Non-accidental features for gesture spotting

Adam Fourney and Richard Mann

David R. Cheriton School of Computer Science
University of Waterloo

May 2009

Non-accidental features

- Non-accidental features are often invoked for image interpretation
  (Lowe, 1985; Jepson and Mann, 1999; Jepson and Richards, 1995)
  - image features, such as collinear, coterminating, or parallel edges, provide strong evidence of regularities in the world

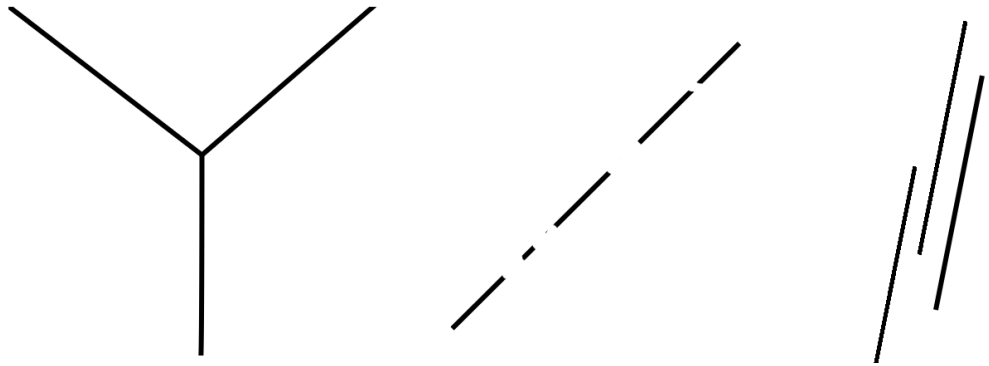
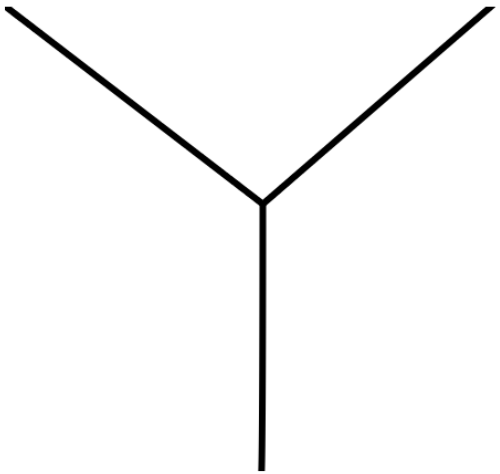


FIGURE 1: Coterminating, collinear and parallel edges.

Non-accidental features



- Consider an "unstructured" world, where line segments are strewn about randomly in 3D space
- In this 2D image, 3 segments appear to coterminate
- Can we infer that these 3 line segments actually coterminate in 3D space?

- What if blocks are regular occurrences in this world?
- We are more willing to consider that these 3 line segments do coterminate in space

Non-accidental features

Non-accidental features:

- best explained by underlying structure or regularities in the world (e.g: expected presence of blocks, polygonal surfaces) rather than by coincidences (e.g: accidental view-point)
- in the absence of measurement noise, non-accidental features occupy a lower dimensional subset of the feature space
    - structures impose constraints on the features and reduce their degrees of freedom

Spotting hand gestures

- Many have explored the visual perception and recognition of hand gestures as a mechanism for HCI
  (Eickeler *et al.* , 1998; Lee and Kim, 1999; Kim and Song, 2007)
- A challenge faced by vision-based gestural interfaces is that of *gesture spotting*
  - Cameras stream observations including both gesture and non-gesture hand motion
  - Systems must "spot" meaningful gestures embedded in these motion sequences

Application of non-accidental features to gesture spotting

	Line Segments	Hand Motion
"unstructured" world	randomly scattered line segments	unconstrained hand motion (e.g: gesticulation)
"structured" world	polygonal surfaces are regular occurrences	gestures are regular occurrences

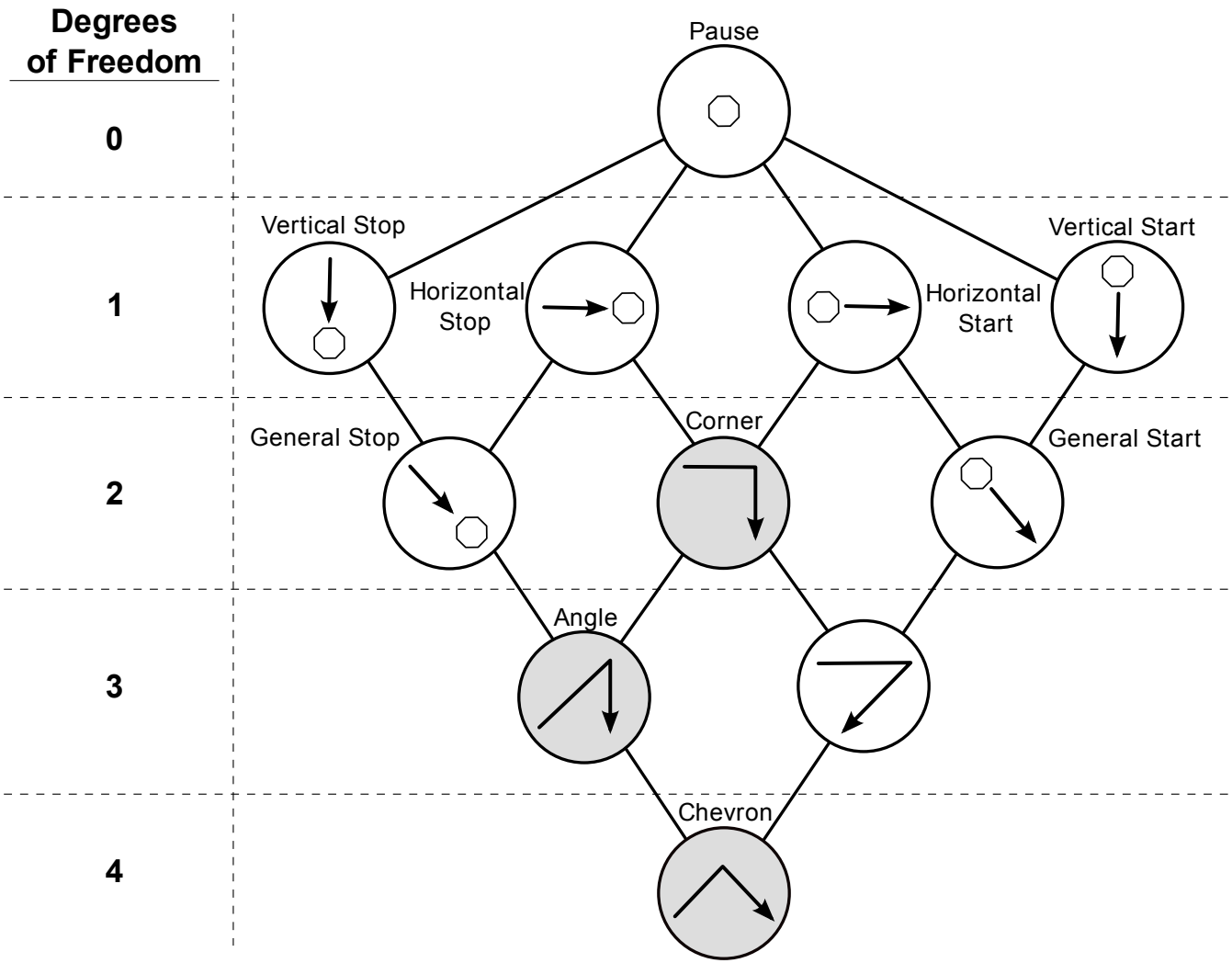
We hypothesize that:

- humans can perform highly structured hand motions that reliably give rise to detectable non-accidental motion features
- gestures that incorporate these motions will be easily differentiated from the background

Designing the gesture language

- Requires some assumptions about the types of regularities that a person can reliably introduce when instructed
  - We assume the following regularities: horizontal movement, vertical movement, and rest
- We consider gestures consisting of two motion segments, each of which can take on one of the three aforementioned categories, or can be "unconstrained" motion

Gesture forms



Features

- We fit a piecewise linear (constant velocity) model to the observed hand trajectories
- "Optimal" piecewise linear segmentation achieved using dynamic programming (Mann, *et al.* 2002)
- Our features describe the hand velocities before and after velocity discontinuities
  - $\mathbf{F} = [X'_-(t) \;\; Y'_-(t) \;\; X'_+(t) \;\; Y'_+(t)]^T$
- Each of our gestures should yield at least one feature

Experiment

- Each gesture performed 100 times
- Each instance performed in a different location in space
  - Gestures performed wrt. a moving target
  - User never repeats an identical trajectory

Observed results

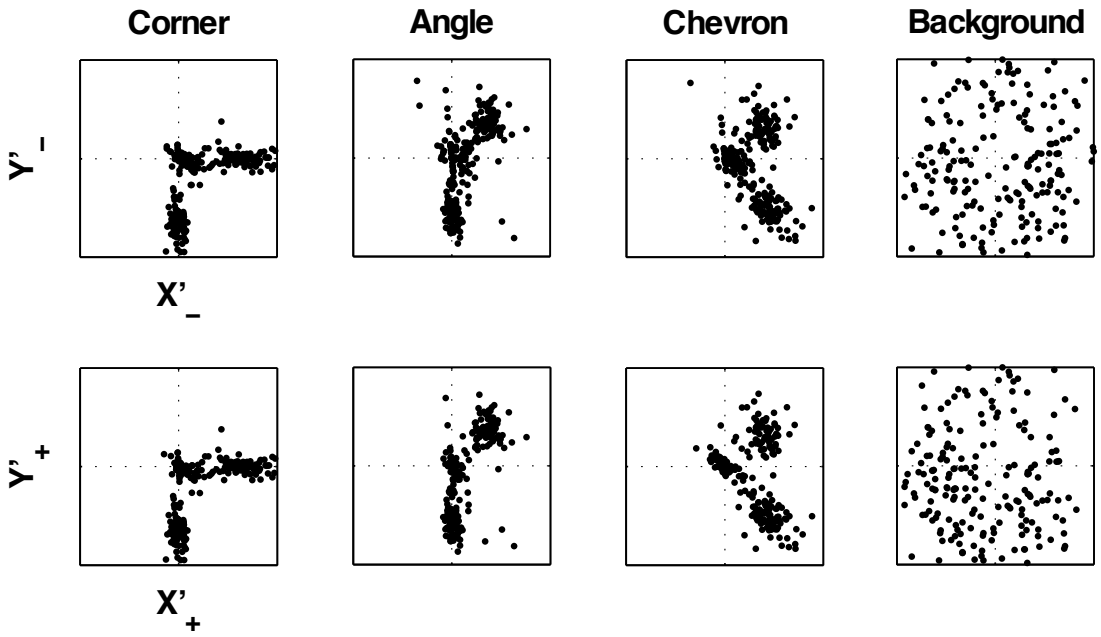


FIGURE 2: Observed features arising from the 3 gestures as well as from unconstrained motion.

Clustering with a Gaussian mixture model

- A Gaussian mixture model is used to cluster the features
- Each gesture G is modelled by a separate mixture model M_G, which generates features f according to the following:

$$P(f|M_G) = \sum_{i=k}^{K} \pi_k \; N(f \; ; \; \mu_k, \Sigma_k)$$

- $\pi_k, \quad \sum_{k=1}^{K} \pi_k = 1$, is the prior probability of generating data from component k
- $N(f \; ; \; \mu_k, \Sigma_k)$ is the Gaussian distribution with mean μ_k and covariance matrix Σ_k for component k

To learn the model parameters from data, we use the expectation maximization (EM) algorithm. (McLachlan and Basford, 1988)

Mixture model: "Corner"

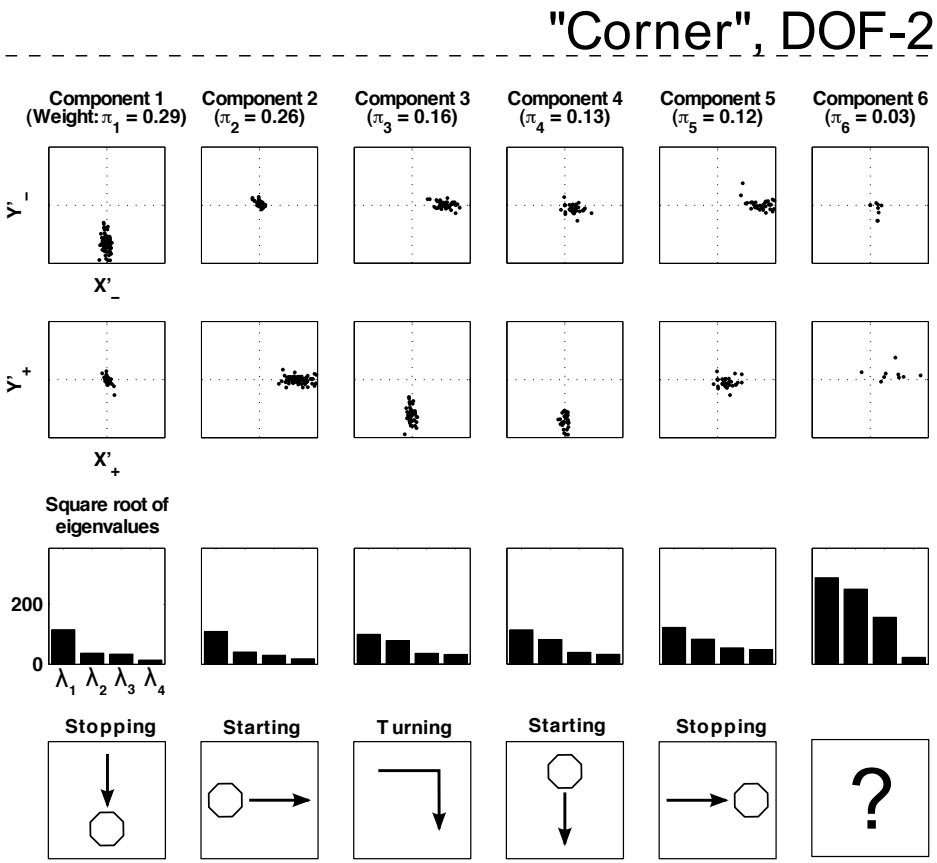


FIGURE 3: Component distributions learned for the "corner" gesture. Each column represents a different component.

Mixture models: "Angle" and "Chevron"

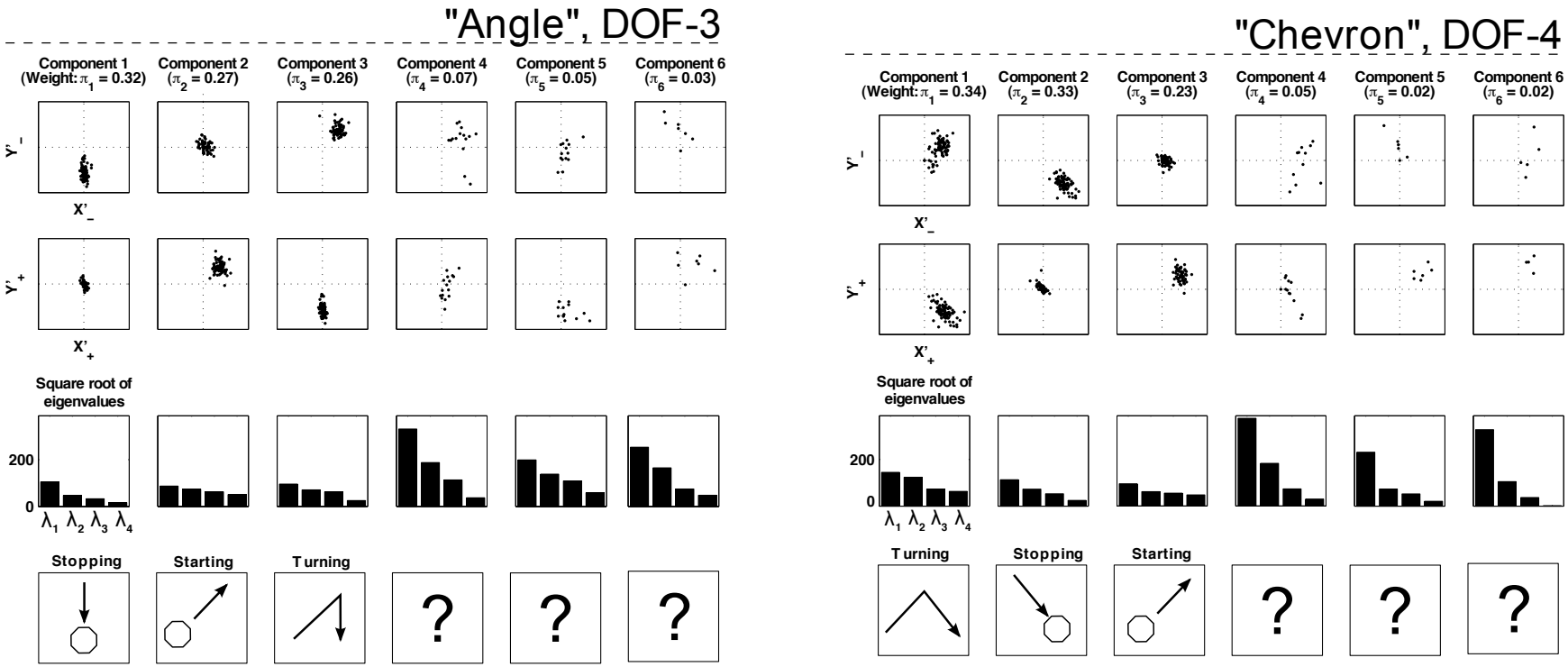


FIGURE 4: Component distributions learned for the "angle" and "chevron" gestures.

Gesture spotting

- Our gestures were motivated by non-accidental features. It it is natural to search for gestures in their vicinity.
- Every feature is classified as either belonging to the background or as arising from a gesture.

Gesture spotting experiment

For each gesture:

- record 20 instances embedded in longer motion sequences
- features that are assigned a high likelihood by a gesture's mixture model are labeled as arising from gesture
- we set the likelihood threshold so that it yields a precision of 95% on held-out training data

Results

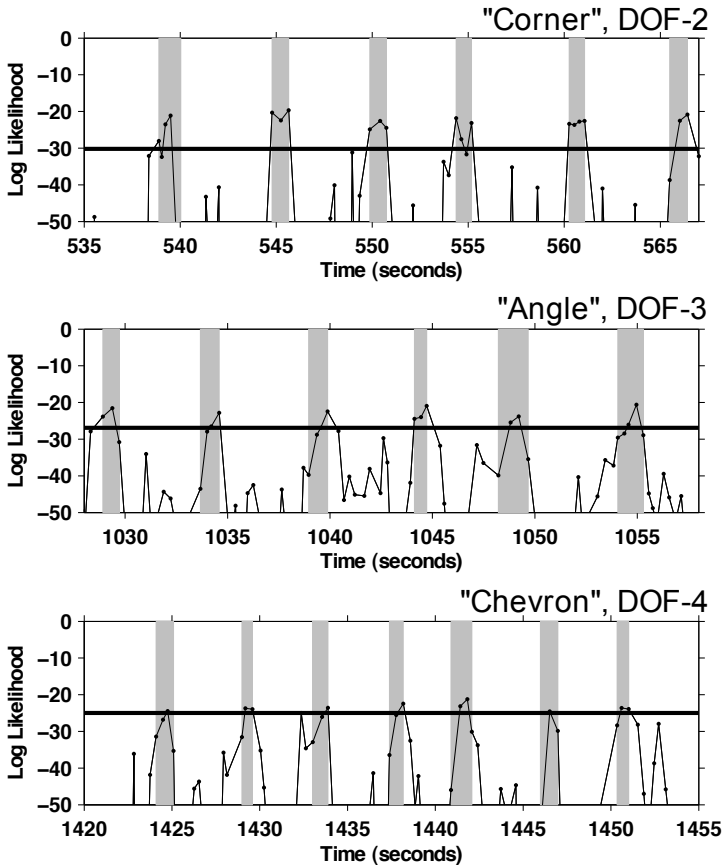


FIGURE 5: Log likelihood $P(f|M_G)$ of the model over time (as new features arrive). Shaded regions indicate intervals in which the gesture is known to have occurred.

Conclusion

- Non-accidental features do occur when people perform certain gestures
- Gestures yielding strong non-accidental features are easier to spot
- Many gestures yield unexpected non-accidental features (such as stopping and then restarting mid-gesture)
  - occur when people hesitate briefly between the strokes of the gesture
  - occur when a gesture is preceded or followed by periods of rest

Future work

- Consider features arising from two-handed (bimanual) gestures
- Exploit temporal constraints between motion segments in more complex gestures

Thank you.

Questions?