DVMM lab
digital video | multimedia laboratary

Concept Detection: Convergence to Local Features and Opportunities Beyond

Shih-Fu Chang[1], Junfeng He[1], Yu-Gang Jiang[1,2], Elie El Khoury[3], Chong-Wah Ngo[2], Akira Yanagawa[1], Eric Zavesky[1]

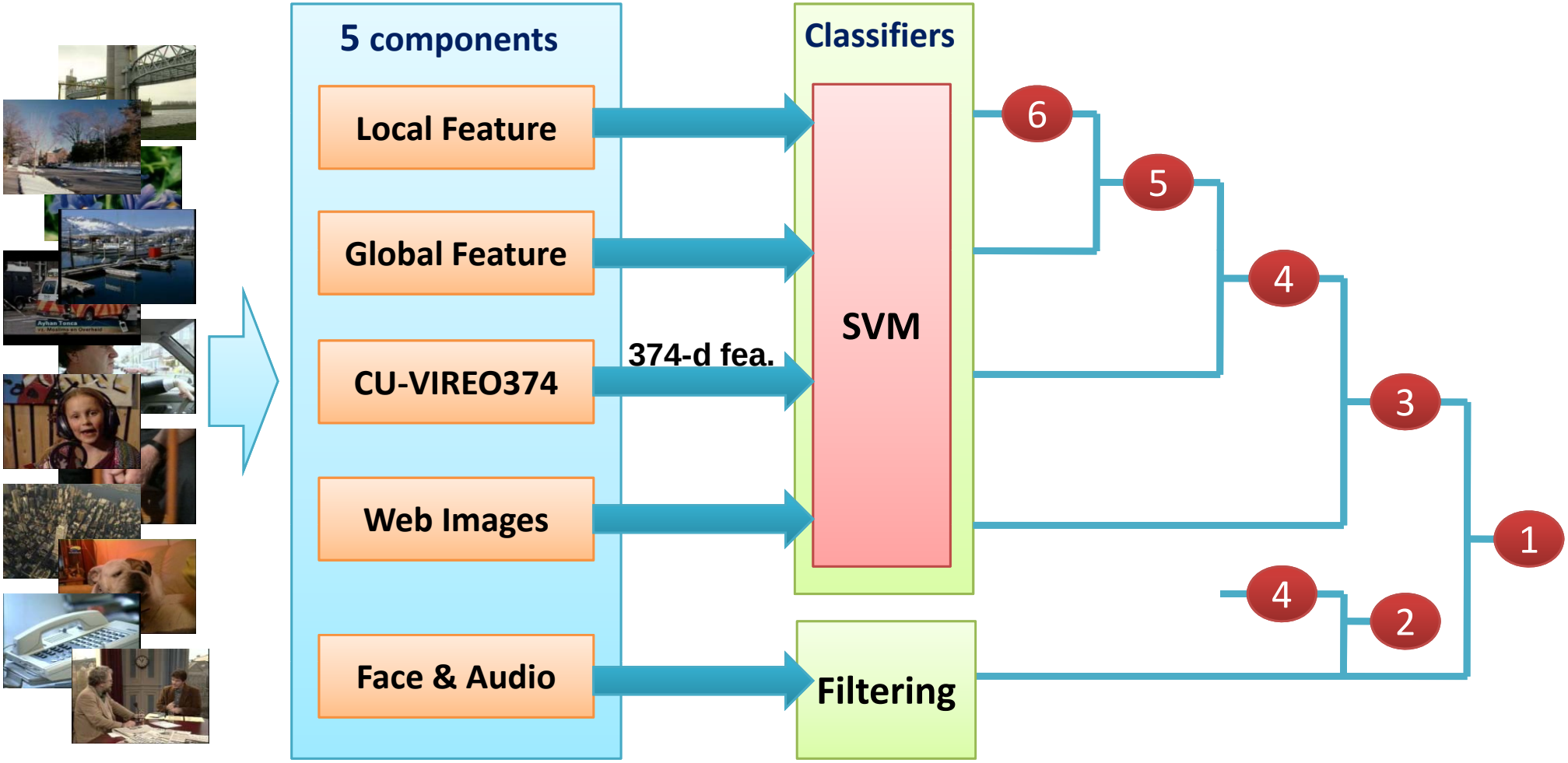
[1] DVMM Lab, Columbia University

[2] City University of Hong Kong

[3] IRIT, Toulouse, France

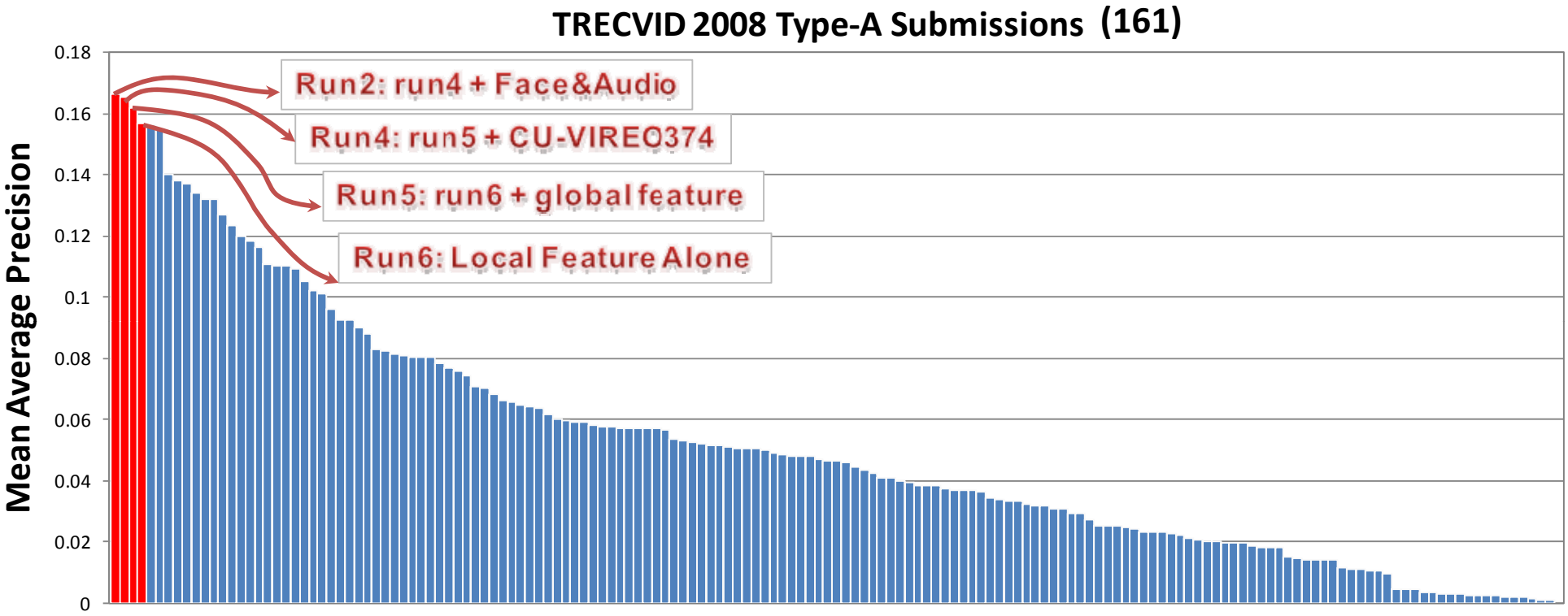
TRECVID 2008 workshop, NIST

Overview: 5 components & 6 runs



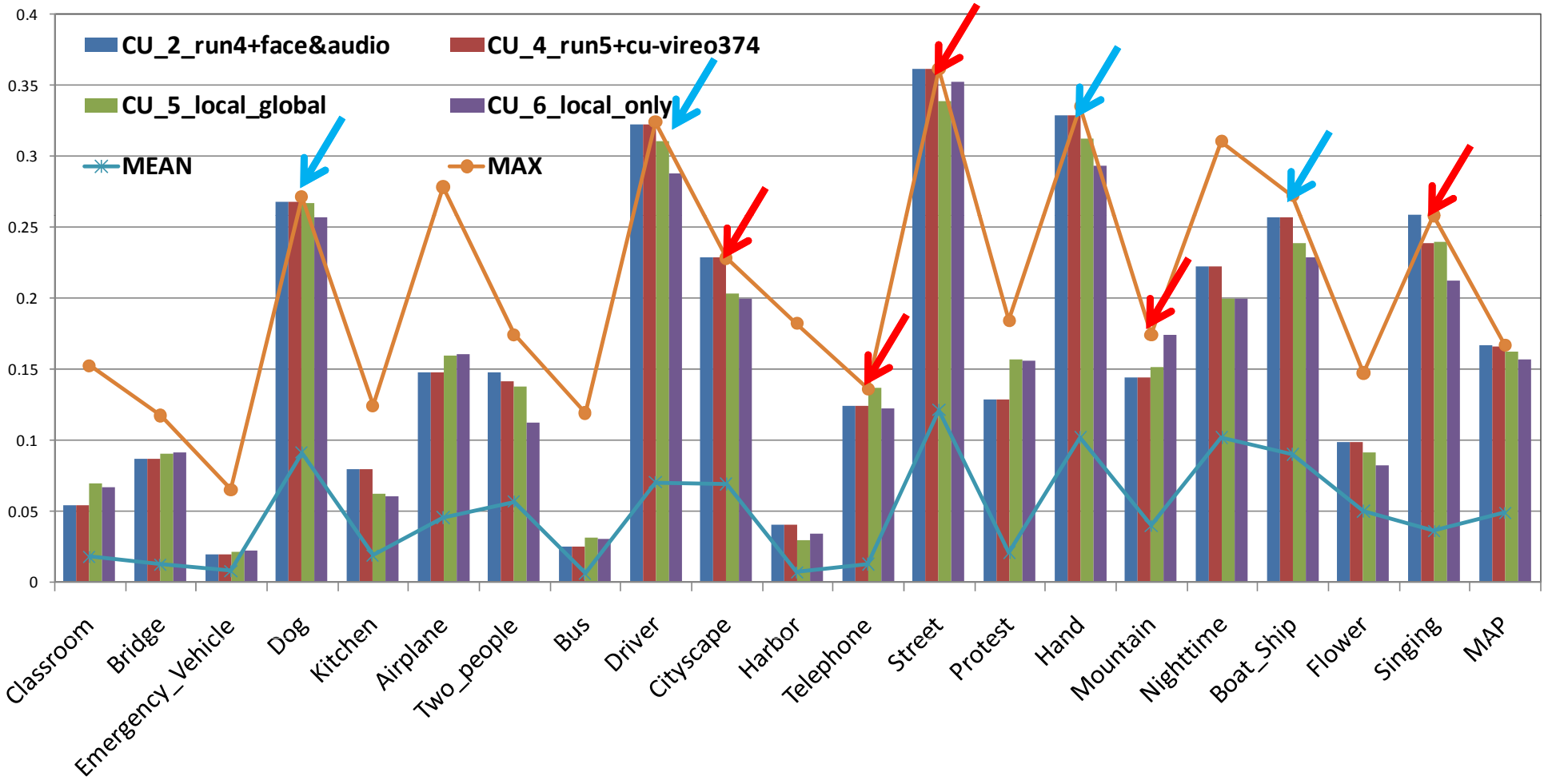
DVMM lab
digital video | multimedia laboratary

Overview: overall performance



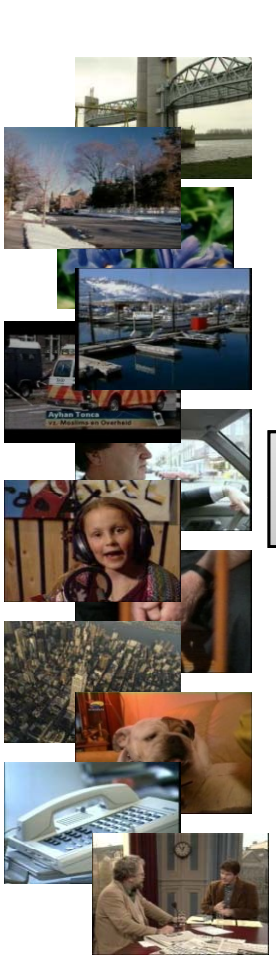
- Local feature alone already achieves near top performance
- Every other component contributes incrementally to the final detection

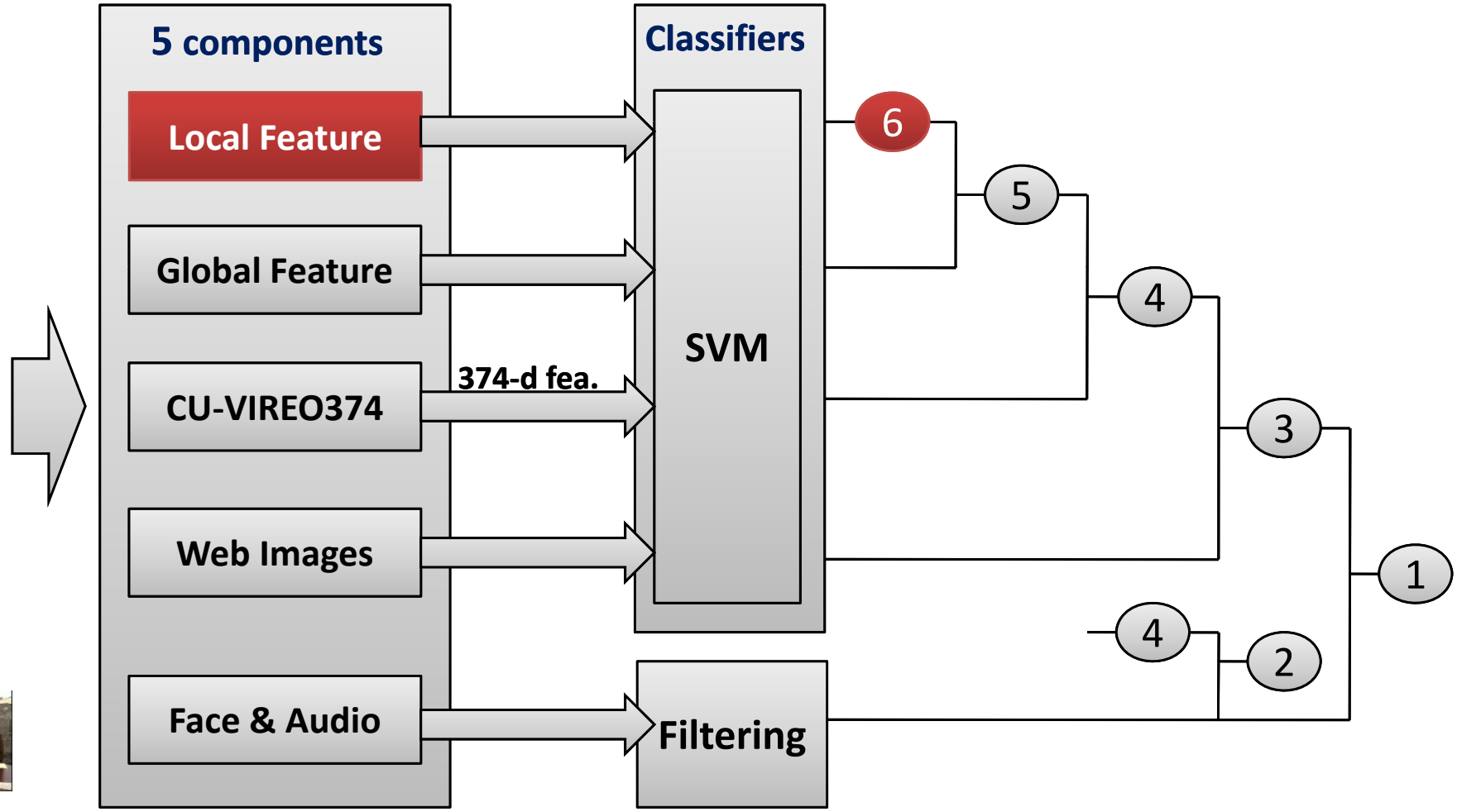
Overview: per-concept performance



DVMM lab
digital video | multimedia laboratary

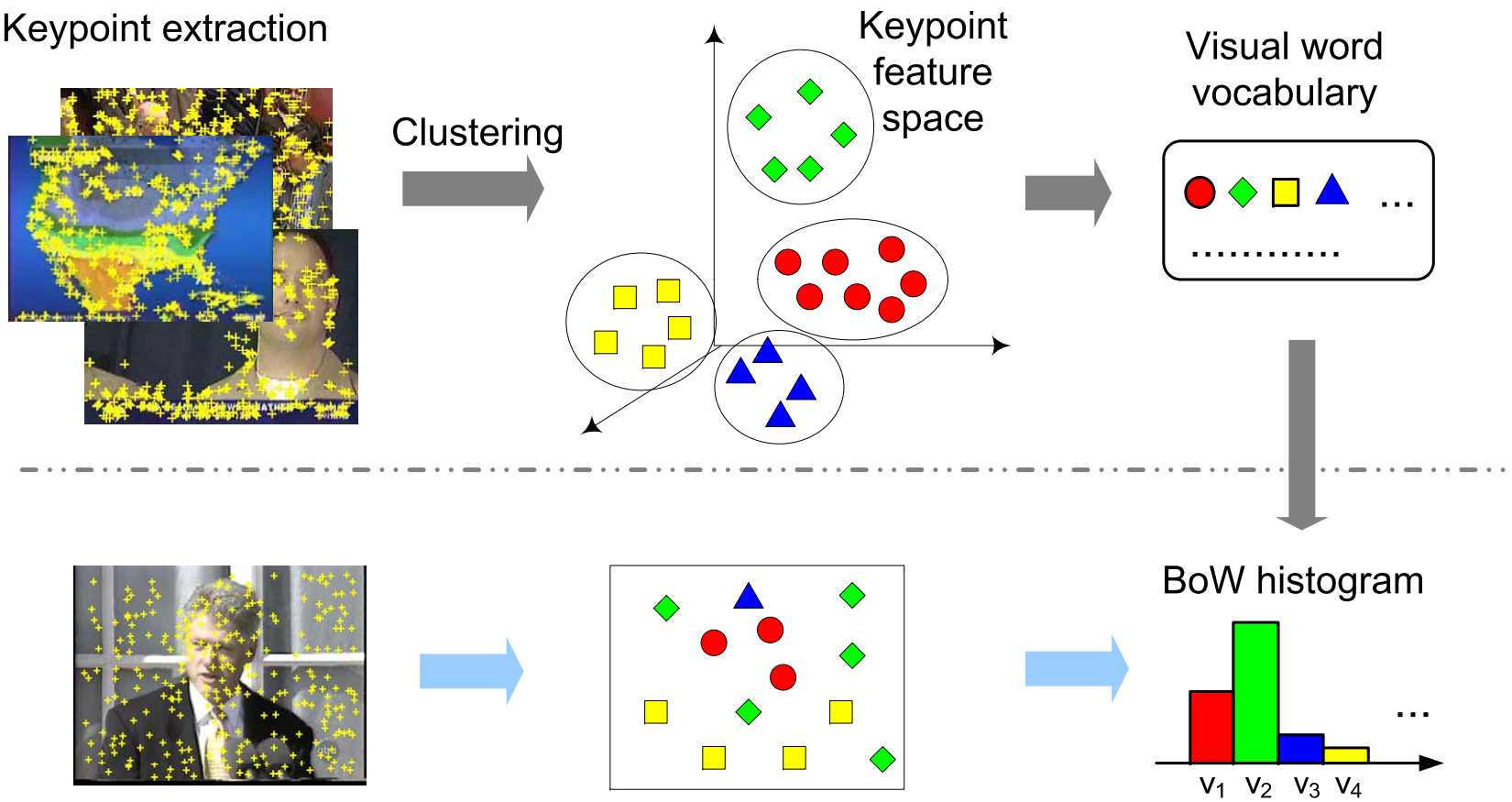
Outline





DVMM lab
digital video | multimedia laboratary

Bag-of-Visual-Words (BoW)

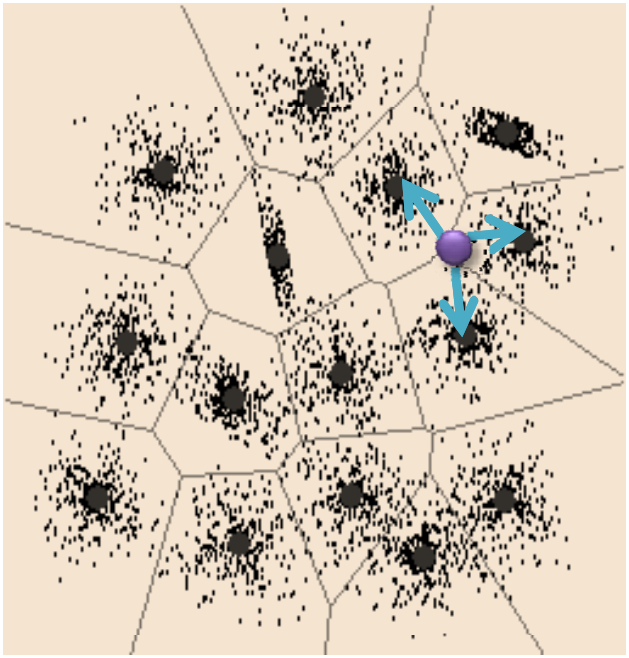


Representation Choices of BoW

- Word weighting scheme
  - *How to weight the importance of a word to an image?*
- Spatial information
  - *Are the spatial locations of keypoints useful?*

Weighting Scheme

- Traditional…
  - Binary, Term frequency (TF), inverse document frequency (IDF)…
- Our method – *soft weighting*



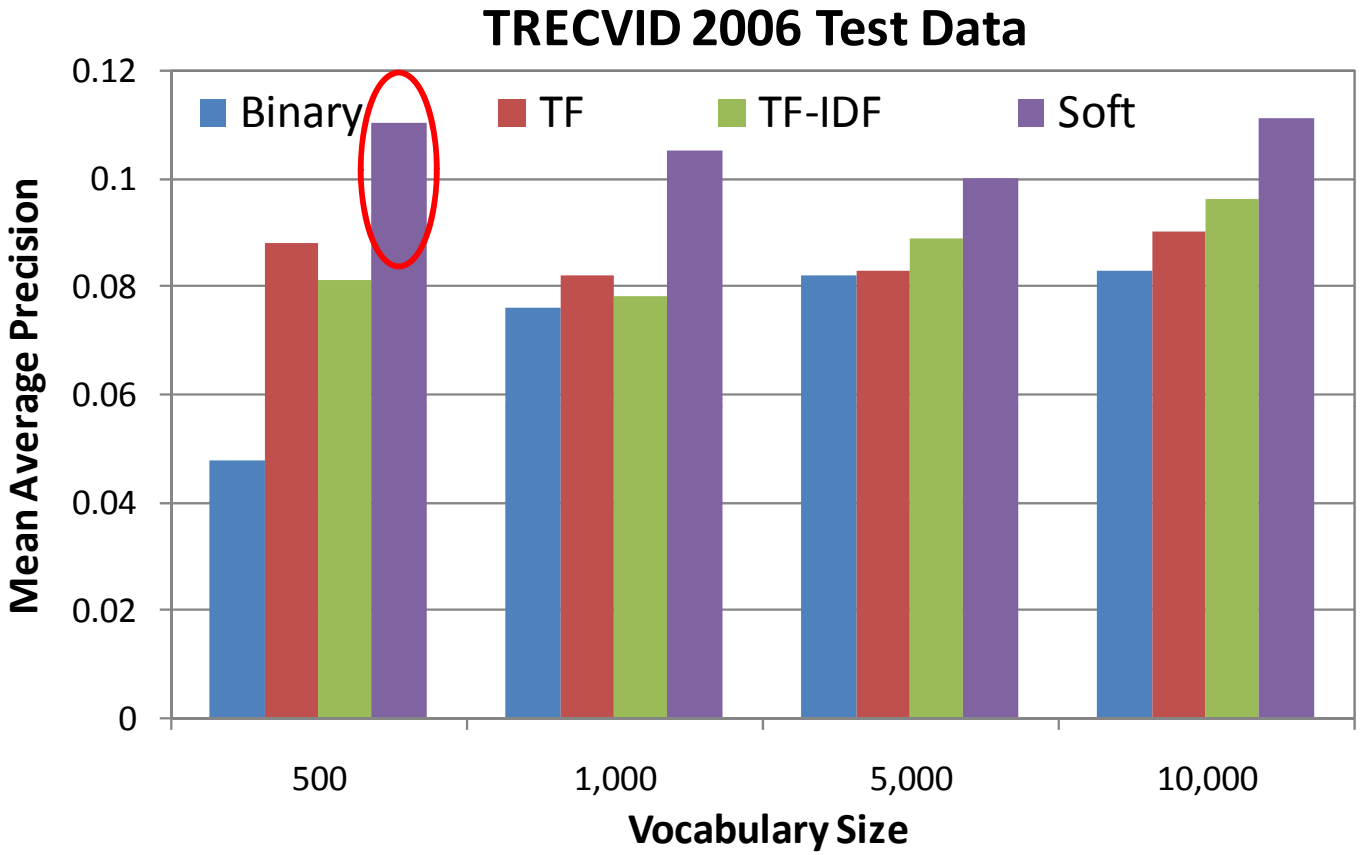
-- Assign a keypoint to multiple visual words

-- weights are determined by keypoint-to- word similarity

Details in:
Jiang et al. CIVR 2007.

Image from http://www.cs.joensuu.fi/pages/franti/vq/lkm15.gif

Vocabulary Size & Weighting Scheme



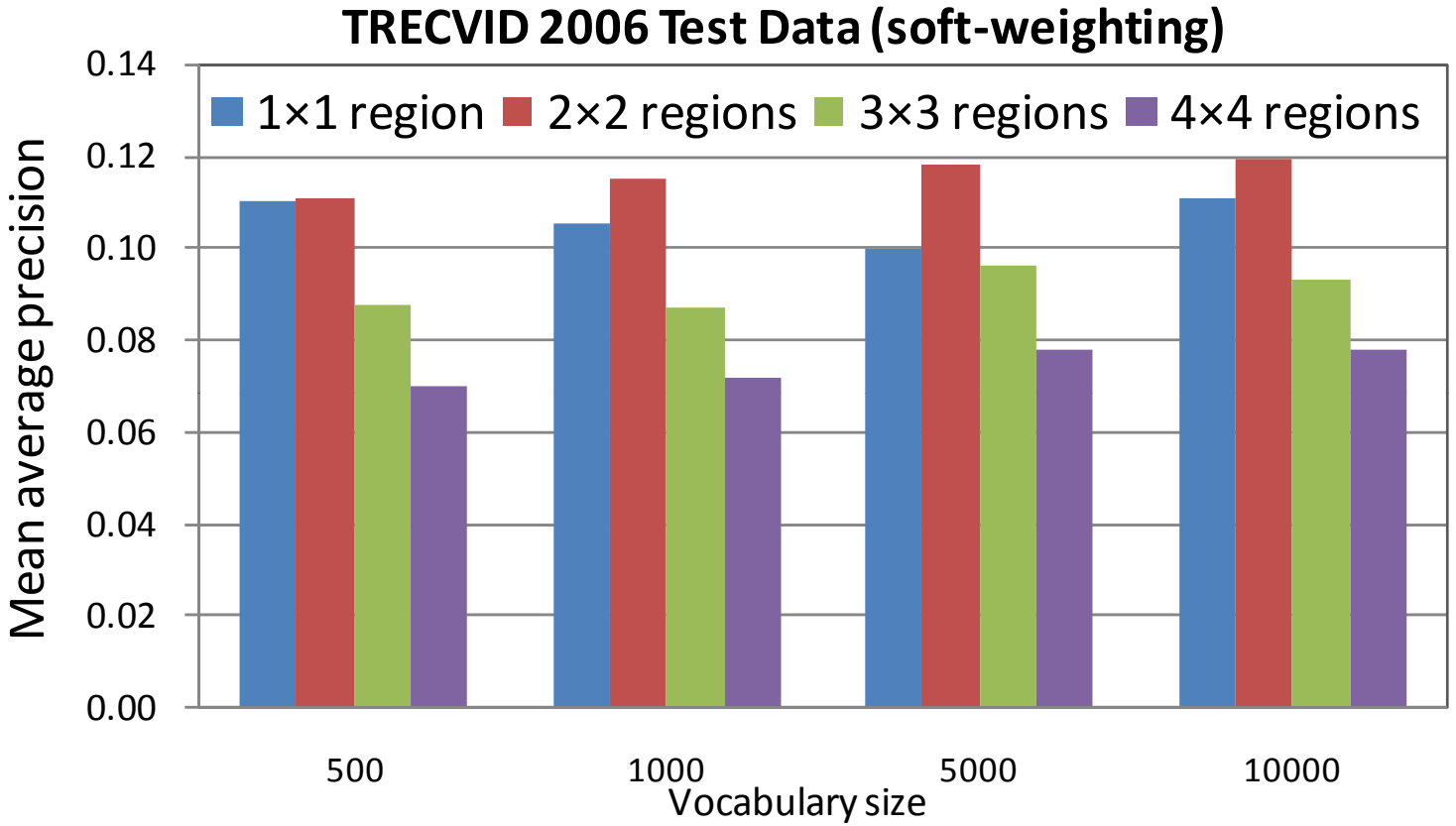
- Soft weighting
  - Improve TF by 10%-20%
    - More accurate to assess the importance of a keypoint

Spatial Information

- Partition image into equal-sized regions
- Concatenate BoW features from the regions
  - Poor generalizability

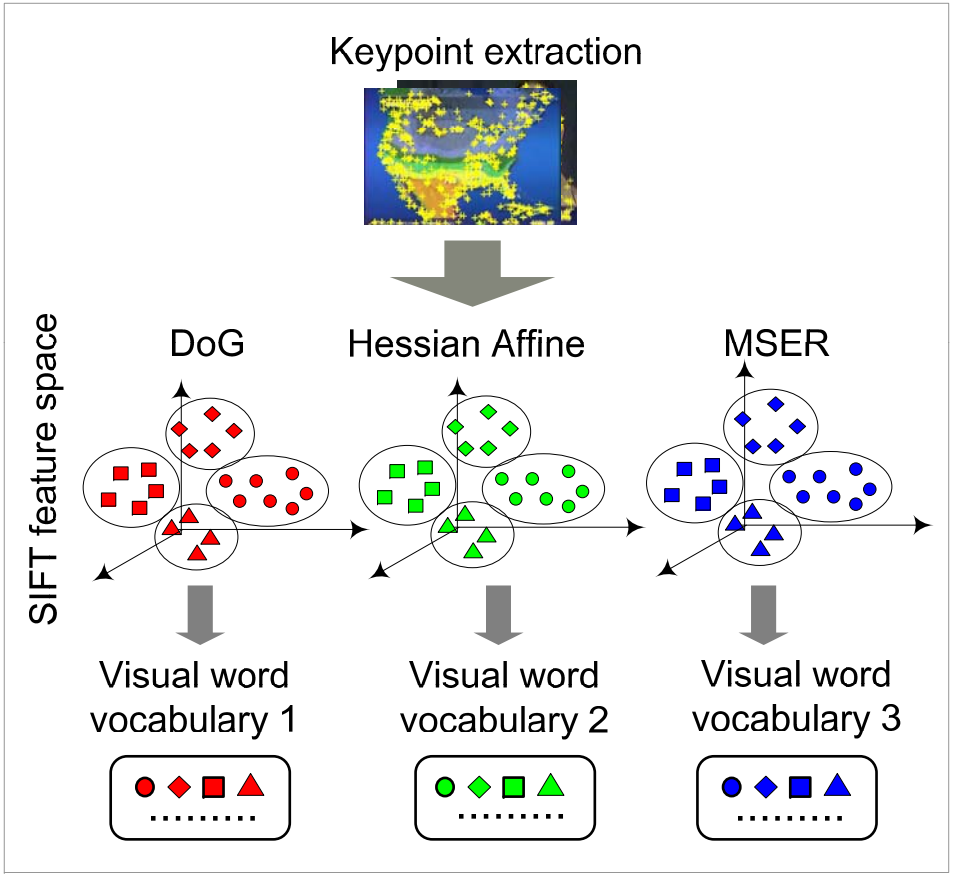
$$\mathbf{F} = (f_{11}, f_{12}, f_{13}, f_{21}, f_{22}, f_{23}, f_{31}, f_{32}, f_{33})$$

Spatial Information

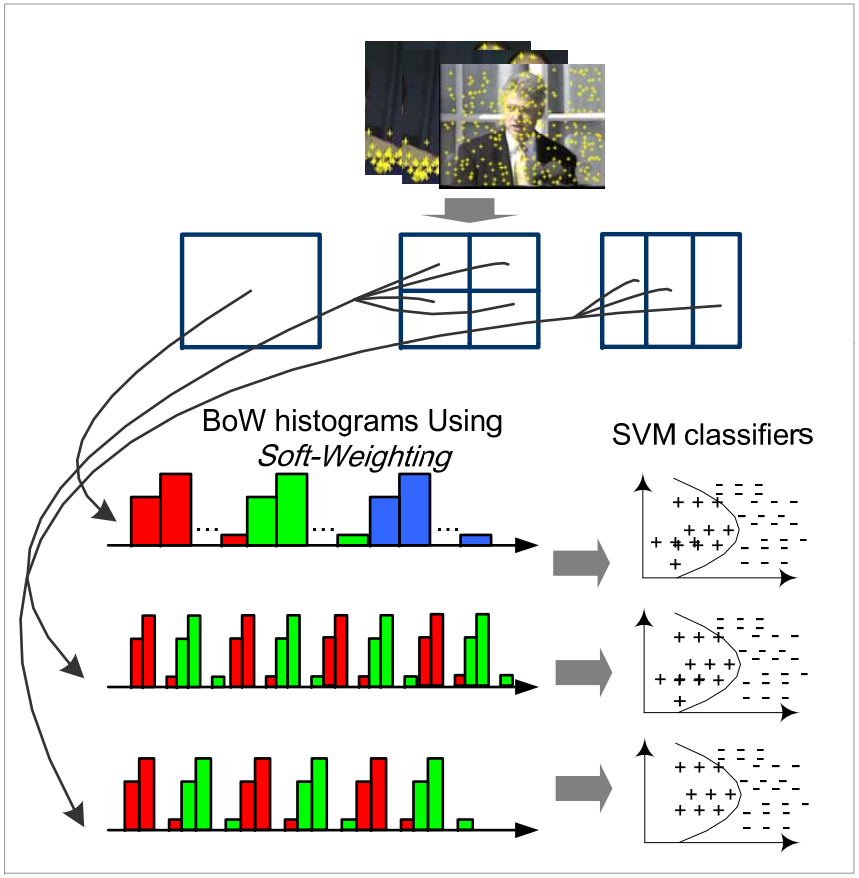


- Spatial Information does not help much for concept detection
  - 2x2 is a good choice
  - 3x3 and 4x4 may cause mismatch problem

Local Feature Representation Framework



Vocabulary Generation

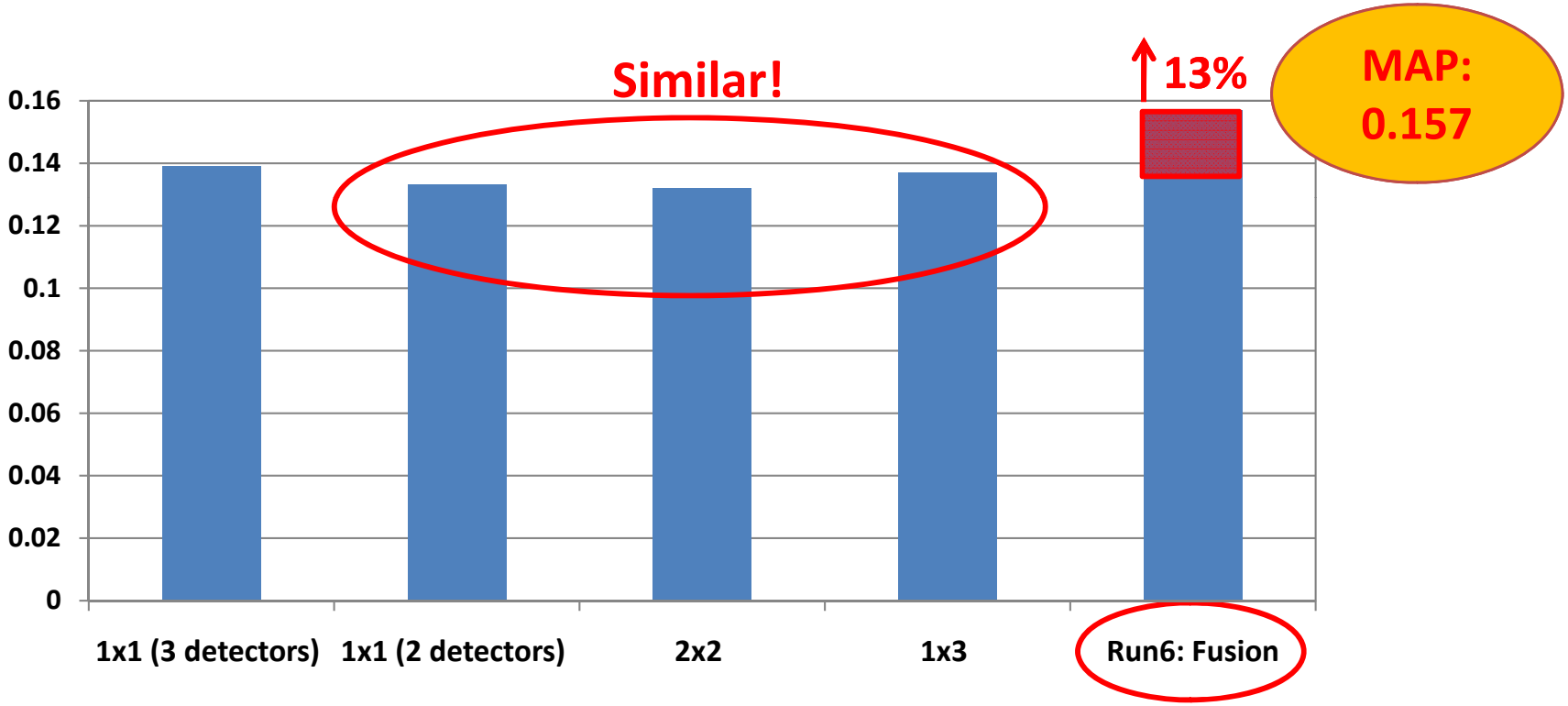


BoW Representation

K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, "A comparison of affine region detectors", IJCV, vol. 65, pp. 43-72, 2005.

Internal Results – Local Features

- Over TRECVID 2008 Test Data



Failure Cases - I

- Flower
  - Small visual area
  - Coloration/texture too similar to background scene

misses

- Possible Solutions
  - Color-descriptor
  - Class-specific visual words

Failure Cases - II

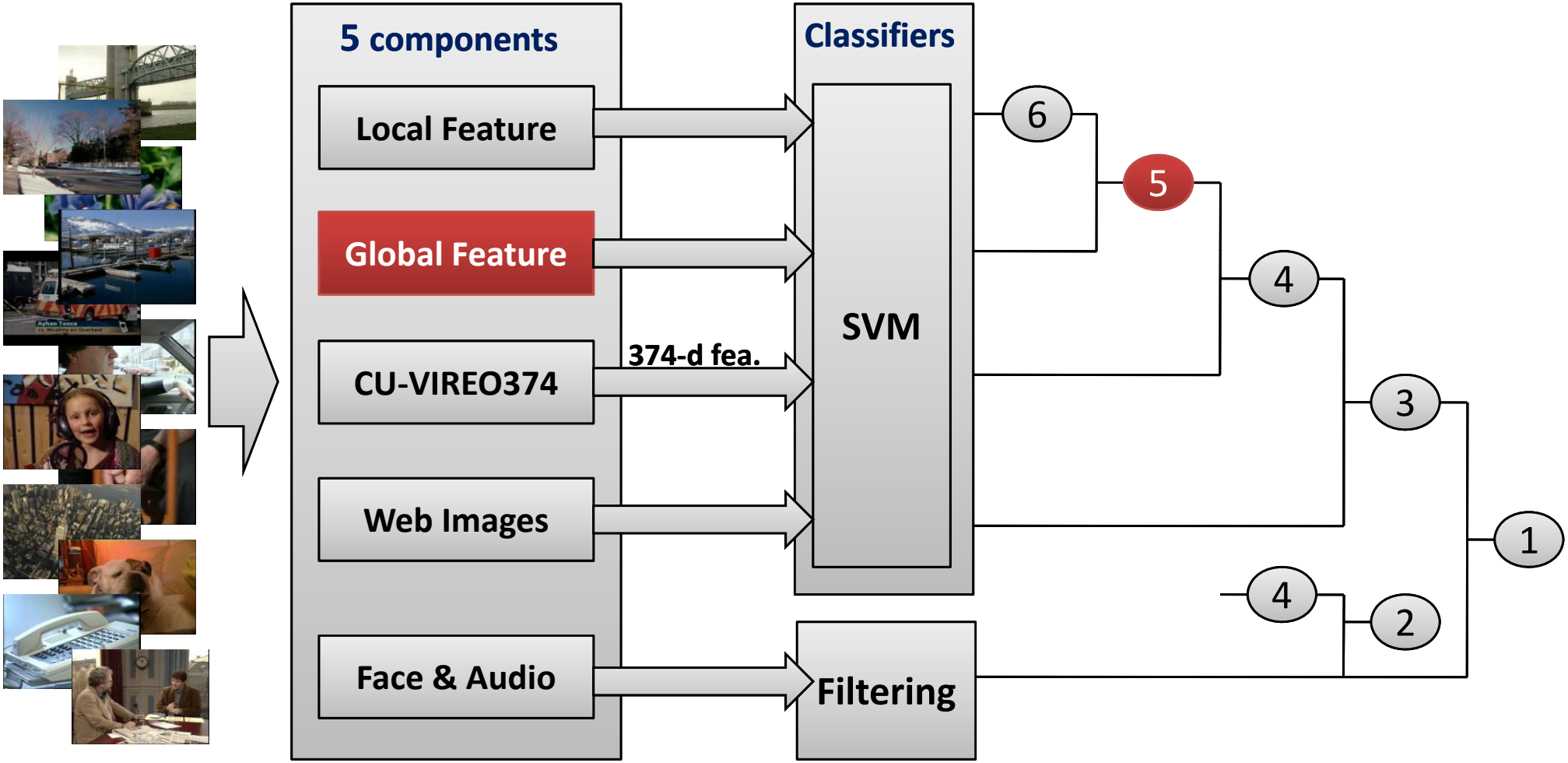
- Boat_Ship, Airplane_flying
  - Learning biased by background scene
  - Difficulty from occlusion
- Possible Solution
  - Feature selection

misses

Summary – Local Features

- BoW with good representation choices achieved very impressive performance
  - Soft-weighting is very effective
  - Multiple spatial layouts are useful
  - Multi-detectors do not help much
- Rooms for future improvement
  - Class-specific visual words, feature selection, color-descriptor etc.

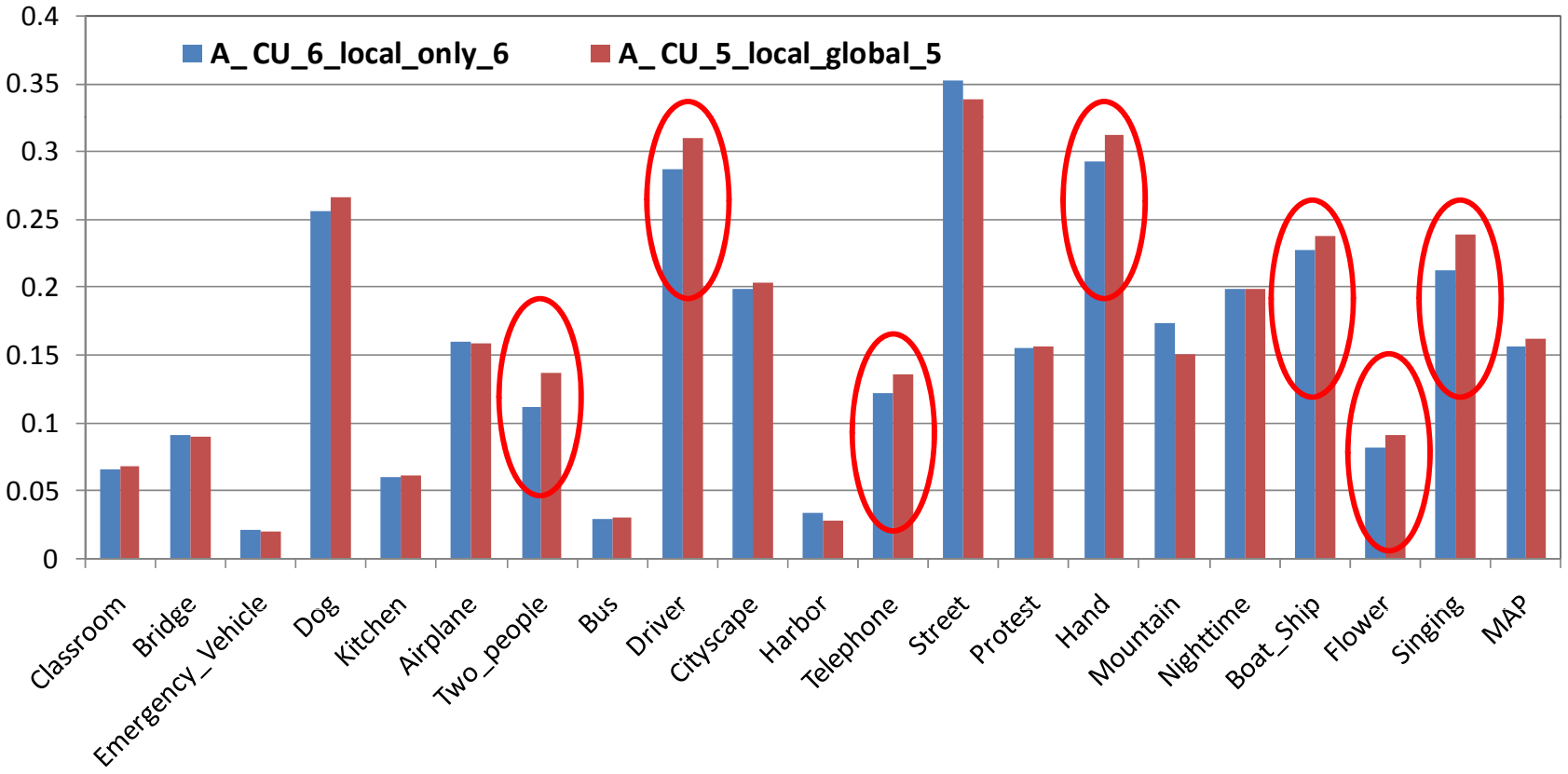
Outline



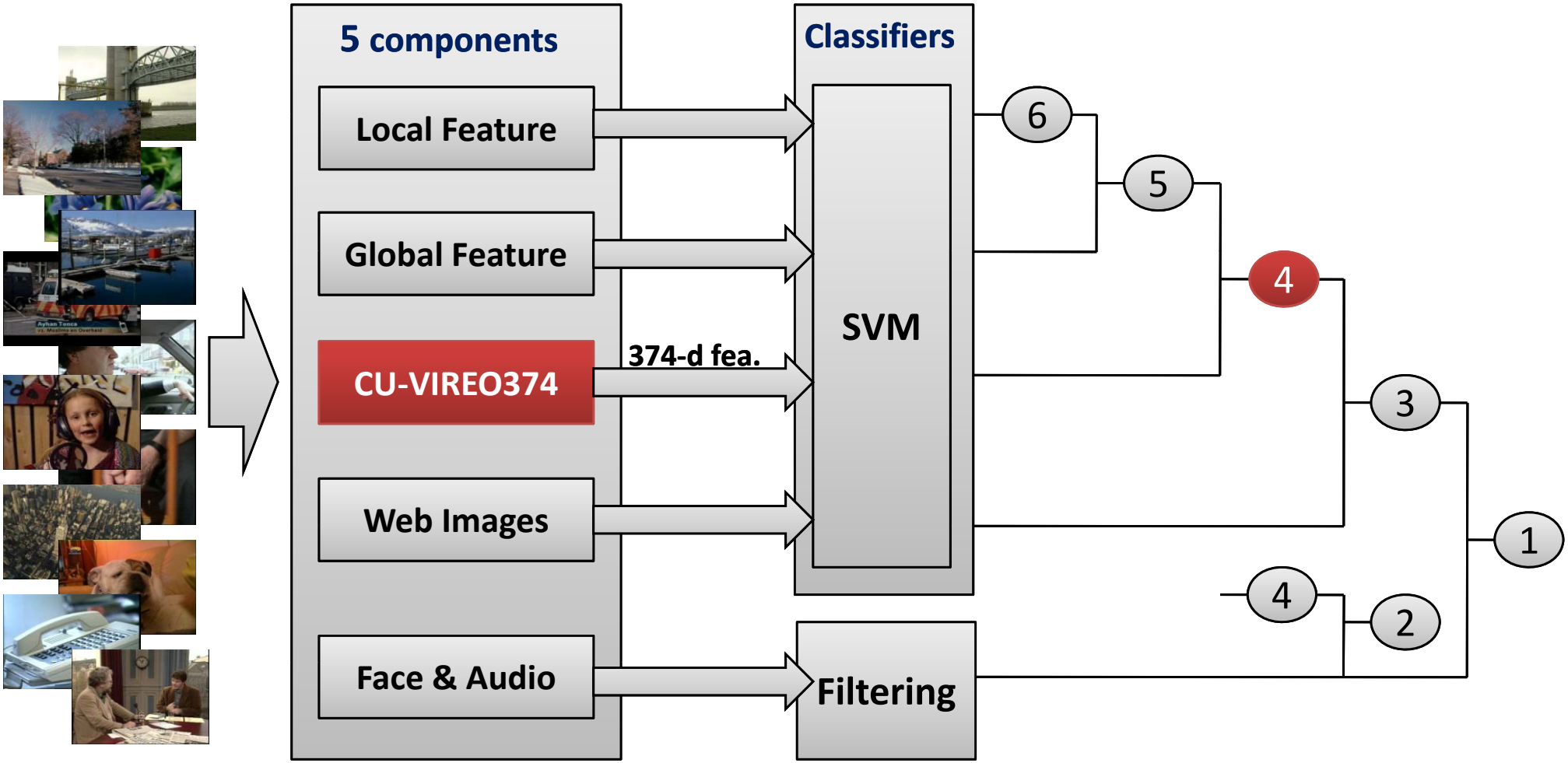
DVMM lab
digital video | multimedia laboratary

Global Features

- Grid-based Color Moments (225-d)
- Wavelet texture (81-d)



Outline

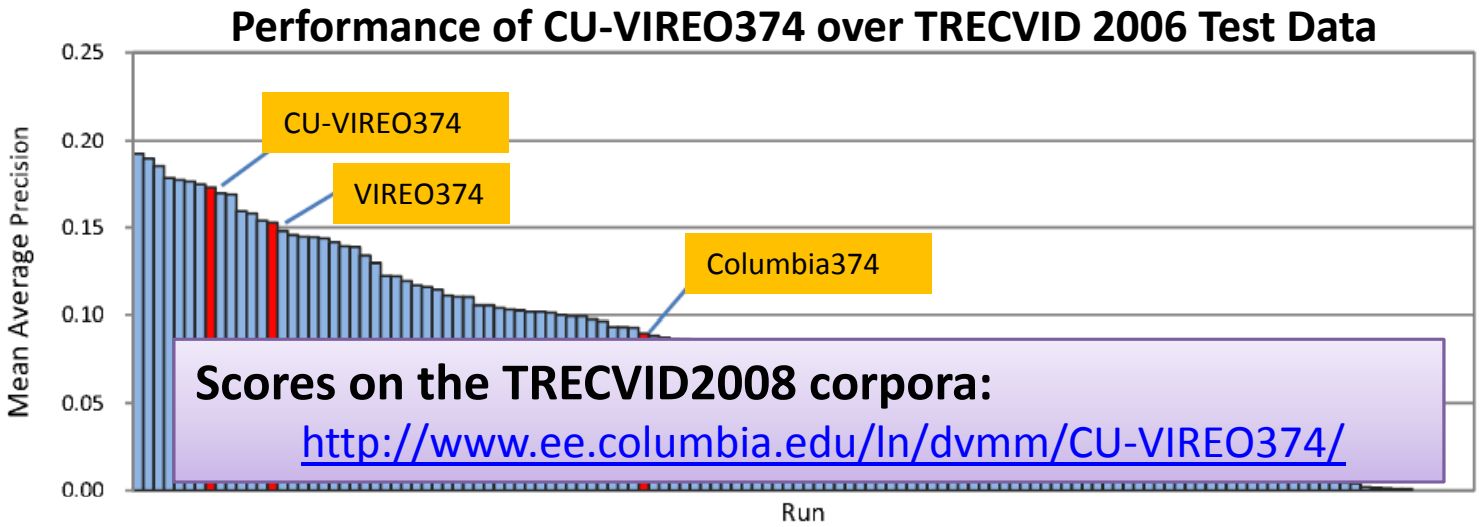


DVMM lab
digital video | multimedia laboratary

CU-VIREO374

- Fusion of Columbia374 and VIREO374

	Feature	Dimension
Columbia374	Grid-based color moment (LUV)	225
	Gabor Texture	48
	Edge Direction Histogram	73
VIREO374	Bag-of-visual-words (soft weighting)	500
	Grid-based Color Moment (*Lab*)	225
	Grid-based Wavelet Texture	81

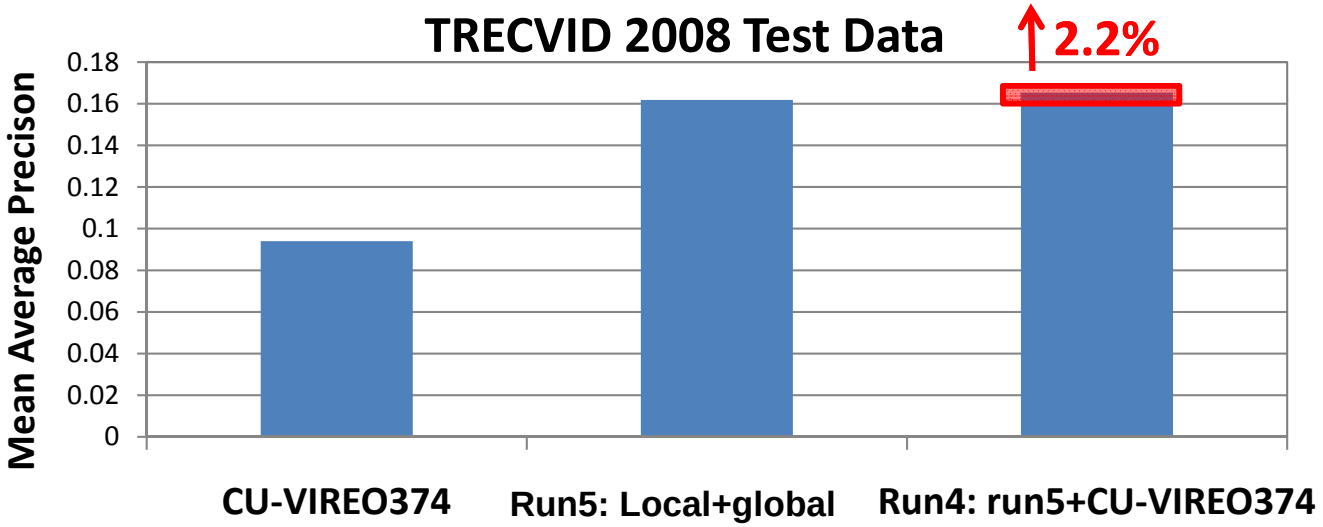


Yu-Gang Jiang, Akira Yanagawa, Shih-Fu Chang, Chong-Wah Ngo, "Fusing Columbia374 and VIREO-374 for Large Scale Semantic Concept Detection", Columbia University ADVENT Technical Report #223-2008-1, Aug. 2008.

DVMM lab digital video | multimedia laboratary

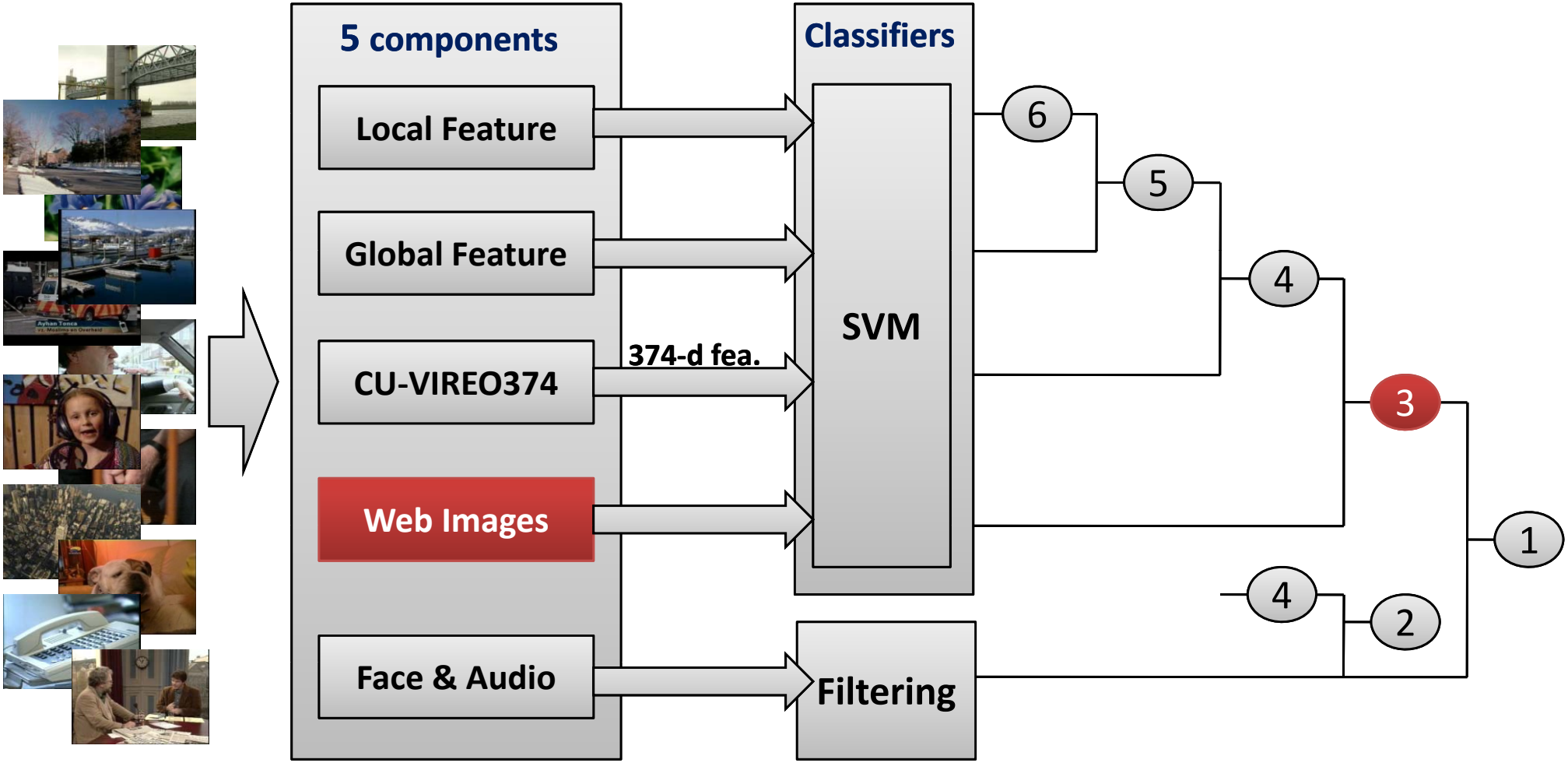
Concept Fusion Using CU-VIREO374

- Train a SVM for each concept
  - Using CU-VIREO374 scores as features



  - Performance improvement is merely 2%
    - Need a better concept fusion model!

Outline



DVMM lab
digital video | multimedia laboratary

Exploring External Images from Web

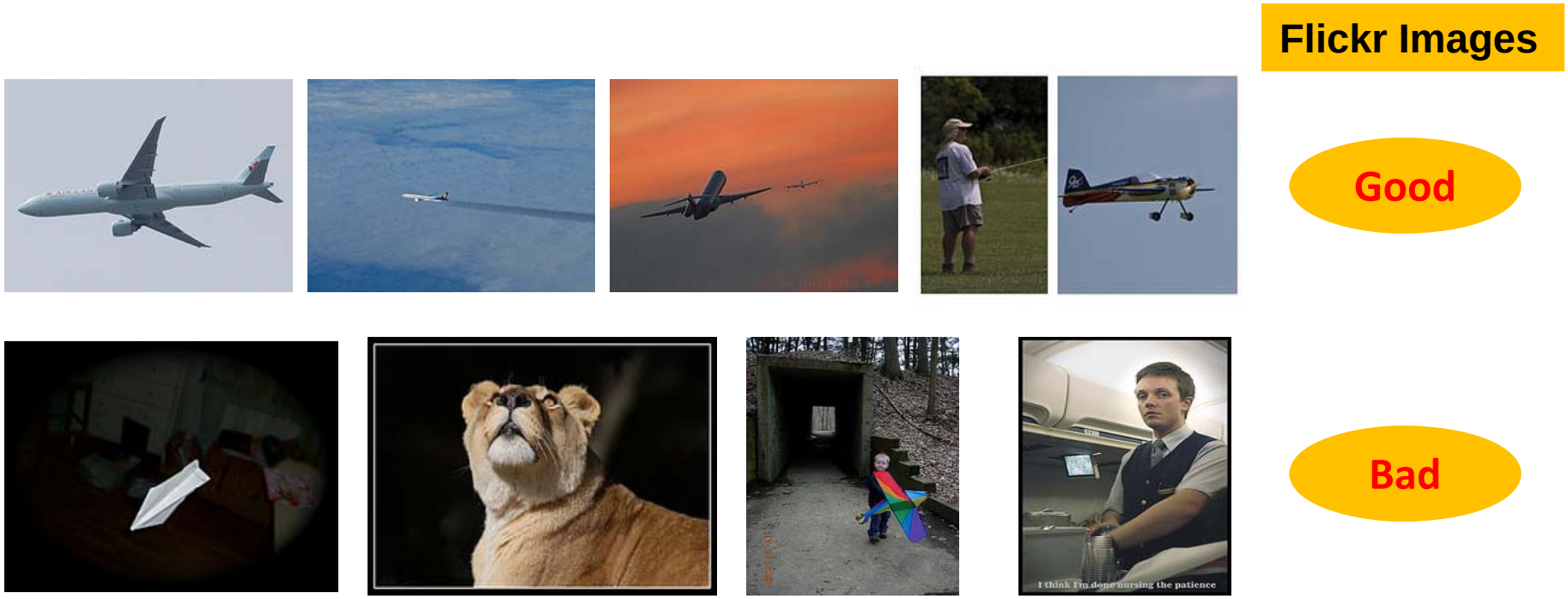
- Problem
  - Sparsity of positive data

Concept Name	# Positive shots	Concept Name	# Positive shots
Classroom	224	Harbor	**195**
Bridge	**158**	Telephone	**184**
Emergency_Vehicle	**88**	Street	1551
Dog	**122**	Demonstration/Protest	**134**
Kitchen	250	Hand	1515
Airplane_flying	**72**	Mountain	239
Two_people	3630	Nighttime	424
Bus	**87**	Boat_Ship	437
Driver	258	Flower	582
Cityscape	288	Singing	366

Total # of shots in TV'08 Dev: 36,262

Challenging Issues

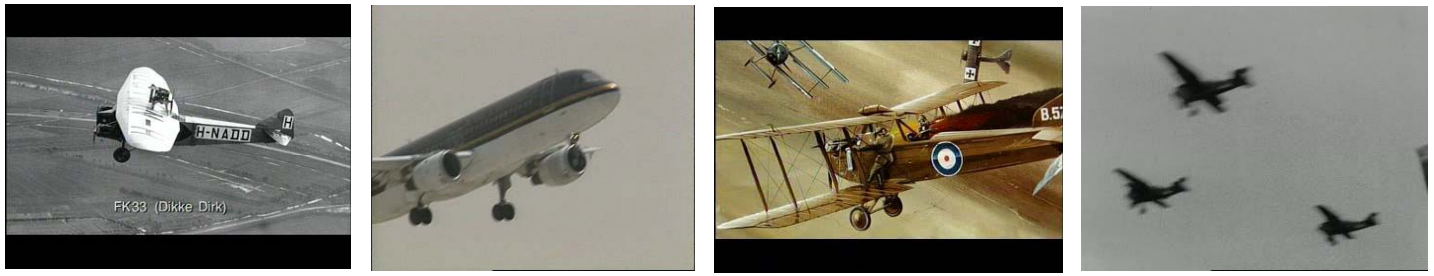
- How to make use of the large amount of “noisily labeled” web images for concept detection?
  - Issue 1: filter the false positive samples



Challenging Issues

- How to make use of the large amount of “noisily labeled” web images for concept detection?
  - Issue 1: filter the false positive samples
  - Issue 2: overcome the cross-domain problem

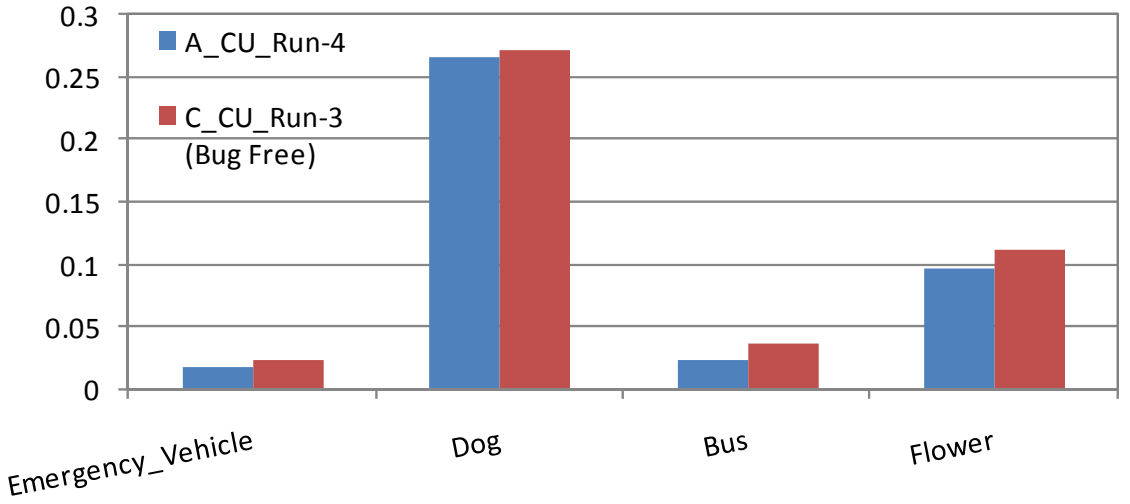
Flickr



TRECVID

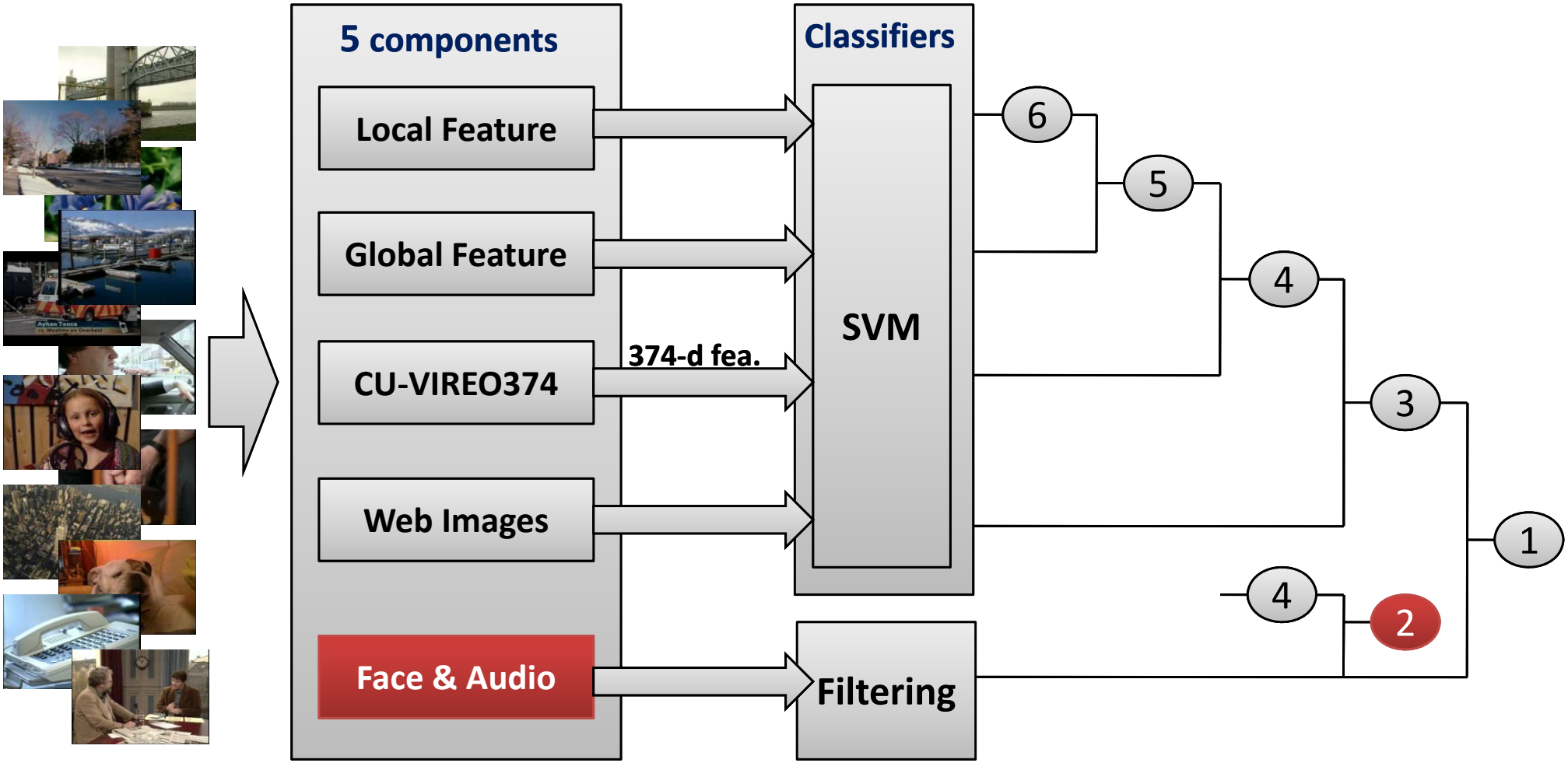
Preliminary Results

- Web image set: 18,000 from Flickr
  - Issue 1: filter the false positive samples
    - ***Graph based semi-supervised learning***
  - Issue 2: overcome the cross-domain problem
    - ***Weighted SVM***
- Results
  - MAP: ***no difference***
  - "Bus": **improve 50%**
- Open Problem!



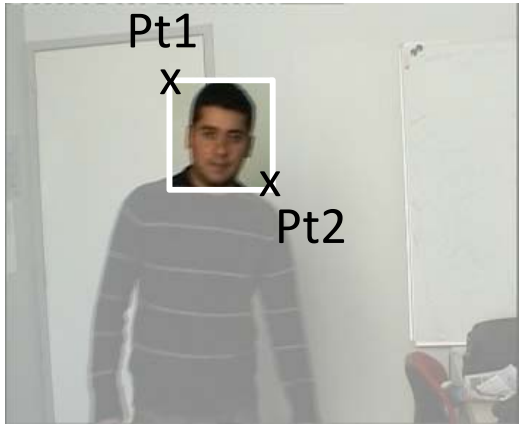
DVMM lab
digital video | multimedia laboratary

Outline

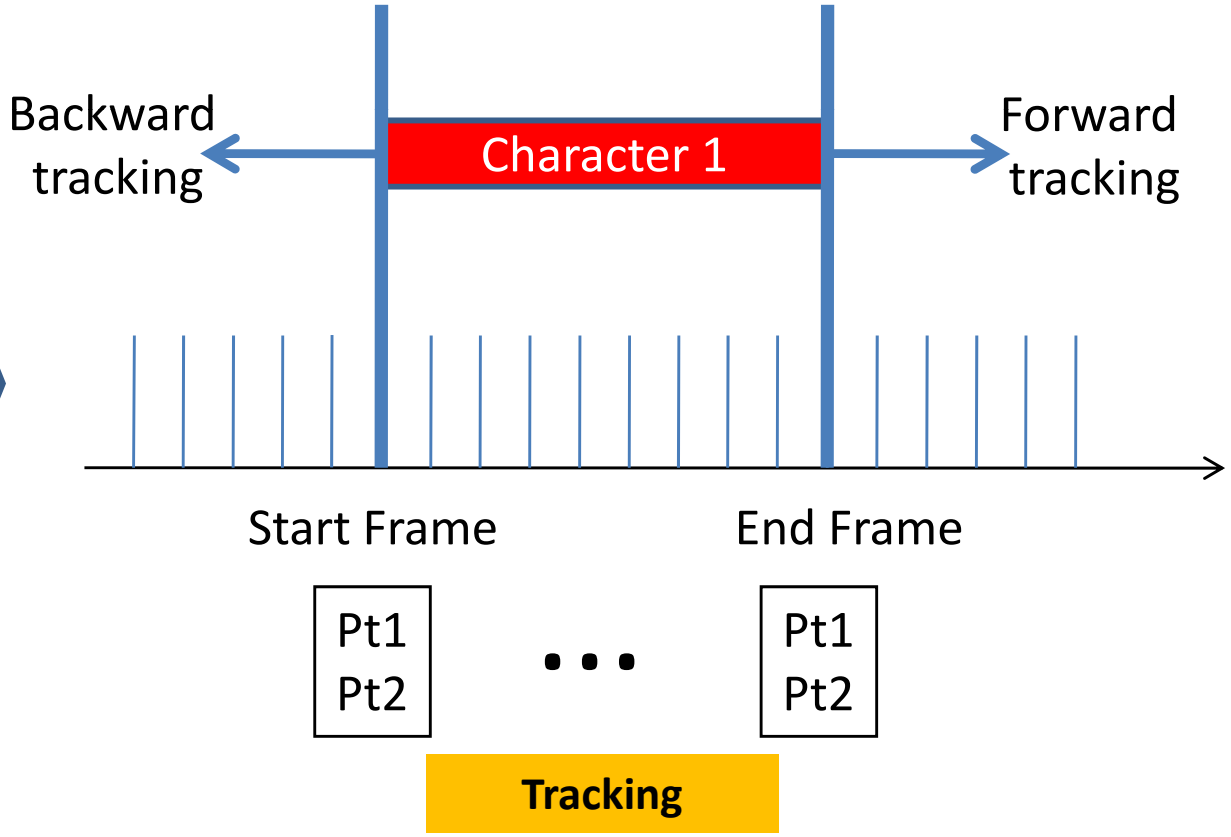


Face Detection and Tracking

- Face Detection (OpenCV Toolbox)
- Tracking based on face location and skin color

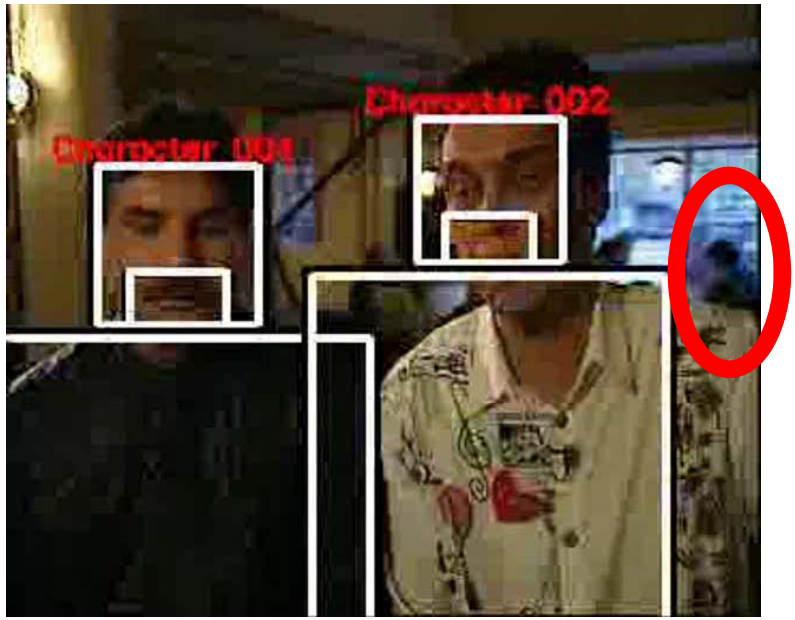


Face Detection



DVMM lab digital video | multimedia laboratary

“Two_people” Detector

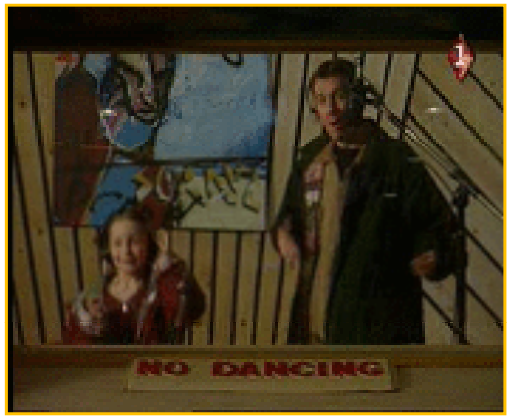


- 250 frames
- 2 people
- 250 frames/person1 & 150 frames/person2
- 100 frames/1 person & 150 frames/2 people

- Drawback
  - Cannot find person when face is too small or invisible

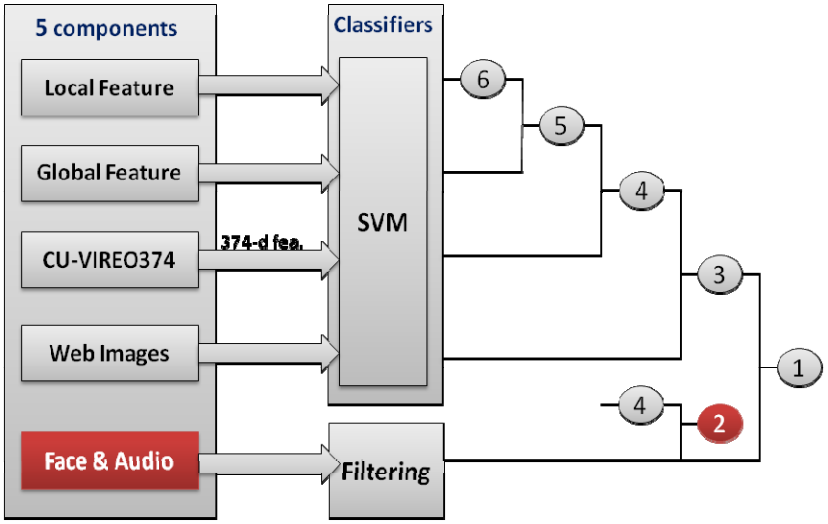
Detecting “Singing” based on Audio

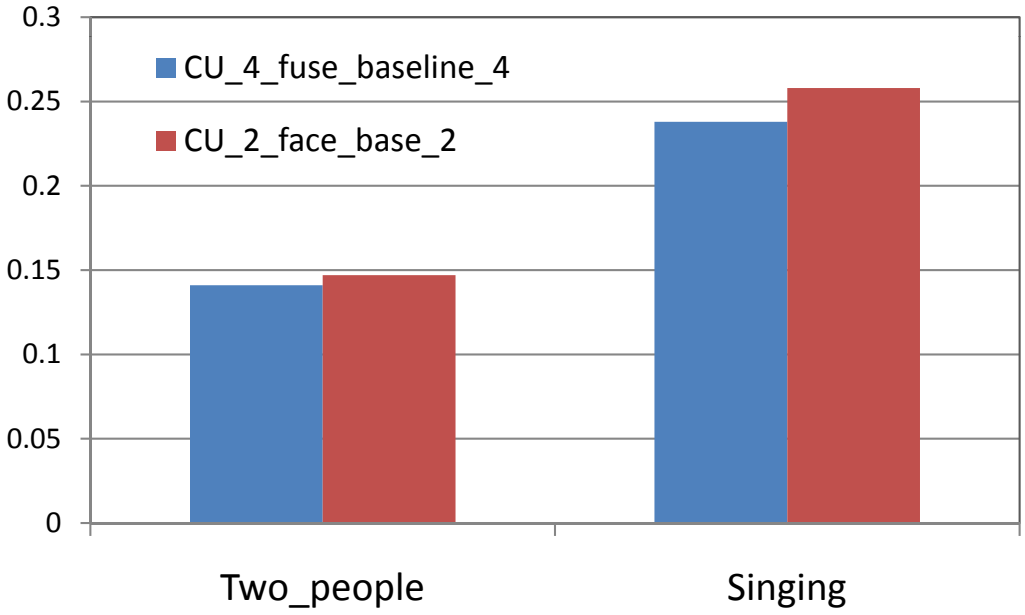
- Vibrato
  - “the variation of the frequency of an musical instrument or of the voice”
- Harmonic Coefficient *Ha*
  - It corresponds to the most important trigonometric series of the spectrum
  - *Ha* is higher in the presence of singing voice



Performance – Face & Audio

- Improve “two_people” by 4% and “singing” by 8%
  - Simple heuristics help detect specific concepts.





Conclusions

- Convergence to Local Features
  - Local feature alone achieved an impressive MAP of 0.157
    - Representation choices are critical for good performance
  - The combination of local features and global features introduces a moderate gain (MAP 0.162)
- CU-VIREO374
  - Useful resource for concept fusion and video search
  - A better fusion model is needed
- Face & Audio detectors
  - Simple heuristics help detect specific concepts
- Training from external web images – **open problem**
  - Useful for concepts lacking in positive training samples
  - Challenges:
    - Unreliable labels
    - Domain differences

DVMM lab
digital video | multimedia laboratary

THANK YOU!

More information at:

http://www.ee.columbia.edu/dvmm/