

Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

1-2008

Concept detection: Convergence to local features and opportunities beyond

Shih-Fu CHANG

Junfeng HE

Yu-Gang JIANG

Elie EL KHOURY

Chong-wah NGO

Singapore Management University, cwngo@smu.edu.sg

See next page for additional authors

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research



Part of the [Data Storage Systems Commons](#), and the [Graphics and Human Computer Interfaces Commons](#)

Citation

CHANG, Shih-Fu; HE, Junfeng; JIANG, Yu-Gang; EL KHOURY, Elie; NGO, Chong-wah; YANAGAWA, Akira; and ZAVESKY, Eric. Concept detection: Convergence to local features and opportunities beyond. (2008). *TREC Video Retrieval Evaluation, TRECVID 2008, Columbia University, October 26*.

Available at: https://ink.library.smu.edu.sg/sis_research/6524

This Conference Paper is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Author

Shih-Fu CHANG, Junfeng HE, Yu-Gang JIANG, Elie EL KHOURY, Chong-wah NGO, Akira YANAGAWA, and Eric ZAVESKY



Concept Detection: Convergence to Local Features and Opportunities Beyond

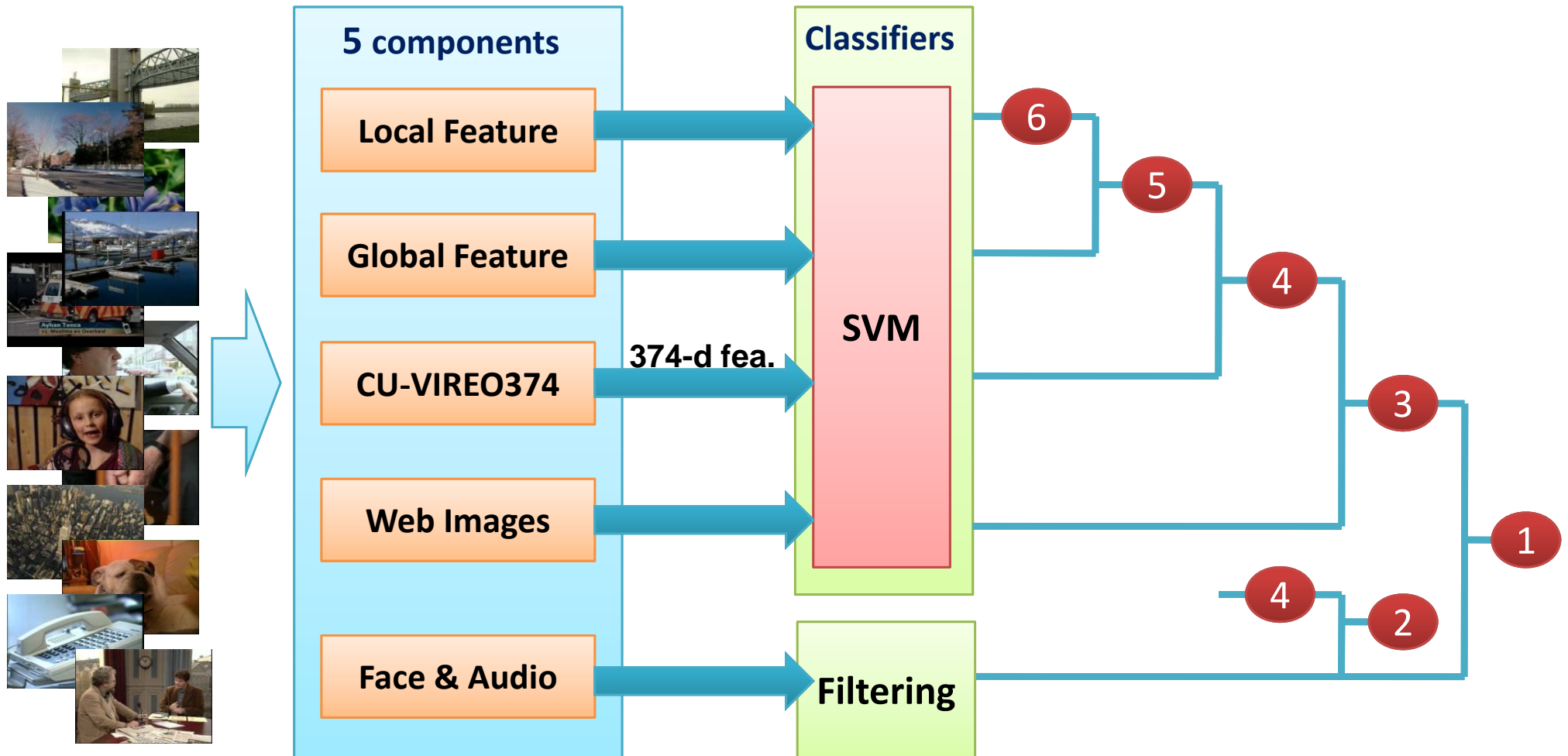
*Shih-Fu Chang¹, Junfeng He¹, Yu-Gang Jiang^{1,2}, Elie El Khoury³,
Chong-Wah Ngo², Akira Yanagawa¹, Eric Zavesky¹*

¹ DVMM Lab, Columbia University

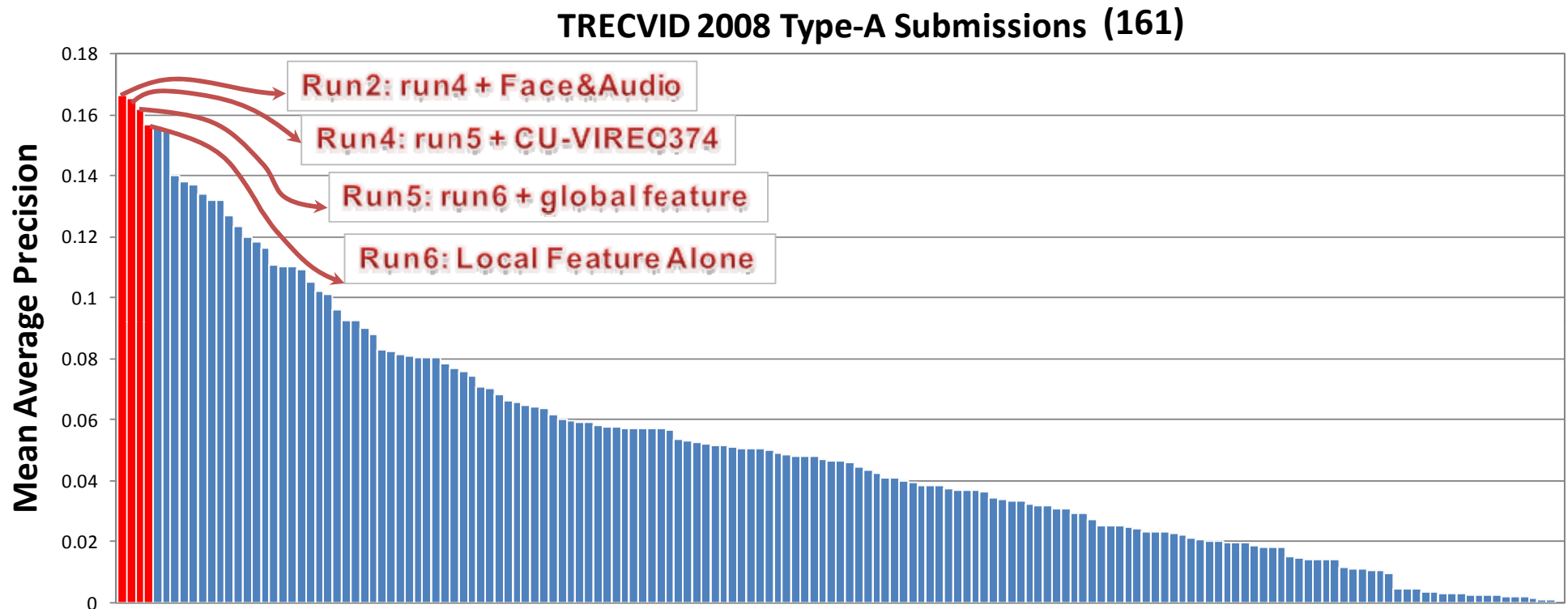
² City University of Hong Kong

³ IRIT, Toulouse, France

Overview: 5 components & 6 runs

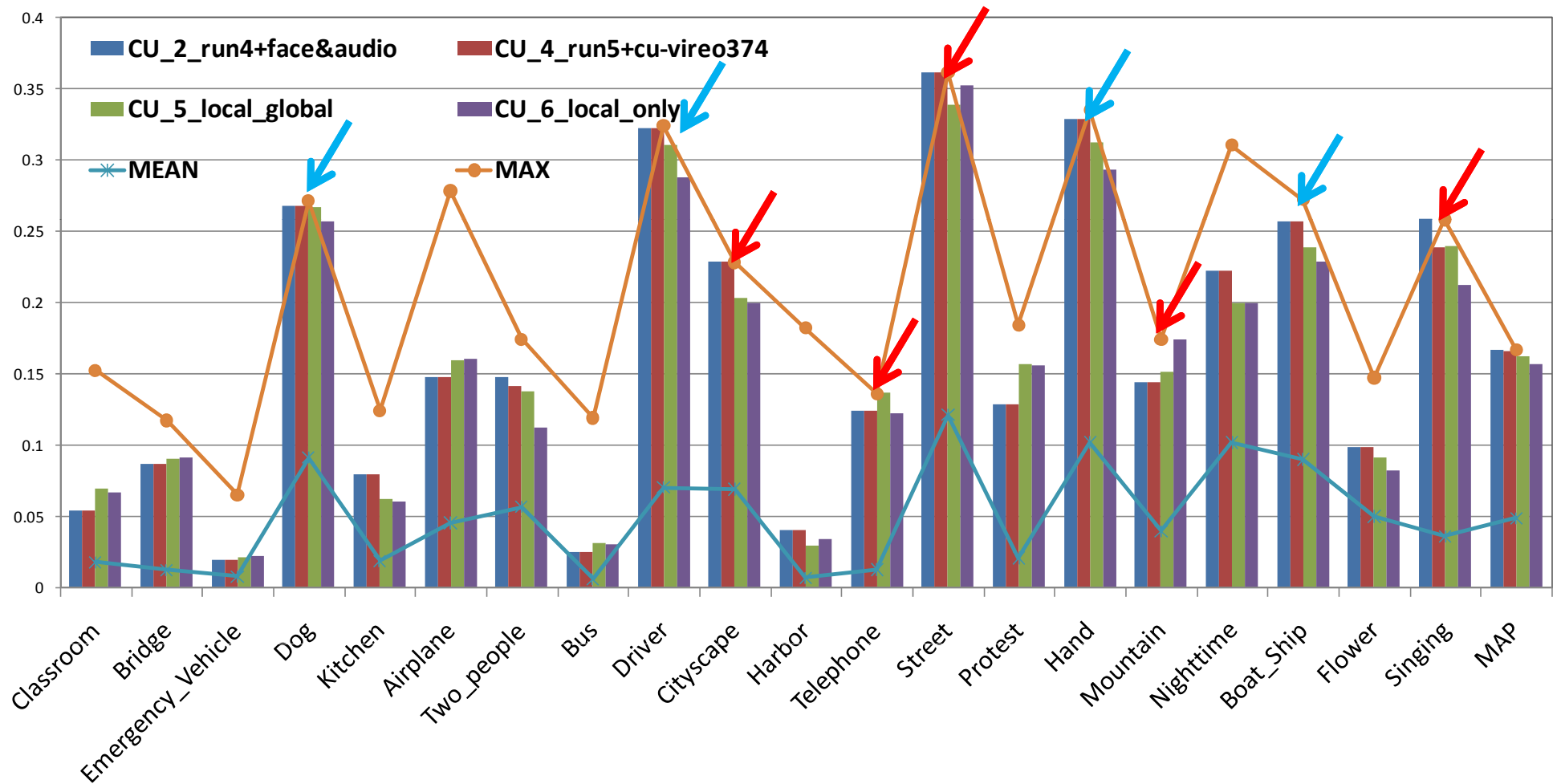


Overview: overall performance

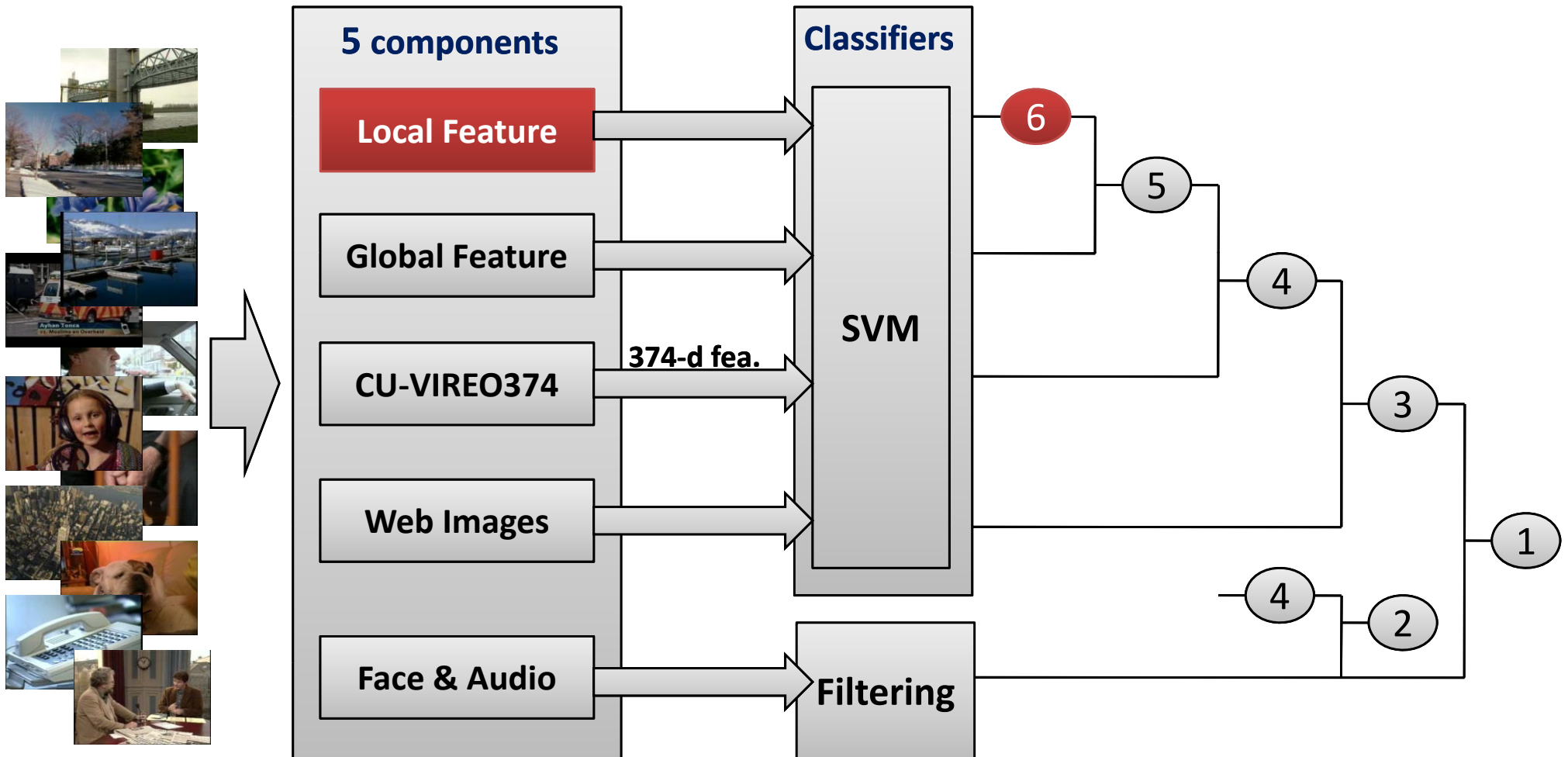


- Local feature alone already achieves near top performance
- Every other component contributes incrementally to the final detection

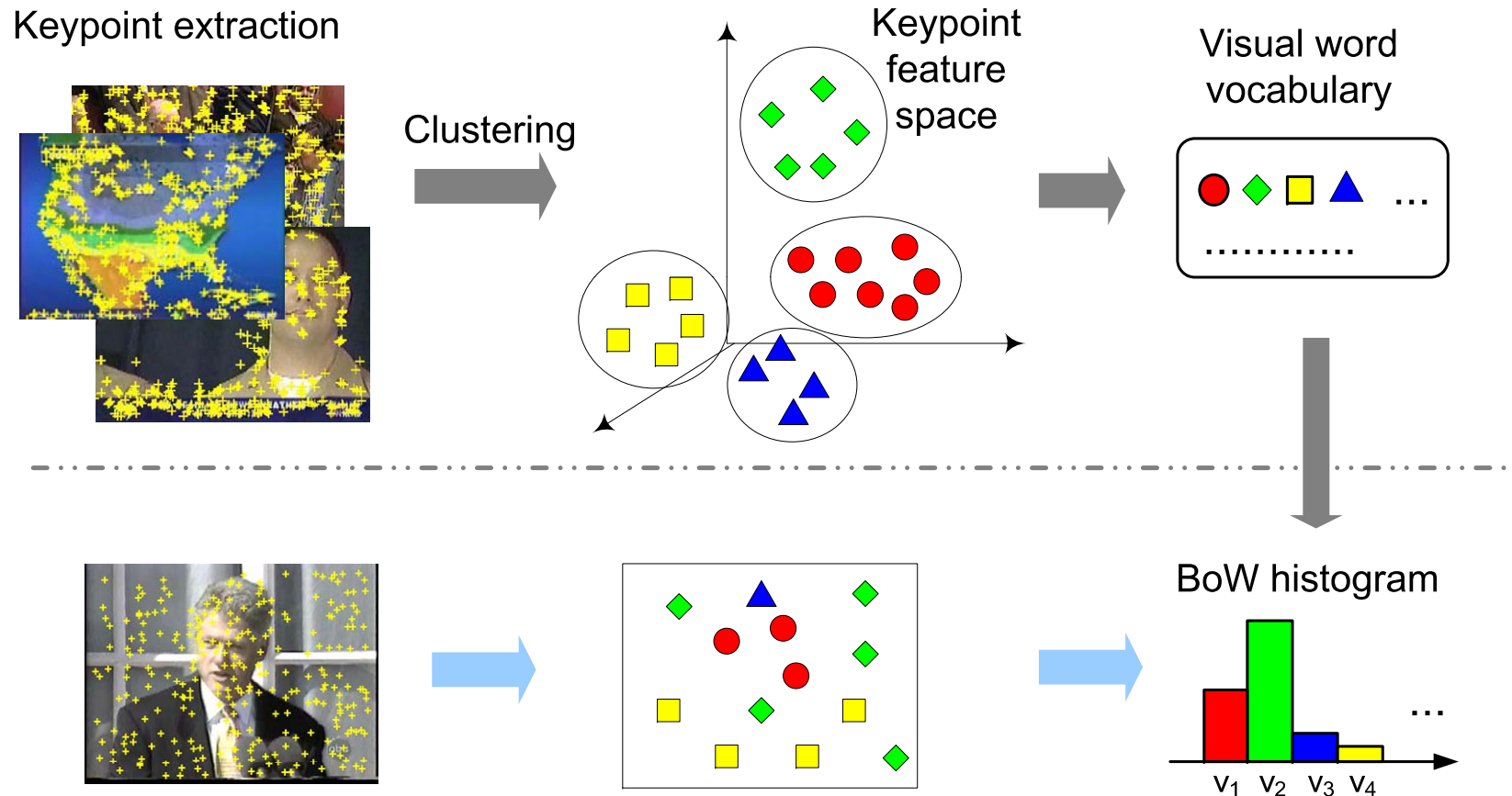
Overview: per-concept performance



Outline



Bag-of-Visual-Words (BoW)

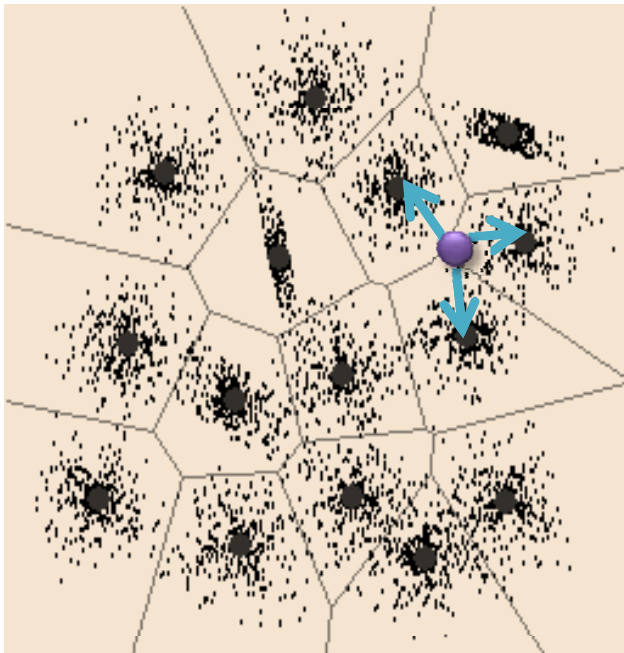


Representation Choices of BoW

- 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*

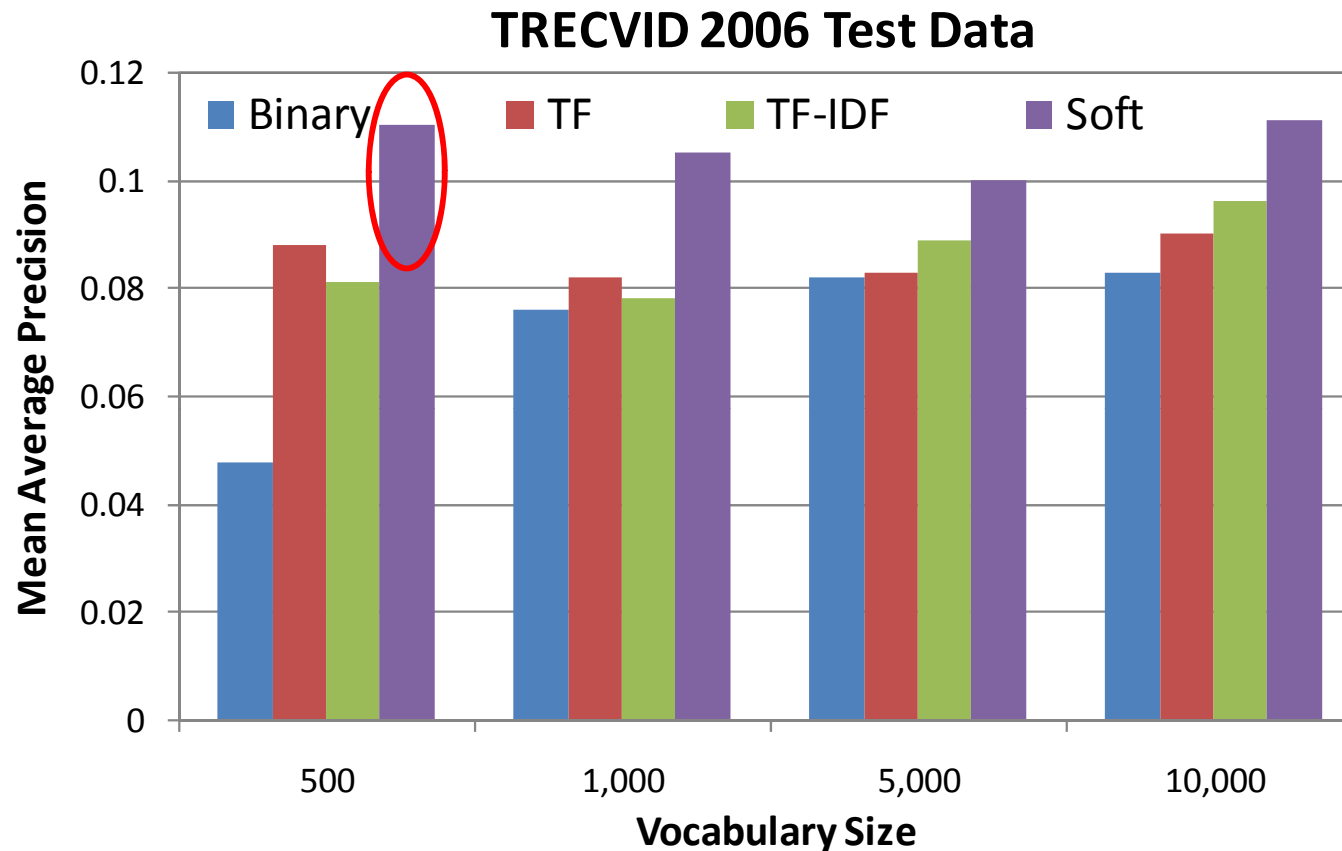


-- *Assign a keypoint to multiple visual words*

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

Details in:
Jiang et al. CIVR 2007.

Vocabulary Size & Weighting Scheme



– Soft weighting

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

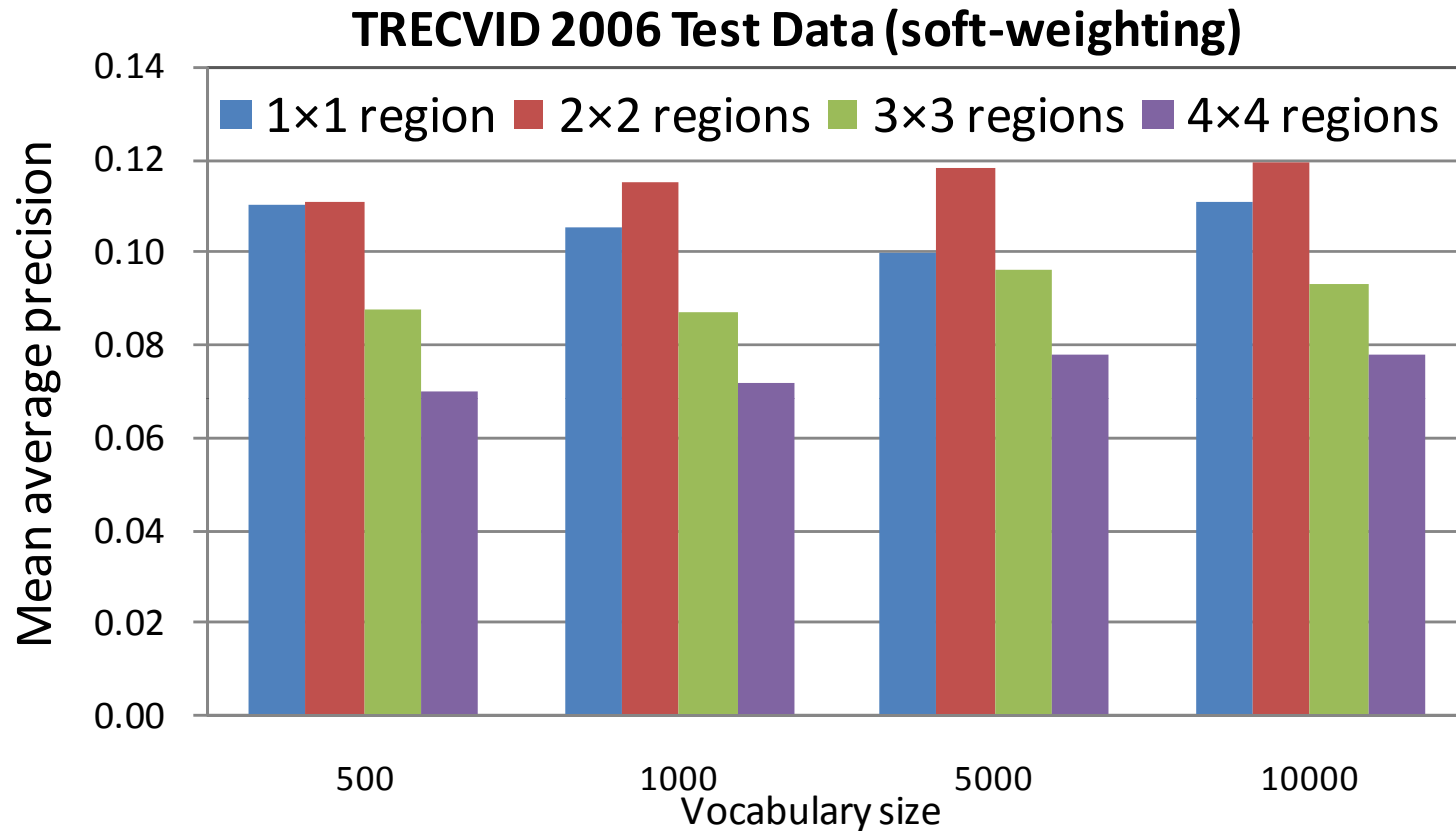
Spatial Information

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



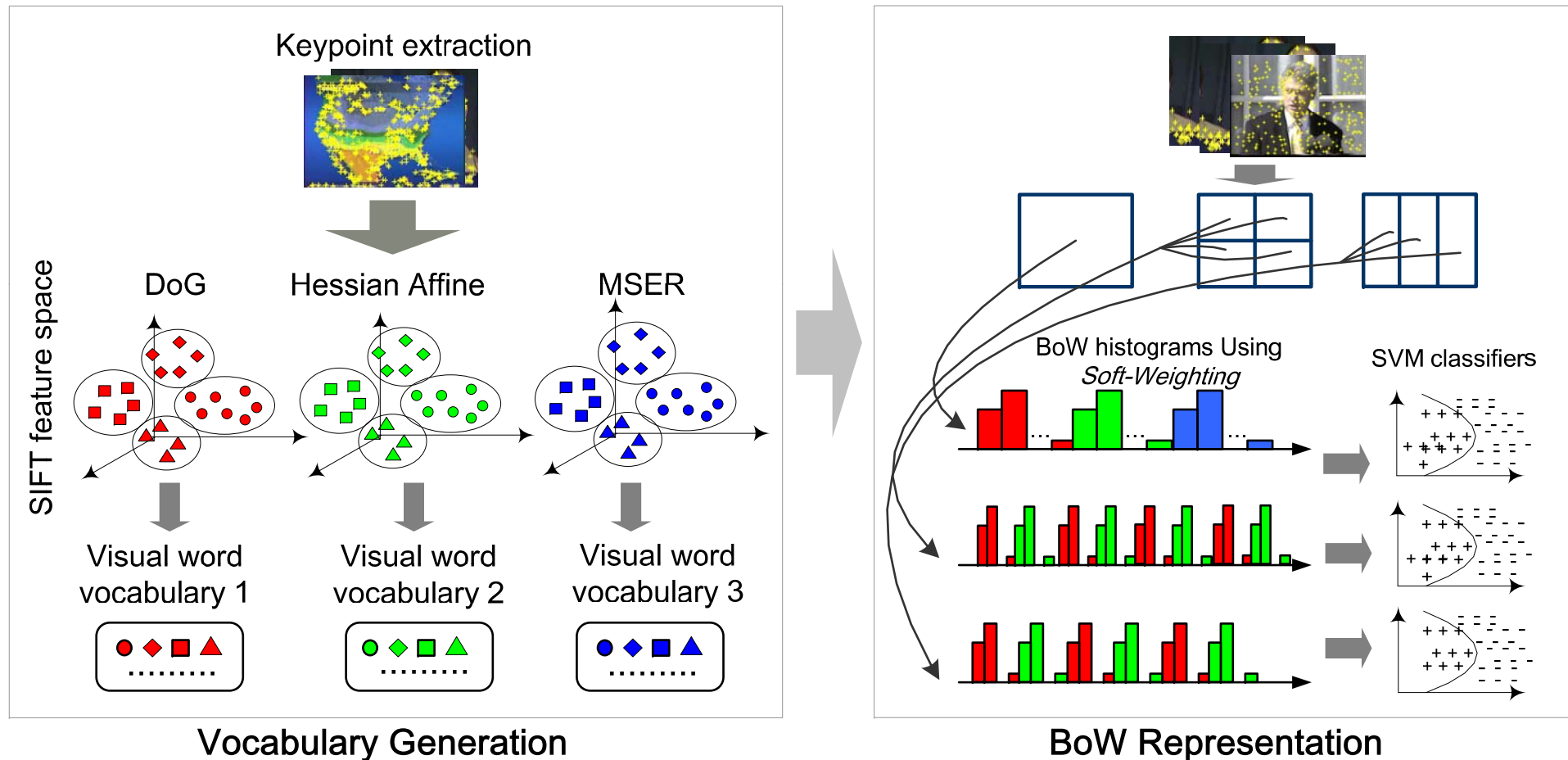
$$F = (f_{11}, f_{12}, f_{13}, f_{21}, f_{22}, f_{23}, f_{31}, f_{32}, f_{33})$$

Spatial Information



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

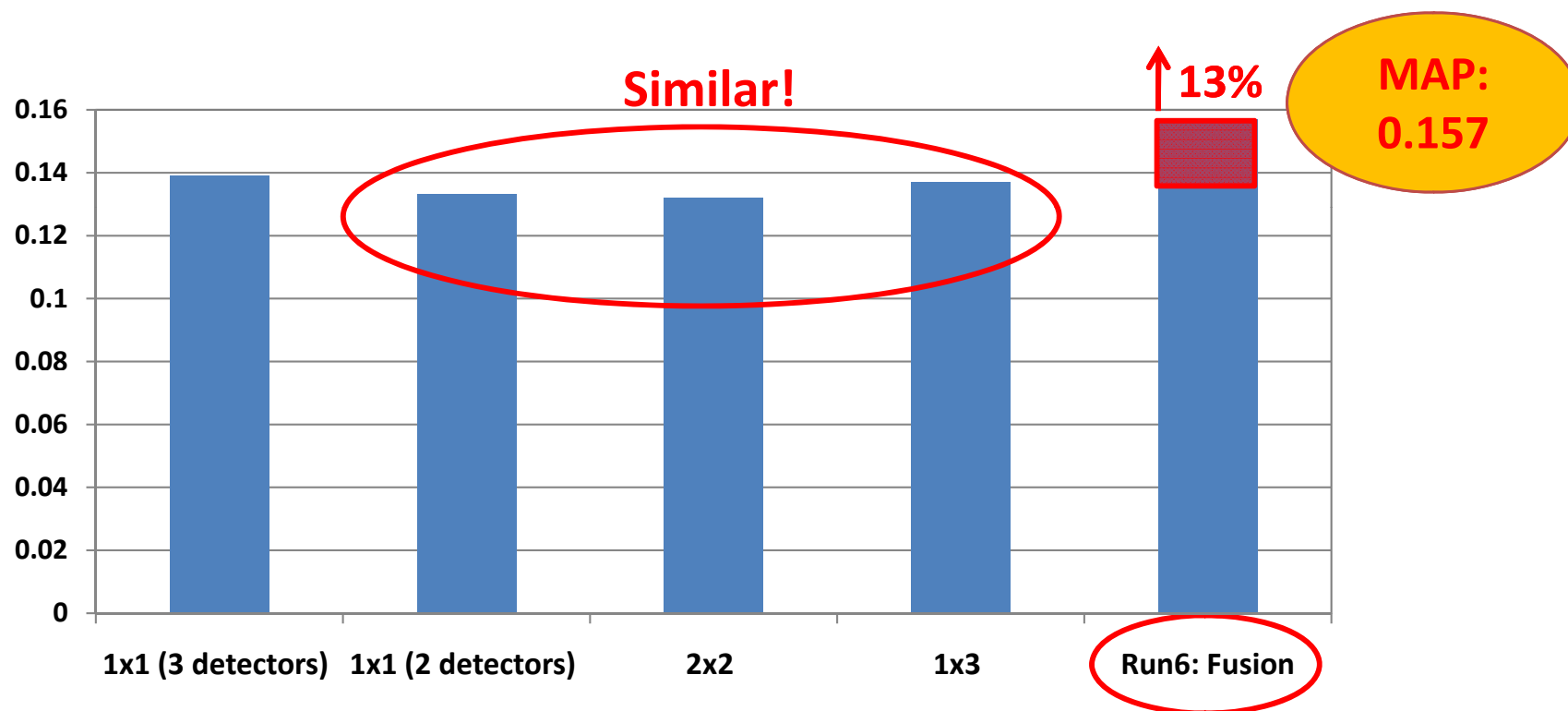
Local Feature Representation Framework



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

misses

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

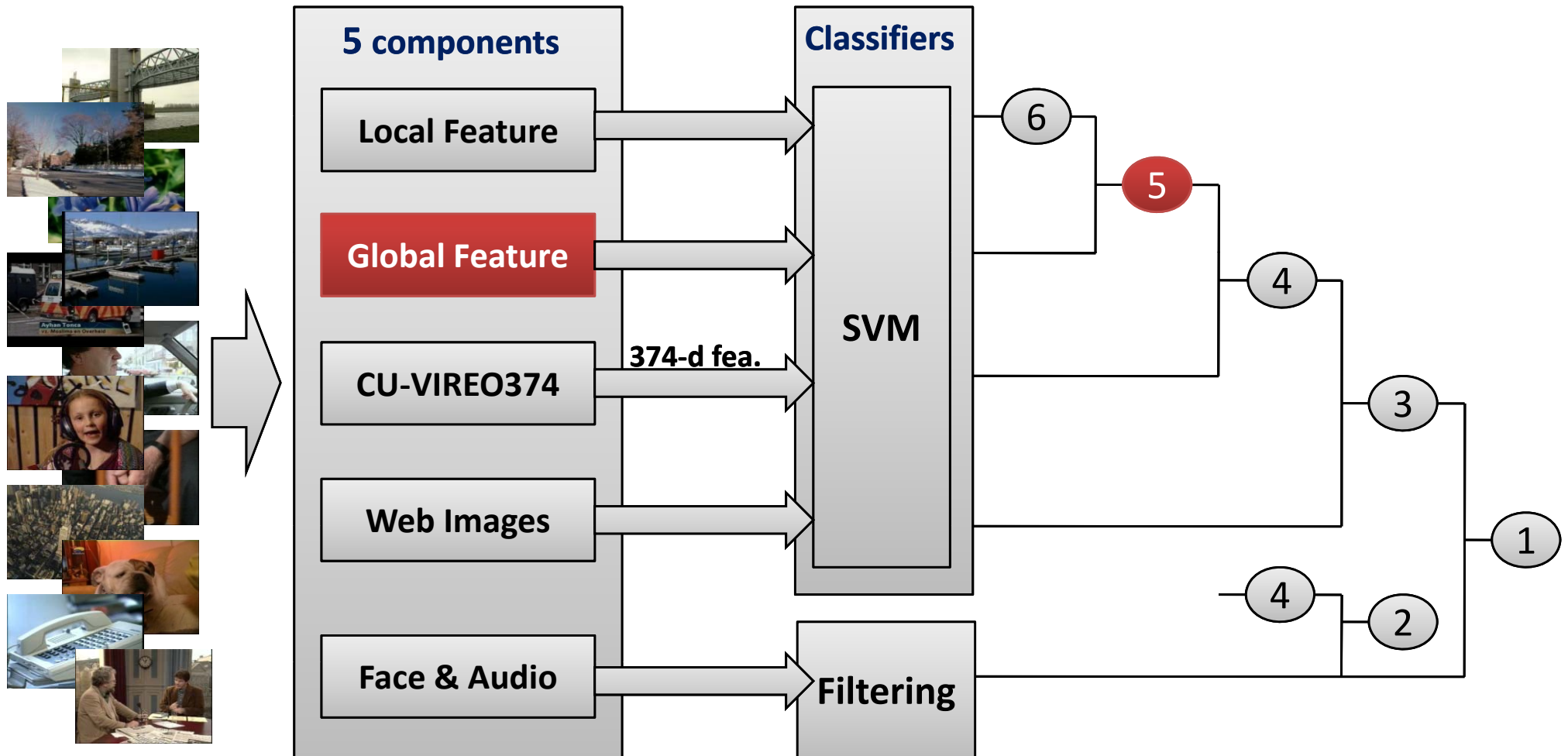
misses



Summary – Local Features

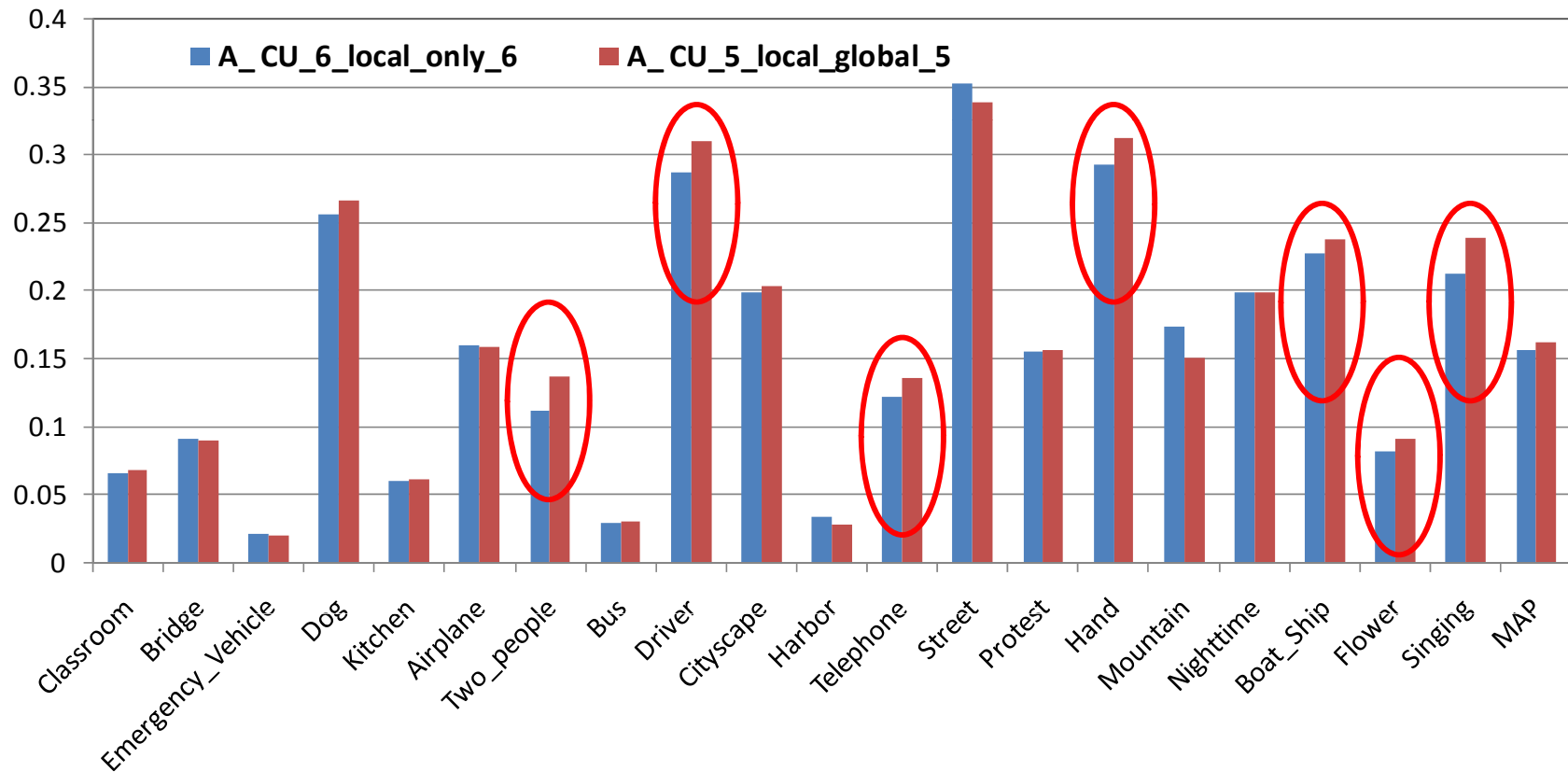
- BoW with good representation choices achieved very impressive performance
 - Soft-weighting is very effective
 - Multiple spatial layouts are useful
 - Multi-detectors do not help much
- Rooms for future improvement
 - Class-specific visual words, feature selection, color-descriptor etc.

Outline

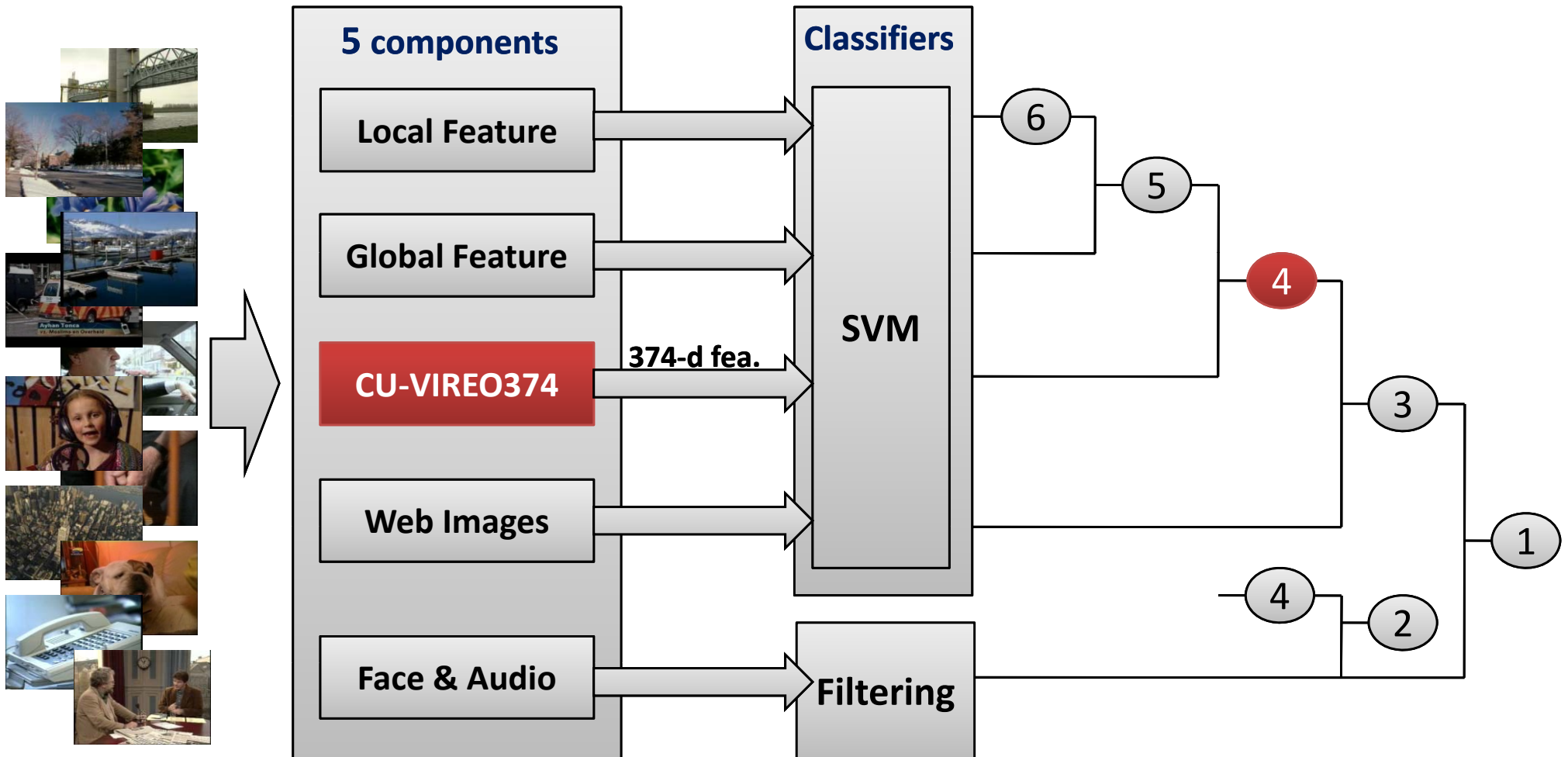


Global Features

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



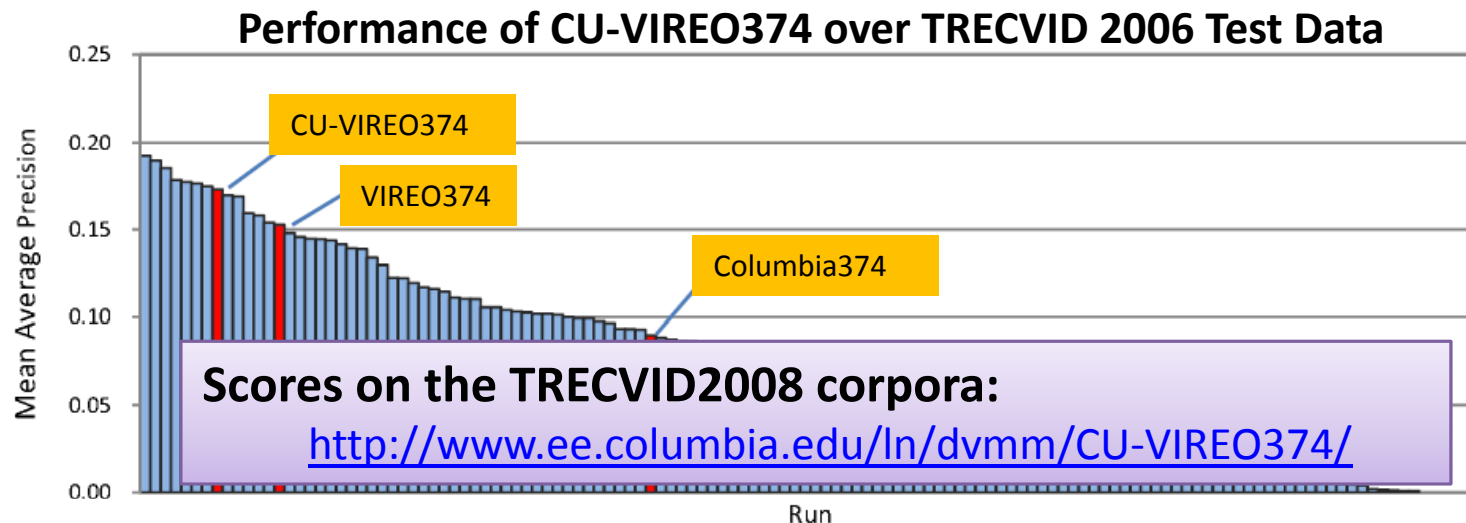
Outline



CU-VIREO374

- Fusion of Columbia374 and VIREO374

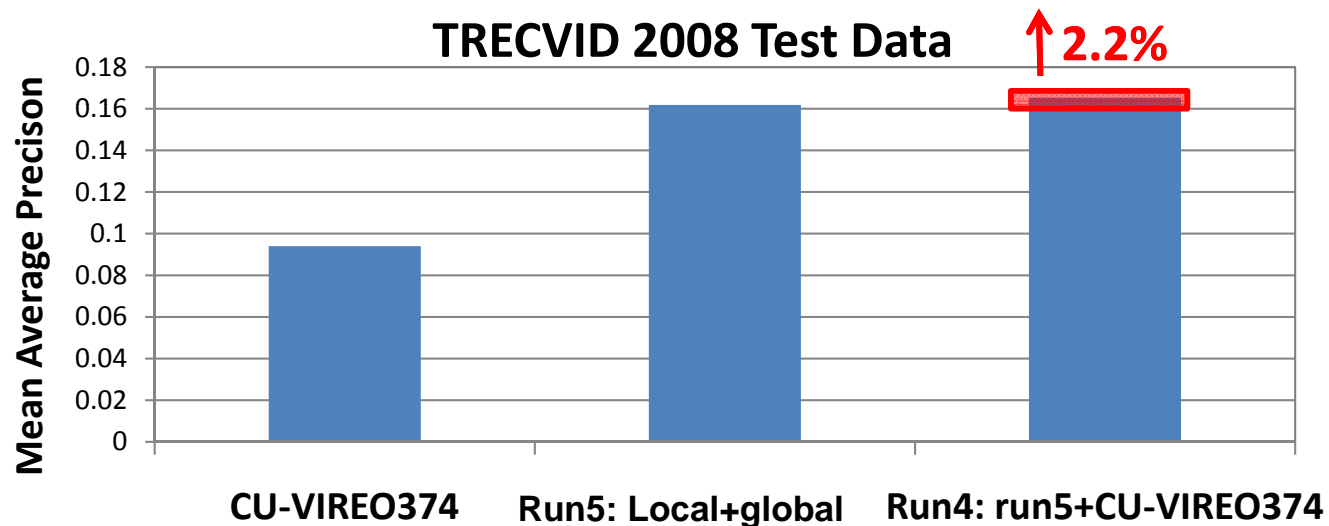
	Feature	Dimension
Columbia374	Grid-based color moment (LUV)	225
	Gabor Texture	48
	Edge Direction Histogram	73
VIREO374	Bag-of-visual-words (soft weighting)	500
	Grid-based Color Moment (<i>Lab</i>)	225
	Grid-based Wavelet Texture	81



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.

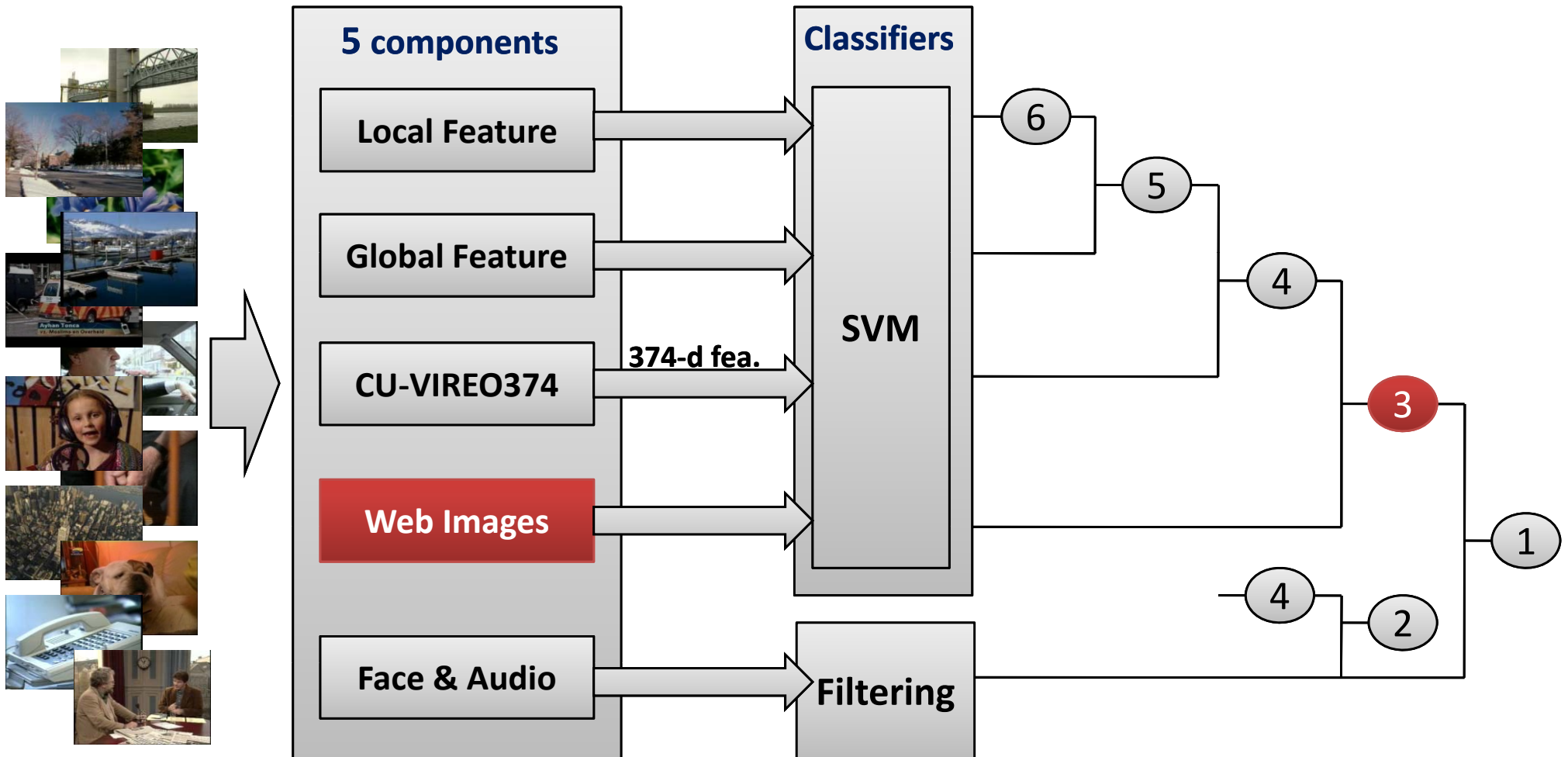
Concept Fusion Using CU-VIREO374

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



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

Outline



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



Flickr Images

Good



Bad

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



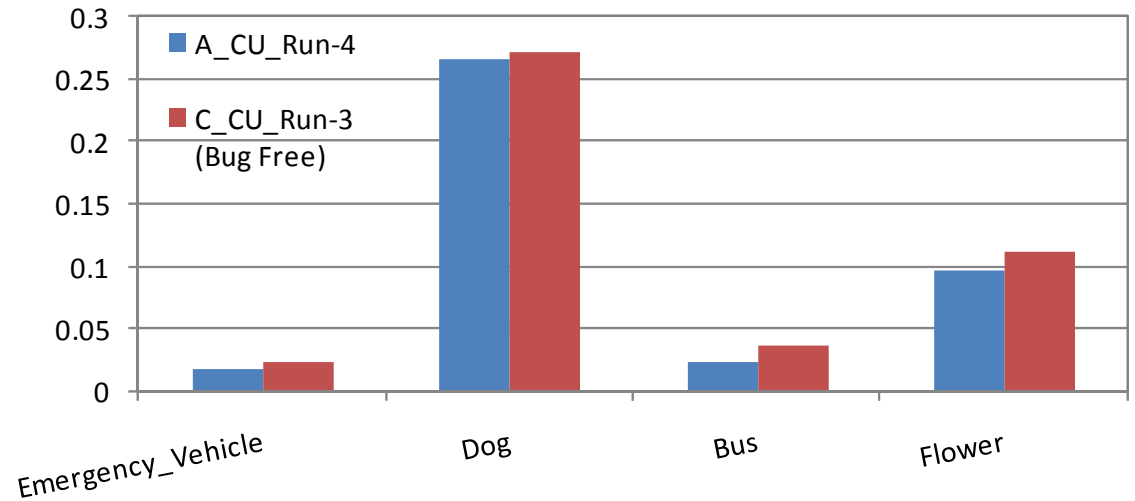
Flickr



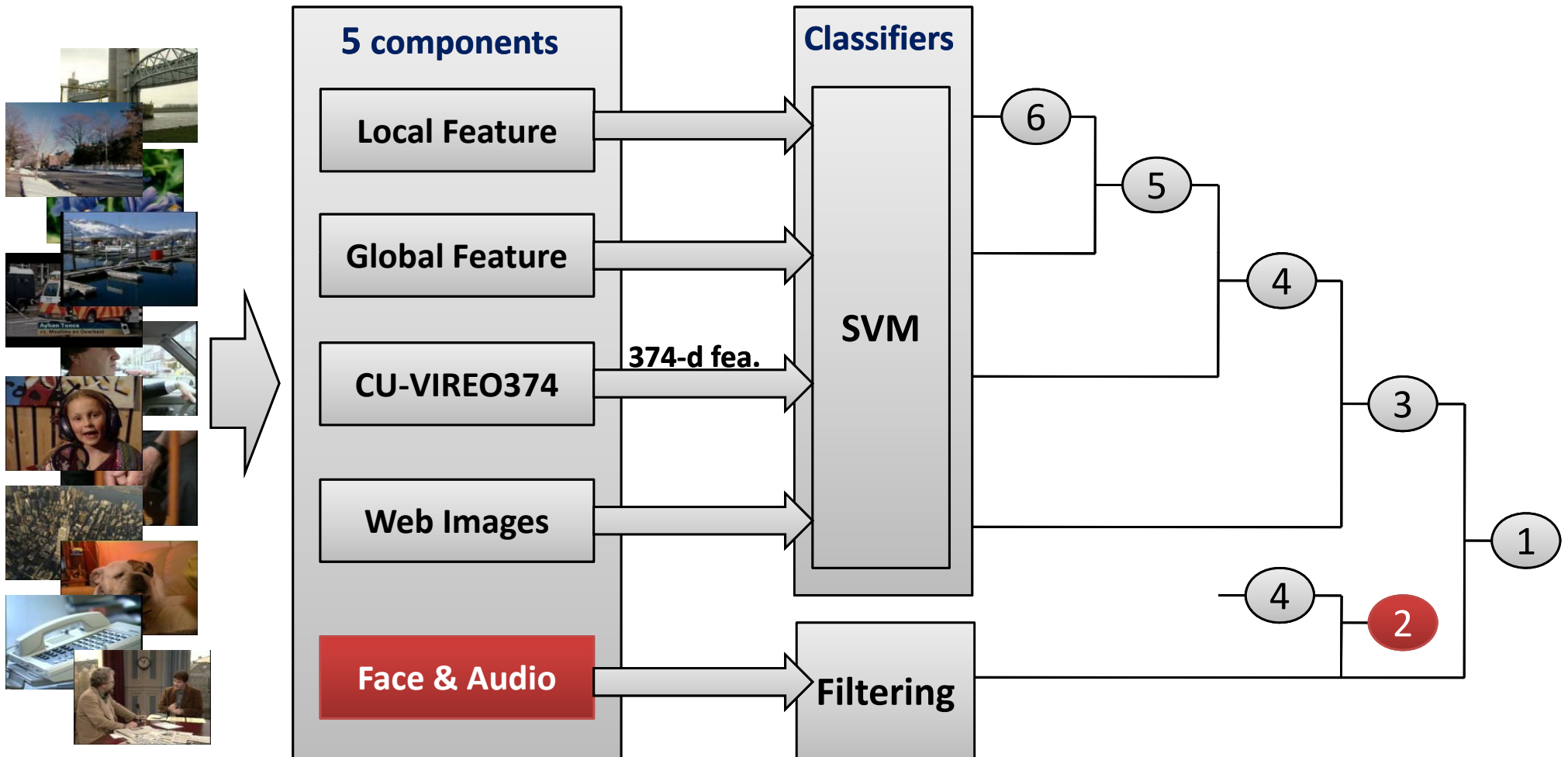
TRECVID

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!

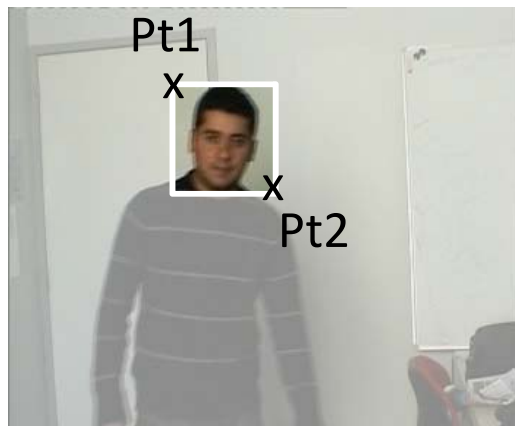


Outline

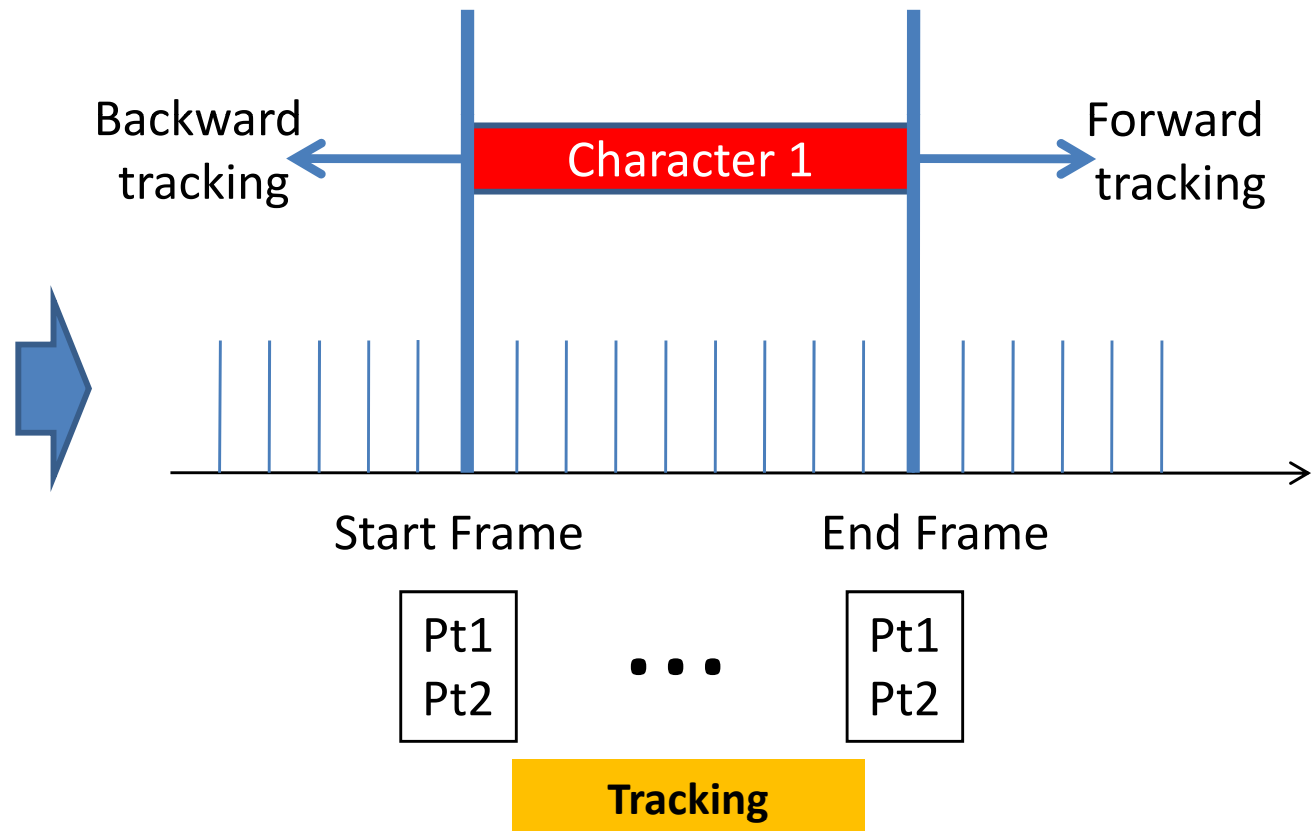


Face Detection and Tracking

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



Face Detection



“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

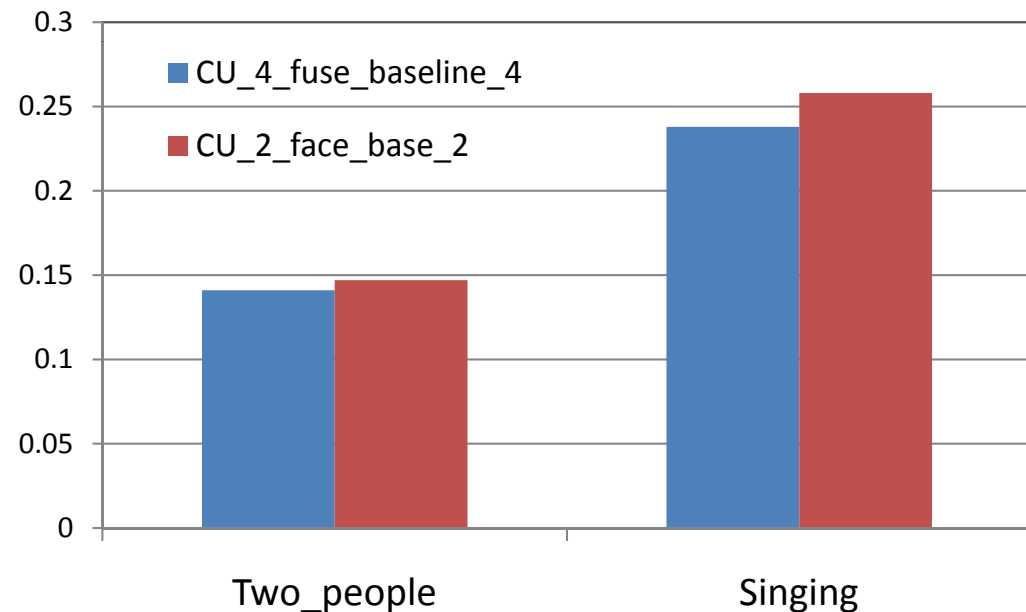
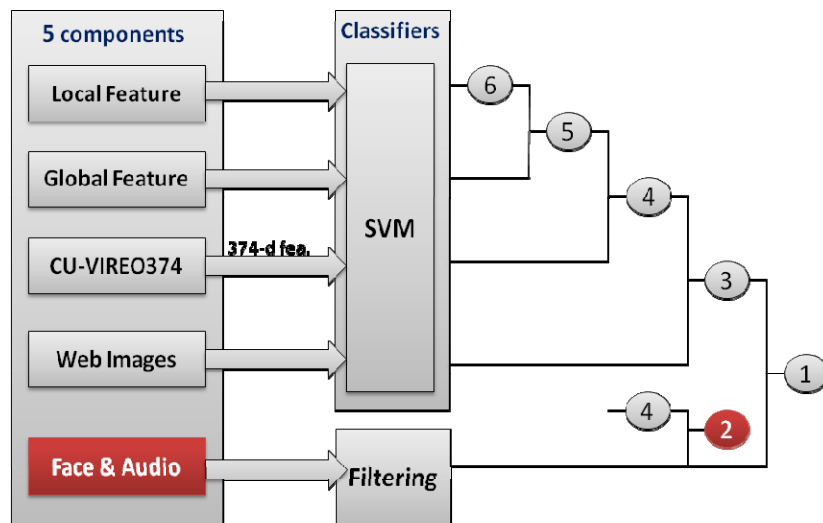
Detecting “Singing” based on Audio

- Vibrato
 - “the variation of the frequency of an musical instrument or of the voice”
- Harmonic Coefficient H_a
 - It corresponds to the most important trigonometric series of the spectrum
 - H_a 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



THANK YOU!

More information at:

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