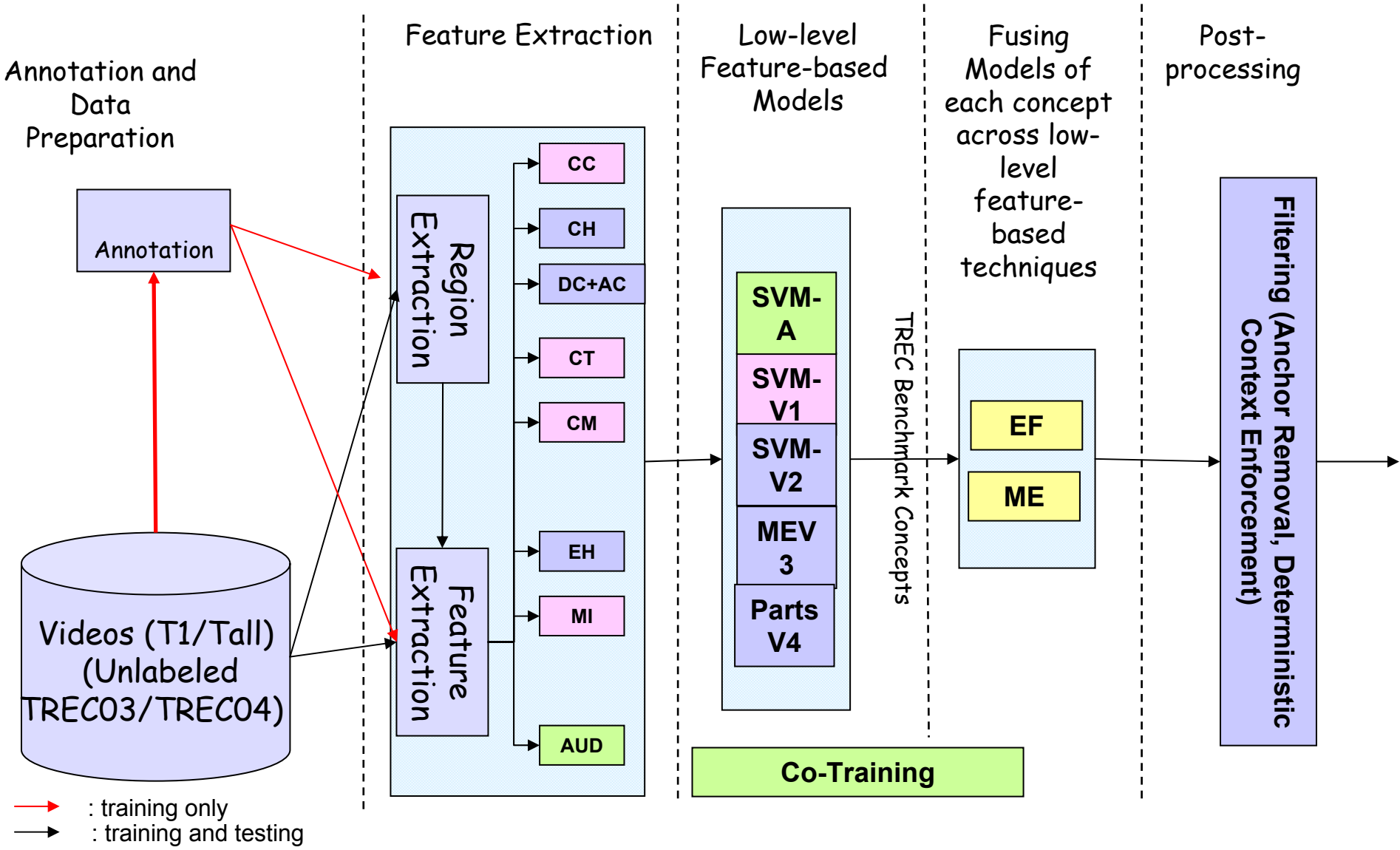
The IBM TRECVID Concept Detection: Some New Directions and Results

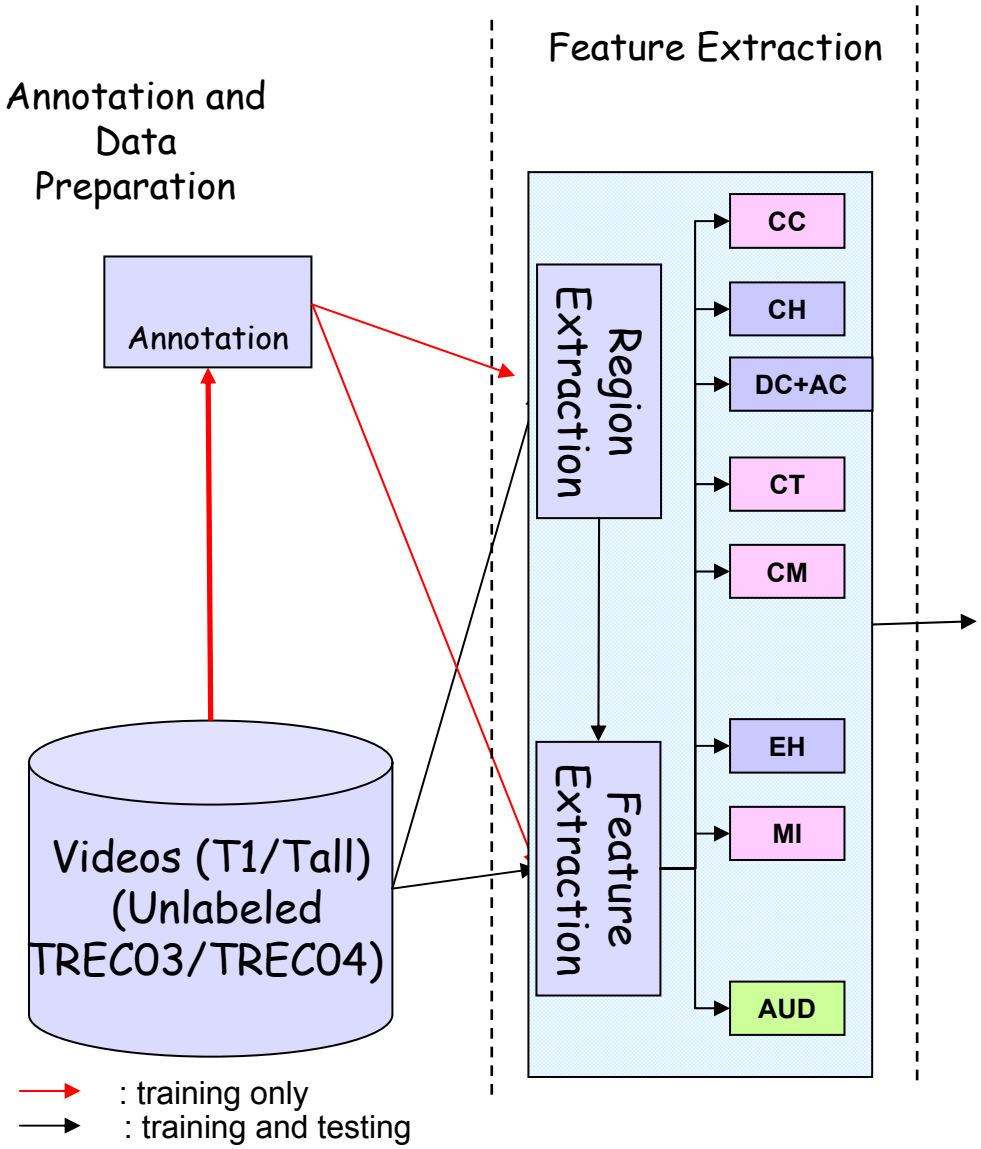
Milind R. Naphade

Contributors:
Milind Naphade, Rong Yan, Janne Argillander, Apostol Natsev, Jelena Tesic, John R. Smith, Arnon Amir, Giridharan Iyengar, Dongqing Zhang, Ching-Yung Lin

The IBM TRECVID 2004 Concept Detection Framework



Feature Extraction



Feature Extraction

Visual

- Color Correlogram (166)
- Color Histogram (166)
- Edge Histogram (64)
- Co-occurrence Texture (96)

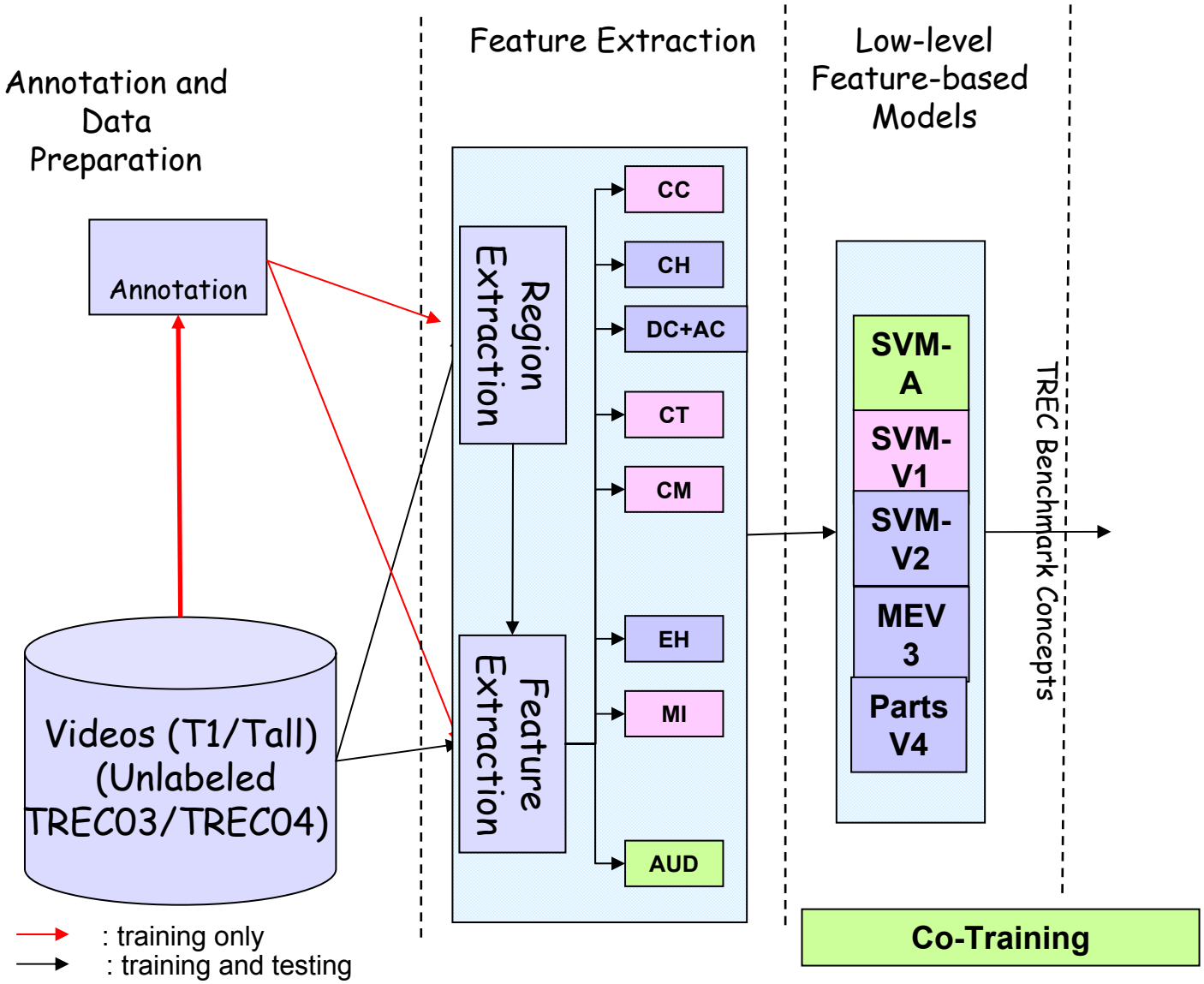
Granularity

- Grid (3x3)
- Layout
- Global
- Segmented Regions-based
- Compressed domain features (DCT DC+AC coefficients

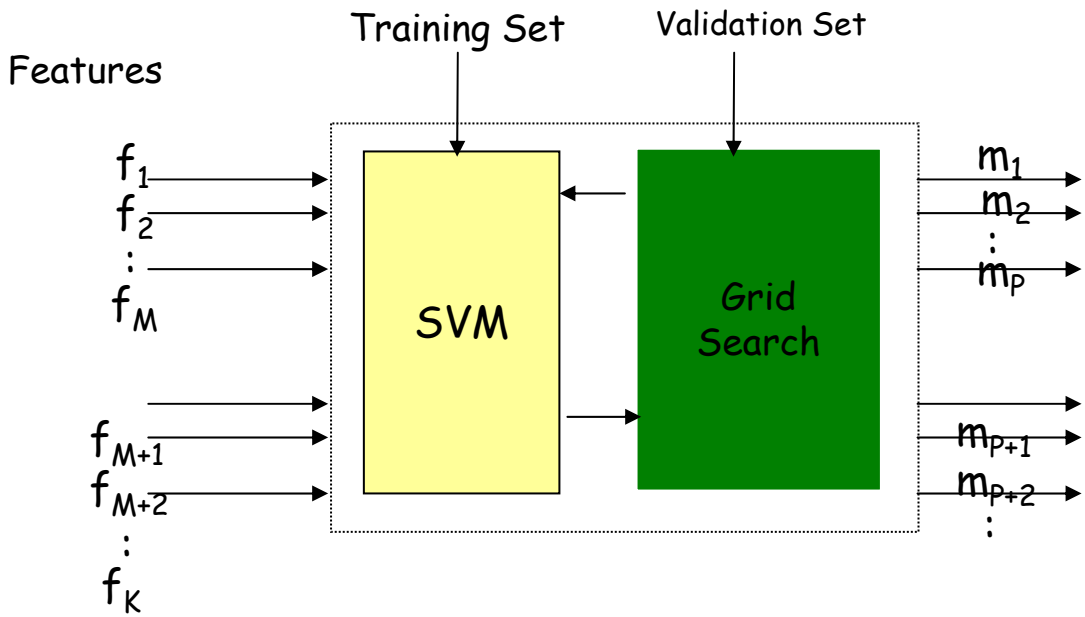
ASR/CC

- LIMSI ASR
- LDC CC
- Alignment provided by CMU
- TF/IDF
- Binary
- Stop-word removal
- Stemming
- Dimensionality reduction

Low-level feature-based Models



Low-level Feature-based Concept Models: Support Vector Machines

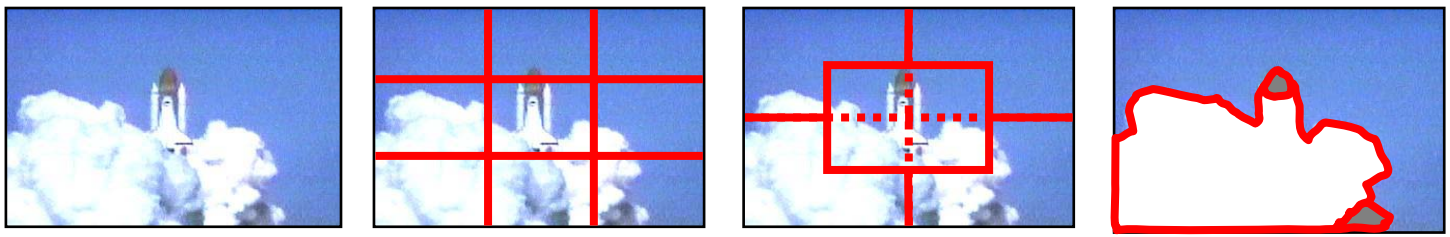


- SVM models for visual and ASR/CC features
- For each concept
  - Avoided early feature fusion due to small number of training samples.
  - Built multiple models for each feature set by varying kernels and parameters.
  - 27 models for different parameter configurations built for each concept
- Validation Set is used to then search for the best model parameters and feature set.

Multi-granular hypotheses testing

Idea

- Using manual annotation, train concept models for regional concepts
- At detection time, identify and test the ***best*** candidate regions:
  - Generate multiple segmentation hypotheses for identifying candidate region set
  - Predict best hypothesis for target concept (e.g., based on performance on independent validation set)
  - Evaluate concept models over regions corresponding to selected segmentation hypothesis
  - Determine final detection confidence score from candidate region scores



Maximum Entropy Approach for Concept detection w/o regional annotation

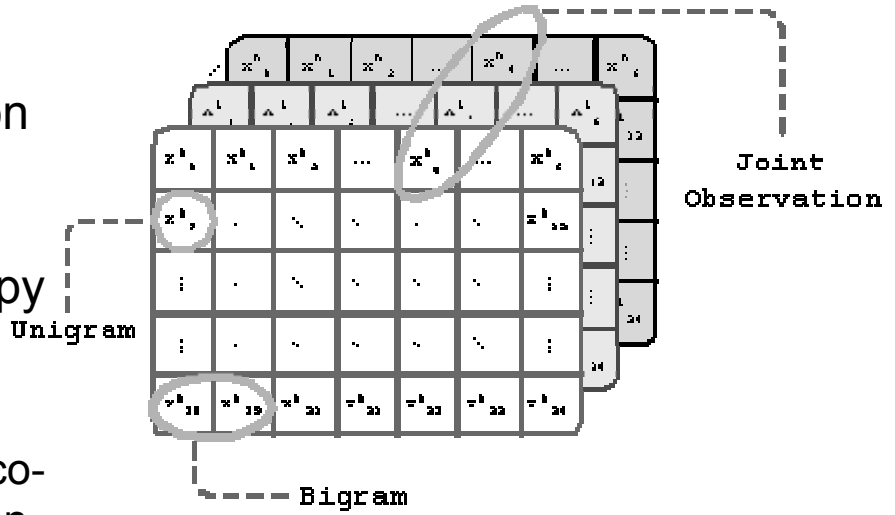
- Key-frame partitioned into regular grid
- Low-level features extracted from each region
- Extracted features are tokenized using K-means.
- Statistical information to the Maximum Entropy model is presented via specially designed predicates:

  Unigram predicates are defined to capture the co-occurrence statistics between manual annotation and tokenized feature.

  Bigram predicates capture the relationships between horizontal and vertical neighboring region.

  Place Dependent predicates are defined to capture location specific statistics.

  Joint Observation predicates are defined to capture interactions between the visual low-level features.

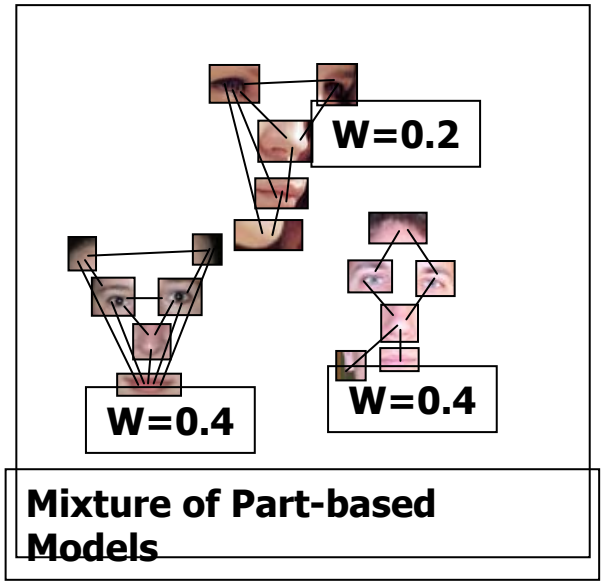


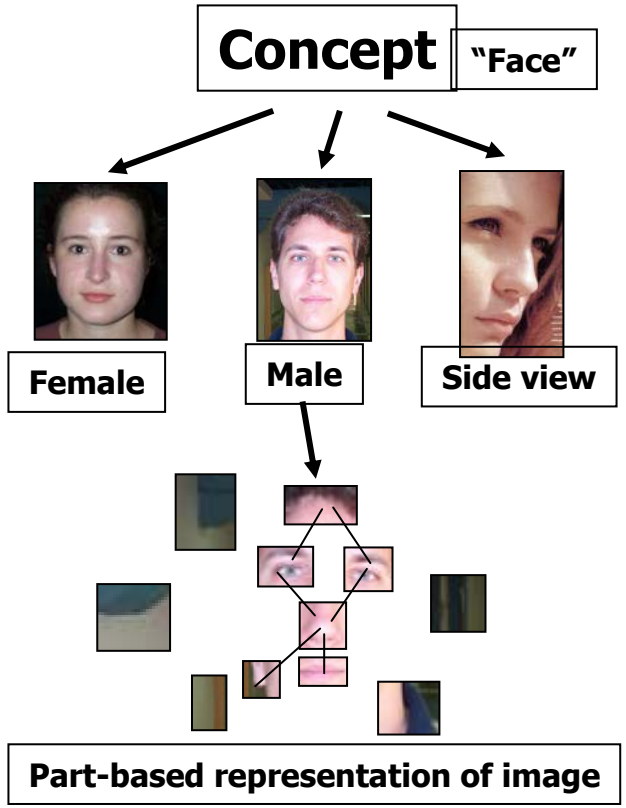
$$p(y|x) = \frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)}{Z_\lambda(x)}$$

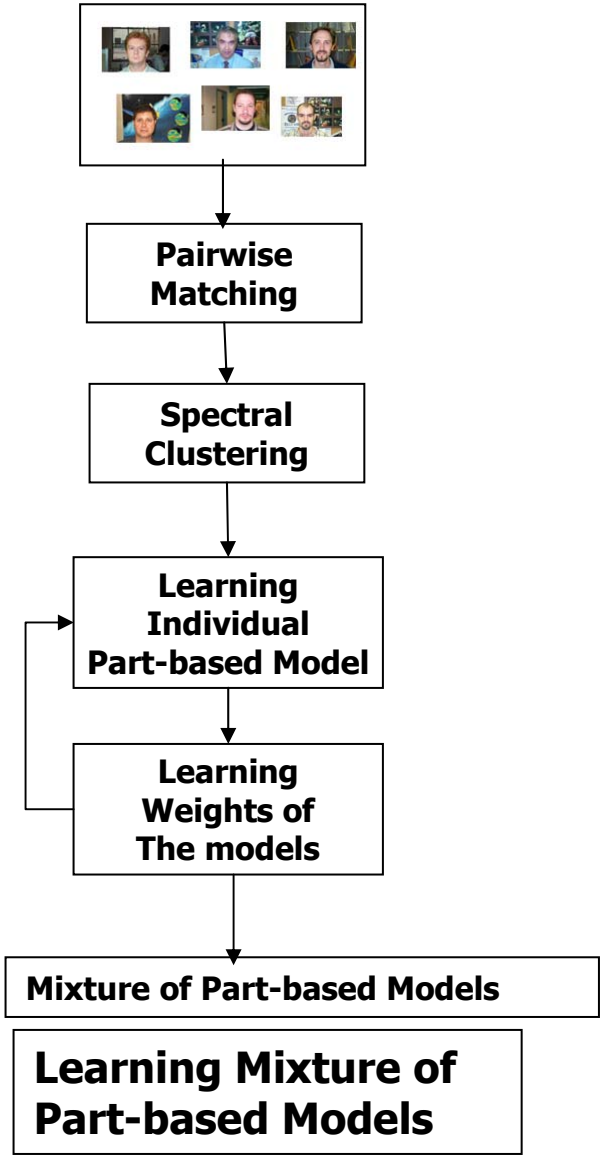
$$Z_\lambda(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x,y)\right)$$

$$\lambda^* = \arg\max_\lambda \Psi(\lambda)$$

Learning Mixture of Part-based Models



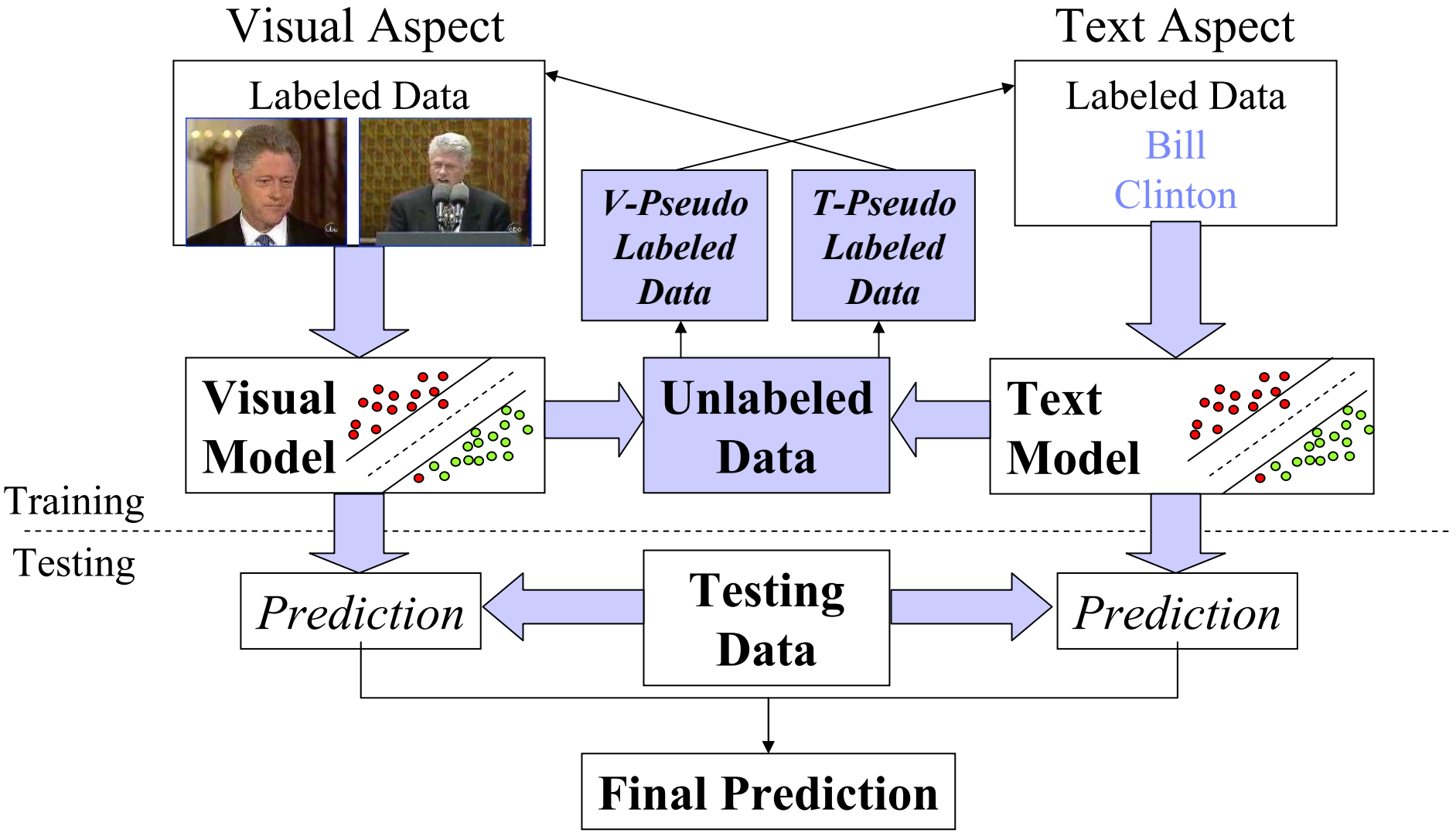




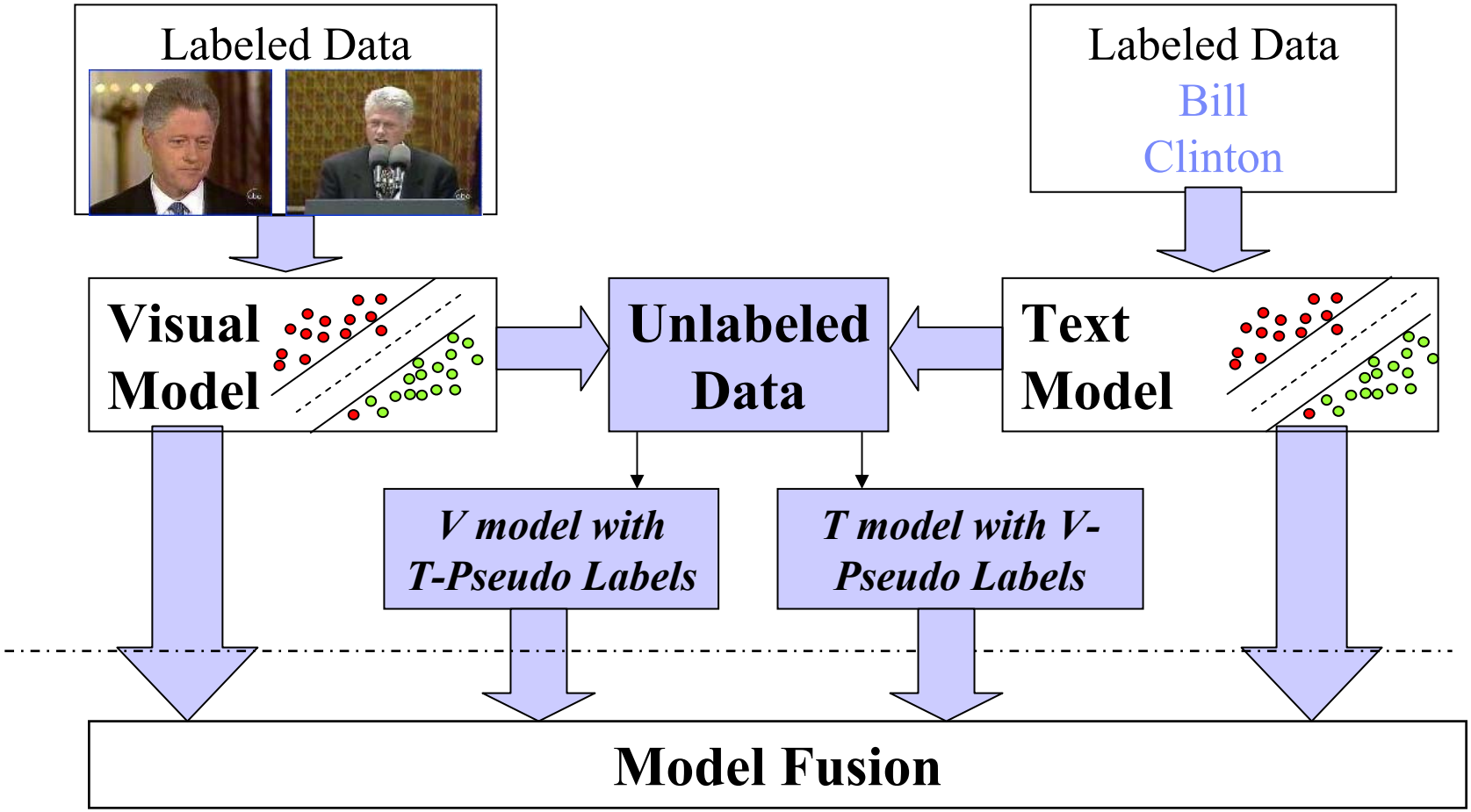
Leveraging Unlabeled Samples

- Experimented with leveraging unlabeled data sets TREC2003, TREC 2004 in conjunction with the labeled common development annotation set.
- Experimented with Co-training and a variation of co-training
- In each case the unimodal classifier was an SVM
- Combination of the multiple modalities and unlabeled and labeled data sets resulted in cross-feature ensemble models.

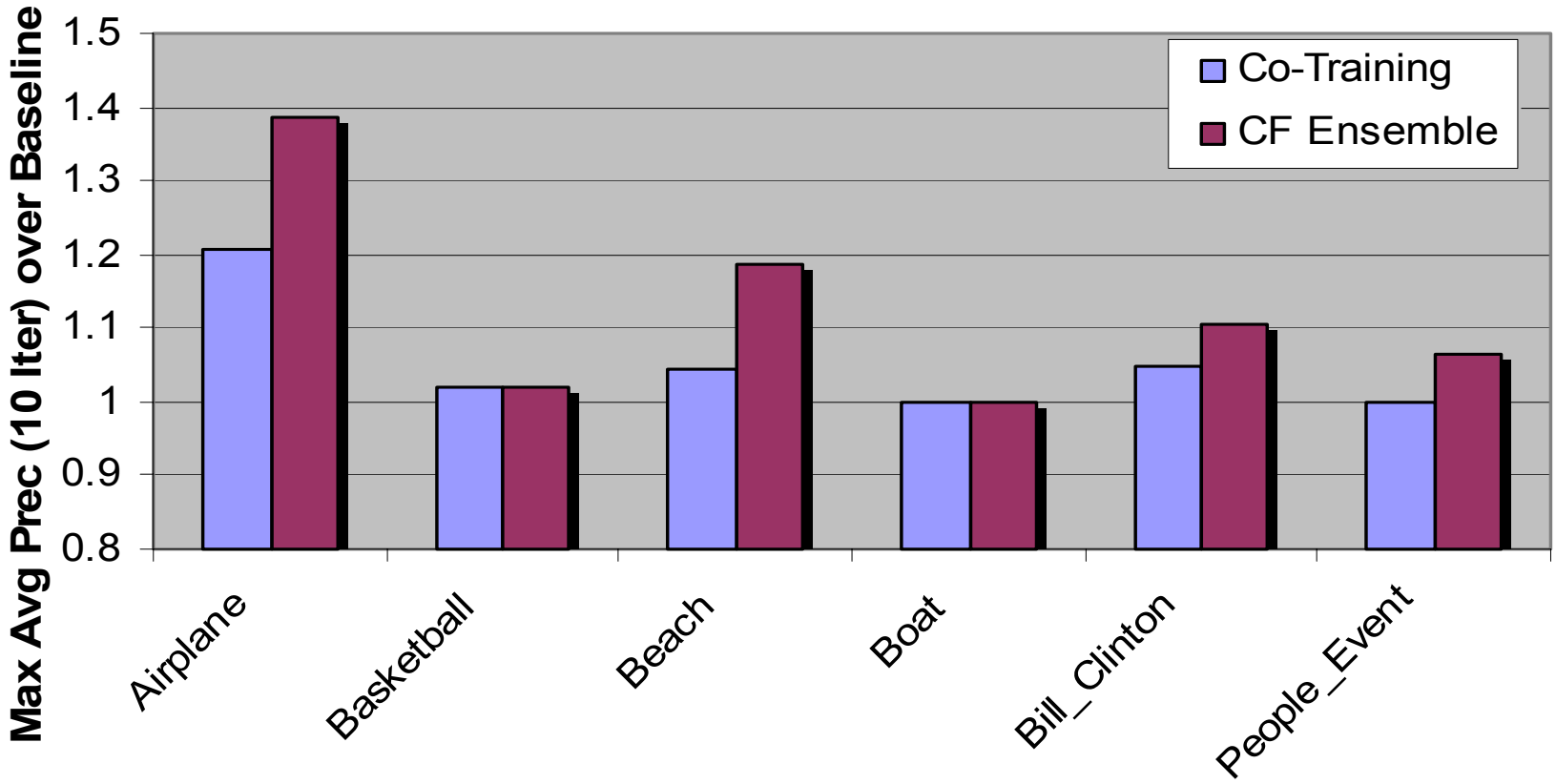
Co-training



Our Approach: Cross Feature Ensemble Learning

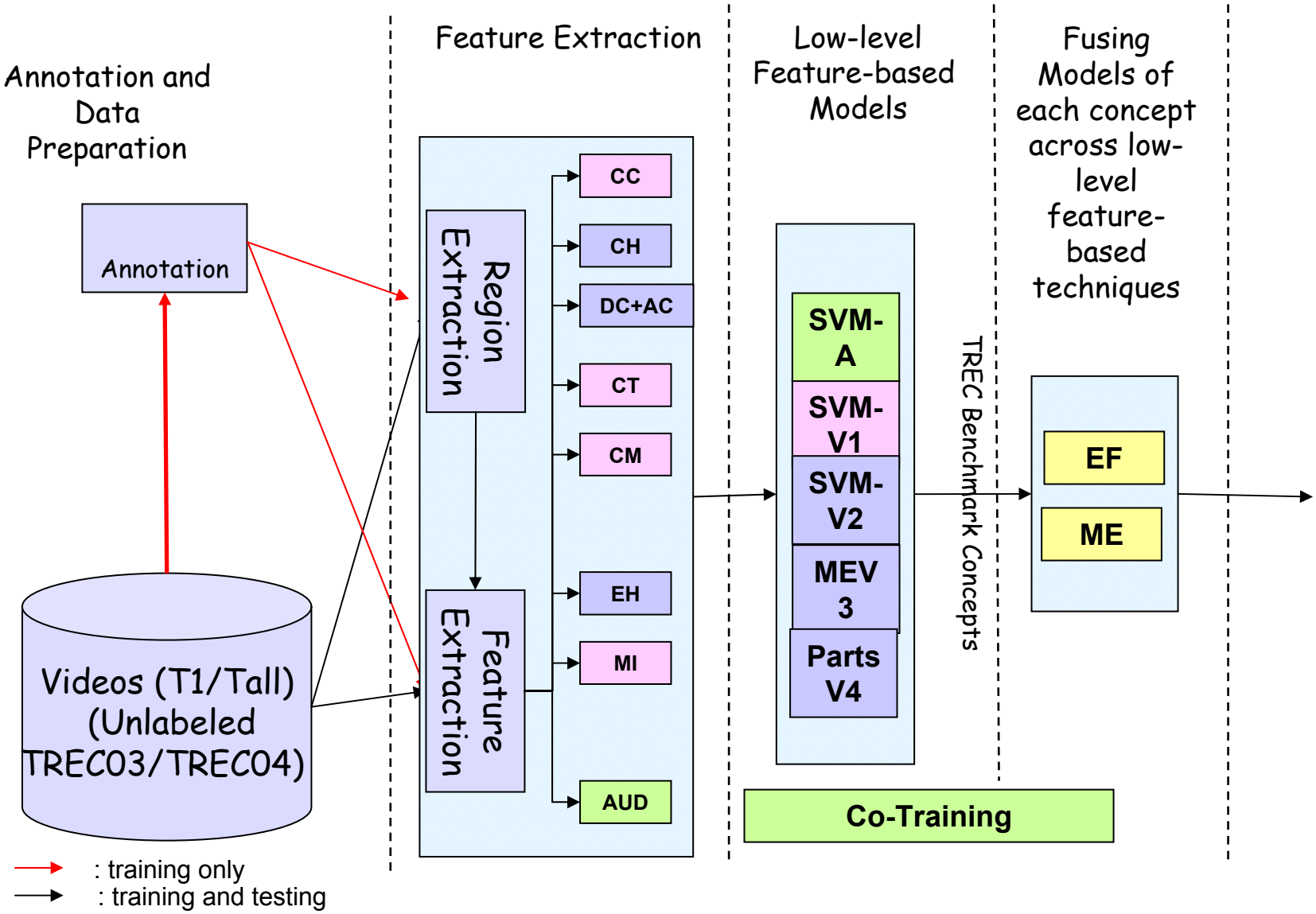


Results: Maximal Average Precision in 10 Iterations

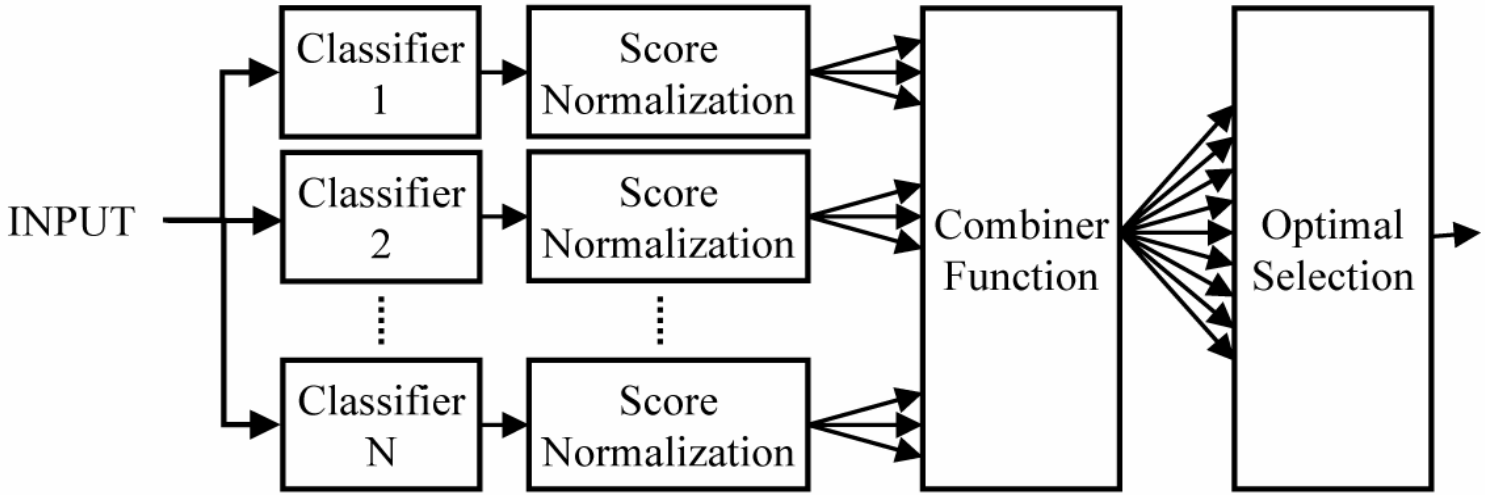


- CFEL has theoretical performance guarantees that Co-training does not.
- Average improvement: CFEL 12 % CT 4 %, Fully Labeled: 14 %

Model-fusion



Multi-Modality/ Multi-Concept Fusion Methods



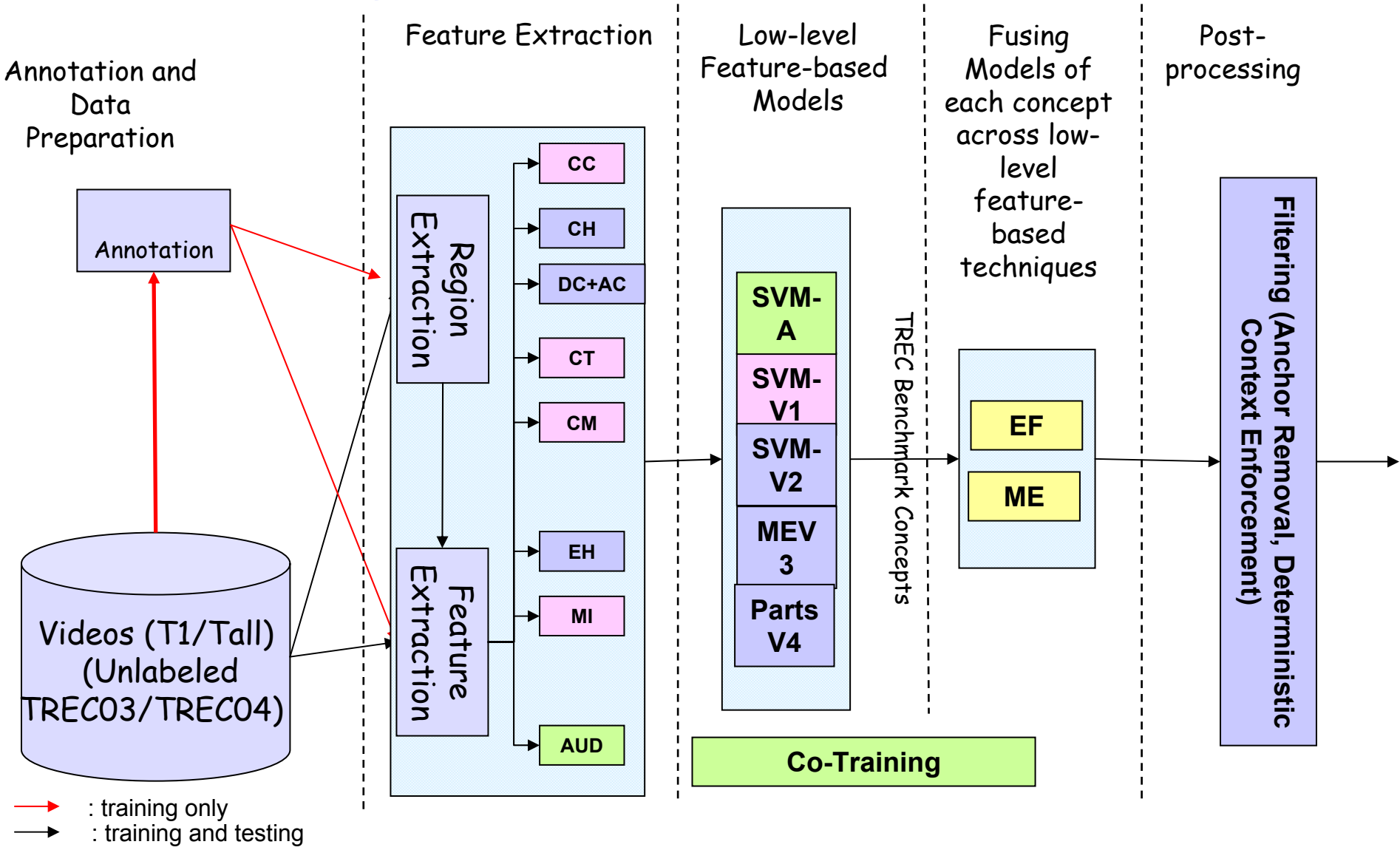
Ensemble Fusion:

- Normalization: rank, Gaussian, linear.
- Combination: average, product, min, max
- Works well for uni-modal concepts with few training examples
- Computationally low-cost method of combining multiple classifiers.

Maximum Entropy Fusion

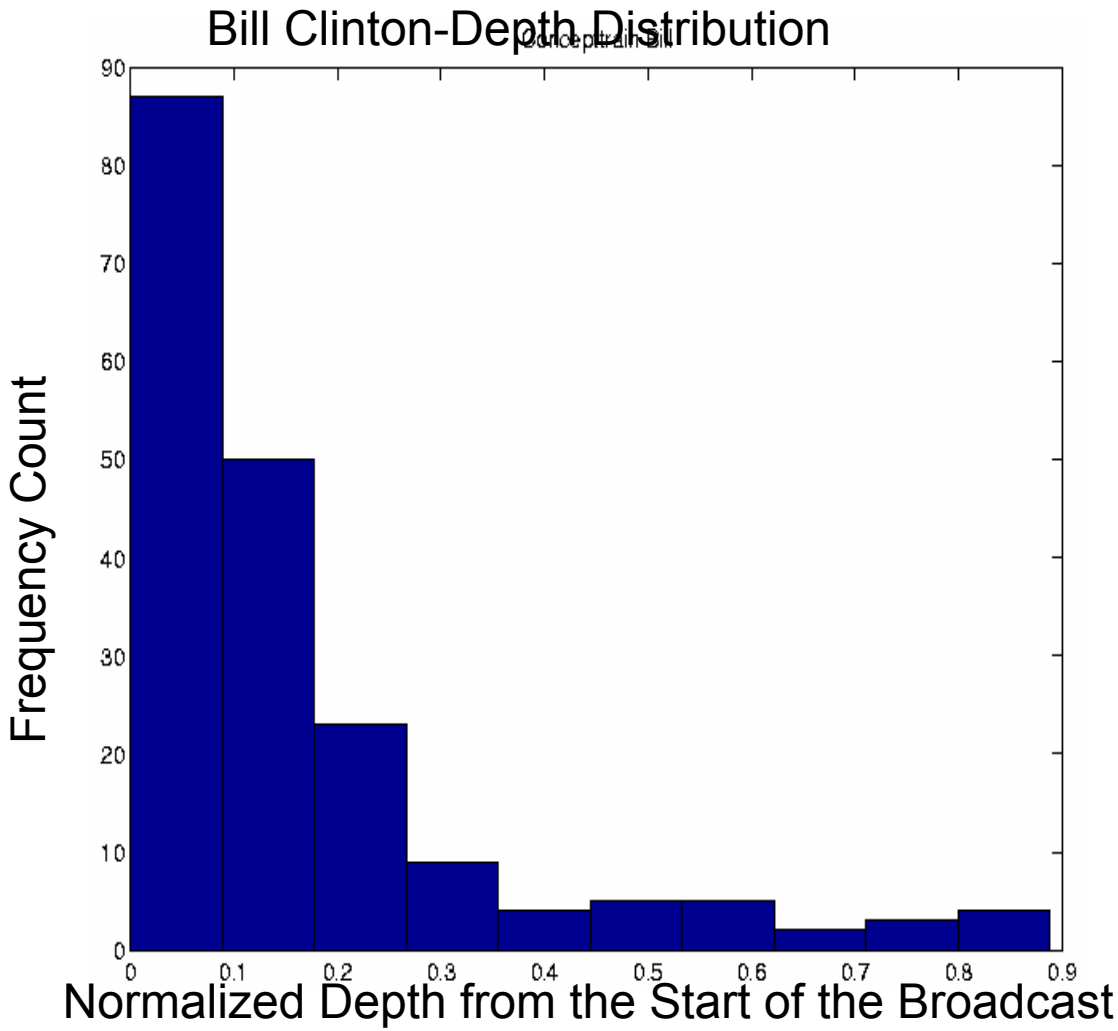
•Similar approach as in ME Detection except that now the supervised ME scheme uses detection results of different models and learns based on joint predicates

Post-processing



Post-processing with Context Filters

- Anchor Detection (Build an anchor detector but can also use Indoors Detector as a surrogate for Anchor Detector)
- Removed top 500 anchor shots from the top of any list
- Modeled the depth of a shot reporting a particular concept, from the beginning of the broadcast. For example Bill Clinton appeared closer to the beginning of the broadcast, while Physical Violence appeared later.
- Built non-parametric density models for the depth distribution of each concept and applied this to provide soft filtering of results
- Also built similar models for shot durations for concepts but not conclusive to help filtering

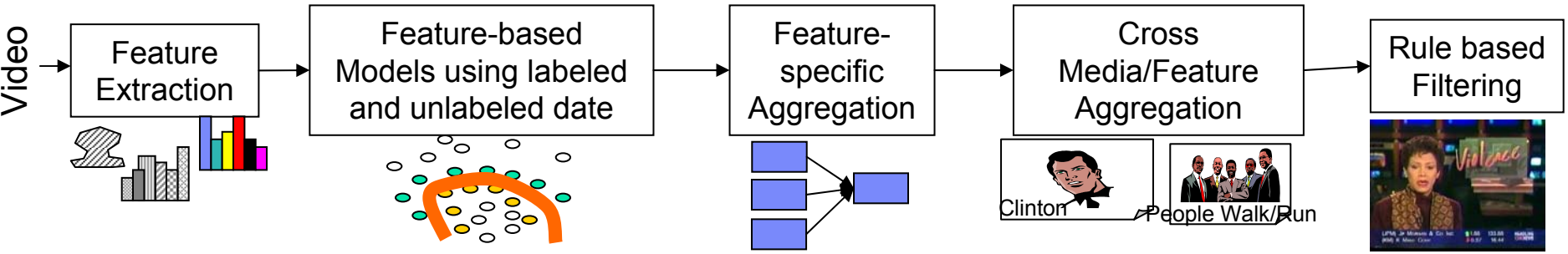


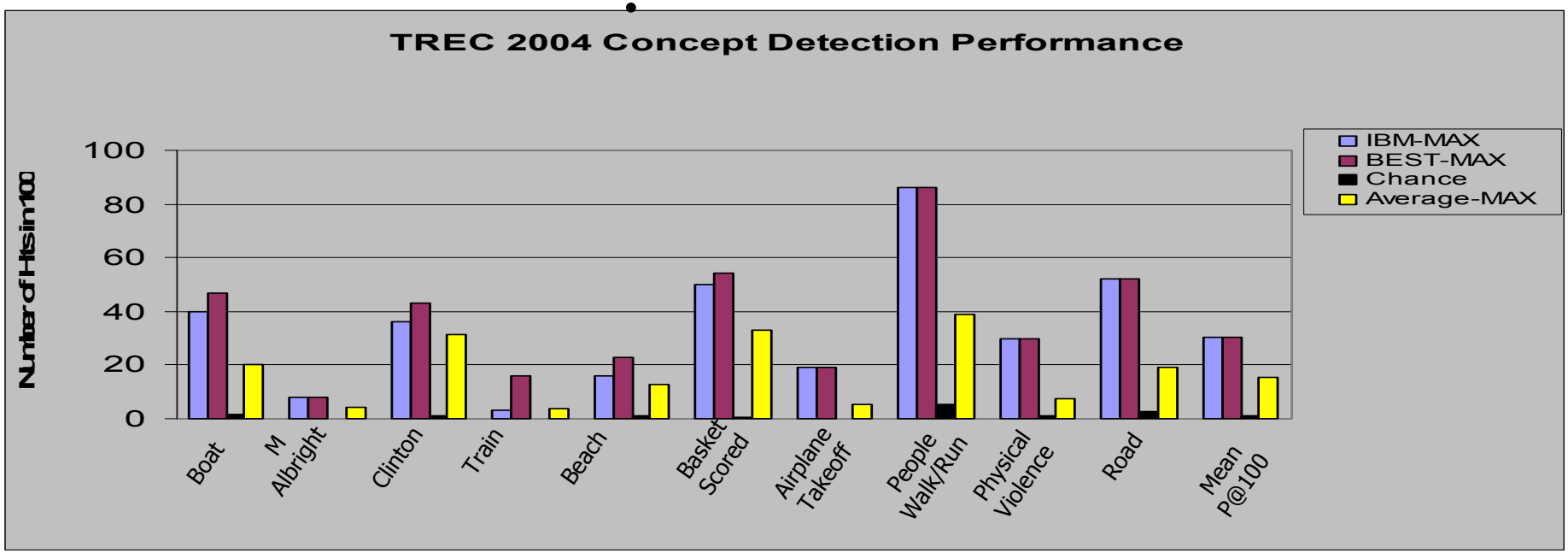
IBM TRECVID 2004 Concept Detection: Logistics

- Submitted runs for all 10 concepts
- All runs were multimodal
- 45 Type A runs
- 82 Runs in all evaluated

BOM: Best combination of single A and V
Mall_T1_EF: All models, Ensemble Fusion
Mall_T1_MEMF: All models, ME Fusion
Mall_Tall_EF: All models, all sets, Ensemble Fusion
CM2all_T1_EF: All models, Co-training, Ensemble Fusion
CM2all_T1_MEMF: All models, Co-training, ME Fusion (TREC03 set as unlabeled set)
CM2all_Tall_EF: All models, Co-training, All sets, EF (TREC03 set as unlabeled set)
CM4all_Tall_EF: All models, Co-training, All sets, EF (TREC04 set as unlabeled set)
Filter1: Mall_T1_EF filtered (w/anchor, depth filtered for 2 concepts)
Filter2: CM2all_Tall_EF filtered (w/anchor, depth filtered for 2 concepts)

TRECVID 2004: Results at a glance

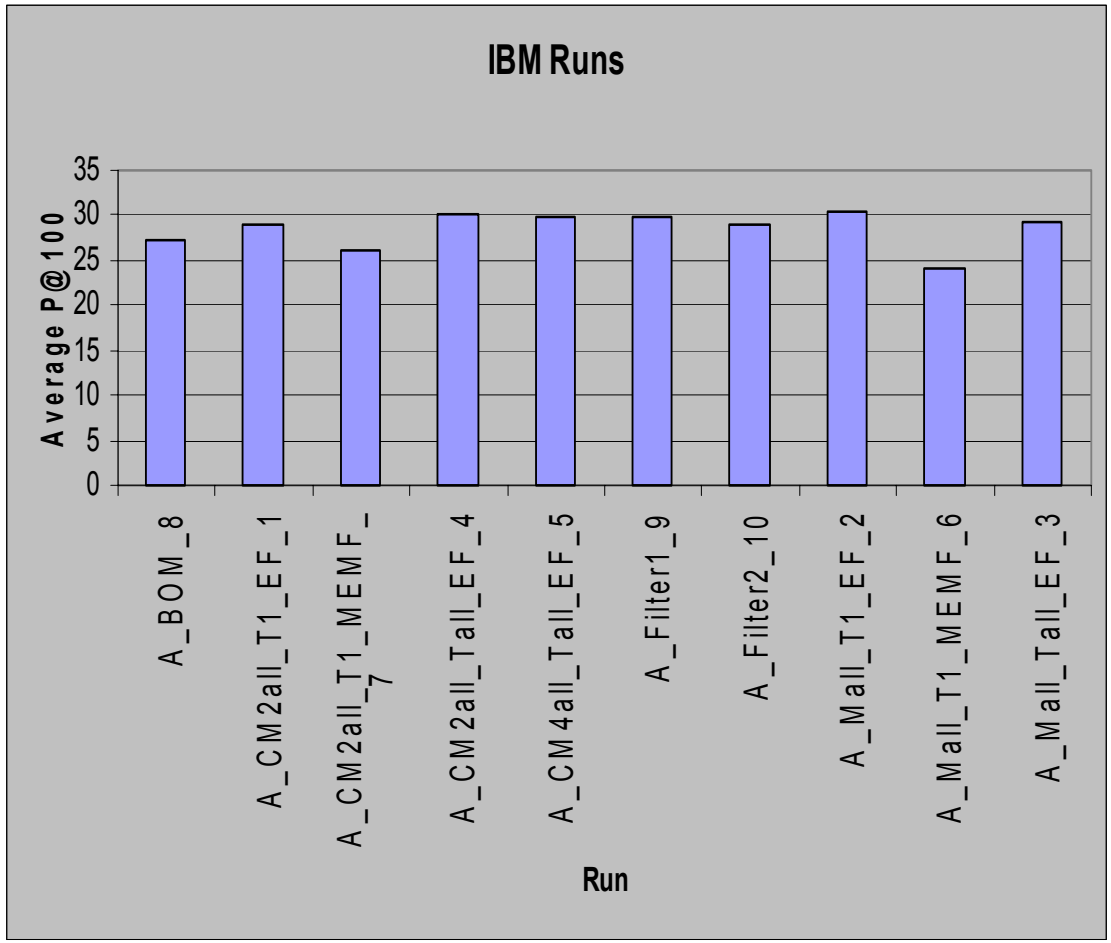




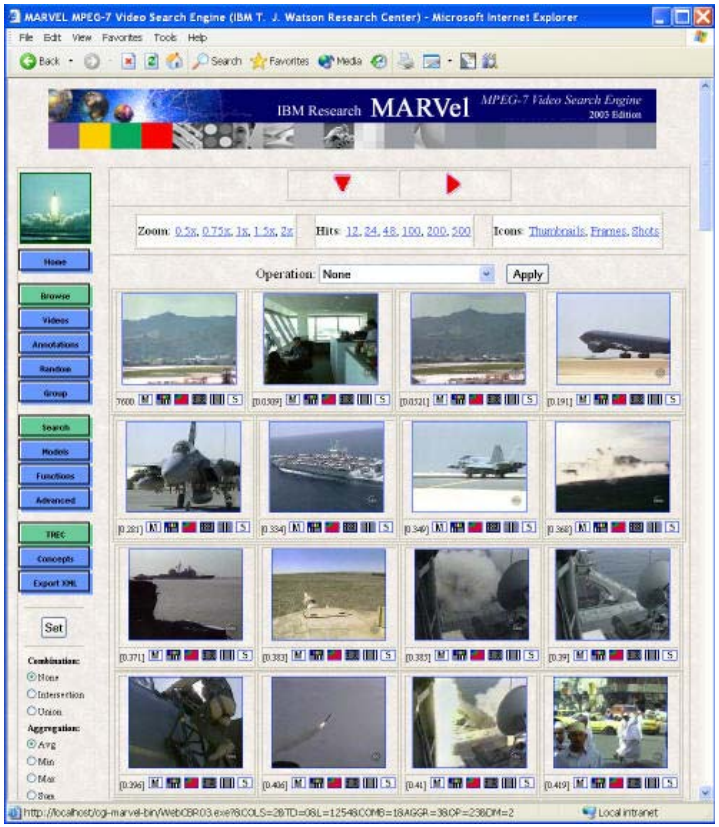
- All IBM runs in the top 11 runs based on MAP
- IBM tops AP in 4/10 concepts, tops P@100, P@1000 and P@2000 in 5/10 concepts
- All IBM concept APs above the median concept AP across all runs

New Directions Explored and Lessons Learnt

- CFEL: Improves precision towards the top by performing re-ranking. Well suited for rare concepts.
- Maximum Entropy Modeling: Works well for concepts with large number of training samples but does not generalize as well on unseen data set. Also does not work well as a fusion strategy at least for infrequent concepts
- Feature Selection: Turned out to be important for choosing smaller but more discriminative features to larger complex features that were suited for frequent concepts. Layout helped in the compressed domain feature based models.
- Classifier: SVM performed better than MaxEnt with or without co-training.
- Filtering improves slightly in P@100 for Physical Violence. For most concepts, there is no significant improvement.



Online Demo



http://mp7.watson.ibm.com