Computer Science @ RIT

B. Thomas Golisano College of Computing & Information Sciences

Action Classification

"Crude classifications and false generalizations are the curse of organized life." - G. B. Shaw

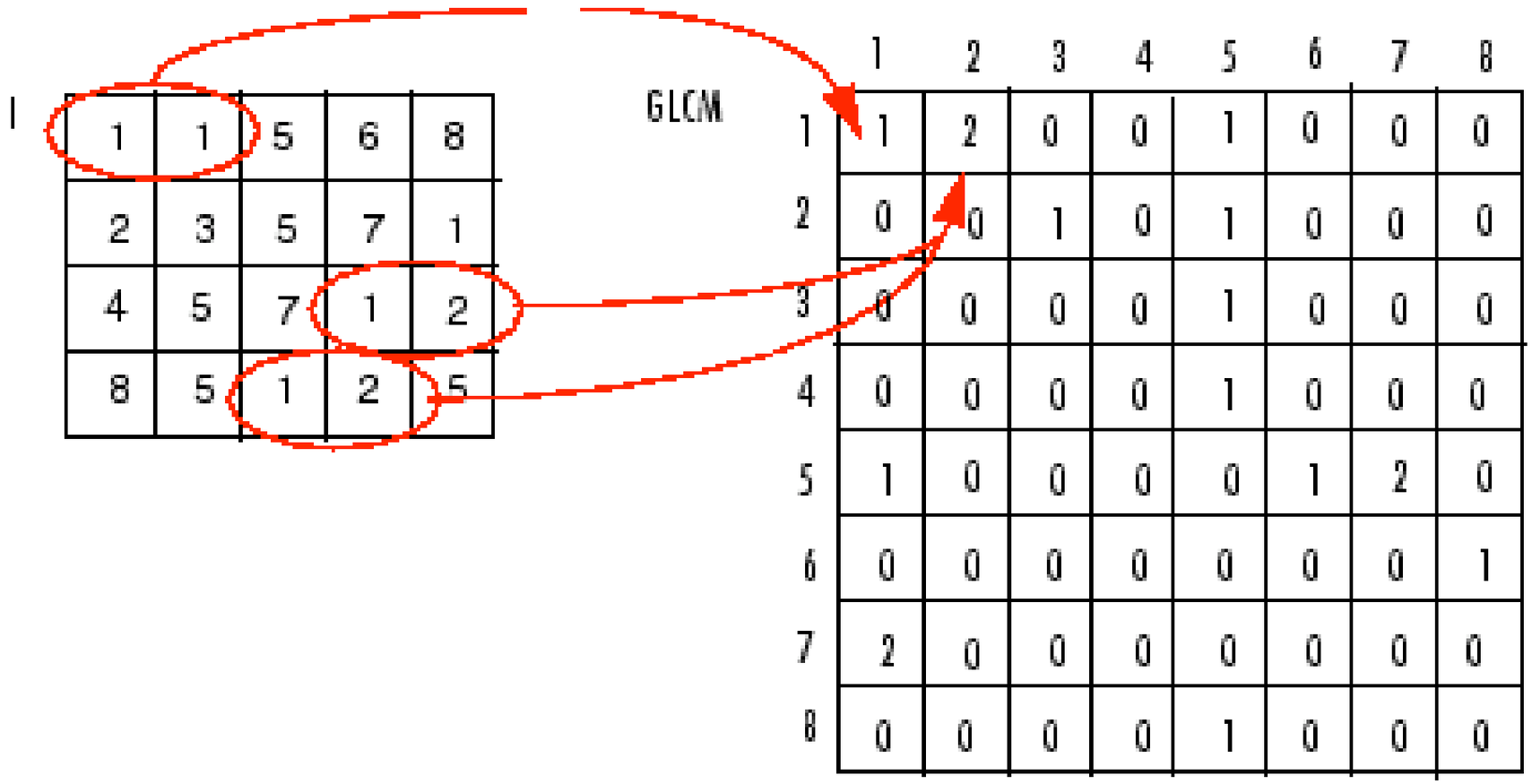
Kurt Alfred Kluever (kurt@kloover.com)
Department of Computer Science
Rochester Institute of Technology

Definitions

- Spatial: relating to space
- Temporal: relating to time
- Codewords: appearance elements
- Video words: like codewords, but associate interest points to the closest element of a vocabulary

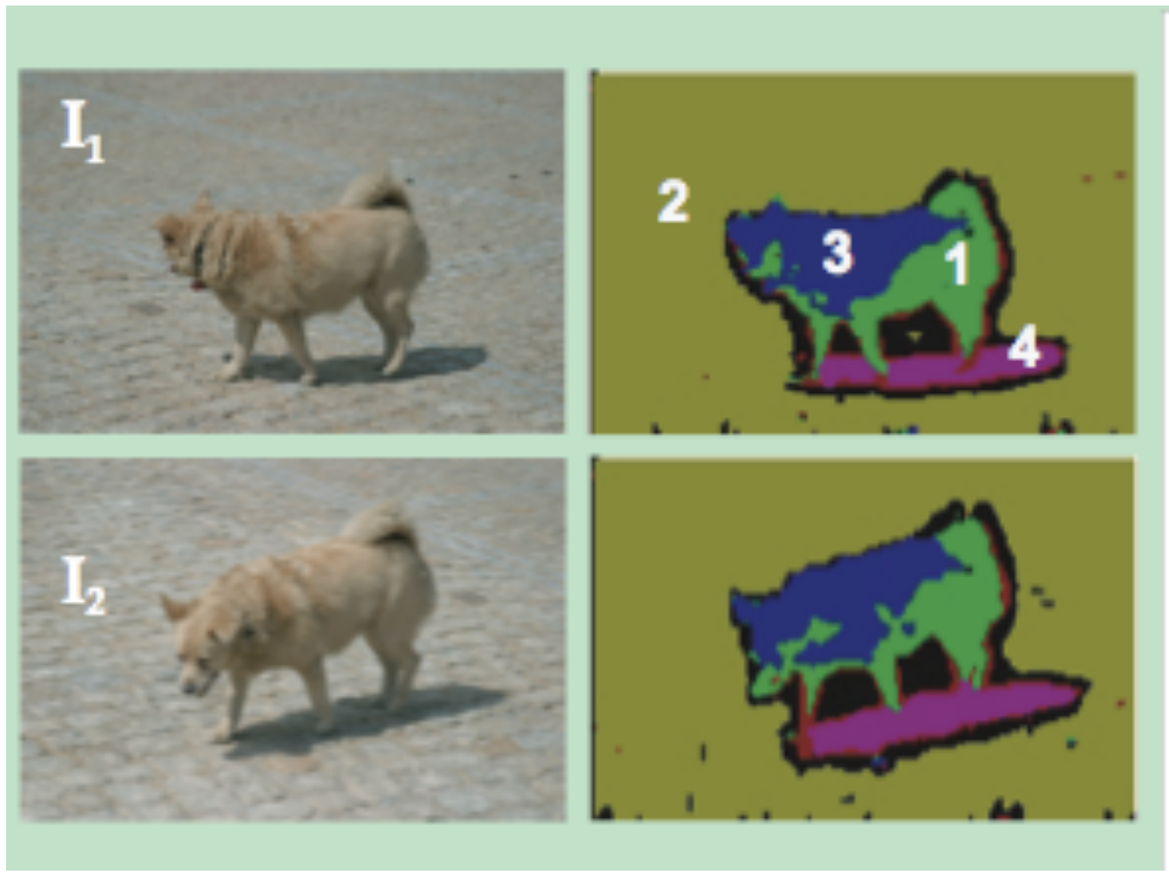
Co-Occurrence Matrix

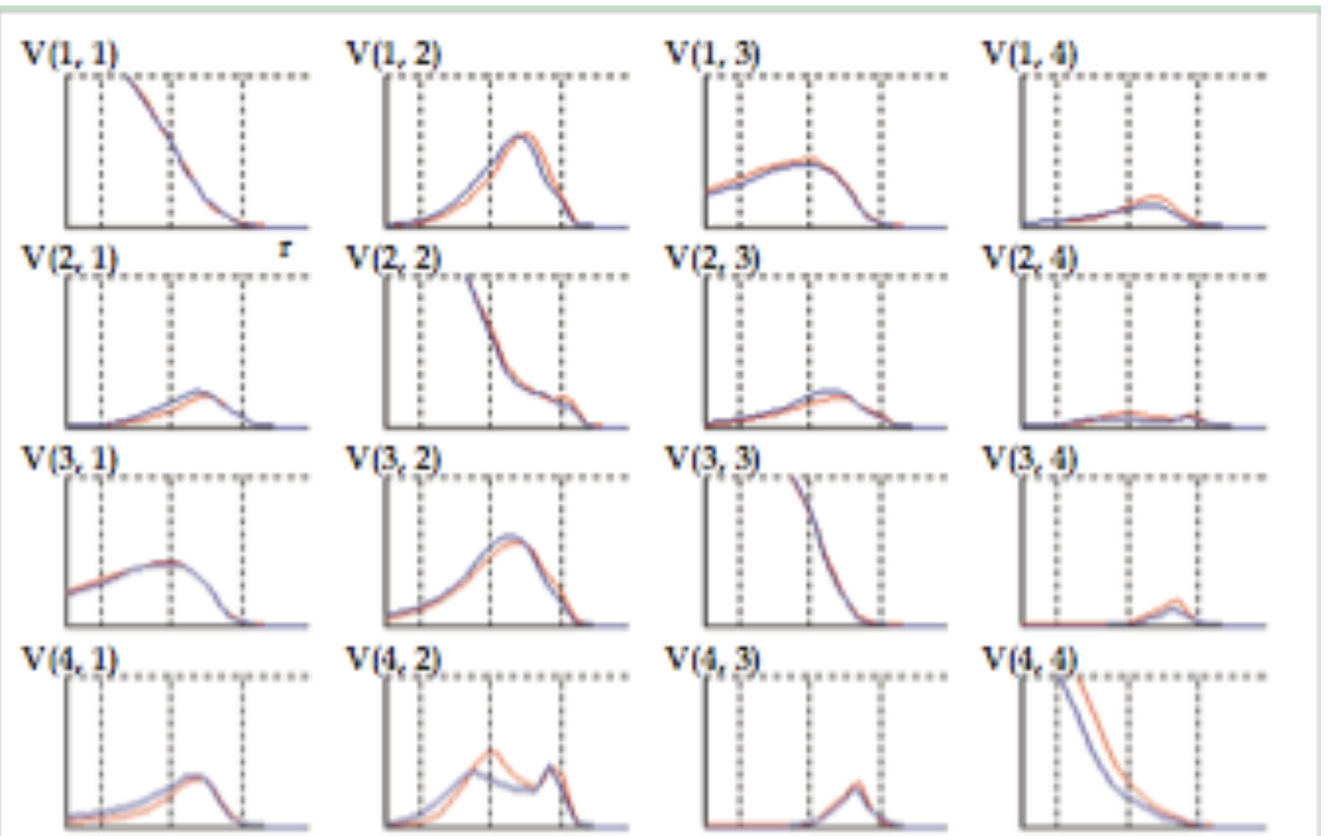
- Captures spatial dependence and angular variations
- Remember from Intro to CV?



Correlograms

- Similar to co-occurrence matrices
- Capture spatial arrangement of codewords
- Tolerate variations in color and shape





Correlatons

- Adaptive vector quantized correlograms
  - ...wait, what?
- Represents the spatial co-occurence of a pair of codewords
  - Correlations of codewords
- Instead of gray levels, we use codewords

Object Models

- "Object models model objects"
- Describes an object
- Histogram of Codewords: capture appearance of information
- Correlatons: capture spatial relationships
- Object Model = Histogram of Codewords + Correlatons

ST-correlatons

- "Spatial-Temporal Correlatons"
- Turn the correlograms into vectors
- Contain spatial and temporal information
- Capture the spatial-temporal correlations of video words
- Video Model = ST-correlaton + distribution of video words
  - very compact representation

Some math

- Local histogram given a kernel: $\mathbf{H}(\Pi)$
  - # of occurrences of label *i* within the kernel π
- Place *K* kernels centered at interest point **p**
- Average Local Histogram:

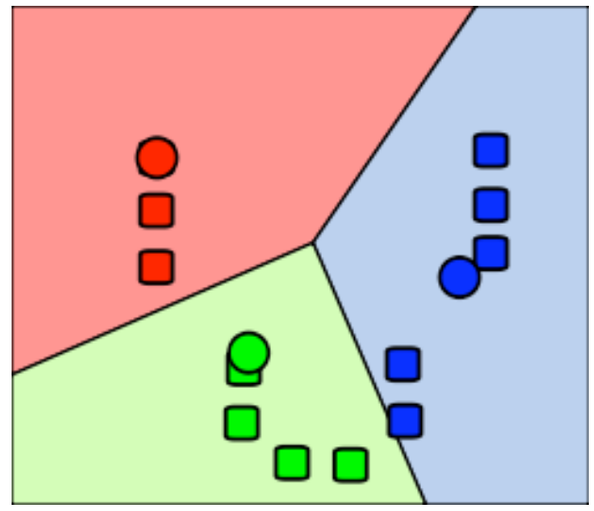
$$\hat{\mathbf{H}}(\Pi_r, i) = \sum_{\mathbf{p} \in \{\mathbf{p}_i\}}^{|\mathbf{p}_i|} \frac{\mathbf{H}(\Pi_r, \mathbf{p})}{|\mathbf{p}_i|}$$

- Concatenate local histograms for all labels and kernels => Correlogram

Kernels

- Using of kernels of various shapes/sizes are more effective in capturing video words
- Spheric kernels
  - Expensive to compute
  - Rotational invariance
- Square kernels
  - Computationally efficient
  - No rotational invariance

Representation



- Detect space-time interest points
- For each point, extract a descriptor (from brightness gradients)
- Cluster the descriptors using K-means
  - Randomly choose centroids, calculate clusters, move centroids, loop until convergence
- Cluster centers form the vocabulary
- Label spatial-temporal interest points
- Cluster ST-correlogram to get ST-correlaton

pLSA

- Probabilistic Latent Semantic Analysis
  - models the probability of each co-occurrence
- Spatial-temporal relationships are maintained due to the correlatons

- Probability of a codeword *w* occurring in a video *d*: $$P(w_i|d_j) = \sum_{k=1}^{K} P(z_k|d_j)P(w_i|z_k)$$
- Use expectation-maximization algo to find the most likely action class *z*

Experiment

- Video words are built from:
  - 2 videos of each action from 3 subjects
- Model built from 24 subject
- Tested against 1 subject
- Testing with kernel sizes from 2 to 40 pixels (spatial) and 2 to 60 frames (temporal)

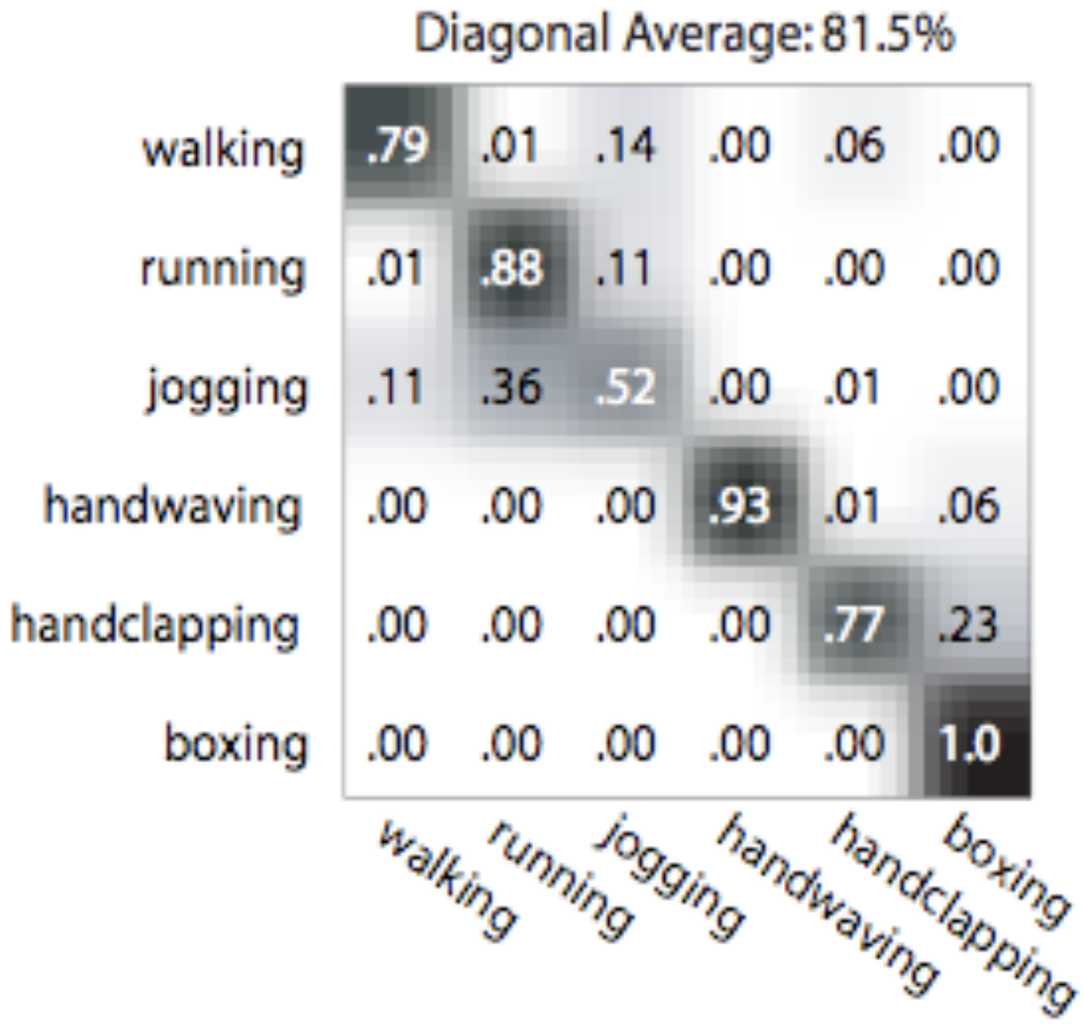
Results

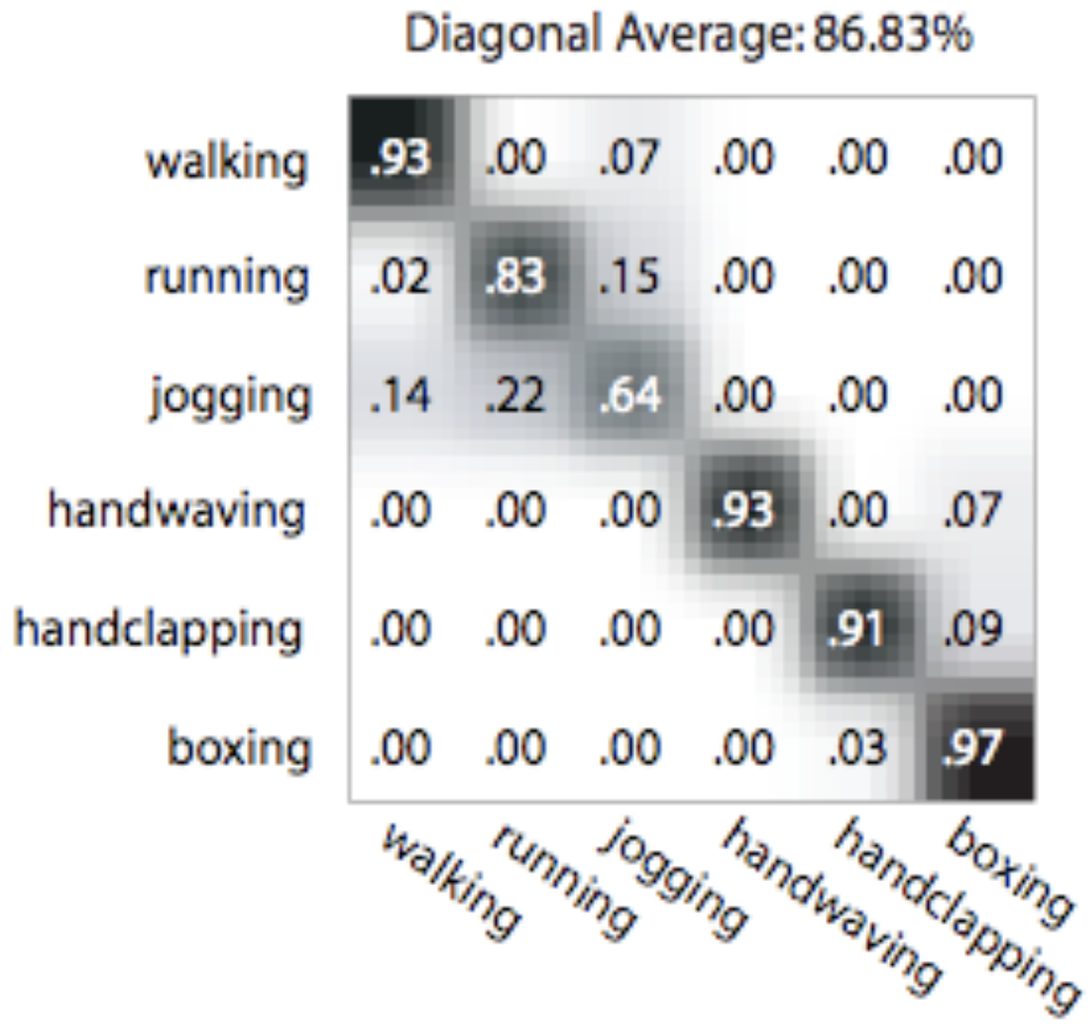
- Models that include ST-correlatons in addition to video words perform better
- Test data comes from KTH database
  - challenging database (moving cameras, variable backgrounds, cluttered scenes)
- Results are dependent on
  - the sizes of vocabulary and ST-correlatons
  - kernel type and configuration

Example from KTH

Results: Confusion Matrix

- Left: Using 500 word vocab
- Right: Using 500 word vocab + ST-correlatons





Conclusion

- Spatial-temporal correlograms capture important relationships
- Highest unsupervised learning accuracy for KTH dataset

References

[1] Silvio Savarese, Andrey DelPozo, Juan Carlos Niebles, and Li Fei-Fei. Spatial-temporal correlations for unsupervised action classification. In *IEEE Workshop on Motion and Video Computing*, Copper Mountain, Colorado, January 2008.