interface that can be forged between man and his artefacts, given that those artefacts have the ability to record and interpret his gestures. In this regard, computer vision-based approaches may provide particularly attractive methods as they have the potential to acquire and interpret gesture information while being minimally obtrusive to the human participant (e.g., without requiring the user to don special devices or otherwise take special actions). In any case, for such methods to be useful they must be accurate in recognition with rapid execution to support natural interaction with a human. Furthermore, scalability to encompass a sizable vocabulary of gestures is of importance.

Currently, we are focused on the representation and recovery of the movement primitives of hand gestures, specifically single-handed rigid movements derived from American sign language (ASL). For the deaf community, interaction via sign language provides the most natural means of access and communication as it employs the person's first language, even for cases where alternative means might be available (e.g., keyboards). Moreover, success in the challenging, yet constrained domain of automated sign language interpretation may provide the groundwork for further developments in flexible, vision-based human computer interaction.

Motivated by the preceding observations, this paper presents an approach to recognizing hand gestures that leverages both linguistic theory and computer vision methods. Linguistic theory defines speech phonemes as the smallest contrastive unit in the sound system of a language. These phonemes have been successfully utilized in speech recognition [1]. Similarly, such primitives have been defined for manual languages (e.g., ASL [2]). A key benefit of the phonemic approach is that modeling, analytically as done here or by training examples [3–5], of a small number of phonemes that represent the generative building blocks of the language is a feasible task, as compared to modeling a large number of gestures as wholes. Thus, this approach ensures that the developed approach is scalable to the size of the language.

To affect the recovery of these primitives, we make use of robust, parametric motion estimation techniques from computer vision to extract signatures that uniquely identify each movement from an 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 [6]. For our case, empirical evaluation suggests that algorithmic instantiation of these ideas has sufficient accuracy to distinguish the target set of ASL movement primitives. Further, since the input to our approach is a monocular video sequence and processing demands are reasonably modest, there is potential to deploy our methods with

* Corresponding author. Address: York University, Department of Computer Science and Engineering, 4700 Keele Street, Toronto, Ont., Canada M3J 1P3. Tel.: +1 416 736 2100x33113; fax: +1 416 736 5872.
E-mail address: kosta@cs.yorku.ca (K.G. Derpanis).

0262-8856/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2008.04.007

minimal invasiveness to signers while using simply a general purpose, off the shelf, computer equipped with a single video camera.

1.2. Related research

Recently, significant effort in computer vision has been marshalled in the investigation of human gesture recognition (see, e.g., [7–10] for general reviews). Here, we highlight several representative approaches. For the specific problem of gesture recognition, the basic approach consists of a feature extractor unit feeding into a recognition unit. In terms of feature extraction, several approaches have been introduced that explicitly attempt to match a rich, stored representation of a hand, namely, a 3D model of the hand [11,12] or an appearance-based model [13], with the image (or images in the case of a multi-camera setup, e.g., [11]) for the purposes of tracking. These approaches have met with limited success due to self-occlusion of the hand, convergence difficulties due to the non-linearity of the model to image feature mappings and the high degrees of freedom of the hand (i.e., modeled as 27 total degrees of freedom, 21 for the finger joint angles plus 6 for global movement of the hand).

Rather than use rich stored models to track the hand the following approaches have used coarser but real-time extractable features, such as colour blobs and optical flow. These contributions are mainly differentiated by their approach to recognition. State-space models have been used to capture the sequential nature of gestures by requiring that a temporal series of states estimated from visual data match the order in time of a model of states [14–19]. In [20,21] the pattern recognition problem has been attacked through application of the dynamic time warping (DTW) method, which is used to temporally align (i.e., match) an input pattern (i.e., series of states) to a stored pattern. An alternative approach has used statistical factored sampling in conjunction with a model of parameterized gestures for recognition [22]; this approach can be seen as an application and extension of the CONDENSATION approach to visual tracking [23]. A main strength of the approach is that it can be adapted to recognize a sequence of gestures without the need of an explicit temporal segmentation. Additionally, it may be possible to leverage the temporal models as constraints on object tracking. A main limitation in the approach lies in the factored sampling step, which is very computationally expensive, making real-time implementation a challenge.

Rule-based approaches have also been applied to the problem of hand gesture recognition [24,25]. Rule-based approaches in general contain a set of encoded predicates that when satisfied indicate that the desired event (gesture) has occurred. As an example, in [24] real-time, view-based gesture recognition is presented for interactive environments. Optical flow is extracted using a feature-based (correlation) approach. Following a subsequent segmentation stage, gestures are recognized using a rule-based approach based on characteristics of the segmented blobs. For each gesture a unique predicate is defined. A gesture is recognized when its predicate is satisfied over a set of consecutive frames.

Neural networks and their extensions, time-delay neural networks (TDNN), have also been applied to the gesture recognition problem [26–28]. A TDNN, like a standard neural network, is a multilayer feedforward network with the addition of delay units between all layers. The addition of the delay units allow the TDNN to represent temporal relationships between events in the sequence. The input layer is a set of features extracted from the video ordered in time, where the time is a fixed length. A main limitation of the approach is that to date, only isolated gestures can be recognized (i.e., temporally segmented).

Further, numerous approaches have made use of the hidden Markov model (HMM) [3,29–41], which previously had been successfully applied to the problem of speech recognition; for a tutorial on the topic of HMMs see, e.g., [42]. A standard application of HMMs to the problem of gesture recognition is to associate an HMM with each gesture; the observation sequence (extracted features from the image) is fed into each HMM and the model returning the highest score (probability) is returned as the match. Advantages afforded by using an HMM are its non-linear time scaling invariance property resulting from recurrent states in the hidden state topology and its ability to handle a continuous input stream without it being explicitly temporally segmented. Disadvantages of HMMs include the following possibilities: the underlying Markov assumption (i.e., hidden state topology) might not hold, the required training stage might overfit the HMM to the training data and the HMM might not capture the essential aspects of the underlying process caused by insufficient training data and/or unrepresentative training data.

A number of the previous approaches (cited above) have been able to achieve potentially useful recognition rates, albeit often with limited vocabularies. Interestingly, many of these approaches analyze gestures without breaking them into their constituent primitives, and thus cannot benefit from the scaling property of a generative model. Instead, gestures tend to be dealt with as wholes, with parameters learned from training sets. This tack limits the ability of such approaches to generalize to large vocabularies as the training task becomes inordinately difficult from the perspective of model building. Also of note is the fact that several of these approaches make use of special purpose devices (e.g., coloured markers, data gloves, electromagnetic trackers) to assist in data acquisition.

In [43,44], two of the earliest efforts using linguistic concepts in the description and recognition of both general and domain specific motion are presented. More recently, in [45] it was shown for the case of automobile traffic scenes how "motion verbs" can be associated with image motion patterns. For the problem of gesture recognition, at least four previous lines of investigations have appealed to linguistic theory as an attack on issues in scaling gesture recognition to sizable vocabularies [3,5,36,46]. Based on the ASL linguistics literature, the authors proposed a phoneme-based modeling of gestures. In [36], the authors use a data glove as the input to their system. Each phoneme from the parameters, hand shape, location, orientation and movement, is modelled by an HMM based on a variety of features extracted from the input stream. The authors report an 80.4% sentence accuracy rate. In [3], 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 (used interchangeably with a electromagnetic tracker). The motion is then fed into parallel HMMs representing the individual phonemes. The authors report that by modeling gestures with phonemes, the word recognition rate was not severely diminished, 91.19% word accuracy with phonemes versus 91.82% word accuracy using word-level modeling. The results thus lend credence to modeling words by phonemes in vision-based gesture recognition. A common drawback of [3,36] is the requirement of special purpose devices in the form of data gloves, mechanical trackers or multiple calibrated camera setups that limit their general deployment.

More recently, a linguistics-based approach to British sign language recognition has appeared [46]. This approach achieves a potentially useful level of performance while working with a single video camera (although illustrated examples are restricted to capture against a uniform, contrastive background). Of particular note in comparison to the approach documented in the current paper is the fact that [46] models a smaller subset of the linguistically defined single hand motion primitives, apparently restricted to image motion that is well characterized by two-dimensional translation. Finally, an extension to our work has been proposed that allows online training [4,5]. Reported empirical results are comparable to those reported in the present paper.

1.3. 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 uncalibrated video camera. Second, we derive ideal mappings between the phonemic movements under consideration and the kinematic description of the visual motion field on the imaging plane. These ideal mappings are used to form unique signatures for each of the gestures. Third, our approach uses the apparent motion of an unmarked hand as input as opposed to fitting a model of an arm/hand or using a mechanical device (e.g., data glove, magnetic tracker). Our current work focuses on isolated occurrences of the phonemic movements irrespective of the hand shape and location of the gesture. The analysis of lexical gestures and their streams in terms of their phonemic constituent parts (i.e., building blocks) are beyond the scope of the present paper. We have evaluated our approach empirically with 592 video sequences and find an 86% phoneme accuracy rate for fully automated processing and 97% for manually initialized processing even as other aspects of the gesture (hand shape and location) vary. A preliminary version of our work has appeared previously [47].

1.4. Outline of paper

This paper is divided into four main sections. This first section has provided motivation for modeling gestures at the phoneme level. Section 2 describes the linguistic-basis of our representation, and derives analytic relationships between our linguistic-basis and the resulting motion field of the hand. Section 3 documents experimental evaluation of our algorithm instantiation. Finally, Section 4 provides a summary.

2. Technical approach

2.1. Linguistics basis

Prior to William Stokoe's seminal work in ASL [2], it was assumed by linguists that the sign was the basic unit of ASL. Stokoe decomposed the basic unit of a sign into units he termed cheremes.[1] These units are analogous to speech phonemes: meaningless (on their own) sub-word patterns that are combined together to define the vocabulary (i.e., the elemental sounds that make up spoken words). 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. It is important to point out that the two-handed movements consist of synchronous and asynchronous combinations of the single-handed movements. The generative power of this representation was displayed in the very first ASL dictionary [2], where over 2000 different signs were described in terms of combinations of Stokoe's phonemes. Current ASL theories still recognize the Stokoe system's basic parameters but differ in their definition of the constituent elements of the parameters [48]. We use Stokoe's definition of the parameters since they are generally 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. Idealized gesture executions

From a purely geometric point of view, the movement of an object from the vantage point of a camera produces a moving image on the camera's image plane. The resulting visual motion field contains valuable information about the movement of the object in the world. In this section we derive the ideal mappings between the phonemic movements as qualitatively described by Stokoe and the kinematic description of the visual motion field on the imaging plane. It is our hypothesis that an idealized definition of the signatures, if not always exact in practice, will provide a sufficiently discriminative feature set such that real movement executions can be classified.

The 3D movement of a point in space is modelled in terms of instantaneous translation, $\vec{T} = (t_x, t_y, t_z)^\top$, and instantaneous rotation, $\vec{\Omega} = (\omega_x, \omega_y, \omega_z)^\top$, about the X, Y and Z Euclidean axes, respectively, with the origin defined at the camera's centre of projection (see Fig. 2). Additionally, we define $\vec{Q} = (q_x, q_y, q_z)^\top$ as the origin in space where the rotation is conducted about (e.g., $\vec{Q} = (0, 0, 0)^\top$ for rotation about the camera coordinate system origin). Under a perspective projection imaging model with focal length equal to 1, the 2D visual motion field, $\vec{v} = (u, v)^\top$ that arises as a 3D point (X, Y, Z) undergoes motion given by $\vec{T}$, $\vec{\Omega}$ and $\vec{Q}$ can be written as

$$\begin{aligned} u &= -\omega_y + \omega_z y + \omega_x xy - \omega_y x^2 \\ &\quad + \frac{-t_x - q_y\omega_z + q_z\omega_y + (-q_y\omega_x + t_z + q_x\omega_y)x}{Z} \\ v &= \omega_x - \omega_z x - \omega_y xy + \omega_x y^2 \\ &\quad + \frac{-t_y - q_z\omega_x + q_x\omega_z + (-q_y\omega_x + t_z + q_x\omega_y)y}{Z} \end{aligned} \tag{1}$$

see [49]. Note the inclusion of $\vec{Q}$ in our formulation; whereas, standard formulations model rotation as about the camera origin (i.e., $\vec{Q} = \vec{0}$).

We model the hand as a planar surface. Given the relatively small depth deviations of the fingers as compared to the distance of the hand relative to the camera such a model is not unreasonable. Formally,

$$\alpha X + \beta Y + \gamma Z = 1 \tag{2}$$

where

$$\begin{aligned} \alpha &= \frac{n_x}{d} \\ \beta &= \frac{n_y}{d} \\ \gamma &= \frac{n_z}{d} \\ d &= n_x X_0 + n_y Y_0 + n_z Z_0 \end{aligned} \tag{3}$$

$\vec{n} = (n_x, n_y, n_z)^\top$ corresponds to the normal of the plane and $(X_0, Y_0, Z_0)^\top$ represents a point in the plane. Given the planar model of a hand, the apparent motion, $(u, v)^\top$, is modelled through first-order in image coordinates by an affine transformation, formally

$$\begin{aligned} u(x, y) &= a_0 + a_1 x + a_2 y \\ v(x, y) &= a_3 + a_4 x + a_5 y \end{aligned} \tag{4}$$

where

[1] The word chereme is derived from the Greek word "$\chi\epsilon\varrho\iota$", the hand. Most linguists today tend to use the term phoneme rather than chereme, in order to highlight the similarities between speech and signing. Another usage one sometimes encounters is *viseme* in reference to visual languages.

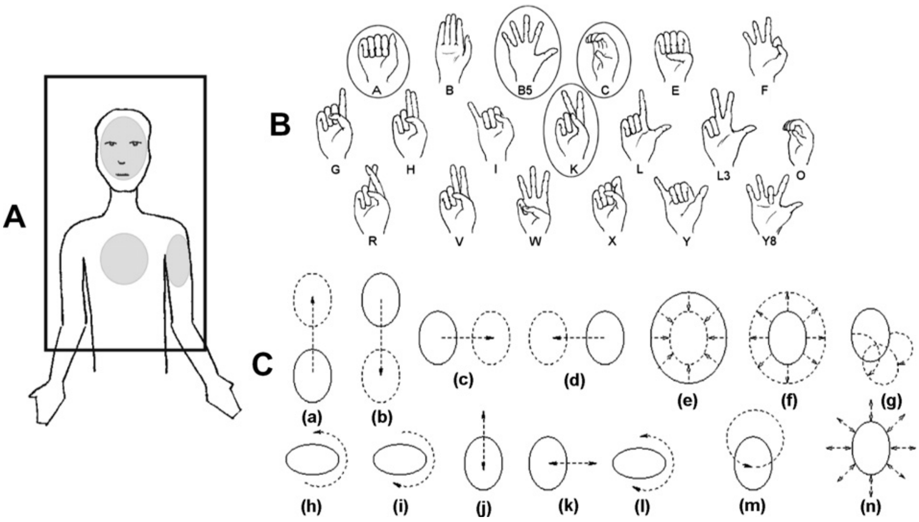

Fig. 1. Stokoe's phonemic analysis of ASL. The left panel (A) depicts the signing space where the locations reside. Shaded regions indicate locations used in our experiments. The upper right panel (B) depicts possible hand shapes. Circled shapes indicate shapes used in our experiments. The lower right 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.

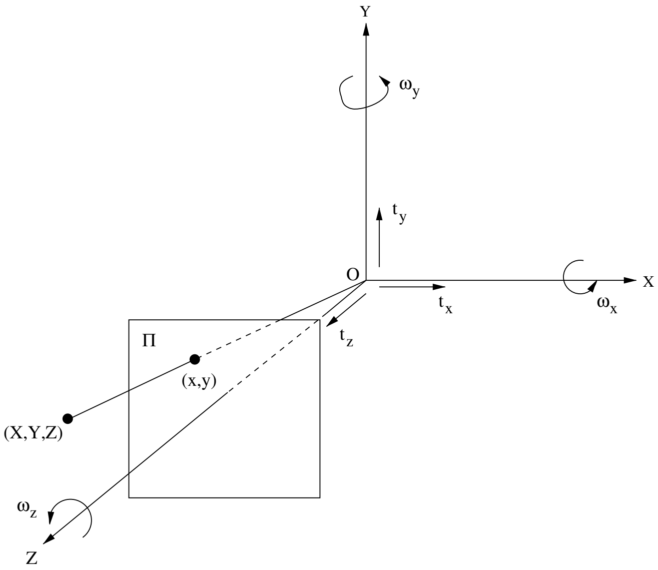

Fig. 2. Camera coordinate system. Depicted is the camera coordinate system, with an image plane Π located at $Z = 1$ and parallel to the X, Y axes. Perspective projection maps a point (X, Y, Z) to (x, y). The parameters t_x, t_y and t_z represent the translational velocities in the X, Y and Z directions respectively, ω_x, ω_y and ω_z represent the infinitesimal angle of rotation about X, Y and Z conducted about the point $\vec{Q} = (0, 0, 0)^\top$ (i.e., camera origin).

$$
\begin{aligned}
a_0 &= -\omega_y + (-\omega_z q_y - t_x + \omega_y q_z)\gamma \\
a_1 &= (-\omega_z q_y - t_x + \omega_y q_z)\alpha + (-\omega_x q_y + \omega_y q_x + t_z)\gamma \\
a_2 &= \omega_z + (-\omega_z q_y - t_x + \omega_y q_z)\beta \\
a_3 &= \omega_x + (-\omega_x q_z - t_y + \omega_z q_x)\gamma \\
a_4 &= -\omega_z + (-\omega_x q_z - t_y + \omega_z q_x)\alpha \\
a_5 &= (-\omega_x q_z - t_y + \omega_z q_x)\beta + (-\omega_x q_y + \omega_y q_x + t_z)\gamma
\end{aligned} \tag{5}
$$

with the coefficients, a_i, derived by substitution of the planar parameterization (2) into the general equations of the visual motion field (1) following by restriction to the affine model (4).

The selection of truncating the analytically correct quadratic flow arising from planar motion after the first-order terms is motivated by the fact that the second-order coefficients are highly sensitive to image noise and are difficult to estimate accurately given a small region of support [50]. Fortunately, this does not pose a problem since the contributions from the second-order terms are small when considered over small image regions [50]. Furthermore, it can be shown both analytically and through numerical simulation that the contribution of the second-order terms for the hand movements of immediate concern are negligible (see [51] for details). It will be shown in this section that the zeroth and first-order terms are sufficient to provide unique signatures for each of the movements under consideration. The affine model for apparent motion has been successfully applied to a variety of applications, examples include: general optical flow [52], 2D tracking [53,54], image registration [55], 3D structure and/or motion estimation [56–60], video partitioning [61] and hand gesture recognition [21].

Inspection of Stokoe's qualitative description of the phonemic movements reveals that the 2D image projection of each movement as captured from a frontoparallel view of a signer may map to a unique subset of the first-order kinematic description measures, (differential) translation, curl (i.e., rotation), divergence (i.e., 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. 5a); cases h, i and l involve rotation; cases e, f and n are characterized by expansion/contraction; case g involves shear and contraction. Owing to their apparent descriptive power in the current context, we rewrite the affine parameters in terms of kinematic quantities corresponding to horizontal (hor) and vertical (ver) translation, divergence (div), curl (curl) and deformation (def) (cf. [62–65]):

$$
\begin{aligned}
\text{hor} &= a_0 \\
&= -\omega_y + (-\omega_z q_y - t_x + \omega_y q_z)\gamma \\
\text{ver} &= a_3 \\
&= \omega_x + (-\omega_x q_z - t_y + \omega_z q_x)\gamma \\
\text{div} &= a_1 + a_5 \\
&= (-\omega_z q_y - t_x + \omega_y q_z)\alpha + 2(-\omega_x q_y + \omega_y q_x + t_z)\gamma \\
&\quad + (-\omega_x q_z - t_y + \omega_z q_x)\beta \\
\text{curl} &= -a_2 + a_4 \\
&= -2\omega_z - (-\omega_z q_y - t_x + \omega_y q_z)\beta + (-\omega_x q_z - t_y + \omega_z q_x)\alpha
\end{aligned} \tag{6}
$$

Table 1
Mappings of the non-periodic movements in the world space to kinematic quantities in the image space

Kinematic quantity	Non-periodic gestures								
	Rightward	Leftward	Upward	Downward	Toward signer	Away signer	Supinate	Pronate	Nod
hor(t)	$-t_x\gamma > 0$	$-t_x\gamma < 0$	0	0	0	0	0	0	0
ver(t)	0	0	$-t_y\gamma < 0$	$-t_y\gamma > 0$	0	0	0	0	0
div(t)	0	0	0	0	$2t_z\gamma > 0$	$2t_z\gamma < 0$	0	0	$-\beta q_z\omega_x > 0$
curl(t)	0	0	0	0	0	0	$-2\omega_z > 0$	$-2\omega_z < 0$	0
def(t)	0	0	0	0	0	0	0	0	$\mid \beta q_z\omega_x \mid$

$|\cdot|$ Represents the absolute value operator.

$$\begin{aligned}\text{def} &= ([a_1 - a_5]^2 + [a_2 + a_4]^2)^{1/2}\\ &= \{[(-\omega_z q_y - t_x + \omega_y q_z)\alpha - (-\omega_x q_z - t_y + \omega_z q_x)\beta]^2\\ &\quad + [(-\omega_z q_y - t_x + \omega_y q_z)\beta + (-\omega_x q_z - t_y + \omega_z q_x)\alpha]^2\}^{1/2}.\end{aligned}$$

In the remainder of this subsection we show that our intuitions regarding the ability of the kinematic quantities to distinguish prototypes for the phonemic movements depicted in Fig. 1C can be backed up analytically. We proceed by systematically instantiating the kinematic parameterizations, (6), to accommodate the prototype movements.

For the leftward, rightward, side to side, upward, downward, up and down, toward signer, away signer, to and fro and circular movements (depicted in Fig. 1) we assume that the plane (i.e., hand) is kept parallel to the image plane throughout the execution of the movement,[2] this is reflected by the surface normal $\vec{n} = (0, 0, -1)^\top$, the plane initially contains the point $(X_0, Y_0, Z_0)^\top = (0, 0, c)^\top$ where $c > 0$, the movements are executed with a constant velocity and $(q_x, q_y, q_z)^\top = (0, 0, c)^\top$. The 3D world velocities for this class of movements are as follows. Corresponding image kinematics are presented in Tables 1 and 2, as derived through substitution of the 3D velocities into the kinematic formulae (6).

> [*Leftward/rightward*] consist of a constant valued t_x throughout the gesture sequence, positive for leftward and negative for rightward, all other world velocities are zero.
> [*Side to side movement*] t_x has a constant magnitude velocity with its sign changing mid-gesture, all other world velocities are zero.
> [*Upward/downward*] consists of a constant value for t_y, positive for upward, negative for downward, all other world velocities are zero.
> [*Up and down*] t_y has a constant magnitude velocity with its sign changing mid-gesture, all other world velocities are zero.
> [*Toward/away signer*] consists of a constant value for t_z, positive for toward signer, negative for away, all other world velocities are zero.
> [*To and fro*] t_z has a constant magnitude velocity with its sign changing mid-gesture, all other world velocities are zero.
> [*Circular*] consists of the plane tracing a circular path parallel to the image plane. The path can be described by the parameterization $(\sin(\omega t), \cos(\omega t))$ where ω represents the frequency. The actual velocity of the plane is described by $(t_x(t), t_y(t)) = (\omega\cos(\omega t), -\omega\sin(\omega t))$, all other world velocities zero.

Unlike the movements described thus far, the orientation of the plane (i.e., hand) for the supinate, pronate and twist wrist movements (depicted in Fig. 1) is not assumed strictly parallel to the imaging plane throughout the execution of the gesture. The normal of the plane is assumed initially to be pointing roughly parallel with the Y-axis, towards the negative direction for supinate and positive for pronate. Strictly speaking, Stokoe's description of the supinate/pronate movements, dictates a normal exactly parallel with the Y-axis (i.e., palm facing initially down for supinate and up for pronate), but under this current analysis such configurations of the hand result in singularities in the kinematic quantities throughout the gesture execution (i.e., viewing plane on edge).

> [*Supinate/pronate*] consist of a constant value for ω_z throughout the gesture, where ω_z is negative for supinating and positive for pronating, all other world velocities are zero.
> [*Twist wrist*] ω_z is constant valued while all other world velocities are zero.

Finally, for the nod movement, initially the surface normal $\vec{n} = (0, 0, -1)^\top$, although this changes throughout the execution of the gesture as the palm rotates. The movement consists of a constant rotation $\omega_x < 0$ about the point $(q_x, q_y, q_z) = (0, 0, c)$ where $c > 0$, all other world velocities are zero.

Visual inspection of Tables 1 and 2 demonstrates that the considered phonemic movements exhibit distinctive kinematic patterns. In the following sections we present a specific approach to tracking and classifying the movement of a hand that exploits the findings of this section.

2.3. Kinematic features

In proposing a set of distinctive prototypical signatures for the phonemic movements, it is useful to exploit the fact that the movements are defined over a finite temporal interval. Correspondingly, we now consider kinematic time series:

$$\begin{aligned}&\text{hor}(t) = a_0(t)\\ &\text{ver}(t) = a_3(t)\\ &\text{div}(t) = a_1(t) + a_5(t)\\ &\text{curl}(t) = -a_2(t) + a_4(t)\\ &\text{def}(t) = \sqrt{(a_1(t) - a_5(t))^2 + (a_2(t) + a_4(t))^2}.\end{aligned} \tag{7}$$

Each of the kinematic time series (7) 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 [66], defined as

$$\hat{z} = \left(\frac{z - \min_1}{\max_1 - \min_1}\right) \times (\max_2 - \min_2) + \min_2 \tag{8}$$

with $\min_1$ and $\max_1$ the minimum and maximum values, respectively in the input data z, while $\min_2$ and $\max_2$ specify the range of the rescaled data taken over the entire population sample. For

[2] We make the planar assumptions on the hand for the sake of simplicity in modeling; nevertheless, in practice our estimation methods tolerate deviations from these idealizations, as evidenced in our experiments where we do not enforce corresponding constraints on our signers (see Section 3).

Table 2
Mappings of the periodic movements in the world space to kinematic quantities in the image space

Kinematic quantity	Periodic gestures				
	Side–side	Up and down	To and fro	Twist wrist	Circular
hor(t)	$-t_x\gamma > 0$, if $t \leqslant N/2$ $-t_x\gamma < 0$, otherwise	0	0	0	$\gamma\omega\cos(\omega t)$
ver(t)	0	$-t_y\gamma > 0$, if $t \leqslant N/2$ $-t_y\gamma < 0$, otherwise	0	0	$-\gamma\omega sin(\omega t)$
div(t)	0	0	$2t_z\gamma > 0$, if $t \leqslant N/2$ $2t_z\gamma < 0$ otherwise	0	0
(t)	0	0	0	$-2\omega_z > 0$, if $t \leqslant N/2$ $-2\omega_z < 0$, otherwise	0
def(t)	0	0	0	0	0

Note that each of the periodic movements have dual definitions realized by swapping the sign of the quantities across the intervals of execution.

scaling ranges, we select $[-1, 1]$ for elements of (7) 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, $\hat{\text{hor}}(t)$, $\hat{\text{ver}}(t)$, $\hat{\text{div}}(t)$, $\hat{\text{curl}}(t)$, $\hat{\text{def}}(t)$ 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,

$$\text{SR}_i = \sum_{t=1}^{T} p_{i,t}$$

where $i \in \{\hat{\text{hor}}, \hat{\text{ver}}, \hat{\text{div}}, \hat{\text{curl}}, \hat{\text{def}}\}$ indexes a time series, T represents the number of frames a gesture spans and $p_{i,t}$ represents the value of 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

$$\text{SAR}_i = \sum_{t=1}^{T} |\overline{p_{i,t}}|$$
$$\overline{p_{i,t}} = p_{i,t} - \text{mean}_i \tag{9}$$

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 (8), 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

$$\text{SRP}_i = \text{SR}_i \Big/ \left(\sum_k |\ \text{SR}_k\ |\right)$$
$$\text{SARP}_i = \text{SAR}_i \Big/ \left(\sum_k \text{SAR}_k\right) \tag{10}$$

with k ranging over $\hat{\text{hor}}, \hat{\text{ver}}, \hat{\text{div}}, \hat{\text{curl}}, \hat{\text{def}}$. 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 (8) and the conversion to proportions via (10) 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, $i \in \{\hat{\text{hor}}, \hat{\text{ver}}, \hat{\text{div}}, \hat{\text{curl}}, \hat{\text{def}}\}$ that encapsulates the kinematics of the imaged gesture.

2.4. Prototype gesture signatures

Given our kinematic feature set, each of the primitive movements for ASL, shown in Fig. 1 has a distinctive idealized signature based on (separate) consideration of the SRP_i and SARP_i values (see Table 3). These signatures are governed by the analytic relationships between the phonemic movements and the kinematic description of the motion field of the hand as derived in Section 2.2 and summarized in Tables 1 and 2.

Distinctive signatures for the constant sign movements (i.e., movements a–i in Fig. 1C) are defined with reference to the SRP_i values. Rightward/leftward movements result in significant response to hor(t) alone, with the resulting signature of $|\ \text{SRP}_{\hat{\text{hor}}}\ |= 1$ while $|\ \text{SRP}_i\ |= 0, i \neq \hat{\text{hor}}$. In order to disambiguate between rightward and leftward movements, the sign of $\text{SRP}_{\hat{\text{hor}}}$ is

Table 3
Gesture signatures

	SRP									SARP				
	Rightward	Leftward	Upward	Downward	Toward signer	Away signer	Supinate	Pronate	Nod	Side to side	Up and down	To and fro	Twist wrist	Circular
hor	+1	−1	0	0	0	0	0	0	0	1	0	0	0	.5
ver	0	0	−1	+1	0	0	0	0	0	0	1	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

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, respectively. The SRP and SARP values are defined with respect to formula (10).

taken into account, positive sign for rightward and negative for leftward. Similarly, upward/downward movements result in responses to $\mathrm{ver}(t)$ alone; hence, of all the SR_i, only $\mathrm{SR}_{\widehat{\mathrm{ver}}}$ should have a non-zero value in (10), leading to a signature of $|\,\mathrm{SRP}_{\widehat{\mathrm{ver}}}\,| = 1$ while $|\,\mathrm{SRP}_i\,| = 0, i \neq \widehat{\mathrm{ver}}$, positive and negative signed $\mathrm{SRP}_{\widehat{\mathrm{ver}}}$ corresponding to downward and upward movements, respectively. The toward/away signer movements are manifest as significant responses in $\mathrm{div}(t)$ alone. Correspondingly, $|\,\mathrm{SRP}_{\widehat{\mathrm{div}}}\,| = 1$ while other values are zero. For this case, positive sign on $\mathrm{SRP}_{\widehat{\mathrm{div}}}$ is indicative of toward, while negative sign indicates away. The supinate/pronate gestures map to significant responses in $\mathrm{curl}(t)$ alone. Here, $|\,\mathrm{SRP}_{\widehat{\mathrm{curl}}}\,| = 1$ while other values are zero with positively and negatively signed $\mathrm{SRP}_{\widehat{\mathrm{curl}}}$ indicating supinate and pronate, respectively. Unlike the other movements described above, nod has two significant kinematic quantities, that have constant signed responses throughout the gesture, namely $\mathrm{def}(t)$ and $\mathrm{div}(t)$. The sign of both $\mathrm{def}(t)$ and $\mathrm{div}(t)$ should be positive. Furthermore, the magnitudes of these two non-zero quantities should be equal. Therefore, we have $|\,\mathrm{SRP}_{\widehat{\mathrm{div}}}\,| = |\,\mathrm{SRP}_{\widehat{\mathrm{def}}}\,| = 0.5$ 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. The side to side movement directly maps to $\mathrm{hor}(t)$, resulting in a value of $\mathrm{SARP}_{\widehat{\mathrm{hor}}}$ equal to 1 with other summed absolute response proportions zero. The up and down movement maps directly to $\mathrm{ver}(t)$, resulting in a value of $\mathrm{SARP}_{\widehat{\mathrm{ver}}}$ equal to 1 while other values are zero. The to and fro movement maps directly to $\mathrm{div}(t)$, resulting in a value of $\mathrm{SARP}_{\widehat{\mathrm{div}}}$ equal to 1 with other summed absolute response proportions zero. The twist wrist movement directly maps to $\mathrm{curl}(t)$, resulting in a value of $\mathrm{SARP}_{\widehat{\mathrm{curl}}}$ equal to 1 with other values zero. The circular movement has two prominent kinematic quantities, $\mathrm{hor}(t)$ and $\mathrm{ver}(t)$. As the hand traces a circular trajectory, these two quantities will oscillate out of phase with each other (see Fig. 5a). Across a complete gesture the two summed absolute responses are equal. The overall signature is thus $\mathrm{SARP}_{\widehat{\mathrm{hor}}} = \mathrm{SARP}_{\widehat{\mathrm{ver}}} = 0.5$, with all other values zero.

2.5. Pattern classification

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

$$\begin{aligned}
\tilde{d}_j &= (1/\,|\,\vec{\mathrm{SR}}\,|) \times d_j; \quad \text{where } j \in \{\text{constant sign distance}\}\\
\tilde{d}_j &= (1/\,|\,\vec{\mathrm{SAR}}\,|) \times d_j; \quad \text{where } j \in \{\text{periodic distances}\}\\
&\text{with}\\
\vec{\mathrm{SR}} &= (\mathrm{SR}_{\widehat{\mathrm{hor}}}, \mathrm{SR}_{\widehat{\mathrm{ver}}}, \mathrm{SR}_{\widehat{\mathrm{div}}}, \mathrm{SR}_{\widehat{\mathrm{curl}}}, \mathrm{SR}_{\widehat{\mathrm{def}}})^{\top}\\
\vec{\mathrm{SAR}} &= (\mathrm{SAR}_{\widehat{\mathrm{hor}}}, \mathrm{SAR}_{\widehat{\mathrm{ver}}}, \mathrm{SAR}_{\widehat{\mathrm{div}}}, \mathrm{SAR}_{\widehat{\mathrm{curl}}}, \mathrm{SAR}_{\widehat{\mathrm{def}}})^{\top}.
\end{aligned}$$

Intuitively, if the norm of $\vec{\mathrm{SR}}$ is greater than that of $\vec{\mathrm{SAR}}$, 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 $|\,\mathrm{SRP}_{\widehat{\mathrm{div}}}\,| \approx |\,\mathrm{SRP}_{\widehat{\mathrm{def}}}\,|$, if not we take the next closest movement. Similarly, for circular we enforce that $\mathrm{SARP}_{\widehat{\mathrm{hor}}} \approx \mathrm{SARP}_{\widehat{\mathrm{ver}}}$. These explicit checks serve to reject misclassifications when noise happens to artificially push estimated feature value patterns toward the nod and circular signatures.

3. Experimental evaluation

3.1. Experimental design

The goal of our experiment was to test the ability of our algorithm to correctly recognize isolated phonemic movements, irrespective of the volunteer and hand location and shape parameters of the complete gesture. 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.

For completeness, we next briefly describe the hand tracker used for recovering the frame-to-frame kinematic description. However, we do not consider this as a novel contribution of our work. In brief, given the initial position of the hand in the first frame (procedure detailed below), a robust affine motion estimator operating over a Gaussian pyramid [67] is applied to regions delineated by the conjunction of a Bayesian skin colour classifier [68] and temporal change, on a frame-to-frame basis. The tracker was initialized in either of two ways: (i) manual outlining the hand region in the initial frame by a rectangular bounding box; (ii) automatic localization based on a combination of a Bayesian skin colour classifier [68] and frame-differencing between adjacent frames to define a map of likely regions where the hand may reside. In both cases, the resulting time series of affine parameters are converted to kinematic time series and then to kinematic signatures SRP_i and SARP_i. The accompanying appendix provides an additional discussion of the employed tracker. For further algorithmic and implementation details see [47,51].

For classification, we use a nearest-neighbour classifier based on a weighted Euclidean distance between our input signatures and their respective stored signatures, where the weights are the reciprocal two-norm of **SR** for the SRP signatures and **SAR** for the SARP signatures, as detailed in Section 2.5.

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.[3] 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 dominant side, 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) × 3 (locations) × 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.

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

Three volunteers each executed all 148 movements while their actions were recorded with a video camera to yield an experimental test set of 3 × 148 = 444. In addition, 12 volunteers executed an approximately equal subset of the gesture space (approximately 14 gestures each). This allowed us to test our approach's robustness to the variability of gesture execution amongst different volunteers without the associated tedium of collecting the full set of gestures from each volunteer. 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). In total, our experimental test set consisted of 592 gestures.

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 15 frames/s. Typically, the imaged hand encompassed a region of 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

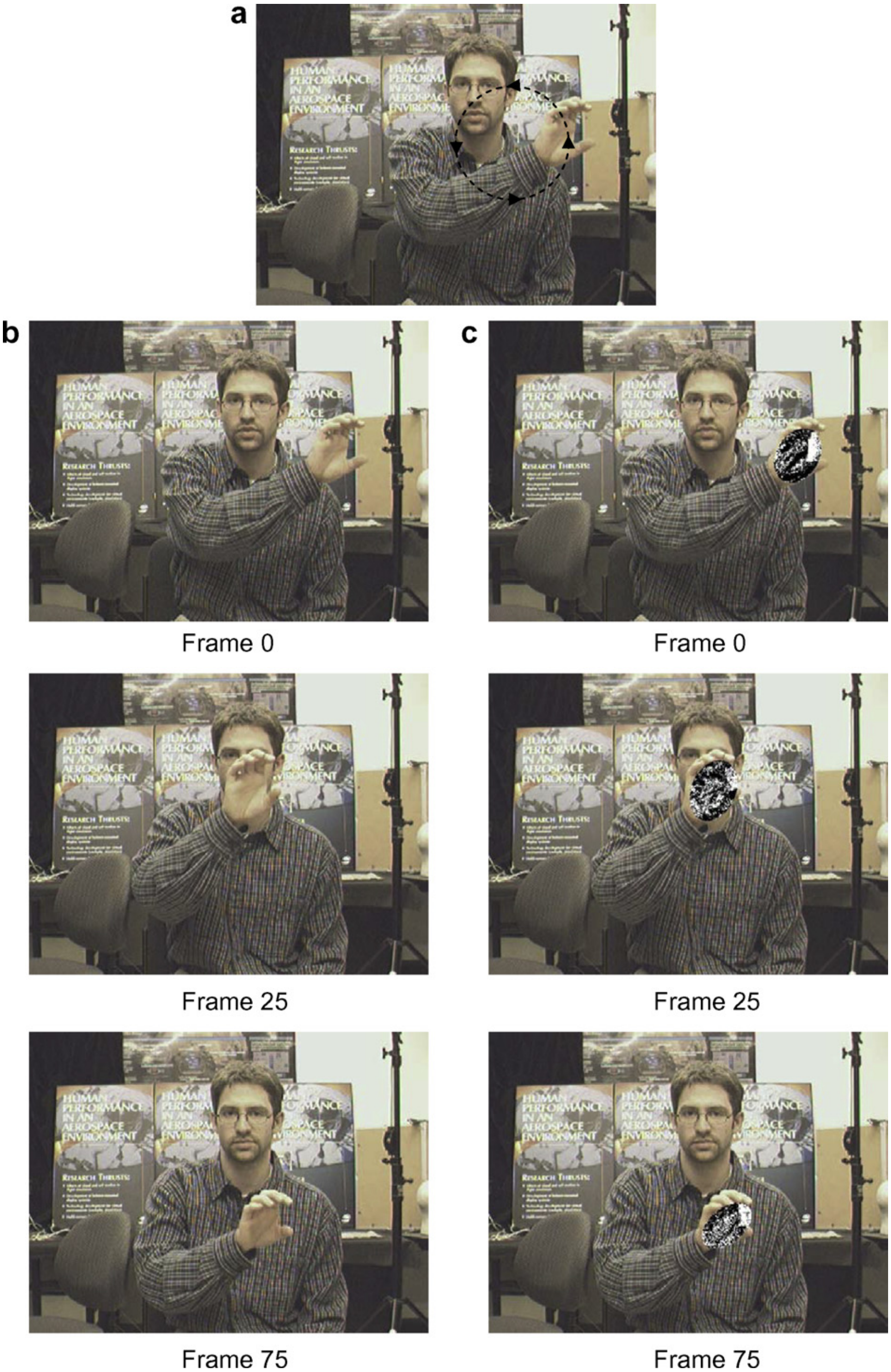

Fig. 3. Circular movement example. (a) The overlayed dotted circle denotes the path of the hand. (b) Selected input frames are shown. (c) Depicts the corresponding frames in (b) with the points on the hand overlayed in black and white denoting inliers and outliers, respectively. Inliers are identified by the skin detection/temporal change/motion estimation processes.

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. 3 and 4 for example sequences and Fig. 5 for representative kinematic time series plots of the circular and nod movements.

3.2. Results

To assess the joint performance of the tracker and classification stages, we conducted two trials. The first set derives from manually initialized tracking; the second set derives from automatic initialization.

Class-by-class results are shown as a confusion matrix in Table 4. Overall, in the manually segmented trials 97% of the 592 test cases were correctly identified. When considering the top two candidate movements, classification performance improved to 99%. While for the automated localization trial an accuracy rate of 86% was achieved and 91% 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

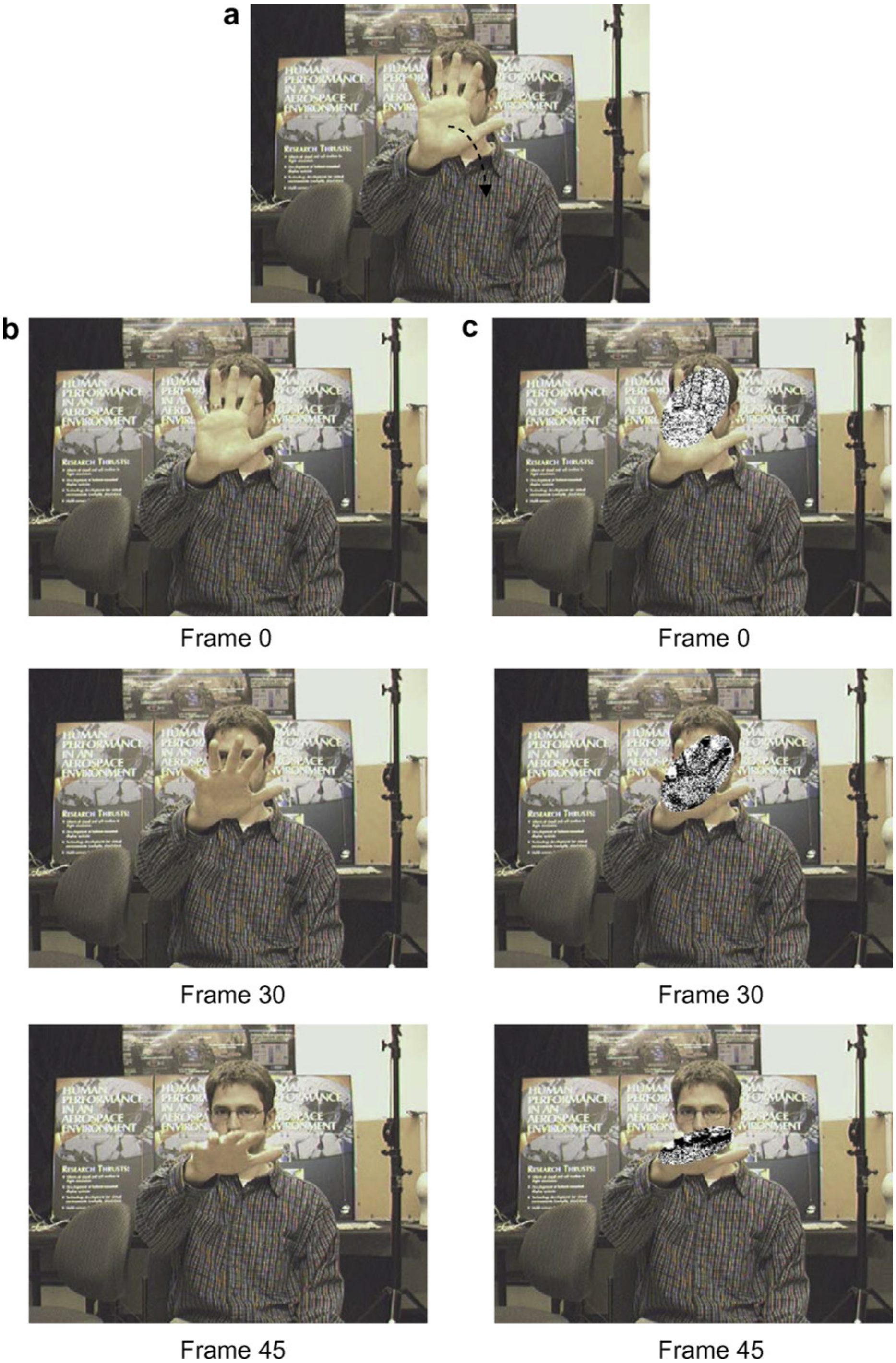

Fig. 4. Nod movement example. (a) The overlayed dotted curve denotes the path of the hand. (b) Selected input frames are shown. (c) Depicts the corresponding frames in (b) with the points on the hand overlayed in black and white denoting inliers and outliers, respectively. Inliers identified by the skin detection/temporal change/motion estimation processes.

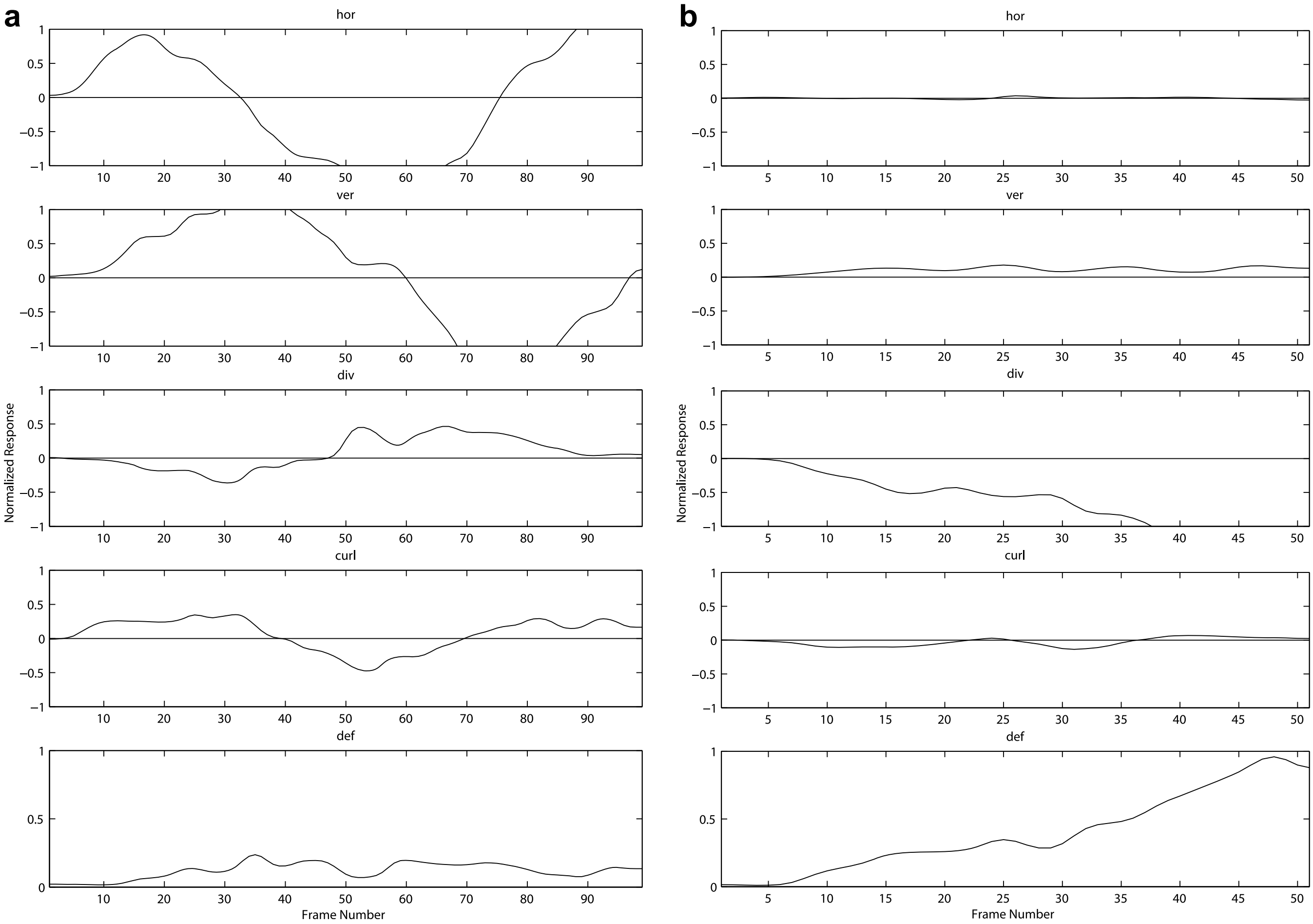

Fig. 5. Example kinematic time series. Plots of the normalized kinematic time series spanning the length of the gesture for the (a) circular and (b) nod movements.

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 skin coloured structure in the scene. Treating these cases as failure to acquire and omitting them from further analysis resulted in an accuracy rate of 92% and an accuracy of 95% when considering the top two candidates. In terms of execution speed, the tracking speed using a Pentium 4 2.1 GHz processor and unoptimized C code was 8 frames/s; the time consumed by all other components was negligible.

3.3. Discussion

Overall, the results demonstrate the ability of our algorithm to recognize correctly the 14 rigid gesture movements that comprise the single-handed movement phonemes of ASL, even while hand location and shape vary widely. This ability to decouple the primitive components of gestures is key to our overall framework, as complex gestures are analyzed in terms of their linguistically defined constituent elements. Furthermore, the results provide credence to our hypothesis that the idealized modeling of the hand movements (described in Section 2.2) if not always exact in practice, provides both a sufficiently discriminative and also extractable (from real video sequences) feature set such that real movement executions can be classified.

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 [3,69,70], restricting the colours in the scene [35,38,39], restricting the type of clothing worn (i.e., long sleeved shirts) [35,38,39], having users hold markers [22], using a priori knowledge of initial gesture pose [23,71], and using multiple, specially configured cameras [3] or magnetic trackers [3,14,26,36]. 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 affine motion estimator. Tracking drift has not been a significant factor 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 use of a motion estimator with a larger capture range (e.g., consideration of a correlation-based, rather than gradient-based method).

Given acceptable tracking, problems in the classification per se arose from non-intentional but significant motions 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. Further improvements in recognition may be found by integrating

Table 4
Gesture movement recognition results

	Upward	Downward	Up and down	Rightward	Leftward	Side to side	Toward signer	Away signer	To and fro	Supinate	Pronate	Twist wrist	Nod	Circular
Upward	**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
Downward	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**

The axes of the table represent the actual input gesture (vertical) vs. 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.

domain specific knowledge into the classification stage (e.g., frequency counts of phonemes in ASL).

4. Summary

We have presented a novel approach to vision-based gesture recognition; the described work particularly concentrates on the representation and recognition of isolated movement phonemes of ASL. Our general approach is based on two key concepts. First, we exploit 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 (e.g., American sign language). Second, we analytically define distinctive signatures for the primitive components that can be recovered from image sequences captured by a single uncalibrated camera. The proposed set of distinctive signatures were arrived at by analyzing the ideal mappings between the phonemic movements and the kinematic description of the visual motion field on the image plane. This is in contrast to most approaches presented in the literature that use machine learning techniques to instantiate models based on training data. 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 movements (movement phonemes) from which ASL symbols are built. The algorithm recovers kinematic features from a single 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. The evaluation of our algorithm showed strong performance on a significant number of test sequences, 592 video sequences in total, which demonstrates its applicability to the analysis of complex gesture videos. Finally, given the descriptive power of the phonemic decomposition and the demonstrated ability to resolve such from image sequences, in future work we intend to use our approach as a building block for the analysis of streams of lexical gestures.

Acknowledgments

Research was funded by the Institute of Robotics and Intelligent Systems one of the Government of Canada's Networks of Centres of Excellence. K. Derpanis holds a National Sciences and Engineering Research Council of Canada PGS B fellowship. J. Tsotsos holds the Canada Research Chair in Computational Vision.

Appendix. Tracking

In this appendix we provide additional details of the tracker that we use to generate the empirical results described in Section 3 of this paper.

Let $I(\vec{\mathbf{x}}, t)$ represent the image brightness at position $\vec{\mathbf{x}} = (x, y)^\top$ and time t. Using the brightness constancy constraint [49], the interframe motion, $\vec{\mathbf{u}}(\vec{\mathbf{x}}) = (u(\vec{\mathbf{x}}), v(\vec{\mathbf{x}}))^\top$, is defined as

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

We employ an affine model to describe the motion

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

The affine model is used for two main reasons. First, as demonstrated in Section 2.2, 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 exists. 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 [67] operating over a Gaussian pyramid [72]. 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 classifier [68]. As a further level of robustness, we restrict consideration to points that experience a significant change in intensity (i.e. $\mathrm{d}I/\mathrm{d}t$). For robustness in motion estimation, we make use of an M-estimator [73] (e.g., as opposed to a more standard least-squares approach, cf., [55]) 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 used is the Geman-McClure [73].

The motion estimator is applied to adjacent frames across an image sequence. 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.

References

[1] L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] W.C. Stokoe, D. Casterline, C. Croneberg, A Dictionary of American Sign Language, Linstok Press, Washington, DC, 1965.

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

[4] S. Wong, R. Cipolla, Real-time adaptive hand motion recognition using a sparse Bayesian classifier, in: Proceedings of the IEEE International Workshop on Human–Computer Interaction, 2005, pp. 170–179.

[5] S. Wong, R. Cipolla, Real-time interpretation of hand motions using a sparse Bayesian classifier on motion gradient orientation images, in: Proceedings of the British Machine Vision Conference, vol. I, 2005, pp. 379–388.

[6] H. Poizner, U. Bellugi, V. Lutes-Driscoll, Perception of American sign language in dynamic point-light displays, Journal of Experimental Psychology: Human Perception and Performance 7 (2) (1981) 430–440.

[7] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Computer Vision and Image Understanding 73 (3) (1999) 428–440.

[8] S. Ong, S. Ranganath, Automatic sign language analysis: a survey and the future beyond lexical meaning, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (6) (2005) 873–891.

[9] V.I. Pavlovic, R. Sharma, T.S. Huang, Visual interpretation of hand gestures for human–computer interaction: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 677–695.

[10] M. Shah, R. Jain, Visual recognition of activities, gestures, facial expressions and speech: an introduction and a perspective, in: M. Shah, R. Jain (Eds.), Motion-Based Recognition, 1997, pp. 1–14.

[11] J.M. Rehg, T. Kanade, Visual tracking of high DoF articulated structures: an application to human hand tracking, in: Proceedings of the European Conference on Computer Vision, vol. B, 1994, pp. 35–46.

[12] B.D.R. Stenger, P.R.S. Mendonca, R. Cipolla, Model-based hand tracking using an unscented Kalman filter, in: Proceedings of the British Machine Vision Conference, 2001.

[13] M. Black, A. Jepson, Eigentracking: Robust matching and tracking of articulated objects using a view-based representation, International Journal of Computer Vision 26 (1) (1998) 63–84.

[14] A.F. Bobick, A.D. Wilson, A state-based approach to the representation and recognition of gesture, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (12) (1997) 1325–1337.

[15] J. Davis, M. Shah, Toward 3-d gesture recognition, International Journal of Pattern Recognition and Artificial Intelligence 13 (3) (1999) 381.

[16] N. Gupta, P. Mittal, S. Dutta Roy, S. Chaudhury, S. Banerjee, Developing a gesture-based interface, IETE Journal of Research 48 (3) (2002) 237–244.

[17] P. Hong, M. Turk, T.S. Huang, Gesture modeling and recognition using finite state machines, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 410–415.

[18] R. Herpers, K. Derpanis, W. MacLean, G. Verghese, M. Jenkin, E. Milios, A. Jepson, J. Tsotsos, SAVI: an actively controlled teleconferencing system, Image and Vision Computing 19 (11) (2001) 793–804.

[19] M. Yeasin, S. Chaudhuri, Visual understanding of dynamic hand gestures, Pattern Recognition 33 (11) (2000) 1805–1817.

[20] T. Darrell, A. Pentland, Space-time gestures, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1993, pp. 335–340.

[21] Y. Zhu, H. Ren, G. Xu, X. Lin, Toward real-time human-computer interaction with continuous dynamic hand gestures, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 544–549.

[22] M.J. Black, A.D. Jepson, A probabilistic framework for matching temporal trajectories: CONDENSATION-based recognition of gesture and expression, in: Proceedings of the European Conference on Computer Vision, vol. II, 1998, pp. 909–924.

[23] M. Isard, A. Blake, CONDENSATION – conditional density propagation for visual tracking, International Journal of Computer Vision 29 (1) (1998) 5–28.

[24] R. Cutler, M. Turk, View-based interpretation of real-time optical flow for gesture recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 416–421.

[25] J. Mammen, S. Chaudhuri, T. Agarwal, A two stage scheme for dynamic hand gesture recognition, in: Proceedings of the National Conference on Communication, 2002, pp. 35–39.

[26] S.S. Fels, G.E. Hinton, Glove-talk. II. A neural network interface which maps gestures to parallel format speech synthesizer controls, IEEE Transaction on Neural Networks 9 (1) (1997) 205–212.

[27] M.B. Waldron, Isolated ASL sign recognition system for deaf persons, IEEE Transactions on Rehabilitation Engineering 3 (3) (1995) 261–271.

[28] M.H. Yang, N. Ahuja, M. Tabb, Extraction of 2D motion trajectories and its application to hand gesture recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (8) (2002) 1061–1074.

[29] B. Bauer, H. Hienz, Relevant features for video-based continuous sign language recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 440–445.

[30] D.A. Becker, Sensei: A real-time recognition, feedback, and training system for T'ai Chi gestures, in: Proceedings of Vismod, 1997.

[31] G. Fang, W. Gao, A SRN/HMM system for signer-independent continuous sign language recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2002, pp. 297–302.

[32] K. Grobel, M. Assan, Isolated sign language recognition using hidden Markov models, in: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 1997, pp. 162–167.

[33] E. Holden, G. Lee, R. Owens, Automatic recognition of colloquial Australian sign language, Proceedings of the in: IEEE Workshop on Motion and Video Computing, vol. II, 2005, pp. 183–188.

[34] C.L. Huang, S.H. Jeng, A model-based hand gesture recognition system, Machine Vision and Applications 12 (5) (2001) 243–258.

[35] H.K. Lee, J.H. Kim, An HMM-based threshold model approach for gesture recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (10) (1999) 961–973.

[36] R.H. Liang, M. Ouhyoung, A real-time continuous gesture recognition system for sign language, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 558–567.

[37] Y. Nam, K. Wohn, H. Lee-Kwang, Modeling and recognition of hand gesture using colored Petri Nets, IEEE Transactions on Sytems, Man, and Cybernetics 29 (5) (1999) 514–521.

[38] J. Schlenzig, E. Hunter, R. Jain, Vision based gesture interpretation using recursive estimation, in: Proceedings of the Asilomar Conference on Signals, Systems and Computers, 1994.

[39] T. Starner, J. Weaver, A.P. Pentland, Real-time American sign language recognition using desk and wearable computer based video, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (12) (1998) 1371–1375.

[40] N. Tanibata, N. Shimada, Y. Shirai, Extraction of hand features for recognition of sign language words, in: Proceedings of the International Conference on Vision Interface, 2002, pp. 391–398.

[41] A.D. Wilson, A.F. Bobick, Parametric hidden Markov models for gesture recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (9) (1999) 884–900.

[42] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

[43] N. Badler, Temporal scene analysis: conceptual descriptions of object movements, Department of Computer Science, University of Toronto, Rep. TR-80, 1975.

[44] J.K. Tsotsos, J. Mylopoulos, H.D. Covvey, S.W. Zucker, A framework for visual motion understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence 2 (6) (1980) 563–573.

[45] H. Kollnig, H.H. Nagel, M. Otte, Association of motion verbs with vehicle movements extracted from dense optical flow fields, in: Proceedings of the European Conference on Computer Vision, vol. B, 1994, pp. 338–347.

[46] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, M. Brady, A linguistic feature vector for the visual interpretation of sign language, in: Proceedings of the European Conference on Computer Vision, 2004, vol. I, pp. 390–401.

[47] K.G. Derpanis, R.P. Wildes, J.K. Tsotsos, Hand gesture recognition within a linguistics-based framework, in: Proceedings of the European Conference on Computer Vision, 2004, Part 2, pp. 282–296.

[48] C. Valli, C. Lucas, Linguistics of American sign language an introduction, Gallaudet University Press, Washington, DC, 2000.

[49] B.K.P. Horn, Robot Vision, MIT Press, Cambridge, MA, 1986.

[50] S. Negahdaripour, S. Lee, Motion recovery from image sequences using first-order optical flow information, in: Proceedings of the IEEE Workshop on Visual Motion, 1991, pp. 132–139.

[51] K. Derpanis, Vision based gesture recognition within a linguistics framework, Tech. rep., York University, Department of Computer Science TR-CS-2004-02, 2004.

[52] S.S. Beauchemin, J.L. Barron, The computation of optical-flow, ACM Computing Surveys 27 (3) (1995) 433–467.

[53] F.G. Meyer, P. Bouthemy, Region-based tracking using affine motion models in long image sequences, Computer Vision, Graphics, and Image Processing 60 (2) (1994) 119–140.

[54] K.Y. Wong, M.E. Spetsakis, Motion segmentation and tracking, in: Proceedings of the International Conference on Vision Interface, 2002, pp. 80–87.

[55] J.R. Bergen, P. Anandan, K.J. Hanna, R. Hingorani, Hierarchical model-based motion estimation, in: Proceedings of the European Conference on Computer Vision, vol. I, 1992, pp. 5–10.

[56] S. Christy, R. Horaud, Euclidean shape and motion from multiple perspective views by affine iterations, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (11) (1996) 1098–1104.

[57] J.J. Koenderink, A.J. van Doorn, Affine structure from motion, Journal of the Optical Society of America: A 8 (2) (1991) 377–385.

[58] J.M. Lawn, R. Cipolla, Robust egomotion estimation from affine motion parallax, in: Proceedings of the European Conference on Computer Vision, vol. A, 1994, pp. 205–210.

[59] C.J. Poelman, T. Kanade, A paraperspective factorization method for shape and motion recovery, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (3) (1997) 206–218.

[60] L.S. Shapiro, Affine Analysis of Image Sequences, Oxford University Press, 1993.

[61] P. Bouthemy, M. Gelgon, F. Ganansia, A unified approach to shot change detection and camera motion characterization, IEEE Transactions on Circuits and Systems for Video Technology 9 (7) (1999) 1030–1044.

[62] J.J. Koenderink, A.J. van Doorn, Local structure of movement parallax of the plane, Journal of the Optical Society of America 66 (7) (1976) 717–723.

[63] H.C. Longuet-Higgins, K. Pradzny, The interpretation of a moving retinal image, Proceedings of the Royal Society of London B 208 (1980) 385–397.

[64] A.M. Waxman, S. Ullman, Surface structure and three-dimensional motion from image flow kinematics, International Journal of Robotics Research 4 (3) (1985) 72–94.

[65] R. Aris, Vectors Tensors and the Basic Equations of Fluid Mechanics, Dover Publications, New York, NY, 1989.

[66] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufman, San Francisco, CA, 2001.

[67] M. Black, P. Anandan, The robust estimation of multiple motions: Parametric and piecewise-smooth flow-fields, Computer Vision and Image Understanding 63 (1) (1996) 75–104.

[68] M.J. Jones, J.M. Rehg, Statistical color models with application to skin detection, International Journal of Computer Vision 46 (1) (2002) 81–96.

[69] S. Lu, D. Metaxas, D. Samaras, J. Oliensis, Using multiple cues for hand tracking and model refinement, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. II, 2003, pp. 443–450.

[70] C. Sminchisescu, B. Triggs, Kinematic jump processes for monocular 3D human tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. I, 2003, pp. 69–76.

[71] A. Elgammal, V. Shet, Y. Yacoob, L. Davis, Learning dynamics for exemplar-based gesture recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. I, 2003, pp. 571–578.

[72] B. Jahne, Digital Image Processing-Concepts, Algorithms, and Scientific Applications, Springer, Berlin, 1991.

[73] P.J. Huber, Robust Statistical Procedures, SIAM Press, Philadelphia, PA, 1977.