Context-based Visual Concept Detection Using Domain Adaptive Semantic Diffusion

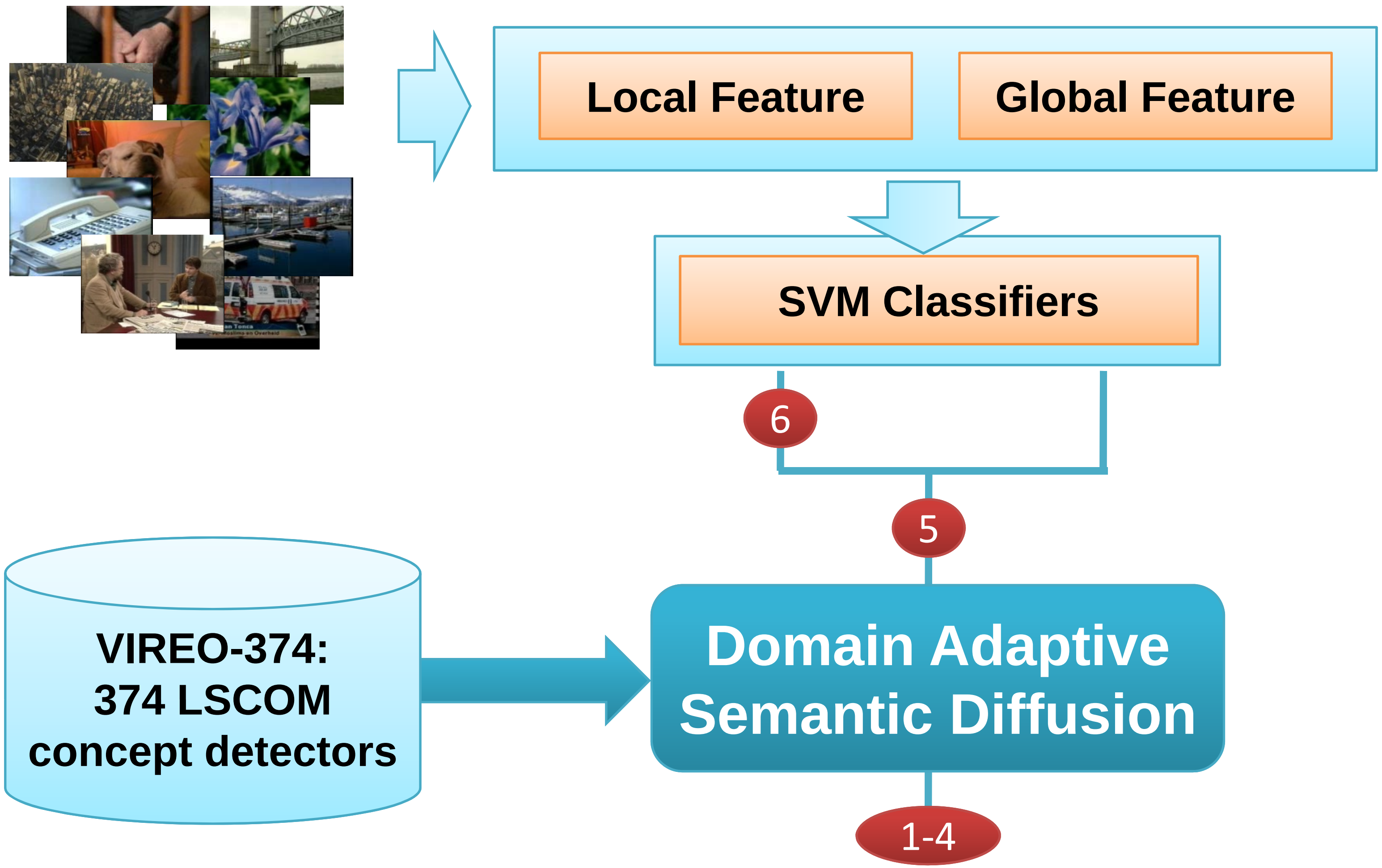
Yu-Gang Jiang[‡], Jun Wang[‡], Shih-Fu Chang[‡], Chong-Wah Ngo[†]

[†] *VIREO Research Group (VIREO), City University of Hong Kong*

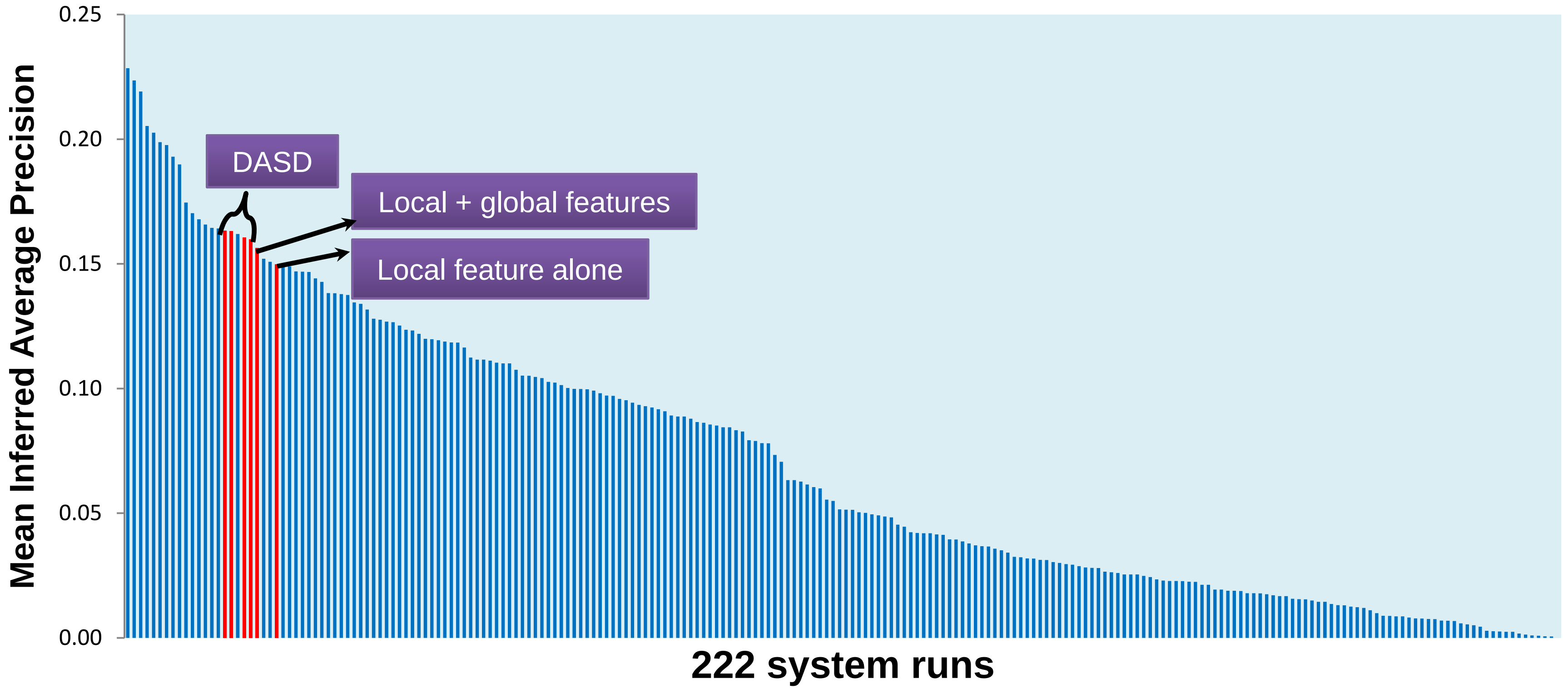
[‡] *Digital Video and Multimedia Lab (DVMM), Columbia University*

NIST TRECVID Workshop, Nov. 2009

Overview: framework

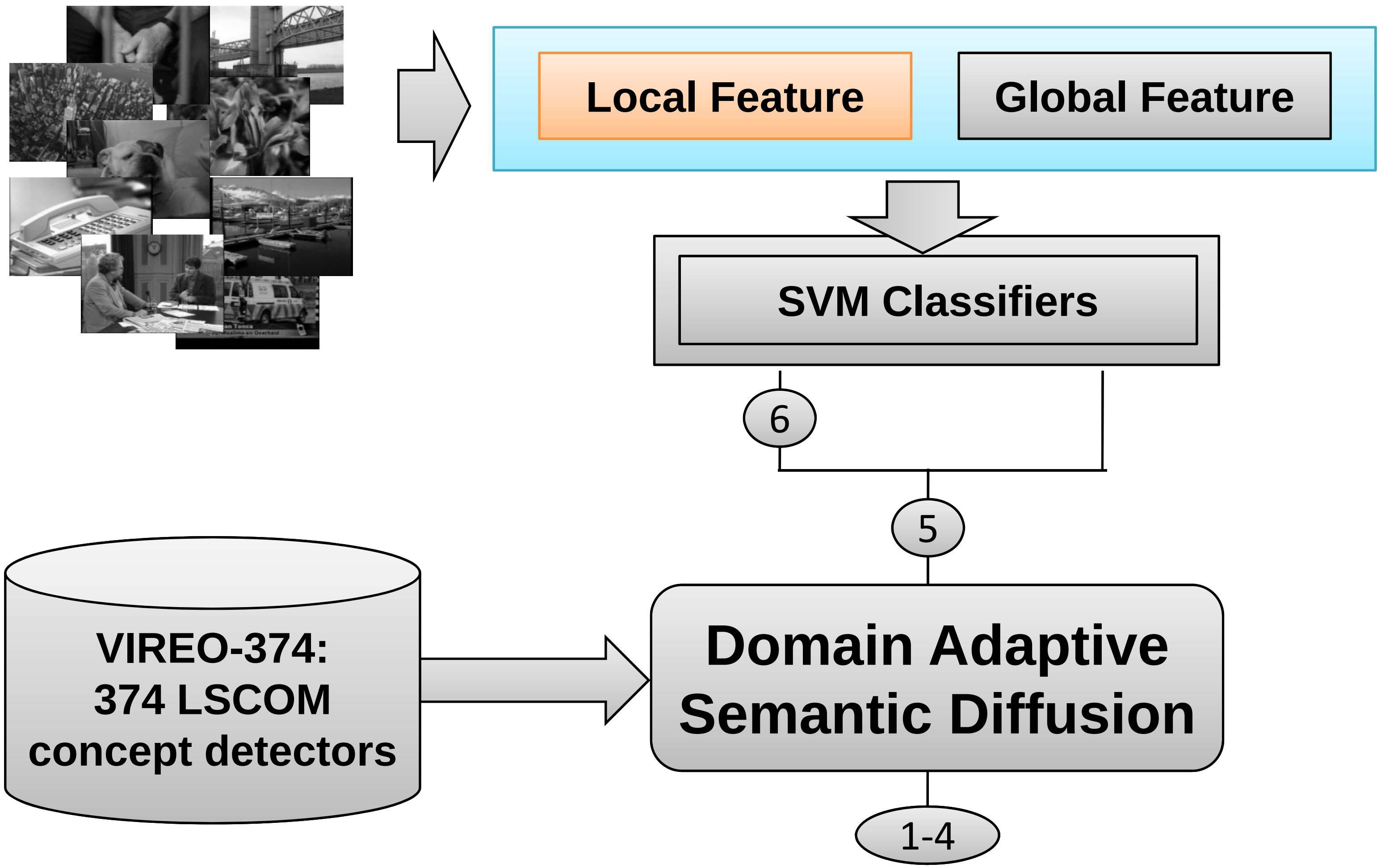


Overview: performance

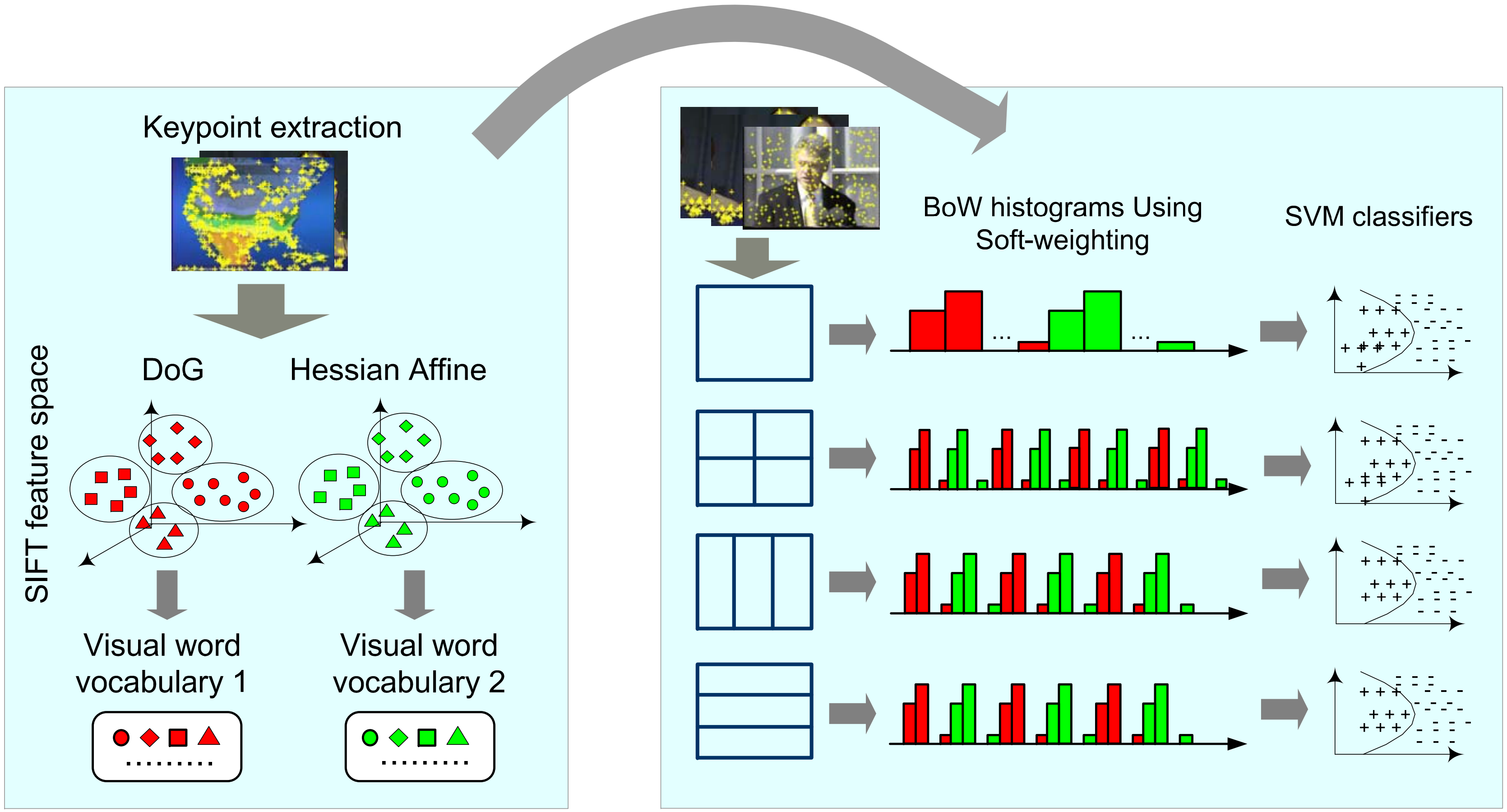


- ❖ Local feature is still the most powerful component (MAP=0.150)
- ❖ Global features help a little bit (MAP=0.156)
- ❖ DASD further contributes incrementally to the final detection

Overview: framework



Local feature representation

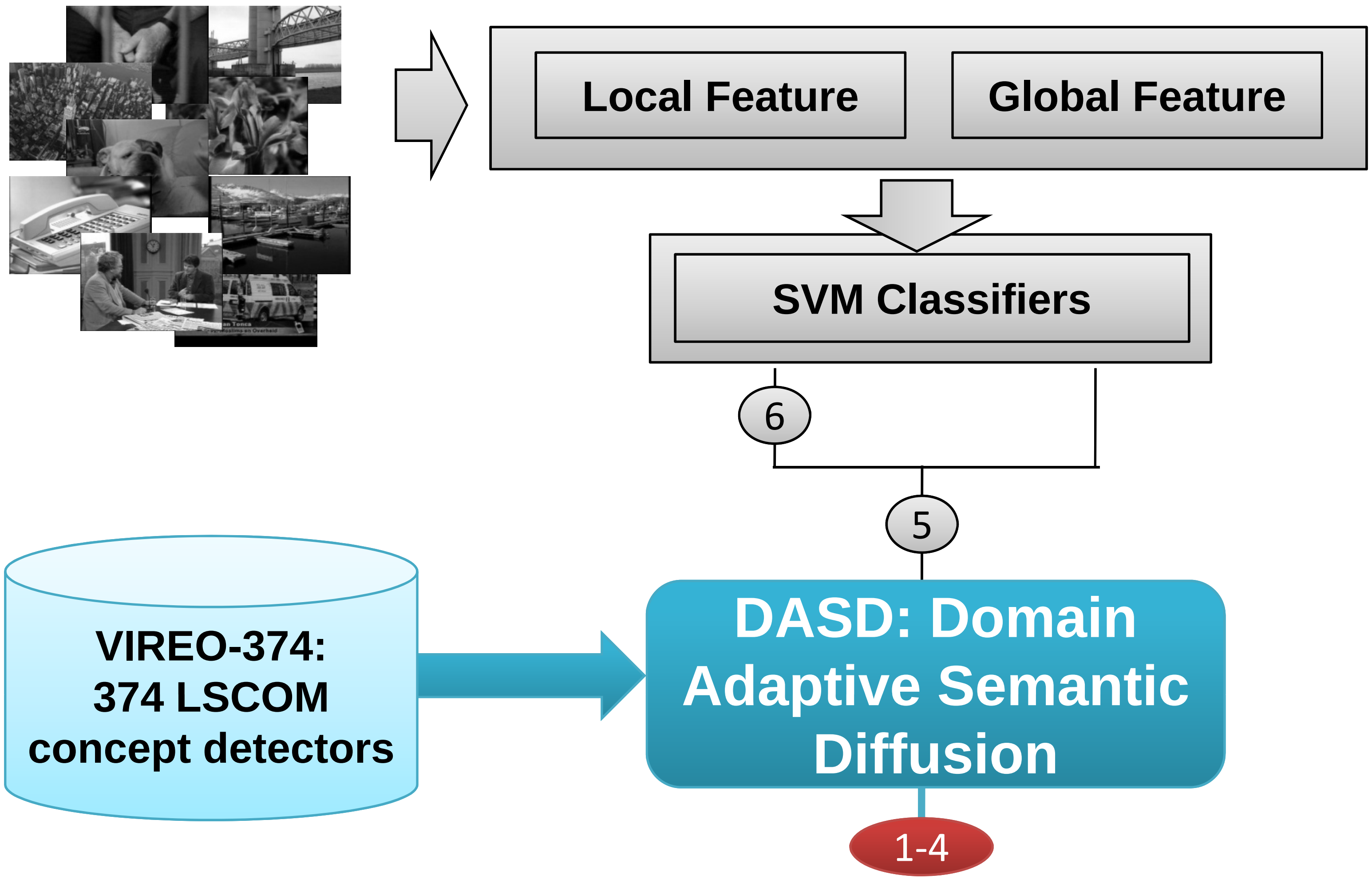


Vocabulary Generation

BoW Representation

Chang et al TRECVID 2008; Jiang, Yang, Ngo & Hauptmann, IEEE TMM, to appear

Context-based concept detection



DASD - motivation

- Most existing methods aim at the assignment of concept labels individually
  - but concepts do not occur in isolation!

DASD - motivation

- Most existing methods aim at the assignment of concept labels individually
  - but concepts do not occur in isolation!
- Domain change between training and testing data was not considered

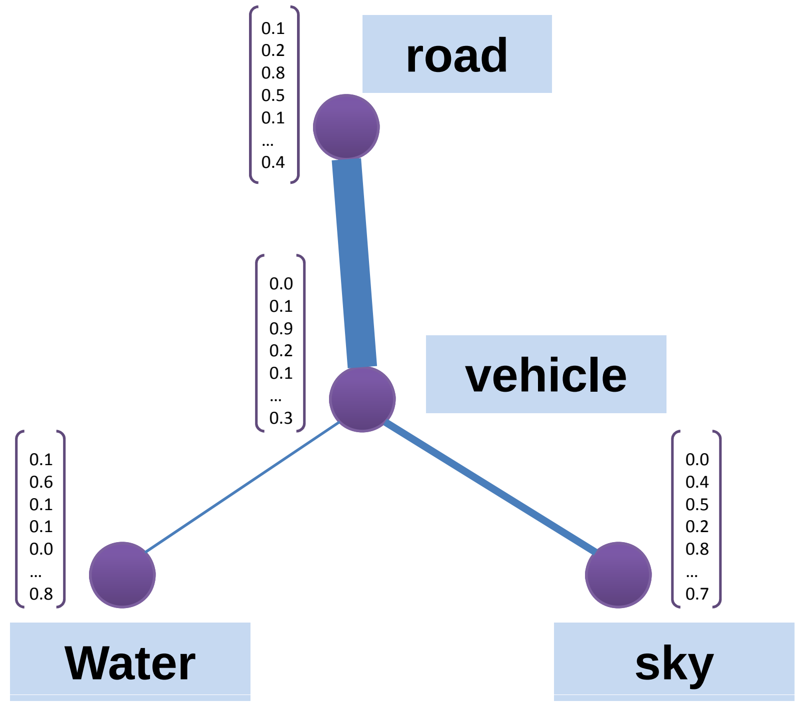
DASD - overview

road	vehicle	water	sky
0.05	0.01	0.11	0.01
0.19	0.12	0.58	0.36
0.80	0.91	0.10	0.53
0.46	0.18	0.13	0.17
0.13	0.05	0.02	0.23

Jiang, Wang, Chang & Ngo, ICCV 2009

DASD - overview

- Domain adaptive semantic diffusion (DASD)
  – Semantic graph
    - Nodes are concepts
    - Edges represent concept correlation
  – Graph diffusion
    - Smooth concept detection scores w.r.t the concept correlation



DASD - formulation

- Energy function

Detection score of concept c_i on test samples

$$\mathcal{E}(g) = \frac{1}{2}\sum_{i=1}^{C}\sum_{j=1}^{C} W_{ij} \left|\left| \frac{g(c_i)}{\sqrt{d(c_i)}} - \frac{g(c_j)}{\sqrt{d(c_j)}} \right|\right|^2$$

Concept affinity

$$(g^*, \tilde{\mathbf{W}}^*) = \arg\min_{g, \tilde{\mathbf{W}}} \mathcal{E}$$

DASD - formulation (cont.)

- Gradually smooth the function makes the detection scores in accordance with the concept relationships

$$\mathcal{E}(g) = \frac{1}{2}\mathrm{tr}(g^T \mathbf{L} g)$$

$$\nabla_g \mathcal{E} = \mathbf{L}g$$

Detection score smoothing process

$$g_t = g_{t-1} - \alpha \nabla_{g_{t-1}} \mathcal{E}$$

DASD - formulation (cont.)

- Graph adaptation

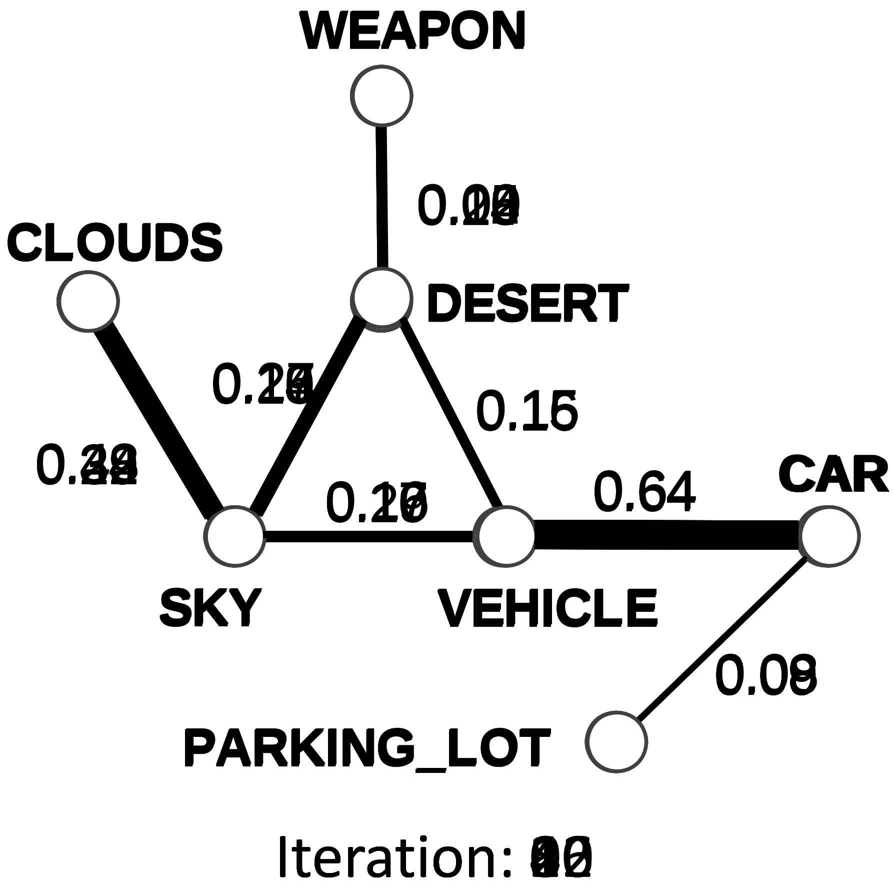
$$\mathcal{E}(g,\tilde{\mathbf{W}}) = \frac{1}{2}\mathrm{tr}(g^T\mathbf{L}g) = \frac{1}{2}\mathrm{tr}\left(g^T\left[\mathbf{I}-\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}\right]g\right)$$
$$= \frac{1}{2}\mathrm{tr}(g^Tg) - \frac{1}{2}\mathrm{tr}\left(g^T\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}g\right)$$
$$= \frac{1}{2}\mathrm{tr}(g^Tg) - \frac{1}{2}\mathrm{tr}\left(g^T\tilde{\mathbf{W}}g\right)$$

$$\frac{\partial\mathcal{E}}{\partial\tilde{\mathbf{W}}} = -gg^T$$

Graph adaptation process

$$\tilde{\mathbf{W}}_t = \tilde{\mathbf{W}}_{t-1} - \beta\frac{\partial\mathcal{E}}{\partial\tilde{\mathbf{W}}_{t-1}}$$

Graph adaptation - example



Broadcast news video domain → Documentary video domain

Experiments on TV ’05-’07

- Baseline detectors
  – VIREO-374
- Graph construction:
  – Ground-truth labels on TRECVID 2005

TRECVID 05/06 (Broadcast News Videos)

TRECVID 07 (Documentary Videos)

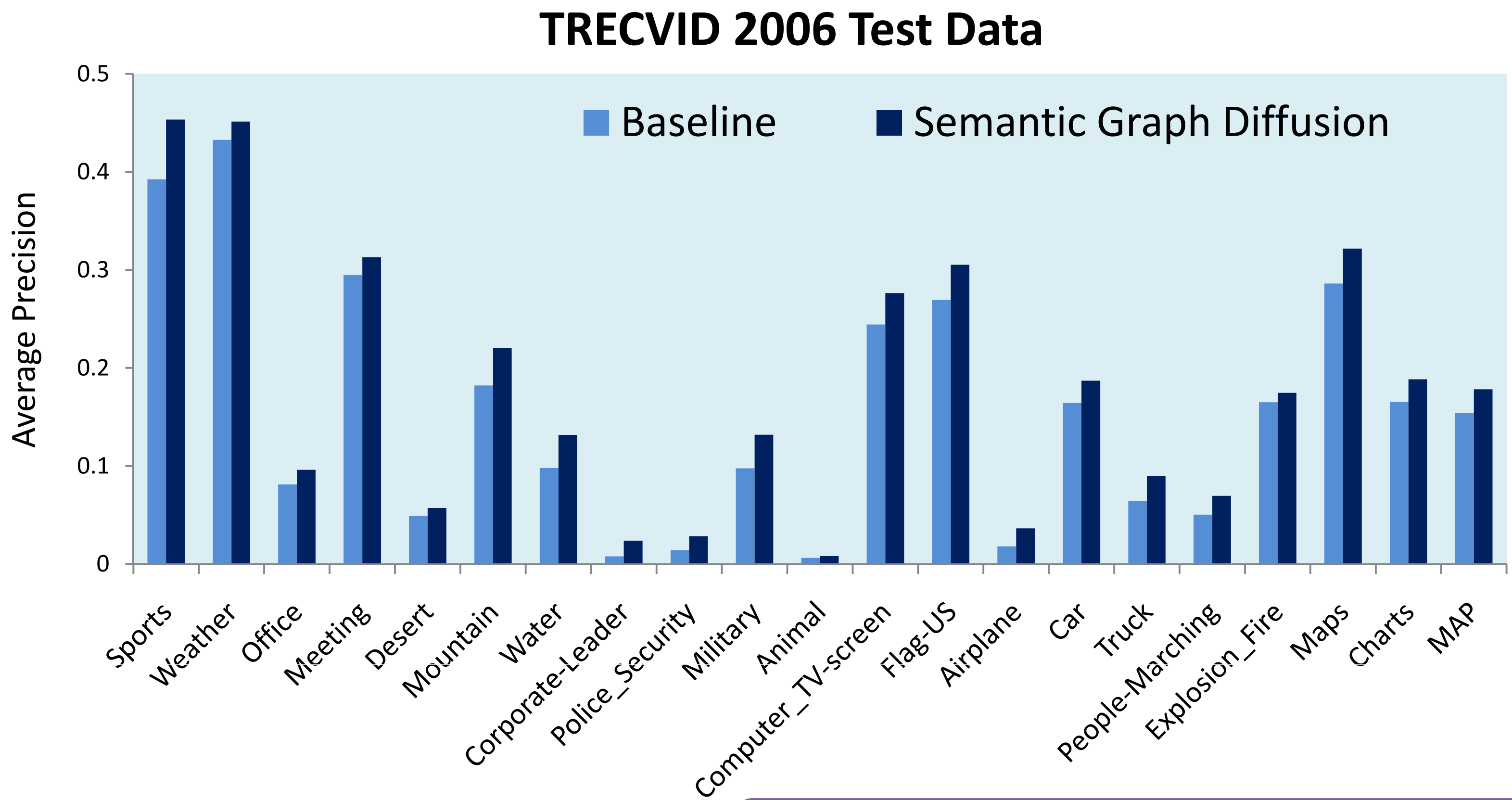
Results on TV ’05-’07

- Performance gain on TRECVID 05-07 Datasets

TRECVID-	2005	2006	2007
# of evaluated concepts	39	20	20
Baseline (MAP)	0.166	0.154	0.099
SD	11.8%	15.6%	12.1%
DASD	<u>11.9%</u>	<u>17.5%</u>	<u>16.2%</u>

- SD: semantic diffusion (without graph adaptation)
  - Consistent improvement over all 3 data sets
- DASD: domain adaptive semantic diffusion
  - Graph adaptation further improves the performance

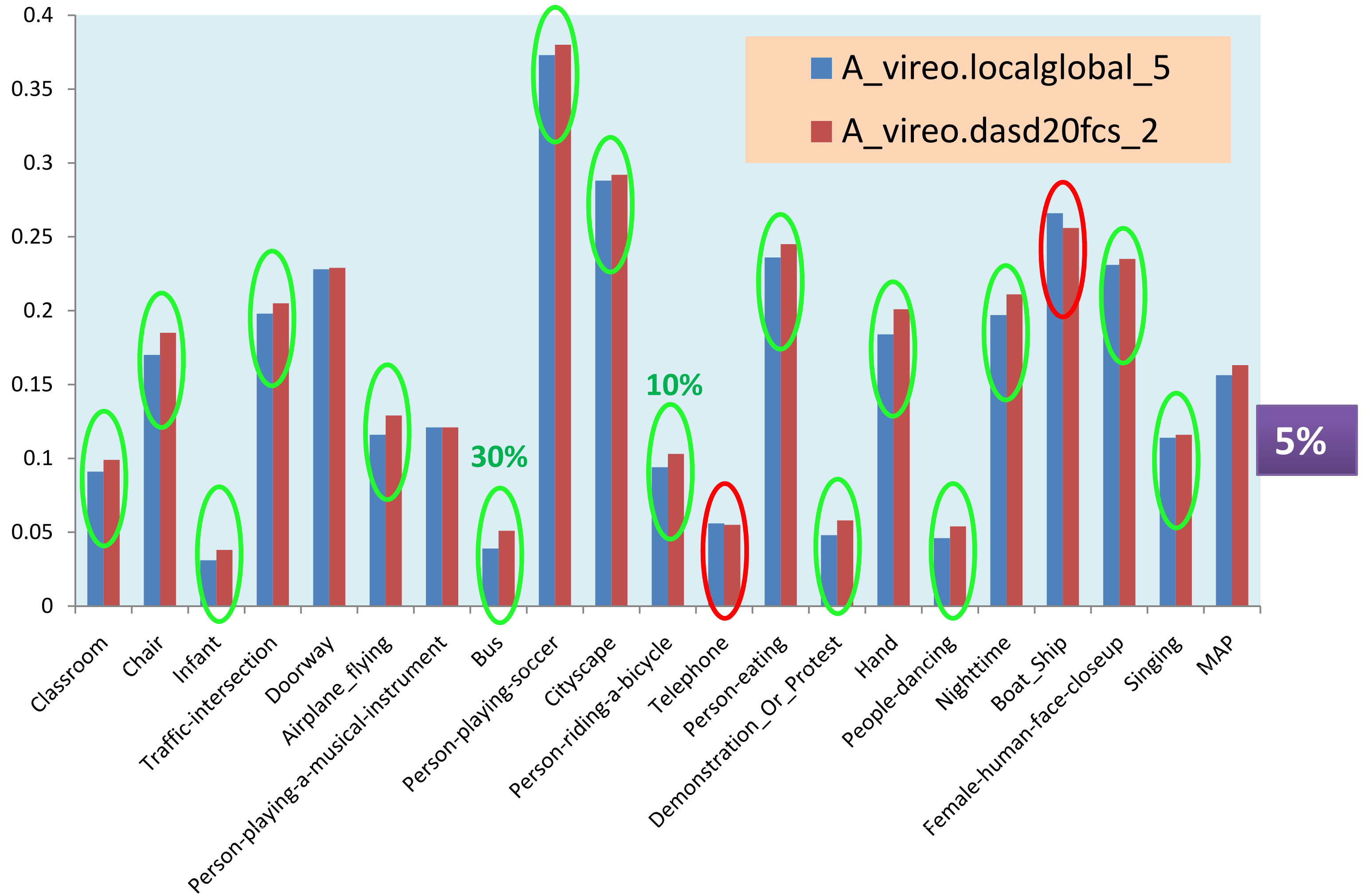
Results on TV '05-'07 (cont.)



Comparison with the state-of-the-arts

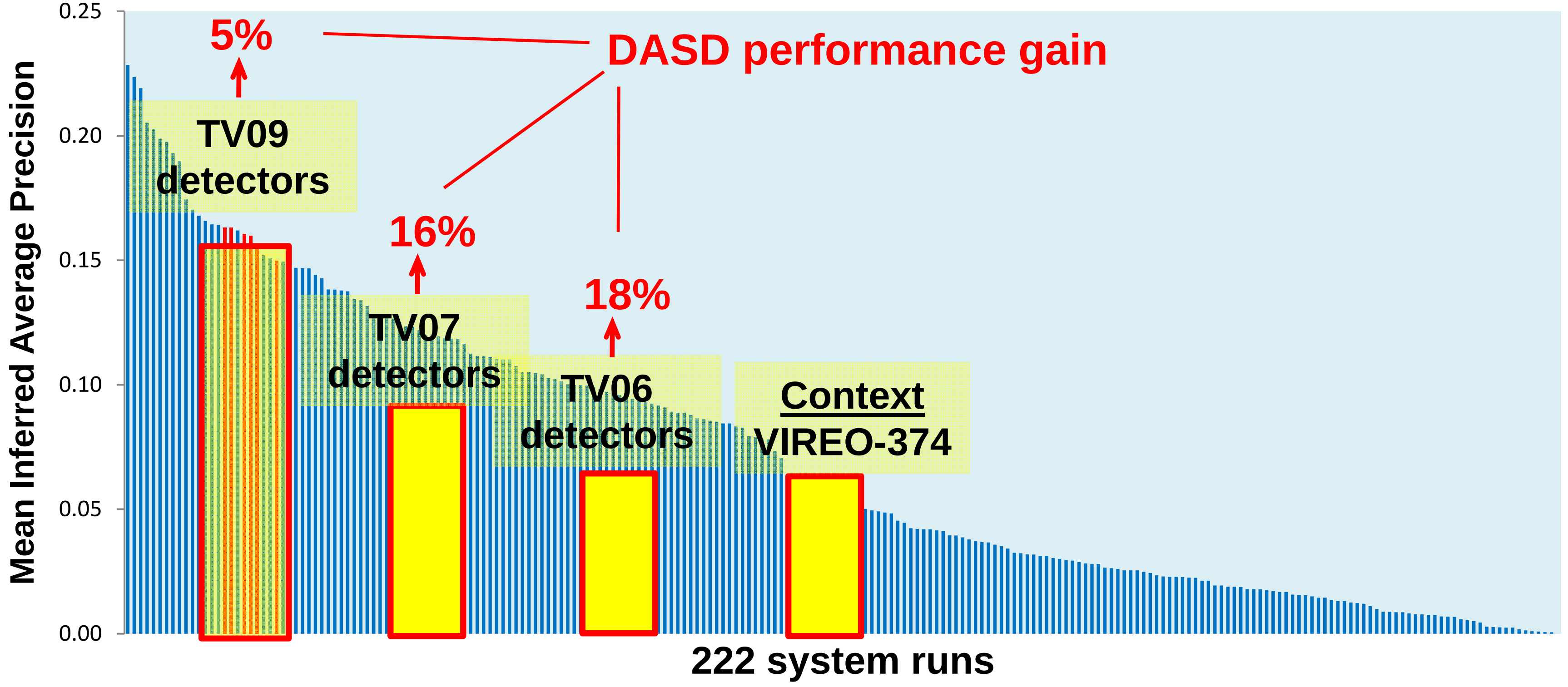
TRECVID	Jiang et al	Aytar et al	Weng et al	DASD
2005	2.2%	4.0%	N/A	11.9%
2006	N/A	N/A	16.7%	17.5%

Results on TRECVID '09



Results on TRECVID '09 (cont.)

- Quality of contextual detectors (VIREO-374)



DASD - computational time

- Complexity is $O(mn)$
  - m: # concepts; n: # video shots
- Only 2 milliseconds per shot/keyframe!

	TRECVID 05	TRECVID 06	TRECVID 07
SD	**59s**	**84s**	**12s**
DASD	**89s**	**165s**	**28s**

Summary

- A well-designed approach using local features achieves good results for concept detection.
- Context information is helpful !
  - Domain adaptive semantic diffusion
    - effective for enhancing concept detection accuracy
    - can alleviate the effect of data domain changes
    - highly efficient !
  - Future directions include:
    - detector reliability: diffusion over directed graph
    - web data annotation: utilize contextual information to improve the quality of tags
  - Source code available for download from DVMM lab research page

THANK YOU!