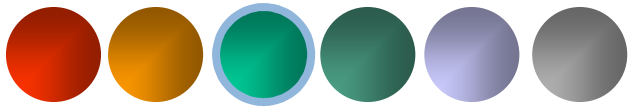
IBM Research – Columbia University

IBM-Columbia TRECVID MED-2011 Experiments

Liangliang Cao, Noel Codella, Leiguang Gong, Matthew Hill, Gang Hua, Apostol Natsev, John R. Smith (IBM T. J. Watson Research Center)

Shih-Fu Chang, Courtenay Cotton, Dan Ellis, John Kender, Michele Merler, Yadong Mu (Columbia University)

Contact: jrsmith@watson.ibm.com
IBM T. J. Watson Research Center
December 5, 2011



Outline

- Partitioning of team and data
- IBM-Columbia TRECVID MED-11 system overview
- Video Feature Extraction:
  - Audio-Visual Features (low-level)
  - Discriminative Semantic Features (high-level)
- Video Event Modeling:
  - Multi-modal fusion, score calibration and thresholding
- Experimental Results:
  - Data partitioning and evaluation setup
  - Run 1: Low-level signal features
  - Run 2: High-level semantic features
  - Run 3: AP-based WAVG fusion of Run 1 and 2
  - Run 4: Linear SVM-based fusion of 14 component runs
- Lessons learned and future directions

IBM-Columbia NIST TRECVID MED-11 Team

IBM Research

Liangliang Cao

Noel Codella

Leiguang Gong

Gang Hua

Matthew Hill

Paul Natsev

John R. Smith

Columbia University

Shih-Fu Chang

Courtenay Cotton

Daniel Ellis

John Kender

Michele Merler

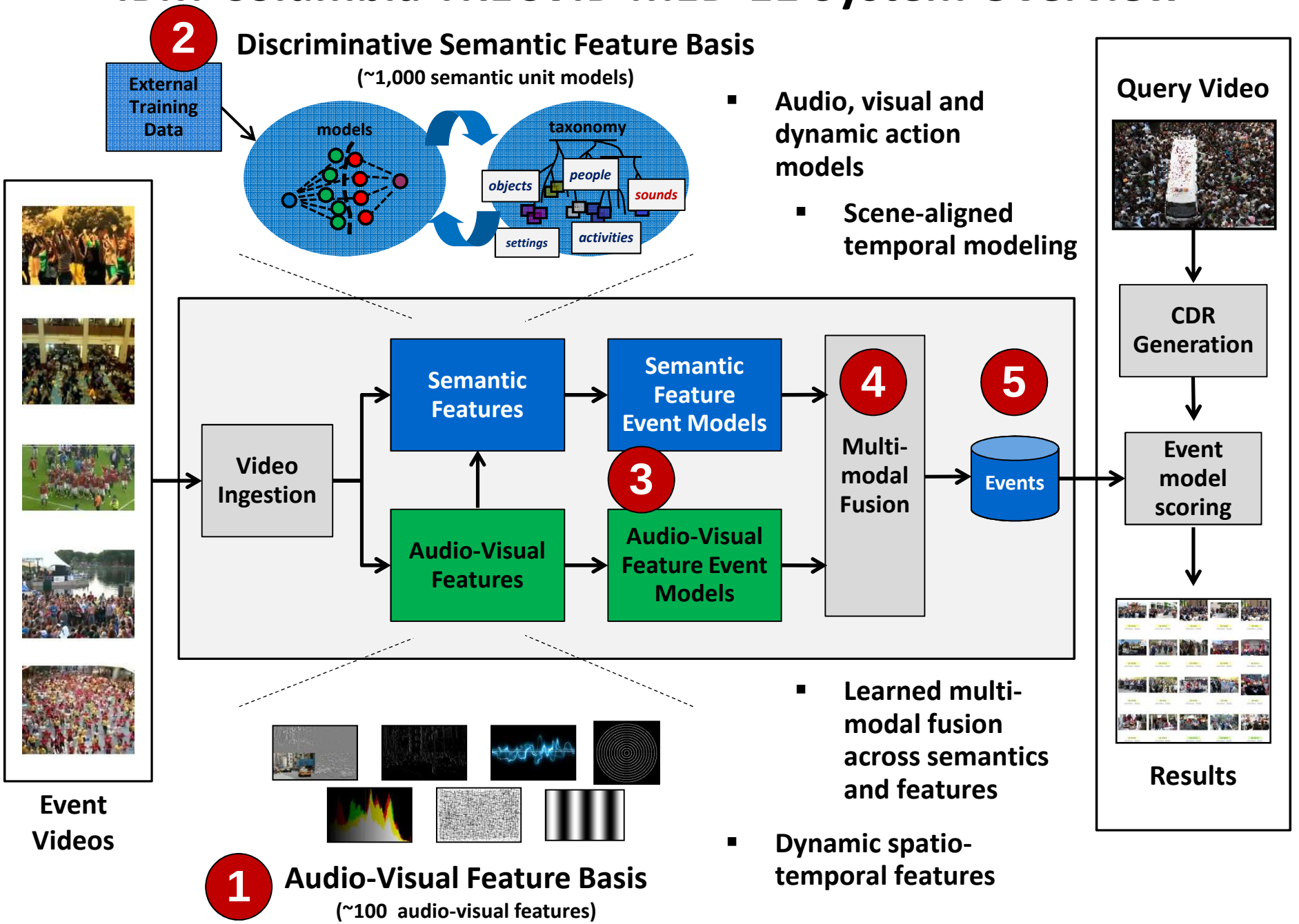
Yadong Mu

TRECVID MED-11 Data Partitioning and Evaluation Setup

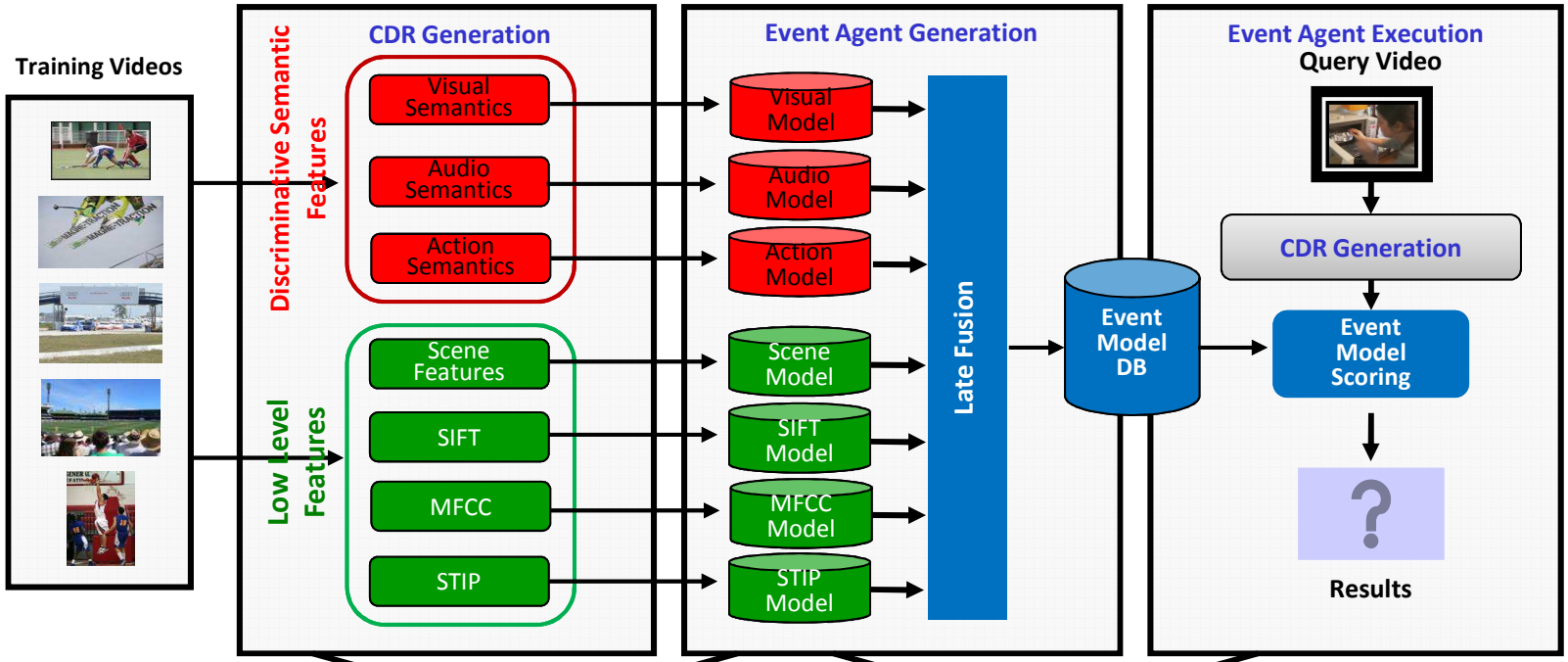
- **Event kits and DevT videos were partitioned in two internal datasets:**
  - **IBM-TEST:** 40 positive examples per event + half of random videos
  - **IBM-TRAIN:** remaining videos from **DevT** and **Event Kits**
- **Training and evaluation setup**
  - All runs trained on **TRAIN** set only (parameters selected using cross-validation)
  - All runs evaluated and fused on **TEST** set
  - Intermediate evaluations based on Average Precision (AP) and Mean AP
  - Final runs thresholded based on target performance metrics (Pmiss, Pfa, NDC)

Dataset	# positive videos	# negative videos	# videos	# keyframes
IBM-TRAIN	2,036	5,216	7,252	547,357
IBM-TEST	600	5,231	5,831	437,251
MED11-TEST	N/A		32,061	1,791,263

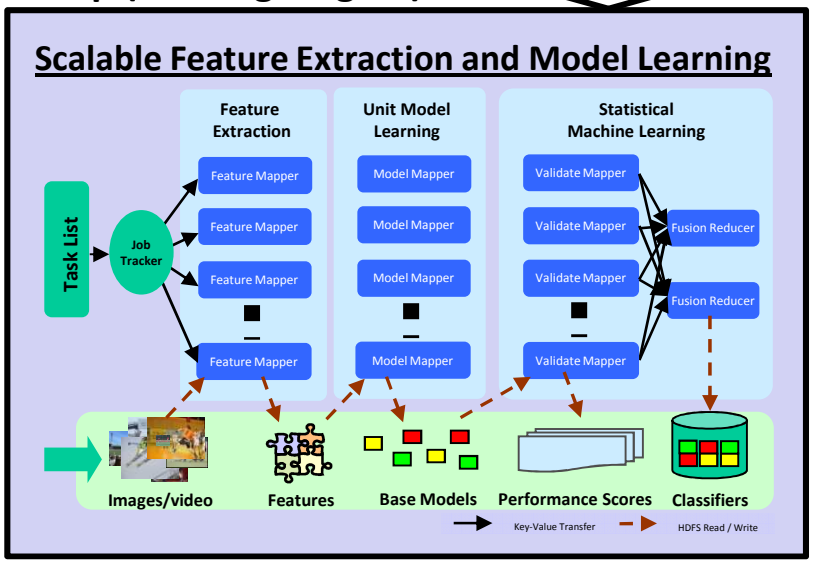
IBM-Columbia TRECVID MED-11 System Overview



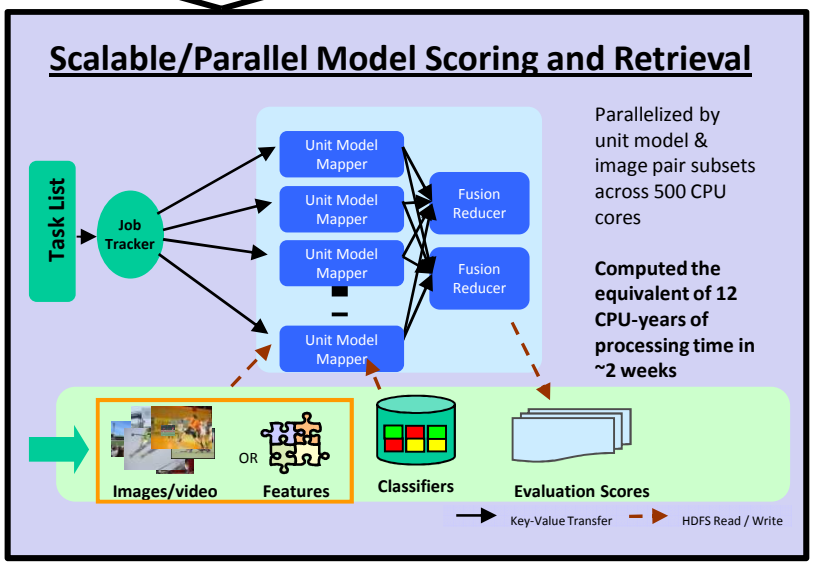
IBM-Columbia: Physical architecture and scalable/parallel approach to computation



- Hadoop (IBM BigInsights)



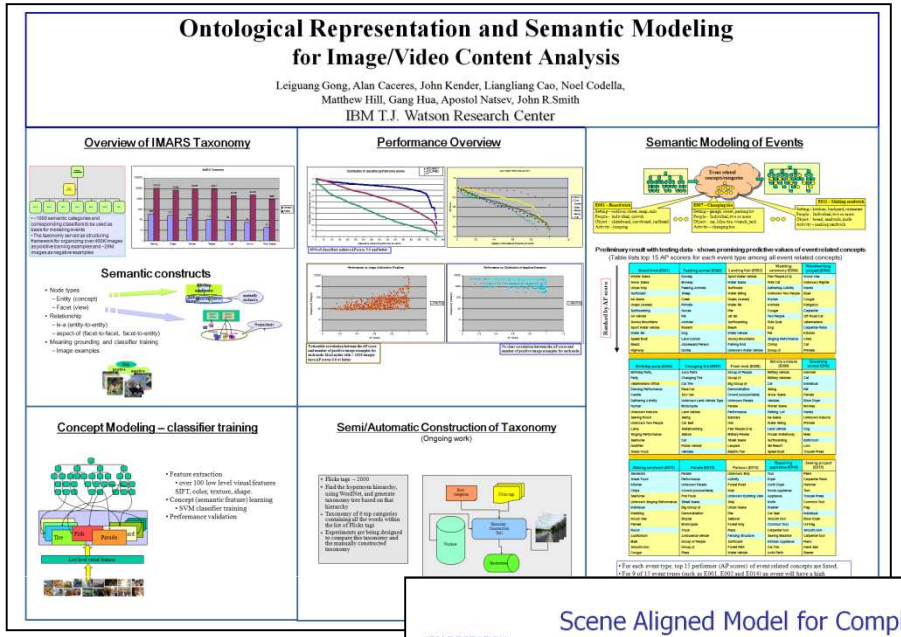
- IBM Streams



Parallelized by unit model & image pair subsets across 500 CPU cores

Computed the equivalent of 12 CPU-years of processing time in ~2 weeks

TRECVID MED-11 Posters – **more details provided*



Ontological Representation and Semantic Modeling for Image/Video Content Analysis

Leiguang Gong, Alan Caceres, John Kender, Liangliang Cao, Noel Codella, Matthew Hill, Gang Hua, Apostol Natsev, John R.Smith
IBM T.J. Watson Research Center

Overview of IMARS Taxonomy

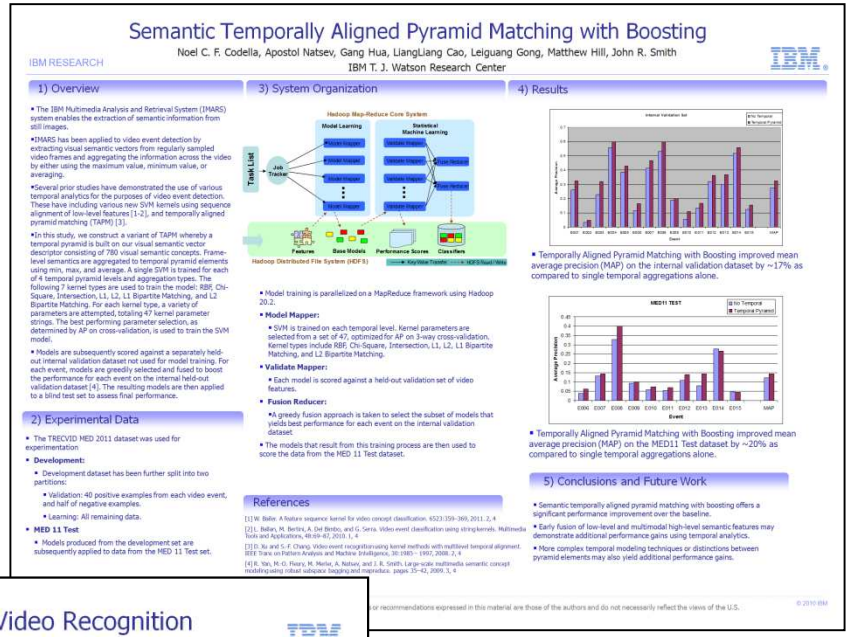
Performance Overview

Semantic Modeling of Events

Semantic constructs

Concept Modeling – classifier training

Semi/Automatic Construction of Taxonomy
(Ongoing work)



Semantic Temporally Aligned Pyramid Matching with Boosting

Noel C. F. Codella, Apostol Natsev, Gang Hua, LiangLiang Cao, Leiguang Gong, Matthew Hill, John R. Smith
IBM T. J. Watson Research Center

IBM RESEARCH

1) Overview

- The IBM Multimedia Analysis and Retrieval System (IMARS) system enables the extraction of semantic information from still images.
- IMARS has been applied to video event detection by extracting visual semantic vectors from regularly sampled video frames and aggregating the information across the video by either using the maximum value, minimum value, or averaging.
- Several prior studies have demonstrated the use of various temporal analytics for the purposes of video event detection. These have including various new SVM kernels using sequence alignment of low-level features [1-2], and temporally aligned pyramid matching (TAPM) [3].
- In this study, we construct a variant of TAPM whereby a temporal pyramid is built on our visual semantic vector descriptor consisting of 780 visual semantic concepts. Frame-level semantics are aggregated to temporal pyramid elements using min, max, and average. A single SVM is trained for each of 4 temporal pyramid levels and aggregation types. The following 7 kernel types are used to train the model: RBF, Chi-Square, Intersection, L1, L2, L1 Bipartite Matching, and L2 Bipartite Matching. For each kernel type, a variety of parameters are attempted, totaling 47 kernel parameter strings. The best performing parameter selection, as determined by AP on cross-validation, is used to train the SVM model.
- Models are subsequently scored against a separately held-out internal validation dataset not used for model training. For each event, models are greedily selected and fused to boost the performance for each event on the internal held-out validation dataset [4]. The resulting models are then applied to a blind test set to assess final performance.

2) Experimental Data

- The TRECVID MED 2011 dataset was used for experimentation
- **Development:**
  - Development dataset has been further split into two partitions:
    - Validation: 40 positive examples from each video event, and half of negative examples.
    - Learning: All remaining data.
- **MED 11 Test**
  - Models produced from the development set are subsequently applied to data from the MED 11 Test set.

3) System Organization

- Model training is parallelized on a MapReduce framework using Hadoop 20.2.
- **Model Mapper:**
  - SVM is trained on each temporal level. Kernel parameters are selected from a set of 47, optimized for AP on 3-way cross-validation. Kernel types include RBF, Chi-Square, Intersection, L1, L2, L1 Bipartite Matching, and L2 Bipartite Matching.
- **Validate Mapper:**
  - Each model is scored against a held-out validation set of video features.
- **Fusion Reducer:**
  - A greedy fusion approach is taken to select the subset of models that yields best performance for each event on the internal validation dataset
- The models that result from this training process are then used to score the data from the MED 11 Test dataset.

References

[1] W. Bailer. A feature sequence kernel for video concept classification. 6523:359–369, 2011. 2, 4
[2] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Video event classification using string kernels. Multimedia Tools and Applications, 48:69–87, 2010. 1, 4
[3] D. Xu and S.-F. Chang. Video event recognition using kernel methods with multilevel temporal alignment. IEEE Trans on Pattern Analysis and Machine Intelligence, 30:1985 – 1997, 2008. 2, 4
[4] R. Yan, M.-O. Fleury, M. Merler, A. Natsev, and J. R. Smith. Large-scale multimedia semantic concept modeling using robust subspace bagging and mapreduce. pages 35–42, 2009. 3, 4

4) Results

- Temporally Aligned Pyramid Matching with Boosting improved mean average precision (MAP) on the internal validation dataset by ~17% as compared to single temporal aggregations alone.

- Temporally Aligned Pyramid Matching with Boosting improved mean average precision (MAP) on the MED11 Test dataset by ~20% as compared to single temporal aggregations alone.

5) Conclusions and Future Work

- Semantic temporally aligned pyramid matching with boosting offers a significant performance improvement over the baseline.
- Early fusion of low-level and multimodal high-level semantic features may demonstrate additional performance gains using temporal analytics.
- More complex temporal modeling techniques or distinctions between pyramid elements may also yield additional performance gains.

Scene Aligned Model for Complex Video Recognition

Liangliang Cao, Yadong Mu, Shih-Fu Chang, Gang Hua, John R. Smith -- IBM Research & Columbia

IBM RESEARCH

Overview

Real-world videos often contain dynamic backgrounds and evolving people activities, especially for those web videos generated by users in unconstrained scenarios.

We propose a new visual representation, namely scene aware concurrent pooling, for the task of event recognition in complex videos. Based on the observation that a video clip is often composed with shots of different scenes, the key idea of scene aware concurrent pooling is to decompose any video features into concurrent scene components, and to construct classification models adaptive to different scenes.

Key Results:

- Consistently improve various features including color histogram, texture, linear binary pattern, SIFT, and semantic model vectors in TRECVID.
- Improve the state-of-the-art on HMDB action dataset by 20%.

Motivation

We propose to use scene information to guide the pooling operation of video features.

- Traditional pooling: average the feature vectors within a spatial neighborhood of images.
- Our method: aggregate video features into concurrent scene components, and then develop scene-dependent modules for the classification task.

Our new model is called *Scene Aware Concurrent Pooling* (SACP).

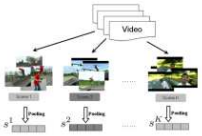


Illustration of Scene Aware Concurrent Pooling

Related Work

Scene recognition:

- UIUC 15 scenes
- MIT SUN dataset
- Scene based recognition (Stanford, MIT, Kodak)

Feature pooling:

- Average pooling
- Max pooling + sparse coding
- Spatial pyramid matching (SPM)

Table 1. Comparing SACP with Spatial Pyramid Matching.

	Subject	Domain	Grid	Assignment
SACP	video	scene	×	soft
SPM [14]	image	spatial	✓	hard

We prefer scene-aware pooling over temporal pooling:

- A video can be of different length and connecting multiple sub-events (five seconds ~ several minutes)
- There is no regularity in the event distribution along temporal axis. A video event can either happen at the beginning part of the video or the second half of the video, or it can last throughout the whole sequence.

Our Model

we choose an unsupervised way to model scenes using K-means on GIST features.

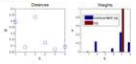
Each frame $\mathbf{x}_t$ is mapped into different scene components

$$\mathbf{x}_t \rightarrow [w_t^1\mathbf{x}_t, w_t^2\mathbf{x}_t, \cdots, w_t^K\mathbf{x}_t]$$

The pooling weights are computed by harmonic concurrent vector quantization

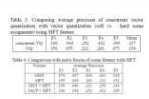
$$w_t^k = \frac{1/(d_t^k)^3}{\sum_l 1/(d_t^l)^3}$$

Unlike naive VQ, our concurrent VQ is a soft weighting strategy instead of hard weighting.

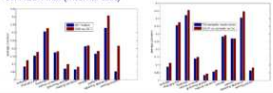


Experiments

On MED Dryrun:



On MED Final (internal test)



On Human Motion DataBase (HMDB):

Comparisons with the results on HMDB

Model	Accuracy
STIP histogram	21.96%
C2	23.18%
SACP + STIP	27.84%

Conclusions and Future Work

SACP can consistently improve different features (color histogram, SIFT, semantic model vectors)for complex video classification.

On HMDB dataset, our model significantly improve the best performance from 23.18% to 27.84%, as a relative 20% increase of accuracy.

Our future work will focus on generalizing the classification model to annotation tasks. Since SACP characterizes the scene structurewithin the video, we expect it can provide both event and scene annotations for a long clip.

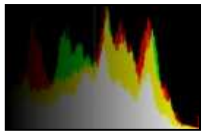
This material is based upon work funded in part by the U. S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

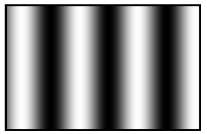
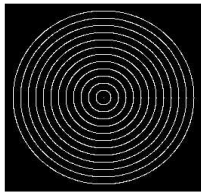
Audio-Visual Feature Extraction

1

Audio-Video Feature Extraction

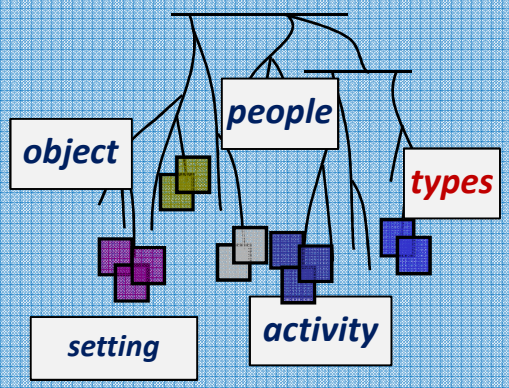
- **Global visual descriptors (13 features x8 granularities):**
  - 166-dim HSV color histogram and HSV color correlogram
  - 225-dim CIE-Lab 5x5 color moments
  - 108-dim Haar wavelet texture
  - 64 dim Sobel filter edge histogram
  - *: and more*
- **Local visual descriptors:**
  - SIFT based on Harris Laplace interest points (1K codebook)
  - SIFT based on DoG and Hessian detectors (5K codebook)
  - 512-dim GIST structure (4x4 grid, 8 orientations, 4 scales)
- **Spatio-temporal visual descriptors:**
  - STIP based on Harris interest points in 3D (space plus time)
- **Audio descriptors:**
  - 1,890-dim MFCC statistics (20-dim, 32 ms window, 16 ms hop)
  - Transient sound events (short duration energy contractions)
  - Perceptually-salient sound textures




2 Discriminative Semantic Features – 946 Semantic Concepts (total)

IBM Visual Taxonomy (core) – 380 Concepts



Facets:

- *Setting*
- *Domain*
- *Object*
- *People*
- *Activity*
- *Type*
- *Color*

IBM Visual Taxonomy (extended) – 400 Concepts

- More training examples
- More categories across facets, e.g.,:
  - *sports*
  - *settings*
- New event-related, e.g.,:
  - *fishing gear*
  - *toolbox*,
  - etc.

Dynamic Visual Actions – 113 Concepts

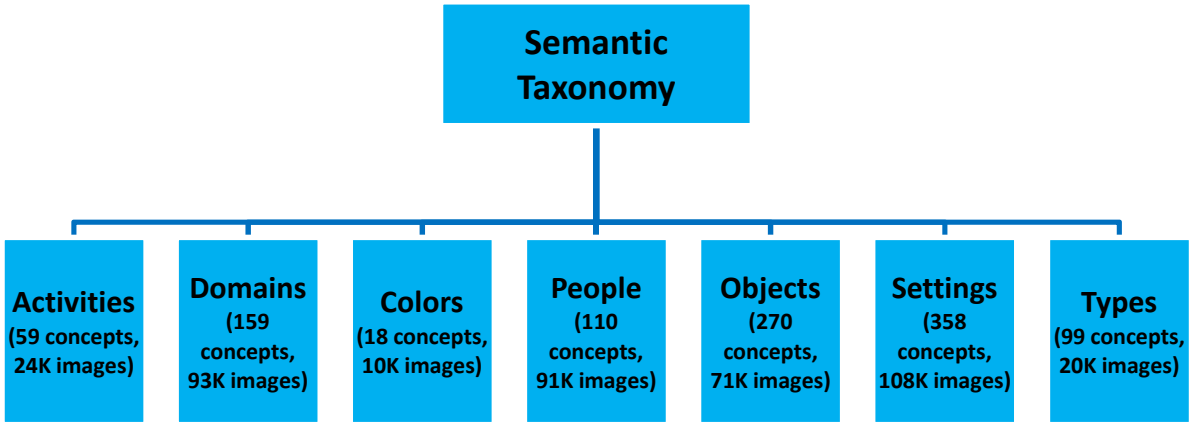
- **UCF50**: 6,681 videos obtained from YouTube and personal videos mainly related to sports (**50 actions**)
- **HMDB**: 6,766 videos from Internet, mostly focusing on human movements (e.g., *kiss, hug, sit up, drink*) (**51 actions**)
- **Hollywood2**: 1,707 videos clips from movies (**12 actions**)

Audio Semantic Models – 55 Concepts

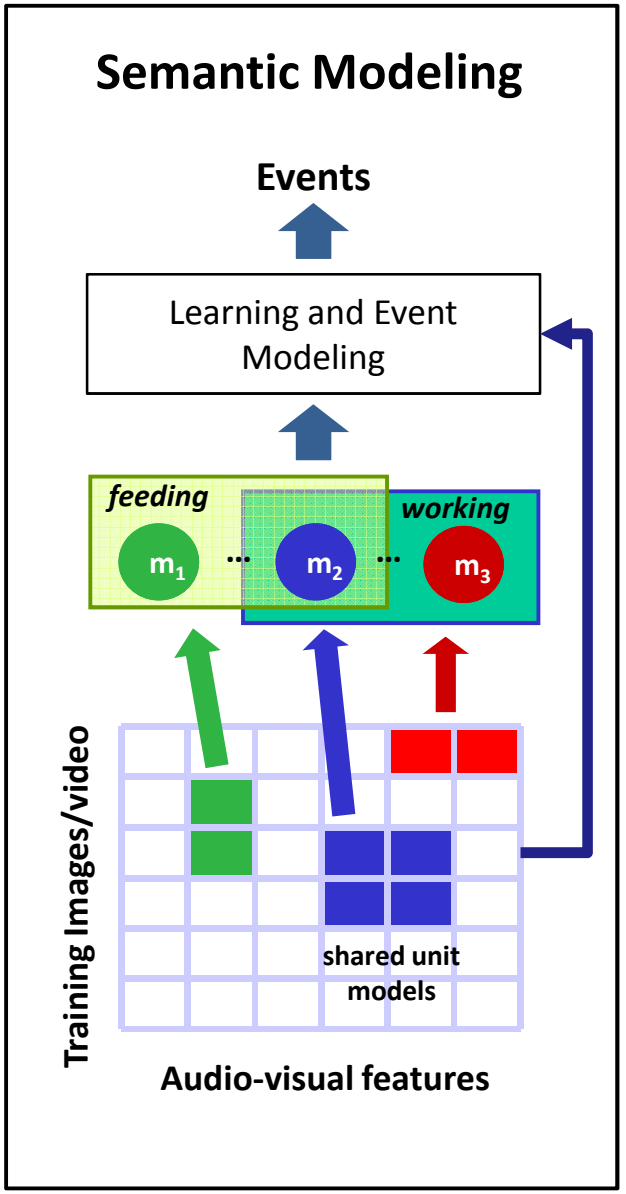
Dataset	# videos	# classes	MAP
YouTube 1873 [Lee & Ellis 2010]	1873	25 (sport, animal, night, beach …)	0.40
Columbia Consumer Video [Jiang et al. 2011]	9413	20 (soccer, cat, birthday, beach …)	0.30
MED2010 [Jiang et al. 2010]	6626 10 sec segments	10 (rural, urban, speech, clap…)	0.48

2 IBM Visual Semantic Modeling – **more details in TRECVID poster*

IBM Visual Semantic Taxonomy (780 concepts)

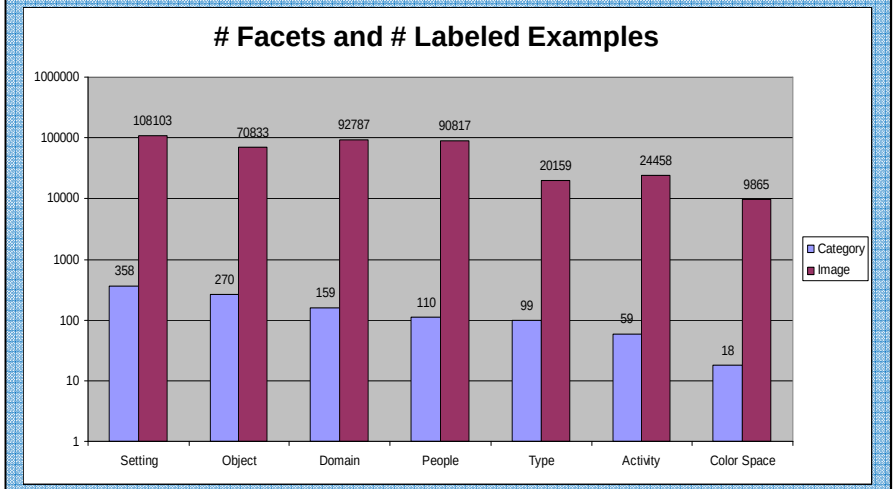


- Novel hierarchical faceted classification scheme supports modeling of semantic concepts and reasoning over labeled data to train classifiers
- Incorporates knowledge from labeled data to create discriminative semantic models that help recognize events
- Complements low-level audio-visual features
- Semantic taxonomy can be extended and adapted (direction, size, shape) as required for specific problem domain

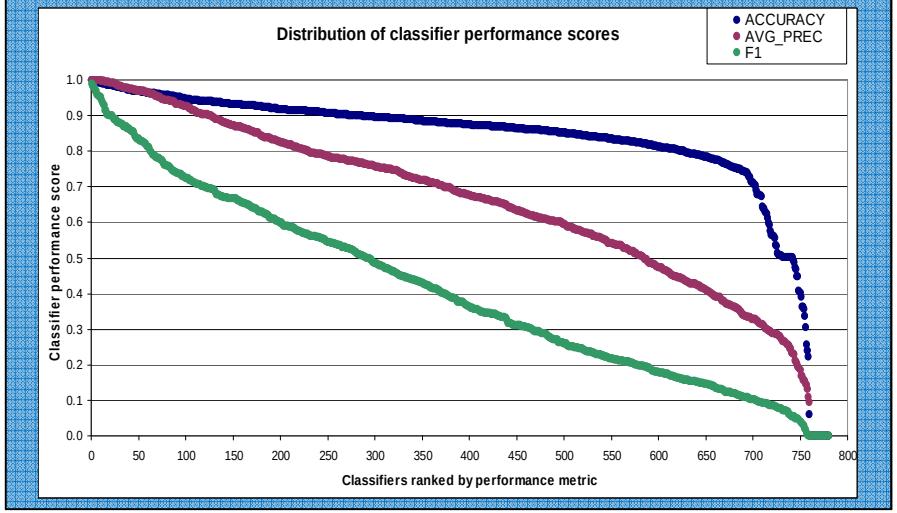


2 IBM Visual Taxonomy (780 concepts) – *Size, Shape and Performance*

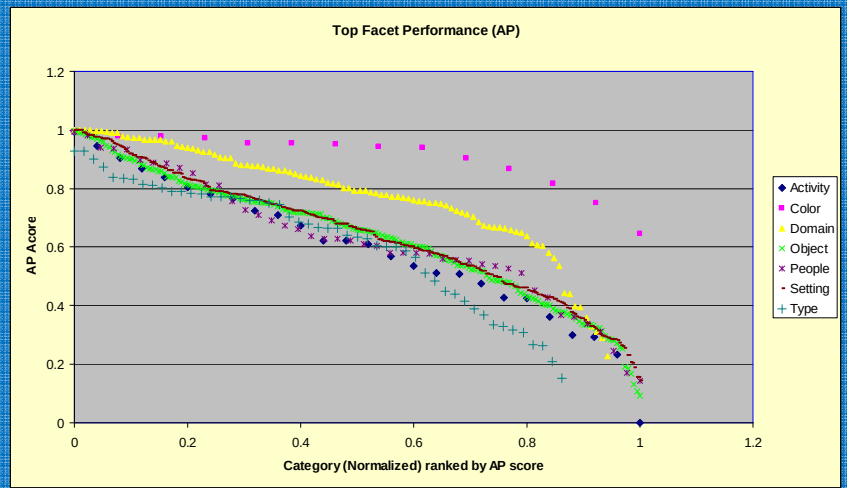
Taxonomy size



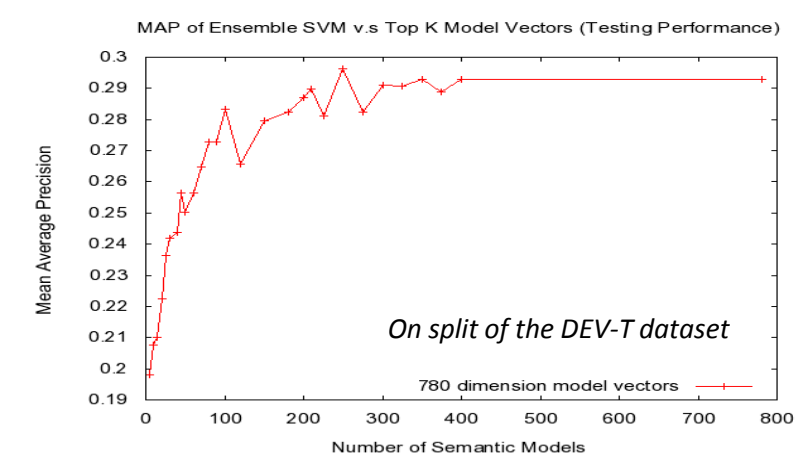
Classification Performance



Classification Performance by Facet



Event Prediction Performance



2

Dynamic Action Semantic Models (113 concepts)

Action Models : 1 vs. All RBF SVM Training

UCF50
50 action categories

Action Category	Train AP	Action Category	Train AP	Action Category	Train AP
BaseballPitch	0.726	JumpRope	0.970	RockClimbingIndoor	0.510
Basketball	0.471	JumpingJack	0.897	RopeClimbing	0.549
BenchPress	0.892	Kayaking	0.679	Rowing	0.420
Biking	0.565	Lunges	0.603	SalsaSpin	0.865
Billiards	1.000	MilitaryParade	0.845	SkateBoarding	0.534
BreastStroke	0.381	Mixing	0.843	Skiing	0.467
CleanAndJerk	0.793	Nunchucks	0.579	Skijet	0.466
Diving	0.574	PizzaTossing	0.345	SoccerJuggling	0.615
Drumming	0.863	PlayingGuitar	0.948	Swing	0.679
Fencing	0.847	PlayingPiano	0.637	TaiChi	0.391
GolfSwing	0.391	PlayingTabla	0.912	TennisSwing	0.456
HighJump	0.617	PlayingViolin	0.554	ThrowDiscus	0.462
HorseRace	0.688	PoleVault	0.640	TrampolineJumping	0.645
HorseRiding	0.647	PommelHorse	0.863	VolleyballSpiking	0.349
HulaHoop	0.595	PullUps	0.820	WalkingWithDog	0.489
JavelinThrow	0.489	Punch	0.920	YoYo	0.584
JugglingBalls	0.766	PushUps	0.719	**AVERAGE**	**0.651**

HMDB
51 action categories

Action Category	Train AP	Action Category	Train AP	Action Category	Train AP
brush_hair	0.558	hit	0.055	shoot_ball	0.232
cartwheel	0.153	hug	0.338	shoot_bow	0.638
catch	0.223	jump	0.280	shoot_gun	0.464
chew	0.510	kick	0.171	sit	0.128
clap	0.389	kick_ball	0.298	situp	0.800
climb	0.162	kiss	0.223	smile	0.380
climb_stairs	0.435	laugh	0.346	smoke	0.163
dive	0.281	pick	0.028	somersault	0.256
draw_sword	0.082	pour	0.365	stand	0.208
dribble	0.559	pullup	0.609	swing_baseball	0.426
drink	0.185	punch	0.563	sword	0.340
eat	0.120	push	0.274	sword_exercise	0.287
fall_floor	0.184	pushup	0.652	talk	0.158
fencing	0.267	ride_bike	0.185	throw	0.343
flic_flac	0.533	ride_horse	0.297	turn	0.200
golf	0.566	run	0.117	walk	0.371
handstand	0.209	shake_hands	0.119	wave	0.133
				AVERAGE	**0.311**

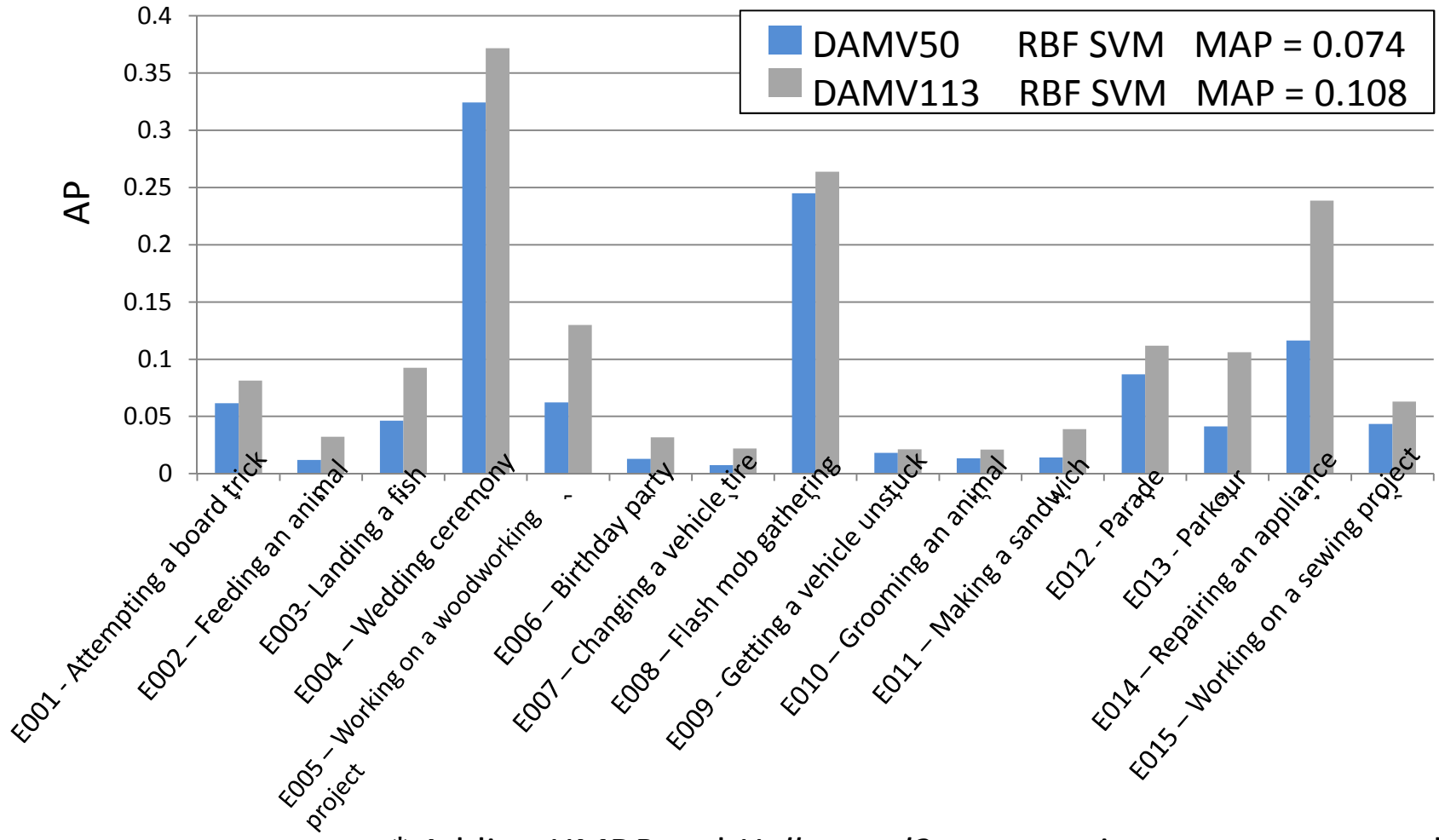
Hollywood2
12 action categories

Action Category	Train AP
AnswerPhone	0.130
DriveCar	0.814
Eat	0.259
FightPerson	0.599
GetOutCar	0.266
HandShake	0.212
HugPerson	0.344
Kiss	0.447
Run	0.559
SitDown	0.492
SitUp	0.130
StandUp	0.412
AVERAGE	**0.389**

DAMV (113 dimensions) = [

2 Dynamic Action Modeling – Event Prediction Performance

IBM data split of MED11 Event + Dev-T sets Training Set : ~7K videos Test Set : ~5K videos



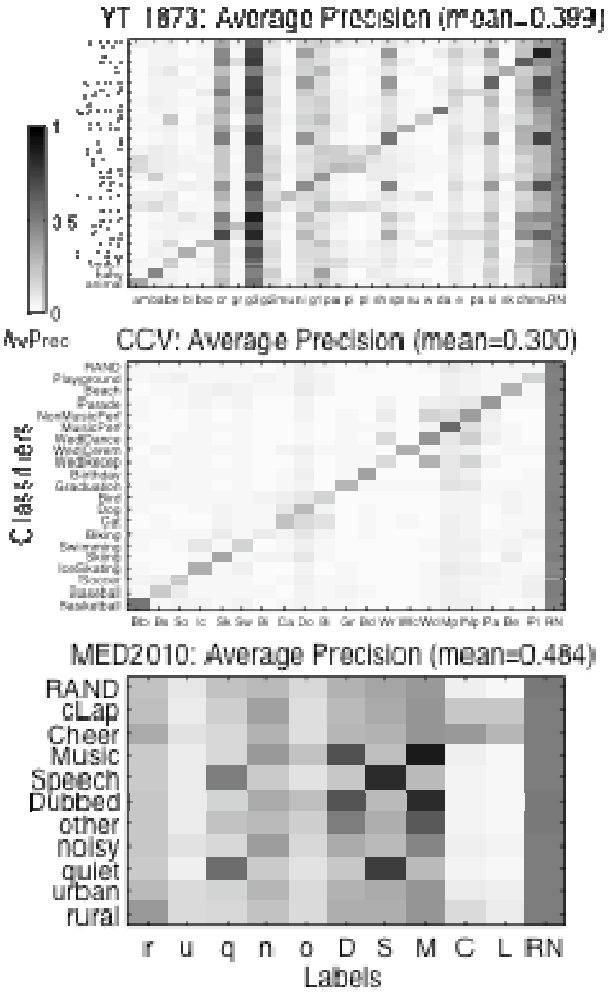
* Adding *HMDB* and *Hollywood2* concepts improves event prediction

2 Audio Semantic Models (55 concepts)

- Audio Classifiers trained from 3 data sets:

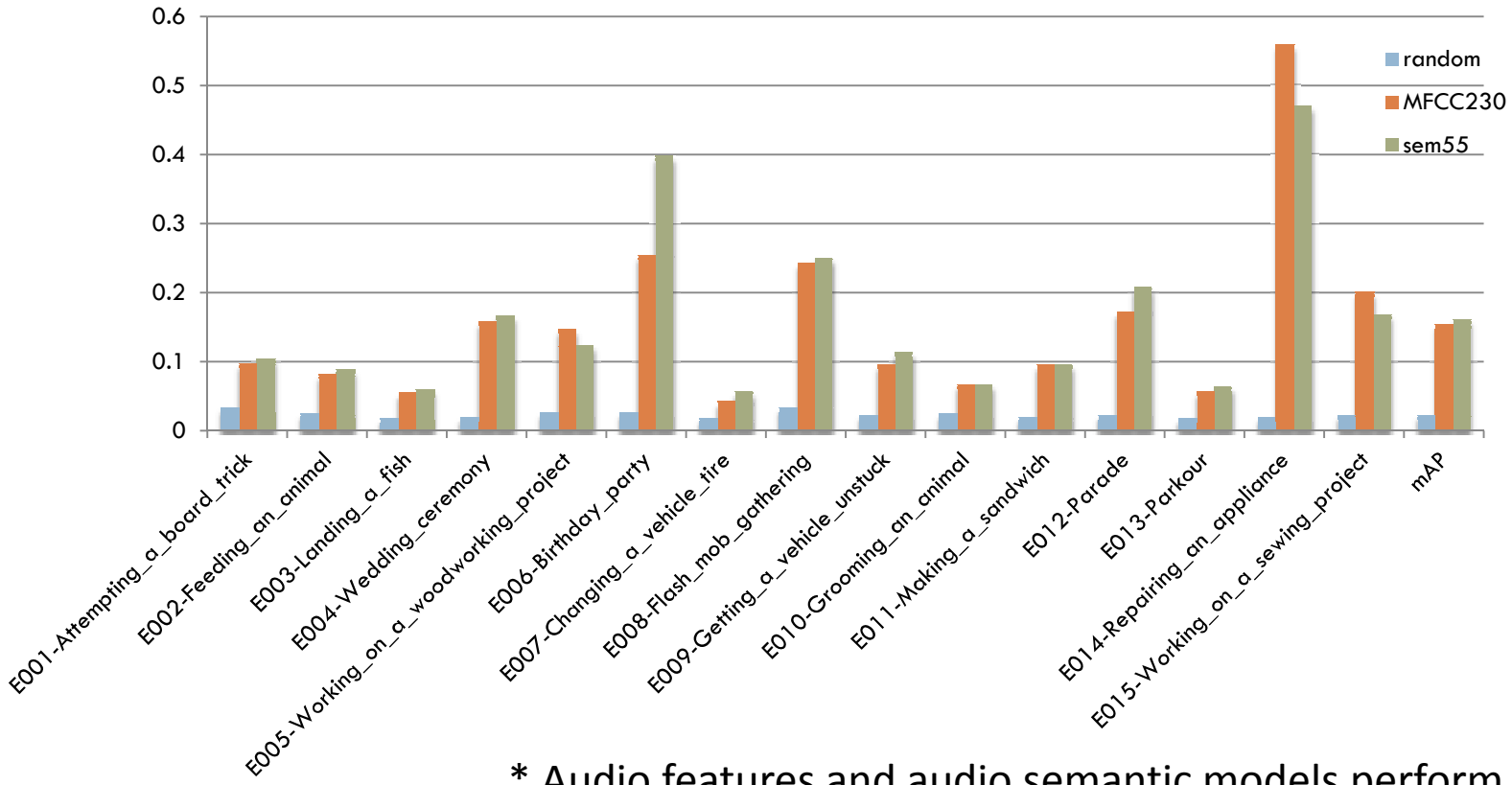
Dataset	# videos	# classes	mAP
YouTube 1873 [Lee & Ellis 2010]	1873	**25** (sport, animal, night, beach …)	0.40
Columbia Consumer Video (CCV) [Jiang et al. 2011]	9413	**20** (soccer, cat, birthday, beach …)	0.30
MED-2010 [Jiang et al. 2010]	6626 10 sec segments	**10** (rural, urban, speech, clap…)	0.48

… for a total of 55 audio semantic models



2 Audio Semantic Models Performance

- Predicting Events E001..E015 with audio classifiers:
  - among pool of 6354 DEVT1 + event kit examples



* Audio features and audio semantic models perform similarly

Event Modeling

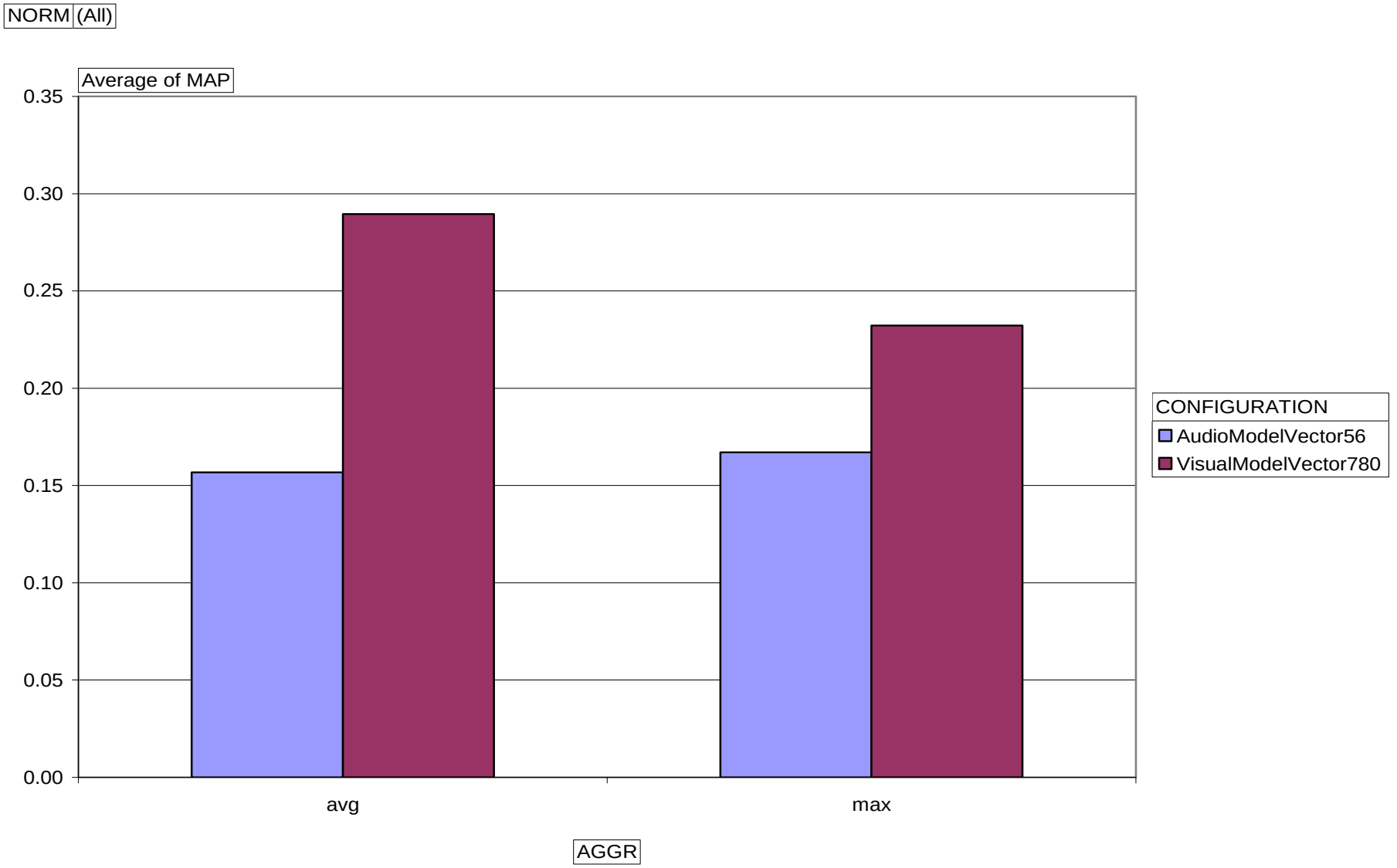
3

Event Agent Generation – Parameter Space

- **Features normalized to calibrate dimensions based on distribution:**
  - Linear: RANGE, STRETCH, STRETCH+L1, STRETCH+L2
  - Gaussian: MEAN+VARIANCE, VARIANCE-ONLY
  - Logistic: SIGMOID, SIGMOID+L1, SIGMOID+L2
- **Features aggregated from frame/segment level to video level using:**
  - AVG or MAX feature pooling per dimension
  - Temporal pyramids
  - Scene-Aligned Modeling (SAM)
- **Event modeling:**
  - Linear regression
  - SVM: linear, RBF, Chi^2, histogram intersection kernels
- **Fusion modeling:**
  - Weighted AVG with uniform/manual/AP-based weights
  - Greedy ensemble fusion with forward model selection
  - AdaBoost, ridge regression, lasso, linear SVM
  - Scene-aligned models: early fusion of static features based on scene alignment
- **Score calibration and threshold selection**
  - Logistic sigmoid score normalization based on collection statistics
  - Threshold selection based on optimal performance at target error ratio

3

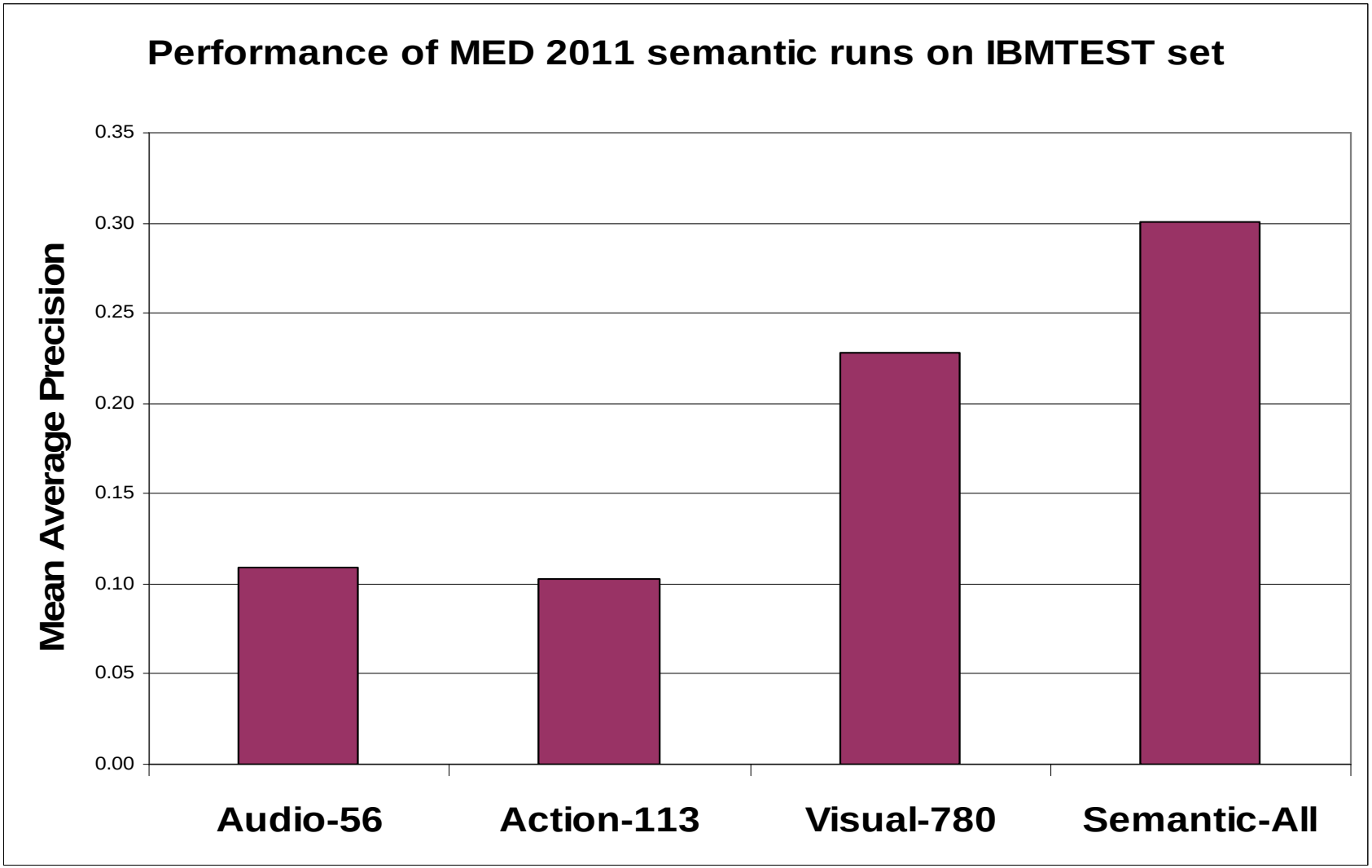
Feature Aggregation Analysis (IBMTRAIN)



AVG aggregation better for visual semantic features, while **MAX** better for audio features

3

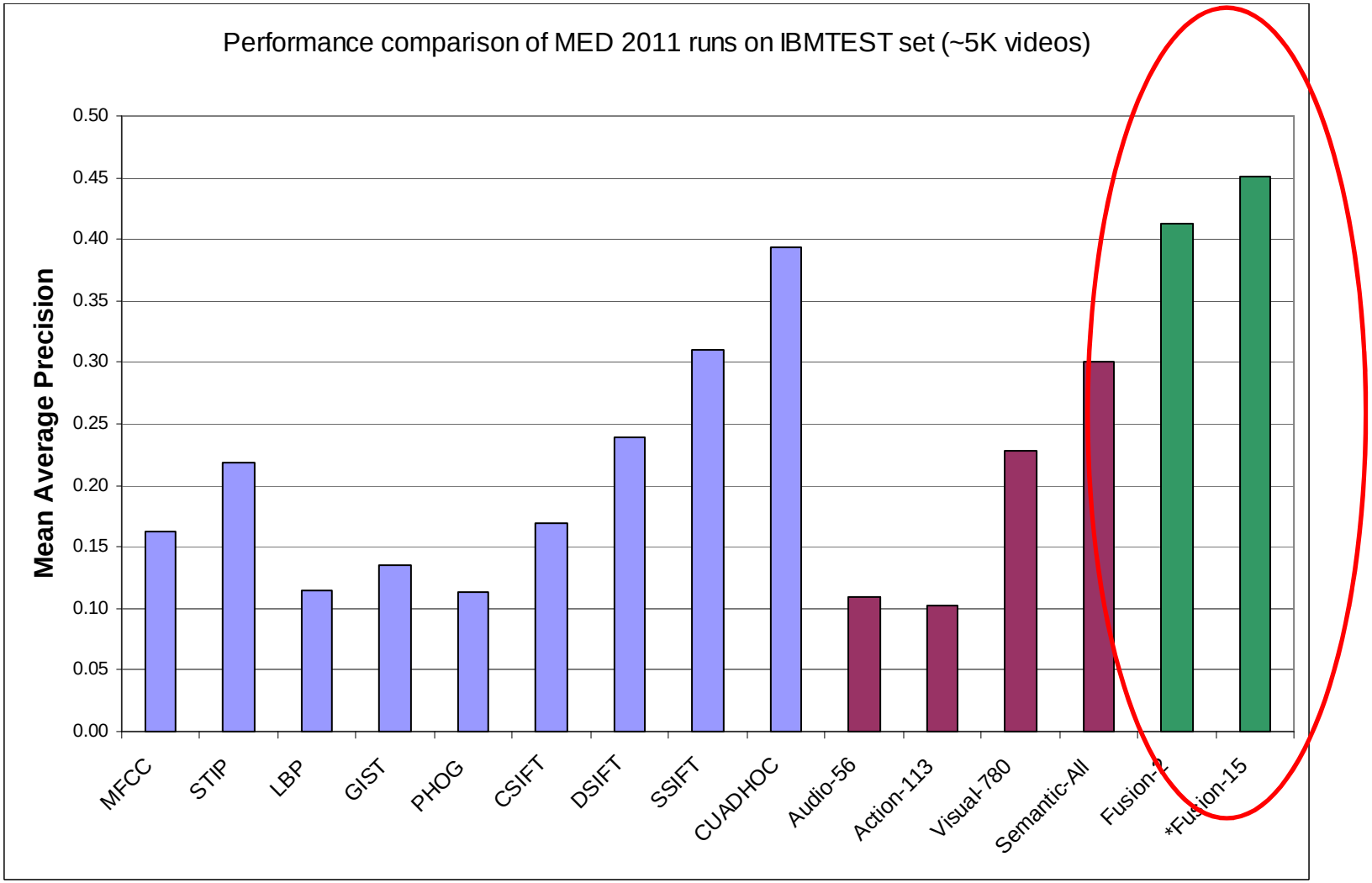
Semantic Features Analysis (IBMTEST)



- **Visual semantic features** significantly outperform **audio** and **action semantic features**
- Fusion across all semantic features further improves MAP score by over 30%

4

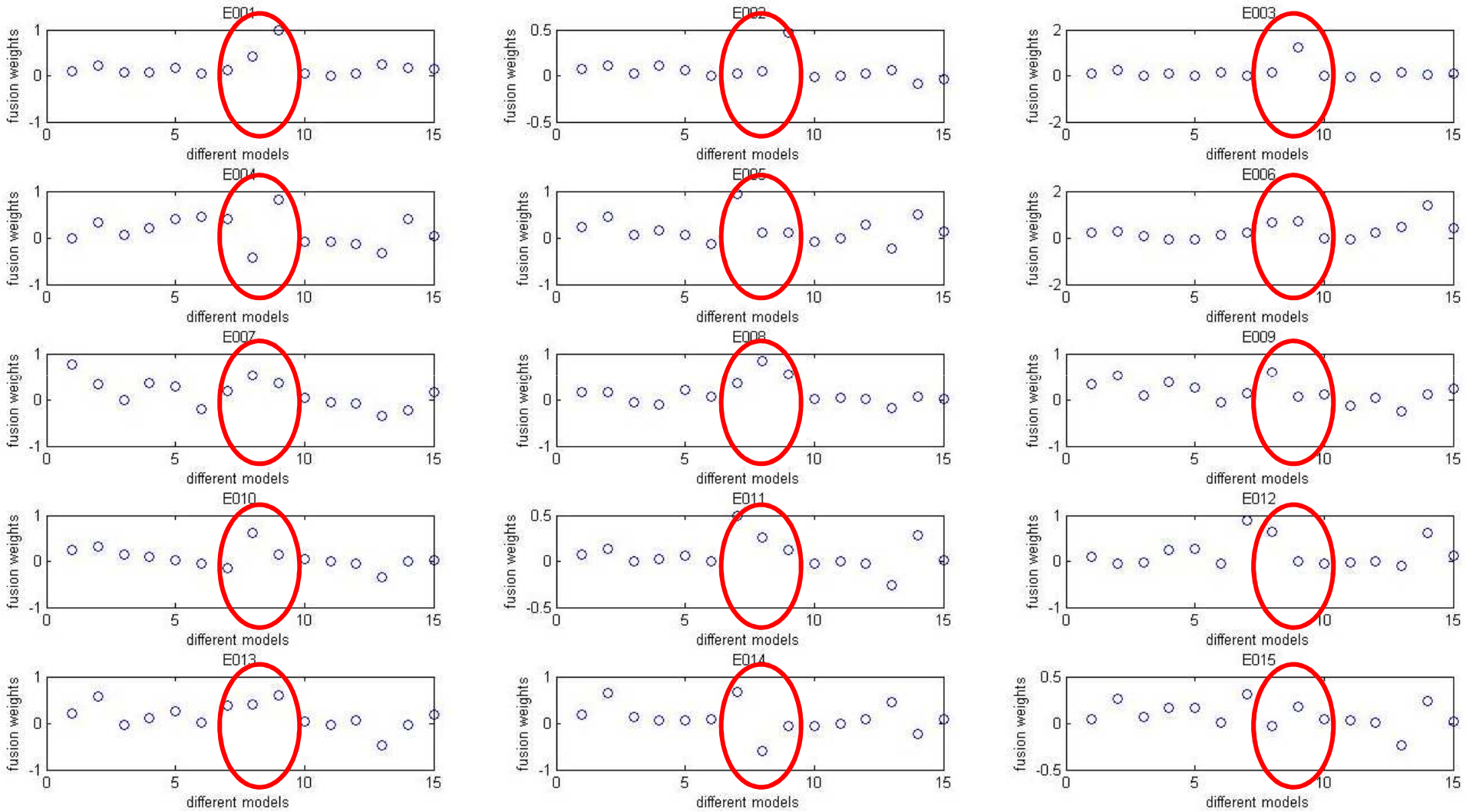
Multi-Modal Fusion Analysis (IBMTEST)



- **Semantic features** did not outperform **low-level features** but fusion of both is best
- **Fusion improves by ~30%** over best single feature, by ~10% over low-level feature fusion

4

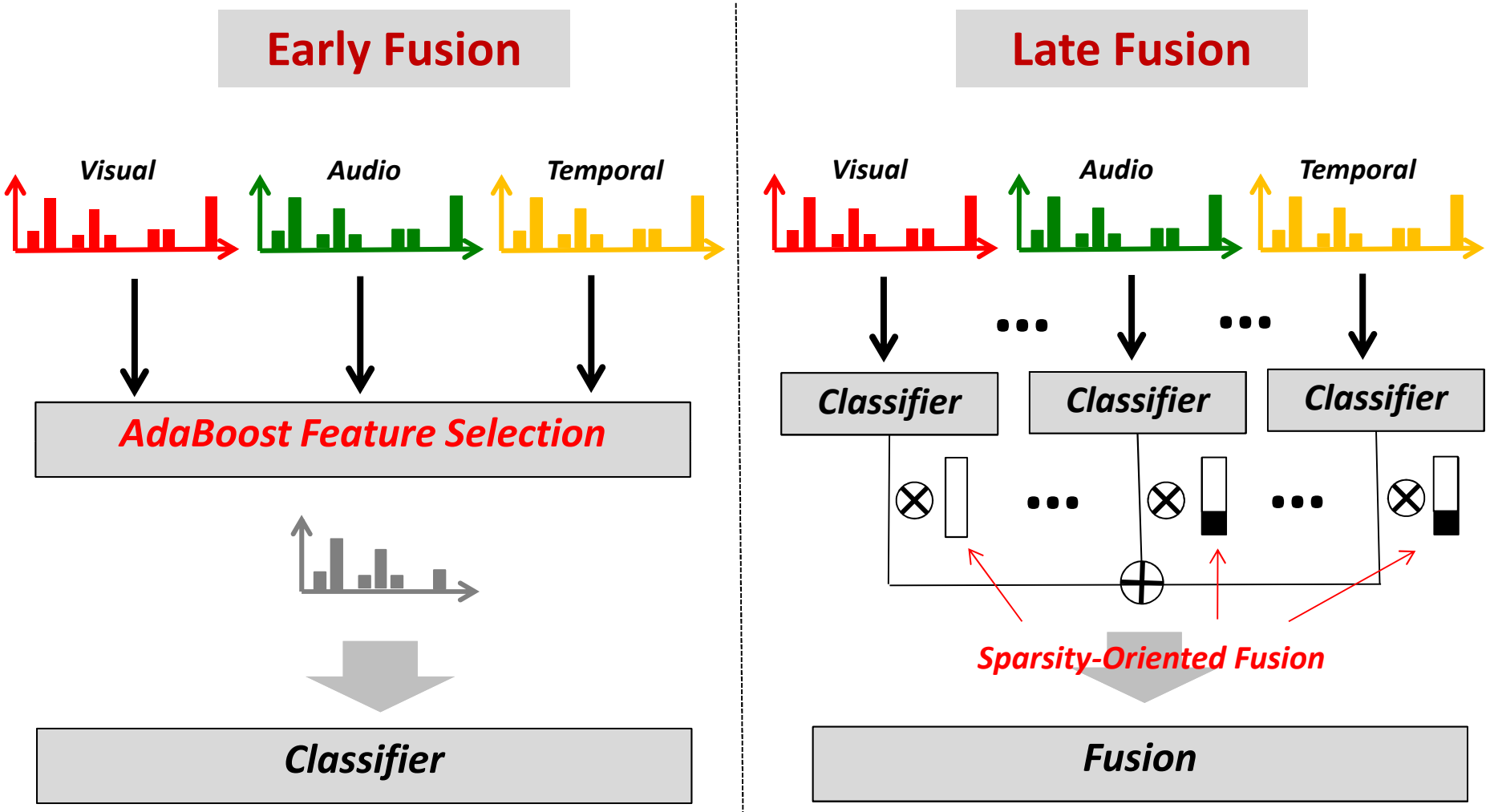
Component Runs Contribution Analysis (IBMTEST)



- **Semantic feature-based runs** appear with highest weights for **12 out of 15 events**
- **A single semantic concept** was a significant predictor for **9 out of 15 events**

4

Experiments with Multi-Modal Fusion



Approach produces 1,500-dim feature that out-performs original 14,000-dim features

More features/components not necessarily better

4

Late Feature Fusion: Comparative Study

- *Comparison on MED'2011 task*
  - *Validation set: 5K videos used to tune the fusion parameters*
  - *Evaluation set: 32K videos kept anonymous during competition*

Fusion Runs	***Validation Set***	***Evaluation Set***	***Sparsity[1]***
AdaBoost (6 components)	0.3850	0.2781	
Uniform (6 components)	0.3743	0.2676	
Ad hoc (6 components)	0.3847	0.2719	
Ridge regression (9 components)	0.4032	0.2786	0.0%
Ridge regression (15 components)	0.4112	0.2838	0.0%
Ridge regression (24 components)	0.4159	0.2799	0.0%
Lasso (9 components)	0.4025	0.2789	43.7%
Lasso (15 components)	0.4132	0.2833	48.4%
Lasso (24 components)	0.4113	0.2792	55.8%
Tree lasso (24 components)	0.4038	0.2781	62.5%

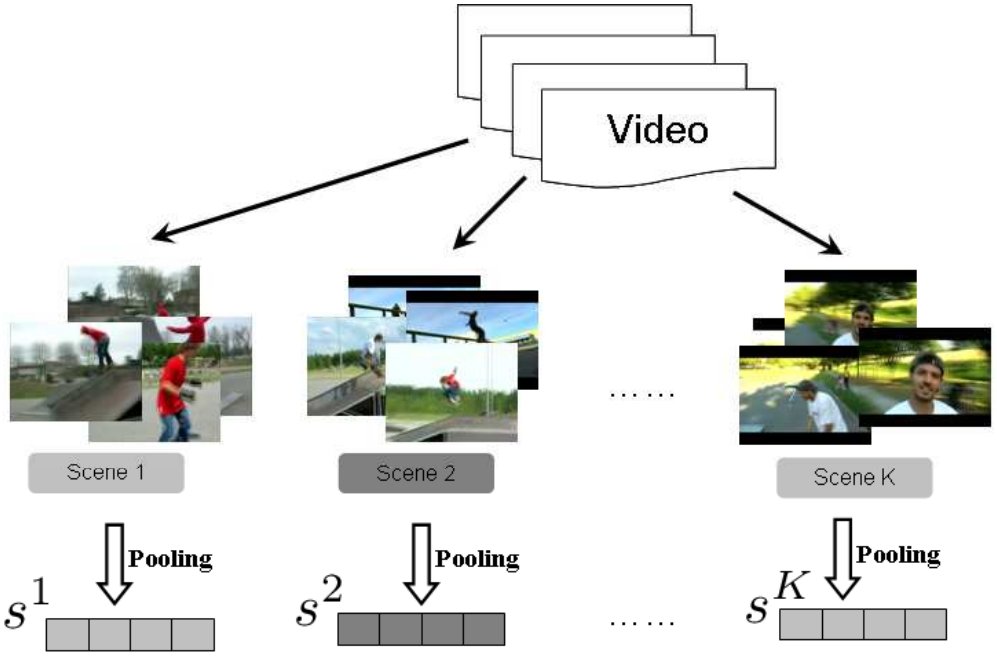
[1] **Sparsity denotes the percentage of zero coefficients.**

More features don't guarantee better performance…
Lasso method attains comparable performance with high sparsity!

Scene Aware Concurrent Pooling (SACP) – **details in TRECVID poster*

$$\mathbf{x}_t \rightarrow [w_t^1 \mathbf{x}_t, w_t^2 \mathbf{x}_t, \cdots, w_t^K \mathbf{x}_t],$$

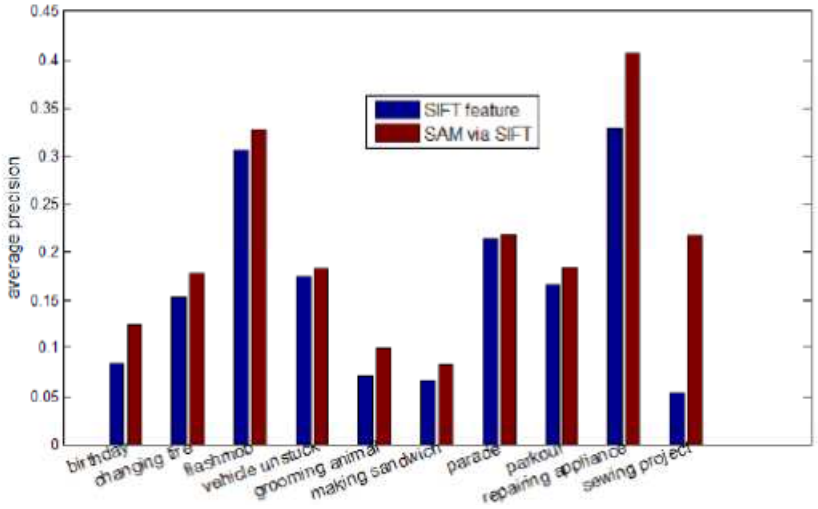
$$w_t^k = \frac{1/(d_t^k)^3}{\sum_l 1/(d_t^l)^3}$$



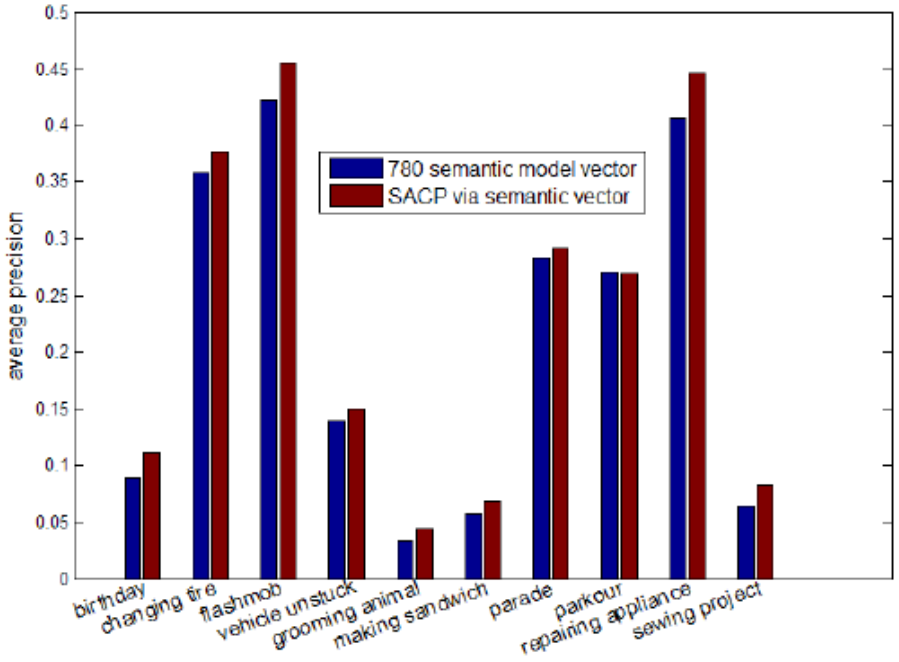
- Traditional pooling averages feature vectors within neighborhood, e.g., spatial regions in images
- SACP aggregates video features (low-level or semantic) into concurrent scene components that support subsequent event classification

Scene Aware Concurrent Pooling (SACP) Results

SACP using SIFT features



SACP using semantic features

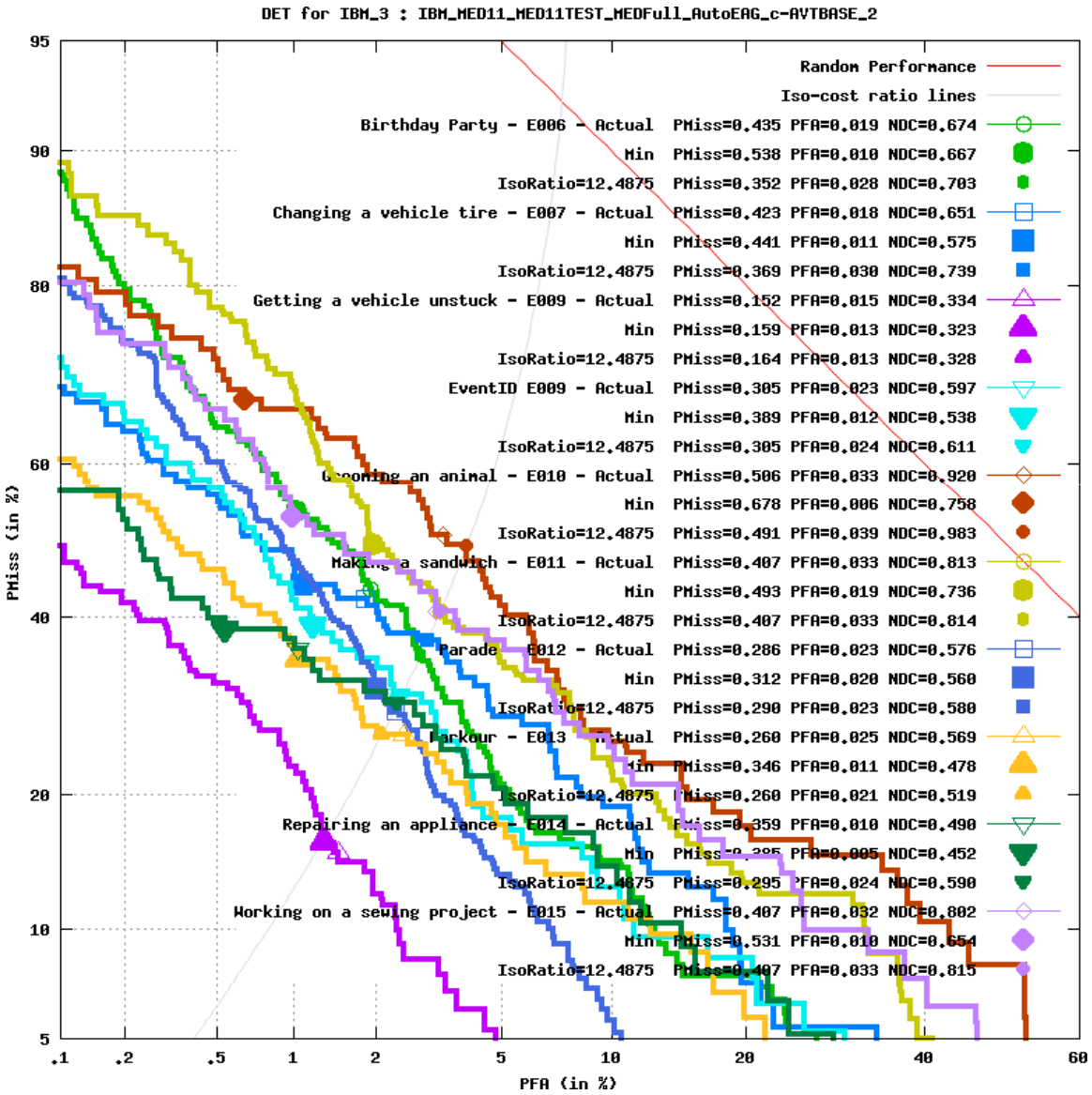


Model	Accuracy
STIP histogram	21.96%
C2	23.18%
SACP + STIP	27.84%

SACP on Brown-MIT's human motion database

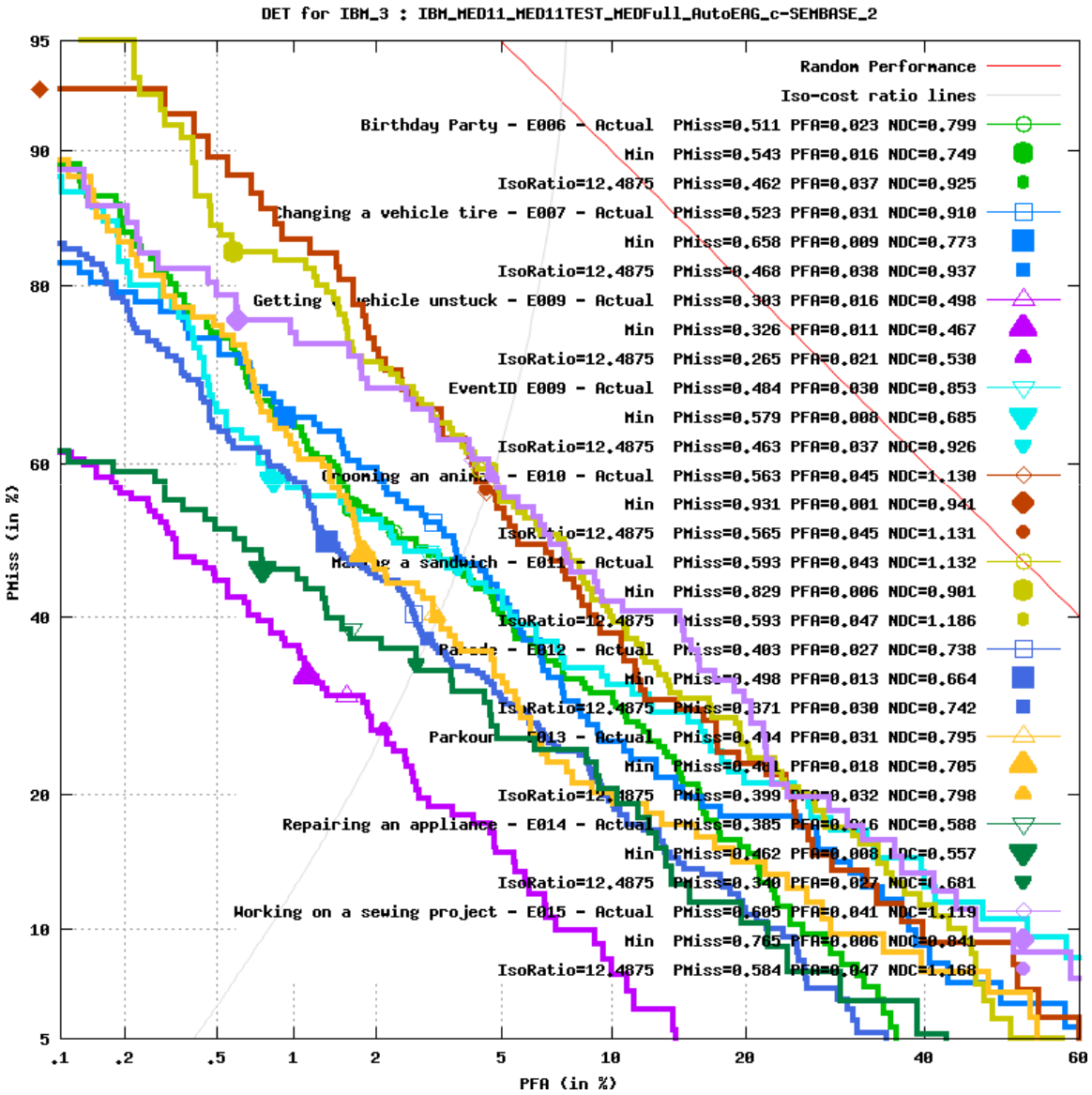
MED Experiments

Run 1 Results – Low-Level Features (MED11TEST)



- Low-level features only
- Sparse SIFT + Dense SIFT + Color SIFT + STIP + MFCC
- SVM with histogram intersection + Chi2 kernels
- Fusion weights based on ridge regression

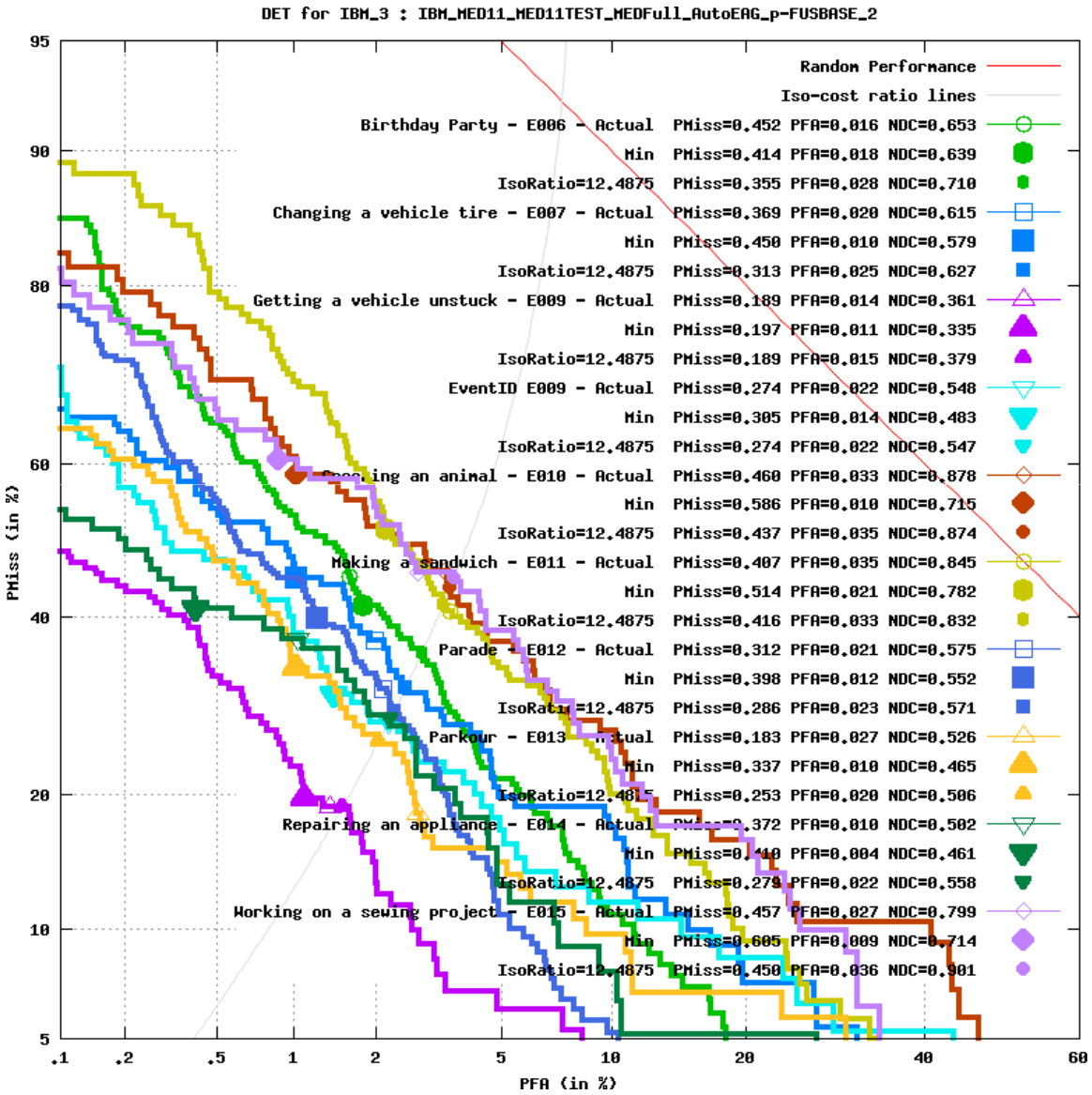
Run 2 Results – Semantic Features (MED11TEST)



- Semantic features only
- 780 visual + 113 action + 55 audio semantic features
- 10 feature normalization and 2 feature aggregation methods
- SVM with RBF, Chi2, and histogram intersection kernels
- Greedy ensemble fusion with forward model selection

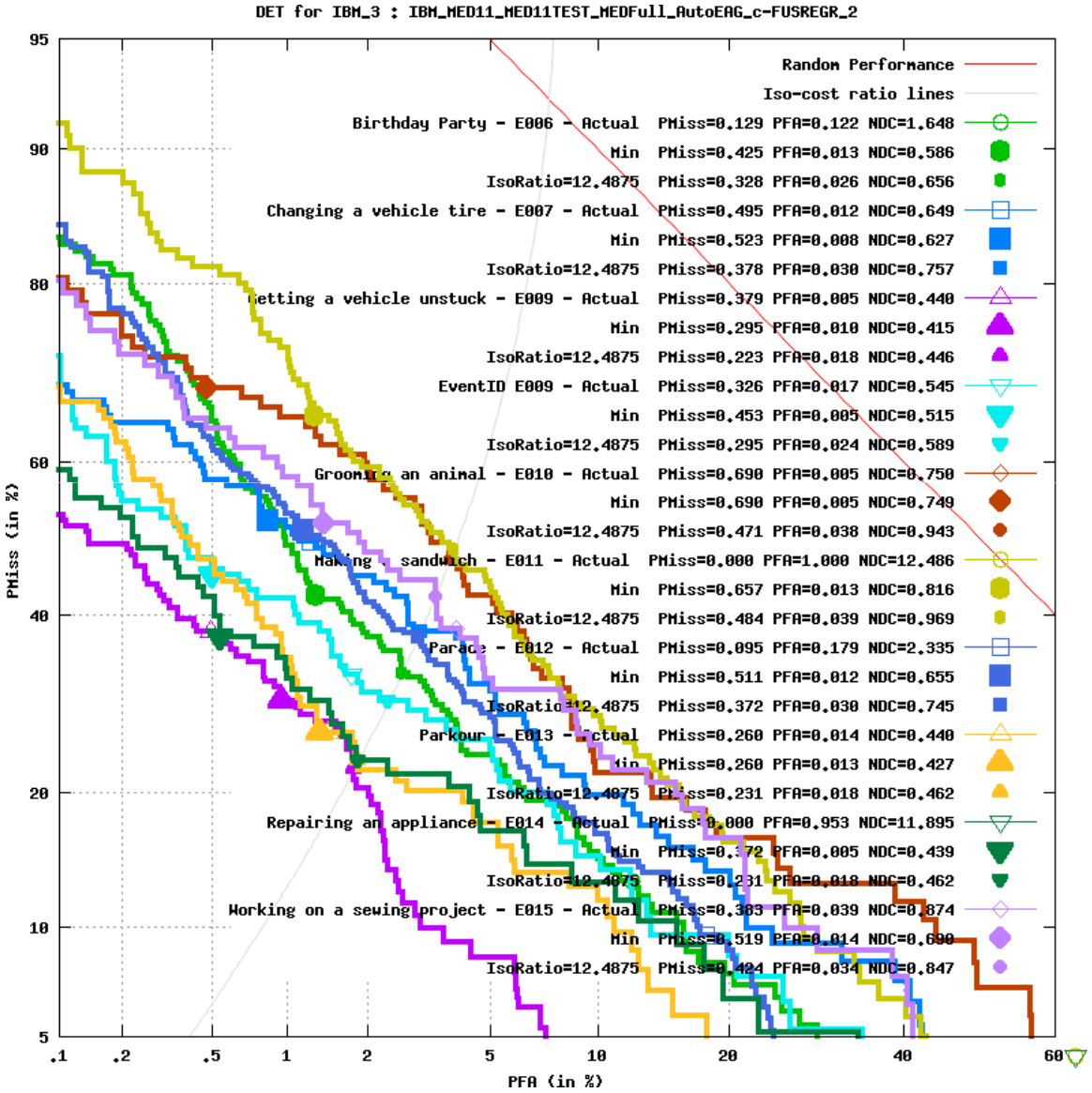
5

Run 3 Results – Multi-modal Fusion (WAVG, MED11TEST)



- Fusion of low-level and semantic features
- Fusion improves 30% over any single component
- Weights proportional to their IBMTEST MAP scores

5 Run 4 Results – Multi-modal Fusion (SVM, MED11TEST)



- Fusion of low-level and semantic features plus scene-aware concurrent pooling
- 14 components runs (8 low-level features + 5 semantic features + SACP)
- Weights learned with linear SVM

Lessons Learned and Next Steps

Lessons Learned

- **Semantic features and low-level features both contribute**
  - Fusion outperforms best single feature by more than 30%
- **Discriminative semantic features complement low-level features:**
  - A single concept was a significant predictor for 9 out of 15 events
  - Modeling events with a **single** semantic feature has respectable MAP of 0.15
  - Experiments show that expanding number of semantic features can help even if trained from noisy data
  - Need better understanding of what semantic concepts to model
- **Semantic feature-based approach lends well to parallelization**
  - Enables easy scale-out to large data sets and continuous data streams
  - Performed 4 CPU-years worth of processing in ~2 days on 800-core system
- **Semantic features support user-friendly event description**
  - Event models and decisions more easily explained through semantic basis

Ongoing and Future Directions

- **Greatly expand discriminative semantic feature space including ability to train from noisy labels**
  - Performance-driven semantic taxonomy expansion (minimize confusion)
  - Data-driven semantic taxonomy expansion (e.g., leverage Web)
  - Event-driven semantic taxonomy expansion (targeted expansion)
  - Multi-modal semantic taxonomy expansion (audio, actions)
  - Explore additional semantic feature selection methods (e.g., PageRank)
- **Improve event modeling based on semantic features**
  - Explore semantic sequence alignment kernels
  - Explore scene-aligned and semantics-aligned feature fusion
  - Explore semantic temporal motifs and structured event modeling
- **Improve fusion across low-level and semantic features**
  - Maximize feature complementarity using AdaBoost-like approach
- **Incorporate more modalities**
  - Speech, text and metadata