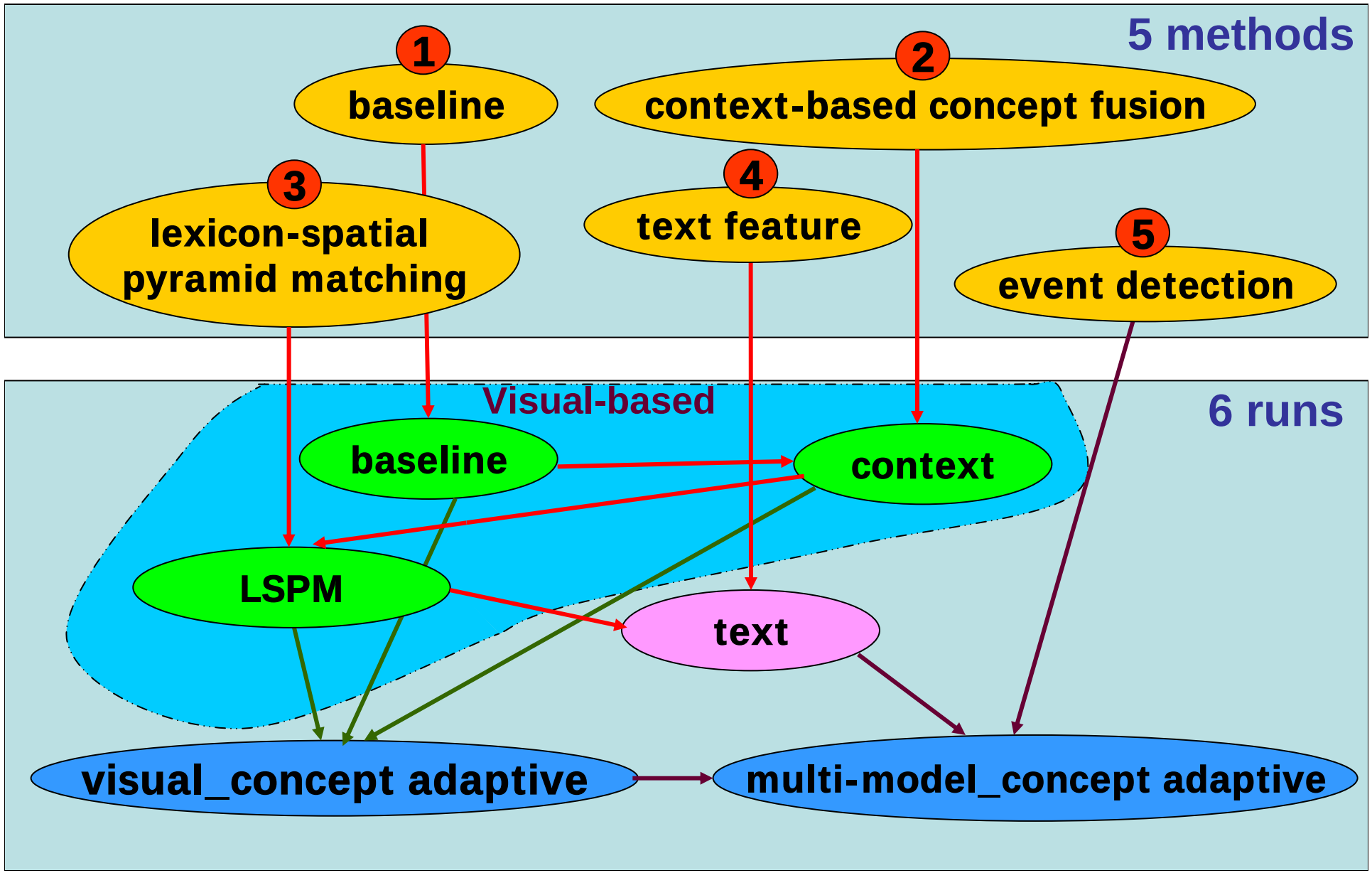
Columbia University TRECVID-2006 High-Level Feature Extraction

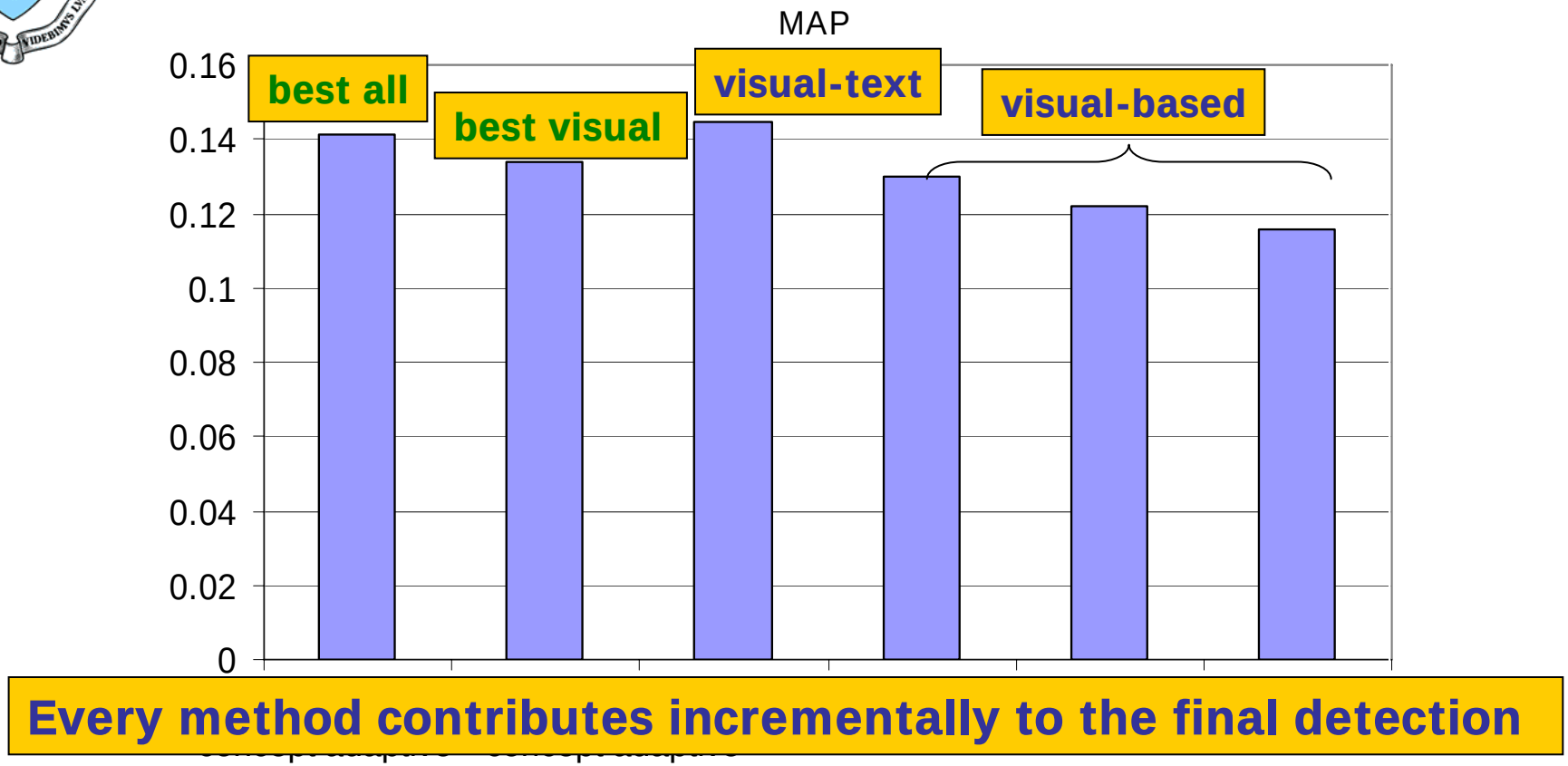
Shih-Fu Chang, Winston Hsu, Wei Jiang,
Lyndon Kennedy, Dong Xu,
Akira Yanagawa, and Eric Zavesky

Digital Video and Multimedia Lab, Columbia University
http://www.ee.columbai.edu/dvmm

Overview – 5 methods & 6 submitted runs



Overview – performance



Every method contributes incrementally to the final detection

- context > baseline

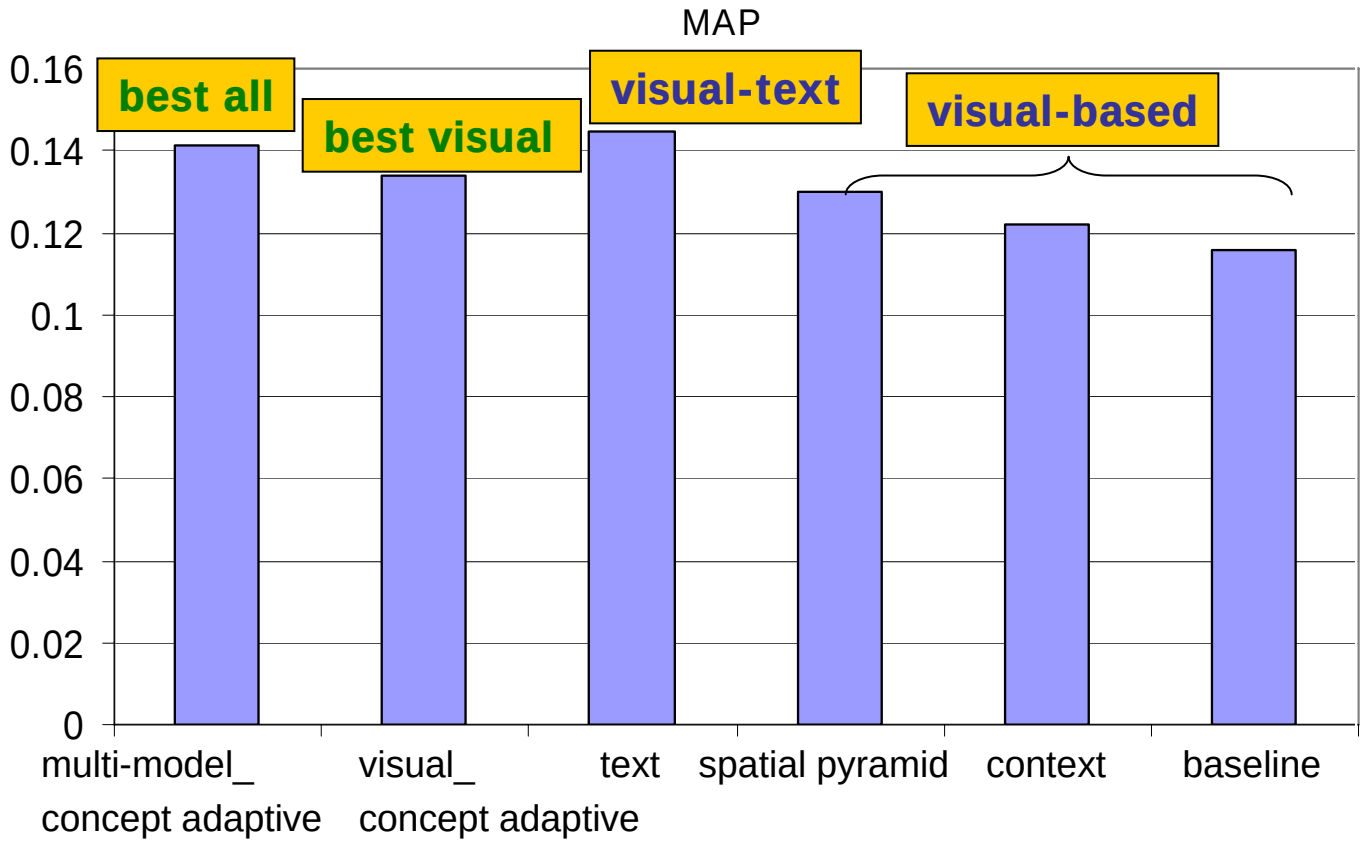
context-based concept fusion (**CBCF**) improves baseline

- LSPM > context

lexicon-spatial pyramid matching (**LSPM**) further improves detection

- text > LSPM: text features improve visual

Overview – performance



visual_concept adaptive > LSPM (also > context > baseline):

best of visual selection works

text > multi-model_concept adaptive:

best of all selection does not work well

probably due to over fitting of text tool

Outline – New Algorithms

- Baseline
- Context-based concept fusion (CBCF)
- Lexicon-spatial pyramid matching (LSPM)
- Text features
- Event detection

Outline – New Algorithms

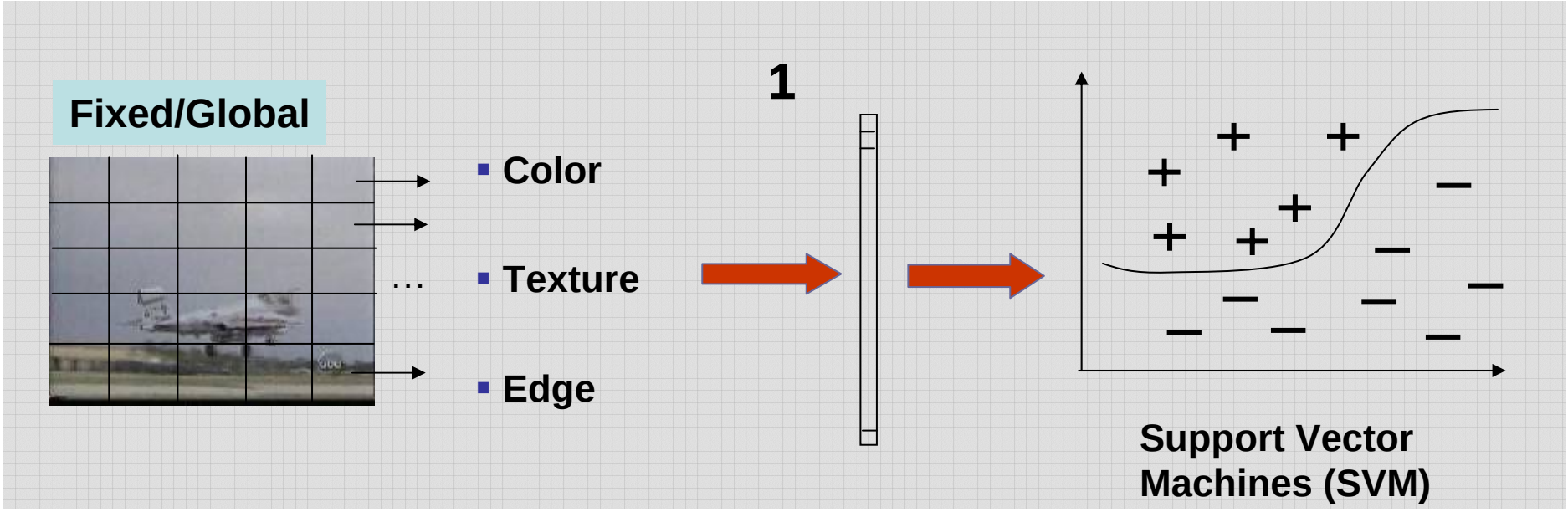
- Baseline
- Context-based concept fusion (CBCF)
- Lexicon-spatial pyramid matching (LSPM)
- Text features
- Event detection

Individual Methods: (1) Baseline

Average fusion of two SVM baseline classification results

Based on 3 visual features

- color moments over 5x5 fixed grid partitions
- Gabor texture
- edge direction histogram from the whole image



coarse local features, layout, and global appearance

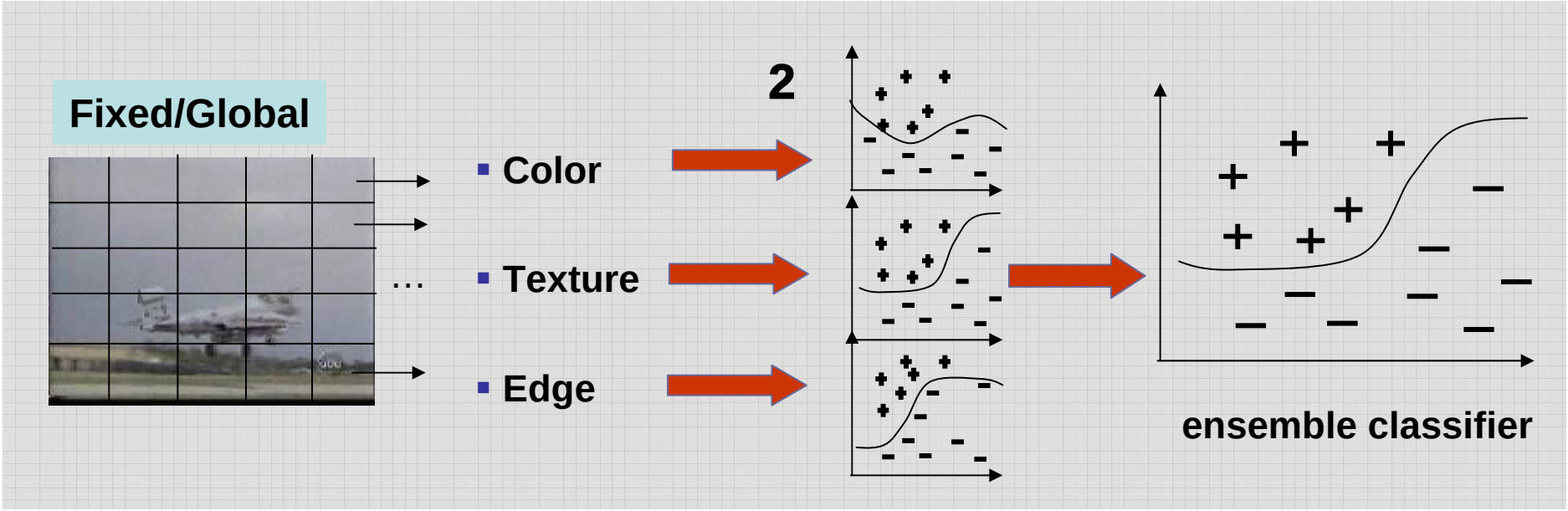
Individual Methods: (1) Baseline

Average fusion of two SVM baseline classification results

Based on 3 visual fea

- color moments o
- Gabor texture
- edge direction histogram fro

Features and models available for download soon!



Yanagawa et al., Tec. Rep., Columbia Univ., 2006 ,
http://www.ee.columbia.edu/dvmm/newPublication.htm

DVMM lab digital video | multimedia laboratary

Outline – New Algorithms

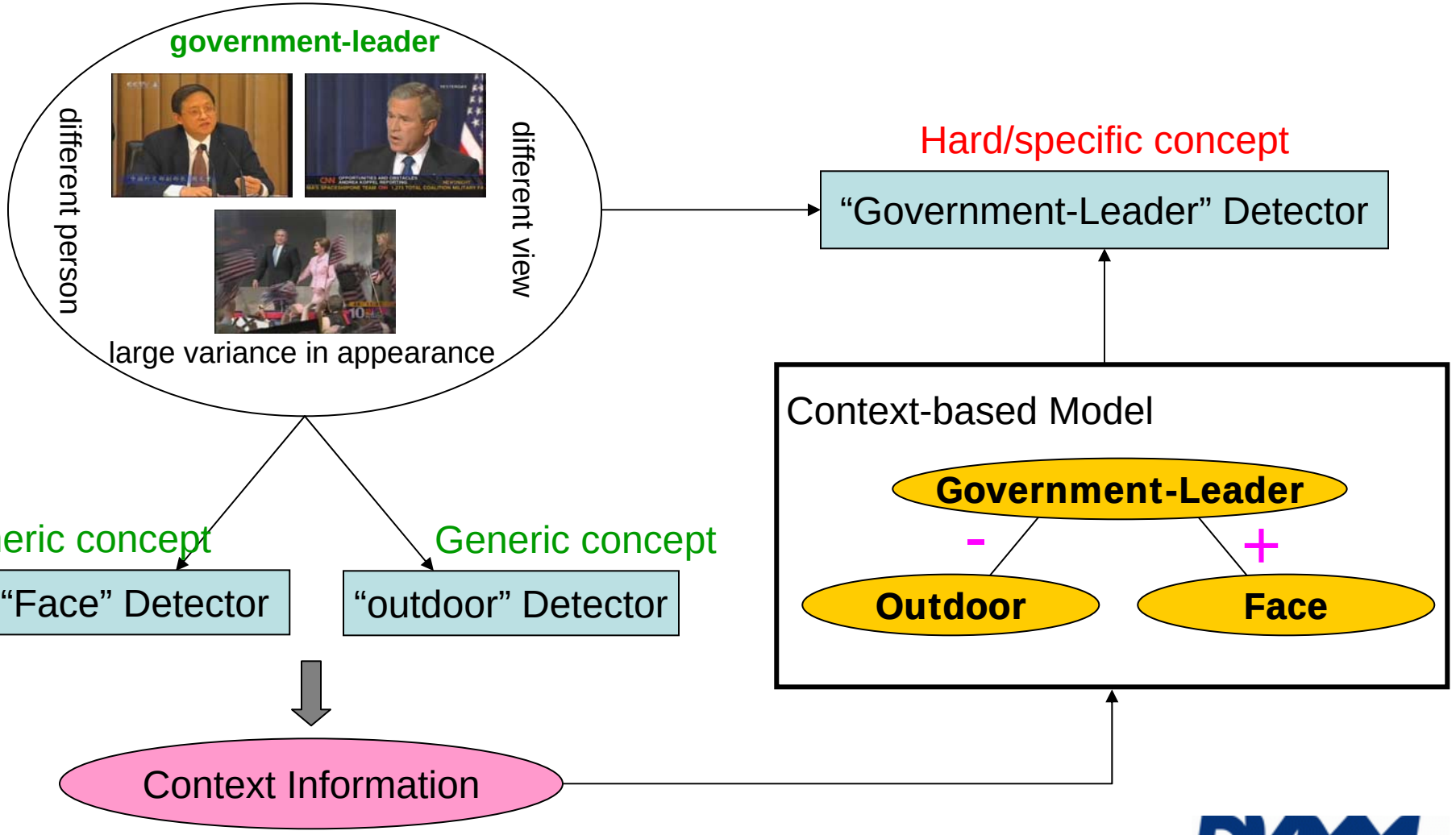
- Baseline
- Context-based concept fusion (CBCF)
- Lexicon-spatial pyramid matching (LSPM)
- Text features
- Event detection

Outline – New Algorithms

- Baseline
- Context-based concept fusion (CBCF)
- Lexicon-spatial pyramid matching (LSPM)
- Text features
- Event detection

Individual Methods: (2) CBCF

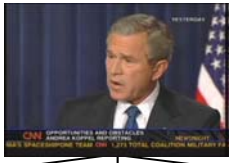
Background on Context Fusion



DVMM lab
digital video | multimedia laboratary

Individual Methods: (2) CBCF

Formulation



outdoor detector | **government-leader detector** | **face detector**

$P(\text{outdoor}|\text{image})$ $P(\text{government-leader}|\text{image})$ $P(\text{face}|\text{image})$

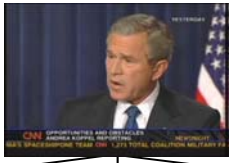
context-based model

(Naphade et al 2002)

$\tilde{P}(\text{outdoor}|\text{image})$ $\tilde{P}(\text{government-leader}|\text{image})$ $\tilde{P}(\text{face}|\text{image})$

Individual Methods: (2) CBCF

Our approach: Discriminative + Generative



I

C_1 C_2 C_3

outdoor detector | **government-leader detector** | **face detector**

$P(\text{outdoor}|\text{image})$ x_1 $P(\text{government-leader}|\text{image})$ x_2 $P(\text{face}|\text{image})$ x_3

observation

Conditional Random Field (Jiang, Chang, et al ICIP 2006)

outdoor **airplane** **office**

updated posteriors

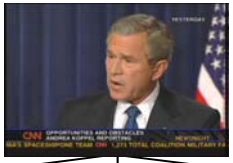
$\tilde{P}(\text{outdoor}|\text{image})$ $\tilde{P}(\text{government-leader}|\text{image})$ $\tilde{P}(\text{face}|\text{image})$

$p(y_1 = 1 | \mathbf{X})$ $p(y_2 = 1 | \mathbf{X})$ $p(y_3 = 1 | \mathbf{X})$

DVMM lab digital video | multimedia laboratary

Individual Methods: (2) CBCF

Our approach: Discriminative + Generative



I

C_1 **outdoor detector**

C_2 **government-leader detector**

C_3 **face detector**

$P(\text{outdoor}|\text{image})$ x_1

$P(\text{government-leader}|\text{image})$ x_2

$P(\text{face}|\text{image})$ x_3

observation

Conditional Random Field

min $J = -\prod_{I}\prod_{C_i} p(y_i = 1|\mathbf{X})^{(1+y_i)/2}\, p(y_i = -1|\mathbf{X})^{(1-y_i)/2}$

iteratively minimized by boosting

updated posteriors

$\tilde{P}(\text{outdoor}|\text{image})$ $p(y_1 = 1|\mathbf{X})$

$\tilde{P}(\text{government-leader}|\text{image})$ $p(y_2 = 1|\mathbf{X})$

$\tilde{P}(\text{face}|\text{image})$ $p(y_3 = 1|\mathbf{X})$

DVMM lab digital video I multimedia laboratary

Individual Methods: (2) CBCF

During each it

two

pt:

1. Using input

min $J = -\prod_i p(y_i = 1 \mid \mathbf{x})^{(1+y_i)/2} p(y_i = -1 \mid \mathbf{x})^{(1-y_i)/2}$

2. Using

ration t-1

iteratively

Classifier 2 keeps updating through iteration
And captures inter-conceptual influences

Without classifier 2, Traditional AdaBoost

Individual Methods: (2) CBCF

Database & lexicon for context

- Predefined lexicon to provide context

  -- 374 concepts from LSCOM ontology (**observation**)
  airplane, building, car, boat, person, outdoor, sports, etc

- Independent detector

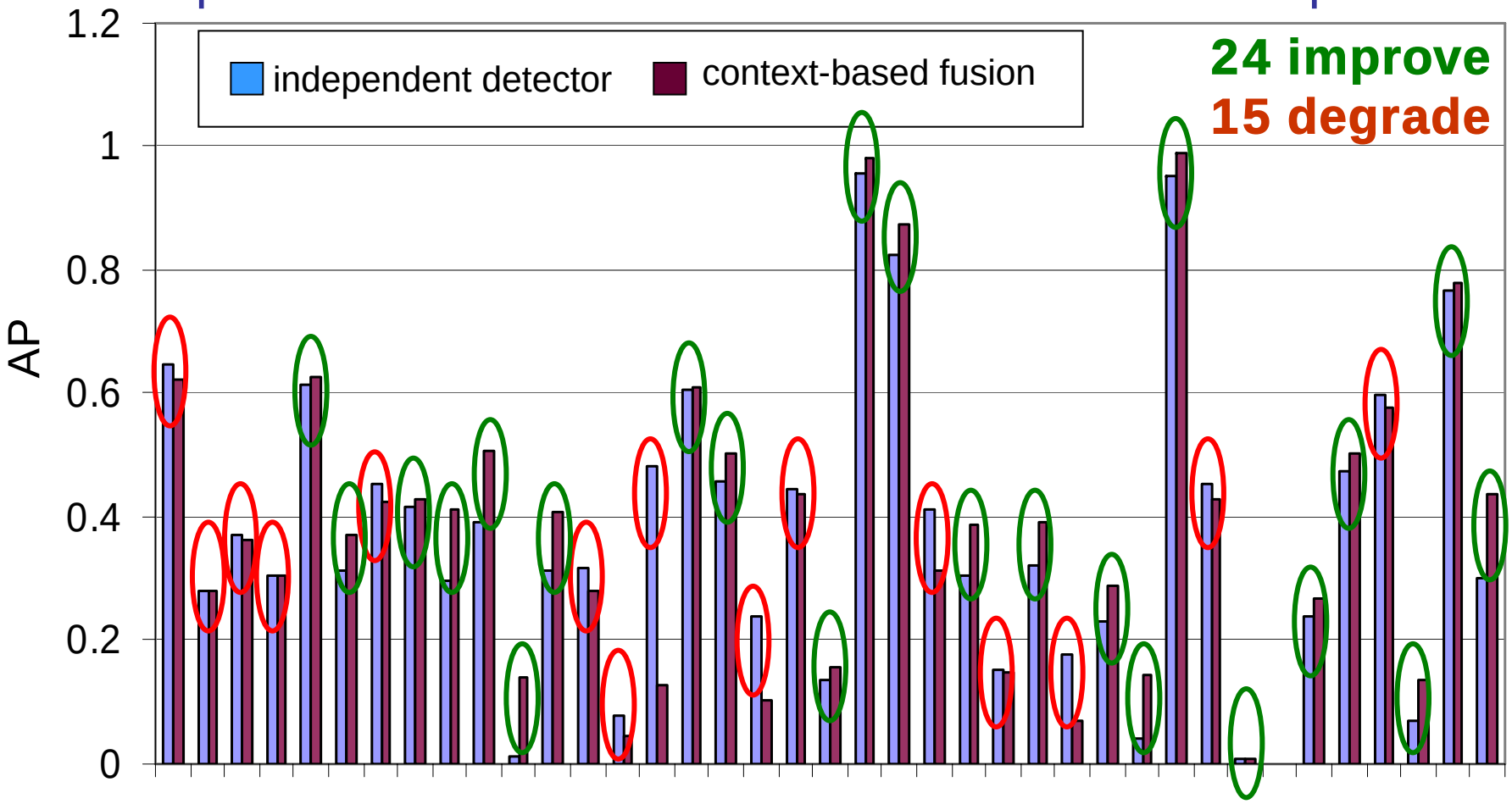
  -- our baseline

- Test concepts

  -- the 39 concepts defined by NIST (**update posteriors**)

Individual Methods: (2) CBCF

experimental results over TRECVID 2005 development set



Selective Application of Context

- Not every concept classification benefits from context-based fusion

  Consistent with previous context-based fusion:

  IBM: no more than 8 out of 17 concepts gained performance
  [Amir et al., TRECVID Workshop, 2003]

  Mediamill: 80 out of 101 concepts
  [Snoek et al., TRECVID Workshop, 2005]

- Is there a way to predict when it works?

Predict When Context Helps

Why CBCF may not help every concept ?

- Complex inter-conceptual relationships vs. limited training samples
- Strong classifiers may suffer from fusion with weak context

Avoid using CBCF for C_i if C_i is strong and with weak context

Use CBCF for concept C_i if C_i is weak or with strong context

$I(C_i;C_j)$ -- mutual information between C_i and C_j

$E(C_i)$ -- error rate of independent detector for C_i

$$\frac{\sum_{C_j, j\neq i} I(C_j;C_i)E(C_j)}{\sum_{C_j, j\neq i} I(C_j;C_i)} < \beta$$

Strong context

or

$$E(C_i) > \lambda$$

weak concept

Predict When Context Helps

Change parameters to predict different number of concepts

# predicted	# concept improved	precision of prediction	MAP gain
39	24	62%	3.0%
20	15	75%	9.5%
16	14	88%	14%
9	9	100%	7.2%

Example

Military

Fighter_Combat

Individual

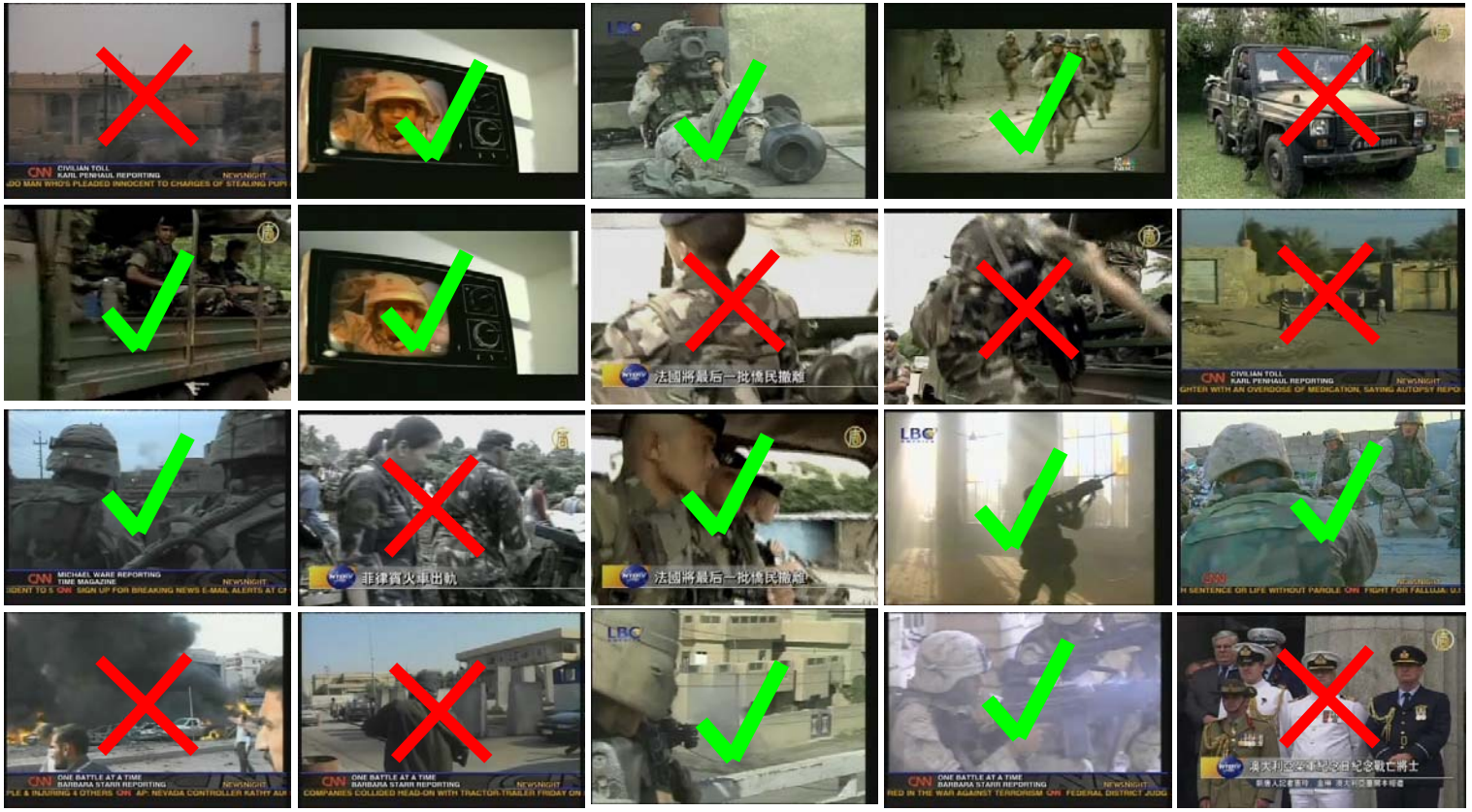
House

. . .

DVMM lab
digital video | multimedia laboratary

Example

Independent Detector

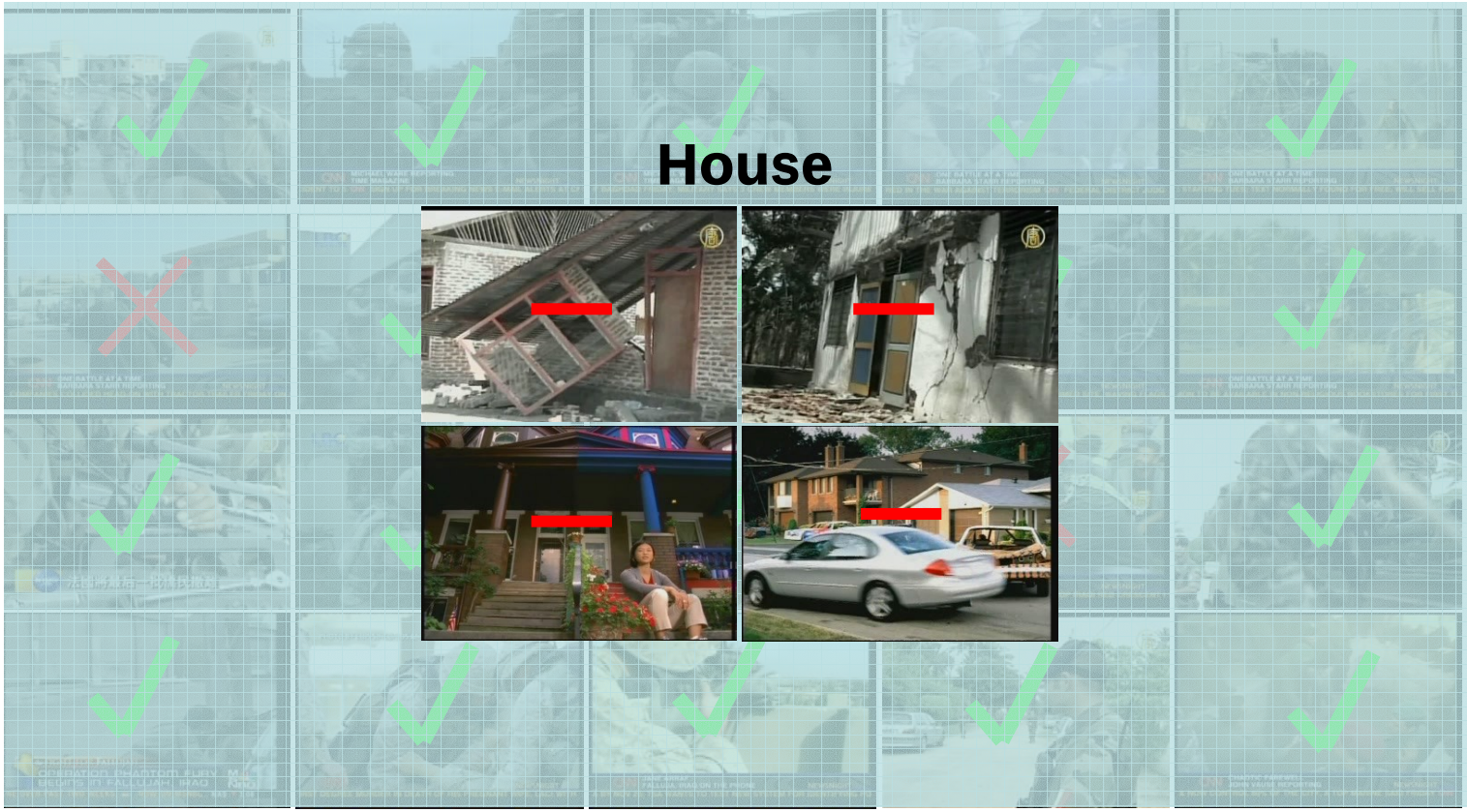


Example

Context-based concept fusion

Example

Context-based concept fusion



DVMM lab
digital video | multimedia laboratary

Example

Context-based concept fusion

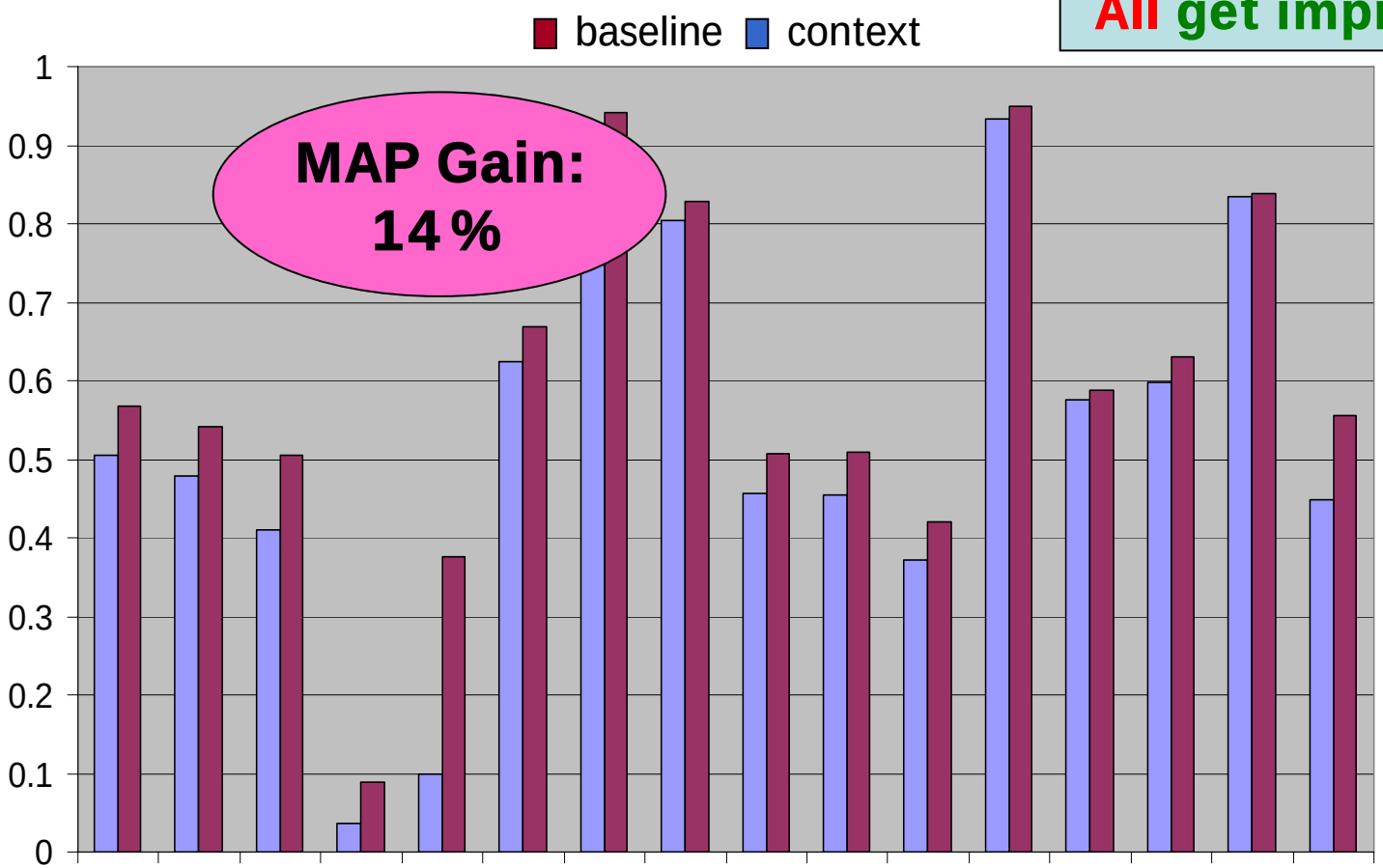
Positive frames are moved forward with the help of Fighter_Combat

DVMM lab
digital video | multimedia laboratary

Context-Based Fusion + Baseline

TRECVID 2005 development set

All get improved !

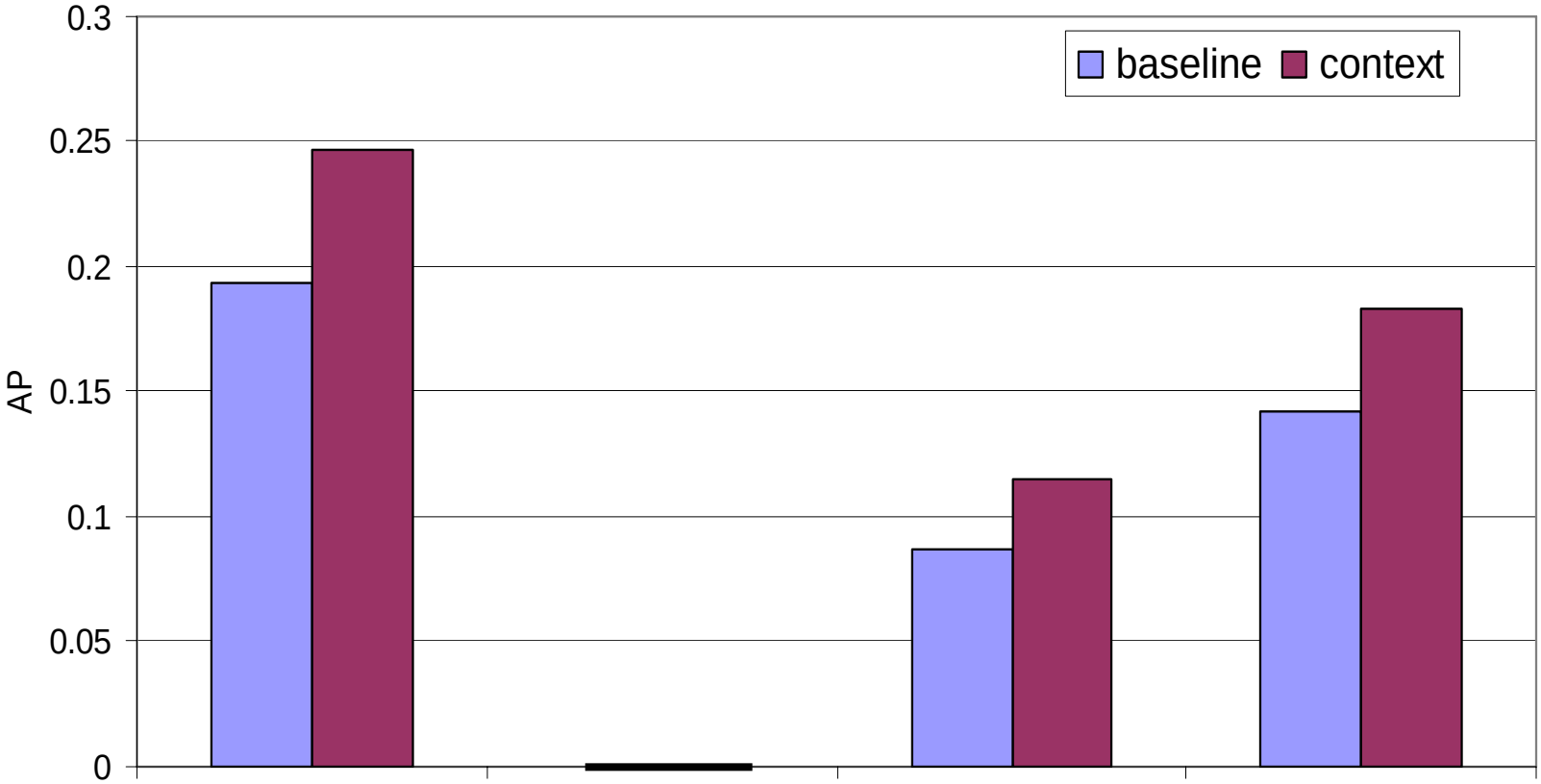


DVMM lab
digital video | multimedia laboratary

Context-Based Fusion + Baseline

TRECVID 2006 evaluation

4 concepts **Similar to results over TRECVID 2005 set !**



DVMM lab
digital video | multimedia laboratary

Discussion

Quality of context:

The smaller the better

$$\frac{\sum_{C_j, j \neq i} I(C_j; C_i) E(C_j)}{\sum_{C_j, j \neq i} I(C_j; C_i)}$$

Concepts with performance improved: 3.23

Concepts with performance degraded: 4.17

Adding context – strong relationship and robust

Outline – New Algorithms

- Baseline
- Context-based concept fusion (CBCF)
- Lexicon-spatial pyramid matching (LSPM)
- Text features
- Event detection

Outline – New Algorithms

- Baseline
- Context-based concept fusion (CBCF)
- Lexicon-spatial pyramid matching (LSPM)
- Text features
- Event detection

DVMM lab
digital video | multimedia laboratary

Individual Methods: (3) LSPM

Local features (SIFT)

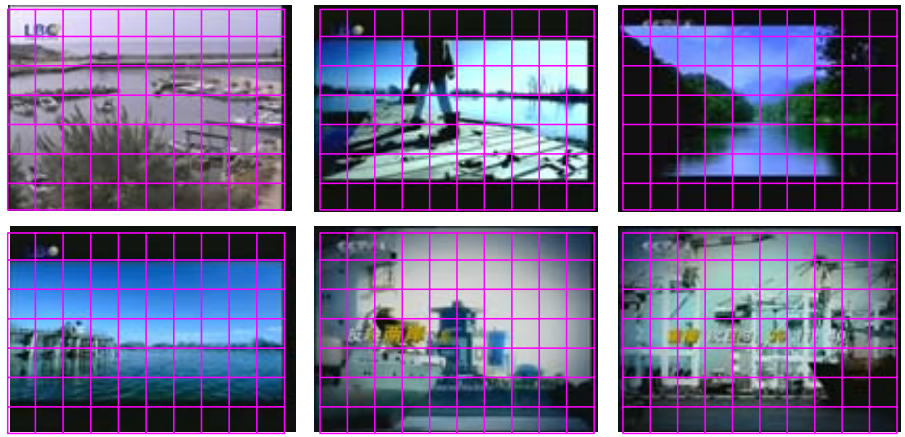
Spatial layout

Spatial Pyramid Matching (SPM) [*Lazebnik et al. CVPR, 2006*]
multi-resolution histogram matching in spatial domain, bags-of-features

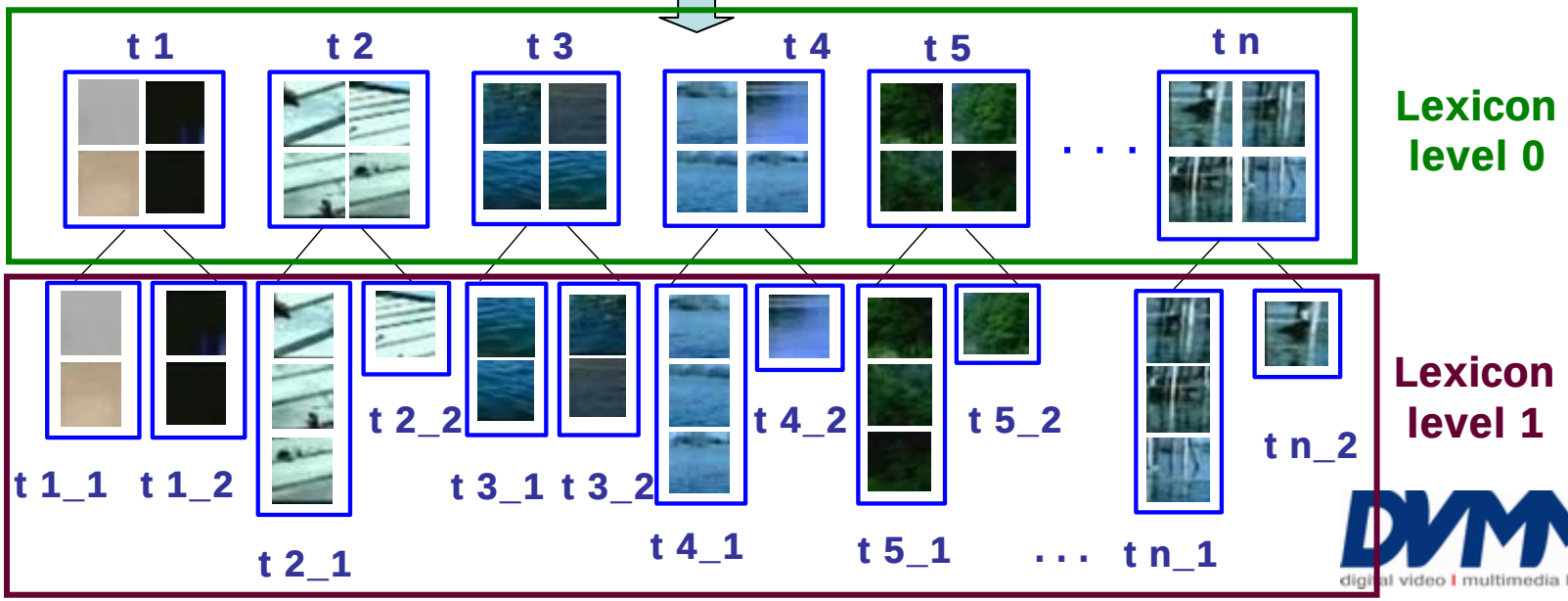
Appropriate size for visual lexicon ?
Lexicon-Spatial Pyramid Matching (LSPM)
SPM matching guided by multi-resolution lexicons

DVMM lab
digital video | multimedia laboratary

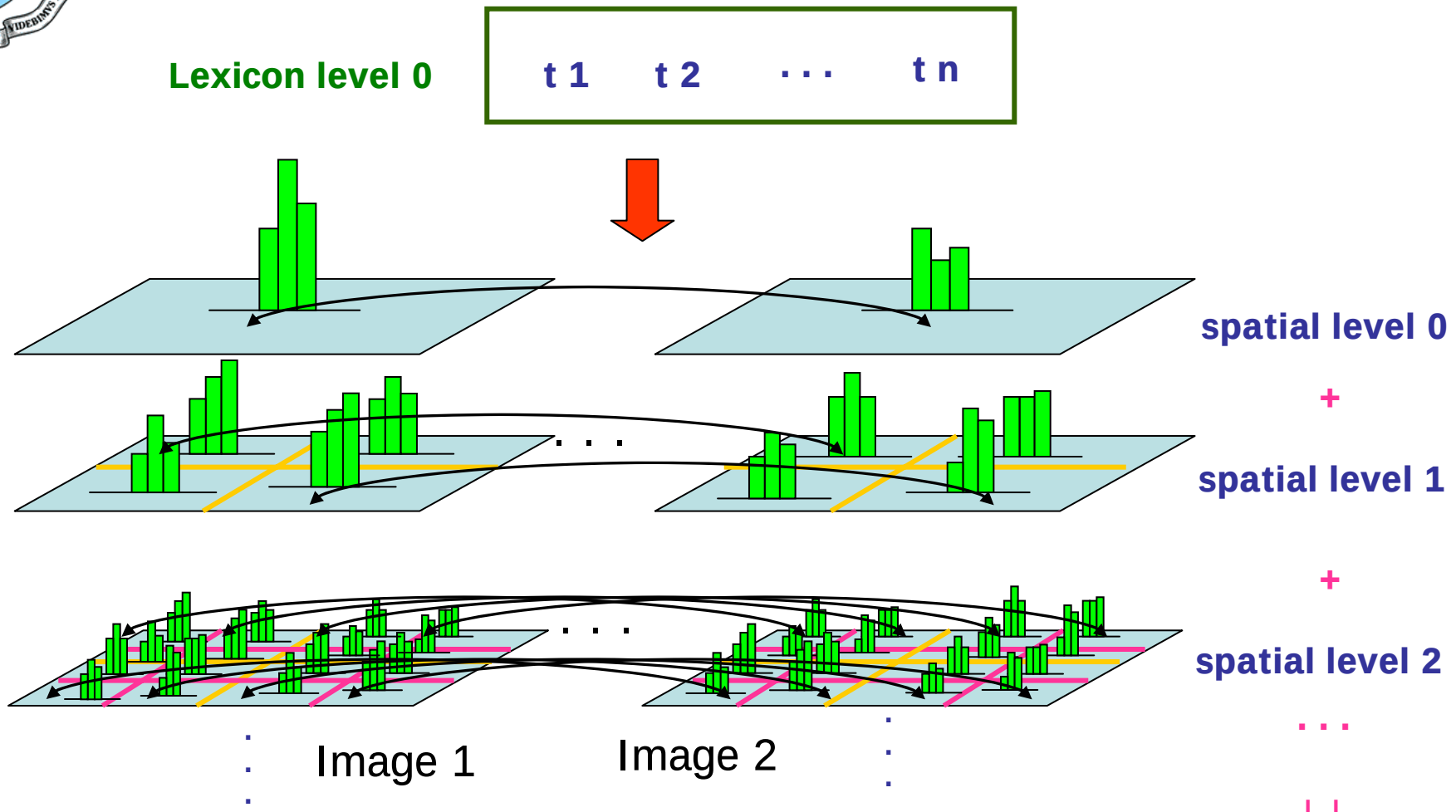
Individual Methods: (3) LSPM



SIFT features

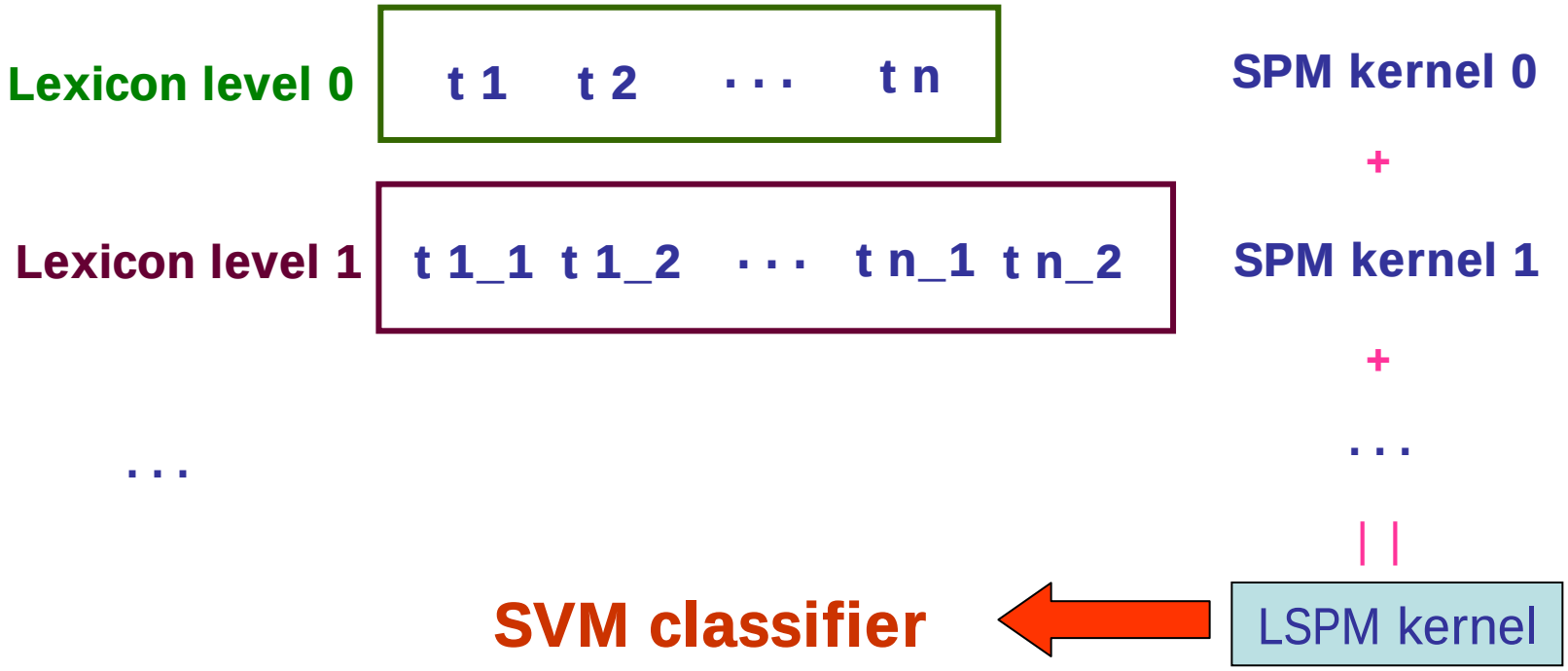


Individual Methods: (3) LSPM



Local features & Spatial layout of local features

Individual Methods: (3) LSPM

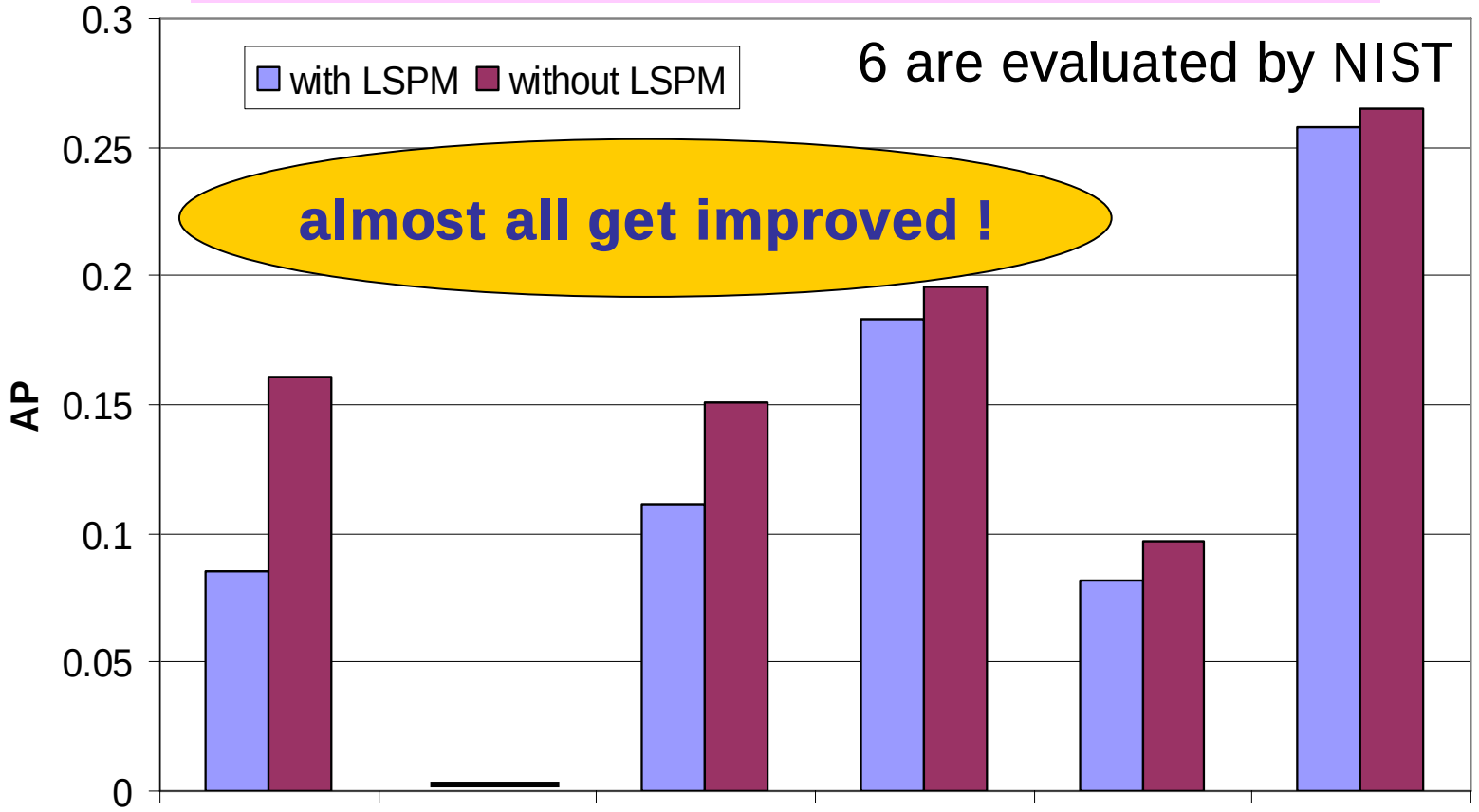


Individual Methods: (3) LSPM

We apply LSPM to 13 concepts:

flag-us, building, maps, waterscape-waterfront, car, charts, urban, road, boat-ship, vegetation, court, government-leader

Complements baseline by considering local features



DVMM lab
digital video | multimedia laboratary

Outline – New Algorithms

- Baseline
- Context-based concept fusion (CBCF)
- Lexicon-spatial pyramid matching (LSPM)
- Text features
- Event detection

Outline – New Algorithms

- Baseline
- Context-based concept fusion (CBCF)
- Lexicon-spatial pyramid matching (LSPM)
- Text features
- Event detection

Individual Methods: (4) Text

Problems:

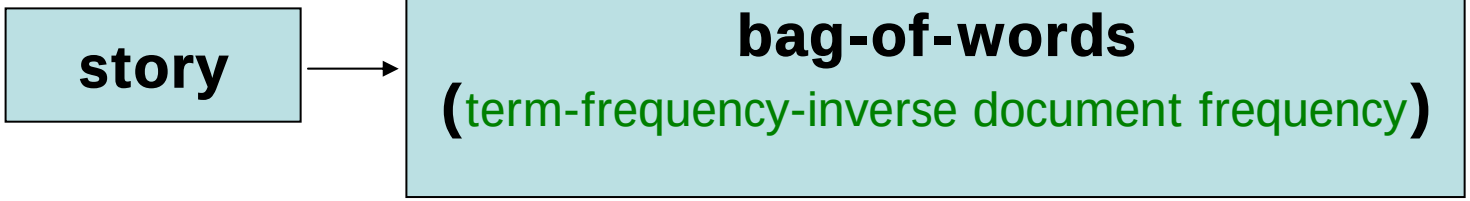
asynchrony between the words being spoken and the visual concepts appearing in the shot

Solution:

incorporate associated text from the entire sto
automatically detected story boundaries

[Hsu et al., ADVENT Technical Report , Columbia Univ., 2005]

by frequency -- top k most frequent words



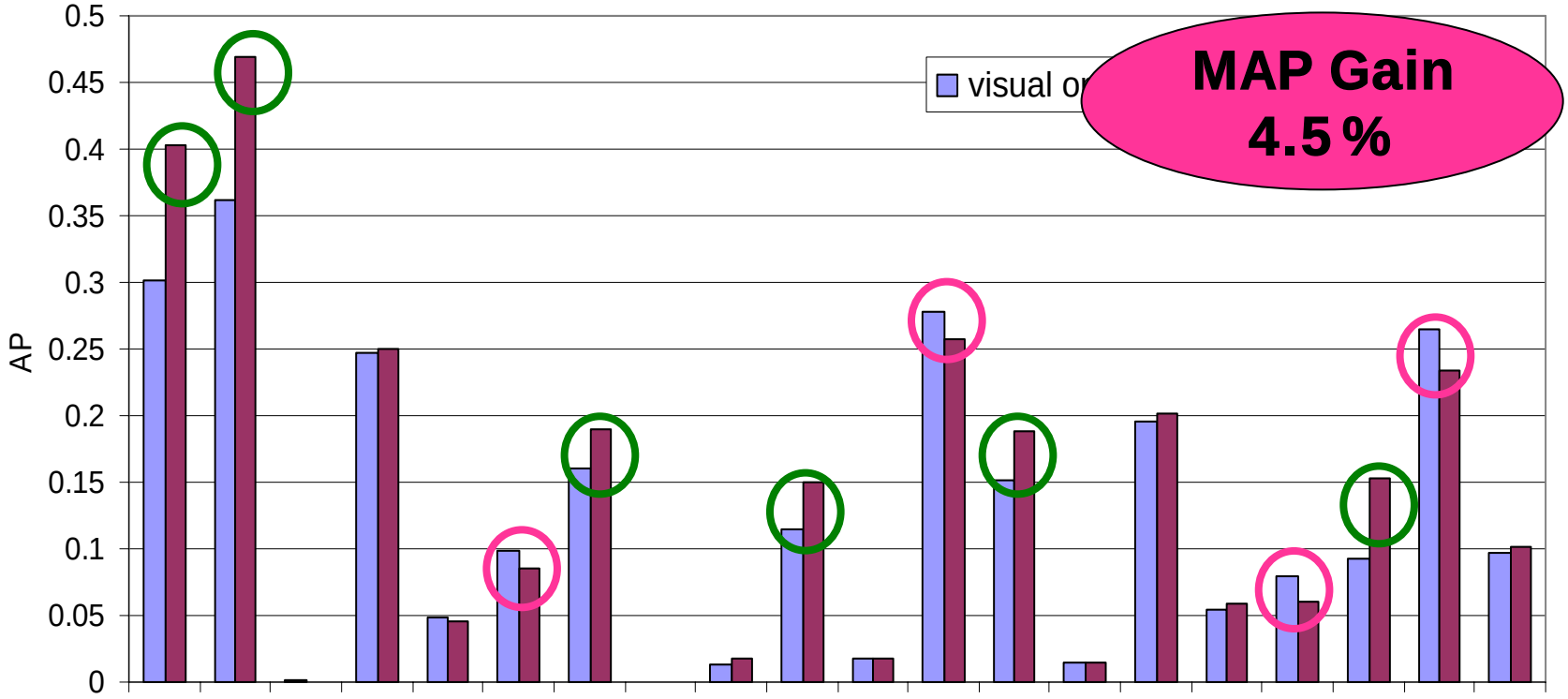
training data: bag-of-words features of stories

ground-truth label: positive – one shot is positive

} SVM

Individual Methods: (4) Text

0.2 text + 0.8 visual



Outline – New Algorithms

- Baseline
- Context-based concept fusion (CBCF)
- Lexicon-spatial pyramid matching (LSPM)
- Text features
- Event detection

DVMM lab
digital video | multimedia laboratary

Outline – New Algorithms

- Baseline
- Context-based concept fusion (CBCF)
- Lexicon-spatial pyramid matching (LSPM)
- Text features
- Event detection

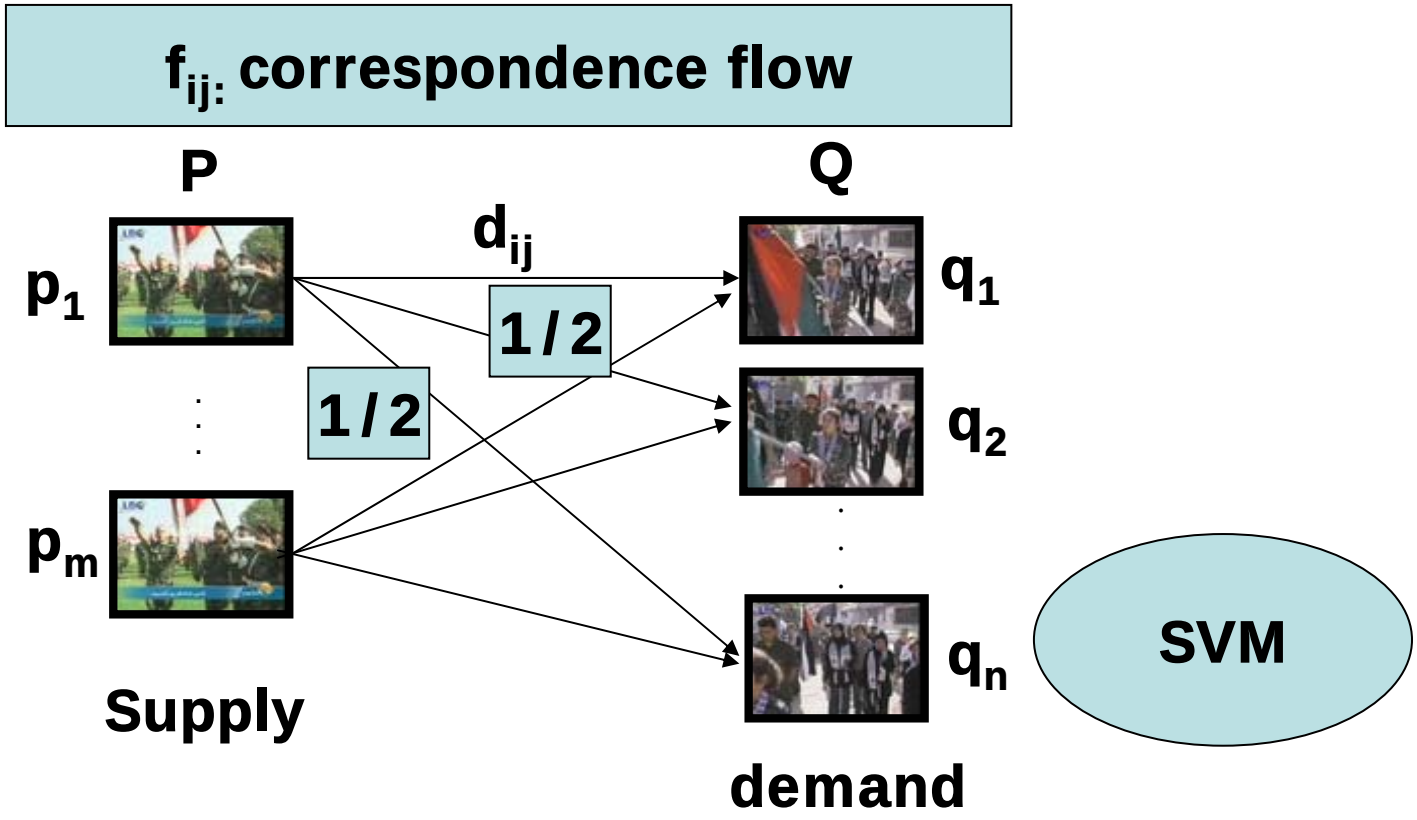
DVMM lab
digital video | multimedia laboratary

Individual Methods: (5) Event

Event detection: Key frame v.s. Multiple frames

Individual Methods: (5) Event

Event detection: Key frame v.s. Multiple frames



Earth Mover's Distance: minimum weighted distance by linear programming

handle temporal shift:

a frame at the beginning of P can map to a frame at the end of Q

Handle scale variations: a frame from P can map to multiple frames in Q

Individual Methods: (5) Event

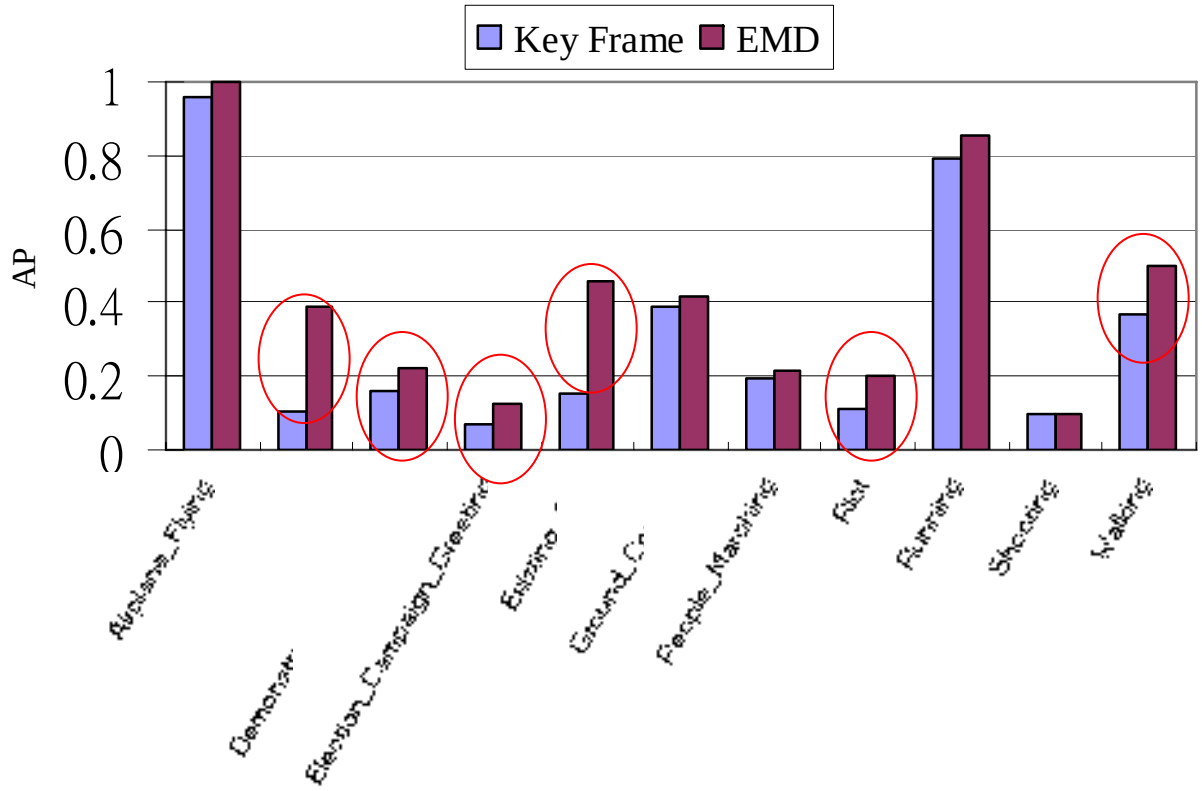
experimental results

Performance over TRECVID 2005 development set

11 events: airplane_flying, people_marching, car_crash,

exiting_car, demonstration_or_protest, election_campaign_greeting,

parade, riot, running, shooting, walking



Conclusion

- TRECVID 2006 offers a mature opportunity for evaluating concept interaction
  - We have built 374 concept detectors
  - Models and feature will be released soon
- Context-Based Fusion
  - Propose a systematic framework for predicting the effect of context fusion
  - (TRECVID 2005) 14 out of 16 predicted concepts show performance gain
  - (TRECVID 2006) 3 out of 4 predicted concepts show performance gain
  - Promising methodology for scaling up to large-scale systems (374 models)
- Results from Parts-based model (LSPM) are mixed
  - But show consistent improvement when fused with SVM baseline
  - 3 out of 6 concepts improve by more than 10%
- Temporal event modeling
  - We propose a novel matching and detection method based on EMD+SVM
  - Show consistent gains in 2005 data set
  - Results in 2006 are incomplete and lower than expected

- More information at
  - http://www.ee.columbia.edu
- Features and models for baseline detectors for 374 LSCOM concepts coming soon