# Hand Gesture Recognition within a Linguistics-Based Framework

Konstantinos G. Derpanis, Richard P. Wildes, and John K. Tsotsos
{kosta, wildes, tsotsos}@cs.yorku.ca
Department of Computer Science
Centre for Vision Research (CVR)
York University, Toronto Canada

## Introduction

**Motivation:** By exploiting natural hand gestures, a powerful interface between man and machine can be realized.
**Objective:** Automated gesture recognition with a single video camera + PC.
**Current focus:** Motion extraction and categorization of hand movements from a frontoparallel view of the signer given the initial location of the hand.
**Requirements:** Accuracy in recognition, real-time performance, unobtrusiveness and scalability to encompass sizable vocabularies (e.g. American Sign Language).

## Contributions

1. Our approach models gestures in terms of their phonemic elements to yield an algorithm that recognizes gesture movements given data captured with a single video camera irrespective of both the location in space the gesture is made and the hand articulation.
2. The apparent motion of an unmarked hand as input is used as opposed to fitting a 3D model of a arm (hand) or using a mechanical device (e.g. magnetic tracker).
3. Our recognition scheme is based on a nearest neighbour match to prototype signatures found analytically; no training is required.

## General approach

**Linguistics basis (Gesture modeling):**
Represent the lexical components of the language based on a small set of elemental movements defined by linguistics theory.

**Tracking (Motion estimation):**
Given a monocular view, extract the affine flow of a hand gesture sequence.

**Gesture signatures (Classification):**
Map motion parameters to kinematic features that yield unique signatures for each elemental movement.

## Linguistics basis: Phonemic analysis

## Motion estimation

**Model:**
An affine model is used to characterize the frame to frame motion of the hand.

$$u(x,y) = a_0 + a_1 x + a_2 y$$
$$v(x,y) = a_3 + a_4 x + a_5 y$$

**Justification:**
The primitive movements may be sufficiently characterized by a first-order kinematic description.

Over the small angular extent that encompasses the hand, small motions of the approximately planar hand can be approximated by an affine model.

**Estimation Technique:**
Robust, hierarchical, gradient-based motion estimate from frame to frame on skin colour regions.

## Kinematic time series

The affine parameter time series are converted into kinematic time series owing to their descriptive power in the current context.

$$hor(t) = a_0(t)$$
$$ver(t) = a_3(t)$$
$$div(t) = a_1(t) + a_5(t)$$
$$curl(t) = -a_2(t) + a_4(t)$$
$$def(t) = \sqrt{[a_1(t) - a_5(t)]^2 + [a_2(t) + a_4(t)]^2}$$

Following initial recovery, each time series is rescaled in amplitude to make comparisons of relative magnitudes meaningful.

## Experimental design

**Goal:** Test ability of our algorithm to correctly recognize movement irrespective of the hand location and hand shape parameters from a frontoparallel view of the signer with respect to the camera.

Consideration of all possible combinations of movement, location and hand shape would lead to an infeasible experiment.
Full space: 14 (movements) x 19 (hand shapes) x 12 (locations) = 3192 combinations for each volunteer!!!

Configuration space is subsampled in location and hand shape by exploiting similarities.
Subsampled space: 14x3x4 =168 combinations

Both manual and automatic initialization of the hand region were used

## Results: Confusion matrix

## Results: Summary statistics

**Manual initialization:**
Overall 97% of test cases were classified correctly.
Correct classification within top 2 movement candidates was 99%.

**Automatic initialization:**
Overall 92% of test cases were classified correctly (omitting 14 failed tracked cases)
Correct classification within top 2 movement candidates was 95%.

Tracking speed averaged 8 frames per second using a Pentium 4 2.1GHz (unoptimized C++ code), classification time is negligible

## Conclusions

Three key concept behind approach:
1. It is desirable to recognize natural gestures with minimal constraints on both the user (i.e., unmarked hand) and the machine (i.e., single video camera and PC)
2. Appeal to linguistics theory to represent complex gestures in terms of primitive components.
3. Define distinctive signatures analytically (as opposed to training) that can be recovered from a monocular image sequence.
Empirical evaluation of the algorithm suggests its applicability to the analysis of complex gesture videos.