

Ad Hoc Video Query and Retrieval using Multi-modal Feature Graphs

CASIS Workshop
May 13th, 2015

Carmen Carrano
Doug Poland

 Lawrence Livermore
National Laboratory



LLNL-PRES-670430

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Motivation - Video is hard

- **Video query and retrieval against user-generated content is very challenging**
 - **Extreme dimensionality and infinite variability of video data**
 - Variability of object appearance
 - Interactivity of multiple objects with each other and with environment
 - Scenes variability with respect to the camera viewpoint both temporally and spatially.

Purpose

- **Perform ad-hoc query by example**
- **Want a quick way to interrogate a large corpus of videos using a rich, flexible and coupled set of appearance and motion features**
 - Image features
 - Motion features
 - Audio features (in progress)
 - Text/Tags (not currently used)
- **We don't want to have to compute distances between the features for each video clip to rank them**
 - Too time consuming at query time

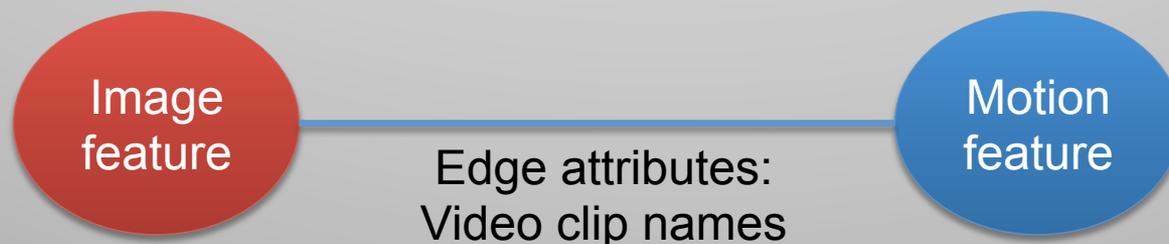
Dataset

- **YLIMED – Yahoo-Livermore-ICSI Multimedia Event Detection**
- **A subset of the YFCC100 videos (Yahoo Flickr Creative Commons 100 Million Dataset)**
- **Intended as an open source replacement for TRECVID MED challenge videos (15 target events), which are not completely open source.**
- **YLIMED has 10 target events (positives) defined along with lots of non-target events (negatives). They are extracted into test and train, positives and negatives.**
- **The initial prototyping work presented here deals with the two positives sets. ~1800 videos**

Ev101 Birthday Party	Ev106 Person Grooming an Animal
Ev102 Flash Mob	Ev107 Person Hand-Feeding an Animal
Ev103 Getting a Vehicle Unstuck	Ev108 Person Landing a Fish
Ev104 Parade	Ev109 Wedding Ceremony
Ev105 Person Attempting a Board Trick	Ev110 Working on a Woodworking Project

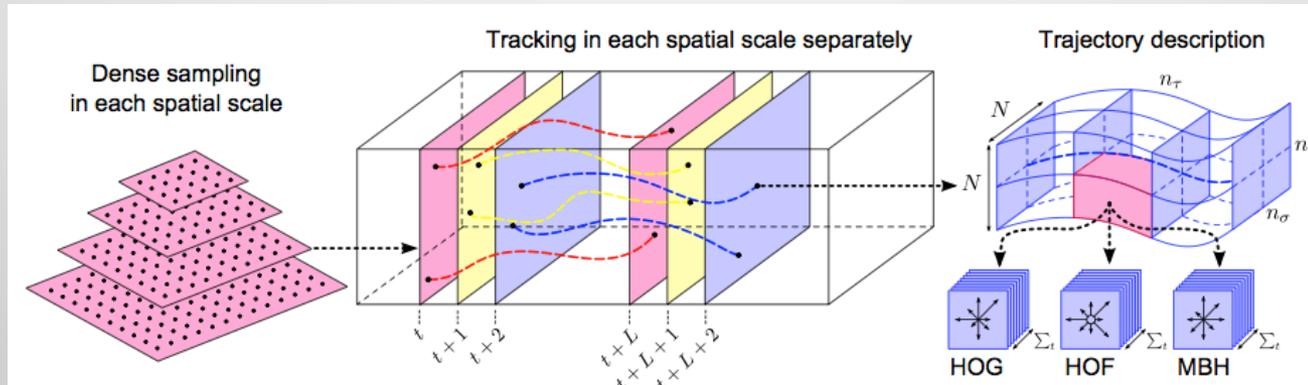
Experimenting with a bipartite graph based approach

- **Nodes: Multimodal Features**
- **Edges: Connect two nodes if a video clip has those two nodes in common**
- **Edge attributes: Video clips**
 - **Fixed to be 60 frame windows (~2 sec)**
 - The 1800 videos result in 35312 video clips



The Motion features

- **Dense Trajectory Features (DTF)** are computed over discretized volumes along many trajectories derived from dense optical flow

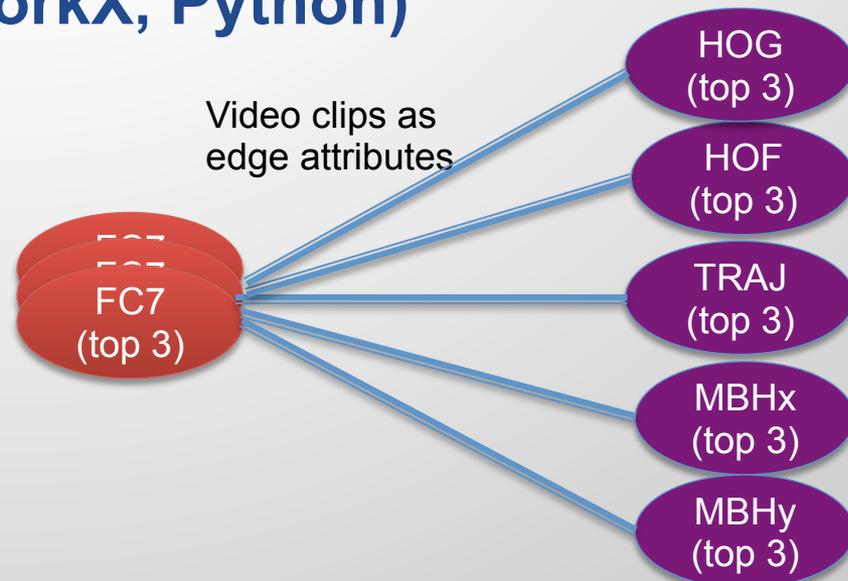


[Dense trajectories and motion boundary descriptors for action recognition](#) Heng Wang et. al. (INRIA/LEAR), IJCV 2013

- **Many DTF features are computed for any given 15 frame time slice**
 - Trajectory shape, HOG, HOF, Motion boundary histograms X and Y
 - Feature encoding is done by using a “bag-of-features” approach (K=2000 here)
- **Histograms of DTF features outperform other approaches on activity recognition**
- **We want *time-resolved* features** – we compute DTF histograms on 60 frame time windows (~2 sec)

Creating the graph (in NetworkX, Python)

- **Left nodes: Image features** – Top 3 ID's from the K-means clustered fc7 features with K=1000 from the few images that fall in the 60 frame window
- **Right nodes: Motion features** – Top 3 ID's from each of the 5 K-means clustered DTF motion (2K) histograms with K=256
- A fully connected bipartite graph is constructed for each video clip and merged to form the full graph; each edge has a list of clip (name, time offset) tuples as an attribute
- 140918 edges, 1986 nodes for this graph



Sample from graph: `G.edges(data=True)`

```
...
('MBHX_29', 'FC7_720',
 {'count': 2,
  'videos': [(('a6412aa127c562251cdc6c8ba23e753c' 60))]}),
('MBHX_29', 'FC7_547',
 {'count': 12,
  'videos': [(('9f152a6b2c6d83e9b44618e4fd88143', 60),
               ('9f152a6b2c6d83e9b44618e4fd88143', 900),
               ('9f152a6b2c6d83e9b44618e4fd88143', 1020),
               ('9f152a6b2c6d83e9b44618e4fd88143', 1380),
               ('9f152a6b2c6d83e9b44618e4fd88143', 1440),
               ('9f152a6b2c6d83e9b44618e4fd88143', 2400))]}),
...
```

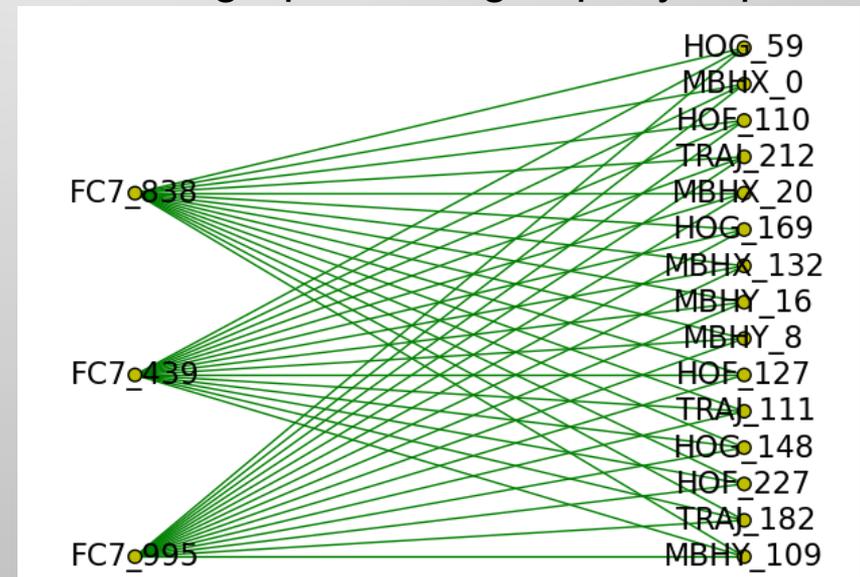
Querying the graph with a single clip

- Retrieve all clips that share one or more edges with the query clip
- Sort by number of shared edges

3 frames from query clip



Subgraph for single query clip



45 connected edges here

TOP 10 CLIPS

```
topclips1_f7d[0:10]
```

```
[('accf27cbcee9c21098cfbd021d75ad4', 60), 45),  
 (('accf27cbcee9c21098cfbd021d75ad4', 360), 21),  
 (('14b051f6784d447bb6898b5756f649', 360), 15),  
 (('14b051f6784d447bb6898b5756f649', 240), 15),  
 (('14b051f6784d447bb6898b5756f649', 780), 15),  
 (('14b051f6784d447bb6898b5756f649', 300), 15),  
 (('14b051f6784d447bb6898b5756f649', 600), 12),  
 (('accf27cbcee9c21098cfbd021d75ad4', 300), 12),  
 (('14b051f6784d447bb6898b5756f649', 660), 9),  
 (('2f554192f275b36a6d3fdab12619749', 60), 8)]
```

Results from single-clip “board-trick” query

- Showing top 18 clips with 3 images per clip
- How to judge results quantitatively?
 - We have “event” ground-truth information
 - Can “score” the top N clips using this

Query Event is Ev105	Event Scoring by Clip	Event Scoring by Video
Top 20	'Ev105': 19 'Ev107': 1 95% correct	'Ev105': 6 'Ev107': 1 86%
Top 50	'Ev105': 42 Others: 8 84%	'Ev105': 19 Others: 7 73%
Top 100	'Ev105': 75 'Ev103': 10 Others: 15 75%	'Ev105': 32 'Ev103': 7 Others: 13 61%



Results from single clip “grooming animal” query

- Showing top 18 clips with 3 images per clip
- How to judge results quantitatively?
 - We have “event” ground-truth information
 - Can “score” the top N clips using this



Query Event is Ev106	Event Scoring by Clip	Event Scoring by Video
Top 20	'Ev106': 19, 'Ev107': 1 95% correct	'Ev106': 10, 'Ev107': 1 91%
Top 50	'Ev106': 37, 'Ev107': 10 Others: 3 74%	'Ev106': 14, 'Ev107': 8 Others: 3 56%
Top 100	'Ev106': 71, 'Ev107': 18, Others: 11 71%	'Ev106': 22, 'Ev107': 12 Others: 10 50%

Going beyond single clip queries

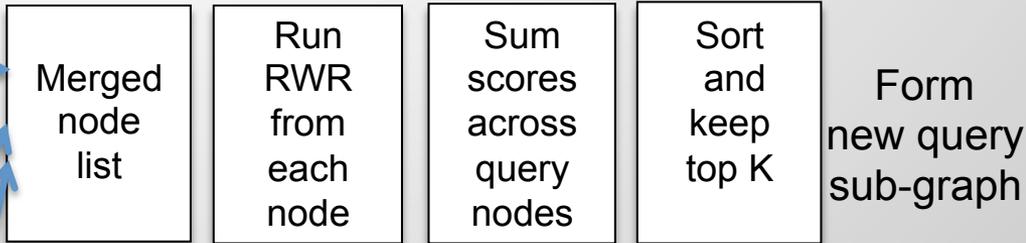
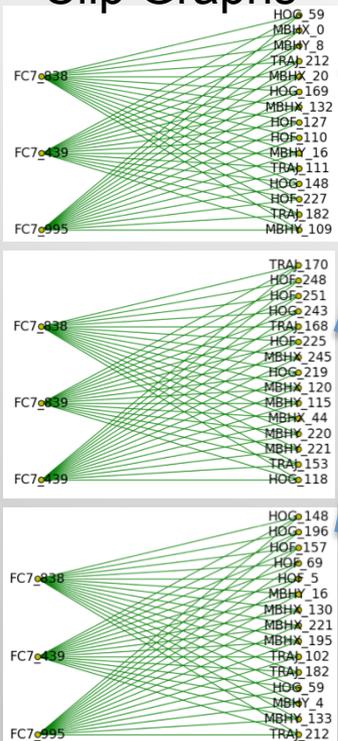
- Want to combine properties of multiple video clips and strengthen the results for a particular query (aka relevance feedback)
 - Can go more diverse or we can go more specific depending on how the query is constructed and what the user wants to do
- In order to extract and prioritize related results from multiple queries, one approach is to use a random walk (with restart) sub-graph (RWSG) approach initialized with multiple clips

RWSG approach by example with 3 different board-trick clips

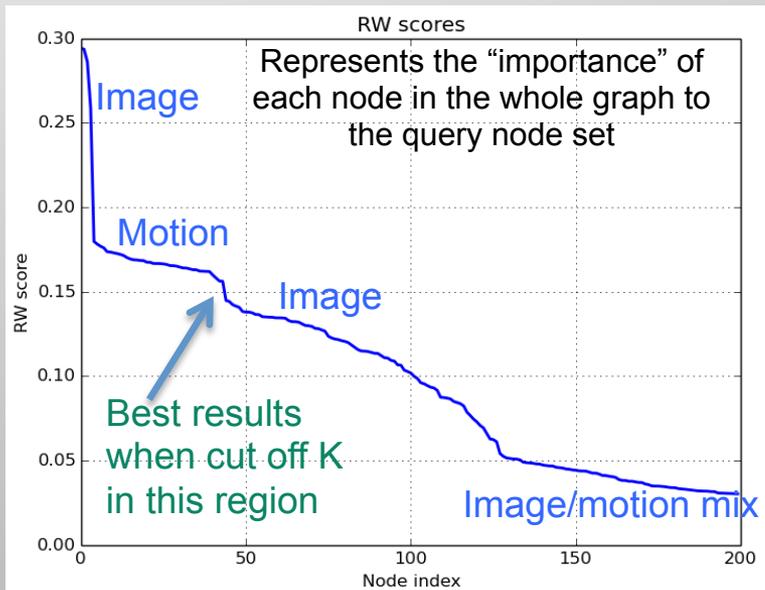
Clips



Clip Graphs



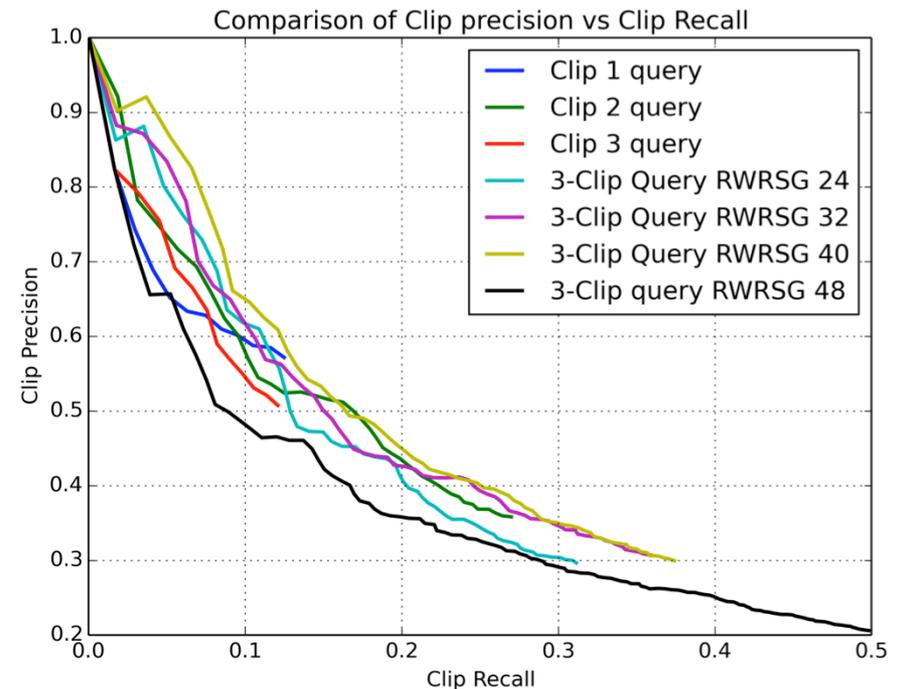
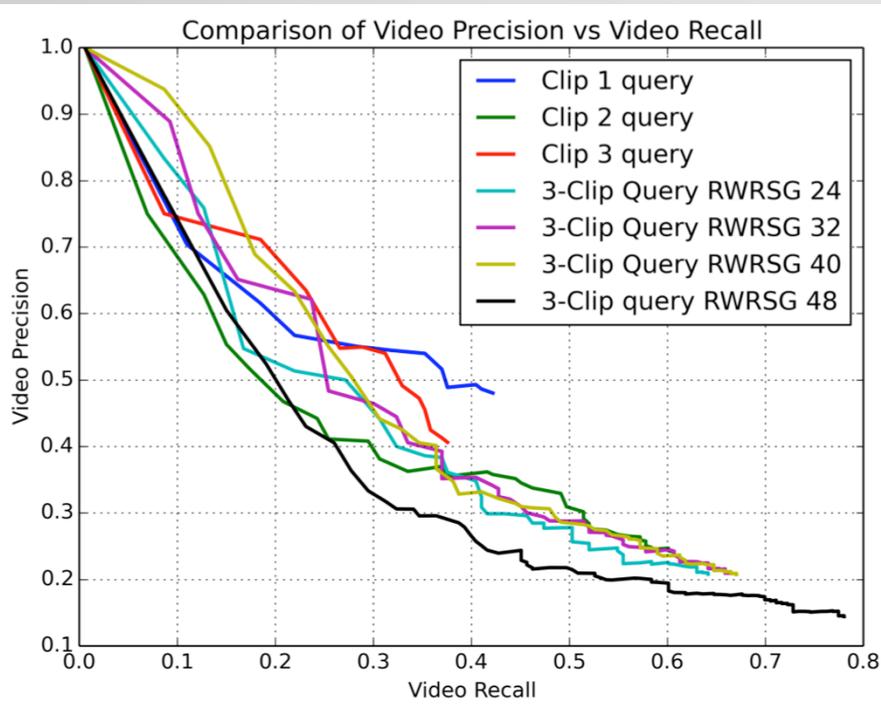
Summed RW scores (top 200)



- [(0.272, 'FC7_839'),
- (0.271, 'FC7_439'),
- (0.267, 'FC7_838'),
- (0.246, 'FC7_995'),
- (0.193, 'HOF_127'),
- (0.188, 'HOF_110'),
- (0.187, 'HOF_5'),
- (0.181, 'HOF_248'),
- (0.181, 'MBHX_221'),
- (0.181, 'MBHY_115'),
- (0.180, 'TRAJ_170'),
- (0.179, 'HOF_69'),
- ...
- (0.168, 'MBHY_4'),
- (0.167, 'TRAJ_182'),
- (0.167, 'MBHY_220'),
- (0.167, 'MBHX_44'),
- (0.166, 'HOG_196'),
- (0.166, 'MBHY_8'),
- (0.165, 'MBHY_109'),
- (0.164, 'MBHY_16'),
- (0.162, 'MBHX_195'),
- (0.162, 'HOG_169'),
- (0.158, 'HOG_59'),
- (0.122, 'FC7_411'),
- (0.121, 'FC7_618'),
- (0.120, 'FC7_506'),

Precision vs Recall – 3-clip board-trick query

- With the right node count choice, RWSG gives us improved clip precision/recall compared to any single clip
- The RWSG 48 nodes plot is telling us that 48 includes too many irrelevant nodes for this event type

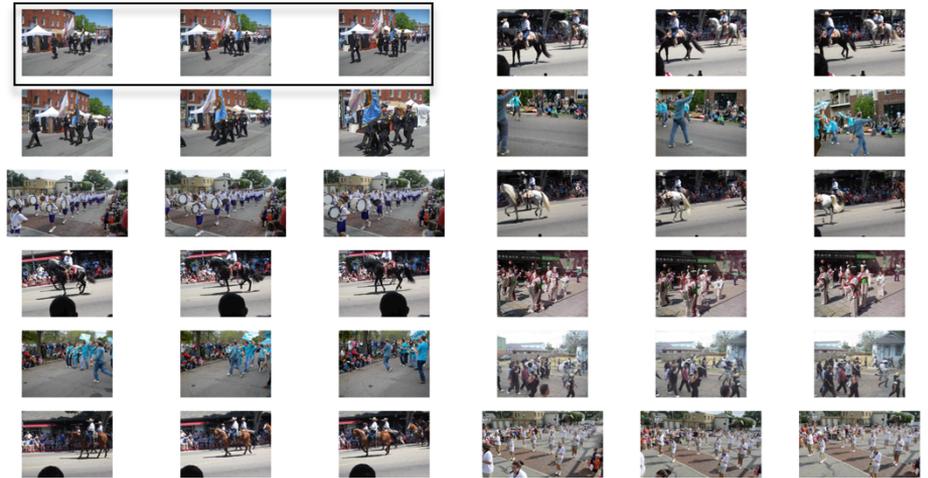


Parade query: Top 10 clips from 3 different parade query clips for submission to RWSG algorithm

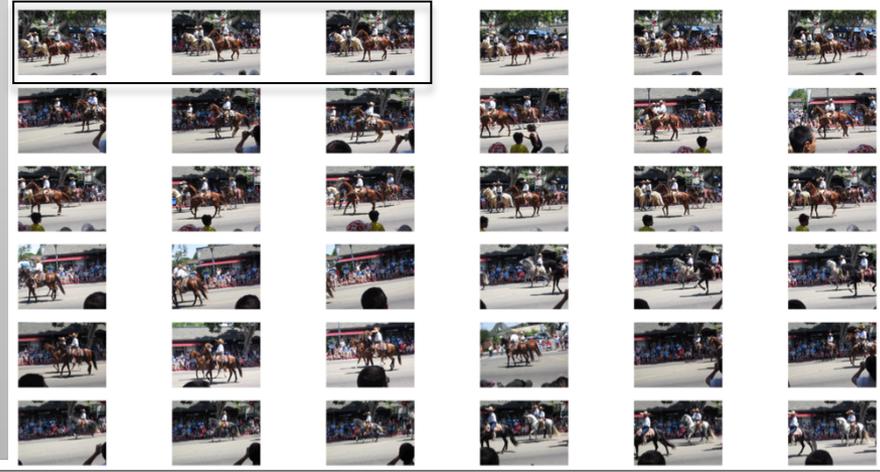
('529fa2a96561969cc71d403d1c256049', 60, 'parade1')
fc7-dtf 1-clip edge query



('a6204677ac60c5c7f9687befac275c', 0, 'parade2')
fc7-dtf 1-clip edge query

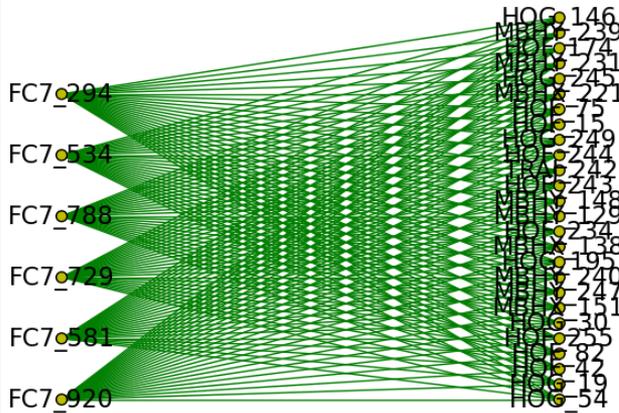


('967467612851e5d0eb5b8ff9eabde17b', 600, 'parade_horse')
fc7-dtf 1-clip edge query

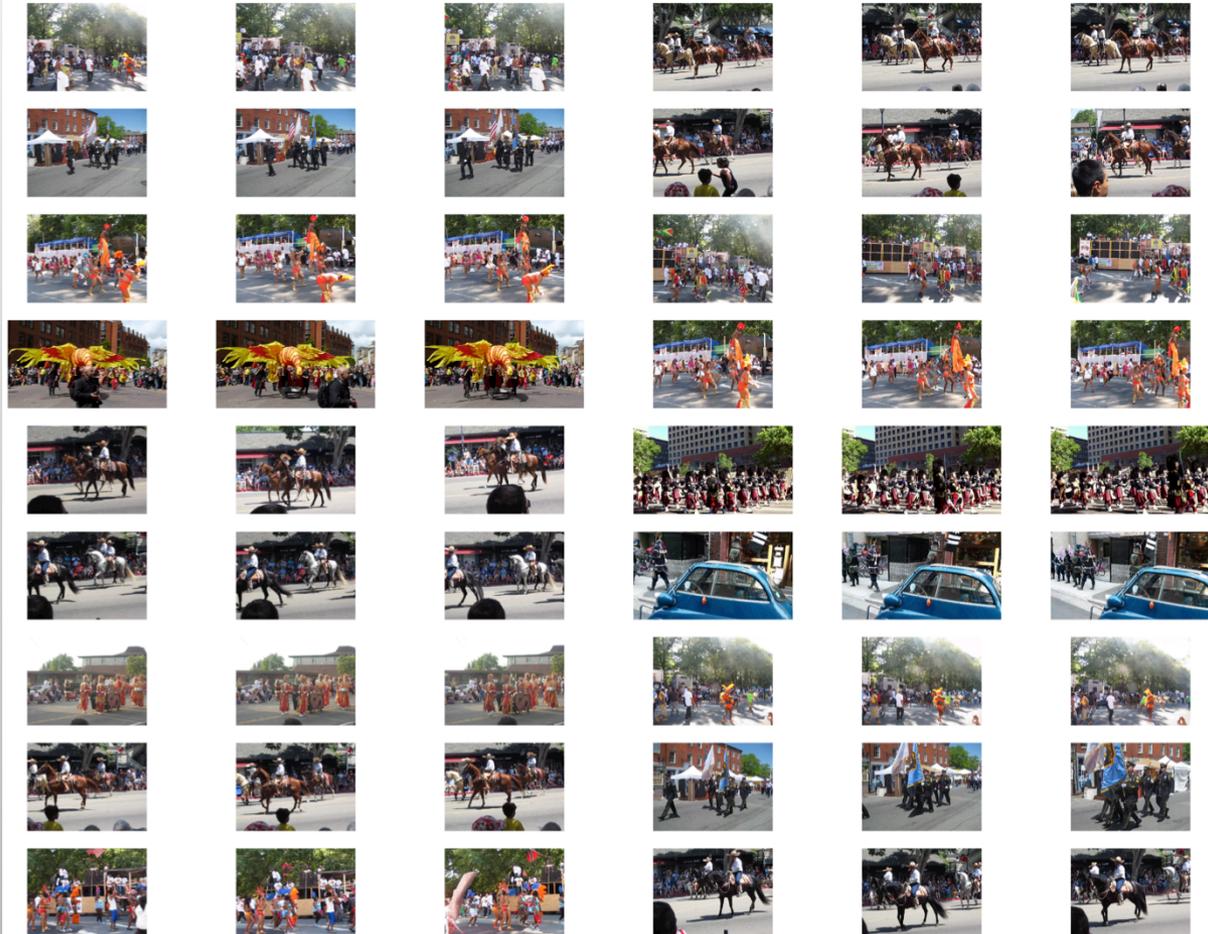


Parade – Result of 3-clip RWSG query, K=32

Subgraph: parade1_and_parade2_and_parade_horse
fc7-dtf RWSIM query seeded by 3-clip query, K=32



parade1 + parade2 + parade_horse
fc7-dtf RWSIM query seeded by 3-clip query, K=32

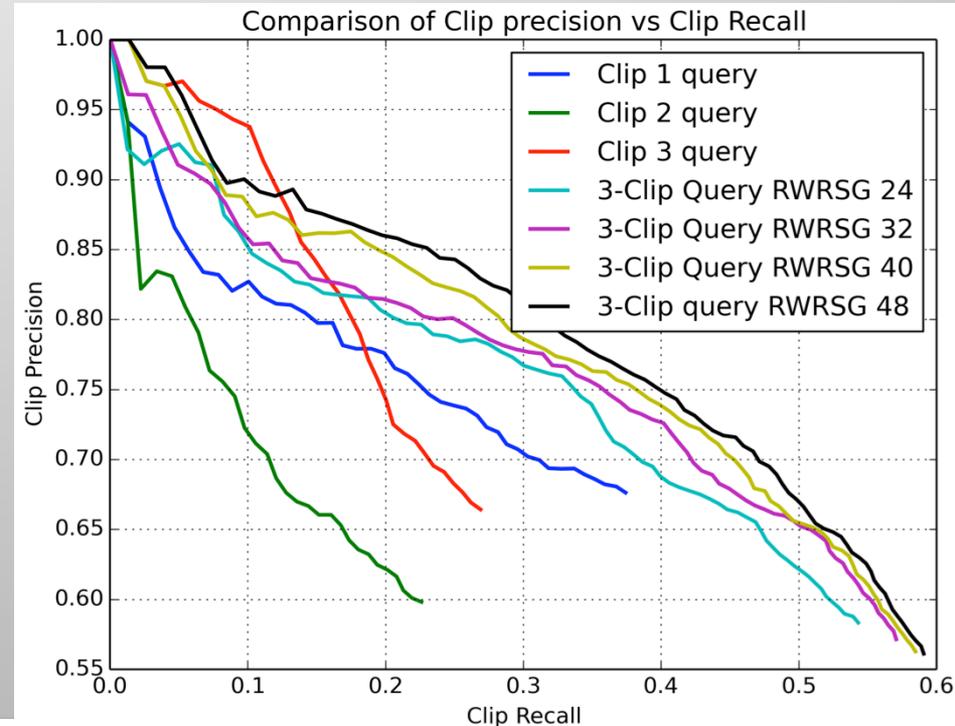
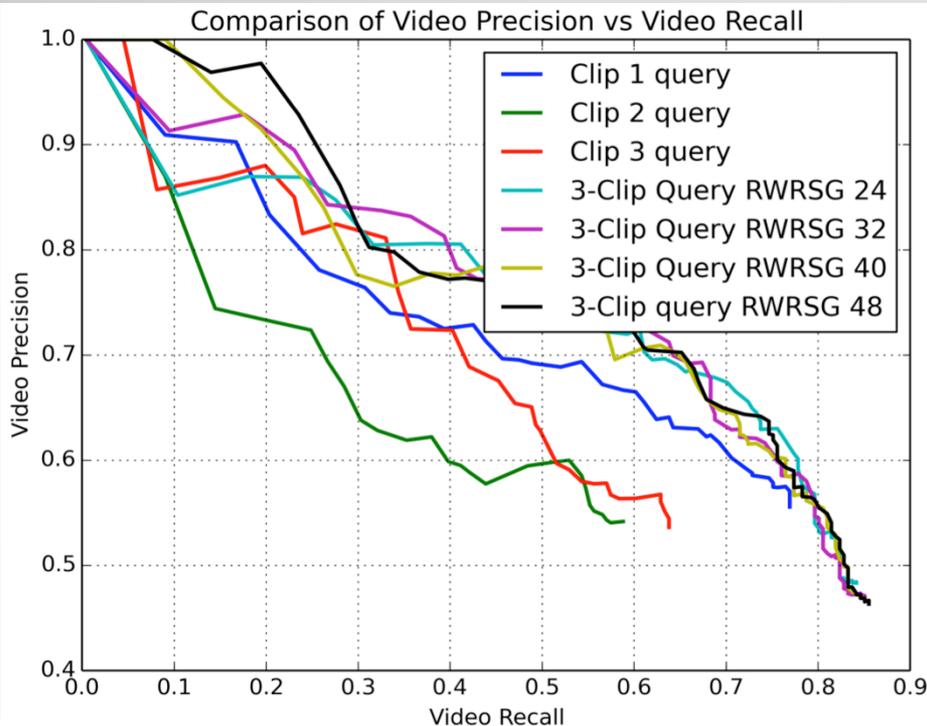


Query Event is Ev104	Event Scoring by Clip	Event Scoring by Video
Top 20	'Ev104': 20 100% correct	'Ev104': 11 100%
Top 50	'Ev104': 48 96%	'Ev104': 20 91%
Top 100	'Ev104': 96 'Ev102': 4 96%	'Ev104': 39 'Ev102': 3 93%

Event #102 is a flashmob, which may look very similar

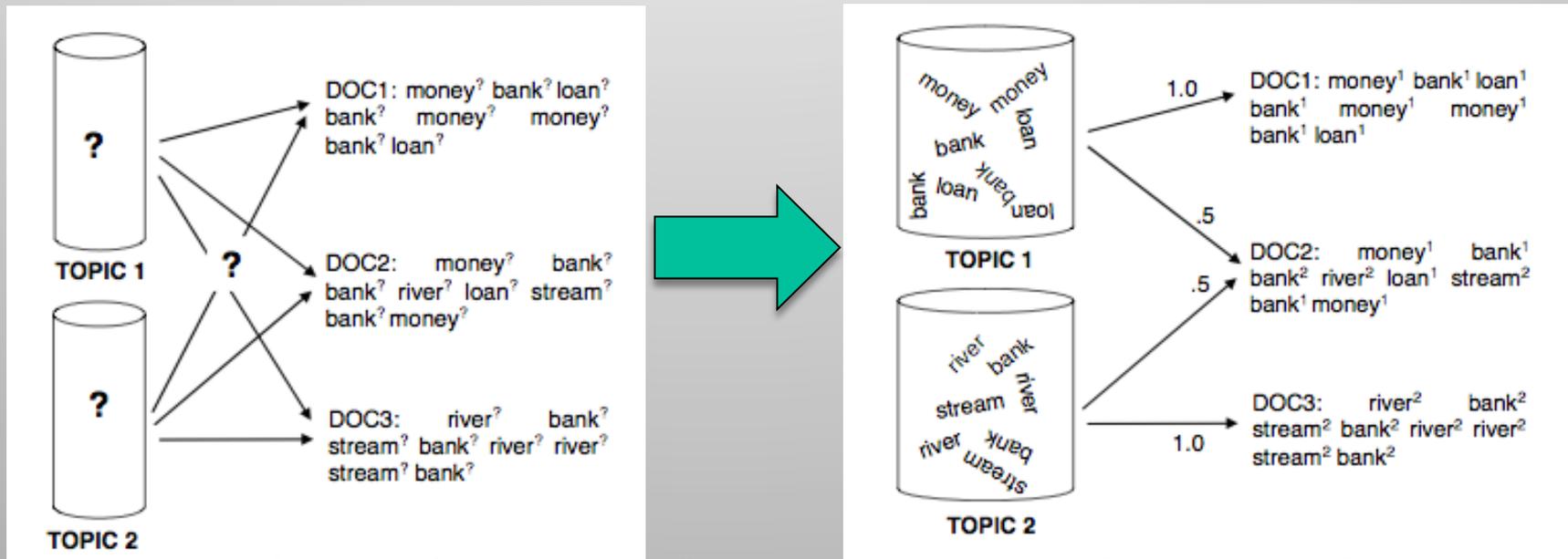
Precision vs Recall – 3-clip parade query

- RWSG with sufficient nodes mostly dominates precision/recall compared to any single clip
- The RWSG-48 nodes case is okay here – “parade” is more diverse event, less well defined



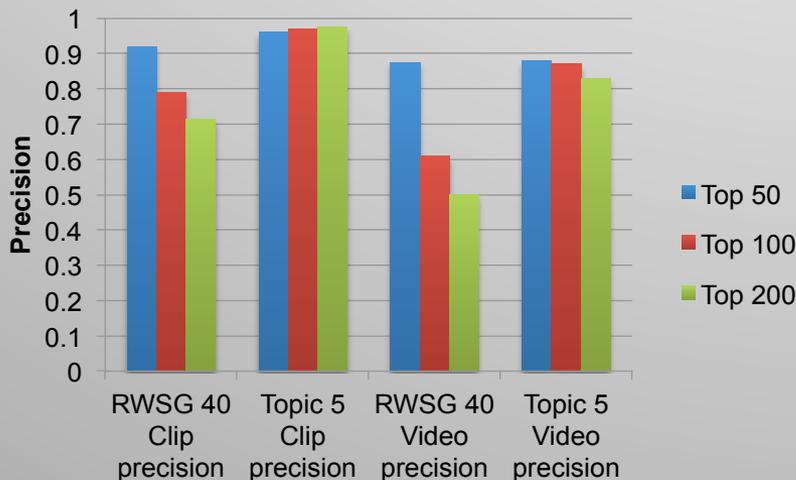
Going beyond the RWSG: Organization into subtopics

- Further organize resulting clip set from RWSG into proposed groups with unsupervised, mixed-membership Latent Dirichlet Allocation (LDA) topic modeling on the associated multimodal graph edges
- Treat each clip as a bag of words (edges)
 - A word is composed of the edges formed from the names of the node pairs.
 - (e.g. FC7_534_TRAJ_144, FC7_788_HOF_234, etc.)
- Use Java application MALLET to do the LDA



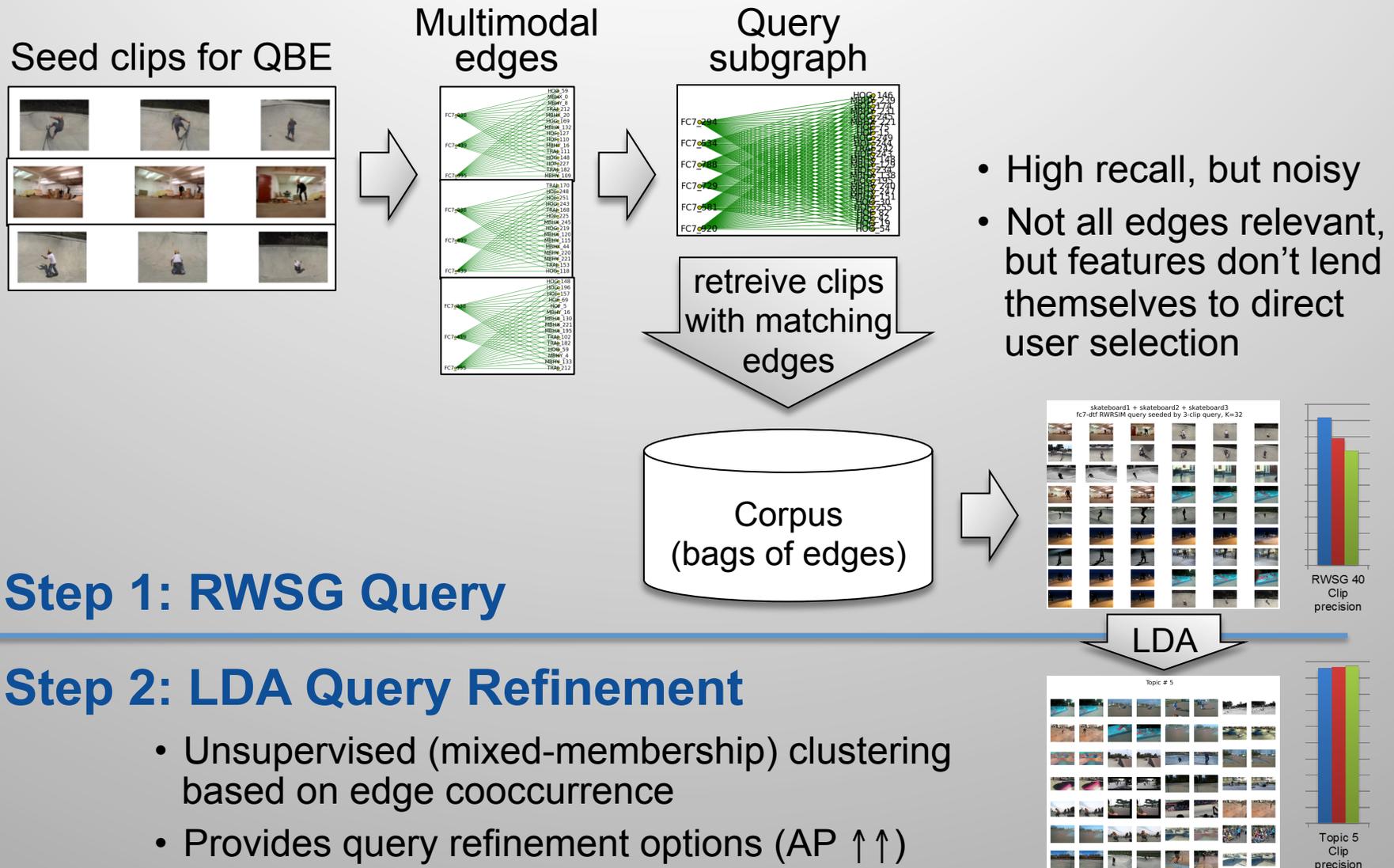
RWSG vs Topics – Board trick

- Feed all the clips returned by RWSG-40 to LDA
- Topic #5 in a 10 topic LDA clustering scored highest
 - Yields higher clip and video precision than the RWSG ordering alone especially in the top 100 or more clips



Now showing only 2 images per clip in order to increase # of videos to display

What's going on here?



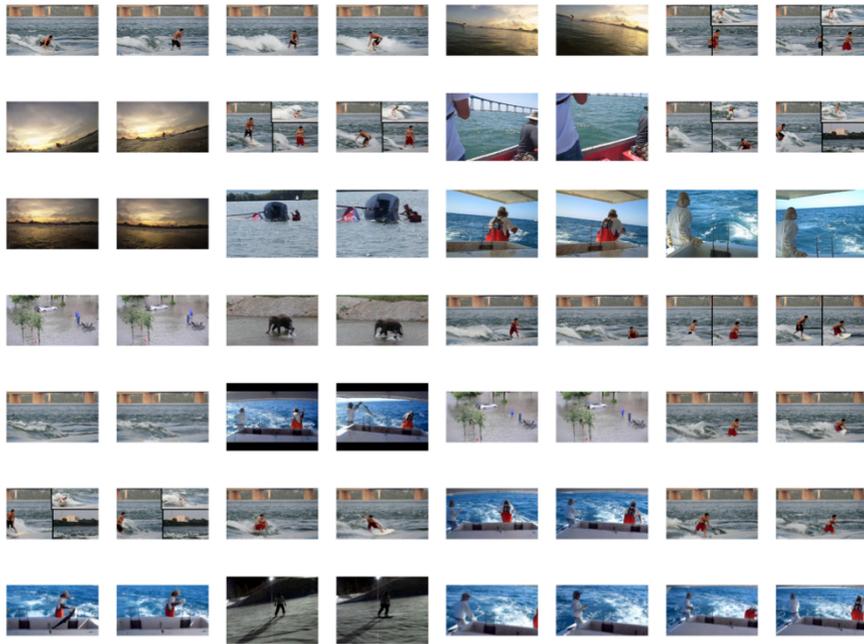
Step 1: RWSG Query

Step 2: LDA Query Refinement

- Unsupervised (mixed-membership) clustering based on edge cooccurrence
- Provides query refinement options (AP ↑↑)

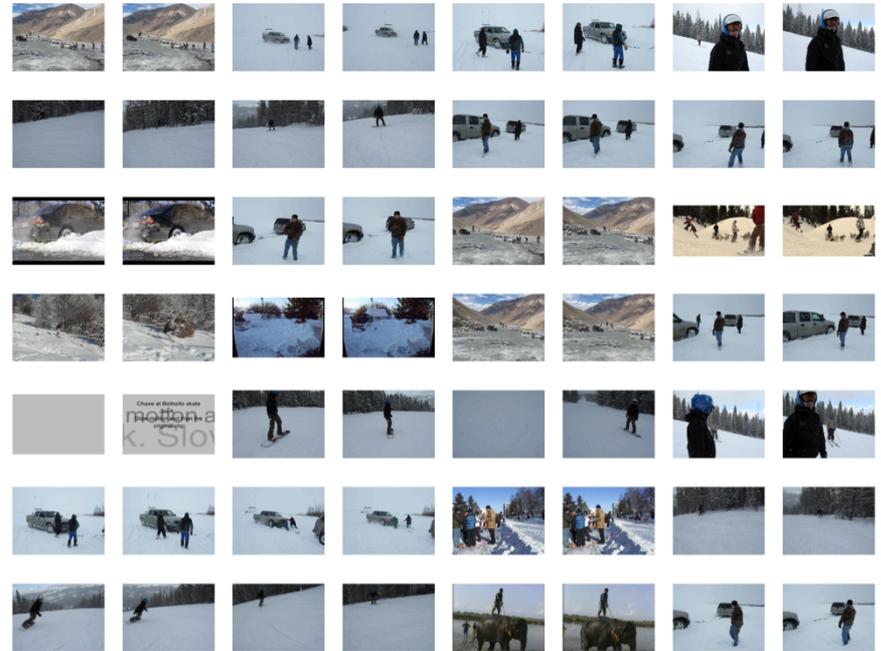
Other sample “topics” from the board trick clip set (using 20 topic LDA)

Topic # 8



Dominated by water

Topic # 16

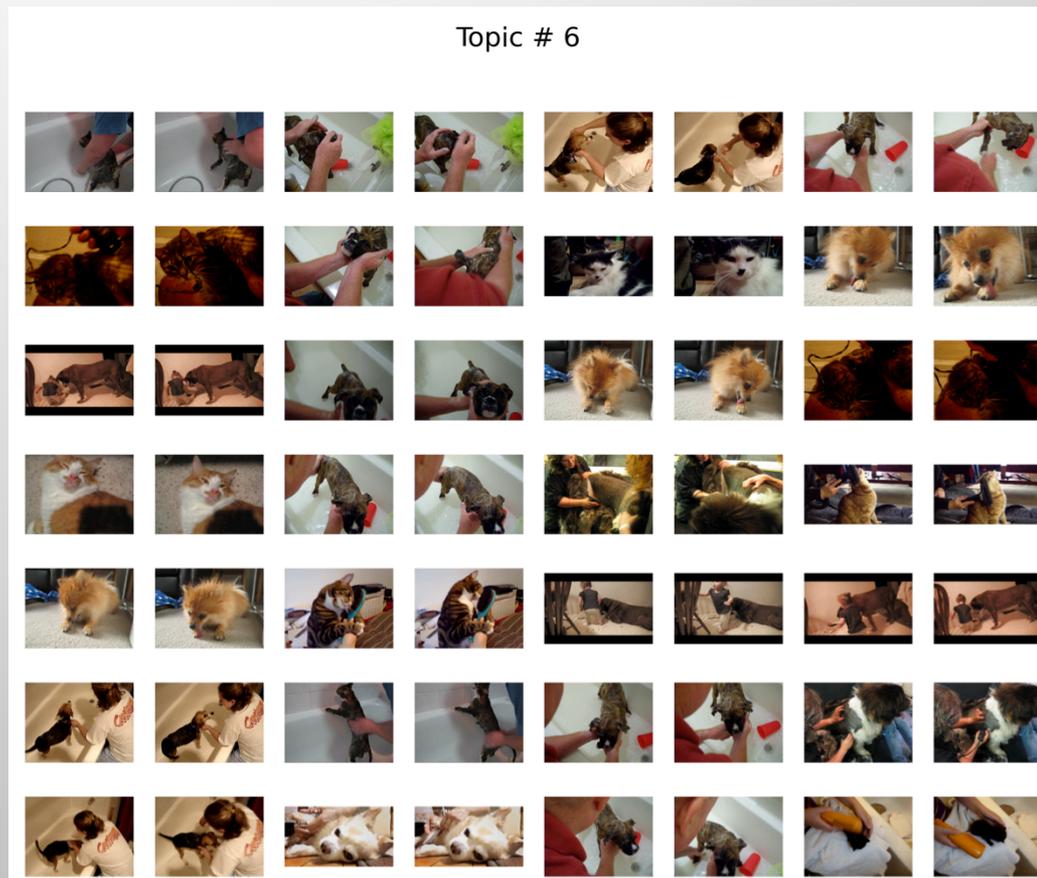
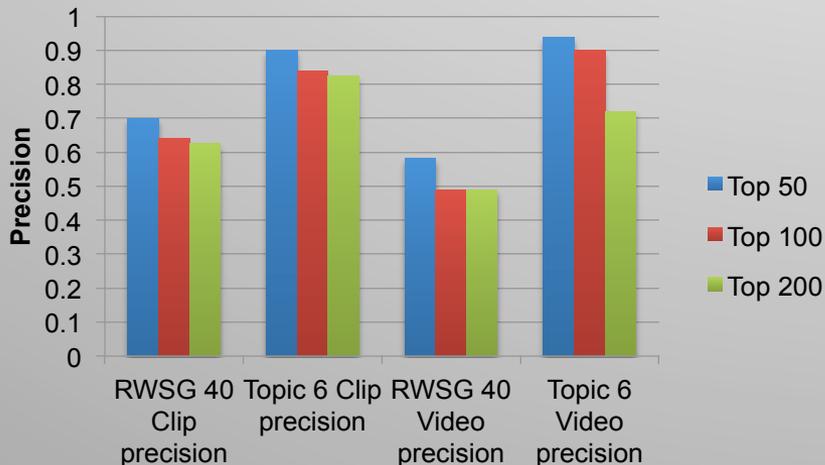


Dominated by snow

Note: There are 194 “board-trick” videos with a ratio of 6:3:1 skate:snow:surf type (only 173 in my graph - some bad videos in set)

RWSG vs Topics – Grooming animal

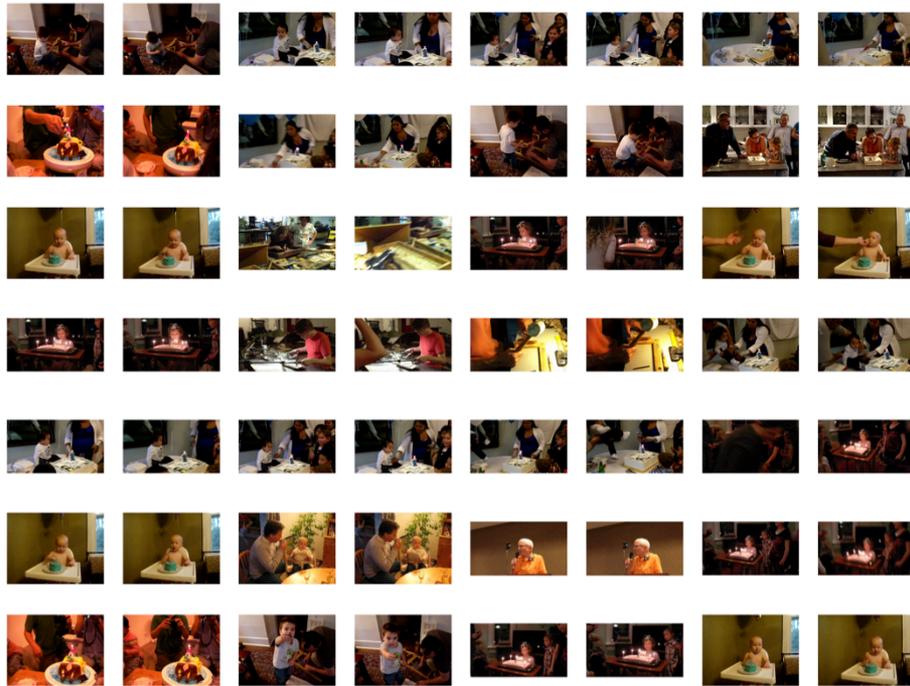
- Feed all the clips returned by RWSG-40 to LDA
- Topic #6 in a 10 topic LDA clustering scored highest for the grooming animal event
- Gives us much higher clip and higher video precision than the RWSG ordering alone



Focused on grooming small animals in a sink or on a table

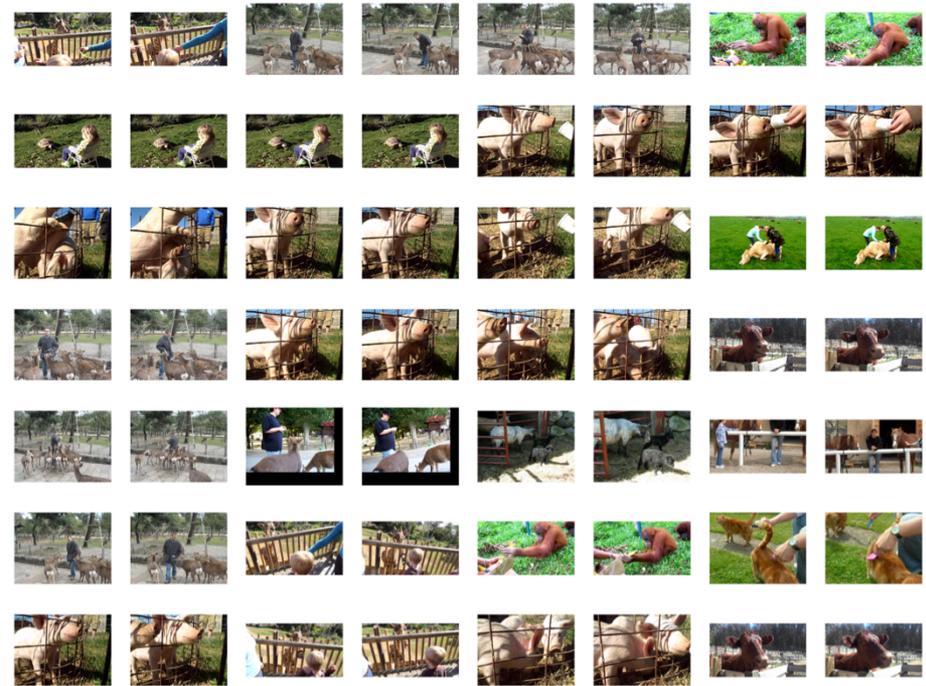
Other sample “topics” from the grooming animal clip set (using 20 topic LDA)

Topic # 1



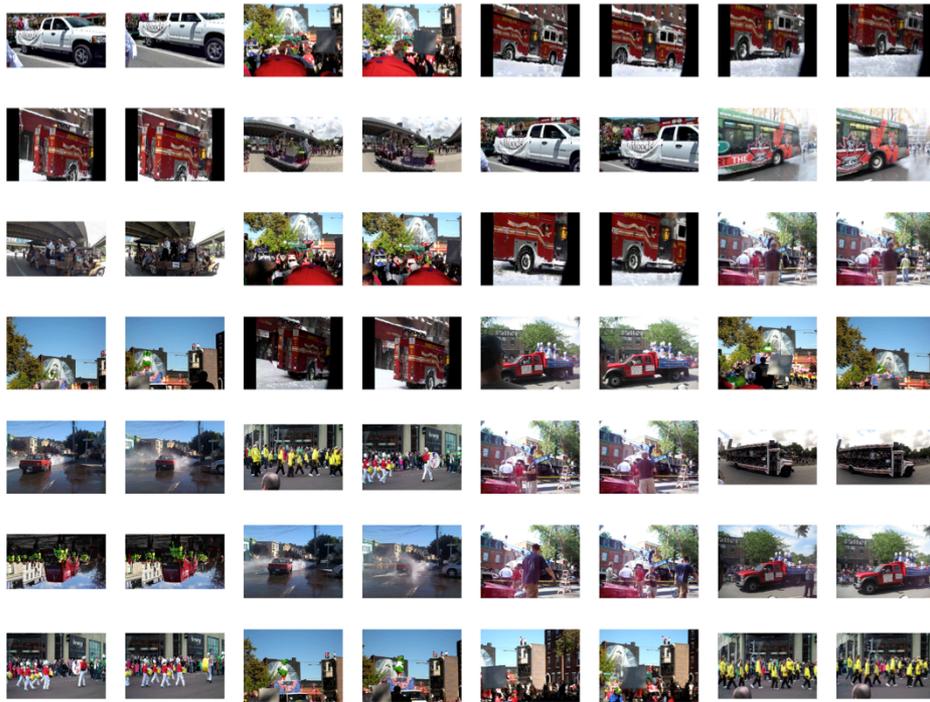
Feeding young human animal
(birthday party event, really)

Topic # 15

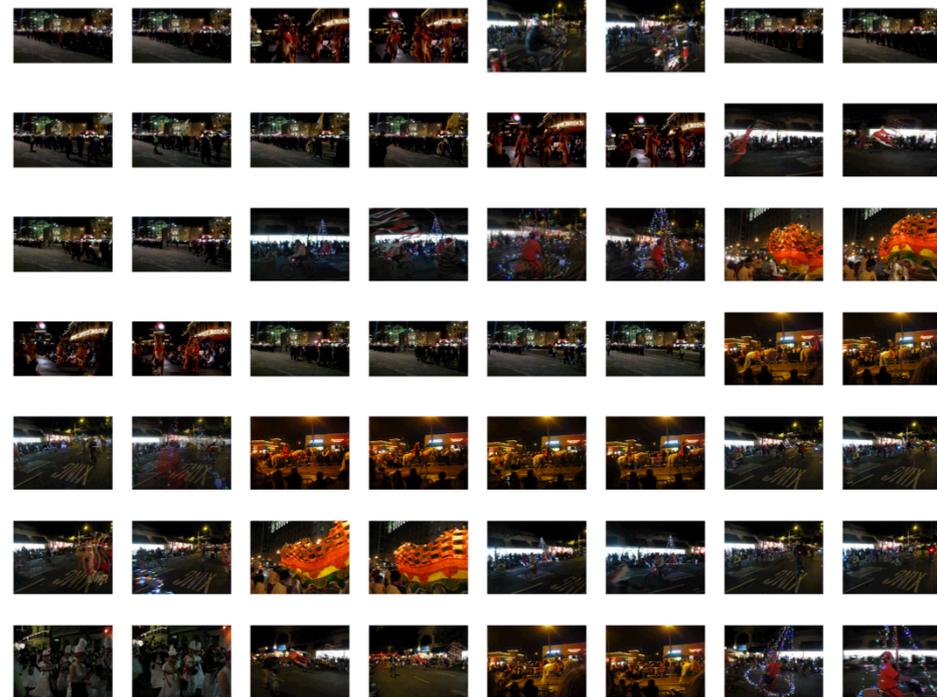


Grooming/feeding animals outside

Some sample “parade” topics seeded by an RWSG-40 clip set



Mostly parades with vehicles



Mostly parades at night

More sample Parade topics



Mostly parades with marching bands



Parades from afar with that big tree/bush thing

Summary

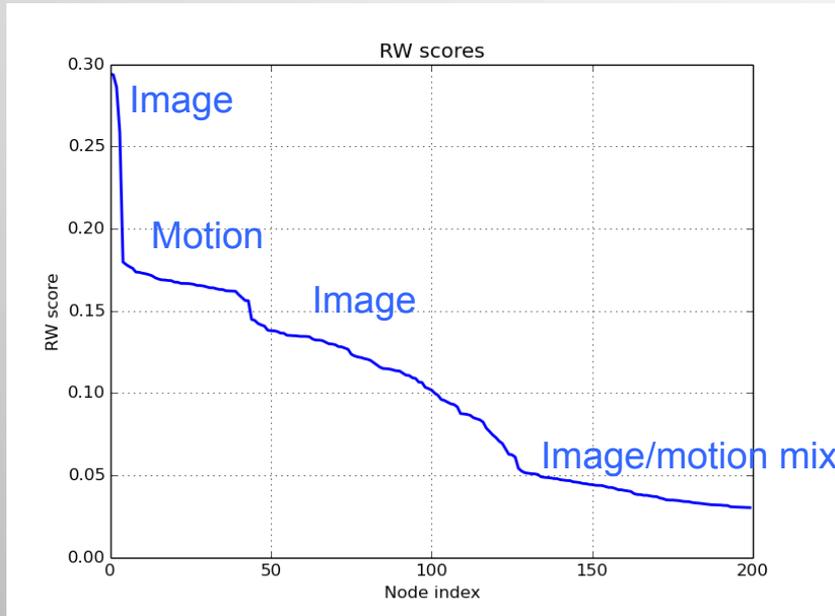
- A graph-based approach can be successfully used to perform ad-hoc multi-modal video retrieval
- Graph-based approaches can be employed to combine multiple clips into a unified, more relevant result set
- LDA can be used to steer and refine a large set of query results by grouping results into possible topics of interest.
- Future work will include more modes, exploration of alternate graph representations and algorithms, GUI-ification of workflow, etc.

Acknowledgements

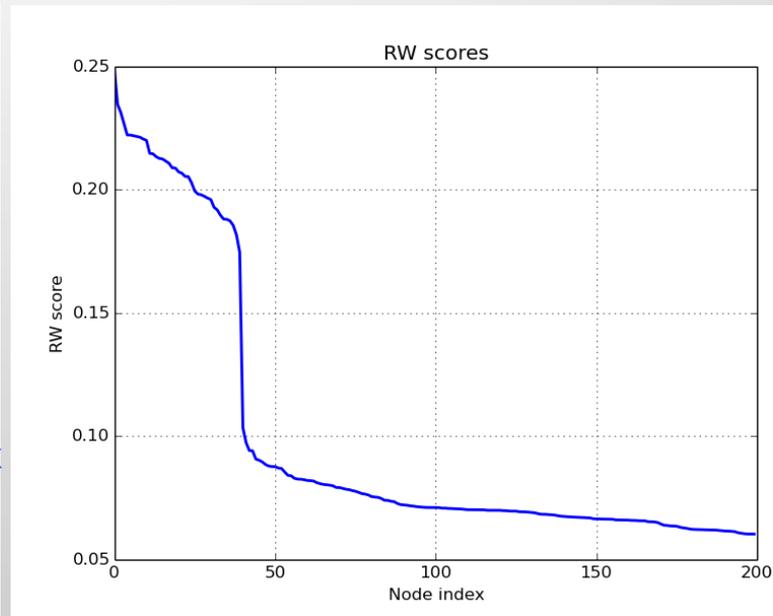
- Karl Ni for the fc7 features
- Jim Brase for the python random walker routine and probably helping make sure this Video LDRD was funded...
- LC for catalyst, edge and surface

Backups

Graph design is important: Effect of DTF cluster size on the 3-clip board-trick random walk scores— 1024 and 64



1024 very similar to 256



64 bumps up the motion importance because if a motion node is larger it will be visited more often, hence RWR will think it is more important

[(0.249, 'HOF_36'),
 (0.234, 'HOF_33'),
 (0.231, 'HOF_4'),
 (0.226, 'HOF_15'),
 (0.222, 'HOF_1'),
 (0.222, 'HOF_0'),
 (0.221, 'FC7_439'),
 (0.221, 'MBHY_7'),
 (0.221, 'FC7_839'),
 (0.220, 'HOF_47'),
 (0.219, 'FC7_838'),
 (0.214, 'FC7_995'),
 (0.214, 'MBHY_32'),
 (0.213, 'TRAJ_30'),
 (0.212, 'HOG_39'),
 (0.212, 'HOF_49'),
 (0.211, 'HOG_2'),
 (0.210, 'TRAJ_1'),
 (0.208, 'MBHX_32'),
 (0.208, 'HOG_13'),
 (0.207, 'MBHY_60'),
 (0.206, 'MBHY_1'),
 (0.205, 'TRAJ_4'),
 (0.205, 'MBHX_11'),
 (0.202, 'MBHY_14'),
 (0.199, 'HOG_54'),
 (0.198, 'TRAJ_63'),
 (0.197, 'TRAJ_55'),
 (0.197, 'MBHY_17'),
 (0.196, 'MBHX_2'),
 (0.195, 'MBHX_56'),
 (0.192, 'HOG_51'),
 (0.191, 'MBHX_36'),
 (0.189, 'HOG_15'),
 (0.188, 'MBHX_7'), ...



**Lawrence Livermore
National Laboratory**