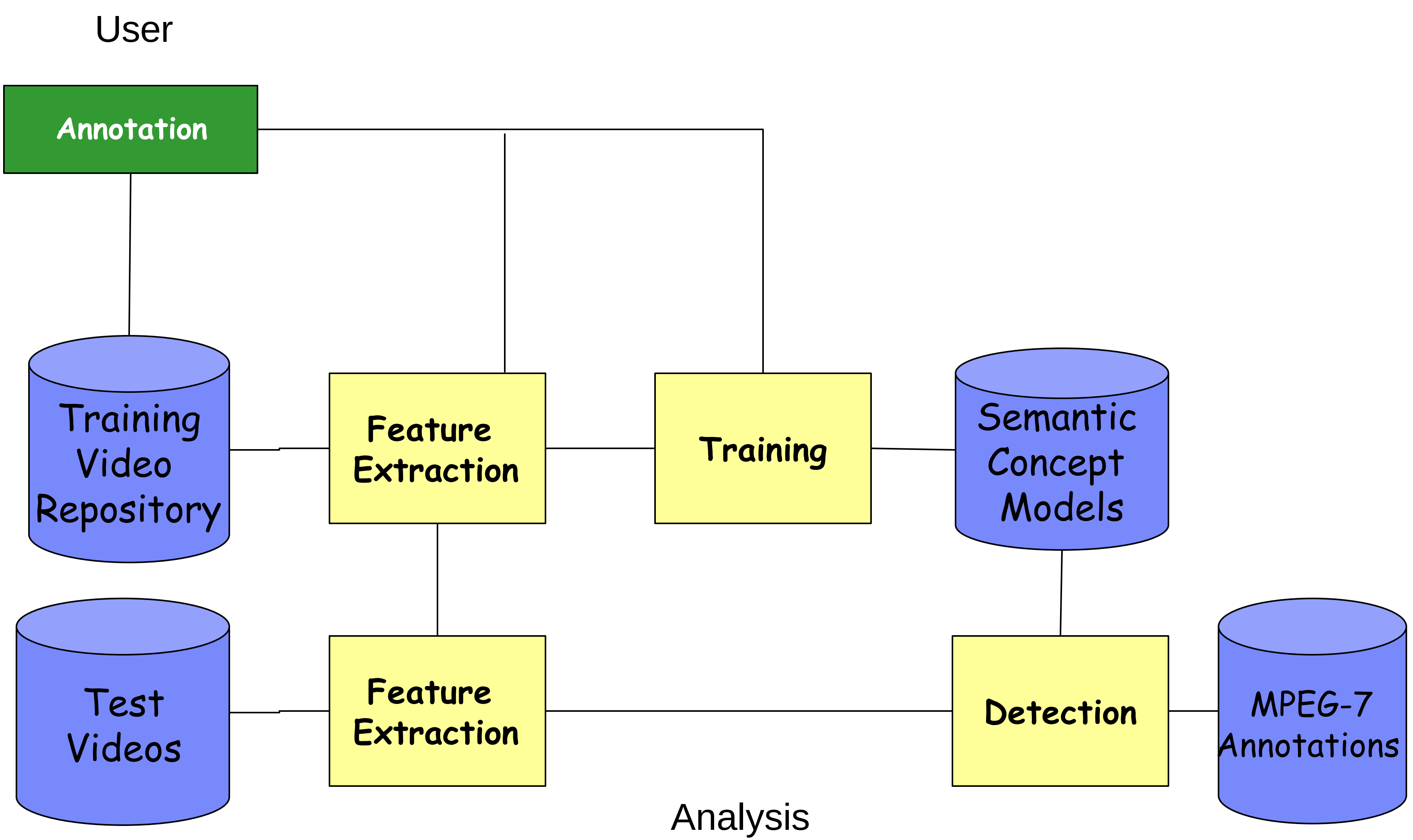
The IBM Semantic Concept Detection Framework

Arnon Amir, Giri Iyengar, Ching-Yung Lin, Chitra Dorai,
Milind Naphade, Apostol Natsev, Chalapathy Neti,
Harriet Nock, Ishan Sachdev, John Smith,
Yi Wu, Belle Tseng, Dongqing Zhang

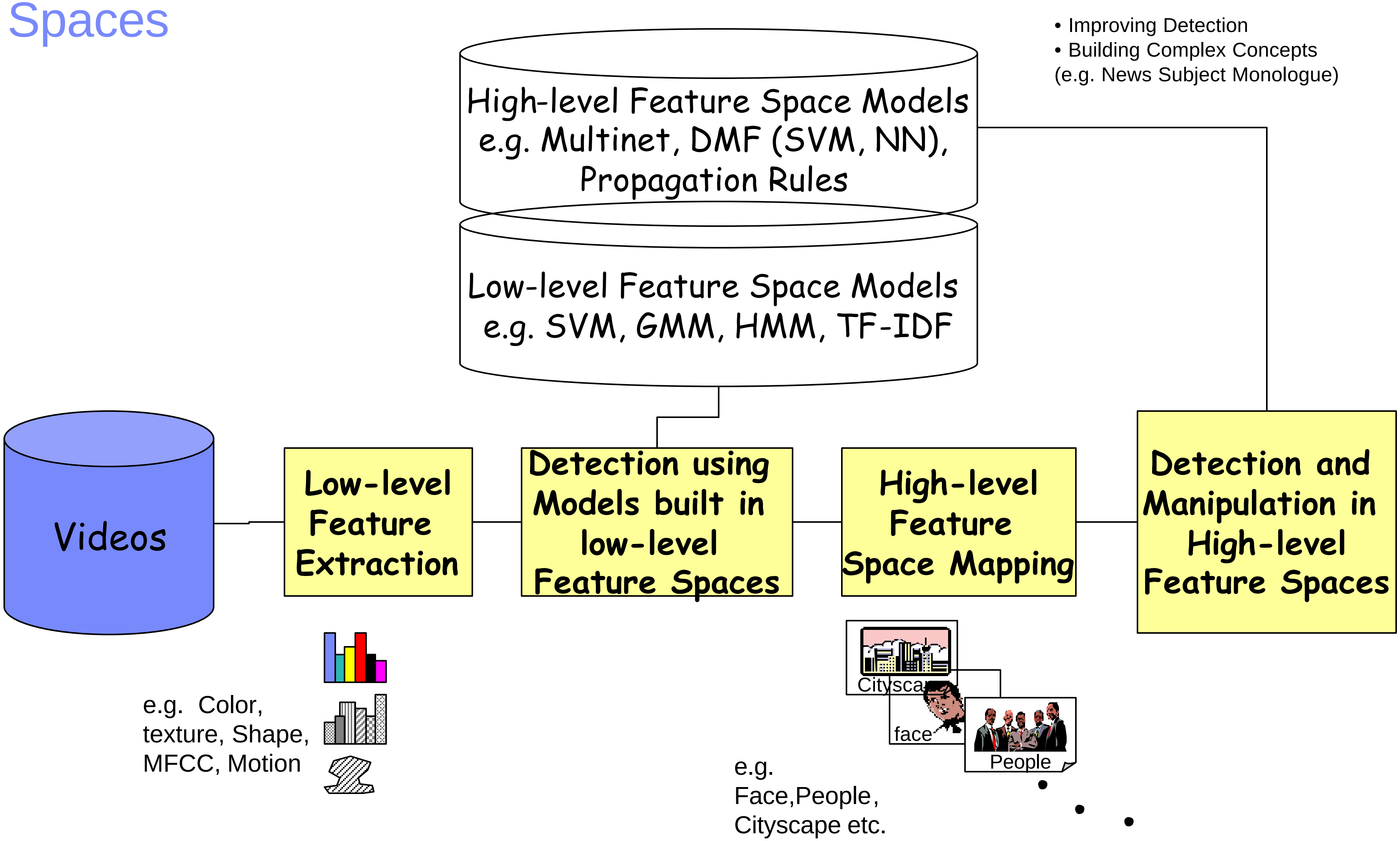
Outline

- Concept Detection as a Machine Learning Problem
- The IBM TREC 2003 Concept Detection Framework
  - Modeling in Low-level Features
  - Multi-classifier Decision fusion
  - Modeling in High-level (semantic) Features
- Putting it All Together: TREC 2003 Concept Detection
- Observations

Multimedia Analytics by Supervised Learning



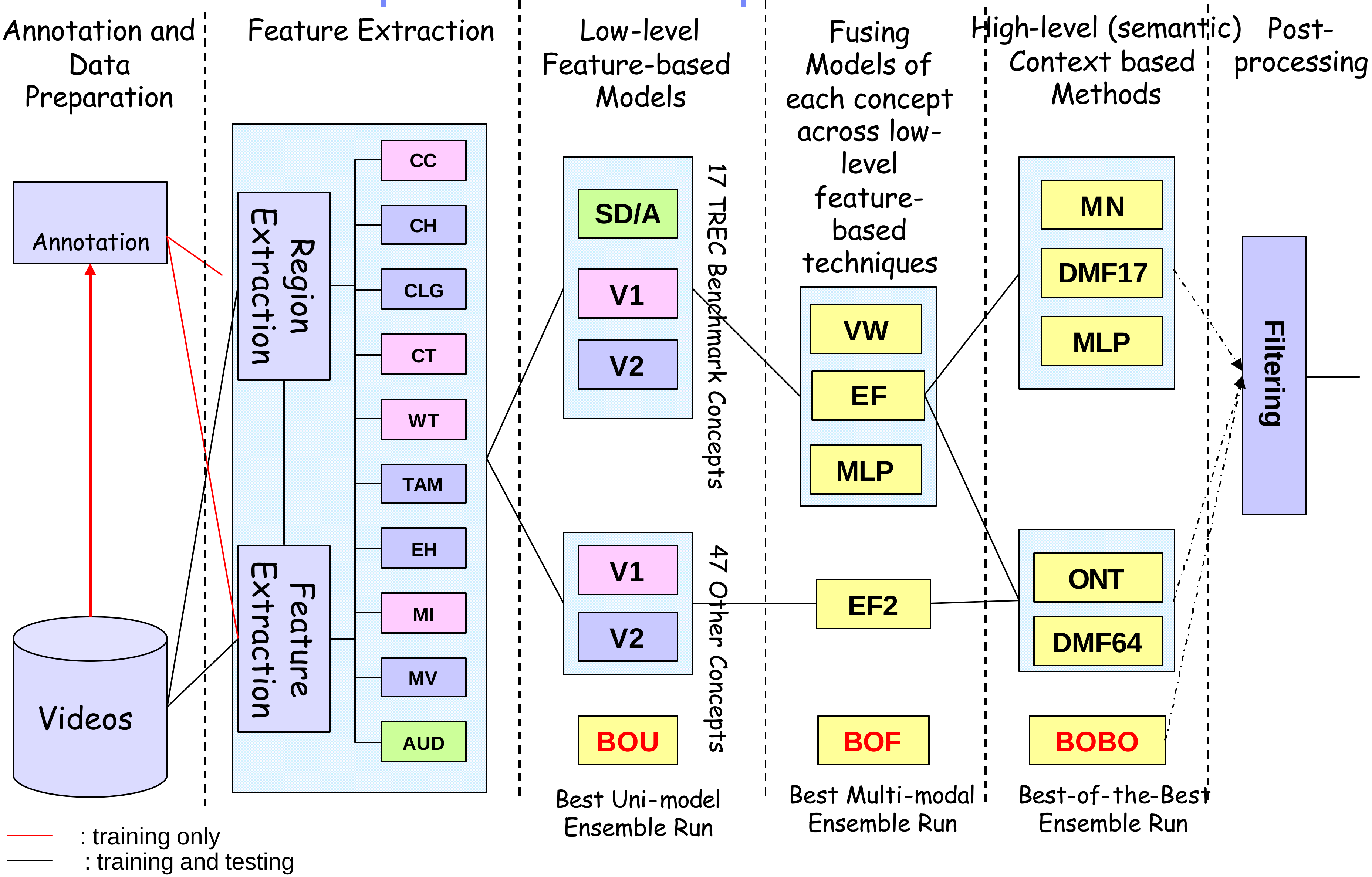
Multi-layered Concept Detection: Working in Increasingly (Semantically) Meaningful Feature Spaces



The Evolving IBM Concept Detection System

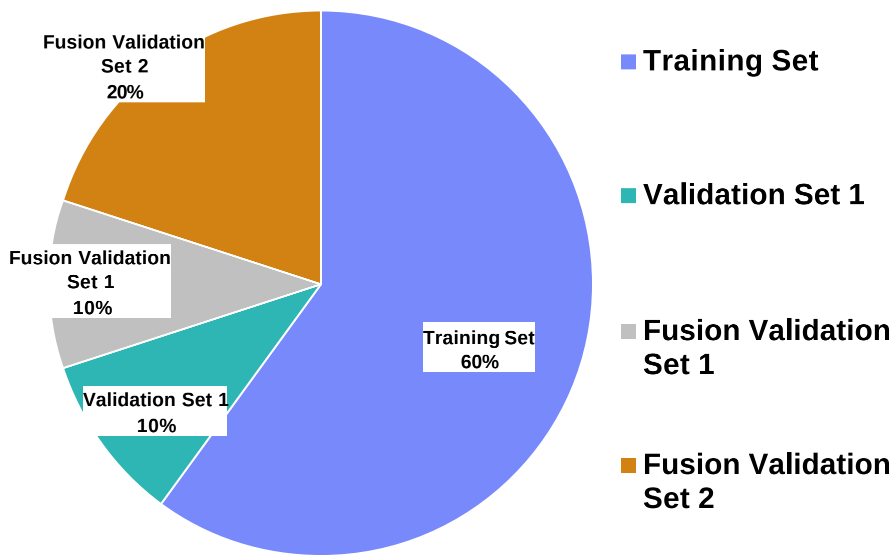
IBM TREC'01, 02	Post TREC' 02 Experiments	IBM TREC'03
Use of SVM, GMM and HMM Classifiers for modeling low-level features	Use of SVM, GMM and HMM Classifiers for modeling low-level and high-level features	Use of SVM, GMM and HMM Classifiers for low-level and high-level features
Ensemble and Discriminant Fusion (TREC02) of Multiple Models of Same Concept Improved performance over single models	Ensemble and Discriminant Fusion of Multiple Models of Same Concept Improved performance over single models	Ensemble and Discriminant Fusion of Multiple Models of Same Concept Improved performance over single models
		Rule-based Preprocessing (e.g. Non-Studio Setting= (**NOT** (Studio_Indoor_Setting)) **OR** (Outdoors))
	Validity Weighted Similarity Improves Robustness	Validity Weighted Similarity Improves Robustness
	Semantic feature based Models (Multinet, DMF) Improves Performance over Single-concept models	Semantic feature based Models (Multinet, DMF-SVMs, NN, Boosting), Ontology Improves Performance over Single-concept models
		Post-Filtering Improves Precision

Video Concept Detection Pipeline



Corpus Issues

- Multi-layered Detection Approach needs multiple sets for cross validation
- Partitioning of Feature Development Set so that each level of processing has a training set and a test set partition that is unadulterated by the processing at the previous level.
- E.g. Low-level feature based concept models built using Training Set and performance optimized over Validation Set.
- Single-Concept, multi-model fusion is performed using Validation Set for training and Fusion Validation Set 1 for testing.
- Semantic-level fusion is performed by using Fusion Validation Set 1 as the training set and Fusion Validation Set 2 as the test set
- Runs submitted to NIST are chosen finally on performance of all systems and algorithms on Fusion Validation Set 2.



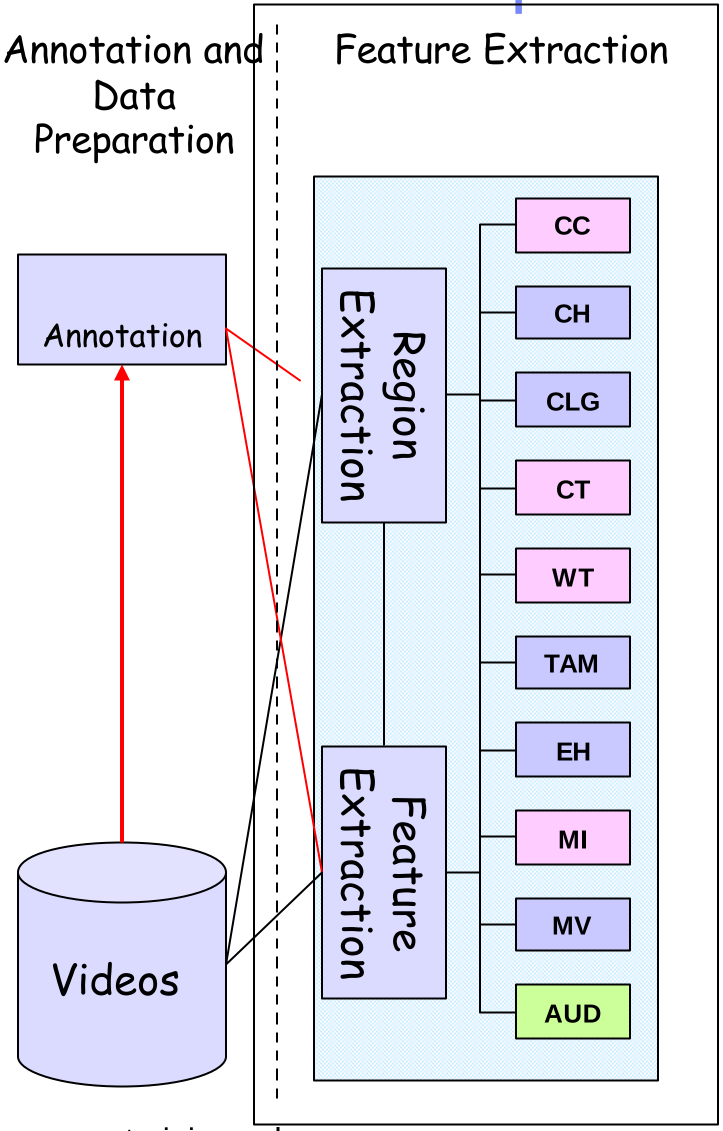
Partitioning procedure

All videos aligned by their temporal order and

For each set of 10 videos

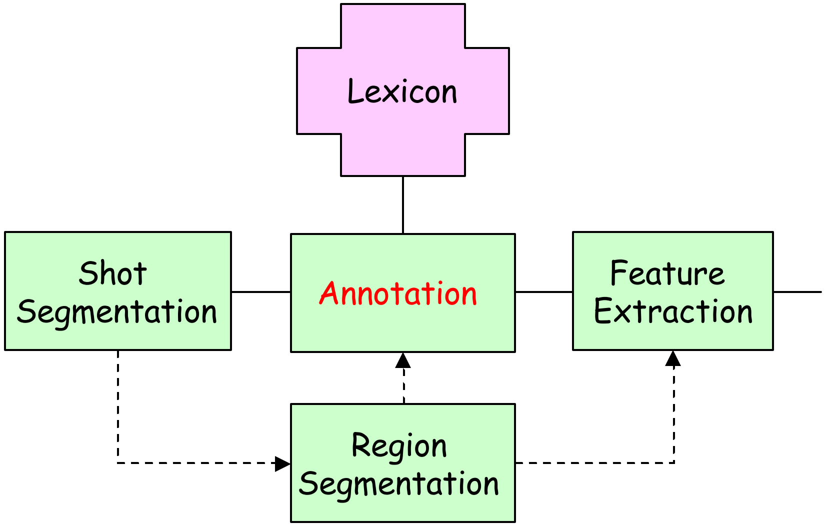
- First 6 -> Training Set,
- 7th -> Validation
- 8th -> Fusion Validation Set 1
- Last 2 ->Fusion Validation Set 2.

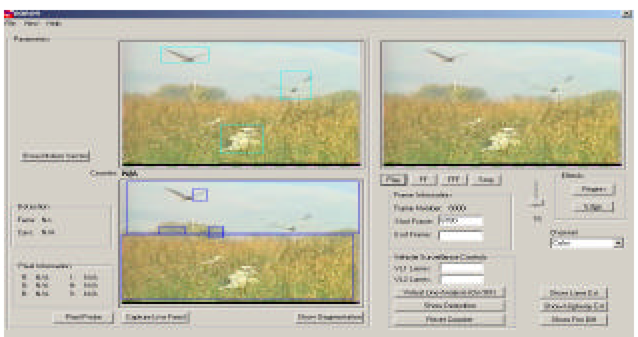
Video Concept Detection Pipeline: Features



: training only
: training and testing

Feature Extraction





Features extracted globally and regionally

Color:

Color histograms (512 dim), Auto-Correlograms (166 dim)

Structure & Shape:

Edge orientation histogram (64 dim), Dudani Moment Invariants (6 dim),

Texture

Co-occurrence texture (96 dim), Coarseness (1 dim), Contrast (1 dim), Directionality (1 dim), Wavelet (12 dim)

Motion

Motion vector histogram (6 dim)

Audio

MFCC

Text

ASR Transcripts

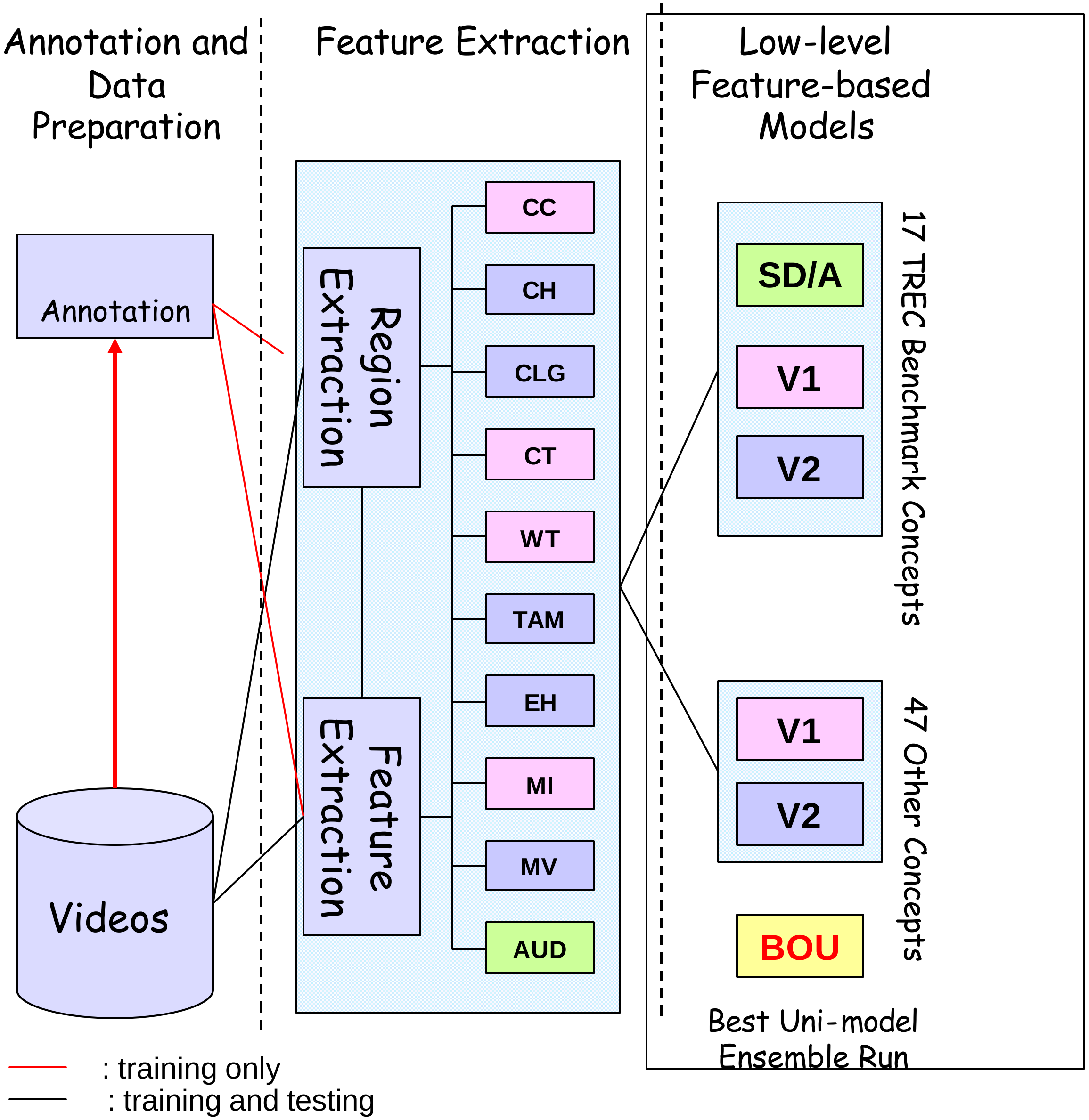
Regions

Object (motion, Camera registration)

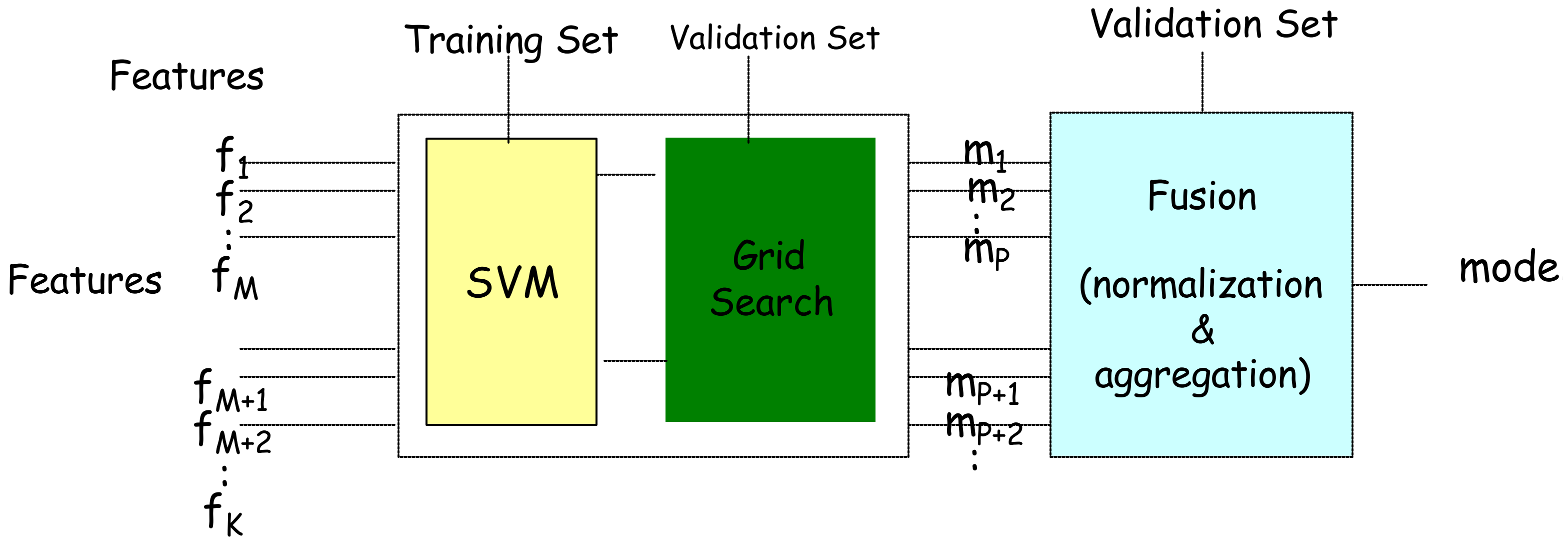
Background (5 regions / shot)

References: Lin (ICME 2003)

Video Concept Detection Pipeline: Low-level Feature Modeling



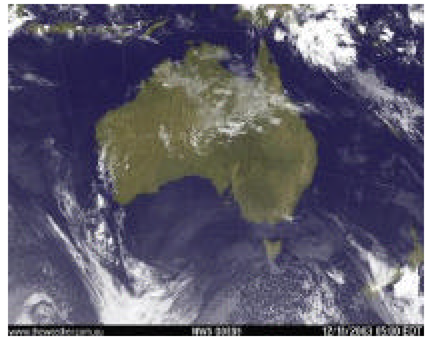
Low-level Feature-based Concept Models
Statistical Learning for Concept Building: SVM



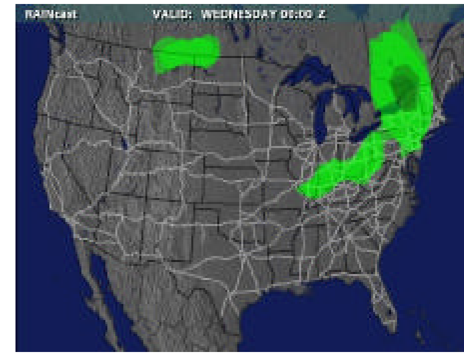
- SVM models used for 2 sets of visual features
  - Combined Color correlogram, edge histogram, cooccurrence features and moment invariants
  - Color histogram, motion, Tamura texture features
- For each concept
  - Built multiple models for each feature set by varying kernels and parameters.
  - Upto 27 models for each concept built for each feature type
- A total of 64 concepts from the TREC 2003 lexicon covered through SVM-based models
- Validation Set is used to then search for the best model parameters and feature set.
- Identical Approach as in IBM System for TREC 2002
- **Fusion Validation Set II MAP: 0.22**
- **References: IBM TREC 2002, Naphade et al (ICME 2003, ICIP 2003)**

Low-level Feature-based Concept Models: Statistical Learning for Concept Building based on ASR Transcripts

TRAINING:
Manually examine examples to find frequently co-occurring relevant words



… some weather news overseas



… update on low pressure storm

WEATHER NEWS QUERY WORD SET:
weather news low pressure storm cloudy mild windy … (etc) …

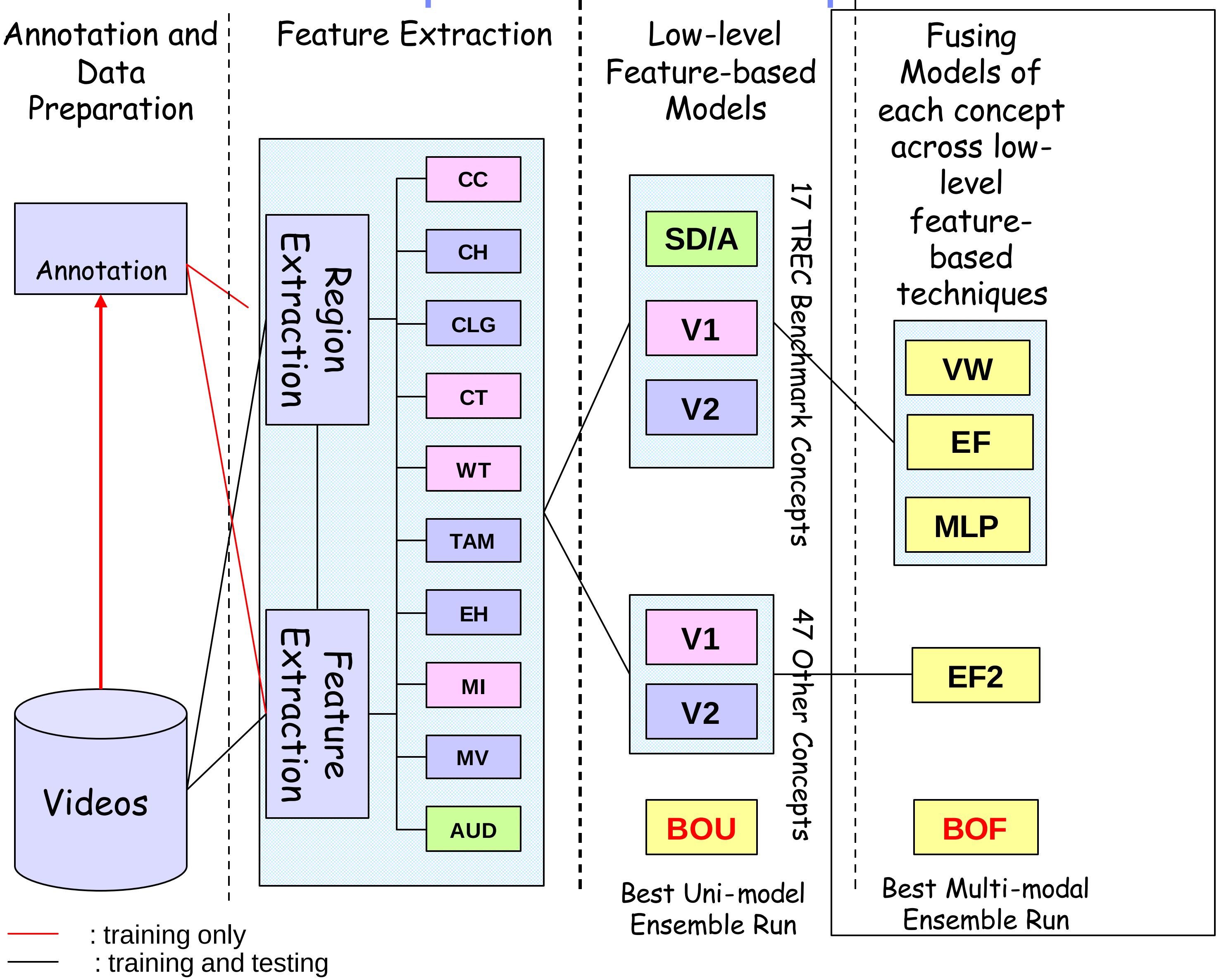
OKAPI SYSTEM FOR SEARCH TEXT ASR TRANSCRIPTS

Ranked Shots

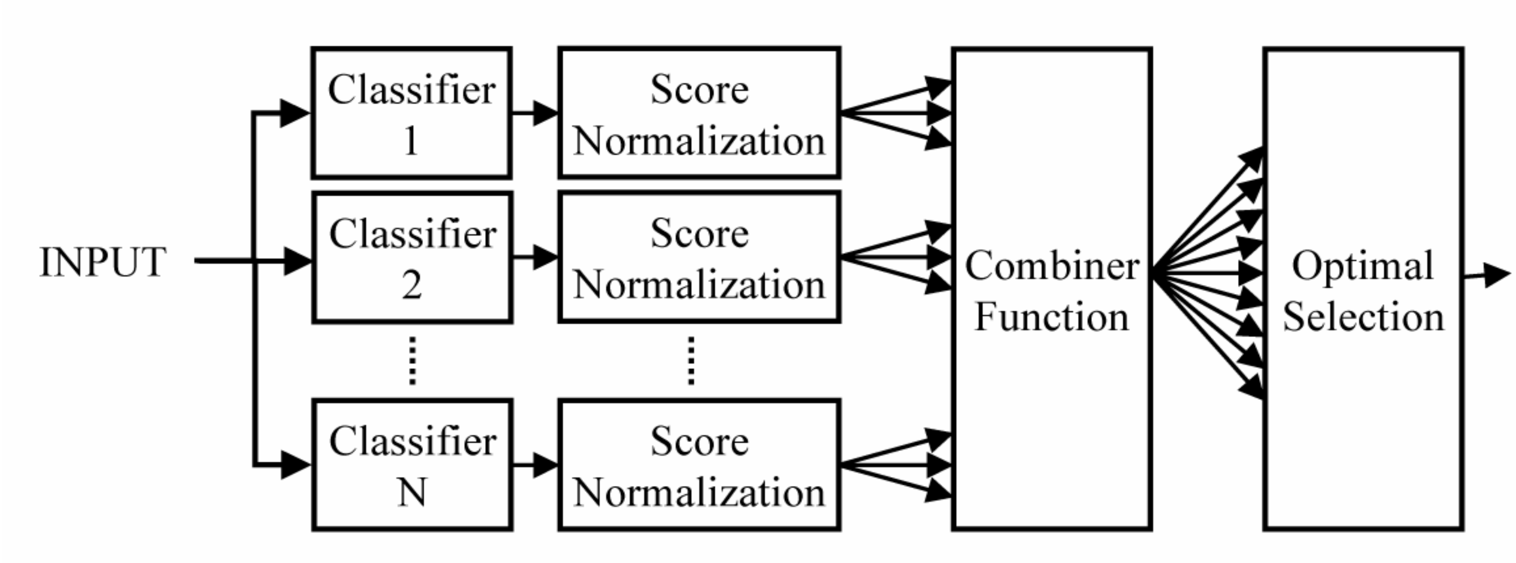
Fusion Validation II MAP = 0.19

References: Nock et al (SIGIR 2003)

Video Concept Detection Pipeline: Fusion I



Multi-Modality/ Multi-Concept Fusion Methods



Ensemble Fusion:

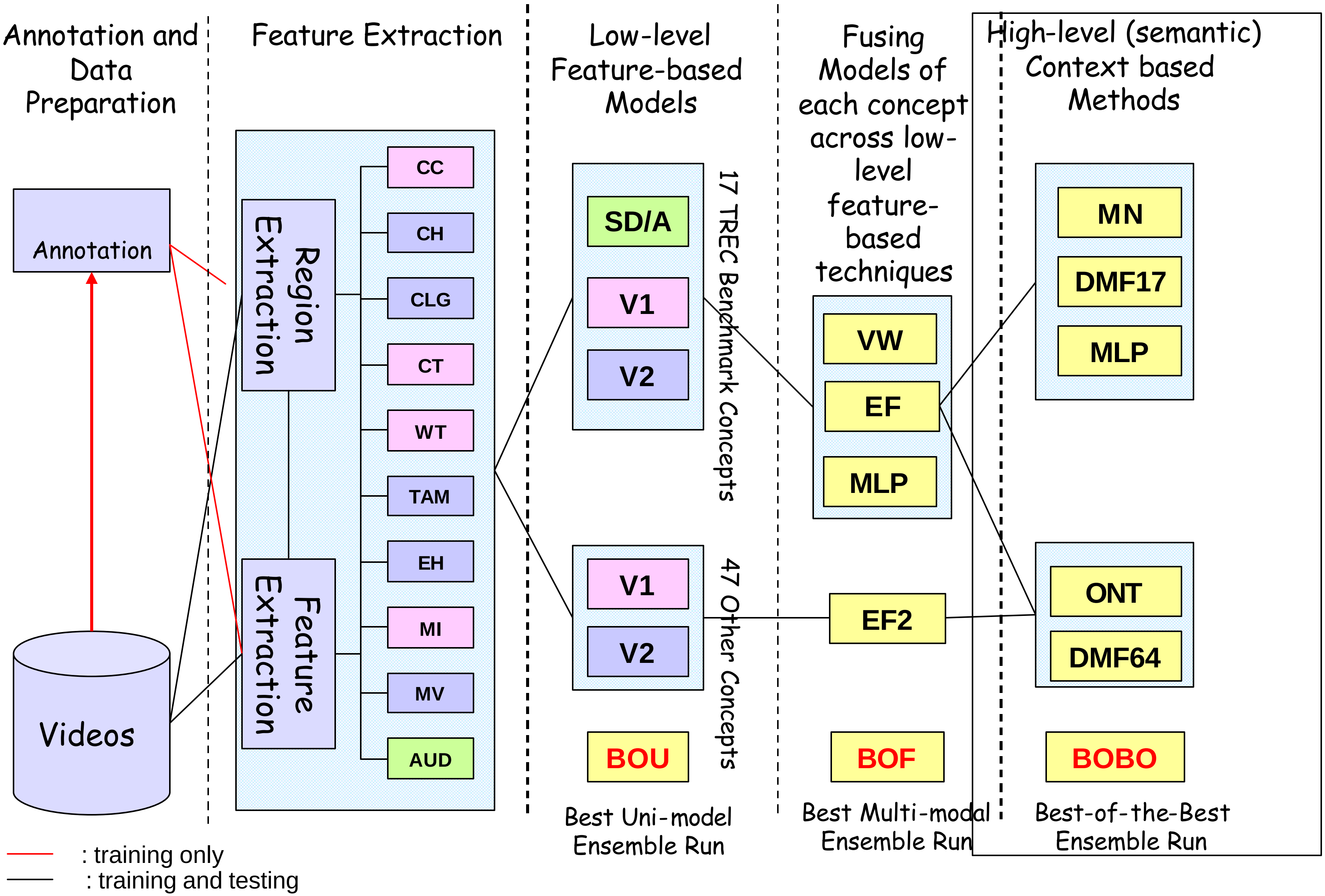
- Normalization: rank, Gaussian, linear.
- Combination: average, product, min, max
- Works well for uni-modal concepts with few training examples
- Computationally low-cost method of combining multiple classifiers.
- Fusion Validation Set II MAP: 0.254
- SearchTest MAP: 0.26
- References: Tseng et al (ICME 2003, ICIP 2003)

Multi-Modality/ Multi-Concept Fusion Methods: Validity Weighting

Validity Weighting:

- Work in the high-level feature space generated by classifier confidences for all concepts
- Basic idea is to give more importance to reliable classifiers.
- Revise distance metric to include a measure of the goodness of the classifier.
- Many fitness or goodness measures
  - Average Precision
  - 10-point AP
  - Equal Error rate
  - Number of Training Samples in Training Set.
- Computationally efficient and low-cost option of merit/performance-based combining multiple classifiers based on
- Improves robustness due to enhanced reliability on high-performance classifiers.
- Fusion Validation Set II MAP: 0.255
- References: Smith et al (ICME 2003, ICIP 2003)

Video Concept Detection Pipeline: Semantic-Feature based Models



Semantic Feature Based Models Incorporating Context

- Multinet: A probabilistic graphical context modeling framework that uses loopy probability propagation in undirected graphs. Learns conceptual relationships automatically and uses this learnt relationships to modify detection (e.g. Uses Outdoor Detection to influence Non-Studio Setting in the right proportion)
- Discriminant Model Fusion using SVMs: Uses a training set of semantic feature vectors with ground truth to learn dependence of model outputs across concepts.
- Discriminant Model Fusion AND Regression using Neural Networks and Boosting: Uses a training set of semantic feature vectors with ground truth to learn dependence of model outputs across concepts. Boosting helps especially with rare concepts.
- Ontology-based processing: Use of the manually constructed annotation hierarchy (or ontology) to modify detection of root nodes based on robust detection of parent nodes. i.e. Use “Outdoor” detection to influence detection

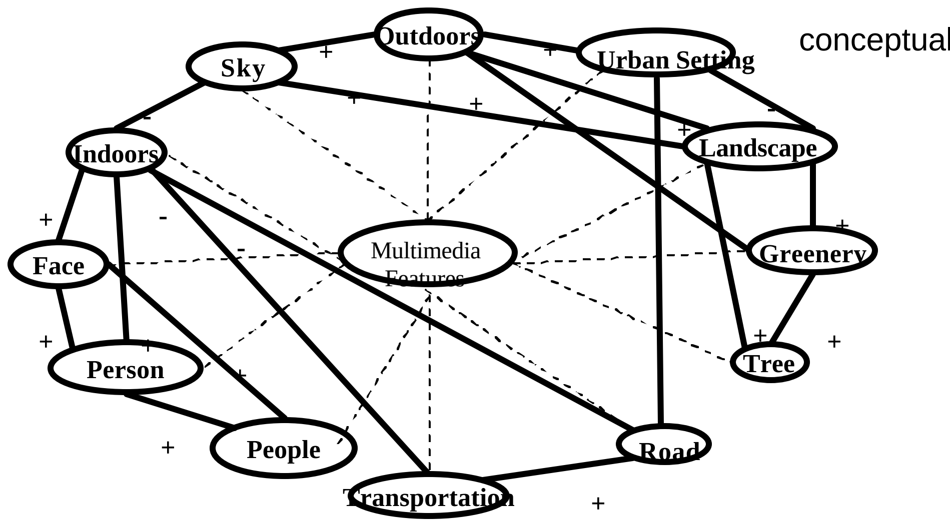
Semantic Context Learning and Exploitation: Multinet

Problem:
Building each concept model independently fails to utilize spatial, temporal and conceptual context and is sub-optimal use of available information.

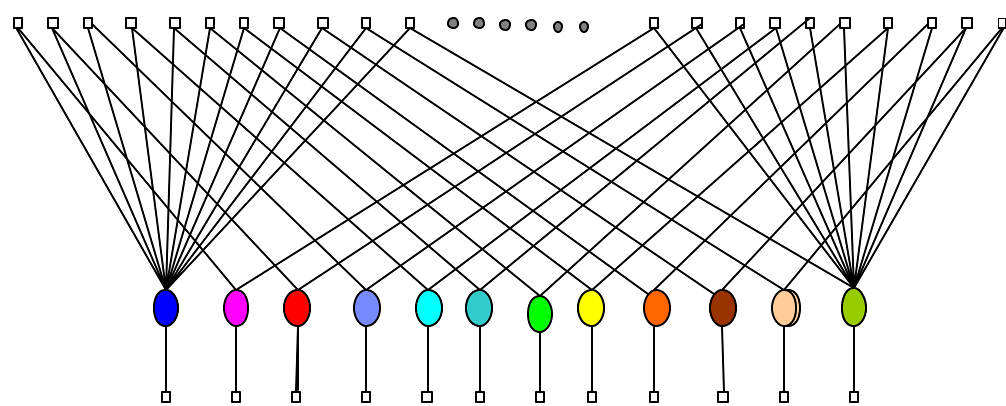
Approach: Multinet:
Network of Concept Models represented as a graph with undirected edges. Use of probabilistic graphical models to encode and enforce context.

Result:

- Factor-graph multinet with Markov chain temporal models improve mean average precision by more than **27% over best IBM Run for TREC 2002 and 36 % in conjunction with SVM-DMF,**
- **Highest MAP for TREC'03**
- Low training cost
- No extra training data needed
- High inference cost
- Fusion Validation Set II MAP: 0.268
- SearchTest MAP: 0.263
- References: Naphade et al (CIVR 2003, TCSVT 2002)



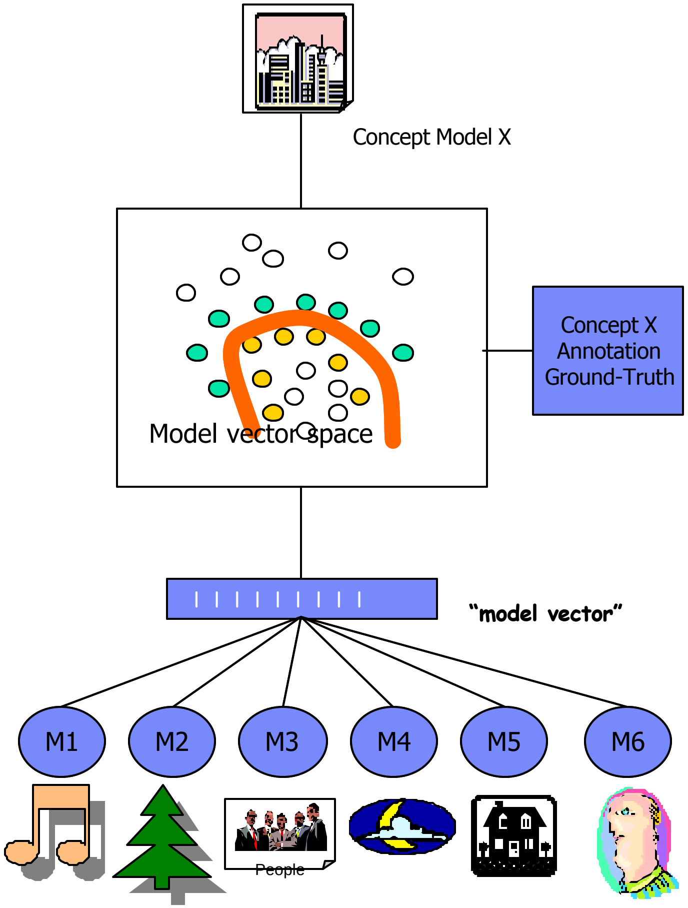
Factor Graph Loopy Propagation Implementation CIVR' 03



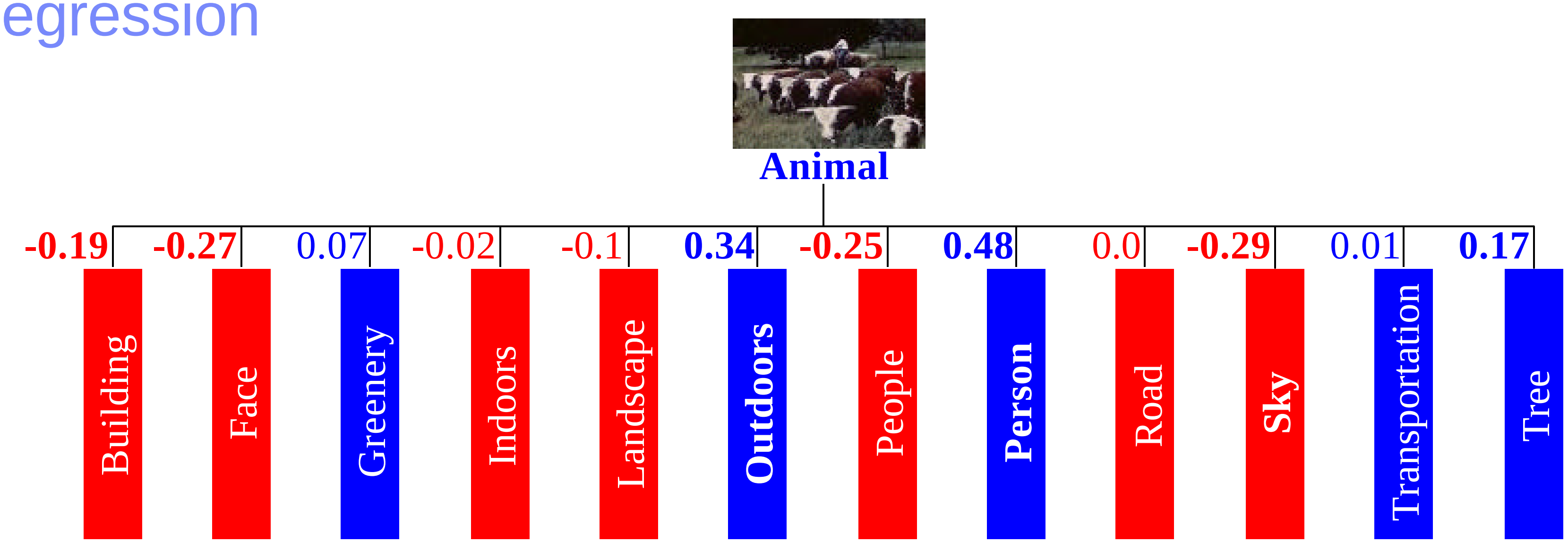
Multi-Modality/ Multi-Concept Fusion Methods: DMF using SVM

Using SVM/NN to re-classify the output results of Classifier 1-N.

- No normalization required.
- Use of Validation Set for training and Fusion Validation Set 1 for optimization and parameter selection.
- Training Cost low when number of classifiers being fused is small (i.e. few tens?)
- Classification cost low
- Used for fusing together multiple concepts in the semantic feature-space methods.
- Fusion Validation Set II MAP: 0.273
- SearchTest MAP: 0.247
- References: Iyengar et al (ICME 2002, ACM '03)



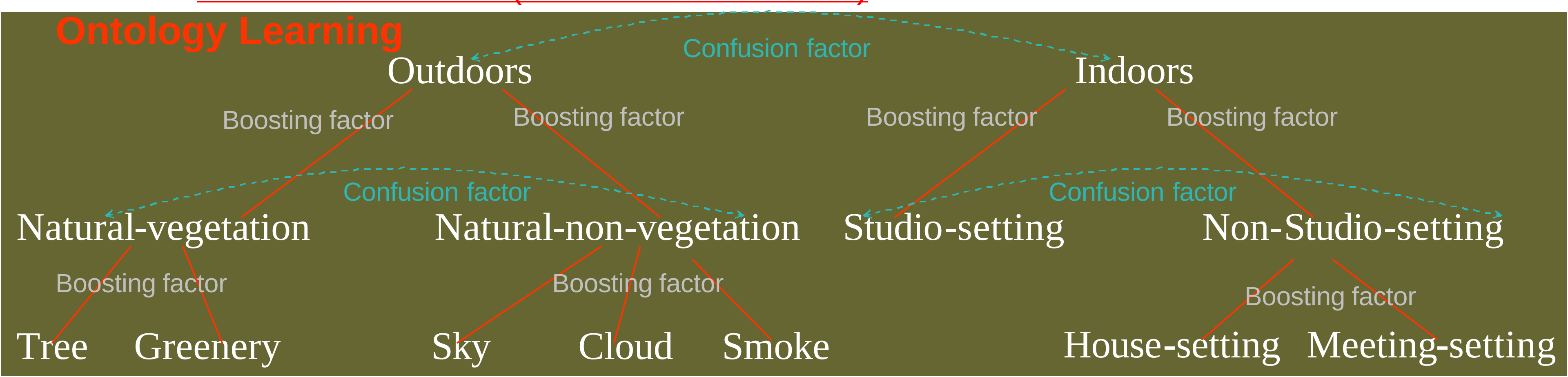
Multi-Concept Fusion: Semantic Space Modeling Through Regression



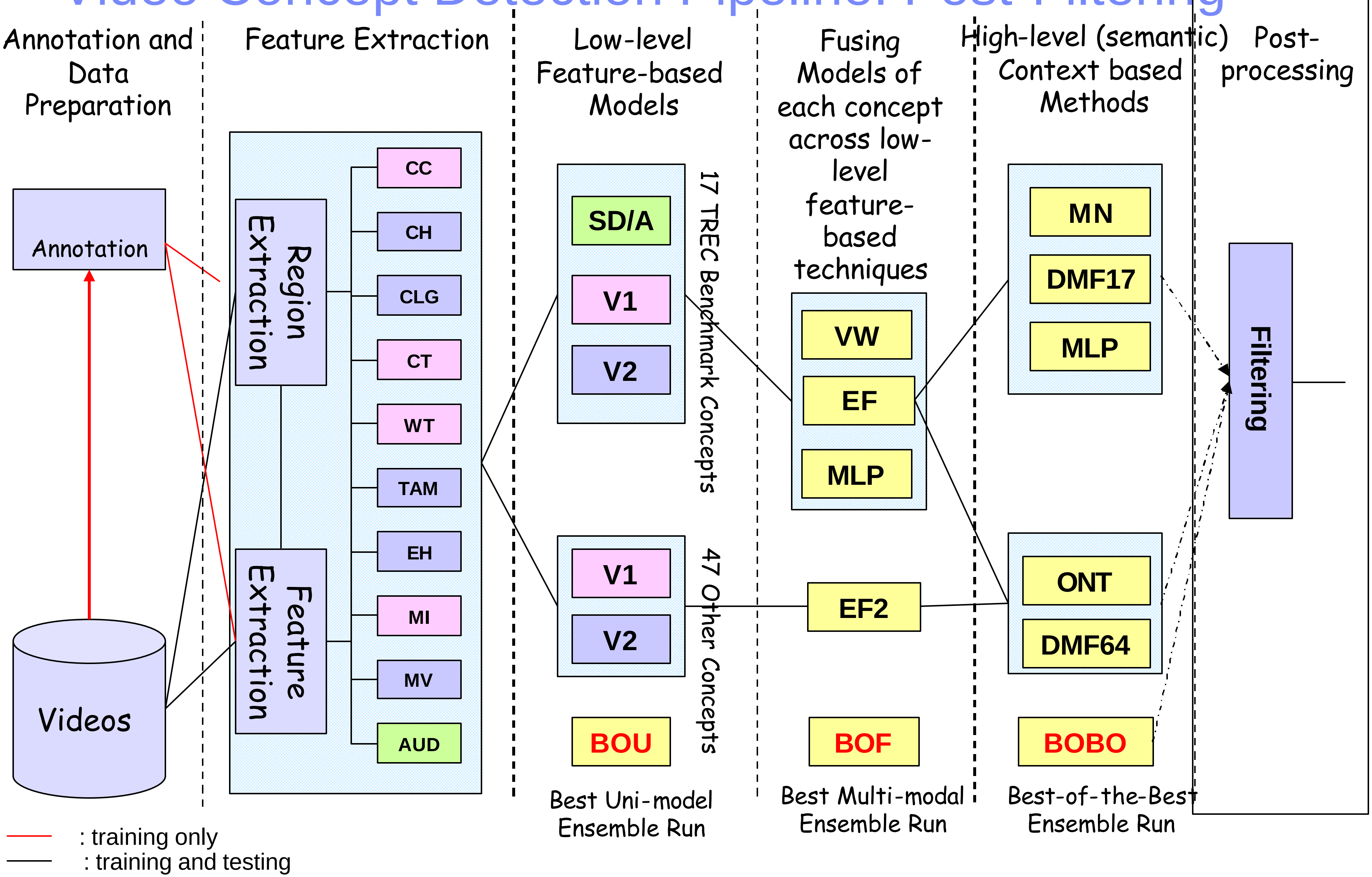
- **Problem:** Given a (small) set of related concept exemplars, learn concept representation
- **Approach:** Learn and exploit semantic correlations and class co-dependencies
  - Build (robust) classifiers for set of basis concepts (e.g., SVM models)
  - Model (rare) concepts in terms of known (frequent) concepts, or anchors
    - Represent images as semantic model vectors, or vectors of confidences w.r.t. known models
    - Model new concepts as sub-space in semantic model vector space
  - Learn weights of separating hyper-plane through regression:
    - Optimal linear regression (through Least Squares fit)
    - Non-linear MLP regression (through Multi-Layer Perceptron neural networks)
  - Can be used to boost performance of basis models or for building additional models
  - Fusion Validation Set II MAP: 0.274
  - SearchTest MAP: 0.252
  - References: Natsev et al (ICIP 2003)

Multi-Concept Fusion: Ontology-based Boosting

- **Basic Idea**
  - Concept hierarchy is created manually based on semantics ontology
  - Classifiers influence each other in this ontology structure
  - Try best to utilize information from reliable classifiers
- **Influence Within Ontology Structure**
  - Boosting factor : Boosting children precision from more reliable ancestors (Shrinkage theory: Parameter estimates in data-sparse children toward the estimates of the data-rich ancestors in ways that are provably optimal under appropriate condition)
  - Confusion factor: The probability of misclassifying C_j into C_i , and C_j and C_i cannot coexist
  - Fusion Validation Set II MAP: 0.266
  - SearchTest MAP: 0.261
  - References: Wu et al (ICME 2004 - submitted)



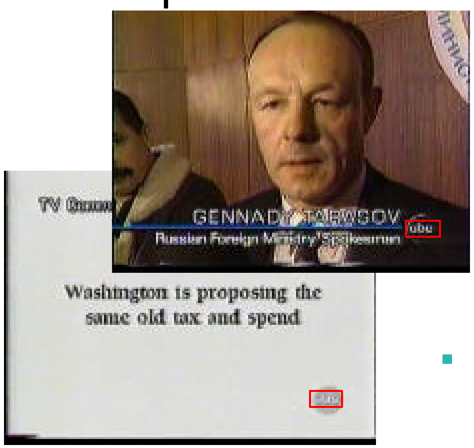
Video Concept Detection Pipeline: Post-Filtering

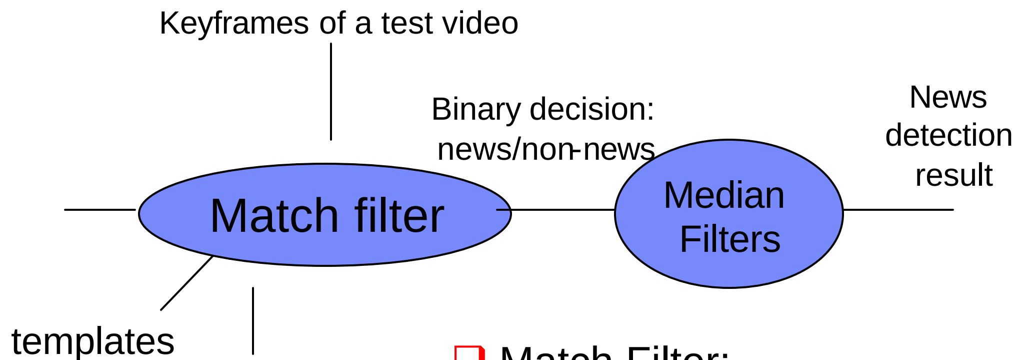


Post Filtering - News/Commercial Detector

CNN template:

ABC templates:





- Match Filter:

  For each template:

  $$S = d(S_C > t'_C) \& d(S_E > t'_E)$$

  where C:Color: E: Edge, and

  $$S_C = \frac{1}{N}\sum_n d(d(P_C, P_{MC}) > t_C)$$

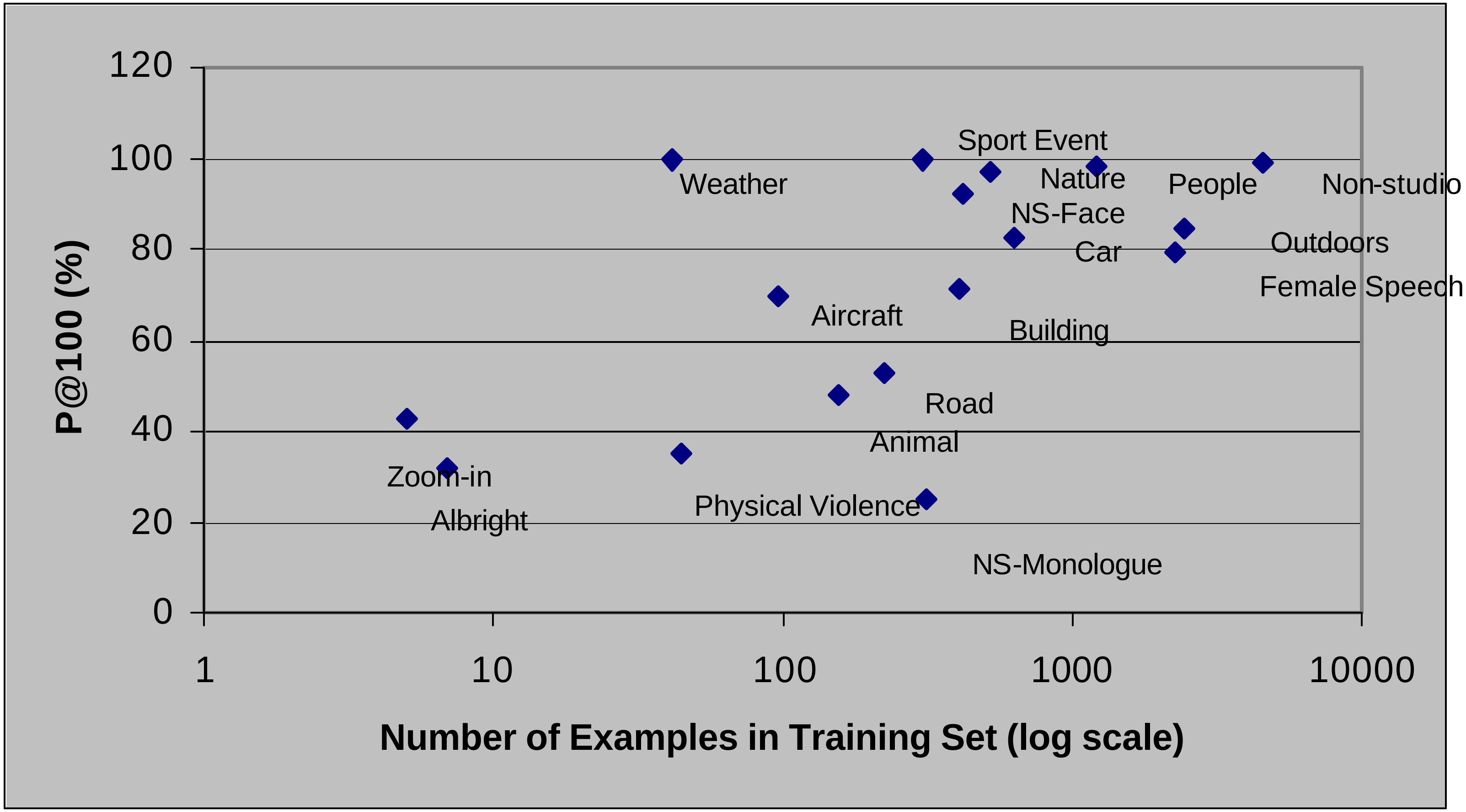
  $$S_E = \frac{1}{N}\sum_n d(d(P_E, P_{ME}) > t_E)$$

  - Thresholds: t_C, t_E, t'_C, t'_E were decided from two training videos. All templates use the same thresholds. Templates were arbitrarily chosen from 3 training videos.

- Performance: *Misclassification (Miss + False Alarm) in the Validation Set :*
  - CNN: 8 out of 1790 shots (accuracy = 99.6%)
  - ABC: 60 out of 2111 shots (accuracy=97.2%)

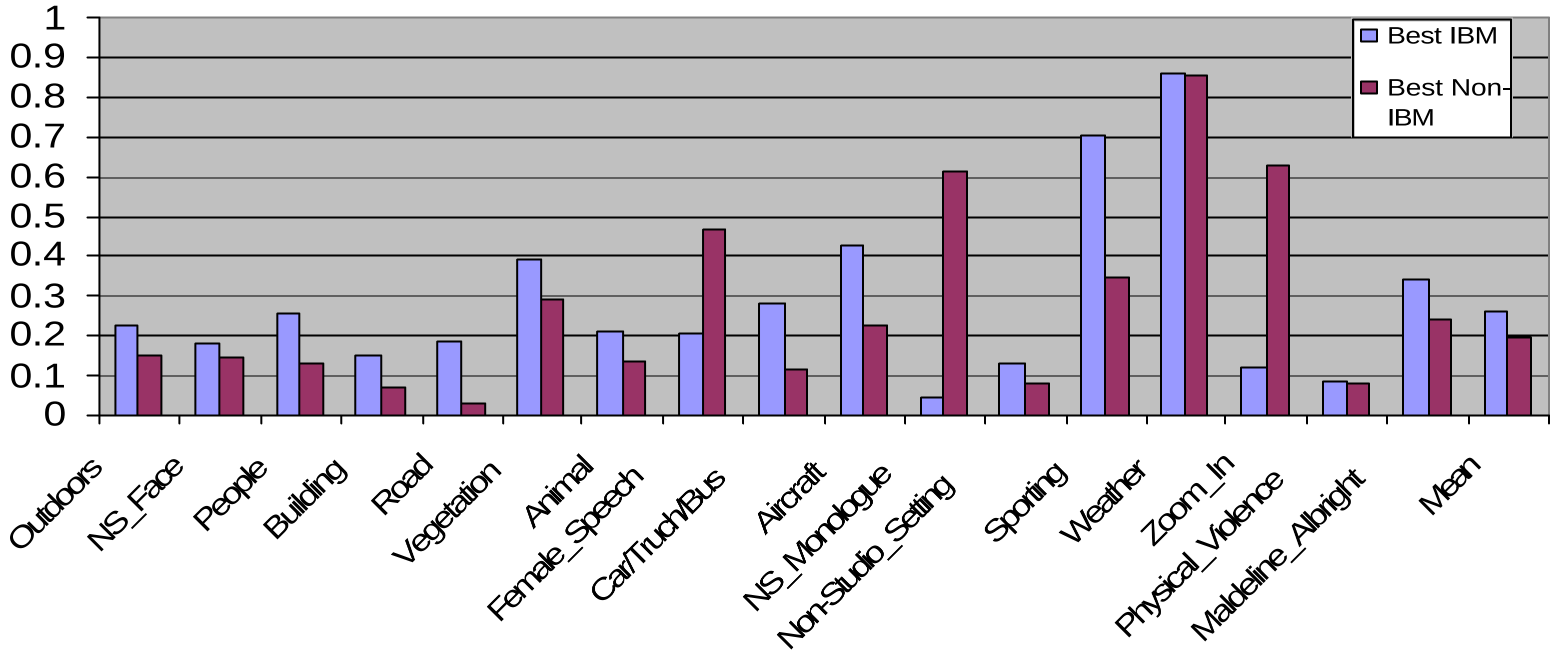
- Our definition of news: news program shots (non-commercial, non-miscellaneous shots)

P@100 vs. Number of examples



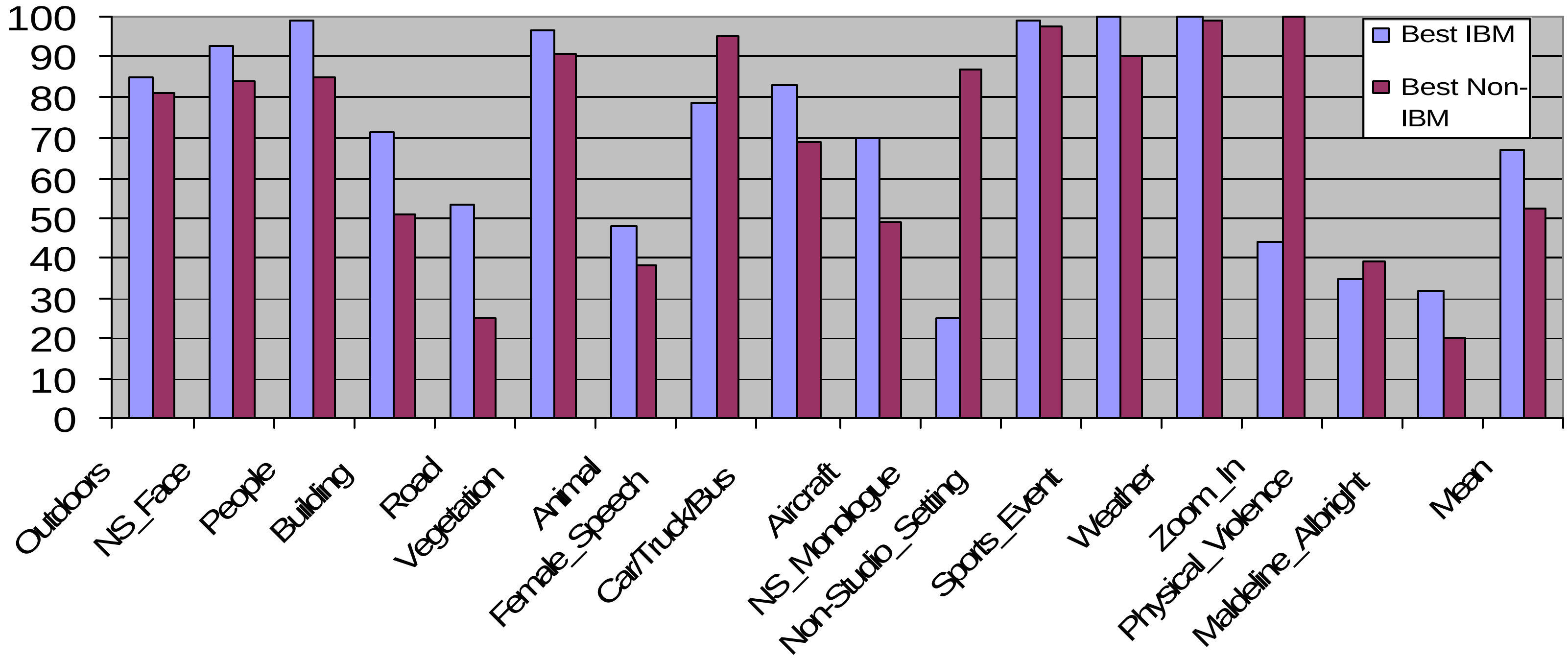
Performance is roughly log linear in terms of number of examples
Yet there are deviations
➔Can Log-linear be considered the default to evaluate concept complexity?

TRECVID 2003 – Average Precision Values



- IBM has the best Average Precision at 14 out of the 17 concepts
- The best Mean Average Precision of IBM system (0.263) is 34 percent better than the second best
- Pooling skews some AP numbers for high-frequency concepts so it makes judgement difficult but can be considered a loose lower bound on performance.
- Bug in Female_Speech model affected second level fusion of Female_Speech, News_Subject_Monologue, Madeleine_Albright among others. This was especially hurting the model-vector-based techniques (DMF, NN, Multinet, Ontology)

TRECVID 2003 -- Precision at Top 100 Returns



- IBM has the highest Precision @ 100 in 13 out of the 17 concepts
- Mean Precision @ 100 of Best IBM System 0.6671
- The best Mean Precision of IBM system is 28 percent better than the other systems.
- Different Model-vector based fusion techniques improve performance for different classes of concepts

Precision of 10 IBM Runs Submitted

	Outdoors	NSFace	People	Building	Road	Vege.	Animal	F_Spee	Vehicle	Aircra	Monol.	NonStudio	Sports	Weather	Zoom_In	Violence	Albright	Mean
BOU	81	80	90	53	46	96	10	46	68	38	24	97	81	79	44	33	32	58.706
EF	67	77	95	60	33	97	47	69	80	63	25	96	99	98	44	28	28	65.059
BOF	71	77	97	71	52	93	47	69	80	47	25	96	98	100	44	35	32	66.706
DMF17	82	93	90	54	49	97	45	35	76	70	1	99	98	99	44	9	28	62.882
DMF64	82	73	79	53	41	96	33	79	56	67	0	93	98	99	44	34	4	60.647
MLP_BOR	78	75	97	61	53	94	47	38	70	65	1	95	100	97	44	27	30	63.059
MLP_EFC	73	67	97	41	33	96	48	19	49	60	3	97	99	99	44	27	27	57.588
MN	85	55	99	52	45	97	47	66	81	63	25	96	99	98	44	22	28	64.824
ONT	67	77	95	56	42	97	47	69	83	69	6	94	99	98	44	28	28	64.647
BOBO	85	73	99	56	52	93	10	66	56	63	0	97	98	99	44	22	32	61.471
Maximum:	85	93	99	71	53	97	48	79	83	70	25	99	100	100	44	35	32	66.706
Average:	76.857	73.857	93.429	55.429	45	95.71	44.857	53.571	70.714	63	8.714	95.71429	98.71	98.5714	44	26	25.286	62.908

- Processing beyond single classifier per concept improves performance
- If we divide TREC Benchmark concepts into 3 types based on frequency of occurrence
  - Performance of Highly Frequent (>80/100) concepts is further enhanced by Multinet (e.g. Outdoors, Nature_Vegetation, People etc.)
  - Performance of Moderately Frequent concepts (>50 & < 80) is usually improved by discriminant reclassification techniques such as SVMs (DMF17/64) or NN (MLP_BOR, MLP_EFC)
  - Performance of very rare concepts needs to be boosted through better feature extraction and processing in the initial stages.
- Based on Fusion Validation Set 2 evaluation, visual models outperform audio/ASR models for 9 concepts while the reverse is true for 6 concepts.
- Semantic-feature based techniques improve MAP by 20 % over visual-models alone.
- Fusion of multiple modalities (audio, visual) improves MAP by 20 % over best unimodal (visual) run (using Fusion Validation Set II for comparison)

Observations and Future Directions

- Generic Trainable Methods for Concept Detection demonstrate impressive performance.
- Need to increase Vocabulary of Concepts Modeled
- Need to improve Modeling of Rare Concepts
- Need Multimodality at an earlier level of analysis (e.g. multimodal model of Monologue (TREC'02) better than fusion of multiple unimodal classifiers (TREC'03)
- Multi-classifier, Multi-concept and Multi-modal fusion offer promising improvement in detection (as measured on TREC'02 and TREC'03 Fusion Validation Set 2 and in part also by TREC SearchTest 03)

Acknowledgements

- Thanks for additional contributions from:
  - Chitra Dorai (IBM) for Zoom-In Detector,
  - Javier Ruiz-del-Solar (Univ. of Chile) for Face Detector,
  - Ishan Sachedv (summer intern – MIT) for helping with Visual uni-models,
  - For collaborative annotation:
    - IBM -- Ying Li, Christrian Lang, Ishan Sachedv, Larry Sansone, Matthew Hill,
    - Columbia U. -- Winston Hsu
    - Univ. of Chile – Alex Jaimes, Dinko Yaksic, Rodrigo Verschae

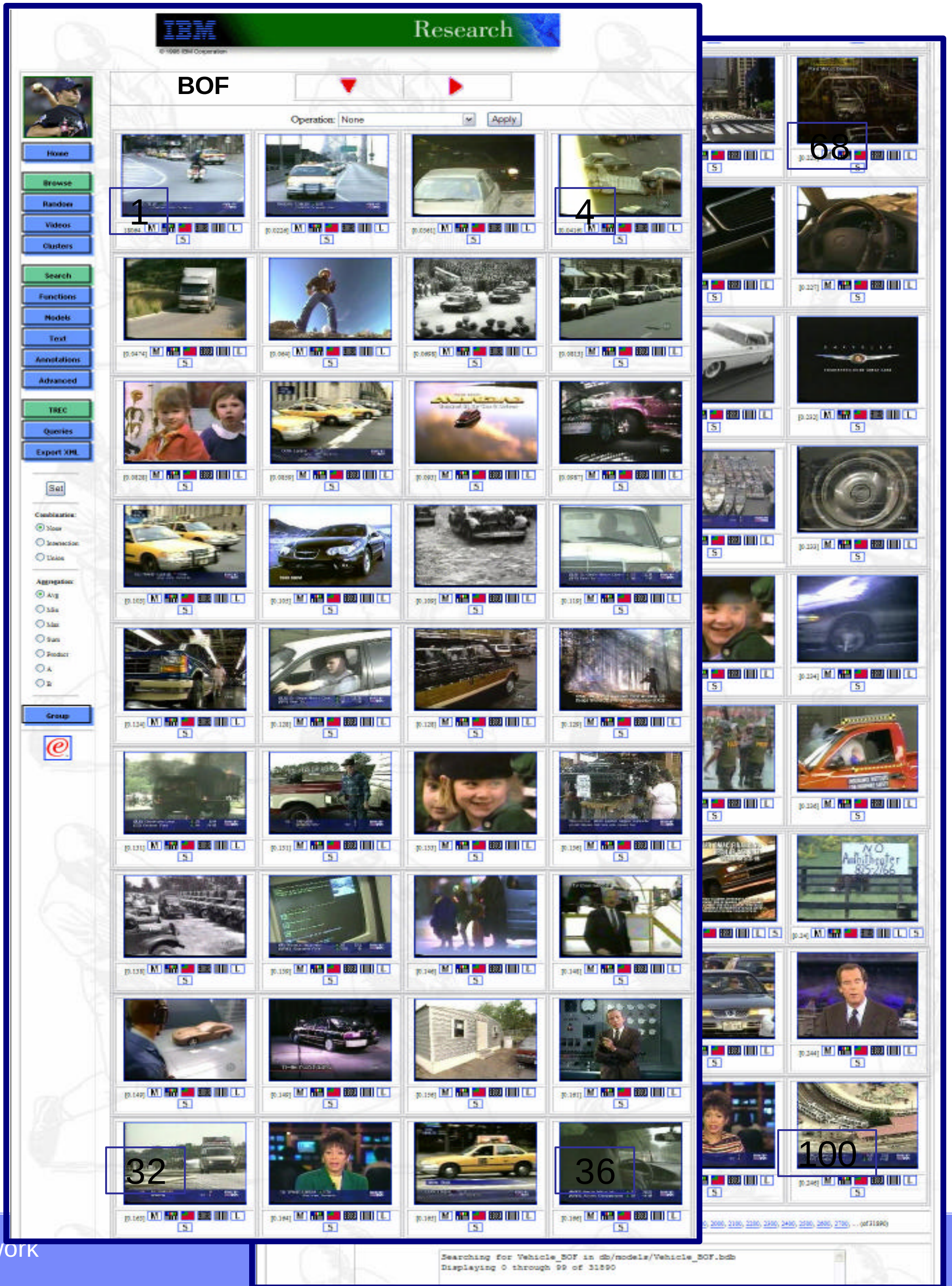
Concept Detection Example: Cars

- ***"Car/truck/bus**: segment contains at least one automobile, truck, or bus exterior"*

- Concept was trained on the annotated training set.

- Results are shown on the test set

Run	Precision @100
Best IBM	0.83



Concept Detection Example: Ms. Albright

- ***"Person X**: segment contains video of person x (x = Madeleine Albright)."*

- Contributions of the Audio-based Models and Visual-based Models
  -- Results at the CF2 (validation set)

Run	Average Precision
Best IBM Audio Models	0.30
Best IBM Visual Models	0.29
Best of Fusion	0.47

- Results are shown on the test set TREC Evaluation by NIST

Run	Precision
Best IBM	0.32