A “String of Feature Graphs” Model for Recognition of Complex Activities in Natural Videos

U. Gaur, Y. Zhu, B. Song, A. Roy-Chowdhury
University of California, USA

Nov 2011 - ICCV

Aytaç ÇAVENT

Motivation

- Activity recognition involves understanding interactions between objects.
  - Kinesics of individual objects
  - Chronemics or temporal aspects
  - Proximics or spatial relationship between objects
  - Haptics
- Complex activities involves multiple objects interacting with each other.

Challenges

- Individual variances
- Spatial and temporal extend of an activity
- Background clutter
- Occlusions
- Illimunation variations
- Viewpoint changes
- ….

Related Work

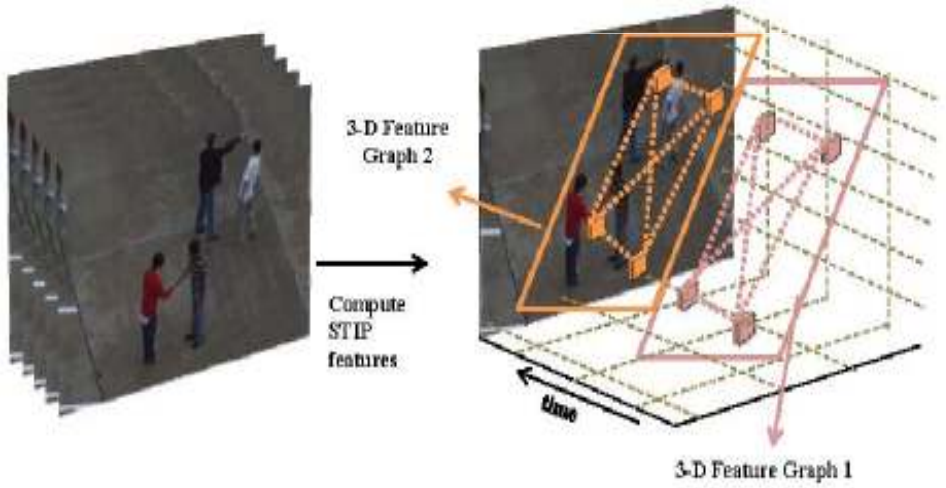
- Local and global features used for activity recognition
  - STIP, Cuboids, Space Time Shapes
- Graphical models, Involves conditional dependencies,
  - Dynamic Bayesian Networks
  - Finite State Model
  - Hidden Markov Model
- Syntactic model
  - Motivated by grammers in language modeling
  - How to construct activities from action primitives using rules as grammers.
- Time series model with temporal alignment
  - Dynamic time warping

Problem Definition

- Model spatial relationship between objects.
- Model temporal evolutions of these relationships.
- Similarity computation in both spatial and temporal domain.
- Should be able to work without large number of examples.
- Database can be complex and the query can be simple.

Feature Graph

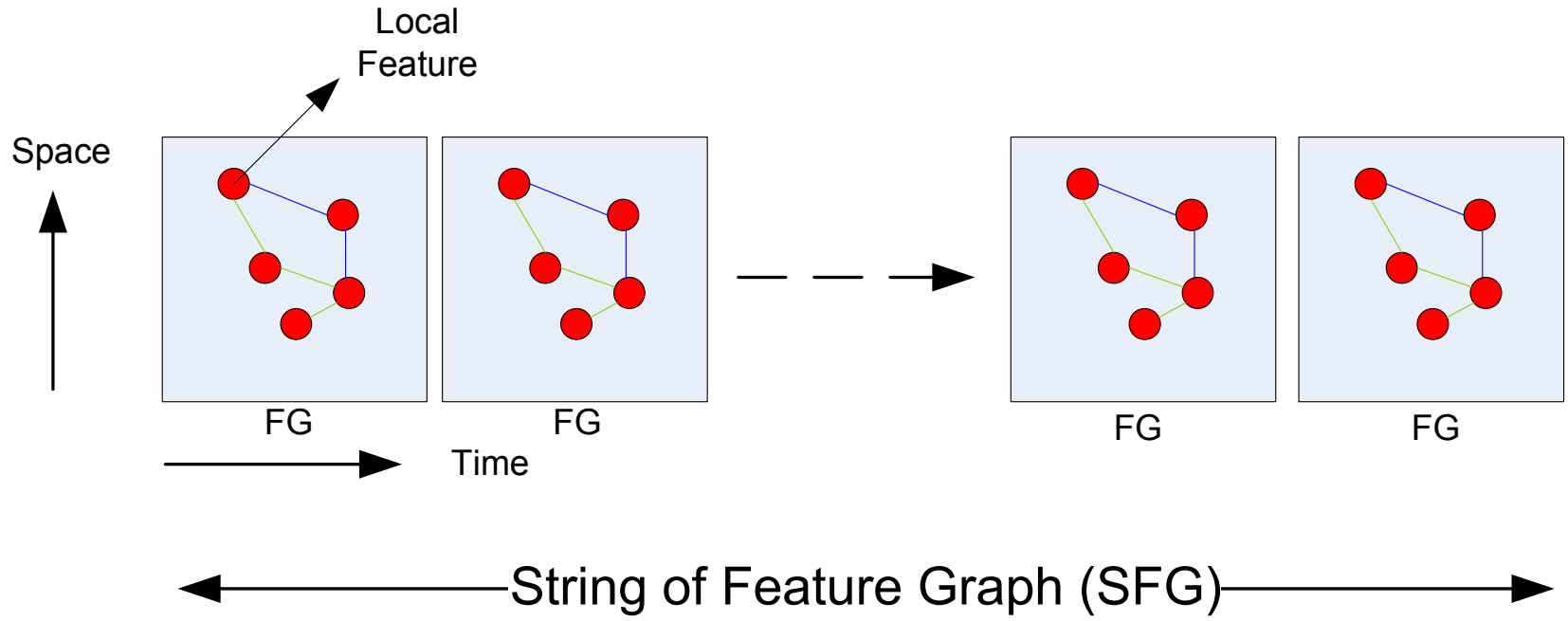
- A video can be thought of as a spatio-temporal collection of primitive features (STIP features).
- Divide the features into small temporal bins and each bin consisting of a 3D graphical structure representing the spatial arrangement of the low-level features



Edges are euclidian distance between two features.

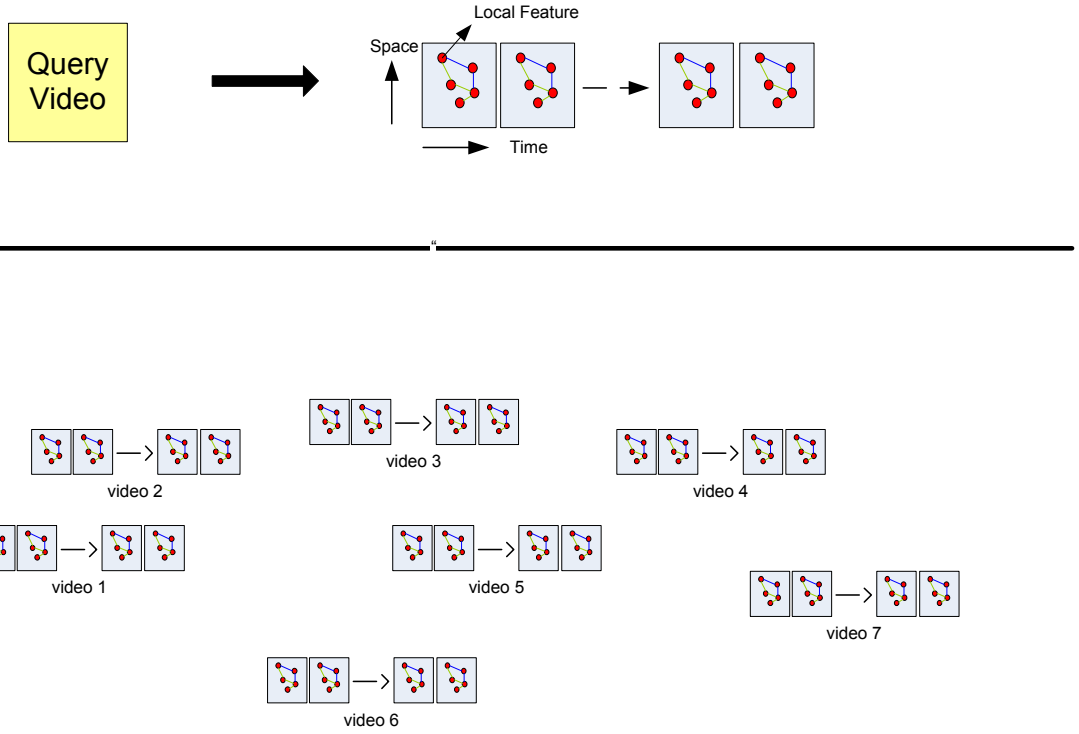
String of Feature Graphs (SFGs)

- Represent the video as a temporally ordered collection of such feature-bins.

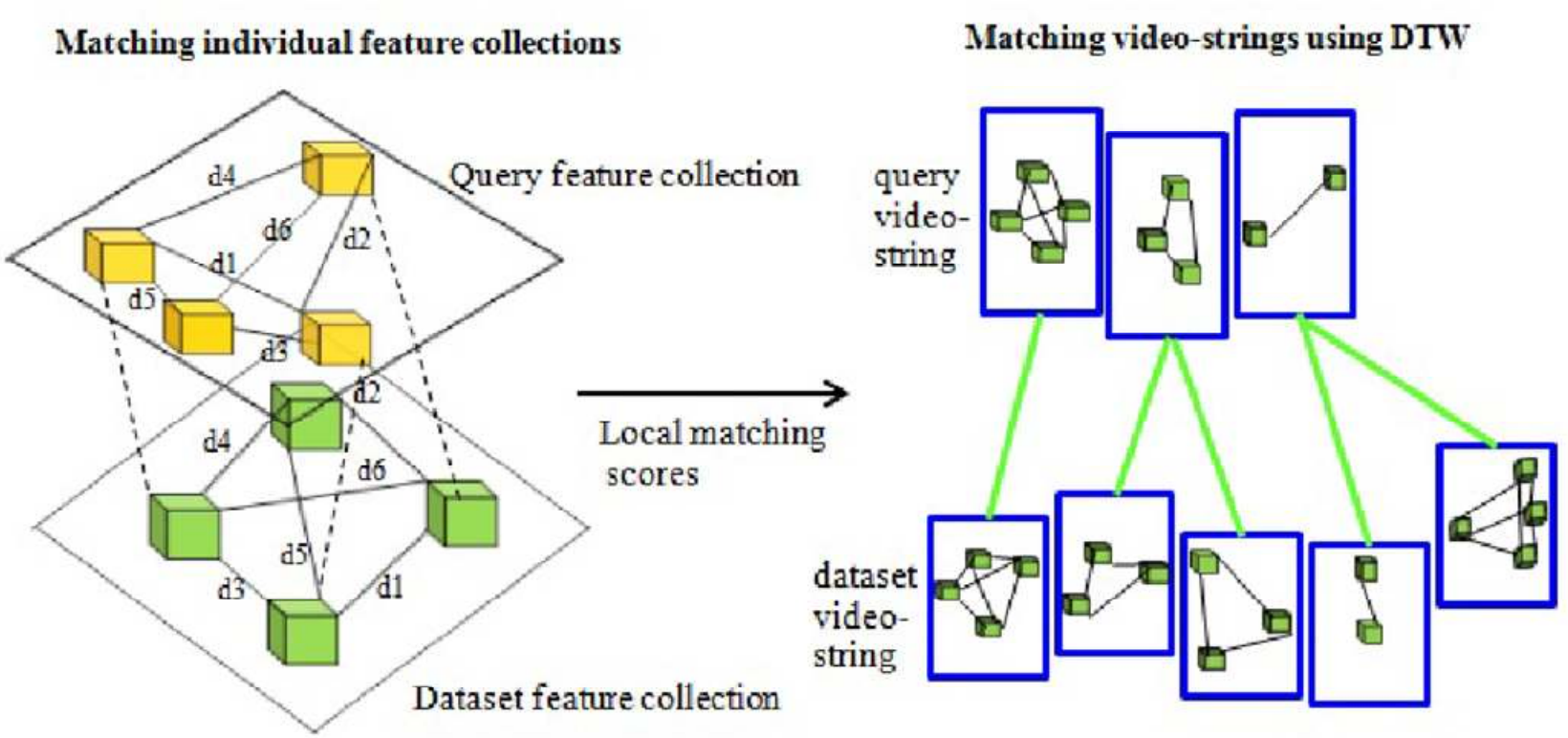


Matching Videos

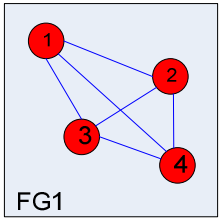
- Find similar videos that contains similar actions.
- Do not label the action type(hug, walk…)

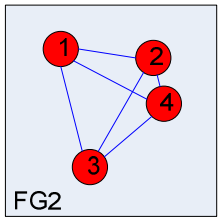


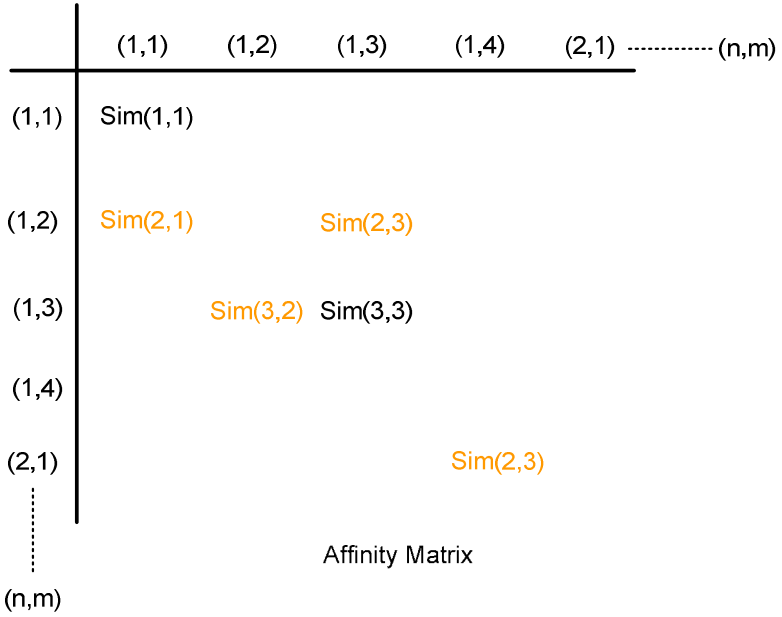
Matching SFGs



Matching FGs







$$\mathbf{M}(a,a) = \begin{cases} \tau_n - d_n(i,i') & d_n(i,i') \le \tau_n \\ 0 & d_n(i,i') > \tau_n \end{cases}$$

$$\mathbf{M}(a,b) = \begin{cases} \tau_e - d_e(\vec{ij}, \vec{i'j'}) & d_e(\vec{ij}, \vec{i'j'}) \le \tau_e \\ 0 & d_e(\vec{ij}, \vec{i'j'}) > \tau_e \end{cases}$$

$d_n(i,i')$, euclidian distance between two nodes,

$d_e(i,i')$, angle between two edges,

Matching FGs

- Find the best correspondences (i,j) using affinity matrix.
  - Find x vector that maximizes Total Score(S)

$$S = \sum_{a,b \in C} M(a,b) = x^T M x \qquad x^* = argmax(x^T M x)$$

- The score depends mainly on three things:
  - the number of assignments in the cluster
  - how interconnected the assignments are (number of links)
  - how well they agree (weights on the links)

- x* that will maximize the cluster score is the principal eigenvector of M.

Matching FGs

- Algorithm for finding best correspondences.

1. Build the symmetric non-negative n x n matrix M

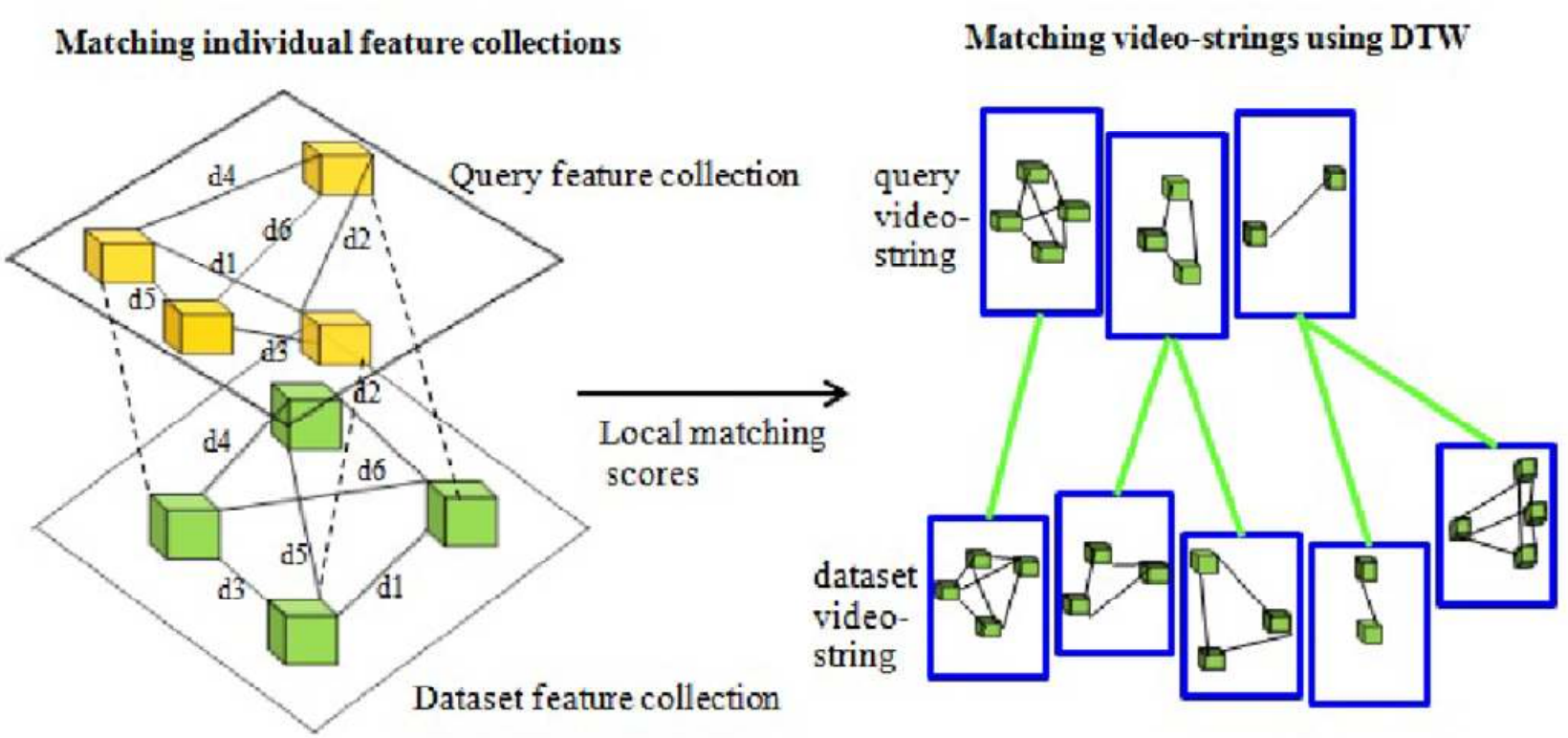
2. Let x* be the principal eigenvector of M. Initialize the solution vector x with the n x 1 zero vector. Initialize L with the set of all candidate assignments.

3. Find a* = argmax $a \epsilon L(x^*(a))$, If x *(a*) = 0 stop and return the solution x. Otherwise set x(a*) = 1 and remove a* from L.

4. Remove from L all potential assignments in conflict with a* = (i, i'). These are assignments of the form (i, k) and (q, i').

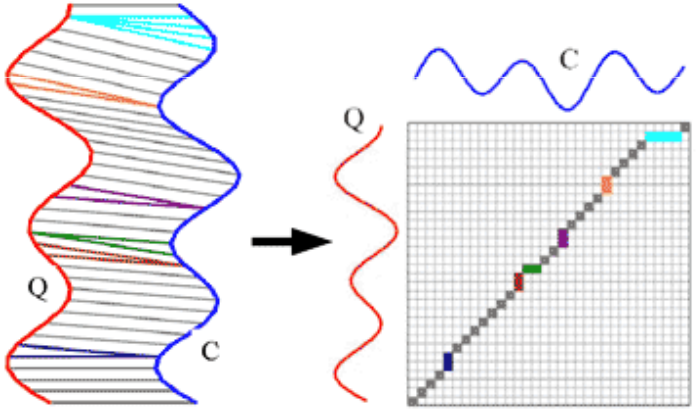
5. If L is empty return the solution x. Otherwise go back to step 3

Matching SFGs



Matching Video Strings using DTW

- DTW finds an optimal match between two sequences of feature vectors which allows for streached and compressed sections of the sequence.



$$sim(Q,P) = (x^*)^T \mathbf{M} x^*,$$

$$d(Q,P) = 1 - \frac{sim(Q,P)}{sim(Q,Q)}$$

Matching Video Strings using DTW

- *Monotonic condition*: the path will not turn back on itself, both the i and j indexes either stay the same or increase, they never decrease.
- *Continuity condition*: The path advances one step at a time. Both i and j can only increase by 1 on each step along the path.
- ***Boundary condition***: the path starts at the bottom left and ends at the top right.
- *Adjustment window condition*: a good path is unlikely to wander very far from the diagonal.
- *Slope constraint condition*: The path should not be too steep or too shallow. This prevents very short sequences matching very long ones.

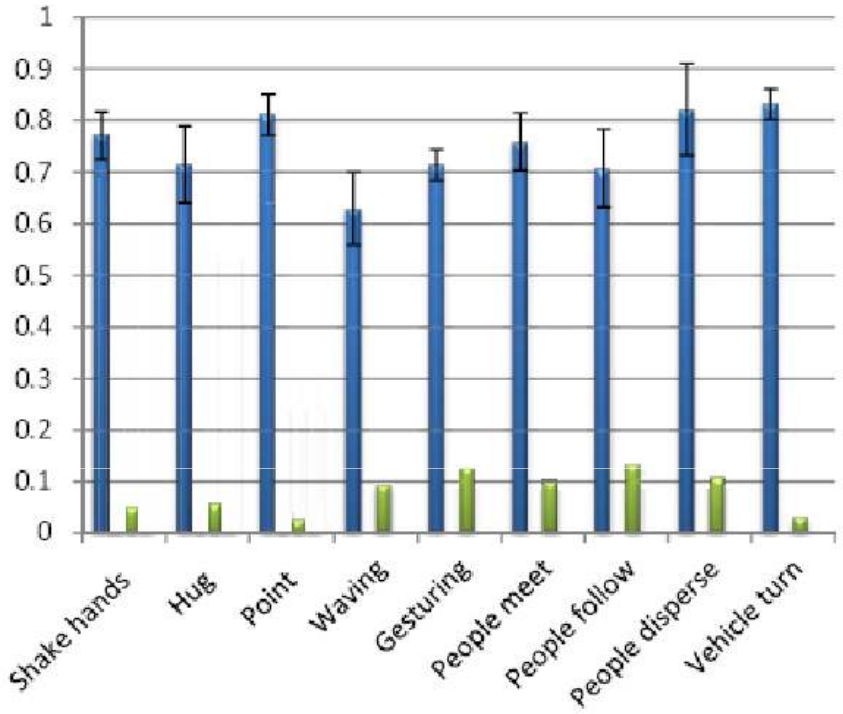
Experimental Results

- UT Interaction and UCR VideoWeb datasets were used in the activity recognition contest at ICPR 2010.
- UCR VideoWeb : very challenging due to the wide variation in he activities and clutter in scene.
- UT Interaction : activities across different background, scale and illumination.

UT Interaction Dataset

UCR VideoWeb Activity Dataset

Retrieval Results on UCR VideoWeb Dataset

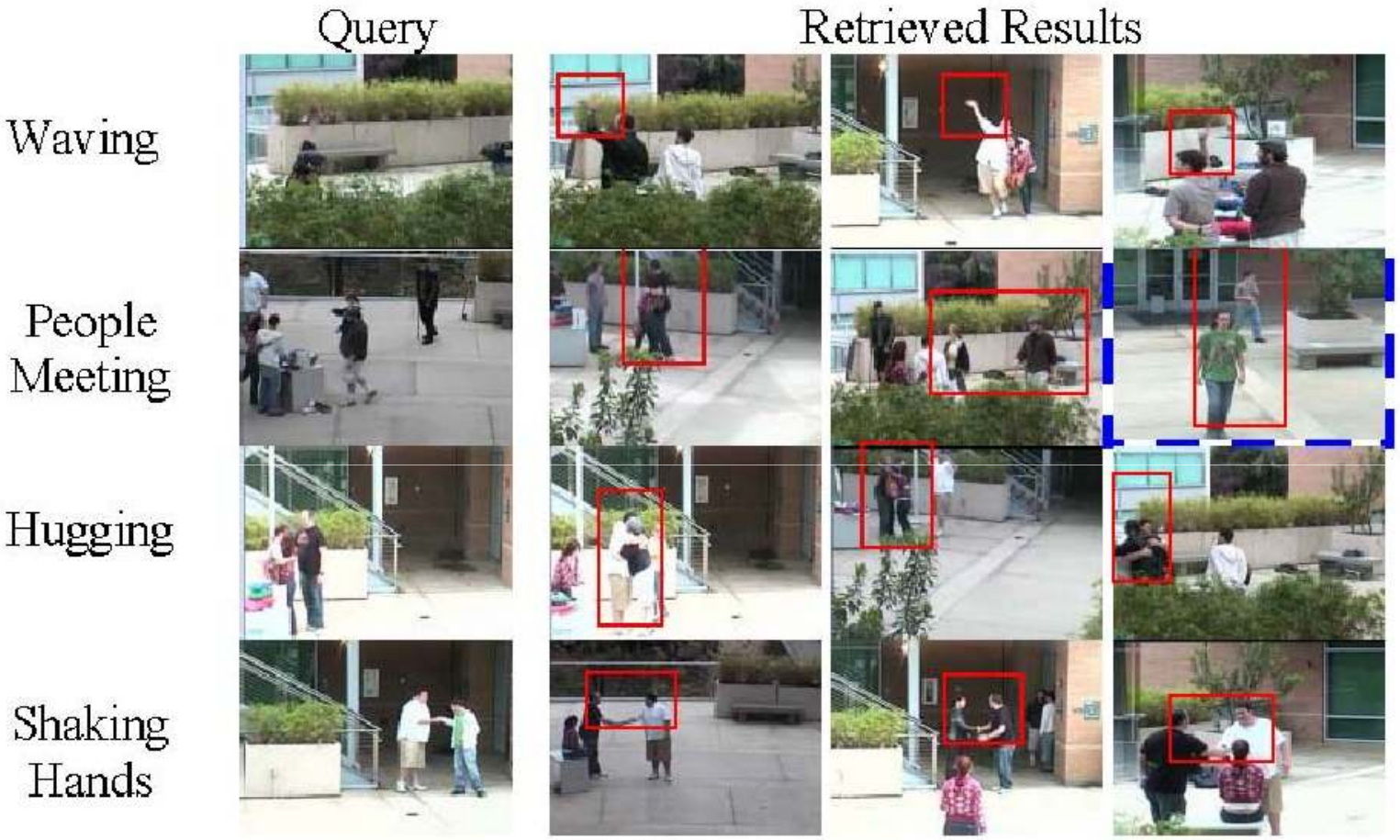


Recognition accuracy and false positives on 9 activities from UCR VideoWeb dataset in query–based retrieval framework. Standard deviation in performance(accuracy) for different queries is marked on the bars.

Retrieval Results on UCR VideoWeb Dataset

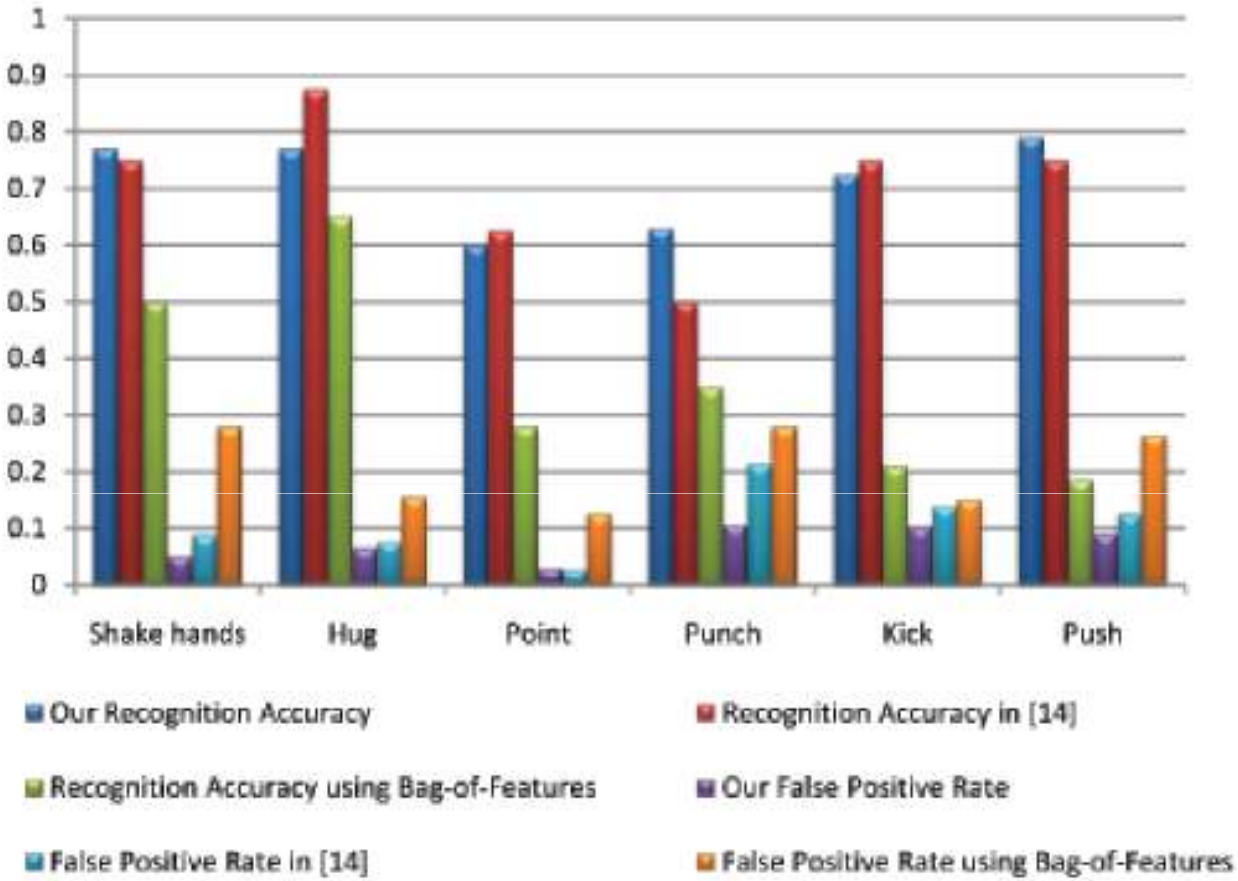
- Length of each segment is set to 20 frames.
- For each activity class, 3 random video clips are selected as queries.
- Recognition rates are better on activities that continue longer time periods.
- Short duration activities are more difficult to recognize.
- Short duration activities such as “pointing” has higher variablity.

Retrieval Results on UCR VideoWeb Dataset



The left column depitcs the query videos and the other columns are the best matches on UCR VideoWeb dataset.

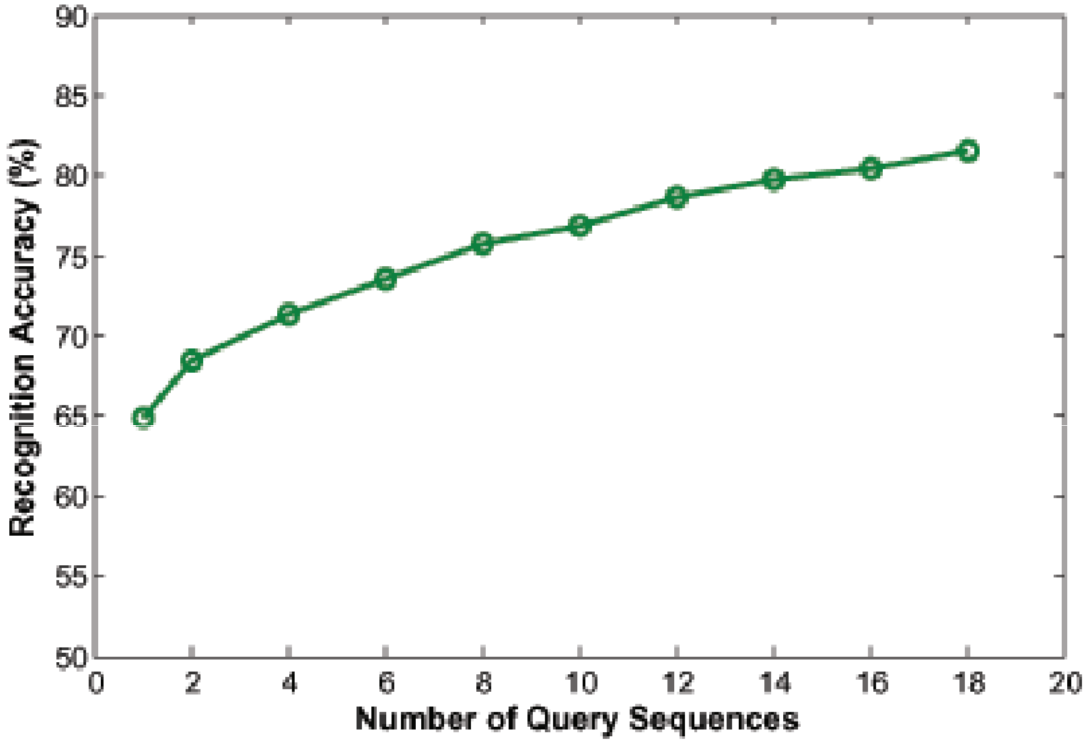
Results on the UT Interaction Dataset



2 videos are randomly selected as query videos and tested on 8 videos while the other methods use one for testing and 9 for training.

[14] M. S. Ryoo and J. K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In IEEE Conf. on Computer Vision and Pattern Recognition, 2006.

Results on the UT Interaction Dataset



Even with just one query, recognition accuracy of the proposed method is %65. This is the major difference with other methods.

Summary

- An activity modeling and matching approach that explicitly quantifies the spatial and temporal relationships between the objects.
- Motivated by graphical models and time sequence alignment.
- No training examples are required.
- Allows matching a short video to a large database through subsequence matching.

Questions

?

Thank you!