Vision Based Gesture Recognition within a Linguistics Framework

Konstantinos G. Derpanis

Richard P. Wildes

John K. Tsotsos

Technical Report CS-2004-02

July 12, 2004

Department of Computer Science

4700 Keele Street North York, Ontario M3J 1P3 Canada

Abstract

An approach to recognizing human hand gestures from a monocular temporal sequence of colour images is presented. Of particular concern in this report is the representation and recognition of hand movements that are used in single handed American Sign Language (ASL). The approach presented 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 subsequently demonstrate that given a monocular gesture sequence, kinematic features can be recovered from the apparent motion that provide distinctive signatures for 14 single-handed rigid 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 97.13%.

Contents

List of Tables

List of Figures

Chapter 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 developed 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.

In this regard, currently we are particularly focused on the representation and recovery of the movement primitives of hand gestures, specifically single-handed movements derived from American Sign Language (ASL). Motivated by

the preceding observations, this report presents an approach to recognizing hand gestures that leverages both linguistic theory and computer vision methods in response to the challenges at hand. Following a path that has been taken in the speech recognition community for the interpretation of vocal data [61], 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 approach ensures that the developed approach is scalable (see Section 2.3 for details).

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 [60]. 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 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., [2, 58, 68] for general reviews).

Here, we highlight several representative approaches. (See Tables 1.1 and 1.2 for a condensed summary of various hand gesture related work.) For the specific problem of gesture recognition, the basic approach taken 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 [63, 71] or an appearance based model [12], with the image (or images in the case of a multi-camera setup, e.g. [63]) for the purposes of tracking. These approaches have met 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. 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 must match the order in time of a model of states [14, 23, 31, 35, 52, 86]. In [22, 89] this problem has been formulated as a pattern recognition problem, where the dynamic time warping (DTW) method is used to temporally align (i.e. match) an input pattern (i.e. series of states) to a stored pattern. An alternative approach has appealed to the use of statistical factored sampling in conjunction

with a model of parameterized gestures to affect recognition [13]; this approach can be seen as an application and extension of the CONDENSATION approach to visual tracking [39]. 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 the 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 [21, 53]. 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 [21] 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 N consecutive frames.

Neural networks and their extensions, time-delay neural networks (TDNN), have also been applied to the gesture recognition problem [25, 79, 85]. 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) [6, 8, 24, 30, 37, 48, 49, 56, 67, 70, 74, 78, 81] which had been previously successfully applied to the problem of speech recognition; for an excellent tutorial on the topic of HMMs see [62]. Hidden Markov Models are a statistical method that relies on the assumption that the output can be well approximated by a sequence of unobservable (hidden) states where the observation is a probabilistic function of the state. In order to instantiate the model parameters, numerous exemplars of the observation sequence are used to train the model. A standard application of HMMs to the problem of gesture recognition is to have each gesture associated with an HMM, 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 is its non-linear time scaling invariance property resulting from recurrent states in the hidden state topology and HMMs can handle a continuous input stream without it being explicitly temporally segmented. Disadvantages of HMMs include the following: possibilities that the underlying Markov assumption (i.e. hidden state topology) does not hold, the training stage may overfit the HMM to the training data and the HMM not capturing the essential aspects of the underlying process caused by insufficient training data and/or unrepresentative training data.

A number of the cited approaches have been able to achieve 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 to represent a large vocabulary from a small set of generative elements. Instead, gestures tend to be 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 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 [5, 75], two of the earliest efforts of using linguistic concepts in the description and recognition of both general and domain specific motion are presented. More recently, in [46] it was shown how "motion verbs" can be associated with image motion patterns. For the problem of gesture recognition, at least two previous lines of investigations have appealed to linguistic theory as an attack on issues in scaling gesture recognition to sizable vocabularies [49, 78]. Based on the ASL linguistics literature, the authors promote a phoneme based modelling of gestures. In [49] 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 [78],

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 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. Here, it is interesting to note that in [54] the authors report on the most extensive video database of ASL to date based on capturing the phonemic elements of ASL[1].

[1]This database was not used in our experiments since to date, the database has not been released publicly due to logistical issues.

Work	Features	Feature Extraction	Recognition Method	Application Domain
Bauer et. al. [6]	hand location, shape orientation, 2D movement	coloured glove	HMM	German Sign Language
Becker [8]	hand location	colour marker tracker	HMM	T'ai Chi
Black et. al. [13]	2D hand trajectory	colour marker tracker	CONDENSATION	white board manipulation
Bobick et. al. [14]	sequence of appearances	eigenspace	finite state machine	generic
Braffort [16]	location, hand shape and movement	data glove	not addressed	French Sign Language
Cutler et. al. [21]	2D hand trajectory	optical flow (correlation)	rule-based	interface for children
Darrell et. al. [22]	sequence of hand views	correlation scores	dynamic time warpintg	generic
Davis et. al. [23]	2D finger trajectories	marker based finger-tip tracking	finite state machine	generic
Fang et. al. [24]	location, hand shape orientation, movement	data glove	HMM	Chinese Sign Language
Fels et. al. [25]	3D hand movement	data glove	neural network	generic
Freeman et. al. [28]	static hand pose	edge map	orientation histogram	generic
Grobel et. al. [30]	location, hand shape orientation	coloured glove	HMM	Sign Language of the Netherlands
Gupta et. al. [31]	sequence of hand appearances	eigenspace	finite state machine	generic
Holden et. al. [34]	hand shape, movement	data glove	fuzzy expert system	Australian Sign Language
Hong et. al. [35]	2D hand trajectory	skin colour	finite state machine	interface for children
Huang et. al. [37]	2D trajectory, hand shape	edge extraction	HMM	generic

Table 1.1: Summary of automated hand gesture recognition work.

Work	Features	Feature Extraction	Recognition Method	Application Domain
Lee et. al. [48]	2D hand trajectory	skin colour tracker	HMM	generic
Liang et. al. [49]	hand shape, orientation, movement	data glove	HMM	Taiwanese Sign Language (phonemic based)
Maclean et. al. [52]	sequence of hand shapes	skeletonization of skin coloured region	finite state machine	camera control
Mammen et. al. [53]	2D trajectory, hand shape	skin colour	rule-based	telerobotic interface
Nam et. al. [56]	hand shape, orientation, movement	data glove	HMM	generic
Sagawa et. al. [65]	location, hand shape, movement	data glove	various specialized algorithms	Japanese Sign Language
Schlenzig, et. al. [67]	sequence of appearances	image moments	HMM	telerobotic control
Starner et. al. [70]	2D hand pos, orientation	skin colour tracker	HMM	ASL
Tanibata et. al. [74]	hand shape, movement	skin colour	HMM	Japanese Sign Language
Vogler et. al. [78]	3D hand position orientation	3D electromagnetic tracker or physics based vision arm tracker	HMM	ASL (phonemic based)
Waldron et. al. [79]	hand shape, location, orientation, movement	data glove	neural network	ASL
Wilson et. al. [81]	3D position	stereo colour blob	HMM	generic
Yang et. al. [85]	2D hand trajectory	multiscale region matching	TDNN	ASL
Yeasin et. al. [86]	2D hand trajectory	difference images and temporal zero crossings	finite state machine	telerobotic control
Zhu et. al. [89]	2D trajectory, hand shape	affine motion parameters	dynamic time warping	generic

Table 1.2: Summary of automated hand gesture recognition work, continued.

1.3 Contributions

In the light of previous research, the main contributions of this report are as follows.

- Our approach makes use of linguistic theory to model gestures in terms of their phonemic elements to yield an algorithm that recognizes gesture movement primitives given data captured with a single video camera.
- We analytically derive the mappings for a subset of single handed movements to a kinematic feature space describing the visual motion field; this analysis is then leveraged in our classification scheme.
- 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, electromagnetic tracker).
- Our recognition scheme is based on a nearest neighbour match to prototype signatures, where each of the movement primitives of ASL under consideration is found to have a distinctive prototype signature in a kinematic feature space.
- We have evaluated our approach empirically with 592 video sequences taken from 15 volunteers situated with a frontoparallel attitude with respect to the camera and find that our algorithm is capable of reliably recognizing movement primitives, 97.13% phoneme accuracy rate, even as other aspects

of the gesture vary, namely, hand shape and location. Additionally, we have conducted a preliminary study on the robustness of our algorithm to changes of the attitude of the signer with respect to the camera.

1.4 Outline of report

This report is subdivided into four main chapters. This first chapter has provided motivation for modelling gestures at the phoneme level. Chapter 2 describes the linguistic-basis of our representation, presents an analysis of the mappings between the phonemic movements and kinematic quantities describing the apparent motion, as well as the algorithmic aspects of the approach. Chapter 3 documents experimental evaluation of a software implementation of our algorithm. Finally, chapter 4 provides a summary of our work, as well as possible future directions.

Chapter 2

Technical Approach

2.1 General technical framework

Our approach to gesture recognition centres around two motivating 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 in as minimally constrained a fashion as possible, e.g., with a standard video camera. The first level of the decomposition is in terms of three sets of primitives, hand shapes, location and movement. Of present concern in this report is the recovery of linguistically defined single hand movement primitives of American Sign Language (ASL). Further levels of decomposition involve the lexical and sentence levels and are part of our plan for future work. For a pictorial overview of our work in a broader perspective see Fig. 2.1.

We take the input to our system to be a temporal sequence of images that depicts a single movement phoneme taken from a roughly frontoparallel view of

the signer with respect to the camera. Currently, we also assume that the region corresponding to the hand in the first frame has been delineated manually. (For a proposed fully automated localization method see Appendix B.) 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, hand shape and signer). To affect the recognition, a robust, affine motion estimator is applied to regions of interest defined by skin colour on a frame-to-frame basis. Though skin colour segmentation is not strictly required given the fact that the hands are currently manually outlined, the inclusion of skin colour segmentation will play an important future role with the inclusion of an automated hand localization scheme which may oversegment the hand region. The resulting time series of affine parameters is mapped to a kinematic time series that is in turn individually accumulated across the sequence to yield a signature that is used for classification of the depicted gesture.

The remainder of this chapter consists of the following: details of the phonemic movement vocabulary (Section 2.2), the strategy pertaining to exploiting ASL linguistics for the problem of vision based hand gesture recognition (Section 2.3), an analytic derivation of mappings of the phonemic movements to kinematic quantities describing their apparent motion (Section 2.4) and details of the various processing stages of a phonemic movement classifier algorithm exploiting the findings of our analytic derivation (Section 2.5, 2.6, 2.7 and 2.8).

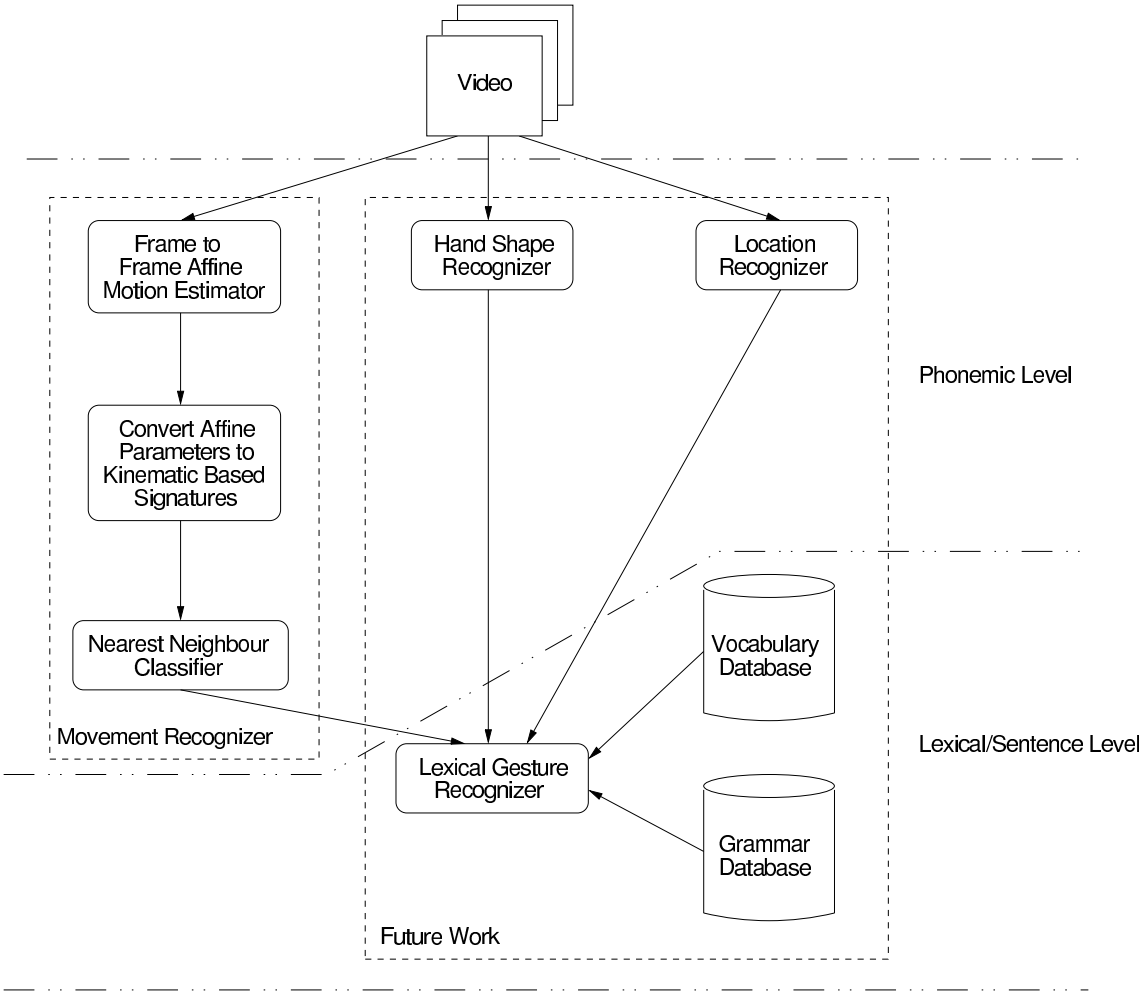


Figure 2.1: System framework

2.2 Linguistics basis

Webster's dictionary [1] defines linguistics as "the study of the nature, structure, and variation of language, including phonetics, phonology, morphology, syntax, semantics, sociolinguistics, and pragmatics". In the current context we leverage the linguistic structure of ASL to yield a scalable framework for sizable gesture vocabularies.

Prior to William Stokoe's seminal work in ASL [72], it was assumed that linguistics was exclusive to the study of human speech. Sign language was regarded by linguists as a series of pictorial gestures with no linguistic structure. Stokoe demonstrated that signing was indeed a rich, linguistically complex language. A fundamental contribution of his work was redefining the basic unit of a sign to units he termed cheremes[1] as opposed to the sign as a whole, these units are analogous to speech phonemes: minimally contrastive patterns that distinguish the symbolic vocabulary of a language. Stokoe's work culminated in the first true dictionary of ASL [73]; the dictionary described over 2000 signs.

Stokoe's system consists of three parameters that are executed simultaneously and sequentially to define a gesture, see Fig. 2.2. The three parameters capture location ("Where on the body or in space is the sign being made?"), hand shape ("How are the fingers extended and bent in this particular sign?") and movement

[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.

("How does the hand(s) move?"). Extensions to the basic Stokoe system include the orientation of the palm and non-manual gestures (e.g. facial expressions). These additions are not considered in this report.

In terms of location, there are 12 elemental locations defined by Stokoe. The locations reside 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 direction and extending close to the extents of the signer's body in the horizontal direction. As for hand shapes, there are 19 possible hand shapes in the Stokoe system. Other authors (e.g. [50]) have indicated that the actual hand shape space may be quite larger, for the most part the additional hand shapes cited in the ASL linguistics literature represent subtle differences from Stokoe's hand shape phonemes. While Stokoe's complete vocabulary of movements consists of 24 primitives (i.e. single and two-handed movements), here, as a starting point, we restrict consideration to 14 *single handed* movements, shown in Fig. 2.2. Current ASL linguistic theories still recognize Stokoe system's three basic parameters but differ in their definition of the constituent elements of the parameters [76].

In ASL each hand has a distinct role. The dominant hand is the hand that performs the one-handed signs and the major component of two-handed signs. The non-dominant hand is the opposite of the dominant hand. For right-handed signers, the dominant hand is typically the signer's right hand the non-dominant hand is the signer's left hand.

We use Stokoe's definition of the parameters for this study as they are well defined by the author, 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 all manual languages and they provide the basis for many more recent developments in manual language linguistics. It is interesting to note that this same linguistic analysis has also been applied to other manual languages.

In the remainder of this , we show how to automatically recognize Stokoe's 14 single handed movements irrespective of signer, hand shape and location. Ability to recognize each parameter (shape, location and movement) independently of each other is key to being able to leverage combinatorics for application to sizeable vocabularies.

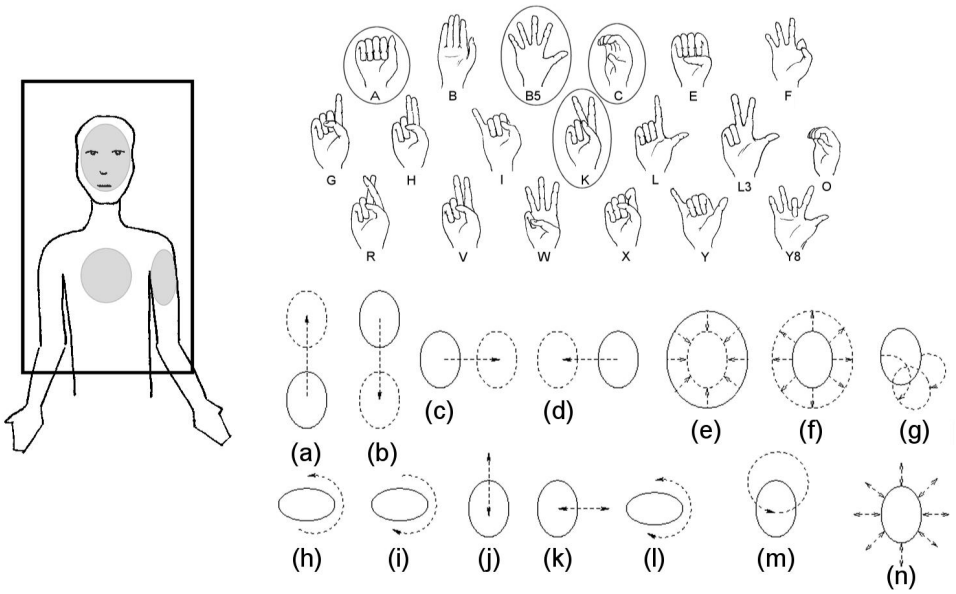


Figure 2.2: Stokoe's Phonemic Analysis of ASL. The left panel depicts the signing space in which the locations reside. Shaded regions indicate locations used in our experiments. The upper right panel depicts possible hand shapes. Circled shapes indicate shapes used in our experiments. The lower right panel depicts possible single handed movements from a frontoparallel pose of the signer with respect to the viewer (i.e. camera) (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 represents the initial hand location, the dashed ellipse represents the final location and the dashed arrow represents the path taken. Our experiments investigate the recognition of movement independent of location and shape.

2.3 Strategy

In this section we outline our motivation for selecting a phonemic modelling of gestures.

Basing a recognition system on phonemes keeps the problem of ASL recognition tractable (i.e. scalable to sizable vocabularies), because there is a small number of phonemes as compared to the unlimited number of lexical gestures that can be formed from the phonemes [78]. Specifically, the tractability problem is related to the impracticality of building individual models, either by training examples (e.g. [78]) or analytically constructing models (as is done in this report), when the gestures are modelled as wholes. On the other hand constructing a recognition system by modelling and recognizing a limited small number of phonemes that in turn can be used to recognize lexical gestures is a feasible task. For instance using the Stokoe model, there are 24 (movements) + 19 (shapes) + 12 (locations) = 55 phonemes. Even using a more current ASL model which include finer phonemic descriptions, the number of phonemes is still quite small numbering approximately 150-200 phonemes as compared to the roughly 4500 signs documented in a recent ASL dictionary [20].

Unlike speech where phonemes are realized in series, in ASL, phonemes are realized both in parallel and in series within a gesture. The number of possible combinations of phonemes formed simultaneously using Stokoe's parameter definitions equals 24 (movements, both single and two handed) x 19 (shapes) x 12

(locations) = 5472 combinations. The Stokoe definitions can be thought of as the minimum required to describe ASL, some authors have proposed far richer phonemic descriptions that have brought the number of combinations to as many as 1.5×10^9 number of phonemic combinations [50]. Modelling all the phonemic combinations brings us back to the initial practicality problem of modelling the gestures. To make the problem tractable we assume that each of the parameters can be recognized irrespective of the others. In terms of linguistics this is a fair assumption since the phonemes are realized from independent parameters, in reality the kinematic structure of the body introduces slight coarticulation affects. 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 right side, a slight rotation is introduced to bring the hand roughly parallel with the camera. We will show that our approach is highly robust to these coarticulation effects.

Faced with noisy input, a portion of which is the non-dominant structured noise due to coarticulation effects, we present in the following sections a qualitative approach to recognizing dominant structures (signatures) related to each of the 14 single handed movements under consideration.

2.4 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 described by Stokoe and the kinematic description of the visual motion field on the imaging plane. This will provide us with a principled approach to recognizing movement phonemes.

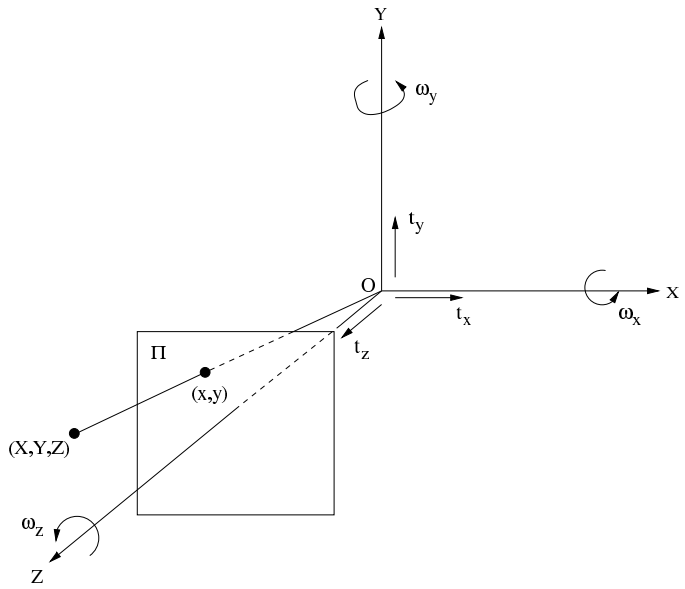


Figure 2.3: Camera coordinate system. Depicted is the camera coordinate system, with an image plane Π located at $Z = 1$. Perspective projection maps a point (X, Y, Z) to (x, y). The parameters t_x, t_y and t_z represent the translational velocities in the X, Y and Z directions respectively, ω_x, ω_y 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).

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, $\Omega = (\omega_x, \omega_y, \omega_z)^\top$, about the X, Y and Z Euclidean axes respectively, with the origin defined at the camera's centre of projection (see Fig. 2.3). 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 [36], 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,

$$\begin{aligned} u &= \omega_y - \omega_z y - \omega_x xy + \omega_y x^2 + \frac{t_x + q_y\omega_z - q_z\omega_y + (q_y\omega_x - t_z - q_x\omega_y)x}{Z} \\ v &= -\omega_x + \omega_z x + \omega_y xy - \omega_x y^2 + \frac{t_y + q_z\omega_x - q_x\omega_z + (q_y\omega_x - t_z - q_x\omega_y)y}{Z} \end{aligned} \tag{2.1}$$

see [36] and Appendices A.1, A.2 and A.2 for this standard derivation.

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,

$$\alpha X + \beta Y + \gamma Z = 1 \tag{2.2}$$

where

$$\begin{aligned} \alpha &= \frac{n_x}{d} \\ \beta &= \frac{n_y}{d} \\ \gamma &= \frac{n_z}{d} \\ d &= n_x X_0 + n_y Y_0 + n_z Z_0, \end{aligned} \tag{2.3}$$

$\vec{n} = (n_x, n_y, n_z)^\top$ 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), is modelled through first-order in image coordinates by an affine

transformation, formally,

$$\begin{aligned} u(x,y) &= a_0 + a_1 x + a_2 y \\ v(x,y) &= a_3 + a_4 x + a_5 y \end{aligned} \tag{2.4}$$

where

$$\begin{aligned} a_0 &= \omega_y + (\omega_z q_y + t_x - \omega_y q_z)\gamma \\ a_1 &= (\omega_z q_y + t_x - \omega_y q_z)\alpha + (\omega_x q_y - \omega_y q_x - t_z)\gamma \\ a_2 &= -\omega_z + (\omega_z q_y + t_x - \omega_y q_z)\beta \\ a_3 &= -\omega_x + (\omega_x q_z + t_y - \omega_z q_x)\gamma \\ a_4 &= \omega_z + (\omega_x q_z + t_y - \omega_z q_x)\alpha \\ a_5 &= (\omega_x q_z + t_y - \omega_z q_x)\beta + (\omega_x q_y - \omega_y q_x - t_z)\gamma \end{aligned} \tag{2.5}$$

for the derivation see Appendix A.1, A.2 and A.3.

The selection of truncating the analytically correct quadratic flow after the first-order terms is necessitated by the fact that the second-order coefficients become sensitive to image noise and cannot be estimated accurately given a small region of support [57]. Fortunately, this does not pose a problem since the contributions from the second-order terms are small, particularly near the image centre ($x, y << f = 1$) [57]; also in Appendix A.4 we present an analytic study as well as a simulation of the contribution of the second-order terms and conclude that indeed the contributions of the second-order terms are negligible for our situation. Moreover, it will be subsequently shown in this section that the zeroth and first-order terms are sufficient in providing 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 [7], 2D tracking [55, 82], image registration [9], 3D structure and/or motion estimation [18, 45, 47, 59, 69], video partitioning [15] and hand gesture recognition [89].

Owing to their 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*) (see [44, 51, 80] and Appendix A.5).

$$
\begin{array}{rcc}
hor & = & a_0 \\
 & = & \omega_y + (\omega_z q_y + t_x - \omega_y q_z)\gamma \\
 & & \\
ver & = & a_3 \\
 & = & -\omega_x + (\omega_x q_z + t_y - \omega_z q_x)\gamma \\
 & & \\
div & = & a_1 + a_5 \\
 & = & (\omega_z q_y + t_x - \omega_y q_z)\alpha + 2(\omega_x q_y - \omega_y q_x - t_z)\gamma \\
 & & +(\omega_x q_z + t_y - \omega_z q_x)\beta \\
 & & \\
curl & = & -a_2 + a_4 \\
 & = & 2\omega_z - (\omega_z q_y + t_x - \omega_y q_z)\beta + (\omega_x q_z + t_y - \omega_z q_x)\alpha \\
 & & \\
def & = & ([a_1 - a_5]^2 + [a_2 + a_4]^2)^{1/2} \\
 & = & ([(\omega_z q_y + t_x - \omega_y q_z)\alpha - (\omega_x q_z + t_y - \omega_z q_x)\beta]^2 \\
 & & +[(\omega_z q_y + t_x - \omega_y q_z)\beta + (\omega_x q_z + t_y - \omega_z q_x)\alpha]^2)^{1/2}
\end{array}
\tag{2.6}
$$

We now show that the kinematic quantities are sufficient for describing the phonemic movements. For the leftward, rightward, side to side, upward, down-

ward, up and down, toward signer, away signer, to and fro and circular movements (depicted in Fig. 2.2) we assume that the plane (i.e. hand) is kept parallel to the image plane throughout the execution of the movement, this is reflected by the surface normal $\vec{n} = (0, 0, -1)^\top$, the plane initially contains the point (0,0,c) where $c > 0$, the movements are executed with a constant velocity and $(q_x, q_y, q_z)^\top = (0, 0, 0)^\top$ (i.e. rotation is about the origin).

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, formally, $t_x(t) = k$, where time instant $t = 1 \ldots N$, N represents the total length of the gesture sequence, k is a constant, $k > 0$ for leftward, $k < 0$ for rightward and $t_y(t) = t_z(t) = \omega_x(t) = \omega_y(t) = \omega_z(t) = 0$. In kinematic space this maps to $hor(t) = t_x(t)\gamma$ throughout the gesture, $hor(t) < 0$ for rightward and $hor(t) > 0$ for leftward; all other quantities are zero, i.e. $ver = div = curl = def = 0$.

side to side movement t_x has a constant magnitude velocity with its sign changing mid-gesture, all other world velocities are zero. This maps in kinematic space to $hor(t) = t_x(t)\gamma < 0$ while $t \leq N/2$ and $hor(t) = t_x(t)\gamma > 0$ while $t > N/2$ or alternatively $hor(t) = t_x(t)\gamma > 0$ while $t \leq N/2$ and $hor(t) = t_x(t)\gamma < 0$ while $t > N/2$; all other quantities are zero.

upward/downward consists of a constant value for t_y, positive for upward, negative for downward, all other world velocities are zero. This maps in

kinematic space to $ver(t) = t_y(t)\gamma$, $ver(t) > 0$ for upward and $ver(t) < 0$ for downward; all other quantities are zero.

up and down t_y has a constant magnitude velocity with its sign changing mid-gesture, all other world velocities are zero. This maps in kinematic space to $ver(t) = t_y(t)\gamma > 0$ while $t \leq N/2$ and $ver(t) = t_y(t)\gamma < 0$ while $t > N/2$ or alternatively $ver(t) = t_y(t)\gamma < 0$ while $t \leq N/2$ and $ver(t) = t_y(t)\gamma > 0$ while $t > N/2$; all other quantities 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. This maps in kinematic space to $div(t) = -2t_z\gamma$, $div(t) > 0$ for away signer, $div(t) < 0$ for toward signer; all other quantities are zero. Note that the magnitude of the $div(t)$ changes throughout the gesture execution since the point defining plane changes (reflected by the γ parameter).

to and fro t_z has a constant magnitude velocity with its sign changing mid-gesture, all other world velocities are zero. This maps in kinematic space to $div(t) = 2t_z(t)\gamma > 0$, for $t \leq N/2$ and $div(t) = 2t_z(t)\gamma < 0$, for $t > N/2$ or alternatively $div(t) = 2t_z(t)\gamma < 0$, for $t \leq N/2$ and $div(t) = 2t_z(t)\gamma > 0$, for $t > N/2$; all other quantities 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. This movement maps directly to both $hor(t) = \gamma\omega cos(\omega * t)$ and $ver(t) = -\gamma\omega sin(\omega * t)$; all other quantities are zero.

Unlike the previous movements described, the orientation of the plane (i.e. hand) for the supinate, pronate and twist wrist movements (depicted in Fig. 2.2) 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, dictate 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 positive for supinating and negative for pronating, all other world velocities are zero. In kinematic space this maps to a constant $curl(t) = 2\omega_z$, positive for supinate and negative for pronating; all other quantities are zero.

twist wrist ω_z is constant valued while all other world velocities are zero. This maps in kinematic space to $curl(t) = 2\omega_z > 0$ for $t \leq N/2$, otherwise

$curl(t) = 2\omega_z < 0$ or alternatively $curl(t) = 2\omega_z < 0$ for $t \leq N/2$, otherwise $curl(t) = 2\omega_z > 0$; all other quantities 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. In kinematic space the nod consists of $div(t) = \beta q_z \omega_x < 0$ and $def(t) = |\beta q_z \omega_x|$, where $|\cdot|$ represents the absolute value operator; all other kinematic quantities are zero.

For a summary of the mappings for each of the phonemic movements to the kinematic quantities see Tables (2.1) and (2.2). Significantly in this section, we have shown that the considered phonemic movements exhibit distinctive patterns as projected on the kinematic quantities. In the following sections we will present a specific approach to tracking and classifying the movement of a hand that exploit the findings of this section.

	Non-Periodic Gestures								
kinematic quantity	*rightward*	*leftward*	*up*	*down*	*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	$\lvert\beta q_z \omega_x\rvert$

Table 2.1: Mappings of the non-periodic movements in the world space to kinematic quantities in the image space. $|\cdot|$ represents the absolute value operator.

	Periodic Gestures				
kinematic quantity	*side-side*	*up down*	*to and fro*	*twist wrist*	*circular*
hor(t)	$t_x\gamma < 0$ and $t \leq N/2$ $t_x\gamma > 0$ otherwise or $t_x\gamma < 0$ and $t \leq N/2$ $t_x\gamma > 0$ otherwise	0	0	0	$\gamma\omega cos(\omega * t)$
ver(t)	0	$t_y\gamma > 0$ and $t \leq N/2$ $t_y\gamma < 0$ otherwise or $t_y\gamma < 0$ and $t \leq N/2$ $t_y\gamma > 0$ otherwise	0	0	$-\gamma\omega sin(\omega * t)$
div(t)	0	0	$-2t_z\gamma > 0$ and $t \leq N/2$ $-2t_z\gamma < 0$ otherwise or $-2t_z\gamma < 0$ and $t \leq N/2$ $-2t_z\gamma > 0$ otherwise	0	0
curl(t)	0	0	0	$2\omega_z > 0$ and $t \leq N/2$ $2\omega_z < 0$ otherwise or $2\omega_z < 0$ and $t \leq N/2$ $2\omega_z > 0$ otherwise	0
def(t)	0	0	0	0	0

Table 2.2: Mappings of the periodic movements in the world space to kinematic quantities in the image space.

2.5 Colour segmentation

A number of gesture recognition and more generally people tracking systems employ skin colour detection as an approximate segmentation step due to its attainable real-time performance and orientation invariant processing under a Lambertian surface reflectance assumption ([10, 26, 41, 70, 85]). In our case, a rough segmentation is crucial because of the motion estimator's requirement that the region of support used (i.e. the hand), exhibit the same 3D motion. In this section we provide a brief introduction to colour theory as it relates to the colour appearance of human skin and provide a description of our approach to skin colour segmentation.

Of particular concern in this report is the colour of human hands (i.e. skin colour). The distribution of skin colour in the visible spectrum depends primarily on the concentration of melanin and hemoglobin in human skin [77]. Numerous studies have reported that skin colour amongst different ethnic groups has a compact distribution in chromaticity space (i.e. a definition of colour omitting brightness) while widely differing in intensity (e.g. [41]), see Fig. 2.5 for an example skin distribution. Given this result various attempts have been made to quantify this distribution for the purpose of skin segmentation.

Existing skin colour labelling methods are generally differentiated by the selected colour space, such as HSV [33, 42, 43, 88], YUV [19, 83], RGB [41], and normalized RG [66, 84] and the selected classification method, which includes

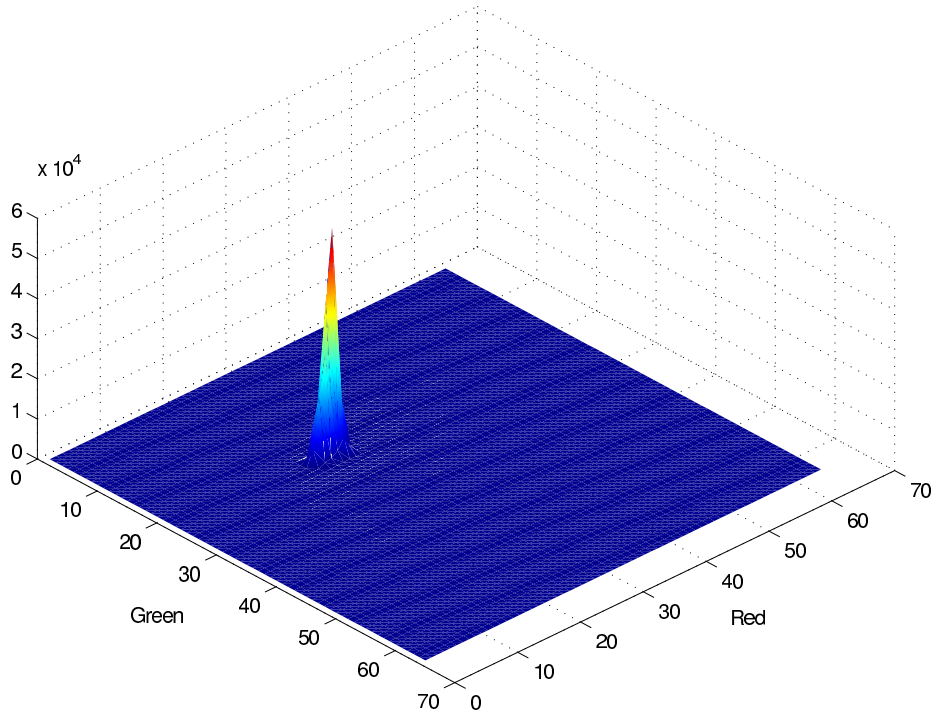


Figure 2.4: Skin colour distribution in normalized RG space. The distribution was constructed from a total of 241,835 pixels obtained by sampling hand regions in our experimental video sequences.

methods such as delineating regions of the colour space associated with the object (i.e. thresholding) [27], histogramming methods [43, 66, 84] and utilizing Bayes' rule to return the likelihood of a pixel being skin by using stored priors of the distribution of skin colour and non-skin colour (approximated by histograms for the respective categories) in the selected colour space [33, 41, 42, 88]. In our present work we have selected the normalized RG colour space since it affords some degree of colour constancy in the case of changes in scene brightness, though not under illuminant changes. In terms of classification we have selected

a Bayesian method due to its computational efficiency and reliable discrimination ability [41, 87].

The normalized RG representation is characterized by the chromaticity tuple (r, g) where each component is defined as,

$$\begin{aligned} r &= R/(R+G+B) \\ g &= G/(R+G+B) \end{aligned} \tag{2.7}$$

The normalized blue channel $b = B/(R+G+B)$ is rendered redundant under the constraint $r + g + b = 1$. Brightness change is characterized as $c(R, G, B)$ where $c > 0$ is a scalar representing change in brightness. It can be seen that the normalized RG model results in the cancellation of the brightness scalar c, thus leaving behind the chroma component of the colour.

In order to utilize the Bayesian approach the priors $P((r,g)|skin)$ and $P((r,g)|\neg skin)$ are estimated off-line by using a set of skin and non-skin histograms as follows,

$$P((r,g)|skin) \approx \frac{num_{skin}((r,g))}{total(skin)} \tag{2.8}$$

$$P((r,g)|\neg skin) \approx \frac{num_{\neg skin}((r,g))}{total(\neg skin)} \tag{2.9}$$

where $num_{skin}((r,g))$ represents the pixel count in bin (r,g) in the skin histogram and likewise $num_{\neg skin}((r,g))$ represents the pixel count in bin (r,g) in the non-skin histograms, $total(skin)$ and $total(\neg skin)$ represent the total number of data points in the skin and non-skin histograms respectively. The histograms were constructed by manually extracting skin and non-skin regions from a set of exemplar

images taken from video sequences used in our experiments. The skin classifier is also a function of the size of the histogram (i.e number of bins in the histogram). In [41] it was reported that too few bins result in significant oversegmentation, while too many bins leads to significant undersegmentation due to over-fitting. To alleviate this problem the authors experimented with different quantization levels (ranging from 16 to 256 bins) and found a histogram of 32 bins returning the best performance. Given that colour segmentation is not the main emphasis of this report only informal experimentation was done on the optimal bin size of the histogram; 32 was found qualitatively to return the best results.

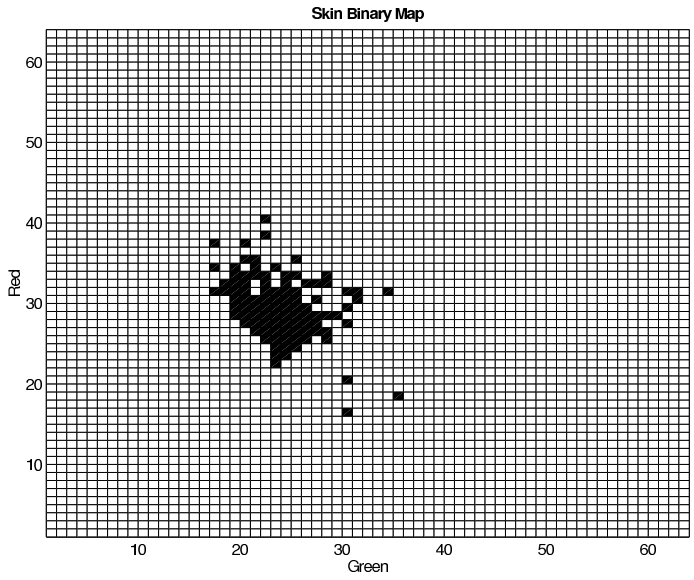


Figure 2.5: Binary skin map. Depicted is the quantized 64×64 normalized RG space. Black denotes bins that when indexed are to be labelled as skin, while white denotes non-skin toned regions in the colour space

Given that we have no prior knowledge of the probability of a pixel containing skin colour, for classification we assume that the probability of a pixel containing

skin colour and the probability of non-skin are equal (i.e. maximum likelihood assumption). This assumption has been evaluated in [87] and found to return reasonable segmentation results. The classification reduces to evaluating the ratio given in (2.10) for each pixel [29]. If the ratio is greater than one the pixel is classified as skin, otherwise the pixel is classified as non-skin.

$$\frac{P(skin|(r,g))}{P(\neg skin|(r,g))} = \frac{P((r,g)|skin)P(skin)/P(r,g)}{P((r,g)|\neg skin)P(\neg skin)/P(r,g)} = \frac{P((r,g)|skin)}{P((r,g)|\neg skin)} \quad (2.10)$$

During an offline stage a binary map of (2.10), depicted in Fig. 2.5, is instantiated and stored, to be indexed during the skin segmentation stage.

2.6 Affine motion estimation

In Section 2.4 we demonstrated that the visual motion field of a hand undergoing each of 14 single handed Stokoe movements can theoretically be described by an affine transformation. In this section we present a robust approach to extracting the affine motion (frame to frame) of the hand from a video sequence.

Rather than resort to estimating motion by looking for correspondences of features, many have analyzed the instantaneous temporal change in the spatial structure of the image over time. This approach is commonly referred to as the gradient approach to finding the optical flow, i.e., apparent motion seen in the image. This is the approach that we have selected given the lack of strong persistent features present on the hand which precludes the use of feature-based methods. Though the optical flow and the visual motion field do not strictly

correspond [36], given the qualitative nature of our approach this does not pose a problem.

The initial assumption made is the brightness constancy constraint [36], which assumes that the brightness structures of local-time varying image regions are unchanging under motion for a short period of time. Formally, this is defined as,

$$I(\vec{x}, t) = I(\vec{x} + \vec{\delta x}, t + \delta t) \tag{2.11}$$

where $\vec{x} = (x, y)^\top$ represents image position in pixel coordinates, $\vec{\delta x} = (\delta x, \delta y)^\top$ represents motion at image position $\vec{x}$ over the time δt and $I(\vec{x}, t)$ represents the image brightness at position $\vec{x}$ and time t. The Taylor series expansion of the right hand side of (2.11) results in,

$$I(\vec{x}, t) = I(\vec{x}, t) + \nabla^\top I \vec{\delta x} + \delta t I_t + O^2 \tag{2.12}$$

where, $\nabla^\top I = (I_x, I_y)$ and I_t are the first order partial derivatives of the image sequence and O^2 denotes the second and higher order terms. Subtracting $I(\vec{x}, t)$ from both sides of (2.12), assuming the second and higher order terms to be negligible and dividing through by δt results in the optical flow constraint equation,

$$\nabla^\top I \vec{u} + I_t = 0 \tag{2.13}$$

where $\vec{u} = (u, v)^\top$ is the apparent velocity.

In accordance with our analysis of idealized gesture movements (2.4) we model

the optical flow parametrically as an affine tranformation,

$$\begin{aligned} u(x,y) &= a_0 + a_1 x + a_2 y \\ v(x,y) &= a_3 + a_4 x + a_5 y. \end{aligned} \tag{2.14}$$

Substituting (2.14) into (2.13) yields the affine constraint,

$$I_x(a_0 + a_1 x + a_2 y) + I_y(a_3 + a_4 x + a_5 y) + I_t = 0 \tag{2.15}$$

Given one image point, the affine flow constraint (2.15) is incapable of fully capturing the motion since the linear constraint involves six unknowns; six independent constraints are required to fully solve for the motion, yet only one constraint equation is given at a point. This problem is commonly referred to as the "aperture problem". Assuming a single motion is present within a local region, a solution is commonly found by taking the least squares solution of the affine flow constraints of each point (at least six points). This approach can be thought of as finding the intersection of the hyper-planes defined by each point constraint. With the presence of Gaussian noise the least squares solution is guaranteed to return the best solution (in the least squares sense), given a largely overdetermined system of equations and sufficient image texture within the region, so that spatiotemporal gradients are well defined. When outlying motions are present the least squares solution performs poorly since the contribution by outliers is not bounded.

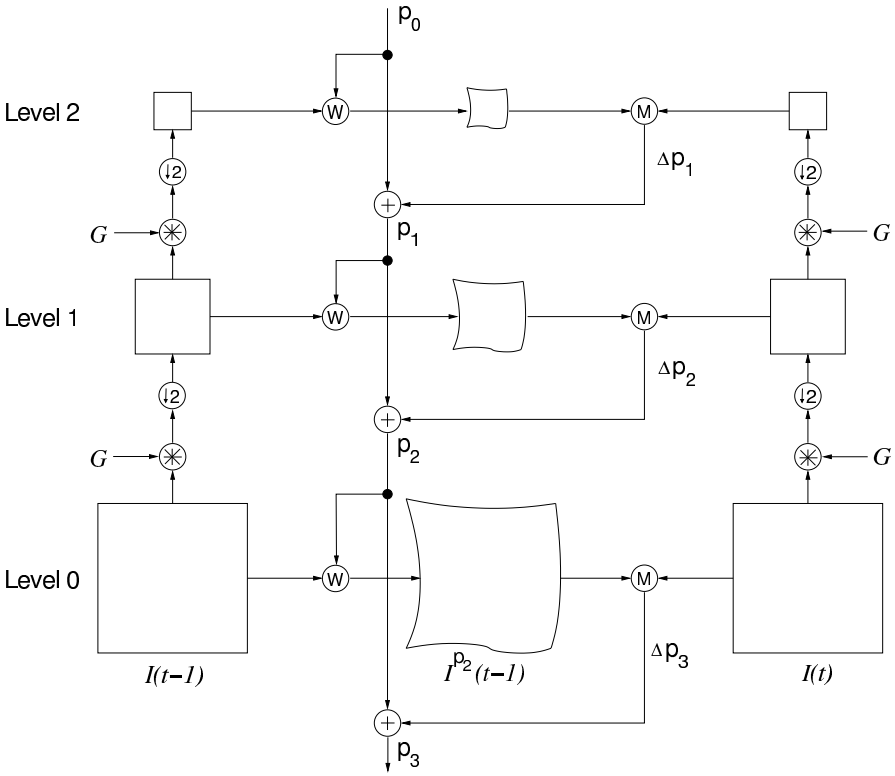


Figure 2.6: Diagram of the hierarchical motion estimation framework. Given two temporally ordered images, the first step consists of constructing Gaussian pyramids ([17, 40] and Appendix B) of each image to level 2, denoted by $I(t-1)$ and $I(t)$ respectively, this is accomplished by a series of convolutions (asterisk symbol) by the kernel G and downsampling by 2 (represented by the symbol with an arrow point down and the number 2). Starting from level 2, each level $i \in \{0, 1, 2\}$ of pyramid $I(t-1)$ is warped (represented by symbol W) by the previous estimate of the affine motion (warp) p_{i-1} plus the residual motion estimate $\triangle p_i$. This is followed by the estimation (represented by M) of the residual motion $\triangle p_{i+1}$ between the warped image and level i of $I(t)$. This diagram is adapted from [9].

To affect the recovery of the affine parameters we make use of a robust estimator embedded within a hierarchical framework [11]. For an illustrative overview of the robust hierarchical motion estimator framework see Fig. 2.6. Benefits afforded by the hierarchical portion of the estimator are, greater capture range (handle motion larger than one pixel), avoids local minima, and computational efficiency.

We use skin colour to restrict the region of support to image data that arises from the hand. A binary skin segmentation map of $I(t-1)$ at level 1 is instantiated and a Gaussian pyramid of the binary map is created up to the same level of the image pyramid. During the motion estimation stage only pixels whose value exceeds a predefined threshold at the corresponding level and position of the binary map pyramid are used in the motion estimation step.

For further robustness, we make use of an M-estimator [38] to allow for operation in the presence of non-Gaussian distributed outlying data in the form of non-hand pixels due skin-color oversegmentation, pixels that grossly violate the surface planarity approximation as well as points that violate brightness constancy. The particular error norm we choose is the Geman-McClure, defined as,

$$\rho(\xi, \sigma) = \frac{\xi^2}{\sigma^2 + \xi^2}$$

with ξ the quantity whose magnitude is to be minimized (in our case departure from brightness constancy, (2.11)) and σ the scale parameter that determines the

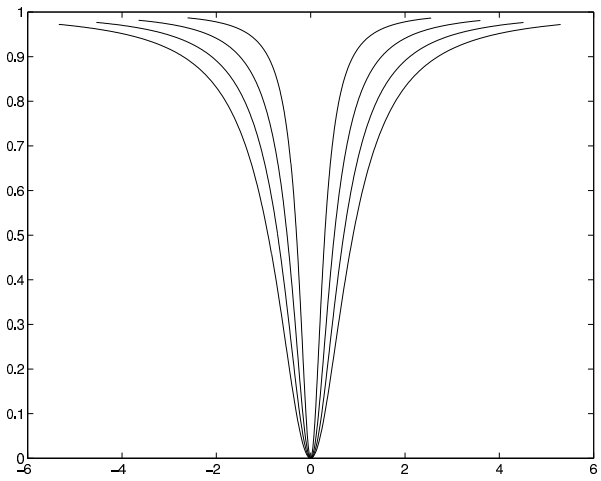


Figure 2.7: Depicted are the family of plots of the Geman-McClure error function with several different instantiations of σ, where $\sigma \in \{0.3, 0.5, 0.7, 0.9\}$.

extent to which the effects of outliers are diminished, see Fig. 2.7.

The minimization problem is formulated as follows:

$$\min_{\{a_0,\ldots,a_5\}} \sum_{\mathbf{x}\epsilon H} \rho(\nabla^\top I\vec{u}(\mathbf{x}) + I_t, \sigma) \tag{2.16}$$

where H represents the set of skin coloured image points, ρ represents the error norm function and σ is a scale parameter defining the point in which the effect of outliers on the estimation begins to diminish. The minimization of (2.16) requires an iterative approach. A standard gradient descent approach was found sufficient for our purposes.

The motion estimator is applied to adjacent frames across an input image sequence. As an initial seed, the hand region in the first frame of the sequence is coarsely outlined manually to define a window for analysis; affine parameters

are initialized identically to zero. Upon recovering the motion between a pair of frames, the analysis window is moved based on the affine parameters found, 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 $(a_0, ..., a_5)$ over the length of the sequence, are realized.

2.7 Kinematic features

Given the analysis done in Section 2.4 we rewrite the affine parameters realized from the affine tracker (described in Section 2.6) in terms of kinematic quantities corresponding to horizontal and vertical translation, divergence, curl and deformation. In particular, from the coefficients in the affine transformation (2.14) we calculate the following time series,

$$\begin{aligned} hor(t) &= a_0(t) \\ ver(t) &= a_3(t) \\ div(t) &= a_1(t) + a_5(t) \\ curl(t) &= -a_2(t) + a_4(t) \\ def(t) &= \sqrt{(a_1(t) - a_5(t))^2 + (a_2(t) + a_4(t))^2} \end{aligned} \tag{2.17}$$

The resulting time series tend to be quite noisy. Since we are interested in the general trends of the time series in the subsequent qualitative analysis step, a

median filter is applied across t time steps followed by a 5-tap binomial filter (i.e. low-pass filter, $\frac{1}{16}[1\ 4\ 6\ 4\ 1]$) is applied.

Each of the kinematic time series (2.17) has an associated unit of measurement (e.g. horizontal and vertical motion are in pixel units) that may differ amongst each other. In order to facilitate comparisons across the time series for the purposes of recognition, a rescaling of responses is appropriate. Here, we make use of min-max rescaling [32], defined as,

$$\hat{z} = \left(\frac{z - min_1}{max_1 - min_1} \right) \times (max_2 - min_2) + min_2 \tag{2.18}$$

with min_1 and max_1 the minimum and maximum values respectively in the input data z, while min_2 and max_2 specifying the desired range of the rescaled data taken over the entire population sample. For scaling ranges, we select $[-1, 1]$ for the elements of (2.17) 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 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. 2.2: 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 ways.

First, to distinguish constant sign movements, we introduce the *summed response*, SR_i,

$$SR_i = \sum_{j=1}^{T} p_{i,j} \tag{2.19}$$

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 $p_{i,j}$ represents the value of time series i at time j. 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 introduce the *summed absolute response*, SAR_i,

$$\begin{aligned} SAR_i &= \sum_{j=1}^{T} |\overline{p_{i,j}}| \\ \overline{p_{i,j}} &= p_{i,j} - mean_i \end{aligned} \tag{2.20}$$

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 (2.18), 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 experimental 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,

$$\begin{aligned} SRP_i &= SR_i/(\sum_k |SR_k|) \\ SARP_i &= SAR_i/(\sum_k SAR_k) \end{aligned} \tag{2.21}$$

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 (2.18) and the conversion to proportions via (2.21) accomplish different goals, both of which are necessary: The former brings all the kinematic variables into generally comparable units and 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.

2.8 Prototype gesture signatures

Given our kinematic feature set derived in the previous section, each of the primitive movements for ASL, shown in Fig 2.2. has a distinctive idealized signature based on (separate) consideration of the SRP_i and $SARP_i$ values, see Table 2.3. These signatures exactly parallel the idealized gesture executions derived in Section 2.4 and summarized in Tables 2.1 and 2.2.

	SRP									**SARP**				
	up	*down*	*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

Table 2.3: 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 (2.21).

Distinctive signatures for the constant sign movements (i.e., movements a-i in Fig. 2.2) are defined with reference to the SRP_i values.

upward/downward 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 (2.21), leading to a signature of $|SRP_{v\hat{e}r}| = 1$ while $|SRP_i| = 0, i \neq v\hat{e}r$, where $|\cdot|$ denotes the absolute value

operator. In order to distiguish between upward and downward movements, the sign of $SRP_{\hat{ver}}$ is taken into account, positive sign for upward and negative for downward.

rightward/leftward 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 leftward and rightward movements, respectively.

toward/away signer 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 away, while negative sign indicates toward.

supinate/pronate 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, respectively.

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. (Note that an additional feature that has not been leveraged is the orientation of the deformation.)

For periodic movements (i.e., movements j-n in Fig. 2.2) 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.

up and down directly maps to $ver(t)$, resulting in a value of $SARP_{\hat{ver}}$ equal to 1 with other summed absolute response proportions zero.

side to side directly maps to $hor(t)$, resulting in a value of $SARP_{\hat{hor}}$ equal to 1 while other values are zero.

to and fro directly maps to $div(t)$, resulting in a value of $SARP_{\hat{div}}$ equal to 1 while other values are zero.

twist wrist directly maps to $curl(t)$, resulting in a value of $SARP_{\hat{curl}}$ equal to 1 with other values zero.

circular 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. 3.2. Across a complete gesture the two summed absolute responses are equal. The overall signature is thus $SARP_{\hat{hor}} = SARP_{\hat{ver}} = 0.5$, with all other values zero.

For classification, comparing the input stream to the prototype signatures 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 consideration of the relative difference between the SR and SAR measures. By construction, only a subset of the SR_i or SAR_i measures will have a significant magnitude at any given time. The determination between a constant sign or periodic movement is done by comparing the magnitudes of the vectors comprising the SR elements and the SAR elements. The vector with the largest L^2-norm determines the type of movement: If the norm of the SR elements is greater than that of the SAR elements, then the movement is constant sign, otherwise the movement is periodic.

Upon determining the type of movement the distance (Euclidean) is computed only amongst the movement signatures of the particular type found. The movement whose signature returns the smallest distance to the input is returned as the classification of the movement. Finally, for movements classified by distance as nod, we explicitly check to make sure $|SRP_{\hat{div}}| \approx SRP_{\hat{def}}$, if not we take the next closest movement. Similarly, for circular we enforce that $SARP_{\hat{hor}} \approx SARP_{\hat{ver}}$. These explicit checks arising from our idealized analysis serve to reject misclassifications when noise happens to artificially push estimated feature value patterns toward the nod and circular signatures.

2.9 Recapitulation

This chapter has outlined our theoretical and algorithmic approach to gesture recognition from a single view (i.e., single video camera). The approach makes

use of the linguistic analysis of manual languages to represent complex gestures in terms of a finite set of primitive components corresponding to hand movement, location and shape. Given a temporal sequence of images that depict a single gesture, our algorithm extracts a set of kinematic features that define distinctive signatures for the primitive movements of ASL. These signatures are used in a simple nearest neighbour classifier that identifies the phonemic movement, irrespective of hand location and shape.

Chapter 3

Experimental Evaluation

3.1 Frontoparallel experiment

3.1.1 Experimental design

In order 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 and to do so irrespective of the volunteer, hand location and hand shape of the complete gesture while the volunteer was seated in a frontoparallel attitude with respect to the camera.

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

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

head, torso and upper arm (left side), see Fig. 2.2. 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 right 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. 2.2. 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 non-native ASL signers volunteers each executed all 148 movements while their actions were recorded with a video camera to yield an experimental test set of 3 (volunteers) $\times$ 148 (phonemic combinations) = 444. In addition, 12 non-native ASL signer volunteers executed an approximate equal subset of the subsampled 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 ges-

tures from each volunteer. It should be noted that the volunteers were fully aware of the camera and their expected position with respect to it (i.e. frontoparallel), 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 30 frames per second. Volunteers were seated approximately 1.45 meters away and instructed to maintain a frontoparallel attitude with respect to the camera during the session. See Fig. 3.1 for a view of our experimental setup.

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 (note that the initial gesture sequence was subsampled temporally by two). Prior to conducting the gesture each volunteer was verbally instructed about 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. Following capture, initial regions of interest for the affine mo-

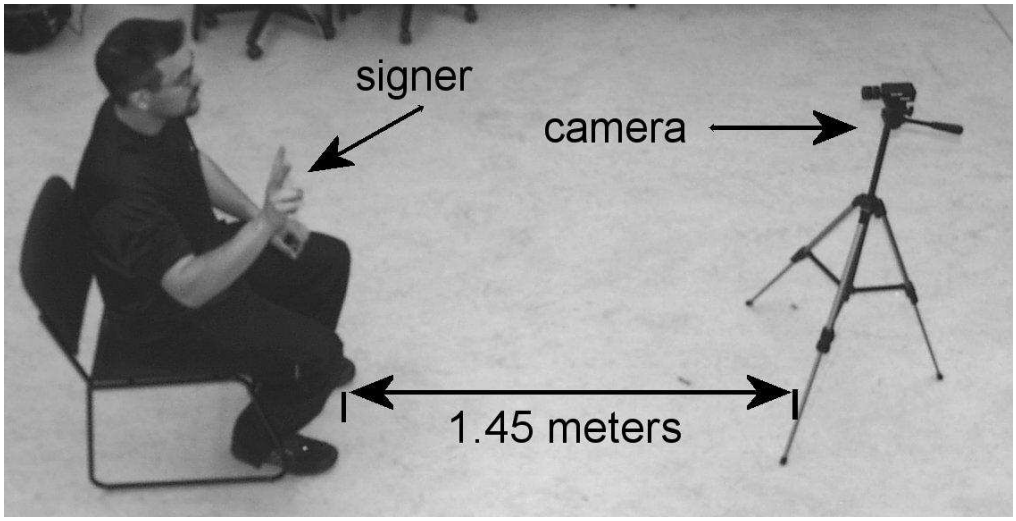
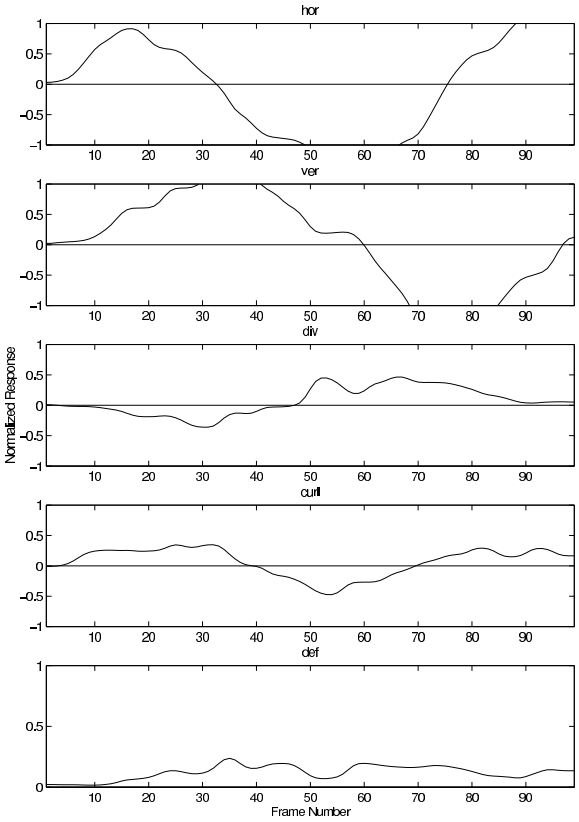


Figure 3.1: Depicted above is a side view of our experimental setup.

tion estimator were manually selected for the first frame of each sequence to seed the automated processing. See Fig. 3.2 for an example sequence; for example sequences of all the movements see Appendix C.

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



(e)

Figure 3.2: Circular movement example. (a)-(d) Four frames of a circular movement image sequence (e) Plots of the normalized kinematic time series spanning the length of the gesture. The frame numbers marked on the graphs correspond to the frame numbers of the image sequence.

3.1.2 Results

The results of the recognition process are shown in Tables 3.1 and 3.2. Overall 97.13% of the 592 test cases were properly identified. Of the failures, three cases arose from failure of the affine motion estimator to correctly track the gesture (quantified by the bounding box about the hand region, as transformed by the affine transformation having more that 95% of its region not occupied by the hand). We deal with these as "failure to acquire"; they were not processed further. Of the cases where the motion estimator correctly tracked the gesture throughout the sequence, the initial classification between constant sign and periodic movements performed without error, and at the subsequent phonemic classification step 97.62% were identified correctly. Additionally, it was found that for 99.49% of the sequences the correct classification was within the top two candidate movements (see Table 3.1 for the classification results 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 averaged 8 frames/second across all gestures sequences; the time consumed by all other components was negligible.

Movement	**Top 1**	**Top 2**
up	100	100
down	100	100
up and down	100	100
rightward	100	100
leftward	97	100
side to side	100	100
toward signer	96	100
away signer	98	100
to and fro	92	100
supinate	97	100
pronate	100	100
twist wrist	100	100
nod	84	91
circular	100	100

Table 3.1: The table depicts the percentage of image sequences of a particular movement that were correctly classified by the first candidate (first column) and the the percentage of sequences that were correctly classified by either of the top two candidates (second column).

	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	0	0	0	0	0	0	0	0	0	0	0	0	0
down	0	100	0	0	0	0	0	0	0	0	0	0	0	0
up and down	0	0	100	0	0	0	0	0	0	0	0	0	0	0
rightward	0	0	0	100	0	0	0	0	0	0	0	0	0	0
leftward	0	0	0	0	97	0	0	0	0	0	0	0	3	0
side to side	0	0	0	0	0	100	0	0	0	0	0	0	0	0
toward signer	0	0	0	0	0	0	96	0	0	0	0	0	4	0
away signer	0	2	0	0	0	0	0	98	0	0	0	0	0	0
to and fro	0	0	0	0	0	0	0	0	92	0	0	0	0	8
supinate	0	0	0	0	0	0	0	0	0	97	0	0	3	0
pronate	0	0	0	0	0	0	0	0	0	0	100	0	0	0
twist wrist	0	0	0	0	0	0	0	0	0	0	0	100	0	0
nod	0	6	0	0	0	0	3	0	6	0	0	0	84	0
circular	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Table 3.2: Gesture movement recognition results. The axes of the table represent the actual input gesture (vertical) versus the classification result. Each cell (i,j) in the table holds the percentage of test cases that were actually i but classified as j. The diagonal (i,j) represents the count of the correctly classified gestures.

3.1.3 Discussion

In terms of tracking, two types of failures may occur, gross errors in the frame-to-frame motion estimate and the drifting of the tracked region away from the hand region due to accumulated tracking errors. The three failed tracking cases encountered were caused by gross colour under-segmentation (region of motion support rendered too small for motion analysis) in a frame, over-segmentation (region of motion support includes substantial non-hand regions that the robust estimator cannot subsequently handle) in a frame or fast movements of the hand that lead to frame-to-frame displacement beyond the capture range of our affine motion estimator. Note that colour segmentation errors substantial enough to cause such breakdowns are rare in our experiments. We did not observe any significant drift, this is due to our use of skin colour to define the region of support and a robust motion estimator to further reject gross outliers, including outliers that skin segmentation admitted.

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 volunteers, 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.

Overall, the results demonstrate the ability of our algorithm to recognize

correctly the 14 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.

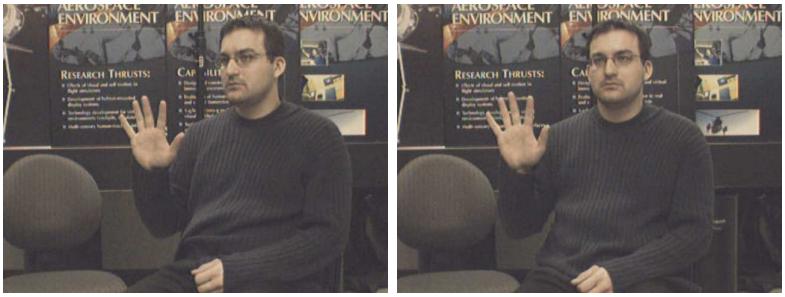
3.2 Attitude experiment

3.2.1 Experimental design

The purpose of this experiment was to test the sensitivity of our phoneme classifier to non-frontoparallel attitudes of the signer with respect to the camera. In this experiment 3 volunteers each executed the 14 movements at attitude positions ± 15 and ± 30 degrees; the angle is measured with respect to the frontoparallel position which is assumed to be at 0 degrees (see Fig. 3.3). Only the torso location and B5 (open hand) hand shape were used in this experiment. In total this experiment consisted of 14 (movements) $\times$ 4 (attitudes) $\times$ 1 (location) $\times$ 1 (hand shape) $\times$ 3 (volunteers) = 168 test sequences.

As with the previous experiment standard indoor, overhead fluorescent lighting, was used with a somewhat cluttered background. Each gesture sequence again was captured at a resolution of 640×480 at 30 frames per second. Volunteers were seated approximately 1.45 meters away from the camera. Representative views of a volunteer at the considered attitudes are shown in Fig. 3.3. Following capture, we manually segment the hand region within the first frame,

to seed the ensuing tracking and classification processes.



(a) -30° (b) -15°

(c) 0° (frontoparallel view)

(d) 15° (e) 30°

Figure 3.3: Camera view of a volunteer at different attitudes.

3.2.2 Results

In terms of tracking three failed cases occurred. All three occurred during the twist wrist movement at 30°. The cause of the failures in all three cases was the hand reaching a parallel configuration with respect to the imaging rays. This resulted in the tracked surface of the hand being lost. The results of the classification at each of the attitudes (less the failed tracked cases) is summarized in Tables 3.3 through 3.7. Inspection of the results reveals that the erroneously classified gestures are for the most part isolated to those gestures containing a purely horizontal (translation parallel to imaging plane) or divergence (translation along the camera axis) component. This is to be expected since the signatures were derived based on the assumption that the signer was frontoparallel with respect to the camera. As the signer deviates from the frontoparallel pose, the gesture containing a purely horizontal or divergence component will consist of both.

Movement	Attitude							
	15°		−15°		30°		−30°	
	Top 1	Top 2	Top 1	Top 2	Top 1	Top 2	Top 1	Top 2
up	3	3	3	3	3	3	3	3
down	2	3	3	3	3	3	3	3
up down	3	3	3	3	3	3	3	3
righward	3	3	3	3	3	3	3	3
leftward	0	2	3	3	0	0	3	3
side to side	3	3	3	3	3	3	1	3
toward signer	3	3	1	2	1	3	0	0
away signer	1	2	1	3	0	2	0	1
to and fro	1	3	0	3	0	3	0	2
supinate	3	3	3	3	3	3	3	3
pronate	3	3	3	3	3	3	3	3
twist wrist	3	3	3	3	*	*	3	3
nod	3	3	3	3	3	3	3	3
circular	3	3	3	3	2	2	3	3

Table 3.3: The table depicts the number of image sequences of a particular movement that were correctly classified by the first candidate (i.e. "Top 1") and the the number of sequences that were correctly classified by either of the top two candidates (i.e. "Top 2") at attitudes 15°, −15°, 30° and −30°. In total there were three image sequences of each gesture for each of the attitudes used in this experiment. The asterisk denotes that zero cases were considered for classification.

	up	*down*	*up and down*	*rightward*	*leftward*	*side to side*	*toward signer*	*away signer*	*to and fro*	*supinate*	*pronate*	*twist wrist*	*nod*	*circular*
up	3	0	0	0	0	0	0	0	0	0	0	0	0	0
down	0	2	0	0	0	0	0	0	0	0	0	0	1	0
up and down	0	0	3	0	0	0	0	0	0	0	0	0	0	0
rightward	0	0	0	3	0	0	0	0	0	0	0	0	0	0
leftward	0	0	0	0	0	0	0	0	0	0	1	0	2	0
side to side	0	0	0	0	0	3	0	0	0	0	0	0	0	0
toward signer	0	0	0	0	0	0	3	0	0	0	0	0	0	0
away signer	0	0	0	0	2	0	0	1	0	0	0	0	0	0
to and fro	0	0	0	0	0	2	0	0	1	0	0	0	0	0
supinate	0	0	0	0	0	0	0	0	0	3	0	0	0	0
pronate	0	0	0	0	0	0	0	0	0	0	3	0	0	0
twist wrist	0	0	0	0	0	0	0	0	0	0	0	3	0	0
nod	0	0	0	0	0	0	0	0	0	0	0	0	3	0
circular	0	0	0	0	0	0	0	0	0	0	0	0	0	3

Table 3.4: Gesture movement recognition results at 15°. The axes of the table represent the actual input gesture (vertical) versus the classification result. Each cell (i,j) in the table holds the number of test cases that were actually i but classified as j. The diagonal (i,j) represents the count 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	3	0	0	0	0	0	0	0	0	0	0	0	0	0
down	0	3	0	0	0	0	0	0	0	0	0	0	0	0
up and down	0	0	3	0	0	0	0	0	0	0	0	0	0	0
rightward	0	0	0	3	0	0	0	0	0	0	0	0	0	0
leftward	0	0	0	0	3	0	0	0	0	0	0	0	0	0
side to side	0	0	0	0	0	3	0	0	0	0	0	0	0	0
toward signer	0	0	0	0	1	0	1	0	0	0	0	0	1	0
away signer	0	0	0	2	0	0	0	1	0	0	0	0	0	0
to and fro	0	0	0	0	0	3	0	0	0	0	0	0	0	0
supinate	0	0	0	0	0	0	0	0	0	3	0	0	0	0
pronate	0	0	0	0	0	0	0	0	0	0	3	0	0	0
twist wrist	0	0	0	0	0	0	0	0	0	0	0	3	0	0
nod	0	0	0	0	0	0	0	0	0	0	0	0	3	0
circular	0	0	0	0	0	0	0	0	0	0	0	0	0	3

Table 3.5: Gesture movement recognition results at $-15°$. The axes of the table represent the actual input gesture (vertical) versus the classification result. Each cell (i,j) in the table holds the number of test cases that were actually i but classified as j. The diagonal (i,j) represents the count 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	3	0	0	0	0	0	0	0	0	0	0	0	0	0
down	0	3	0	0	0	0	0	0	0	0	0	0	0	0
up and down	0	0	3	0	0	0	0	0	0	0	0	0	0	0
rightward	0	0	0	3	0	0	0	0	0	0	0	0	0	0
leftward	0	0	0	0	0	0	0	0	0	0	1	0	2	0
side to side	0	0	0	0	0	3	0	0	0	0	0	0	0	0
toward signer	0	0	0	2	0	0	1	0	0	0	0	0	0	0
away signer	0	0	0	0	3	0	0	0	0	0	0	0	0	0
to and fro	0	0	0	0	0	3	0	0	0	0	0	0	0	0
supinate	0	0	0	0	0	0	0	0	0	3	0	0	0	0
pronate	0	0	0	0	0	0	0	0	0	0	3	0	0	0
twist wrist	0	0	0	0	0	0	0	0	0	0	0	0	0	0
nod	0	0	0	0	0	0	0	0	0	0	0	0	3	0
circular	0	0	0	0	0	0	0	0	1	0	0	0	0	2

Table 3.6: Gesture movement recognition results at 30°. The axes of the table represent the actual input gesture (vertical) versus the classification result. Each cell (i,j) in the table holds the number of test cases that were actually i but classified as j. The diagonal (i,j) represents the count 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	3	0	0	0	0	0	0	0	0	0	0	0	0	0
down	0	3	0	0	0	0	0	0	0	0	0	0	0	0
up and down	0	0	3	0	0	0	0	0	0	0	0	0	0	0
rightward	0	0	0	3	0	0	0	0	0	0	0	0	0	0
leftward	0	0	0	0	3	0	0	0	0	0	0	0	0	0
side to side	0	0	0	0	0	1	0	0	2	0	0	0	0	0
toward signer	0	0	0	0	3	0	0	0	0	0	0	0	0	0
away signer	0	0	0	3	0	0	0	0	0	0	0	0	0	0
to and fro	0	0	0	0	0	3	0	0	0	0	0	0	0	0
supinate	0	0	0	0	0	0	0	0	0	3	0	0	0	0
pronate	0	0	0	0	0	0	0	0	0	0	3	0	0	0
twist wrist	0	0	0	0	0	0	0	0	0	0	0	3	0	0
nod	0	0	0	0	0	0	0	0	0	0	0	0	3	0
circular	0	0	0	0	0	0	0	0	0	0	0	0	0	3

Table 3.7: Gesture movement recognition results at $-30°$. The axes of the table represent the actual input gesture (vertical) versus the classification result. Each cell (i,j) in the table holds the number of test cases that were actually i but classified as j. The diagonal (i,j) represents the count of the correctly classified gestures.

3.2.3 Discussion

As noted in the previous section, the classification of gestures containing a purely horizontal or divergence component are affected when the pose of the signer is greater than a $\pm 15^{\circ}$ offset from the frontoparallel pose. To ameliorate this problem future work may include extracting the relative attitude of the signer with respect to the camera and incorporating it in the classifier. Additionally, future work may include varying the pose by smaller increments in order to uncover the breakdown point in classification with respect to the signer's pose.

3.3 Overall Discussion

In this chapter we have outlined our experimental evaluation of a software realization of our algorithm. The goal of our experimental evaluation was to verify that our algorithm reliably recognizes the phonemic movements under consideration irrespective of the signer, hand location, hand shape and deviations from the frontoparallel pose. With a frontoparallel pose (Section 3.1), high recognition accuracy was achieved, this was imperative since the frontoparallel pose is the most natural signing pose. In our second experiment (Section 3.2), deviation from the frontoparallel pose, a subset of the gestures were found to be misclassified in predictable ways, while all others were classified reliably. The predictability of the breakdowns suggest extensions that may ameliorate this problem. Finally, our current experimental evaluation used only non-native ASL signers, given that

one of the primary users or our system are native ASL signers future work will include native signer's in our experimental evaluation. In this direction, a possible candidate video database is [54], which the author's report to be the most extensive video database of ASL database. A key feature of this database is that special attention was given to capturing the phonemic elements of ASL.

Chapter 4

Conclusion

4.1 Summary

In this report we have presented a novel approach to vision-based hand gesture recognition. Our approach can be summarized by the following three steps. First, we appeal to linguistic theory to represent complex gestures in terms of their primitive (phonemic) 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 analytically derive the ideal mappings between each of 14 single handed phonemic movements and a subset of the kinematic parameters describing the apparent motion of the hand. Third, using these ideal mappings 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, beyond a general purpose computer equipped with a single video camera, our approach has the potential for use in a wide range of human computer inter-

faces. 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 simple nearest neighbour recognition system. Empirical evaluation of the algorithm suggests its applicability to the analysis of complex gesture videos. Finally, it should be noted that though the full hierarchy for recognizing continuous ASL has not been addressed in this report, the movement module we presented may be readily employed in simple gesture based interfaces (e.g. camera control).

4.2 Future work

A logical immediate extension of this report is to remove the following assumptions: the initial location of the hand is known and the hand movement has been temporally segmented. Interestingly, it is common in the gesture recognition literature to make special accommodations for initialization and segmentation [58]. To relax these assumptions future work may appeal to spatiotemporal conjunctions of temporal change and skin colour (some preliminary work on this approach is outlined in Appendix B) to automatically isolate the initial hand location and detecting discontinuities in the kinematic feature time series to temporally segment the gestures (e.g. [64]).

Though tracking failure was not found to be unduly problematic under our experimental conditions, it should be noted that realistic hand speeds of those signing in ASL far exceed those used in our experiments. Those larger displacements may be accommodated by increasing the number of levels used in the pyramid up to a point or replacing the gradient-based motion estimator (e.g., consideration of a correlation-based, rather than gradient-based method); however beyond a certain hand speed significant motion blur sets in which in turn diminishes tracking ability in both the gradient and correlation-based methods. Possible solutions to the motion blur problem is the use of a higher frame rate camera to decrease inter-frame motion

The long term goal is to complete the ASL recognition framework depicted in Fig. 2.1. At the phonemic level, the outstanding problems of location and hand shape pose basic vision problems. Additionally, the inclusion of two-handed gestures with possibly occluding (i.e. one hand obscures the other) scenarios remains to be addressed. As we move up the hierarchy the lexical level will require dictionary knowledge and the ability to provide the sentence level with alternative candidates due to misclassifications that may occur at the phonemic level. Finally, the sentence level will require the incorporation of ASL grammar in order to further constrain possible candidate sentences.

Appendix A

Velocity Description

A.1 World velocity description

The mathematical description of a 3D point (world point in camera coordinates) undergoing a rigid transformation about the camera axes follows.

Let ω_x, ω_y and ω_z represent the angle of rotation about the X, Y and Z axes respectively (see Fig. 2.3). An arbitrary rotation $\mathbf{R}$ is represented as:

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos(\omega_x) & -sin(\omega_x) \\ 0 & sin(\omega_x) & cos(\omega_x) \end{pmatrix} \begin{pmatrix} cos(\omega_y) & 0 & sin(\omega_y) \\ 0 & 1 & 0 \\ -sin(\omega_y) & 0 & cos(\omega_y) \end{pmatrix} \begin{pmatrix} cos(\omega_z) & -sin(\omega_z) & 0 \\ sin(\omega_z) & cos(\omega_z) & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{A.1}$$

Assuming infinitesimal rotations, the zeroth order terms of the Taylor series expansion of the trigonometric functions *sin* and *cos* provide the following approximations,

$$cos(\theta) \approx 1, \quad sin(\theta) \approx \theta \tag{A.2}$$

Using the approximations in (A.2), $\mathbf{R}$ can be approximated as follows in terms

of angular velocity,

$$\mathbf{R} \approx \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -\omega_x \\ 0 & \omega_x & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \omega_y \\ 0 & 1 & 0 \\ -\omega_y & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -\omega_z & 0 \\ \omega_z & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \approx \begin{pmatrix} 1 & -\omega_z & \omega_y \\ \omega_z & 1 & -\omega_x \\ -\omega_y & \omega_x & 1 \end{pmatrix} \tag{A.3}$$

Let the vector $\vec{T} = (t_x, t_y, t_z)^T$ represent the translational velocity, where the elemental components t_x, t_y and t_z represent the translational velocities in the X, Y and Z directions respectively. The velocity $\vec{V} = (\dot{X}, \dot{Y}, \dot{Z})^T$ of a point in the world $\vec{P} = (X, Y, Z)^T$ with respect to camera coordinates undergoing a rigid transformation is represented as,

$$\vec{V} = (\mathbf{R} - \mathbf{I})\vec{P} + \vec{T} = \vec{T} + \Omega \times \vec{P} = \begin{pmatrix} \omega_y Z - \omega_z Y + t_x \\ \omega_z X - \omega_x Z + t_y \\ \omega_x Y - \omega_y X + t_z \end{pmatrix} \tag{A.4}$$

where $\mathbf{I}$ represents a 3 by 3 identity matrix.

In the case where the rotation of the object is about an arbitrary point in space $\vec{Q} = (q_x, q_y, q_z)^T$ (assuming the coordinate systems are aligned) the transformation is represented by the following,

$$\vec{V} = \mathbf{R}(\vec{P} - \vec{Q}) + \vec{Q} + \vec{T} - \vec{P} = \begin{pmatrix} \omega_y(Z - q_z) - \omega_z(Y - q_y) + t_x \\ \omega_z(X - q_x) - \omega_x(Z - q_z) + t_y \\ \omega_x(Y - q_y) - \omega_y(X - q_x) + t_z \end{pmatrix} \tag{A.5}$$

A.2 Image velocity description I

The mathematical description of the image velocity follows for a rigid transformation about the camera coordinate system as defined in A.1. Assuming a perspective projection onto a plane parallel the X, Y axes and located at Z = 1 (without loss of generality the focal length $f = 1$), the relationship between an image point (x, y) and a scene point (X, Y, Z) is

$$x = \frac{X}{Z}, \quad y = \frac{Y}{Z} \tag{A.6}$$

Differentiating (A.6) with respect to time yields,

$$\begin{aligned} \dot{x} = u &= \frac{\dot{X}}{Z} - \frac{X\dot{Z}}{Z^2} \\ \dot{y} = v &= \frac{\dot{Y}}{Z} - \frac{Y\dot{Z}}{Z^2} \end{aligned} \tag{A.7}$$

Substituting (A.4) and (A.6) into (A.7) results in the following,

$$\begin{aligned} u &= -\omega_x xy + \omega_y(x^2+1) - \omega_z y + \frac{t_x - t_z x}{Z} \\ v &= \omega_y xy - \omega_x(y^2+1) + \omega_z x + \frac{t_y - t_z y}{Z} \end{aligned} \tag{A.8}$$

Assuming that the imaged surface is a plane with surface normal $\vec{n} = (n_x, n_y, n_z)^\top$ and containing the point (X_0, Y_0, Z_0), provides the following constraint,

$$\alpha X + \beta Y + \gamma Z = 1 \tag{A.9}$$

or equivalently as

$$\alpha x + \beta y + \gamma = \frac{1}{Z} \tag{A.10}$$

where

$$\begin{aligned} \alpha &= \frac{n_x}{d} \\ \beta &= \frac{n_y}{d} \\ \gamma &= \frac{n_z}{d} \\ d &= n_x X_0 + n_y Y_0 + n_z Z_0 \end{aligned} \tag{A.11}$$

Substituting the planar constraint (A.10) into (A.8) leads to following formulation for the instantaneous velocities,

$$\begin{aligned} u &= a_0 + a_1 x + a_2 y + a_7 xy + a_6 x^2 \\ v &= a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2 \end{aligned} \tag{A.12}$$

where

$$\begin{aligned} a_0 &= t_x\gamma + \omega_y \\ a_1 &= t_x\alpha - t_z\gamma \\ a_2 &= t_x\beta - \omega_z \\ a_3 &= t_y\gamma - \omega_x \\ a_4 &= t_y\alpha + \omega_z \\ a_5 &= t_y\beta - t_z\gamma \\ a_6 &= \omega_y - t_z\alpha \\ a_7 &= -t_z\beta - \omega_x \end{aligned} \tag{A.13}$$

Through first order in image coordinates, we have the following affine model for the instantaneous velocity,

$$\begin{aligned} u &= a_0 + a_1 x + a_2 y \\ v &= a_3 + a_4 x + a_5 y \end{aligned} \tag{A.14}$$

A.3 Image velocity description II

The mathematical description of the image velocity for a rigid transformation consisting of a rotation about an arbitrary point as defined in A.1 follows.

Assuming a perspective projection onto a plane parallel the X, Y axes and located at Z=1 (without loss of generality the focal length $f = 1$), the relationship between an image point (x, y) and a scene point (X, Y, Z) is given by (A.6) and the velocity of each image point is give by (A.7).

Substituting (A.5) and (A.6) into (A.7) results in the following,

$$\begin{aligned} u &= \omega_y - \omega_z y - \omega_x xy + \omega_y x^2 + \frac{t_x + q_y\omega_z - q_z\omega_y + (q_y\omega_x - t_z - q_x\omega_y)x}{Z} \\ v &= -\omega_x + \omega_z x + \omega_y xy - \omega_x y^2 + \frac{t_y + q_z\omega_x - q_x\omega_z + (q_y\omega_x - t_z - q_x\omega_y)y}{Z} \end{aligned} \tag{A.15}$$

Substituting the planar constraint (A.10) into (A.15) and once again limiting, consideration through first order terms leads to the following affine model for the instantaneous velocities,

$$\begin{aligned} u &= a_0 + a_1 x + a_2 y \\ v &= a_3 + a_4 x + a_5 y \end{aligned} \tag{A.16}$$

where

$$\begin{aligned}
a_0 &= \omega_y + (\omega_z q_y + t_x - \omega_y q_z)\gamma \\
a_1 &= (\omega_z q_y + t_x - \omega_y q_z)\alpha + (\omega_x q_y - \omega_y q_x - t_z)\gamma \\
a_2 &= -\omega_z + (\omega_z q_y + t_x - \omega_y q_z)\beta \\
a_3 &= -\omega_x + (\omega_x q_z + t_y - \omega_z q_x)\gamma \\
a_4 &= \omega_z + (\omega_x q_z + t_y - \omega_z q_x)\alpha \\
a_5 &= (\omega_x q_z + t_y - \omega_z q_x)\beta + (\omega_x q_y - \omega_y q_x - t_z)\gamma
\end{aligned} \tag{A.17}$$

A.4 First-order accuracy

In this section we summarize the results of analytical as well a numerical investigation we conducted to compare the accuracy of the first-order visual motion field description of a planar surface (approximation assumed in this report) to the analytically globally correct second order representation. Furthermore, we omit the effects of lens distortion and assume noiseless data, as our concern is with the residual error between the first and and second-order models.

For this analysis the plane (representing the hand in our model) undergoes each of the following movements (for detailed descriptions of the movements see Section 2.4): upward, away signer, nod and supinate. Similar results hold for the remaining phonemic movements by symmetry. For each of the movements it is assumed that its constituent 3D instantaneous movements are constant throughout the sequence. Additionally, the plane is assumed to be located initially at

a distance of 1.45 meters (taken from our experimental setup) away from the camera.

The perspective projection camera (depicted in Fig. 2.3) is assumed to have a focal length $f = 6$ millimeters, with equal horizontal and vertical scaling factors of $s = 1/(5.6 \times 10^{-3})mm$. Both the focal length and scaling factors are taken from our experimental setup.

The following is the second order description of the motion field (in pixel units) of a planar surface with the inclusion of the focal length f and scale factor s,

$$\begin{aligned} u &= a_0 + a_1 x + a_2 y + a_7 xy + a_6 x^2 \\ v &= a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2 \end{aligned} \tag{A.18}$$

where

$$\begin{aligned} a_0 &= \gamma f_s t_x + f_s \omega_y - \gamma f_s q_z \omega_y + \gamma f_s q_y \omega_z \\ a_1 &= \alpha t_x - \gamma t_z + \gamma q_y \omega_x - \gamma q_x \omega_y - \alpha q_z \omega_y + \alpha q_y \omega_z \\ a_2 &= \beta t_x - \beta q_z \omega_y - \omega_z + \beta q_y \omega_z \\ a_3 &= \gamma f_s t_y - f_s \omega_x + \gamma f_s q_z \omega_x - \gamma f_s q_x \omega_z \\ a_4 &= \alpha t_y + \alpha q_z \omega_x \omega_z + \omega_z - \alpha q_x \omega_z \\ a_5 &= \beta t_y - \gamma t_z + \gamma q_y \omega_x + \beta q_z \omega_x - \gamma q_x \omega_y - \beta q_x \omega_z \\ a_6 &= \frac{-\alpha t_z + \alpha q_y \omega_x + \omega_y - \alpha q_x \omega_y}{f_s} \\ a_7 &= \frac{-\beta t_z - \omega_x + \beta q_y \omega_x - \beta q_x \omega_y}{f_s} \end{aligned} \tag{A.19}$$

where $f_s = f \times s$. For a detailed derivation and definitions of all the terms less the inclusion of the focal length and scaling factors see Appendix A.1, A.2 and A.3.

The first-order model consists of truncating the second-order model up to the first-order terms, thus the associated error $(u_{error}, v_{error})^\top$ between the models is,

$$\begin{aligned} u_{error} &= a_7 xy + a_6 x^2 \\ v_{error} &= a_6 xy + a_7 y^2 \end{aligned} \tag{A.20}$$

From (A.19 and A.20) it can be seen that the error associated with the first-order model (i.e. quadratic terms of second-order model) are independent of t_x, t_y and ω_z, thus the motion of the hand undergoing upward and supinate movements and their symmetric movements are fully characterized by the affine model (i.e. zero associated error). Further, note that for rotation about $\vec{Q} = (0, 0, q_z)^\top$, which occurs for the nod movement, the error is

$$\|(u_{error}, v_{error})\| = \|(\frac{-xy\omega_x + x^2\omega_y}{f_s}, \frac{xy\omega_y - y^2\omega_x}{f_s})\| \tag{A.21}$$

since $\omega_y = 0$ for all our movements the error reduces to the following,

$$\|(u_{error}, v_{error})\| = \|(\frac{-xy\omega_x}{f_s}, \frac{-y^2\omega_x}{f_s})\| \tag{A.22}$$

Taking, $\|\cdot\|$ as the l_2-norm we find,

$$\|(u_{error}, v_{error})\| = \frac{|\omega_x|}{f_s}|y|[x^2 + y^2]^{1/2} \tag{A.23}$$

The error thus depends linearly on the magnitude of the rotational velocity ω_x, inversely on the scaled focal length f_s, modulated by both the magnitude of the distance of the point $(x, y)^\top$ from the image origin (i.e. intersection of image plane and camera axis) and the the magnitude of y.

Similarly, if $(\alpha, \beta) = (n_x/d, n_y/d) = (0,0)$, i.e., the plane is parallel with the imaging plane, then $||(u_{error}, v_{error})|| = 0$ for pure t_z cases (i.e. away signer, towards signer and to-fro). From these analytical developments we anticipate little difficulty in relying on the first-order motion model in our work.

To provide further backing, for out selection of the affine model, we have conducted a numerical simulation. Given an initial $110 \times 180\ mm^2$ region of the planar surface (size of author's hand) centered about the camera axis in the world space, we uniformly sample this region by selecting every fifth point and record the second order components of the visual motion (i.e. affine error as in A.23) of each point as it undergoes a nod movement, since this is the only movement with non-zero analytic case. In terms of 3D instantaneous movements this is accomplished by rotating the plane by the instantaneous rotational velocity,

$$\omega_x = \left(-\frac{\pi}{2} radians\right)\left(\frac{1}{number of frames}\right) = \left(-\frac{\pi}{2}\right)\left(\frac{1}{35}\right) = -\frac{\pi}{70} radians/frame \tag{A.24}$$

(where a frame equals 1/15 of a second, differs by a factor of 2 from video rate capture due to temporal subsampling during preprocessing) about the point $\vec{Q} = (0, 0, q_z)^\top$ where $q_z = 1450$ (distance of plane in millimeters). Note that the number of frames used in our estimate of the instantaneous rotational velocity was estimated from our experimental image sequences. Figure A.1 shows a plot of the maximum error versus time (i.e. frame number). In Fig. A.2 the plots of the magnitude of the velocity of the point with maximum affine error

(over time) is plotted, under both the affine and quadratic models. Note that these maximum errors correspond to the extremal points of the planar patch. In both cases it can be seen that in the worst case (i.e. points furthest from the origin) the contribution of the second-order terms (i.e. affine error) is negligible. Thus leading us to the conclusion that the affine model used in this report is a reasonable approximation for the cases under consideration.

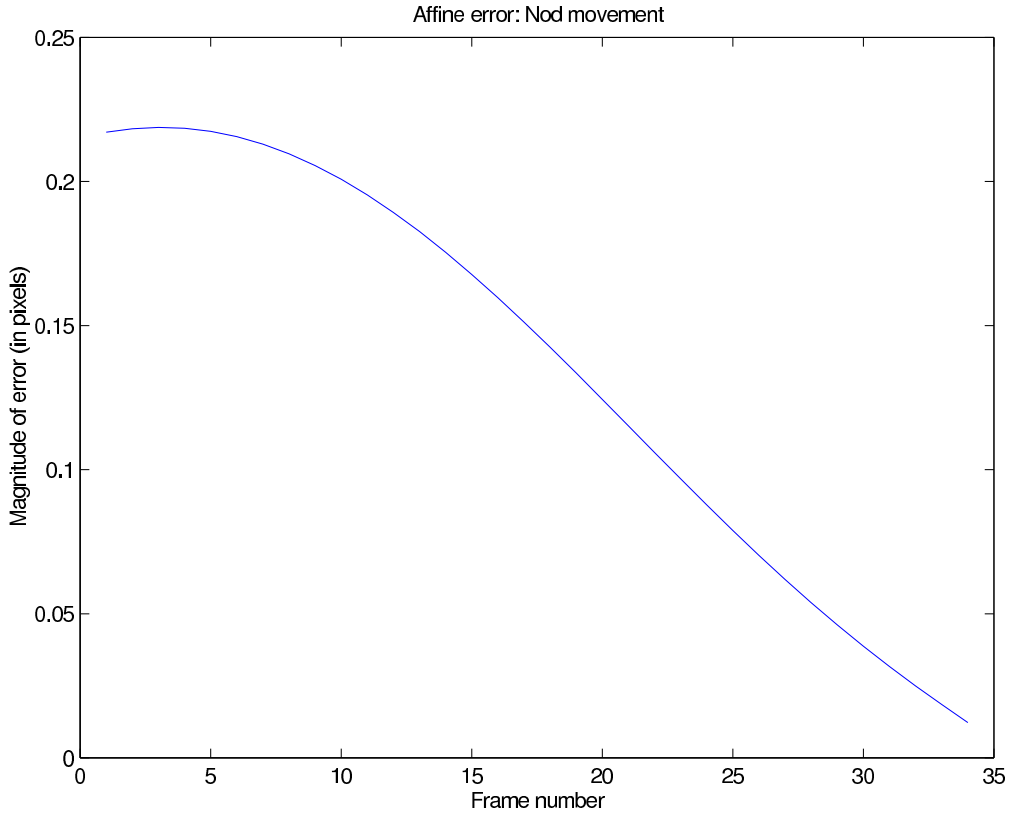


Figure A.1: This plot depicts the magnitude of the maximum affine error of the visual motion field of a 110×180 mm^2 central region of a plane undergoing a nod movement.

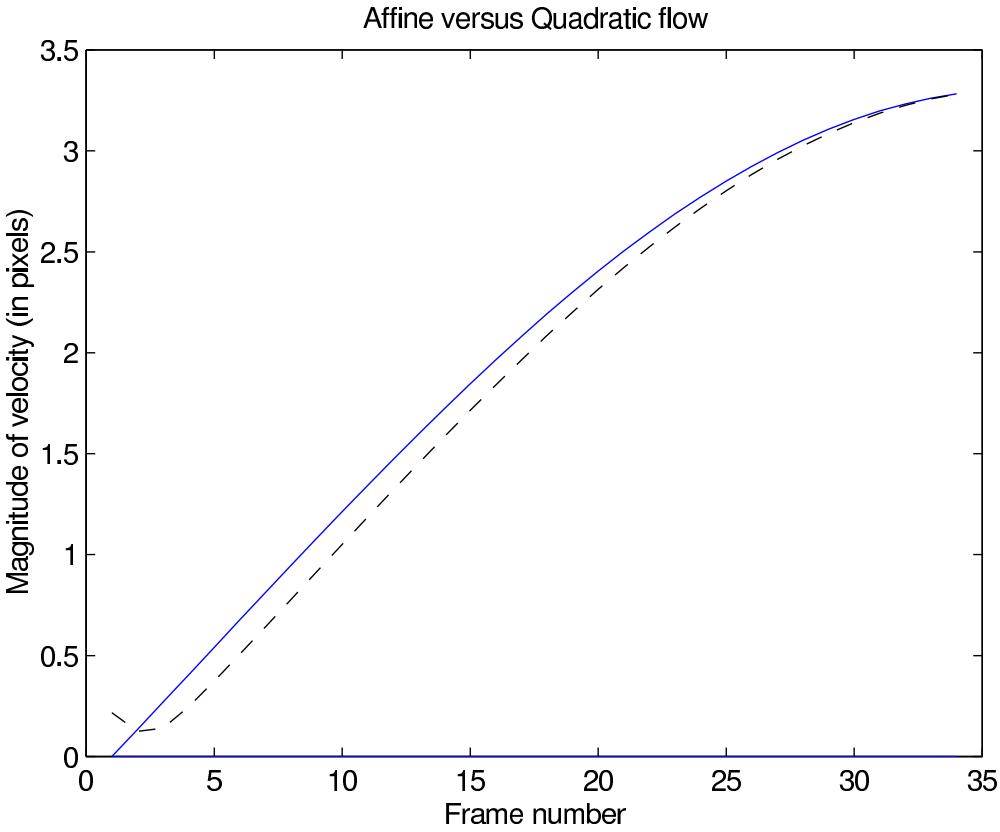


Figure A.2: This plot depicts the magnitude of the velocity of the point exhibiting maximum affine error over time, under the affine (solid plot) and quadratic models (dashed plot).

A.5 Kinematic parameter definitions

Eq. (2.4) can be rewritten in matrix form as an affine transformation $\mathbf{A}$ plus a translation $\vec{t}$, formally,

$$\vec{u} = \mathbf{A}\vec{x} + \vec{t} \tag{A.25}$$

where

$$\mathbf{A} = \begin{pmatrix} a_1 & a_2 \\ a_4 & a_5 \end{pmatrix} \tag{A.26}$$

$\vec{u} = (u, v)^T$ represents the 2D displacement, $\vec{x} = (x, y)^T$ represents the point and $\vec{t} = (a_0, a_3)^T$ represents the x and y translational components respectively.

In understanding the ramifications of the the transformation embodied in $\mathbf{A}$, it is advantageous to recast it in terms of kinematic quantities that capture (infinitesimal) rotation (curl), expansion/contraction (divergence), and shear (deformation) ([4, 44, 51, 80]). Toward that end $\mathbf{A}$ can be rewritten as a sum of a symmetric and antisymmetric matrix as follows,

$$\mathbf{A} = \frac{1}{2}[(\mathbf{A} + \mathbf{A}^T) + (\mathbf{A} - \mathbf{A}^T)] \tag{A.27}$$

The antisymmetic part can be expressed as,

$$\mathbf{A} - \mathbf{A}^T = \begin{pmatrix} 0 & a_2 - a_4 \\ -a_2 + a_4 & 0 \end{pmatrix} = (-a_2 + a_4) \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \tag{A.28}$$

The symmetric part can be further decomposed into two components, the sum of a scalar multiple of the identity matrix and a symmetric matrix

$$\mathbf{A}+\mathbf{A}^T=\begin{pmatrix} 2a_1 & a_2+a_4 \\ a_2+a_4 & 2a_5 \end{pmatrix}=\begin{pmatrix} a_1+a_5 & 0 \\ 0 & a_1+a_5 \end{pmatrix}+\begin{pmatrix} a_1-a_5 & a_2+a_4 \\ a_2+a_4 & -a_1+a_5 \end{pmatrix} \tag{A.29}$$

With the above decompositions, $\mathbf{A}$ can be expressed as,

$$\mathbf{A}=\frac{1}{2}curl\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}+\frac{1}{2}div\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}+\frac{1}{2}def\mathbf{S} \tag{A.30}$$

where

$$\begin{aligned} div &= a_1+a_5 \\ curl &= -a_2+a_4 \\ def &= \sqrt{(a_1-a_5)^2+(a_2+a_4)^2} \end{aligned} \tag{A.31}$$

and $\mathbf{S}$ is a traceless matrix defining the direction of the area preserving deformation. It is interesting to note that the divergence, curl and magnitude of the deformation are invariant to rotations of the image coordinate frame.

Appendix B

Hand Localization

In our current work it has been assumed that the hand region in the image has been manually delineated. Such an assumption will have to be removed in order to achieve the goal of a fully automated hand gesture recognition system. In this section we describe a possible strategy to automatically localize the hand in the first frame of a gesture sequence. Key assumptions leveraged in the proposed approach are, the hand is the dominant moving object in the scene and the hand can be distinguished from most objects in the scene by its colour.

The basic approach we have selected consists of motion detection via localizing frequency tuned change energy within a pyramid framework [3]. Specifically, we isolate the dominant moving region, assumed to be the hand, by detecting significant energy within a specific region in the spatio-temporal frequency domain. As a final step we apply skin detection to refine the segmentation.

The first step consists of localizing the moving region in the temporal frequency domain. This is accomplished by taking a frame to frame difference image D (crude high pass temporal filter) by subtracting the initial frame $I(0)$

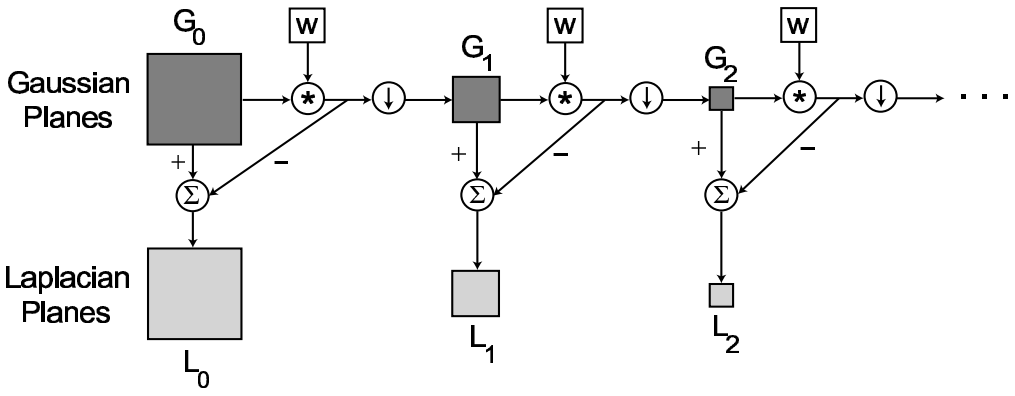


Figure B.1: A summary of constructing Laplacian and Gaussian pyramids is depicted. Let G_0 represent the original image, and G_1 be the result of convolving (denoted by the asterisk symbol) a low-pass filter w (in our case Gaussian filter approximated by a 5×5 binomial filter) with G_0 and downsampling the result by 2 (denoted by the downward pointing arrow). The first level of the Laplacian pyramid L_0 is given by $L_0 = G_0 - G_0 * w$, subsequent levels of the pyramid are given by $L_i = G_i - G_i * w$.

from a subsequent frame $I(t)$, formally, $D = I(t) - I(0)$.

Given the difference image D a Laplacian pyramid is constructed [17] (see Fig. B.1 for details on constructing the Laplacian pyramid); each level of the Laplacian pyramid represents pass bands in the spatial domain spaced at one octave intervals (some overlap does occur). A particular level (L_1) of the Laplacian pyramid is selected for further analysis. The selection of L_1 is dictated by the fact that the object of interest, the human hand, is rather textureless, thus movements over a small number of frames will result in edge structures at the occlusion boundaries of the hand, these edge structures are most evident in the high frequencies; L_0 is another possible candidate but not considered since a significant portion of the image noise is captured at this level.

The next step consists of locally integrating values of the squared result of L_1 (i.e. L_1^2) to form energy measures. This is accomplished through the construction a Gaussian pyramid to a height of $k = 3$ (for details on constructing a Gaussian pyramid see Fig. B.1). We localize the moving region (i.e. the hand region) by first finding the peak energy within G_k. This step is extremely computationally efficient given the reduced resolution. Next we take the weighted average of each image position (i.e. centroid) within a 40 by 40 region of G_{k-1} centred about the peak found at level G_k; the weight is determined by the energy at each point divided by the sum of the energy of all points within the region being examined. The process of finding the centroid is iterated down the pyramid until G_0 is reached. At this point it is assumed that the hand is bounded by the analysis region. The use of a weighted average is necessitated by the fact that as we move down the Gaussian pyramid the outline of the hand becomes the only prominent structure; if the maximum (i.e. peak) response at each level were only considered then localization may only occur about an outline structure rather than consuming the whole hand region.

Finally, we apply skin colour segmentation (see Section 2.5 for details) to refine the localization of the hand.

The general outline of the algorithm is as follows:

1. Form difference image D between initial image and a subsequent image.

2. Decompose the difference image D into a set of spatial frequency bands

through the construction of a Laplacian pyramid.

3. Construct a Gaussian pyramid to level k using L_1^2 as the base level.

4. Recursively localize energy centroid at each level of the Gaussian pyramid starting at level k.

5. Apply skin colour segmentation to refine the localization of the hand.

In Fig. B.2 and Fig. B.3 we present representative successful output of the various stages. In Fig. B.4 we present a failed localization case. The source of the failure is the violation of our assumption that the hand is the dominant moving object in the scene. In fact, in Fig. B.4 it can be seen that both the hand and arm regions contain significant energy, with the arm region containing slightly more energy. Note that in this case the localized region contains very little skin tone. In such cases, the region may be discounted as erroneous and the next largest response in the Gaussian pyramid should be localized.

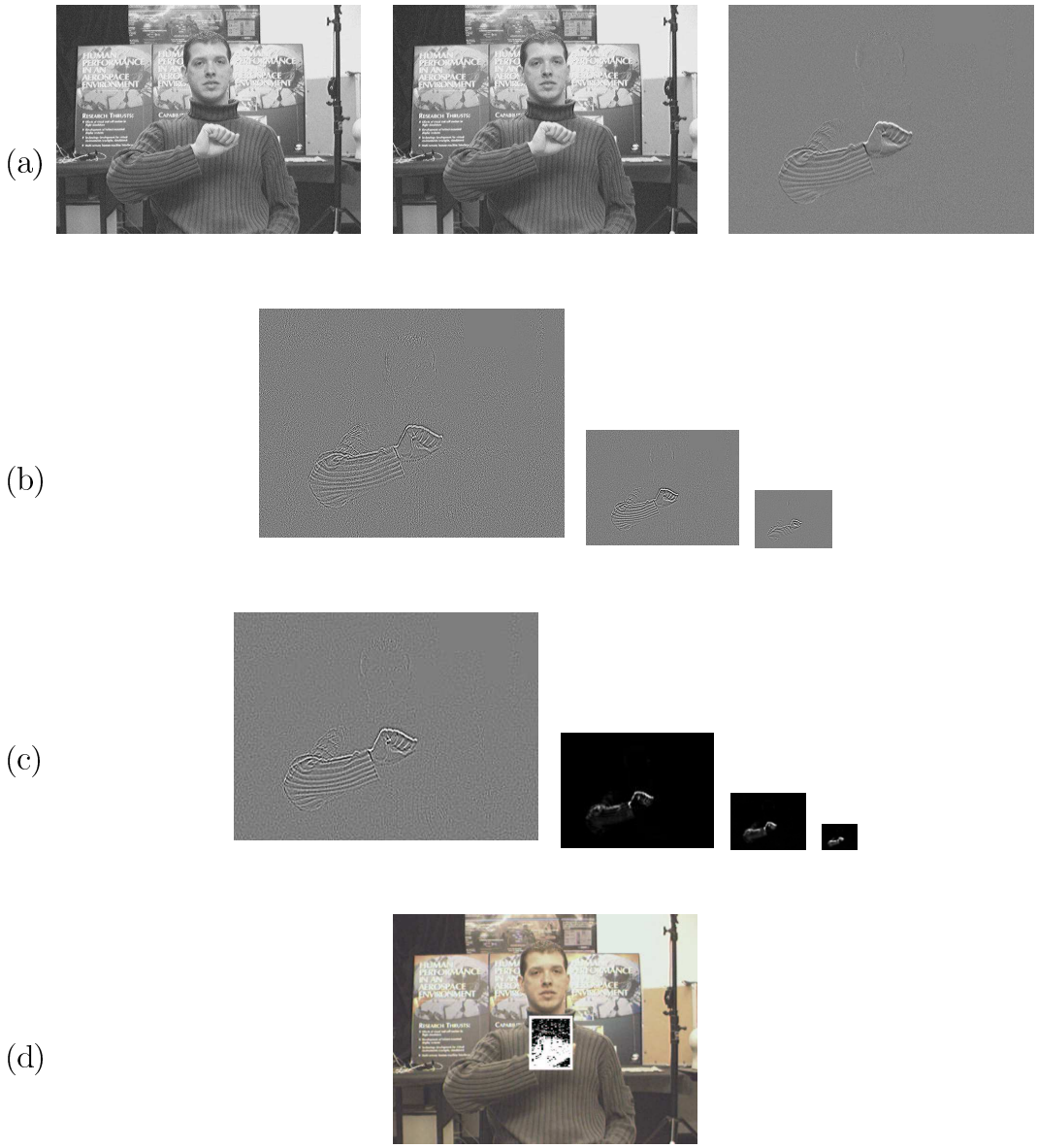


Figure B.2: Localization example 1. Example output of localization steps. Depicted are the outputs of the various stages of the proposed localization scheme for two gesture sequences. (a) difference image (b) Laplacian pyramid decomposition of difference image D (c) Gaussian pyramid with L_1^2 (pointwise squaring of pixels of L_1) as the base level (for clarity (c) is scaled by a factor of 2) (d) $I(0)$ with white box denoting the foveated motion region. Within the delineated region skin segmentation is applied, white denoting a skin pixel and black a non-skin pixel.

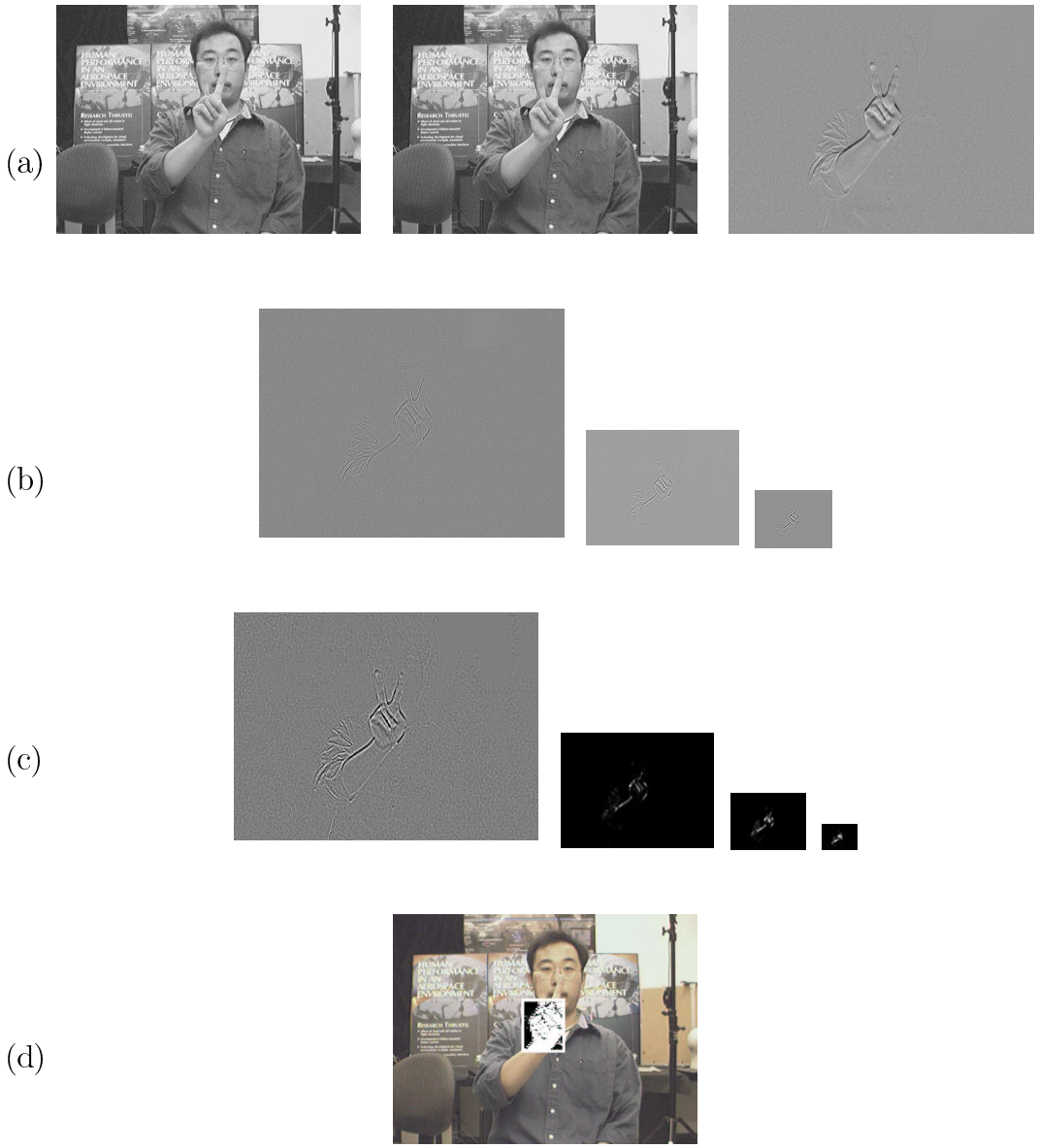


Figure B.3: Localization example 2. Example output of localization steps. Depicted are the outputs of the various stages of the proposed localization scheme for two gesture sequences. (a) difference image (b) Laplacian pyramid decomposition of difference image D (c) Gaussian pyramid with L_1^2 (pointwise squaring of pixels of L_1) as the base level (for clarity (c) is scaled by a factor of 2) (d) $I(0)$ with white box denoting the foveated motion region. Within the delineated region skin segmentation is applied, white denoting a skin pixel and black a non-skin pixel.

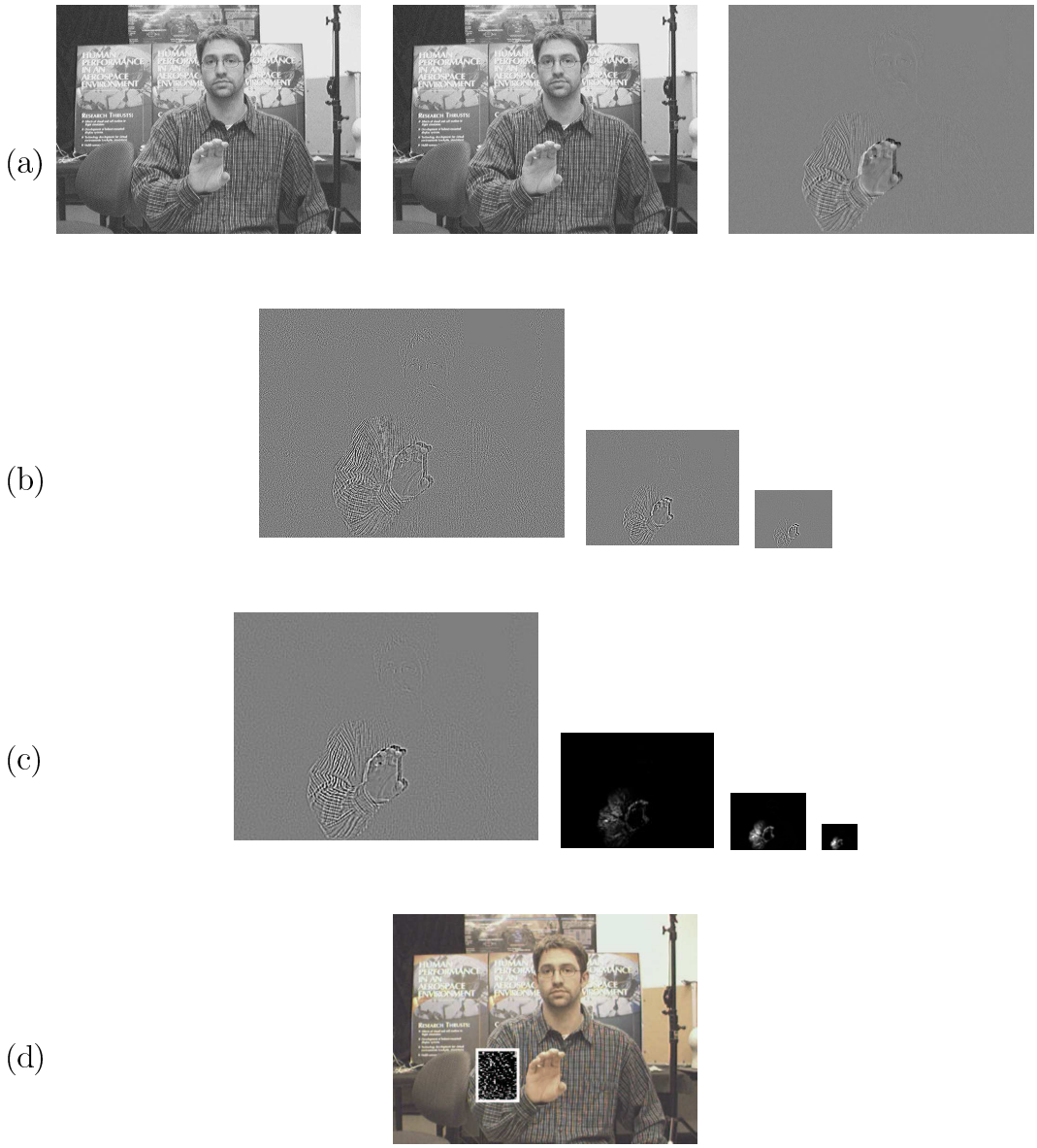


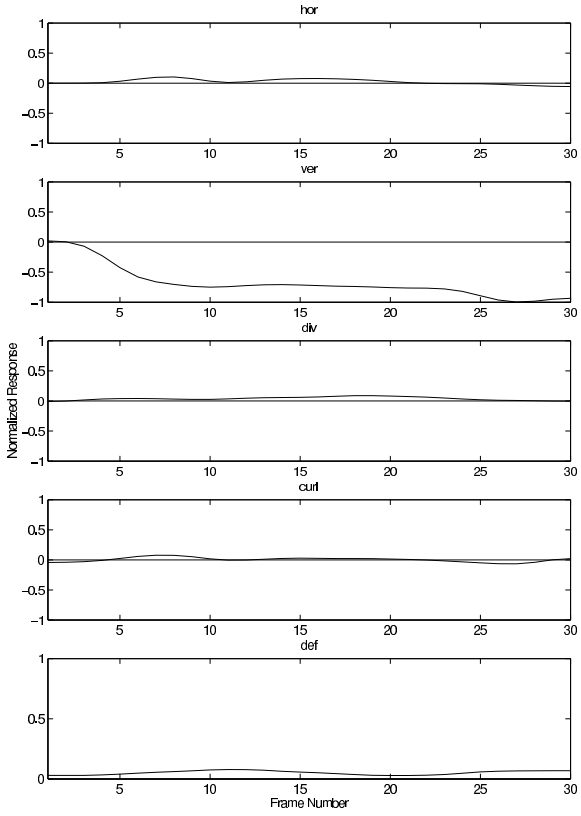
Figure B.4: Failed localization example. Example output of localization steps. Depicted are the outputs of the various stages of the proposed localization scheme for two gesture sequences. (a) difference image (b) Laplacian pyramid decomposition of difference image D (c) Gaussian pyramid with L_1^2 (pointwise squaring of pixels of L_1) as the base level (for clarity (c) is scaled by a factor of 2) (d) $I(0)$ with white box denoting the foveated motion region. Within the delineated region skin segmentation is applied, white denoting a skin pixel and black a non-skin pixel.

Appendix C

Example Sequences

In this section we provide example sequences taken from our experimental data set of each of the 14 movements considered in this report with accompanying plots of each of the kinematic time series. Note that the affine motion estimator that we employ adopts the convention that the vertical axis points downwards; hence normalized *ver* component signs are reversed in the following plots in comparison to the analysis presented elsewhere in this report.

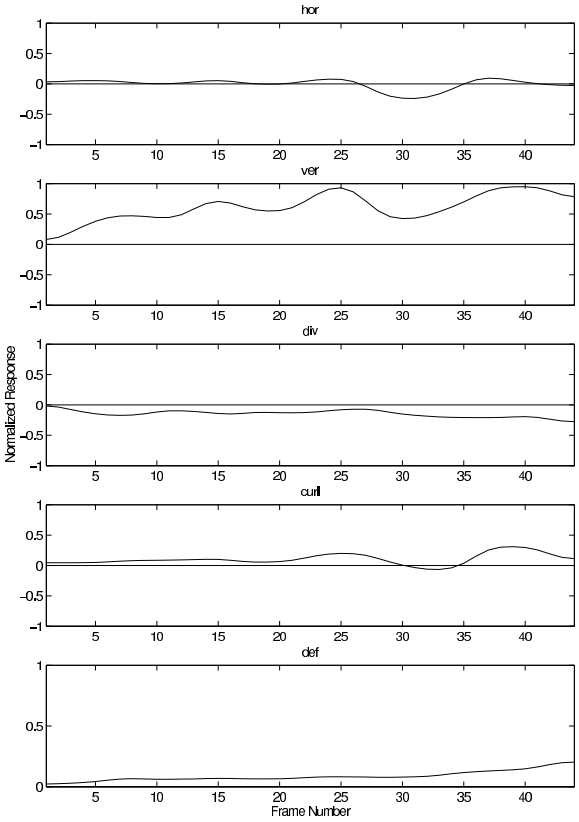
(a) Frame 0 (b) Frame 8 (c) Frame 16 (d) Frame 24



(e)

Figure C.1: Upward movement example. (a)-(d) example frames (e) normalized kinematic time series plots. Note that the vertical axis points downwards in these plots as opposed to upwards as assumed elsewhere in this report. Thus, the normalized *ver* component signs are reversed.

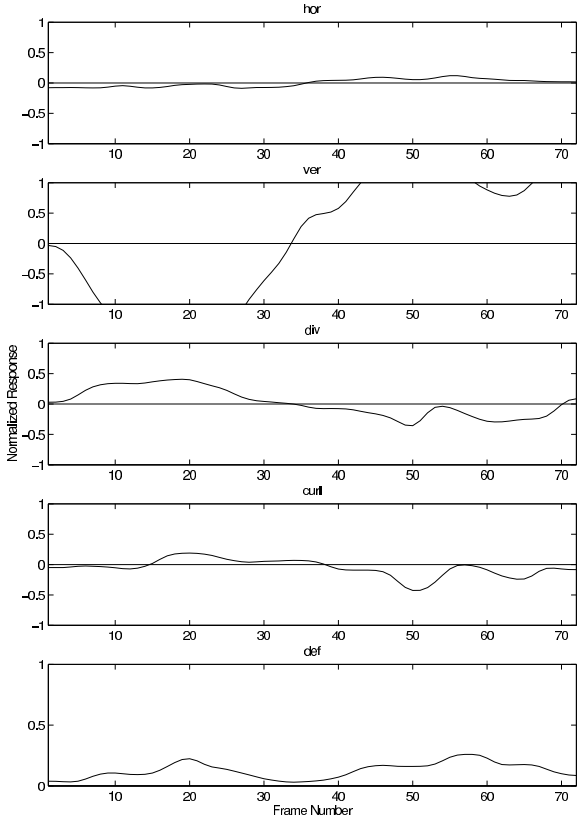
(a) Frame 0 (b) Frame 11 (c) Frame 22 (d) Frame 33



(e)

Figure C.2: Downward movement example. (a)-(d) example frames (e) normalized kinematic time series plots. Note that the vertical axis points downwards in these plots as opposed to upwards as assumed elsewhere in this report. Thus, the normalized *ver* component signs are reversed.

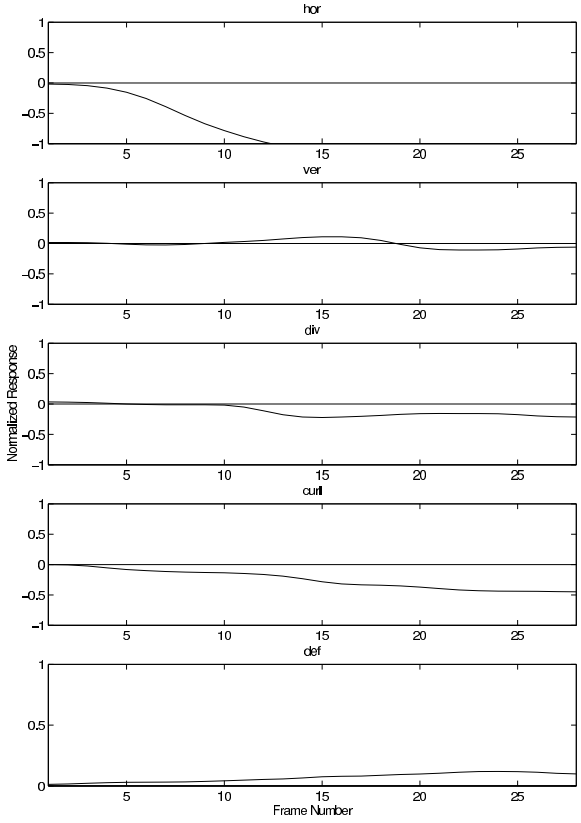
(a) Frame 0 (b) Frame 18 (c) Frame 36 (d) Frame 54



(e)

Figure C.3: Up and down movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

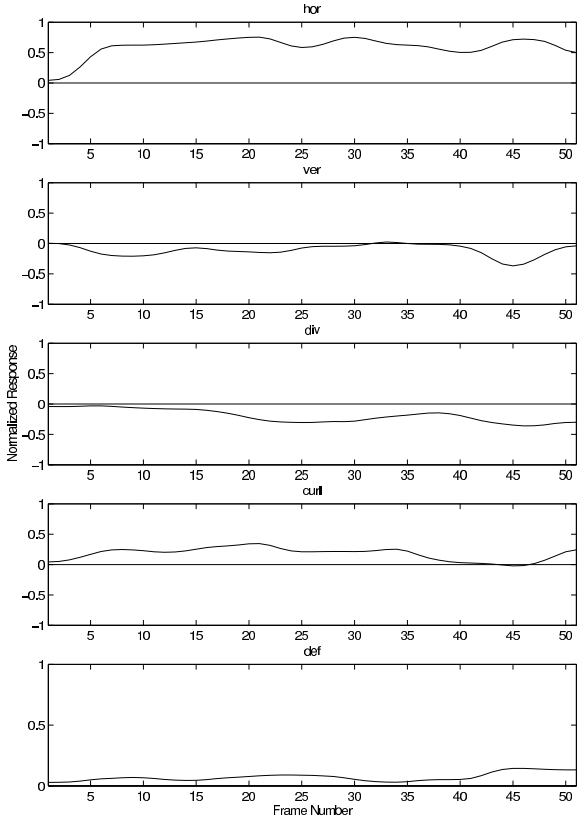
(a) Frame 0 (b) Frame 7 (c) Frame 14 (d) Frame 21



(e)

Figure C.4: Rightward movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

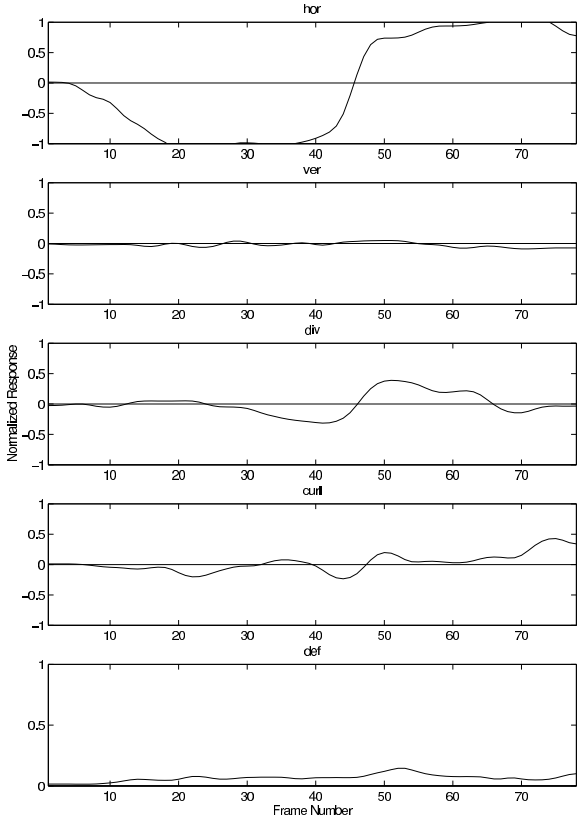
(a) Frame 0 (b) Frame 13 (c) Frame 26 (d) Frame 39



(e)

Figure C.5: Leftward movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

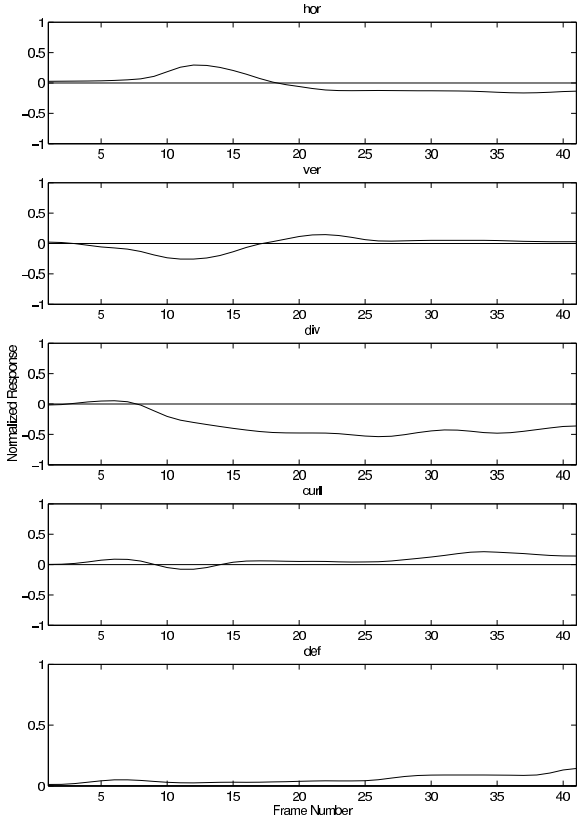
(a) Frame 0 (b) Frame 20 (c) Frame 40 (d) Frame 60



(e)

Figure C.6: Side to side movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

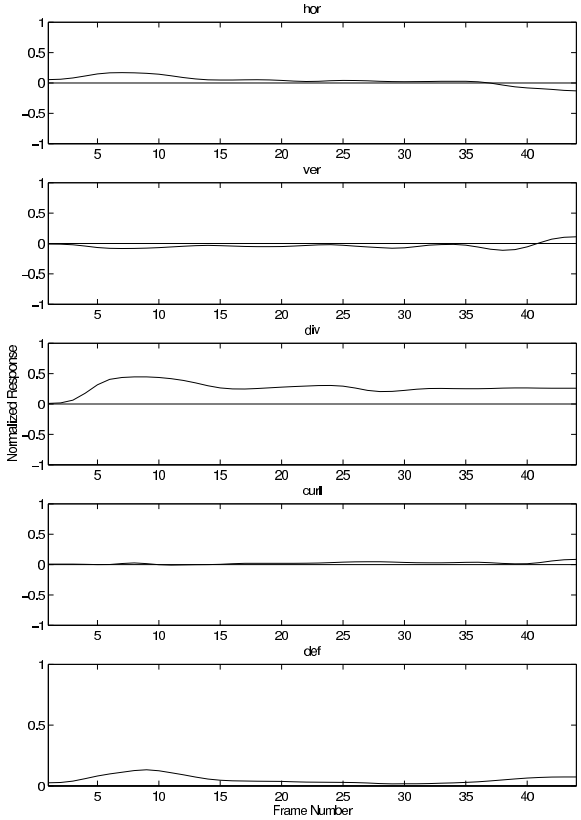
(a) Frame 0 (b) Frame 10 (c) Frame 20 (d) Frame 30



(e)

Figure C.7: Toward signer movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

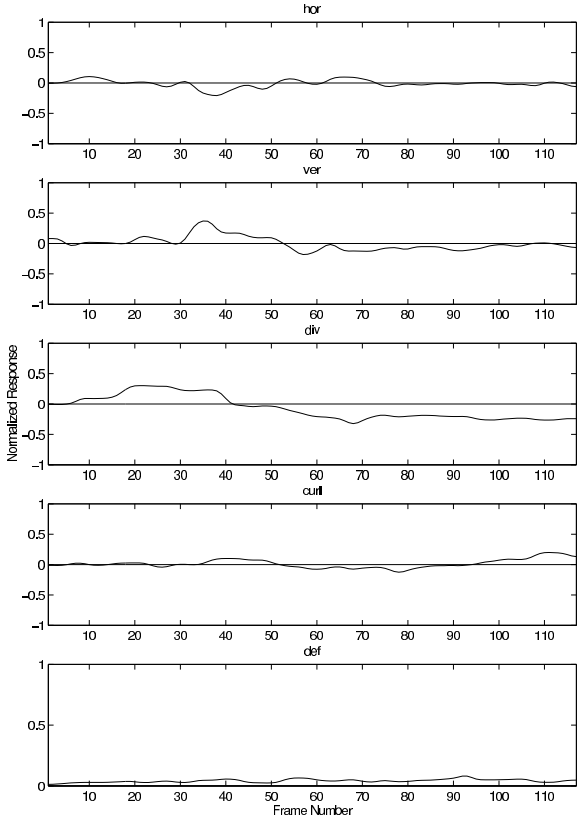
(a) Frame 0 (b) Frame 11 (c) Frame 22 (d) Frame 33



(e)

Figure C.8: Away signer movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

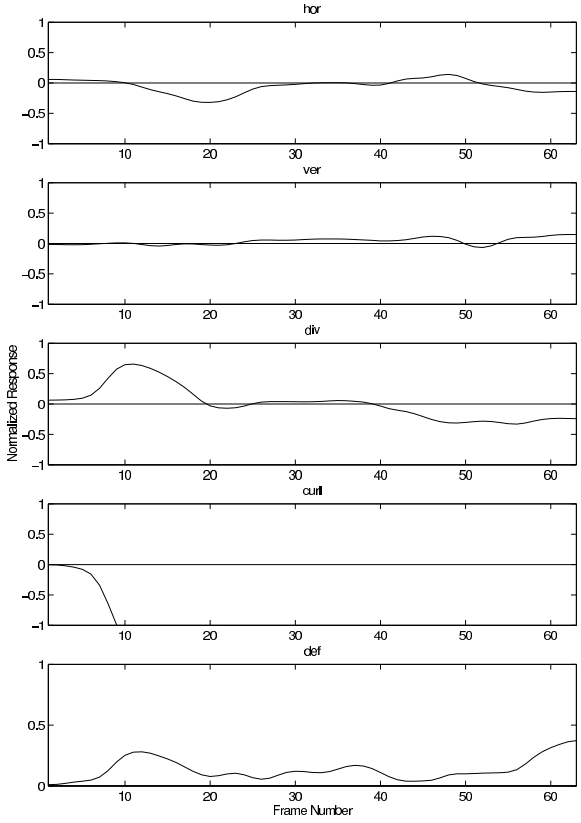
(a) Frame 0 (b) Frame 29 (c) Frame 58 (d) Frame 87



(e)

Figure C.9: To and fro movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

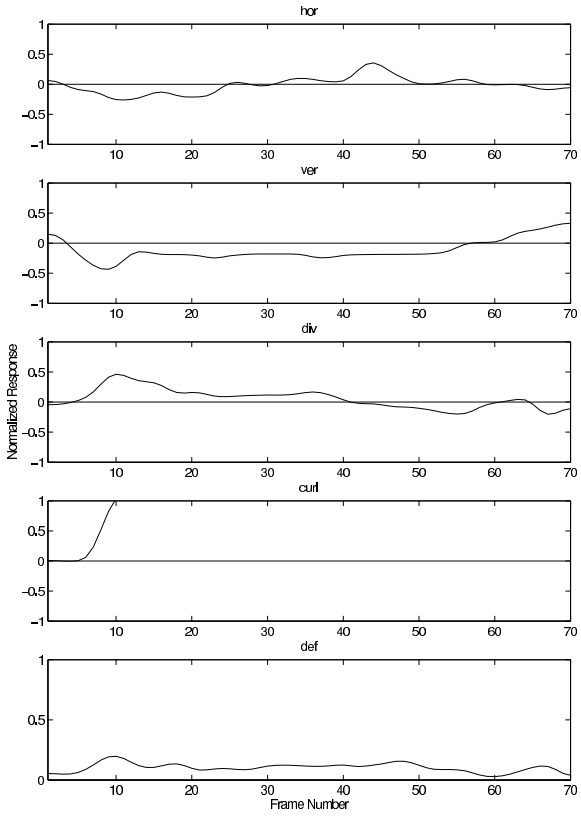
(a) Frame 0 (b) Frame 16 (c) Frame 32 (d) Frame 48



(e)

Figure C.10: Supinate movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

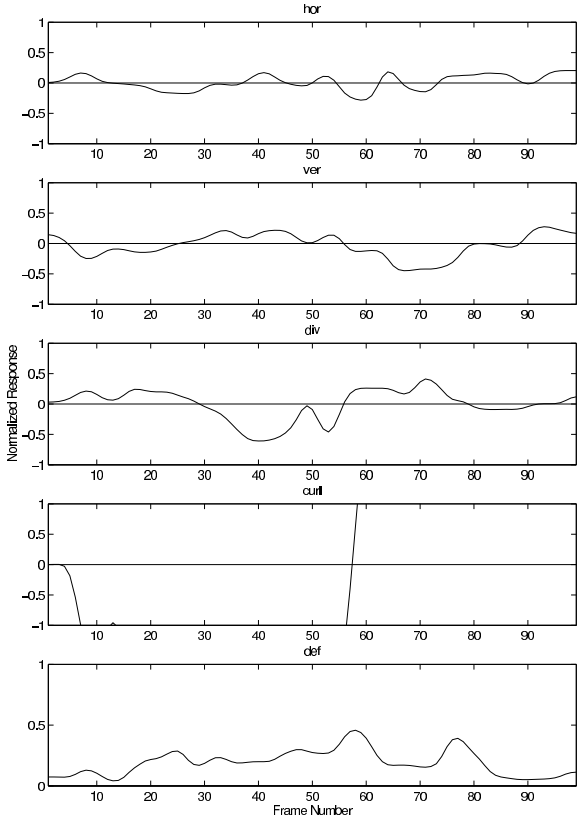
(a) Frame 0 (b) Frame 18 (c) Frame 36 (d) Frame 54



(e)

Figure C.11: Pronate movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

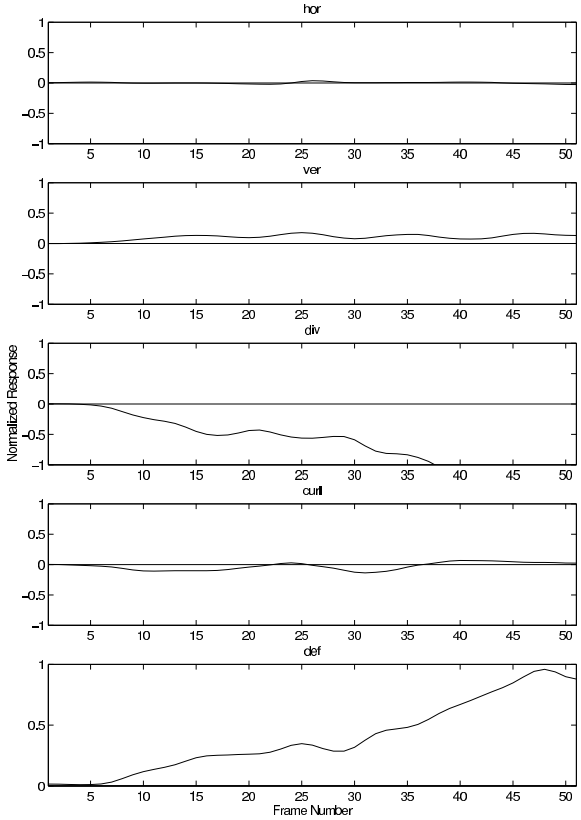
(a) Frame 0 (b) Frame 25 (c) Frame 50 (d) Frame 75



(e)

Figure C.12: Twist wrist movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

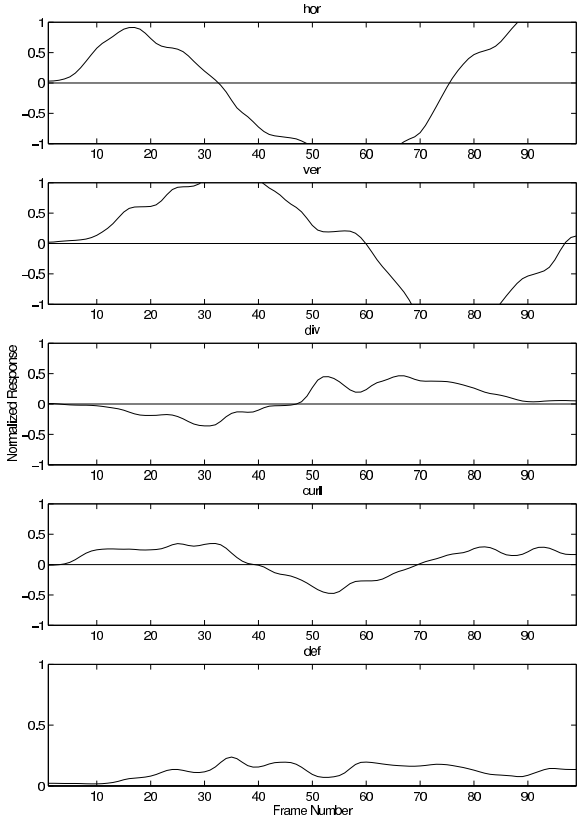
(a) Frame 0 (b) Frame 13 (c) Frame 26 (d) Frame 39



(e)

Figure C.13: Nod movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

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



(e)

Figure C.14: Circular movement example. (a)-(d) example frames (e) normalized kinematic time series plots.

Bibliography

[1] Webster's revised unabridged dictionary. http://www.dictionary.com, 1988.

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

[3] C.H. Anderson, P.J. Burt, and G.S. van der Wal. Change detection and tracking using pyramid transform techniques. In *SPIE Conference on Intelligent Robotics and Computer Vision*, pages 72–78, 1985.

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

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

[6] B. Bauer and H. Hienz. Relevant features for video-based continuous sign language recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 440–445, 2000.

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

[8] D.A. Becker. Sensei: A real-time recognition, feedback, and training system for T'ai Chi gestures. In *Vismod*, 1997.

[9] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *European Conference on Computer Vision*, pages I:5–10, 1992.

[10] S.T. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 232–237, 1998.

[11] M.J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *IEEE International Conference on Computer Vision*, pages 231–236, 1993.

[12] M.J. Black and A.D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *European Conference on Computer Vision*, pages I:329–342, 1996.

[13] M.J. Black and A.D. Jepson. A probabilistic framework for matching temporal trajectories: CONDENSATION-based recognition of gesture and expression. In *European Conference on Computer Vision*, pages II:909–924, 1998.

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

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

[16] A. Braffort. A gesture recognition architector for sign language. In *ACM Conference on Assistive Technologies*, pages 102 – 109, 1996.

[17] P.J. Burt and E.H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(12):532–540, Dec 1983.

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

[19] M. Collobert, R. Feraud, G. Le Tourneur, D. Bernier, Vaiallet, Y. J.E., Mahieux, and D. Collobert. Listen: A system for locating and tracking individual speakers. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 283–288, 1996.

[20] E. Costello and P.M. Setzer. *Random House Webster's American Sign Language Dictionary*. Random House, 1997.

[21] R. Cutler and M. Turk. View-based interpretation of real-time optical flow for gesture recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 416–421, 1998.

[22] T. Darrell and A. Pentland. Space-time gestures. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 335–340, 1993.

[23] J. Davis and M. Shah. Recognizing hand gestures. In *European Conference on Computer Vision*, pages A:331–340, 1994.

[24] G. Fang and W. Gao. A SRN/HMM system for signer-independent continuous sign language recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 297–302, 2002.

[25] S.S. Fels and 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):205–212, 1997.

[26] P. Fieguth and D. Terzopoulos. Color based tracking of heads and other mobile objects at video frame rates. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–22, 1997.

[27] M.M. Fleck, D.A. Forsyth, and C. Bregler. Finding naked people. In *European Conference on Computer Vision*, pages II:593–602, 1996.

[28] W.T. Freeman and M. Roth. Orientation histograms for hand gesture recognition. In *Mitsubishi Electric Research Labs, Rep. TR94-03*, 1994.

[29] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, San Diego, CA, 1972.

[30] K. Grobel and M. Assan. Isolated sign language recognition using hidden markov models. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 162–167, 1997.

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

[32] J. Han and M. Kamber. *Data Mining: Concepts and Techniques.* Morgan Kaufmann, San Francisco, CA, 2001.

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

[34] E. Holden, G.G. Roy, and R. Owens. Adaptive classification of hand movement. In *IEEE International Conference on Neural Netwoks*, pages 1373–1378, 1995.

[35] P. Hong, M. Turk, and T.S. Huang. Gesture modeling and recognition using finite state machines. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 410–415, 2000.

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

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

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

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

[40] B. Jahne. *Digital Image Processing-Concepts, Algorithms, And Scientific Applications.* Springer, Berlin, 1991.

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

[42] B. Kapralos, M.R.M. Jenkin, E. Milios, and J.K. Tsotsos. Eyes 'n ears face detection. In *IEEE International Conference on Image Processing*, pages 66–69, 2001.

[43] R. Kjeldsen and J.R. Kender. Finding skin in color images. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 312–317, 1996.

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

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

[46] H. Kollnig, H.H. Nagel, and M. Otte. Association of motion verbs with vehicle movements extracted from dense optical flow fields. In *European Conference on Computer Vision*, pages B:338–347, 1994.

[47] J.M. Lawn and R. Cipolla. Robust egomotion estimation from affine motion parallax. In *European Conference on Computer Vision*, pages A:205–210, 1994.

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

[49] R.H. Liang and M. Ouhyoung. A real-time continuous gesture recognition system for sign language. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 558–567, 1998.

[50] S. K. Lidell and E. Johnson. *American Sign Language: The Phonological Base*, volume 64, pages 195–277. Lindstok Press, 1989.

[51] H. Christopher Longuet-Higgins and Kiroslav Pradzny. The interpretation of a moving retinal image. *Proceedings of the Royal Society of London B.*, 208:385–397, 1980.

[52] W.J. MacLean, R. Herpers, C. Pantofaru, C. Wood, K.G. Derpanis, D. Topalovic, and J.K. Tsotsos. Fast hand gesture recognition for real-time teleconferencing applications. In *International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, pages 133–140, 2001.

[53] J. Mammen, S. Chaudhuri, and T. Agarwal. A two stage scheme for dynamic hand gesture recognition. In *National Conference on Communication*, pages 35–39, 2002.

[54] A.M. Martinez, B. Wilbur, R. Shay, and A.C. Kak. Purdue RVL-SLLL ASL database for automatic recogition of American Sign Language. In *IEEE International Conference on Multimodal Interfaces*, pages 167–172, 2002.

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

[56] Y. Nam, K. Wohn, and H. Lee-Kwang. Modeling and recognition of hand gesture using colored petri nets. *IEEE Transactions on Sytems, Man, and Cybernetics*, 29(5):514–521, September 1999.

[57] S. Negahdaripour and S. Lee. Motion recovery from image sequences using first-order optical flow information. In *IEEE Workshop on Visual Motion*, pages 132–139, 1991.

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

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

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

[61] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

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

[63] J.M. Rehg and T. Kanade. Visual tracking of high DoF articulated structures: An application to human hand tracking. In *European Conference on Computer Vision*, pages B:35–46, 1994.

[64] Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 111–118, 2000.

[65] H. Sagawa, M. Takeuchi, and M. Ohki. A two stage scheme for dynamic hand gesture recognition. In *International Conference on Intelligent User Interfaces*, pages 97 – 104, 1997.

[66] B. Schiele and A. Waibel. Gaze tracking based on face-color. In *International Workshop on Automatic Face- and Gesture-Recognition*, 1995.

[67] J. Schlenzig, E. Hunter, and R. Jain. Vision based gesture interpretation using recursive estimation. In *Asilomar Conference on Signals, Systems and Computers*, 1994.

[68] M. Shah and R. Jain. Visual recognition of activities, gestures, facial expressions and speech: An introduction and a perspective. In M. Shah and R. Jain, editors, *Motion-Based Recognition*, pages 1–14, 1997.

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

[70] T. Starner, J. Weaver, and 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):1371–1375, December 1998.

[71] B.D.R. Stenger, P.R.S. Mendonca, and R. Cipolla. Model-based hand tracking using an unscented kalman filter. In *British Machine Vision Conference*, 2001.

[72] W.C. Stokoe. *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf.* Linstok Press, Silver Spring, MD, 1960.

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

[74] N. Tanibata, N. Shimada, and Y. Shirai. Extraction of hand features for recognition of sign language words. In *International Conference on Vision Interface*, pages 391–398, 2002.

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

[76] C. Valli and C. Lucas. *Linguistics of American Sign Language: An Introduction.* Gallaudet University Press, Washington, D.C., 2000.

[77] M.J.C. Van Gemert, S.L. Jacques, H.J.C.M. Sterenborg, and W.M. Star. Skin optics. *IEEE Transactions on Biomedical Engineering*, 36(12):1146–1154, 1989.

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

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

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

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

[82] K.Y. Wong and M.E. Spetsakis. Motion segmentation and tracking. In *International Conference on Vision Interface*, pages 80–87, 2002.

[83] C.R. Wren, A. Azarbayejani, T.J. Darrell, and A.P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.

[84] J. Yang and A. Waibel. A real-time face tracker. In *IEEE Workshop on the Application of Computer Vision*, pages 142–147, 1996.

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

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

[87] B. Zarit, B.J. Super, and F. Quek. Comparison of five color models in skin pixel classification. In *International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, pages 58–63, 1999.

[88] X. Zhu, J. Yang, and A. Waibel. Segmenting hands of arbitrary color. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 446–453, 2000.

[89] Y. Zhu, H. Ren, G. Xu, and X. Lin. Toward real-time human-computer interaction with continuous dynamic hand gestures. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 544–549, 2000.