Tsinghua & ICRC @ TRECVID 2007.HFE

New Dataset, New Challenge

- Varied content

- Varied concept occurrence

Feature	1. Sports	12. Mountain	23. Police_Security	28. Flag-US	36. Explosion_Fire	37. Natural-Disaster	38. Maps	39. Charts
% Posit.	1.25	0.69	1.45	0.06	0.25	0.26	0.64	0.63

One team, One mind

- Team members from Intelligent multimedia Group, State Key Lab on Intelligent Tech. and Sys., National Laboratory for Information Science and Technology (TNList), Tsinghua University

  Dong Wang, Xiaobing Liu, Cailiang Liu, Shengqi Zhu, Duanpeng Wang, Nan Ding, Ying Liu, Jiangping Wang, Xiujun Zhang, Yang Pang, Xiaozheng Tie, Jianmin Li, Fuzong Lin, Bo Zhang

- Team members from Scalable Statistical Computing Group in Application Research Lab, MTL, Intel China Research Center

  Jianguo Li, Weixin Wu, Xiaofeng Tong, Dayong Ding, Yurong Chen, Tao Wang, Yimin Zhang

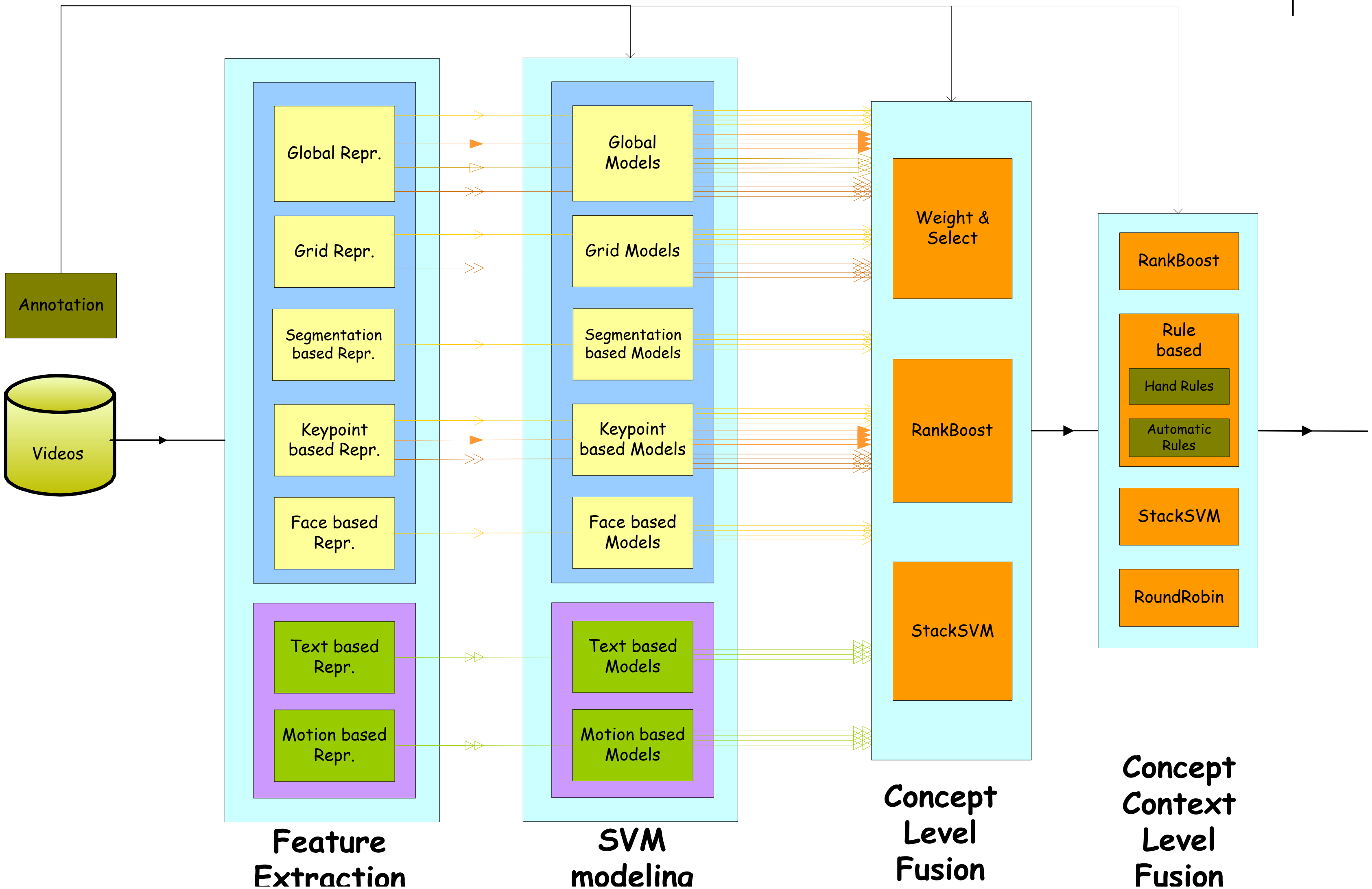
Outline

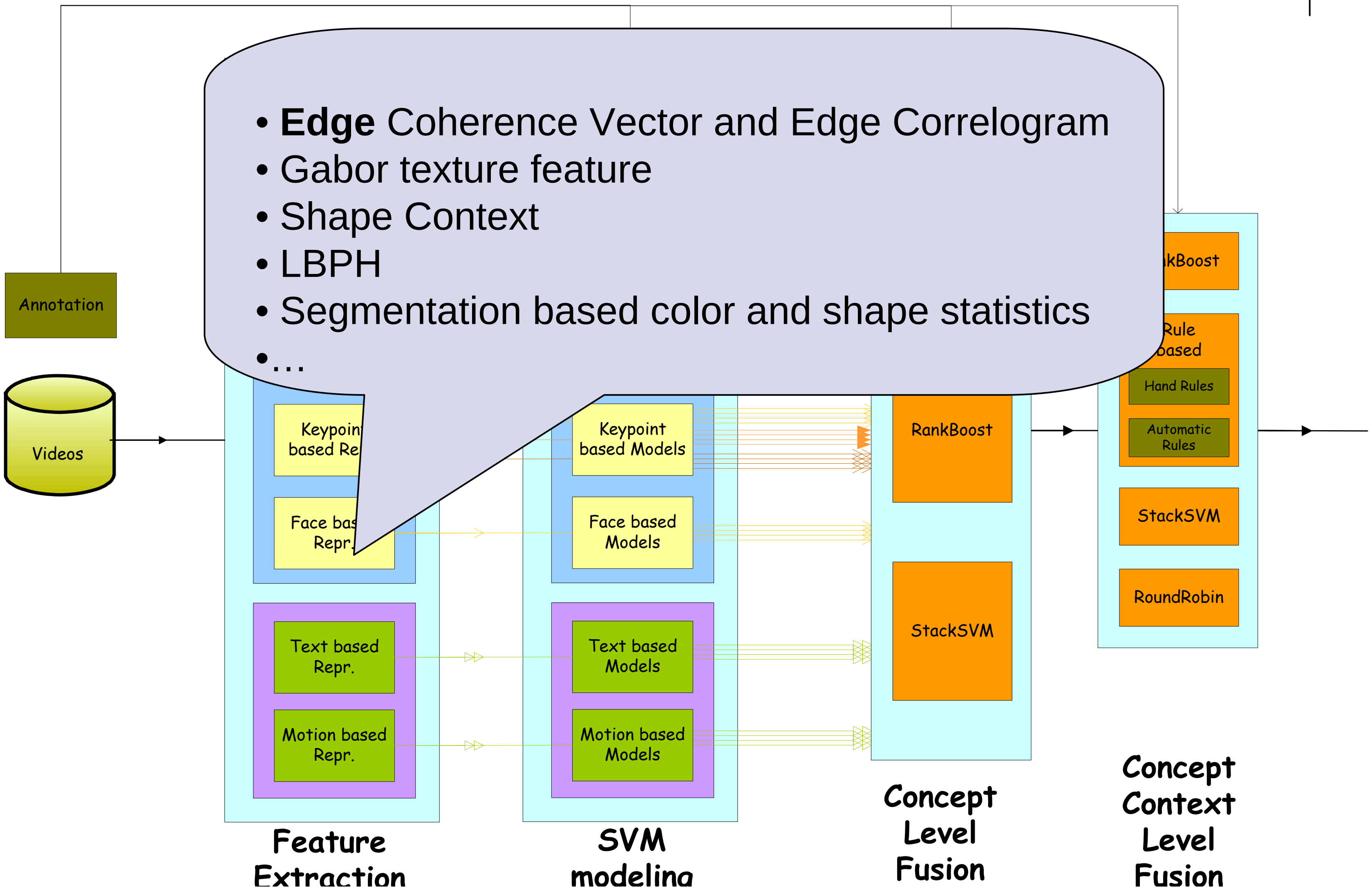
- Overview
- Domain adaptation
- Multi-Label Multi-Feature learning (MLMF)
- New features and other efforts
- Results and discussion

Outline

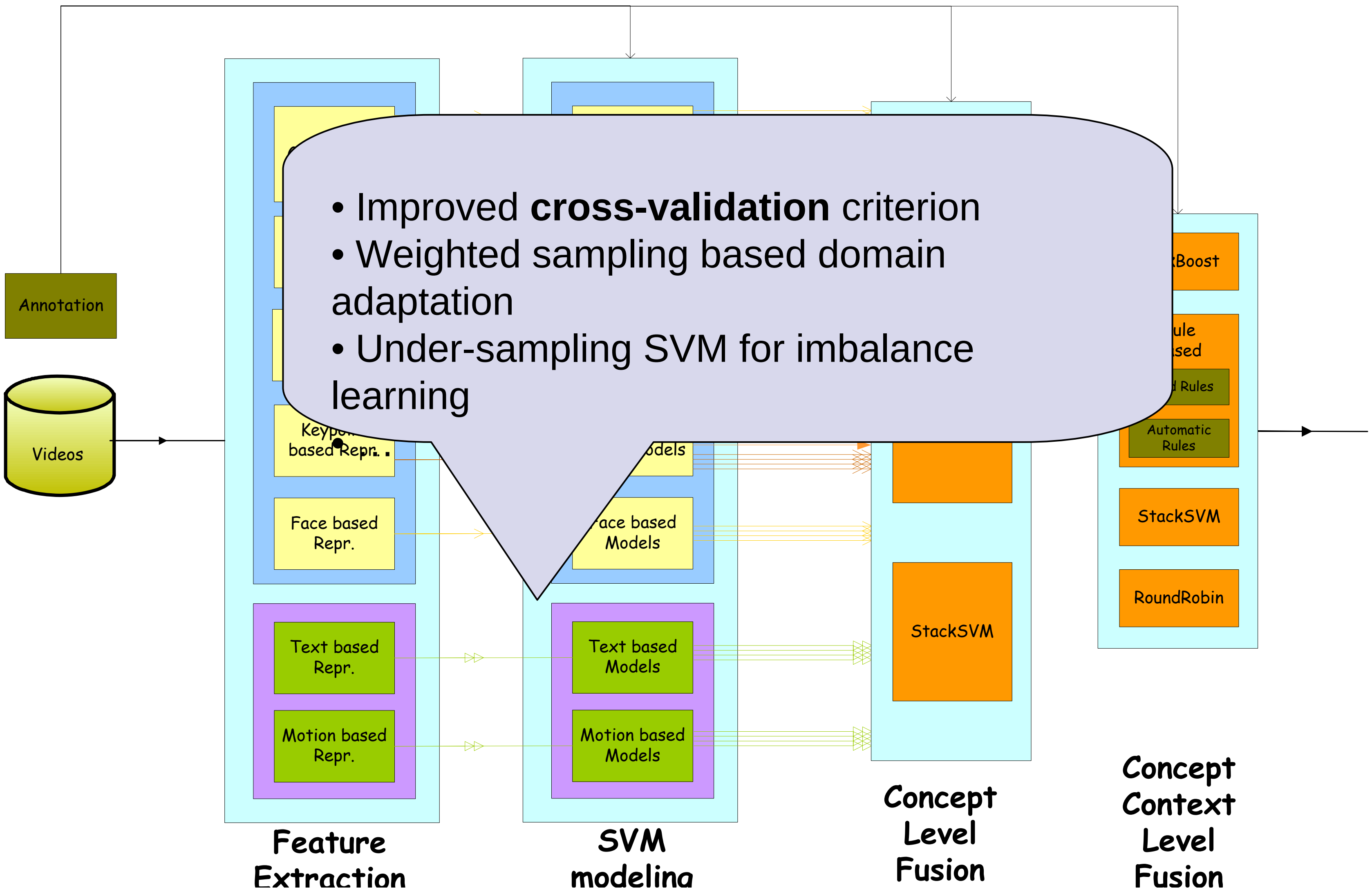
- **Overview**
- Domain adaptation
- Multi-Label Multi-Feature learning (MLMF)
- New features and other efforts
- Results and discussion

Look at the start point

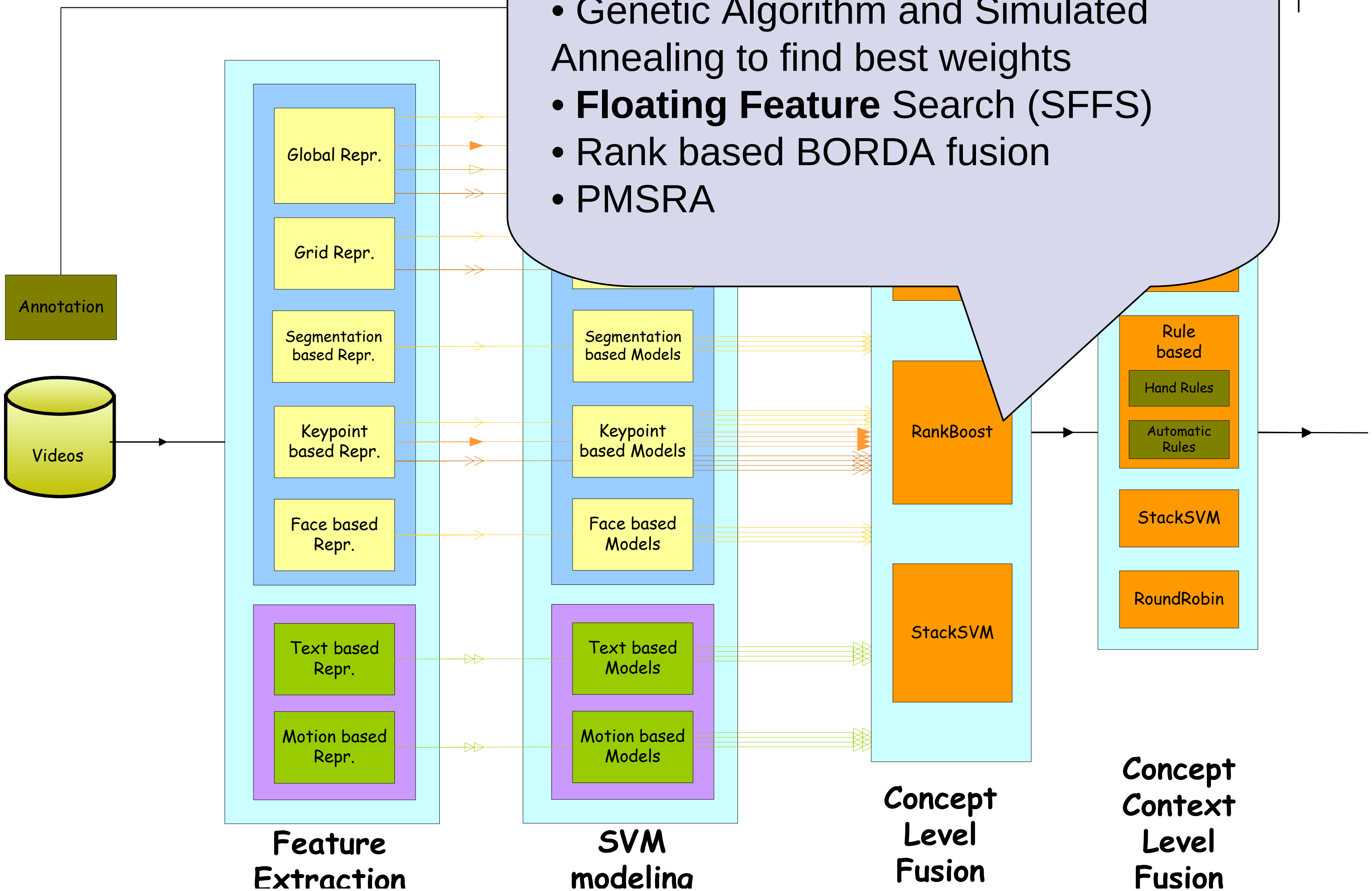



• Edge Coherence Vector and Edge Correlogram
• Gabor texture feature
• Shape Context
• LBPH
• Segmentation based color and shape statistics
•…
Annotation
Videos
Text based Repr.
Motion based Repr.
Keypoint based Models
Face based Models
Text based Models
Motion based Models
RankBoost
StackSVM
Hand Rules
Automatic Rules
StackSVM
RoundRobin
Feature Extraction
SVM modeling
Concept Level Fusion
Concept Context Level Fusion

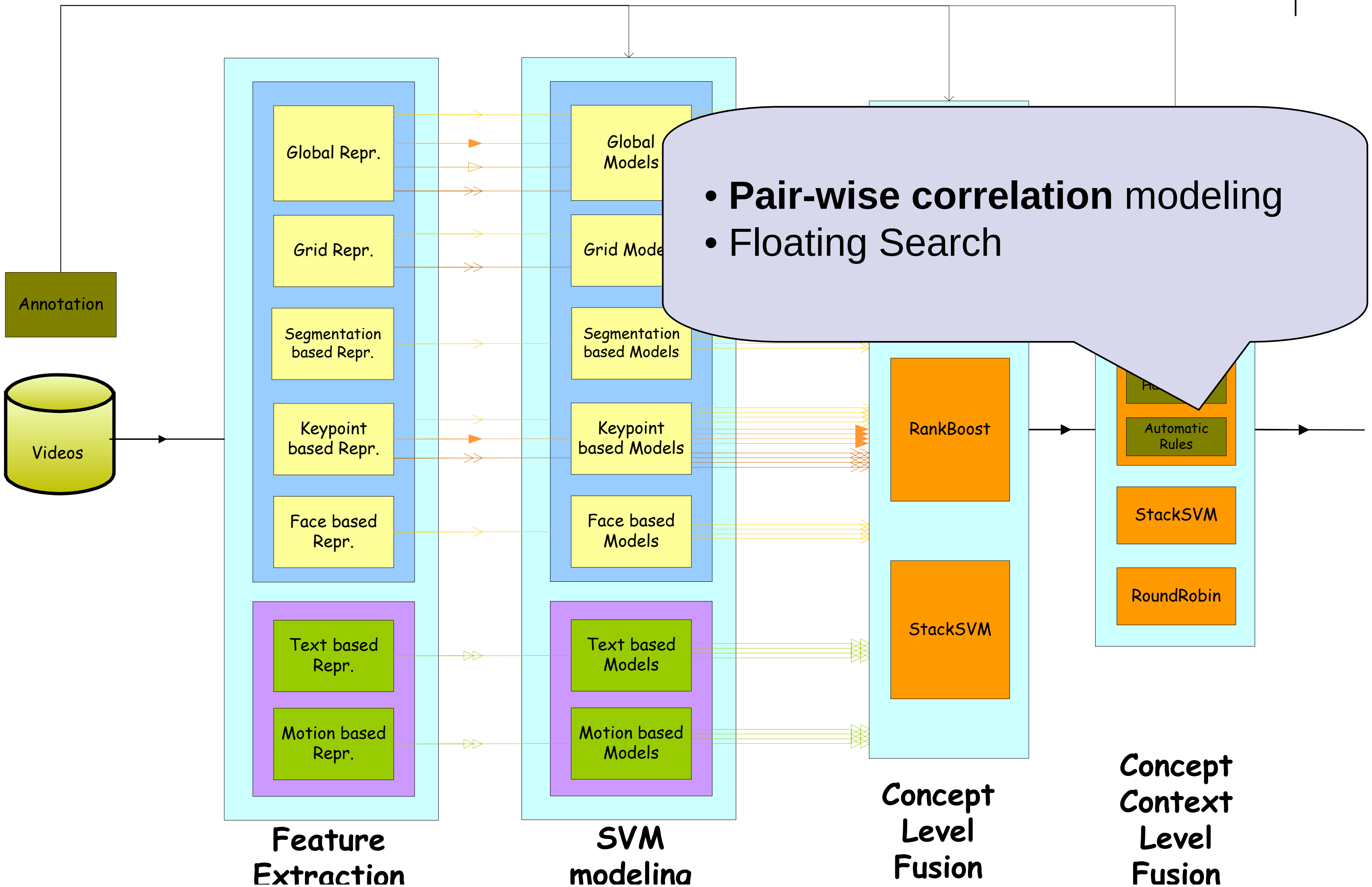
- Improved **cross-validation** criterion
- Weighted sampling based domain adaptation
- Under-sampling SVM for imbalance learning

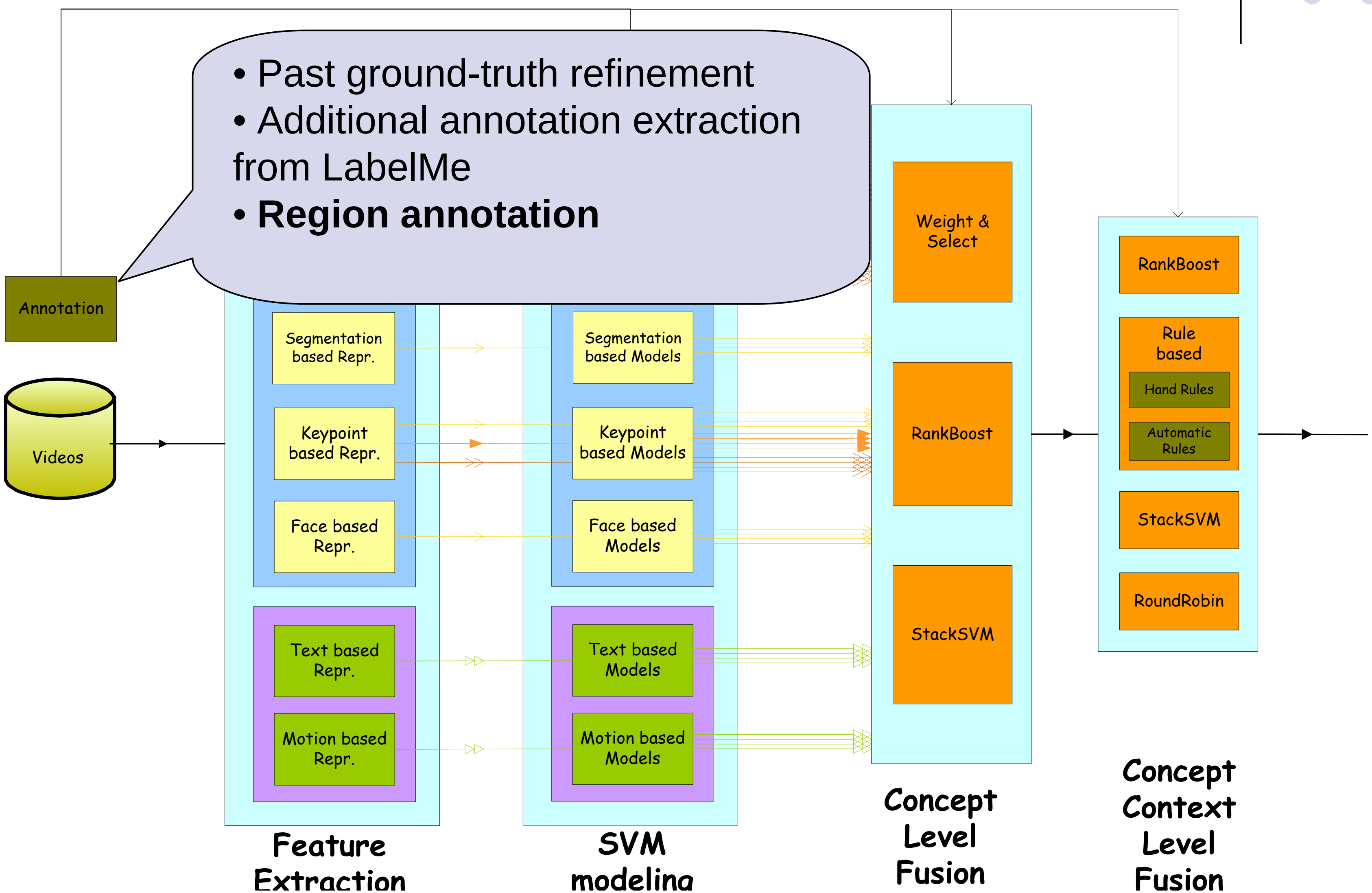


- Boosting at increasing AP
- Genetic Algorithm and Simulated Annealing to find best weights
- **Floating Feature** Search (SFFS)
- Rank based BORDA fusion
- PMSRA



- **Pair-wise correlation** modeling
- Floating Search




• Past ground-truth refinement
• Additional annotation extraction from LabelMe
• **Region annotation**
Annotation
Videos
Segmentation based Repr.
Keypoint based Repr.
Face based Repr.
Text based Repr.
Motion based Repr.
Segmentation based Models
Keypoint based Models
Face based Models
Text based Models
Motion based Models
Weight & Select
RankBoost
StackSVM
RankBoost
Rule based
Hand Rules
Automatic Rules
StackSVM
RoundRobin
Feature Extraction
SVM modeling
Concept Level Fusion
Concept Context Level Fusion

Outline

- Overview
- **Domain adaptation**
- Multi-Label Multi-Feature learning (MLMF)
- New features and other efforts
- Results and discussion

Domain adaptation

- Basic idea: Capture the common characteristics of two related datasets, be able to apply knowledge and skills learned in previous domains to novel domains
- Why: training and testing data often have different distributions
- Advantage:
  - re-use old labeled data to save costs and learn faster

Generalization and adaptation on new data

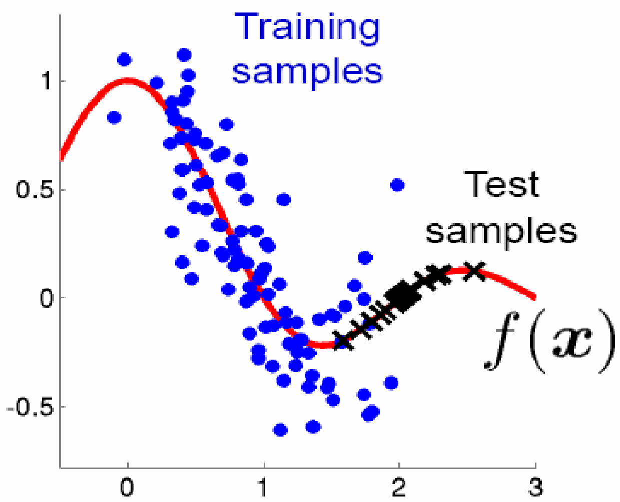
- covariate shift by IWCV (M. Sugiyama in JMLR)

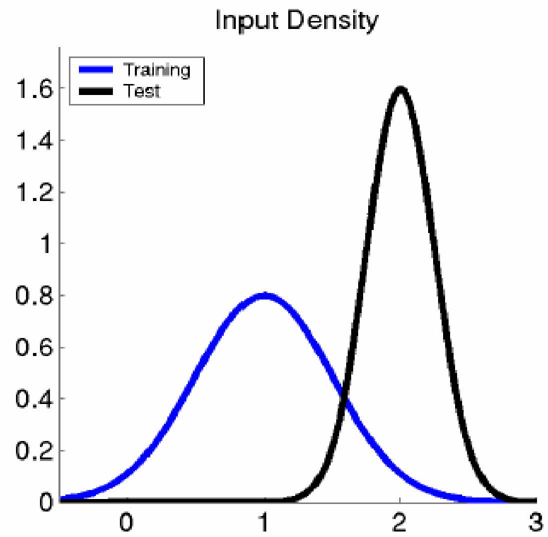
Input distribution changes:

$$P_{train}(\boldsymbol{x}) \neq P_{test}(\boldsymbol{x})$$

Functional relation remains unchanged:

$$P_{train}(y|\boldsymbol{x}) = P_{test}(y|\boldsymbol{x})$$





Importance weighted cross validation

- Under covariate shift, ERM is no longer consistent

$$\lim_{n\to\infty}\left(\mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i=1}^n}\left[\widehat{\boldsymbol{\theta}}_{ERM}\right]\right)\neq\boldsymbol{\theta}^*, \qquad \boldsymbol{\theta}^*=\mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\Theta}\left(\mathbb{E}_{\boldsymbol{t},u}\left[\ell(\boldsymbol{t},u,\widehat{f}(\boldsymbol{t};\boldsymbol{\theta}))\right]\right).$$

- Importance weighted ERM is consistent

$$\widehat{\boldsymbol{\theta}}_{IWERM}=\mathop{\mathrm{argmin}}_{\boldsymbol{\theta}\in\Theta}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{p_{test}(\boldsymbol{x}_i)}{p_{train}(\boldsymbol{x}_i)}\ell(\boldsymbol{x}_i,y_i,\widehat{f}(\boldsymbol{x}_i;\boldsymbol{\theta}))\right], \qquad \lim_{n\to\infty}\left(\mathbb{E}_{\{\boldsymbol{x}_i,y_i\}_{i=1}^n}\left[\widehat{\boldsymbol{\theta}}_{IWERM}\right]\right)=\boldsymbol{\theta}^*.$$

- IWCV (GMM for density estimation)

$$\widehat{R}^{(n)}_{kIWCV}=\frac{1}{k}\sum_{j=1}^{k}\frac{1}{|\mathcal{T}_j|}\sum_{(\boldsymbol{x},y)\in\mathcal{T}_j}\frac{p_{test}(\boldsymbol{x})}{p_{train}(\boldsymbol{x})}\ell(\boldsymbol{x},y,\widehat{f}_{\mathcal{T}_j}(\boldsymbol{x})),$$

$$\widehat{R}^{(n)}_{LOOIWCV}=\frac{1}{n}\sum_{j=1}^{n}\frac{p_{test}(\boldsymbol{x}_j)}{p_{train}(\boldsymbol{x}_j)}\ell(\boldsymbol{x}_j,y_j,\widehat{f}_j(\boldsymbol{x}_j)).$$

Covariate Shift simplified: Combination of tv05d and tv07d

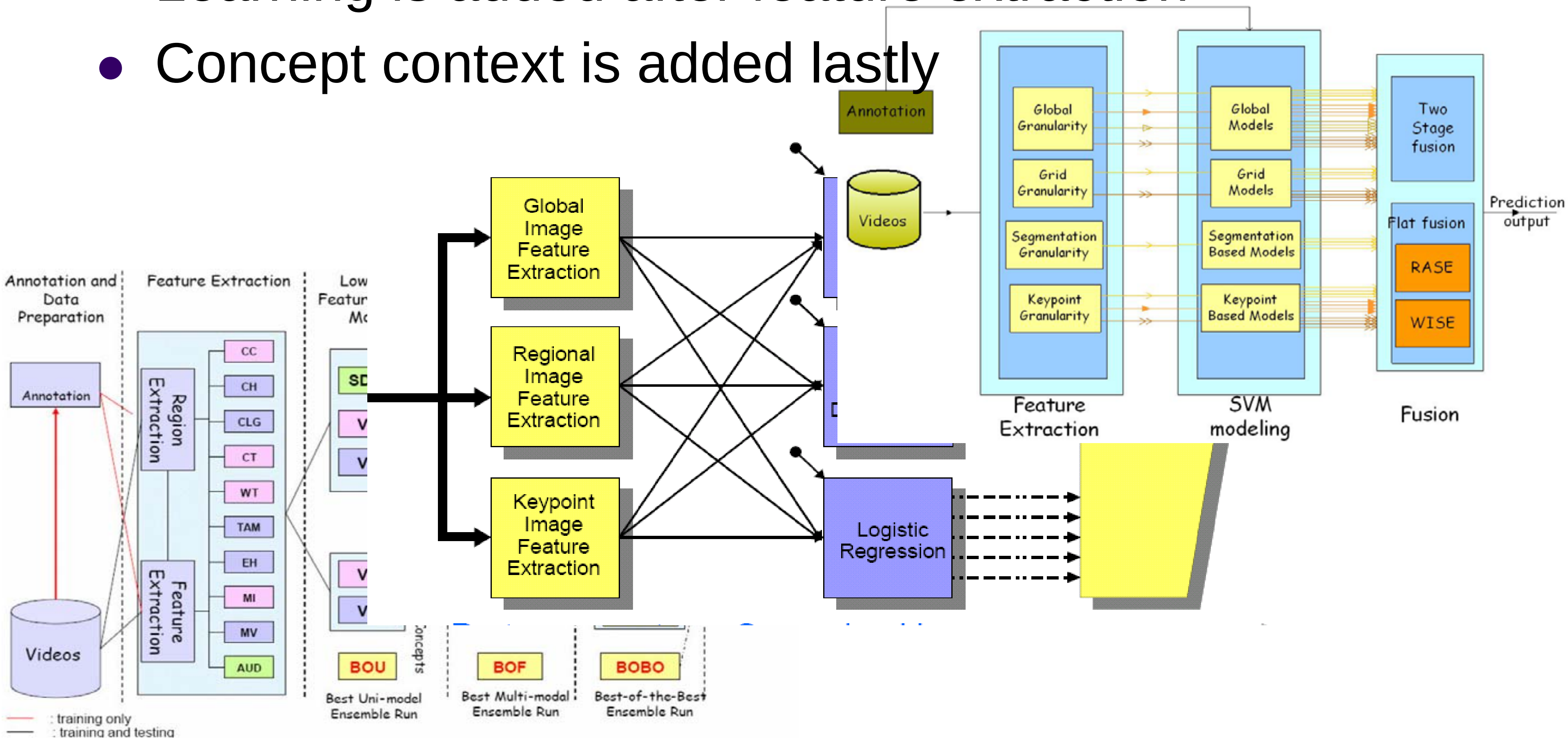
- Devel 05 (05d)/ Devel 07(07d)
  - train classifier C_{07} on 07d
  - predict the positive examples on 05d by C_{07}
  - according to the output of C_{07}, give a weight for 05d positive samples using boosting strategy
  - train C_{05+07} with weighted samples
- Following steps are the same as general framework
- No obvious performance improvement
- Need thorough study and new approach!

Outline

- Overview
- Domain adaptation
- **Multi-Label Multi-Feature learning (MLMF)**
- New features and other efforts
- Results and discussion

The well-accepted pipeline architecture

- Single feature/single concept decomposition
- Learning is added after feature extraction
- Concept context is added lastly

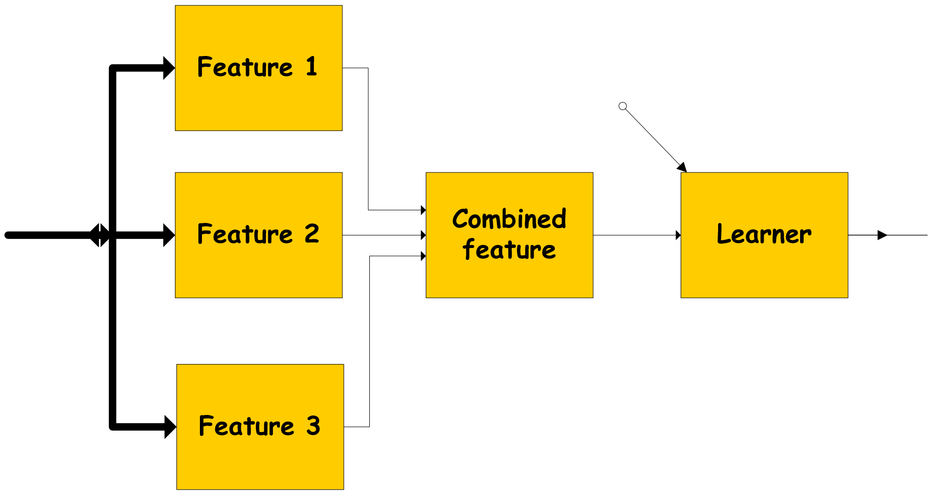


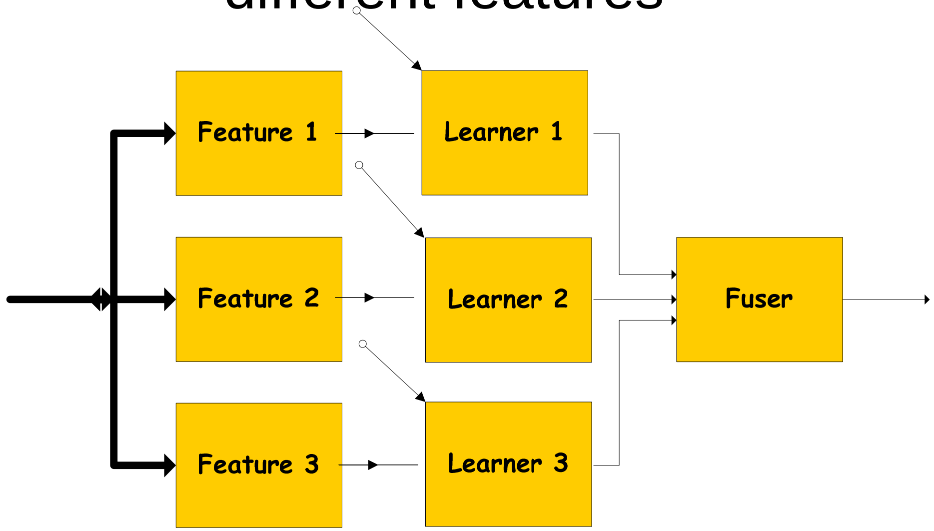
Return to the old debate of Early vs. Late fusion

- Early fusion
- Pro: can count for correlations between different features
- Con: Small example size vs. higher dim

- Late fusion
- Pro: robust
- Con: small example size prevents learning of stable combination weights;

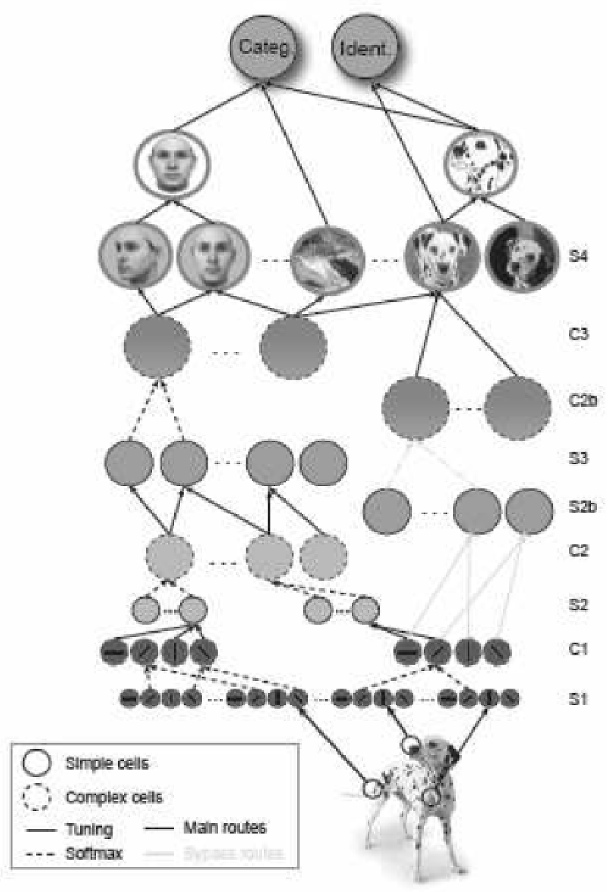
  CANNOT count for correlations between different features





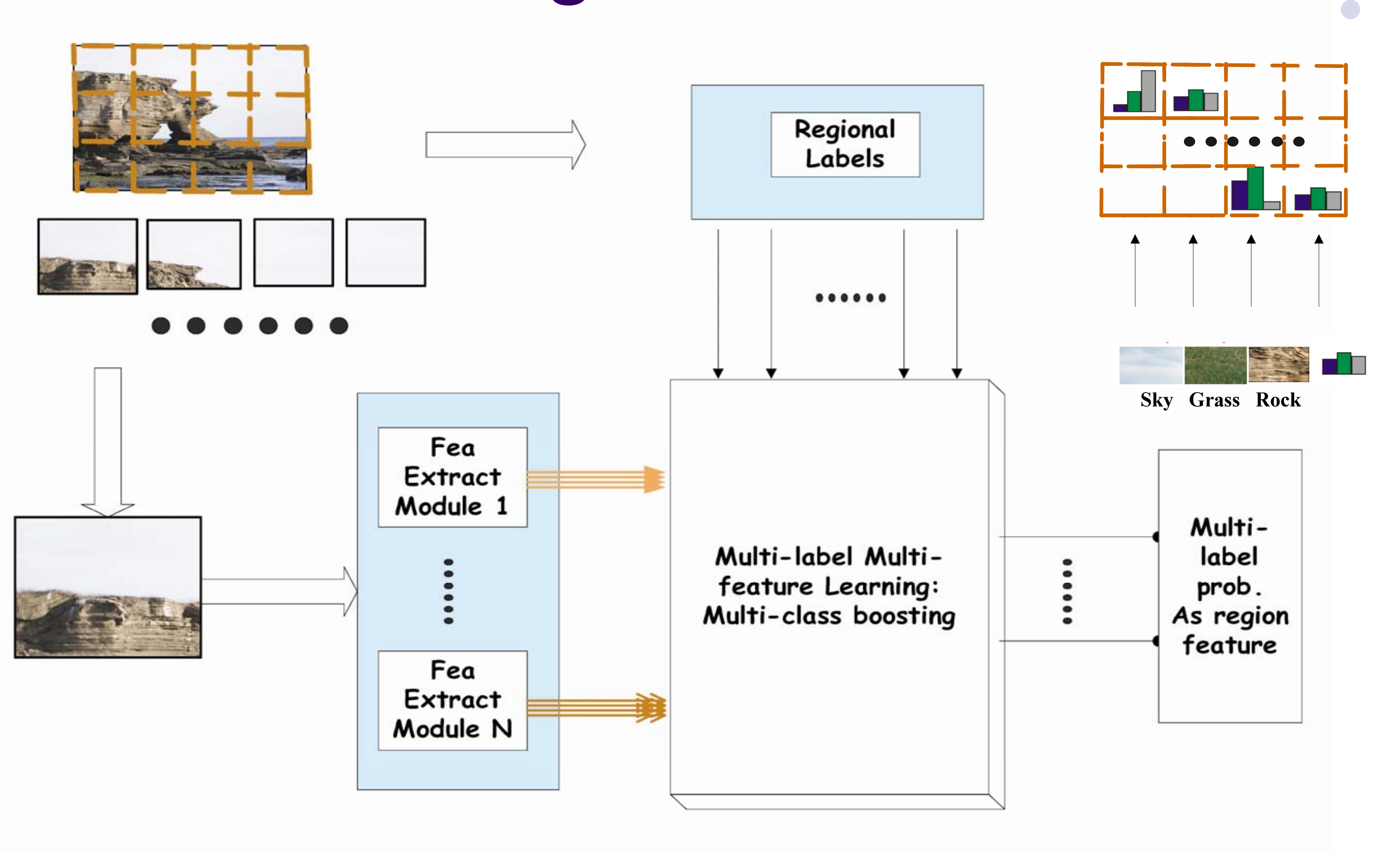
Why human can adapt easily?

- Visual perception of human beings
  - Multi-layer, hierarchical learning
  - From simple cell to complex cell
  - Feed forward processing
- Will human extract lots of specific features for different concepts? No!
- Where fusion takes place in the brain? Distributed!
- Our motivation
  - Hard to map raw feature to complex concepts
  - Try to extract feature hierarchically with learning involved
  - Small scale brings better invariance



After [M. Riesenhuber and T. Poggio]

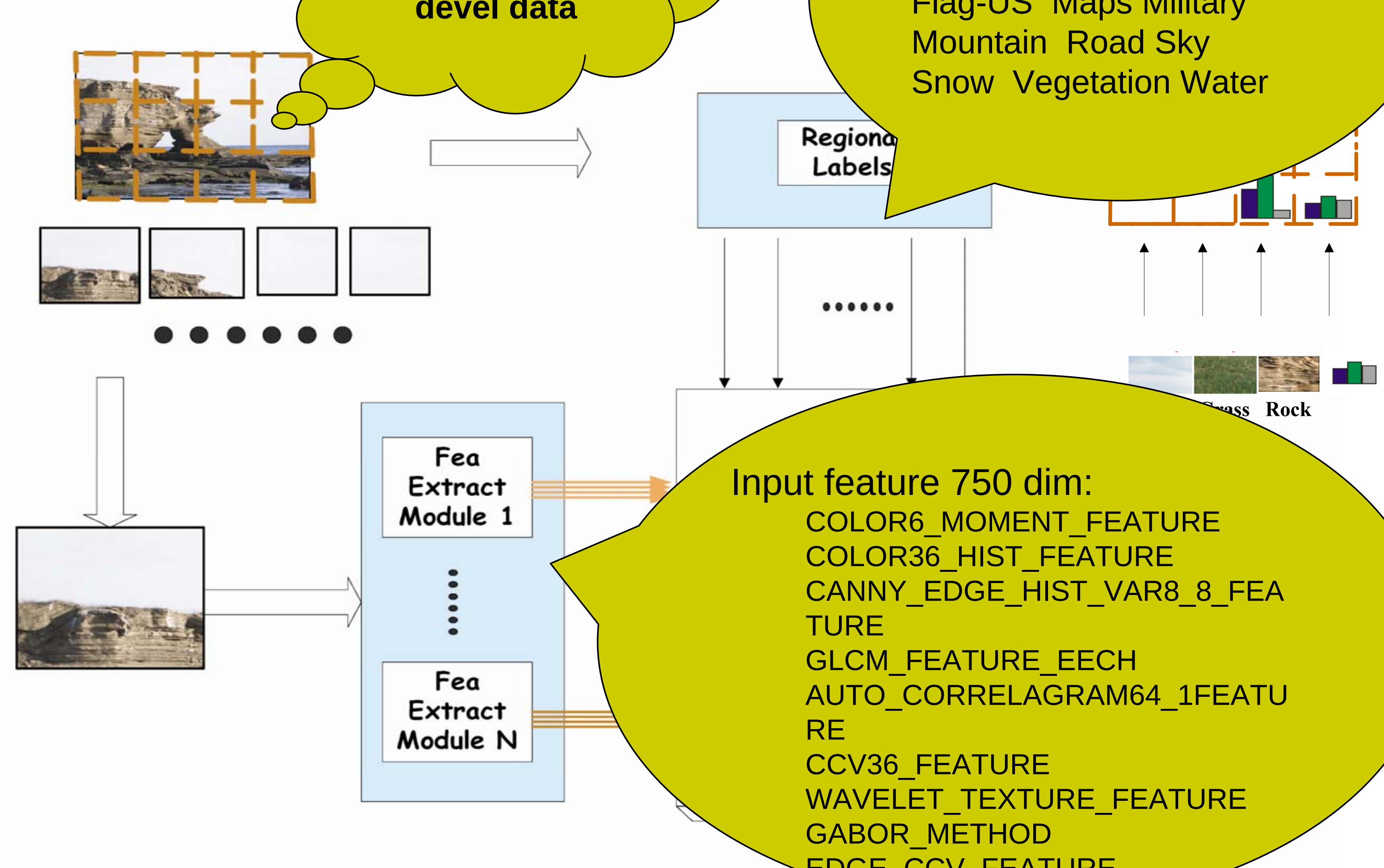
MLMF learning



MLMF

Labelme
TRECVID 2005
TRECVID 2007
devel data

Scene concepts:
Building Charts Crowd
Desert Explosion-Fire
Flag-US Maps Military
Mountain Road Sky
Snow Vegetation Water

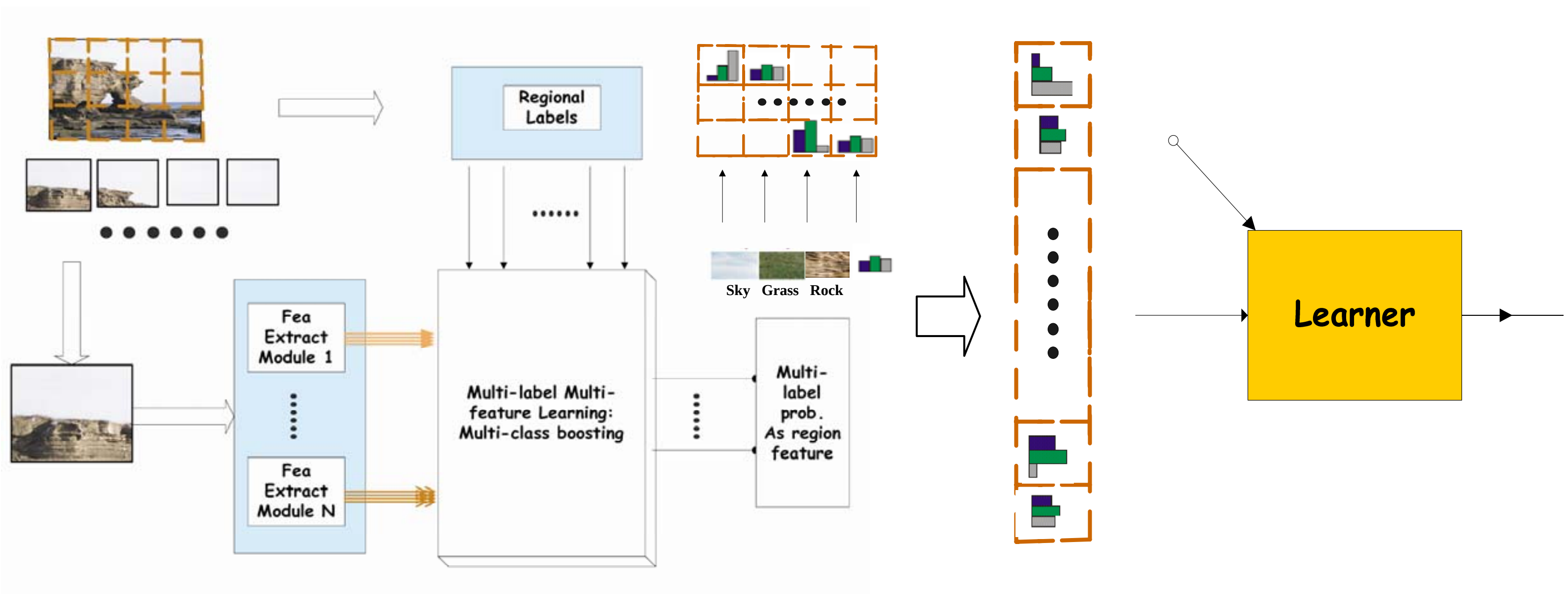


Input feature 750 dim:

COLOR6_MOMENT_FEATURE
COLOR36_HIST_FEATURE
CANNY_EDGE_HIST_VAR8_8_FEATURE
GLCM_FEATURE_EECH
AUTO_CORRELAGRAM64_1FEATURE
CCV36_FEATURE
WAVELET_TEXTURE_FEATURE
GABOR_METHOD
EDGE_CCV_FEATURE
EDGE_CORRELAGRAM_FEATURE

MLMF learning details

- Multi-class boosting for modeling the label correlation and feature correlations.
- Overlapping regional outputs like sliding window
- Then regional scene-concept outputs are concatenated as SVM learner input.



MLMF: Pros and Cons

- improve over the early fusion approach by selecting a few discriminative feature
- improve the late fusion approach by counting the feature correlations properly
- alleviate the semantic gap from raw features to complex concepts. It is also more robust to domain changes.
- The drawback is that it requires regional annotations.

Outline

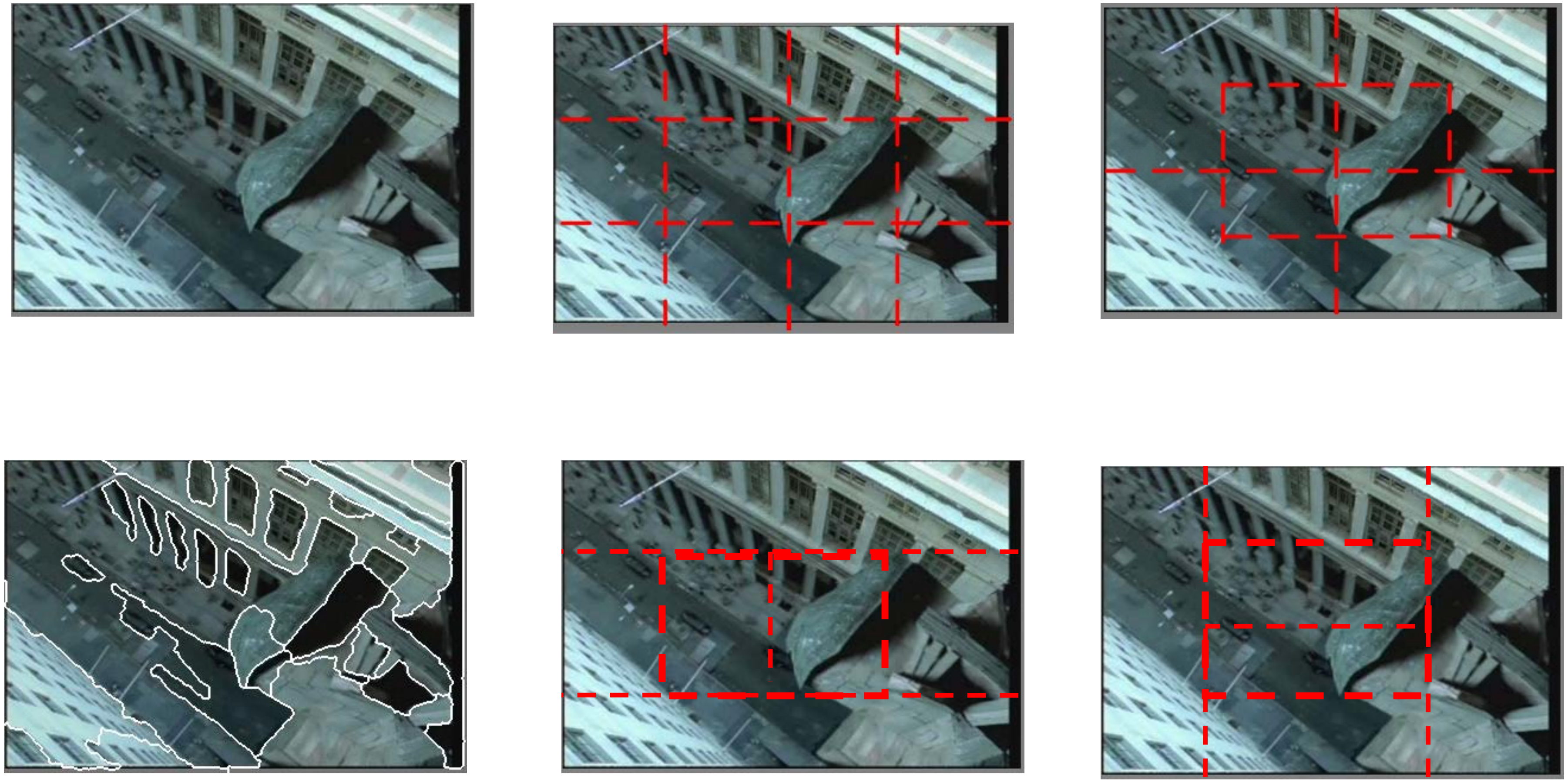
- Overview
- Domain adaptation
- Multi-Label Multi-Feature learning (MLMF)
- **New features and other efforts**
- Results and discussion

Let’s talk about features

- 26 types of various color, edge and texture features,
- Newer features
  - JSeg shape + color statistics
  - Auto correlagram of edges, and coherence vectors for edges
  - Additional implimentation of Gabor, Shape Context, LBPH and MRSAR

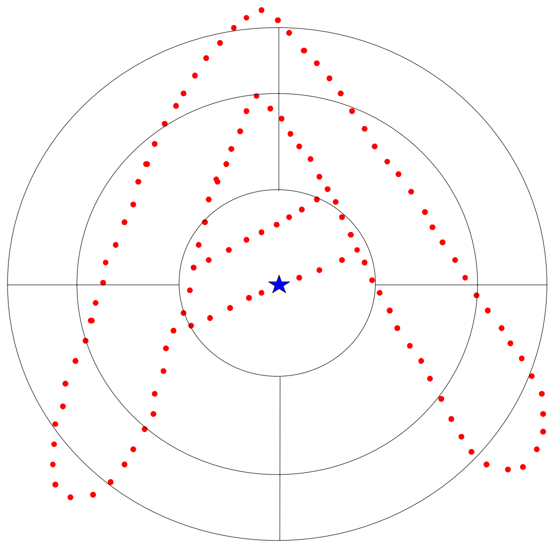
- The effective features: edge and texture
- Keypoint (SIFT) does not work as well as last year.

The partitions used



JSegShape+Color

- JSeg or any segmentation algorithm for image segmentation
- feed the segmentation boundary into the Shape-Context feature extraction
- Quantize in each log-polar region
- Compute color moments in each log-polar region
- Combine the shape-context with the color moments as the final representation.



Modeling Objects

- Man-made Object detector by Boosting + BFM (Boundary fragment model)
- human detection for "crowd", "marching", "person", "walkrun“ with
  - face detection;
  - boosted histograms of oriented gradients,
  - color-texture segmentation
  - probabilistic SVM score.
  - This approach works well for the person concept but bad for crowd and marching concepts due to small human size, occlusion, and noise background disturbance etc.

Person role categorization

- Based on face bounding box
- Boundary fragment model
  - extract up-shoulder bounding box
- Extract feature in up-shoulder region

Parallel computing

- HFE is highly compute intensive
- Computing optimization
  - Parallelize most low-level feature extraction
  - Resampling or undersampling to decompose the large-scale SVM training and testing task into many small jobs, and adopt a cluster/p2p platform to parallel execute those small jobs
  - Use highly (parallel) optimized Intel's library especially OpenCV, and also MKL…

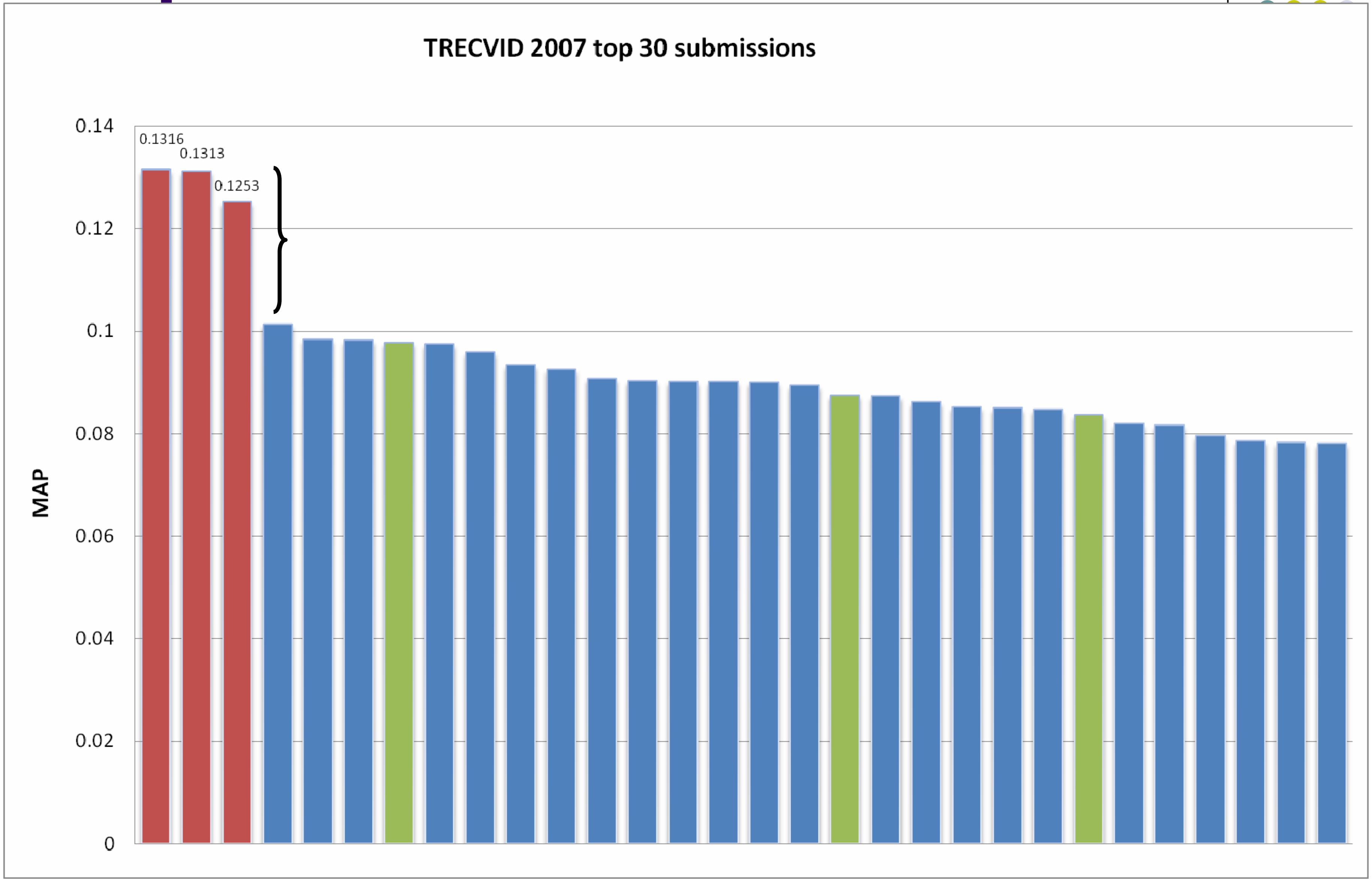
Outline

- Overview
- Domain adaptation
- Multi-Label Multi-Feature learning (MLMF)
- New features and other efforts
- **Results and discussion**

Results

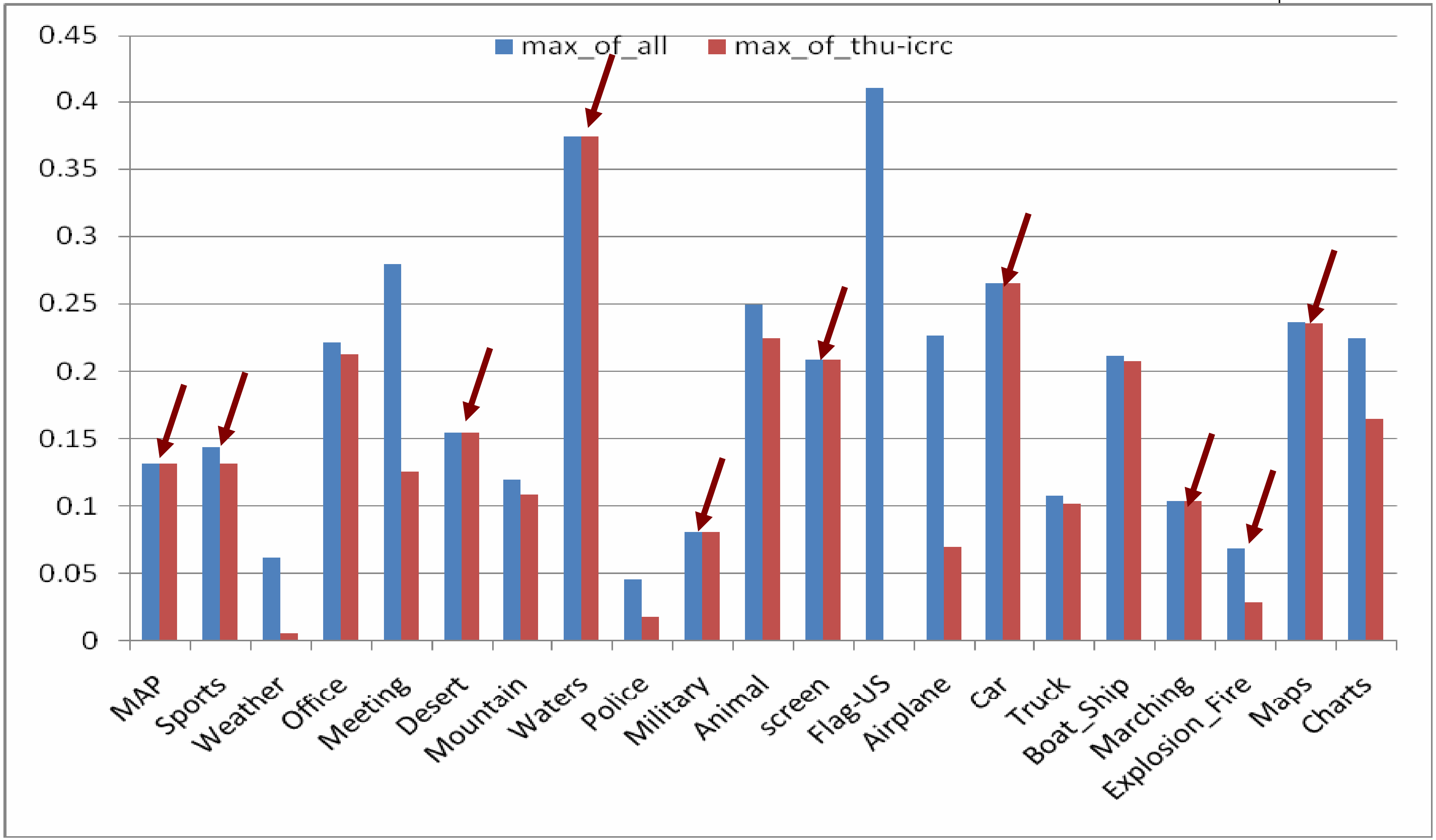
- Benchmarking results
  - Per run, per concept and per feature details
- Further experiments
  - Dataset adaptation and MLMF learning
  - The impact of keyframe sampling rate

Top 30 runs



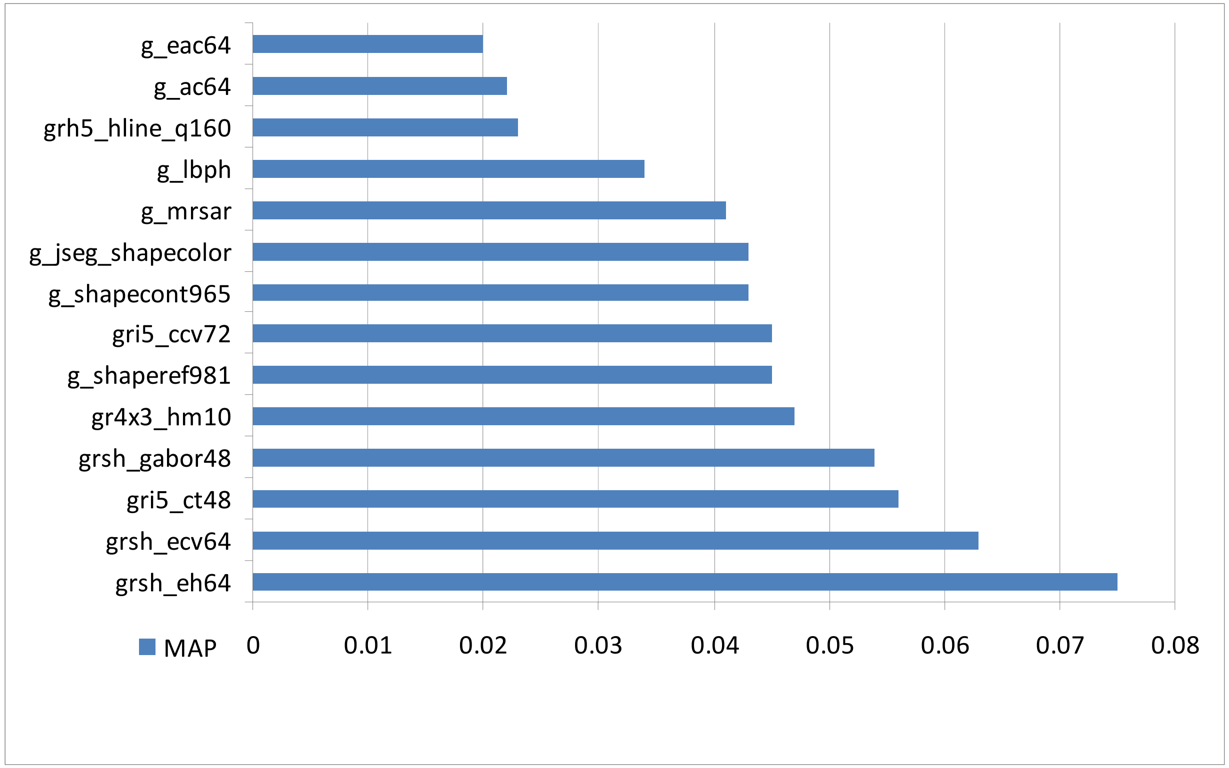
TRECVID 2007 top 30 submissions
MAP
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0.1316
0.1313
0.1253

Per concept results



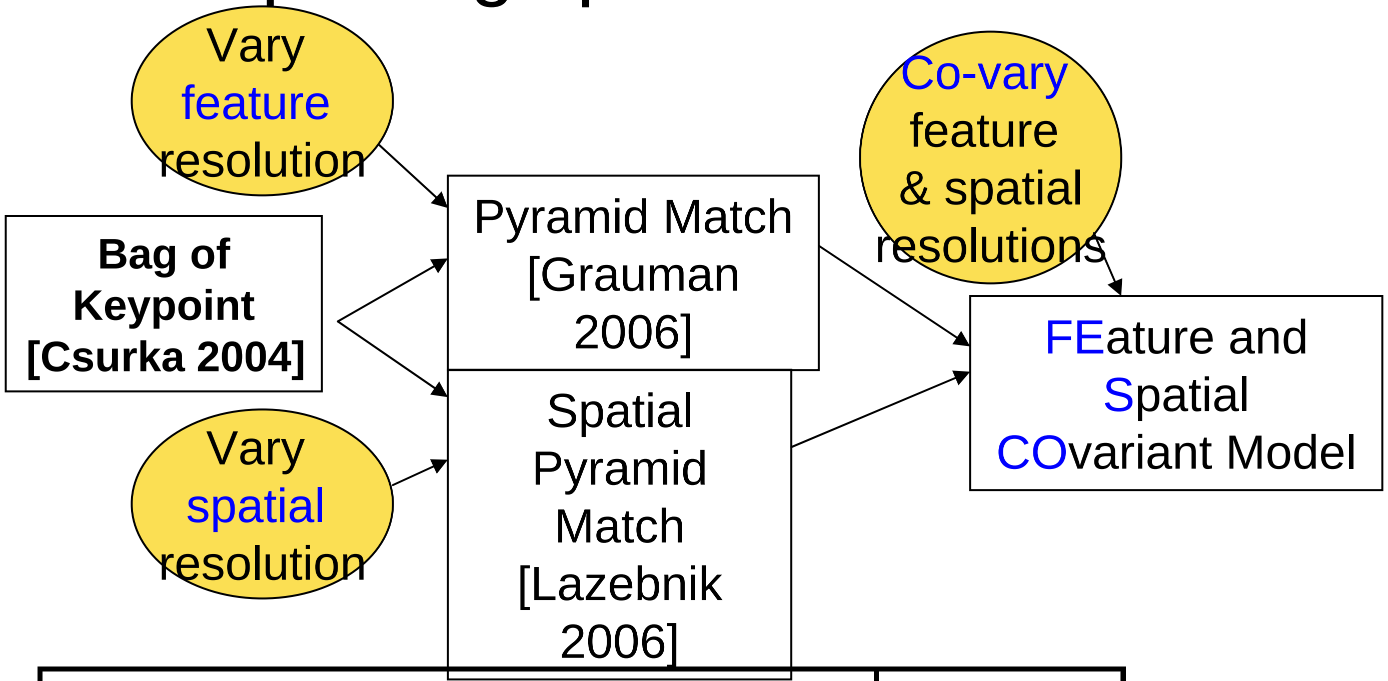
Per-feature analysis

- Edge based features are robust, followed by textures



Per-feature analysis-FESCO

- FESCO: exploiting spatial information



Feature Name	MAP
Combined fesco	0.053
g_hsurf_kmlocal_q288	0.036
gr2x2_hsurf_kmlocal_q72	0.04
gr4x4_hsurf_kmlocal_q18	0.036

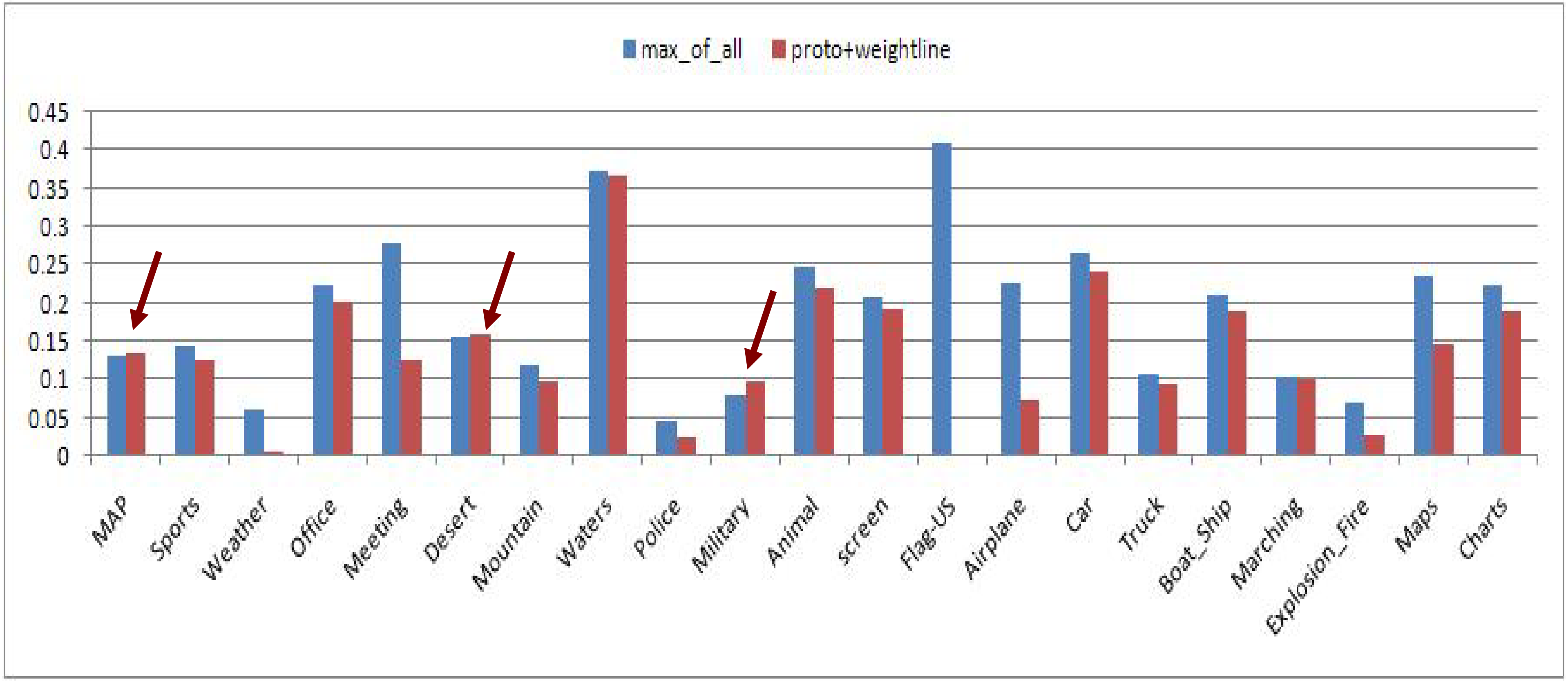
Evaluating dataset adaptation

- MAP: baseline 0.131
- MAP: MLMFline 0.108
- MAP: rerun the last year model 0.065

- Large performance gap!
- MLMF learning generalize better across domains

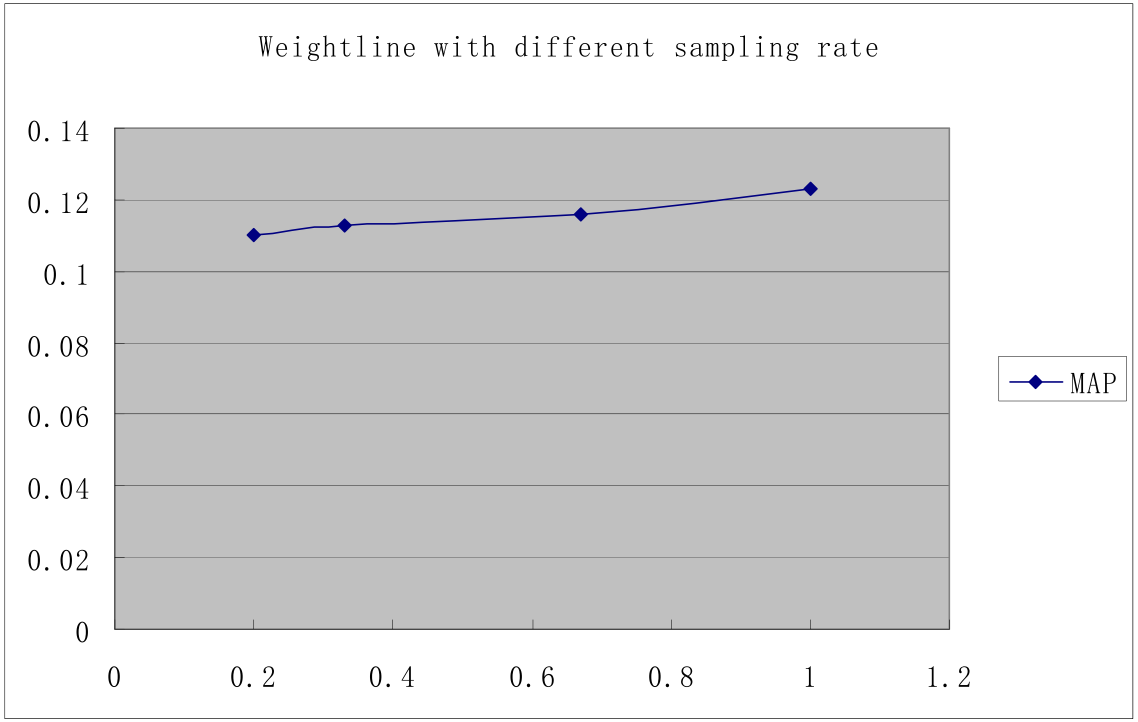
MLMFline+Baseline: MAP (0.1341)

- Type-B system
- MLMFline+Baseline only



Impact of practical issues

- Frame fusion can affect the shot-level AP performance.
- Keyframe sample rate is not so important.



Wrap-up message

- Meaningful features are vital to success
- Spatial information is of additional value
- MLMF is a promising
- Resampling is efficient, USVM is also good
- Simple fusion works pretty well
- As two sides of one coin, fusion and dataset adaptation remains difficult
- Vision based object detection depends on the data

Further work

- Upgrading the MLMF learning framework
- Pushing other new features
- Incorporating temporal information
- Comparing other datasets and image datasets
- Effective domain transfer method

Acknowledgements

- NIST for organizing
- LIP/CAS and the community for annotation
- D. Lowe for SIFT binary
- H. Bay for SURF binary
- C.-J. Lin for LIBSVM
- Computation Platform from NLIST

Thanks! ☺

Any further questions, please contact:

wdong01@mails.tsinghua.edu.cn

jianguo.li@intel.com