CMU Informedia @ TRECVID Multimedia Event Detection

Shoou-I Yu, Lu Jiang, Zexi Mao, Xiaojun Chang, Xingzhong Du, Chuang Gan, Zhenzhong Lan, Zhongwen Xu, Xuanchong Li, Yang Cai, Anurag Kumar, Yajie Miao, Lara Martin, Nikolas Wolfe, Shicheng Xu, Huan Li, Ming Lin, Zhigang Ma, Yi Yang, Deyu Meng, Shiguang Shan, Pinar Duygulu Sahin, Susanne Burger, Florian Metze, Rita Singh, Bhiksha Raj, Teruko Mitamura, Richard Stern, and Alexander Hauptmann

Carnegie Mellon University

Nov. 11, 2014

Carnegie Mellon University

Acknowledgement

- This work was partially supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.
- This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, it used the Blacklight system at the Pittsburgh Supercomputing Center (PSC).

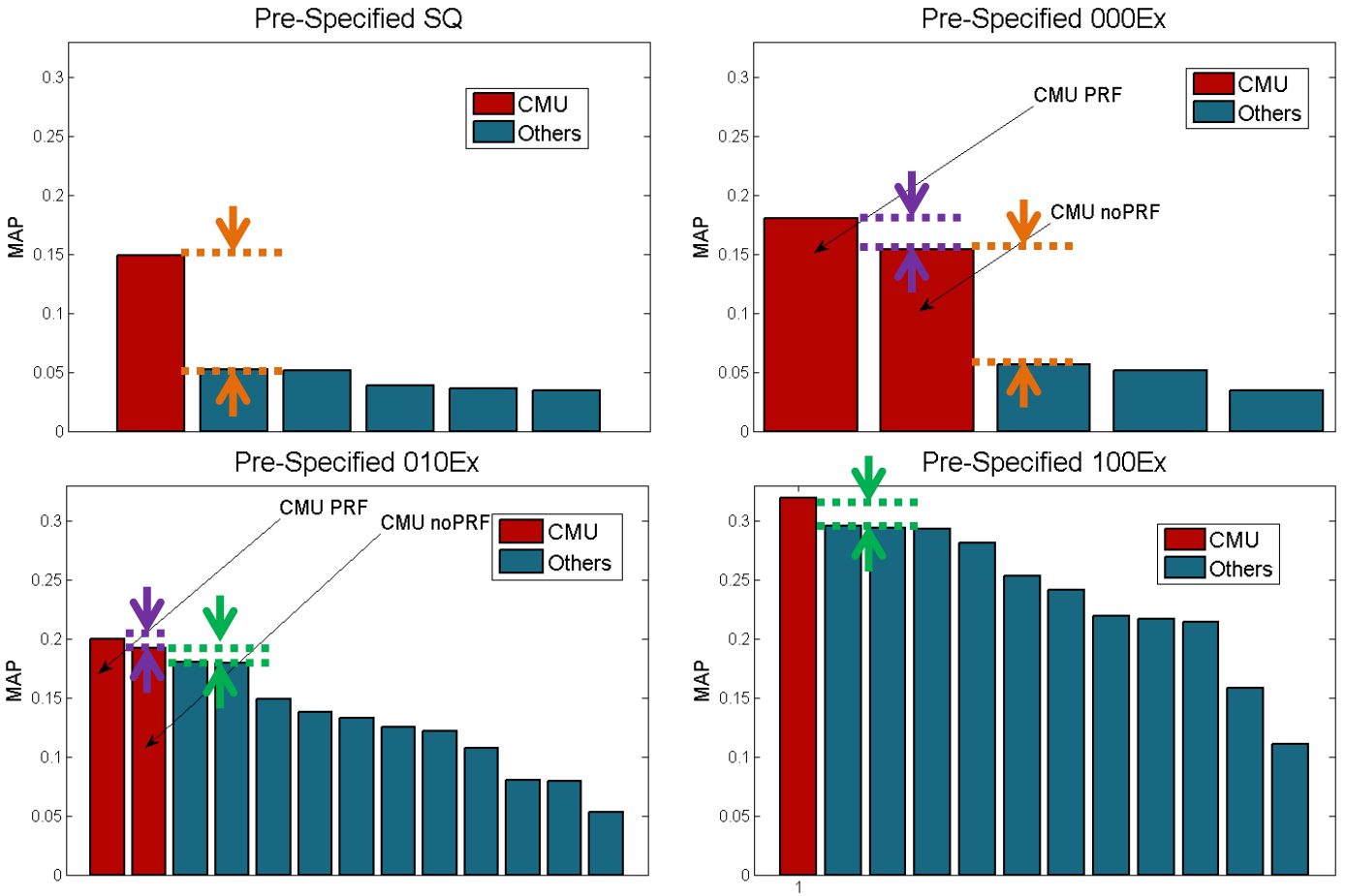
Recap of MED14 Results

Better semantics

Better features & fusion

Good reranking

Training done in < 16 minutes. Prediction done in < 5 minutes.



Preliminary performance on Pre-Specified events on MED14-Eval Full. Adhoc results have similar trend. Red bars are CMU runs.

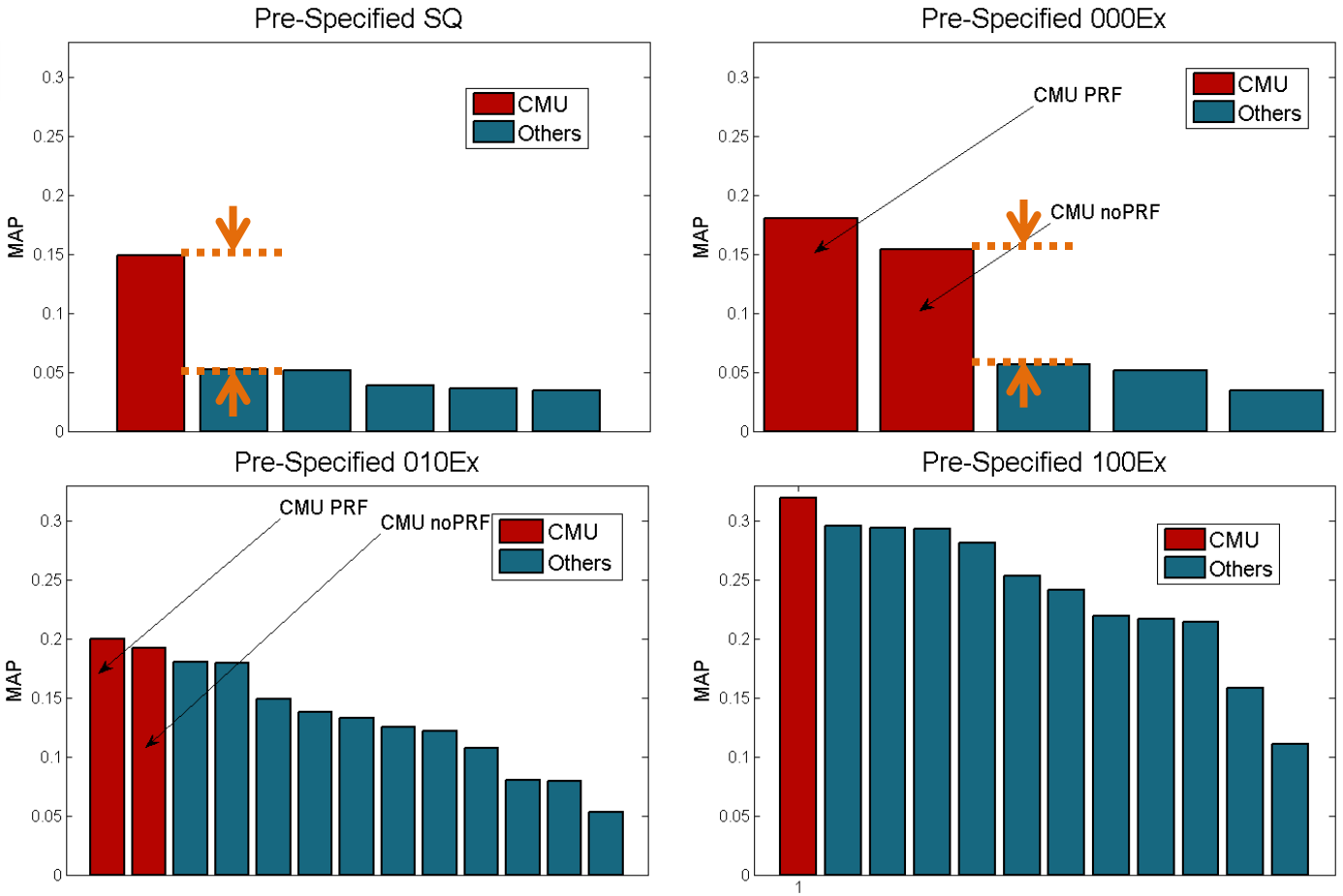
Recap of MED14 Results

Better semantics

Better features & fusion

Good reranking

Training done in < 16 minutes. Prediction done in < 5 minutes.



Performance on Pre-Specified events on MED14-Eval Full set. Adhoc results have similar trend. Red bars are CMU runs.

Semantic Concepts for SQ/000Ex Run

- Two sources of concepts, both very useful
  - **Shot-level video concepts (e.g. Semantic Indexing)**
    - Captures objects and action in context of videos
    - **Main focus of current section of the talk**
  - Static image concepts (e.g. ImageNET 1000)
    - Only capture static information (talk about this later)

MED14 Improvements Toward Better & Faster Shot-Level Semantics

- Better semantics
  - Used improved trajectories [1] as low-level feature
  - Increased vocabulary from 346 -> 3000
  - Self-paced learning for large-scale unbalanced training of semantic concepts
- Faster semantics
  - Fast training on large shared-memory machines
  - Linear classifiers for fast prediction of semantic concepts

[1]: H. Wang, C. Schmid, et al. Action recognition with improved trajectories. In ICCV, 2013.

Dataset Overview

Shot-based Semantic Concepts Datasets Overview

Dataset	#samples	#classes
HMDB	5k	51
CCV	8k	20
UCF101	10k	101
MED Research	20k	55
DIY	72k	1601
SIN346	0.5m	346
Yahoo YFCC100M	0.8m	644
Google Sports	1.1m	487

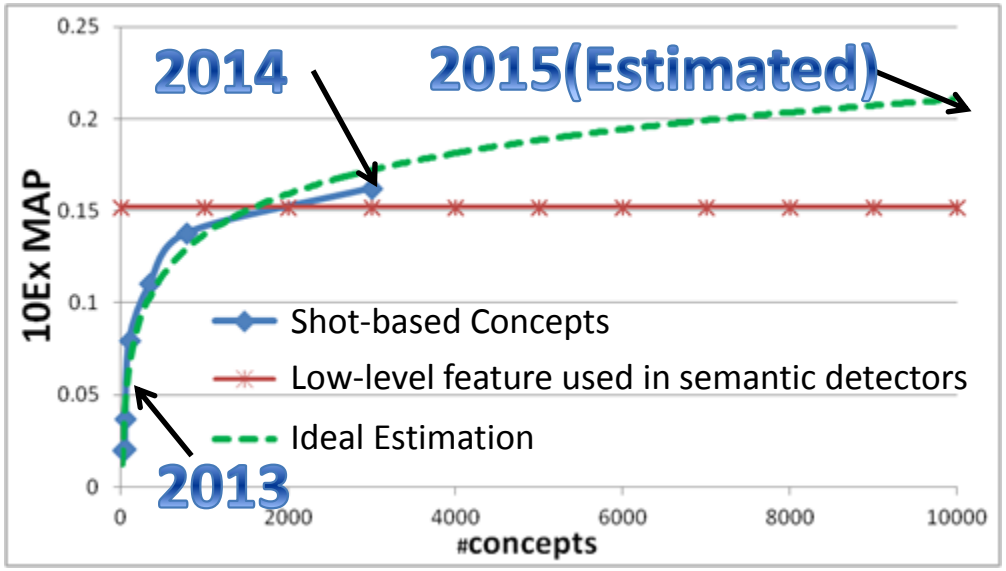
- We annotated over 55 concepts on the MED Research dataset
- Similar to Object Bank which have detectors available for download, Google Sports and DIY detectors were also downloaded from the Internet

http://vision.stanford.edu/projects/objectbank/index.html#software
http://gr.xjtu.edu.cn/web/dymeng/4
http://staff.itee.uq.edu.au/xue/sports.html

HMDB http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/
CCV http://www.ee.columbia.edu/ln/dvmm/CCV/
UCF101 http://crcv.ucf.edu/data/UCF101.php
Google Sports https://code.google.com/p/sports-1m-dataset/
Yahoo YFCC http://labs.yahoo.com/news/yfcc100m/
SIN http://www-nlpir.nist.gov/projects/tv2014/tv2014.html

Shot-level Semantics Outperforms Best Low-Level Feature

E021-E030 Performance



(Assuming we know optimal concept dataset for events)

- This is the first time we observed that semantic features **outperform the low-level feature it was trained on**.

How to train classifiers on highly-imbalanced data?

- We study the **self-paced learning** which provides theoretically justification for the concept training.
- Self-paced learning is inspired by the learning process of humans and animals.
- The samples are not learned randomly but organized in a meaningful order which illustrates from easy to gradually more complex ones.

Easy examples of "bus"

Learning from easy to more complex in a self-paced fashion.

Complex examples of "bus"

Jiang et al., "Self-paced Learning with Diversity," in *NIPS*, 2014.

Efficient Large-Scale Concept Detector Training

- Train linear SVM detectors on improved trajectories
  - Highly parallelized kernel matrix computation
  - In SVM training, requires **fast random access of kernel matrix**
    - A 330000 * 330000 matrix has size 200GB
    - Utilize **shared memory machines** such as PSC Blacklight*
    - Shared memory can be seen by all machines
    - Enables parallel training of many concept detectors

*: Pittsburgh Supercomputing Center Blacklight cluster. Max RAM-DISK 16T.

Carnegie Mellon University

Efficient Large-Scale Concept Detector Training & Prediction

- Training time for 644 semantic detectors
  - RAM-DISK up to 640GB utilized to achieve **8x speedup in the training**
  - **56 hours** to train **644 concept detectors** with **330,000 training examples** using 640 cores
- Prediction: **linear SVM prediction**, very fast

Shot-level Semantic Concept Summary

- Larger vocabulary & better features main reason of SQ/000Ex performance improvement
  - Other sources of improvement
    - Better query-concept mapping
    - Appropriate retrieval algorithms

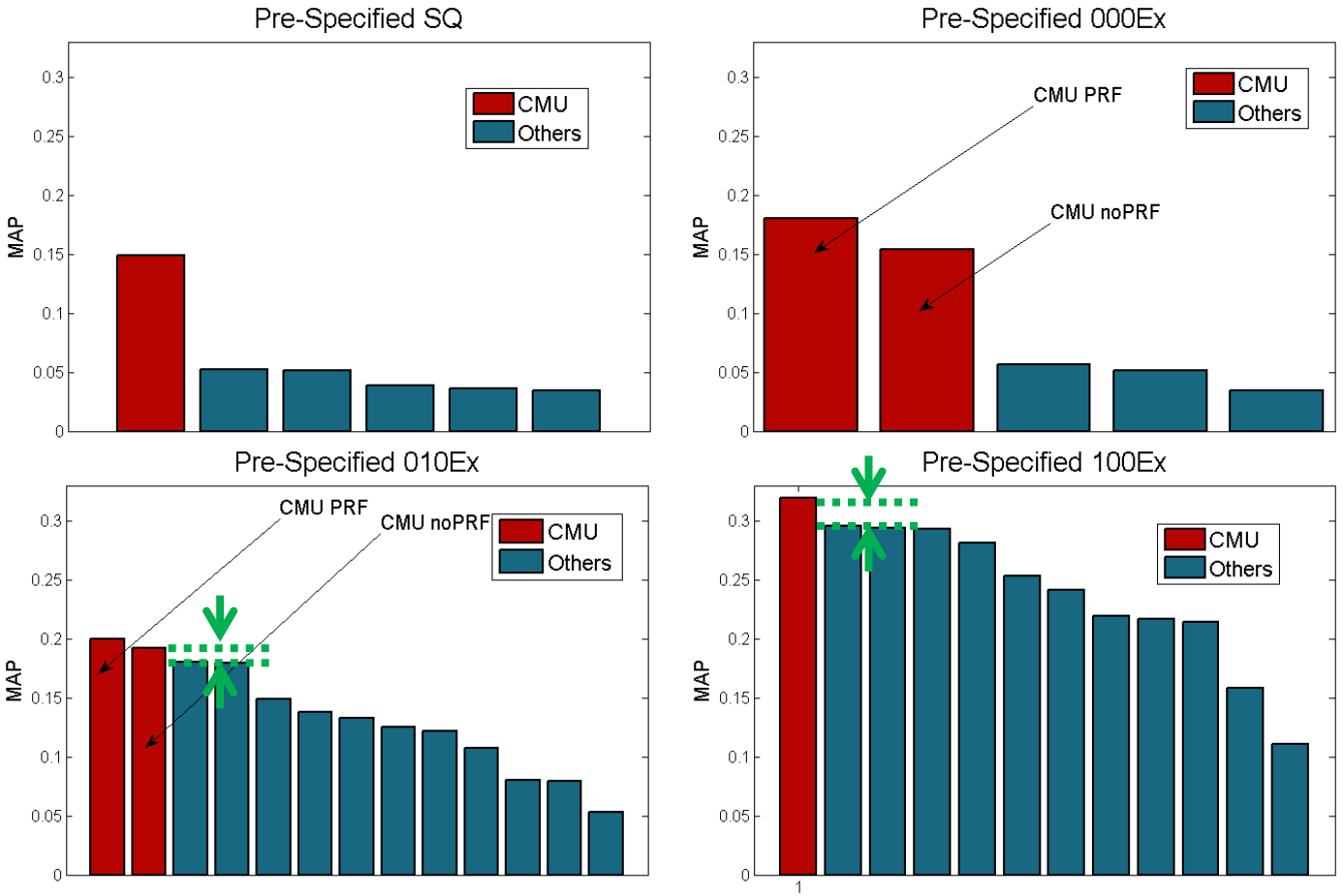
Recap of MED14 Results

Better semantics

Better features & fusion

Good reranking

Training done in < 16 minutes. Prediction done in < 5 minutes.



Performance on Pre-Specified events on MED14-Eval Full set. Adhoc results have similar trend. Red bars are CMU runs.

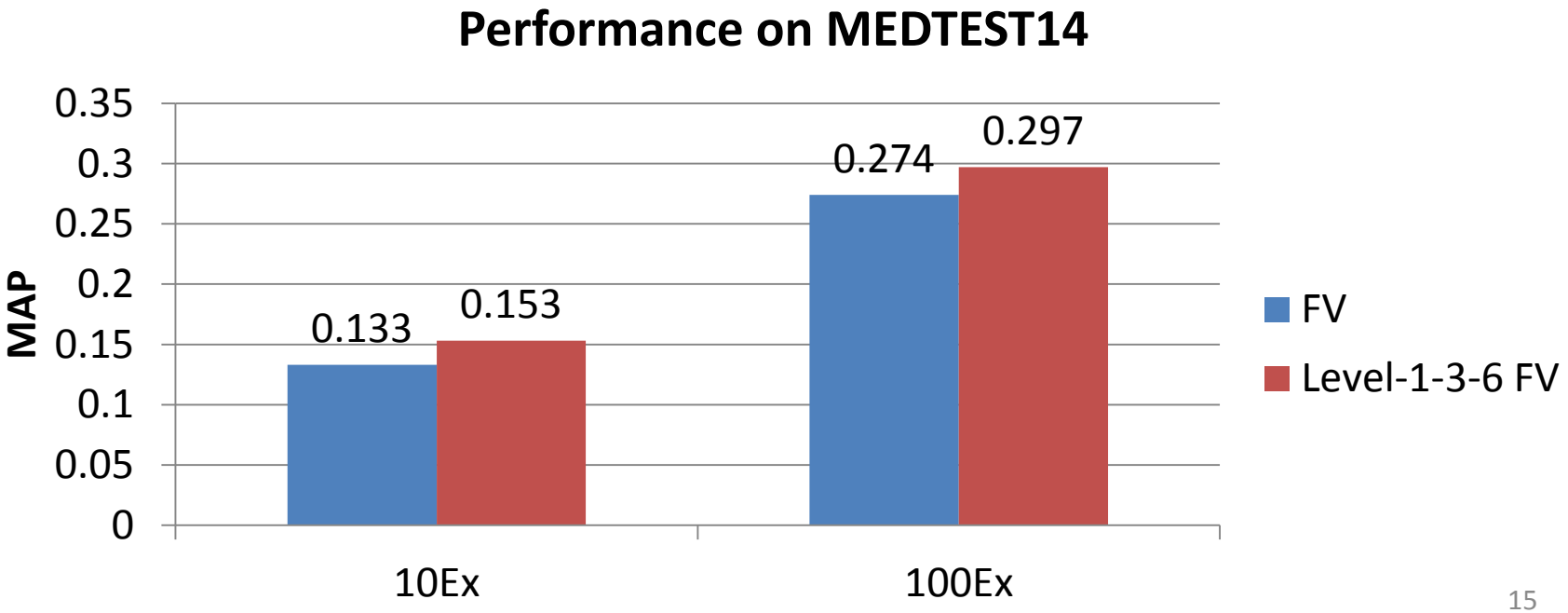
Overview of Improvements in Features and Fusion

Carnegie Mellon University

- Feature improvements
  - CMU improved trajectory
  - Better deep learning features
- Novel robust fusion method
  - Multistage Hybrid Late Fusion

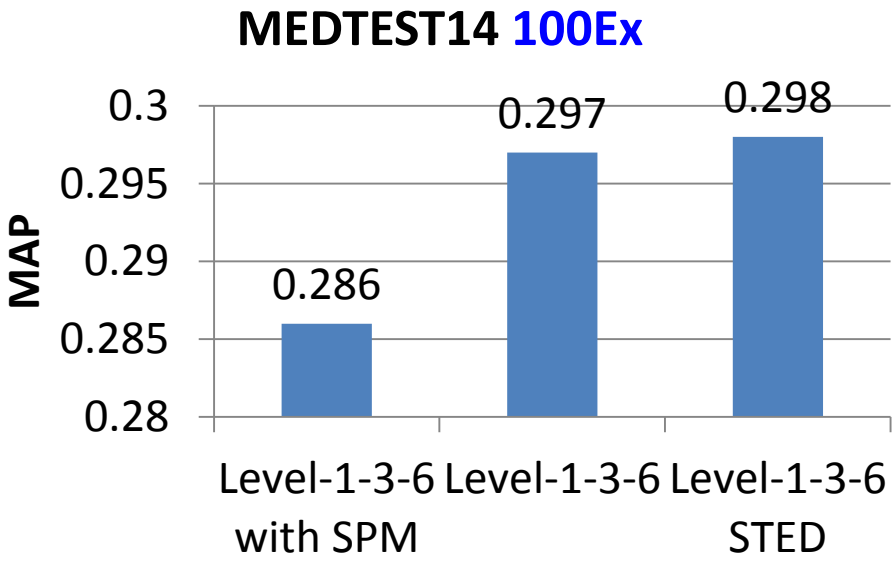
CMU Improved Trajectories 1: Modeling Multi-Temporal Scale

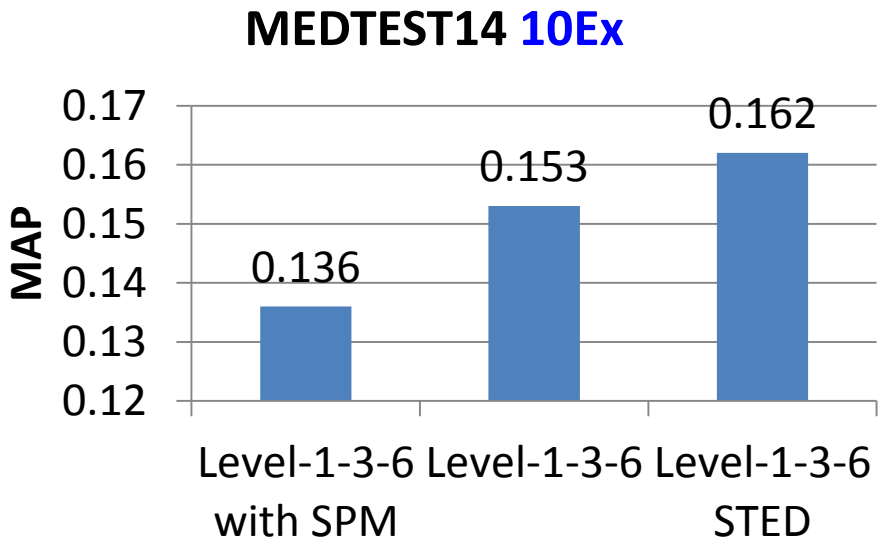
- The same action could be done at **different speeds**
  - Model action speed by multi-temporal-scale representation
  - Extract features by **playing the video at 1x, 3x, 6x speed**
    - Referred to as Level-1-3-6 FV (fisher vector)



CMU Improved Trajectories 2: Encoding Spatial & Temporal Info.

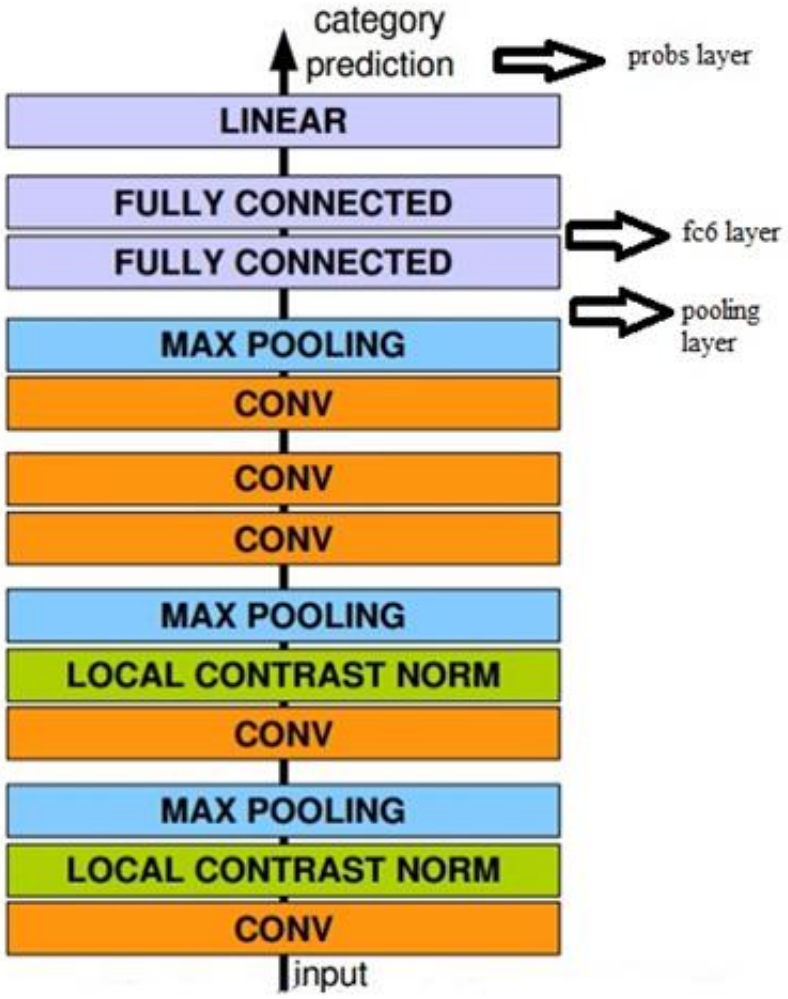
Feature	Description	Dimensions
SPM Fisher vector	Spatial pyramid matching	870K (too large!)
Spatial & Temporal Extension Descriptor (STED)	Add spatial (x,y coordinates) and timestamp t directly into raw feature	110K





CMU Improved Trajectories total improvement:
8.7%, **21.8%** relative on 100Ex, 10Ex respectively.

Better Deep Convolutional Neural Network (DCNN) Features



- Features are extracted from the pooling, fully connected (fc6) and probability (probs) layers.
- Features are pooled into video-level features and given to the classifier.
- Improvements
  - Trained on ImageNet Full 21841 classes
  - 1000 classes, but trained with more layers, more filters, smaller strides & multiview

Figure from A. Krizhevsky, I. Sutskever and G. E. Hinton., "Imagenet classification with deep convolutional neural networks," in NIPS, 2012.

Performance of ImageNet Full DCNN

ImageNet Full: 14.2 Million images with 21,841 object concepts

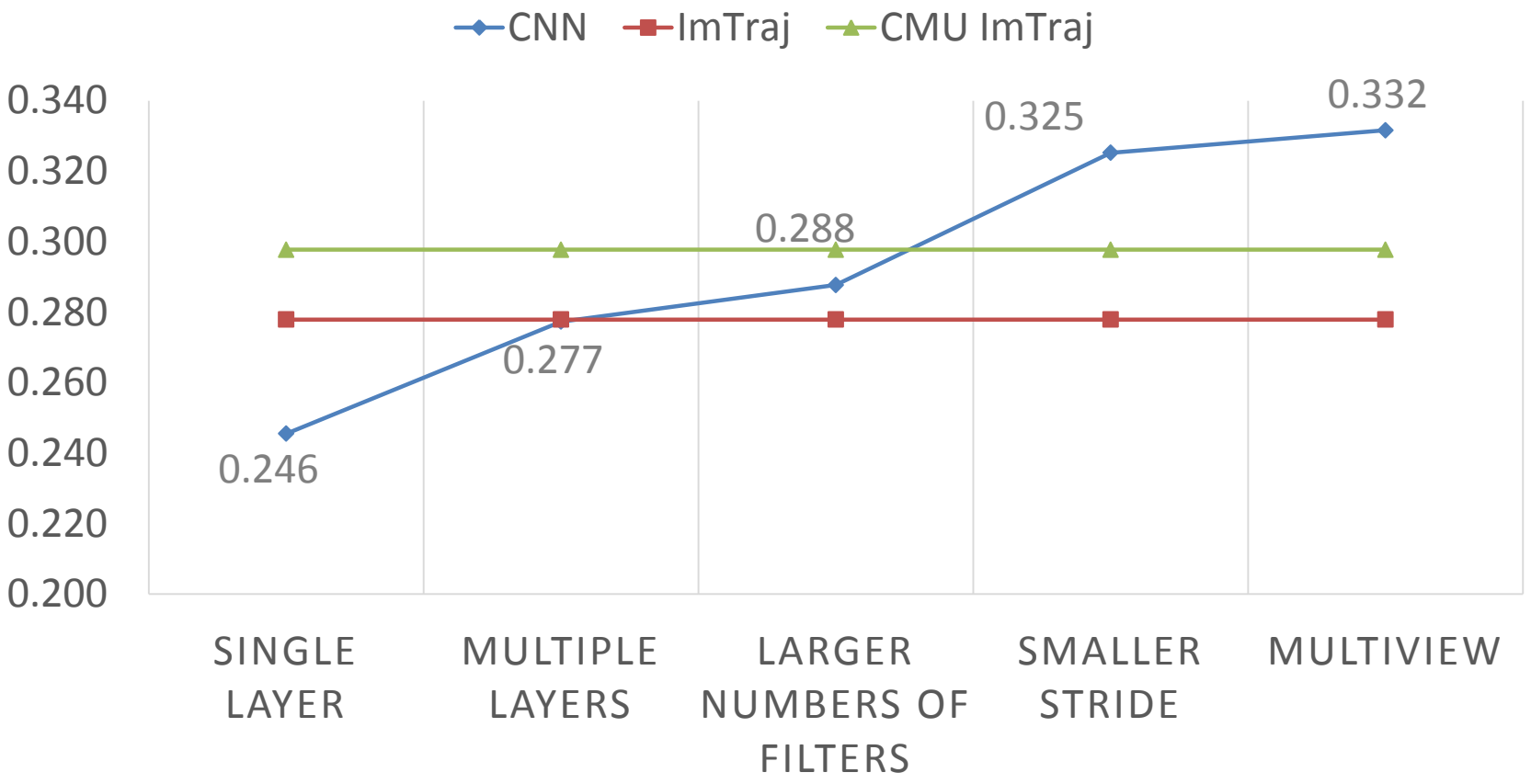
ImageNet Full (Accuracy)	MEDTEST14 100Ex (MAP)		
0.152	Avg Pooling	Max Pooling	Fusion
	0.279	0.273	**0.280**

No motion information used!

MAP of other features:

- SIFT: 0.157
- CMU Improved Trajectories: 0.298

Evolution of 1000 Class DCNNs on MEDTEST14 100Ex MAP



Performance of DCNN models outperforms improved trajectories as more model variations are added.

Overview of Improvements in Features and Fusion

- Feature improvements
  - CMU improved trajectory
  - Better deep learning features
- Novel robust fusion method
  - Multistage Hybrid Late Fusion

Carnegie Mellon University

Multistage Hybrid Late Fusion

- Simple ways of describing a feature
  - Prediction scores
  - Prediction rank
- In addition, find **key features** by clustering
  - Many features are correlated
  - Prediction scores contain large amount of noise
  - Simultaneously find key features and suppress noise with PCA clustering
- Fusion of late fusion methods

Fusion Strategies in Hybrid Fusion

- Hybrid fusion: a fusion of "late fusion"
  - no single fusion strategy does a better job than others on all events
- Average fusion of individual fusion strategies
  - Average
  - Leave-one-feature-out performance drop as fusion weights
  - **SGD-MAP:** use MAP as loss function in stochastic gradient descent
  - …
- Strong theoretical guarantee:
  - Provable optimal bounds for the proposed fusion.

The new fusion provides on average 2% absolute improvement over naïve fusion.

Summary of Feature and Fusion Improvements

ID	Features	100Ex	10Ex
1	MFCC bag-of-words	0.112	0.051
2	MoSIFT fisher vector	0.184	0.084
3	CMU improved trajectories	0.298	0.162
4	Multiview DCNN	0.332	0.197
5	3 + 4	0.369	0.238
6	Average fusion, all features	0.404	0.260
7	MHLF, all features	**0.417**	**0.283**

MAP on MEDTEST14

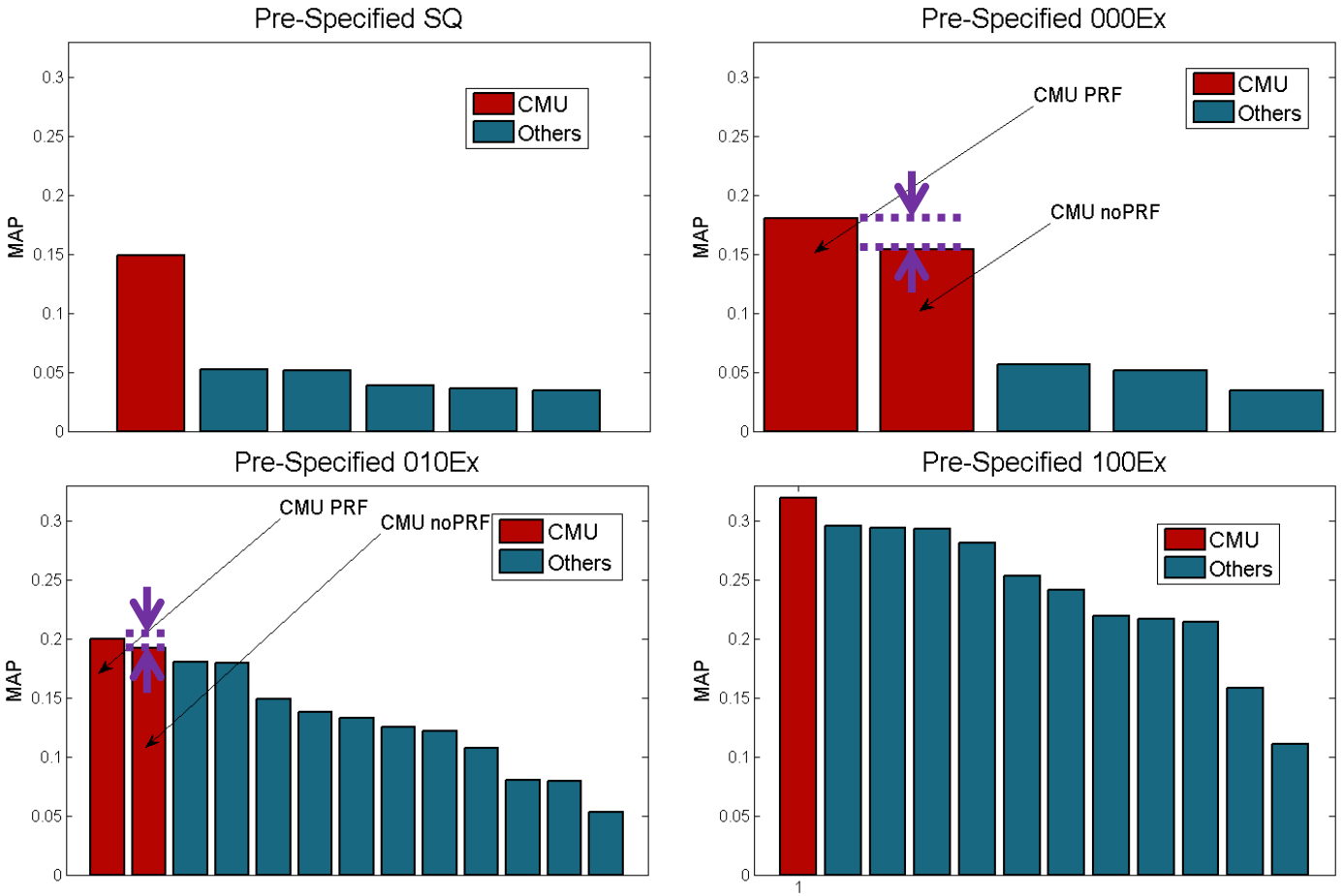
Recap of MED14 Results

Better semantics

Better features & fusion

Good reranking

Training done in < 16 minutes. Prediction done in < 5 minutes.



Performance on Pre-Specified events on MED14-Eval Full set. Adhoc results have similar trend. Red bars are CMU runs.

Boosting Performance with Reranking

- Reranking/Pseudo Relevant Feedback (PRF)
  - Assumes top ranked videos are correct, add them into training set and retrain
  - CMU used MultiModal PRF [1] in MED13, which was effective
- Problem with MMPRF
  - Existing methods assign either binary or predefined weights to the video, which may not faithfully reflect their latent importance.
  - We propose to **automatically learn the weights in a self-paced fashion**.

1. Jiang et al.. Zero-Example Event Search using MultiModal Pseudo Relevance Feedback. In ICMR 2014.

Self-Paced Reranking (SPaR)

- SPaR has a **mathematical formulation** which is **theoretically sound.**
- SPaR boosts our MMPRF by absolute 2%.

- Experiments on MEDTEST 13:

Table 1: MAP (× 100) comparison with the baseline methods across 20 Pre-Specified events.

Method	NIST's split
Without Reranking	3.9
Rocchio	5.7
Relevance Model	2.6
CPRF	6.4
Learning to Rank	3.4
MMPRF	10.1
SPaR	**12.9**

The results are **statistically significant** at the p-level of 0.05

Jiang et al., "Easy Samples First: Self-paced Reranking for Zero-Example Multimedia Search," in ACM MM, 2014.

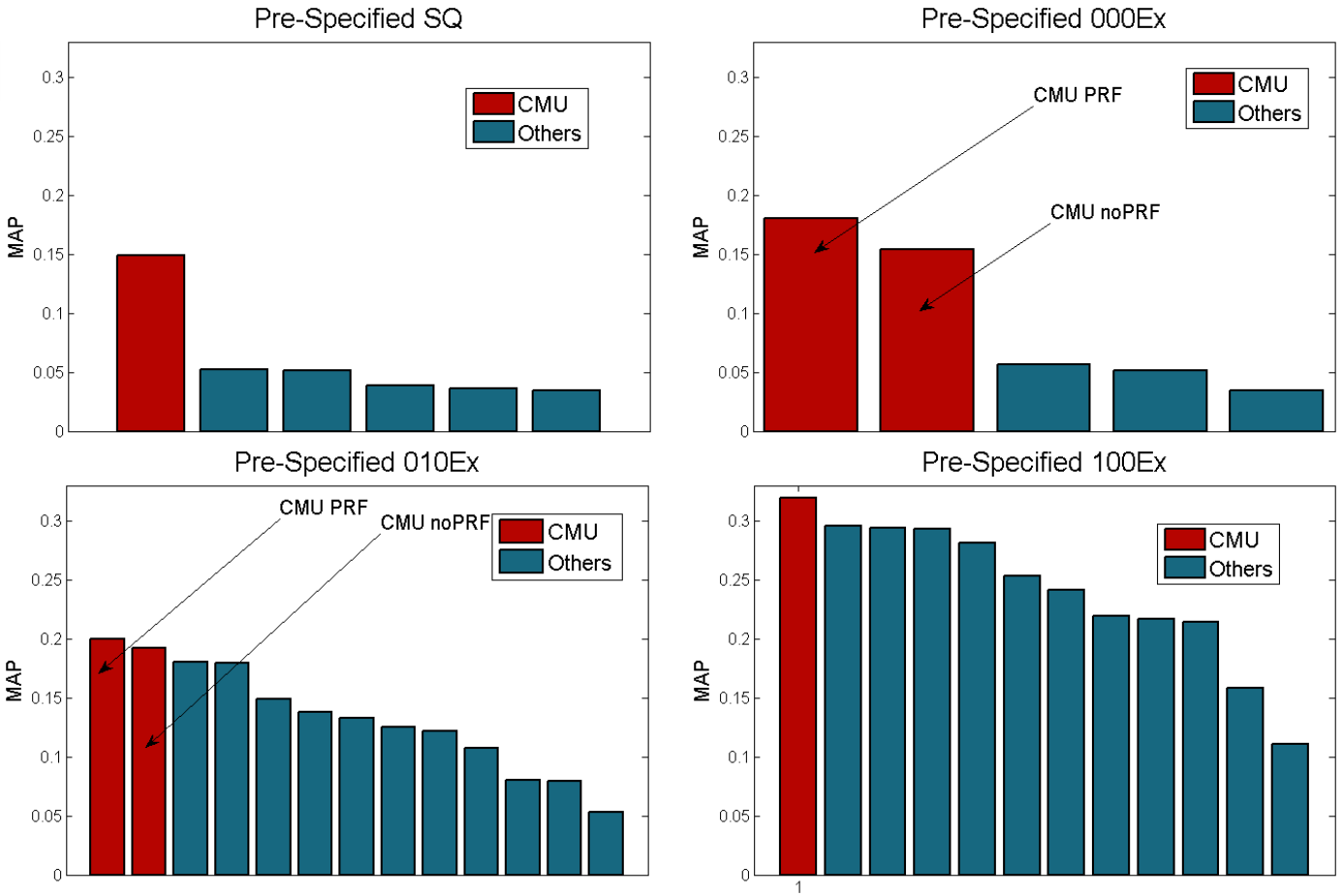
Recap of MED14 Results

Better semantics

Better features & fusion

Good reranking

Training done in < 16 minutes. Prediction done in < 5 minutes.



Performance on Pre-Specified events on MED14-Eval Full set. Adhoc results have similar trend. Red bars are CMU runs.

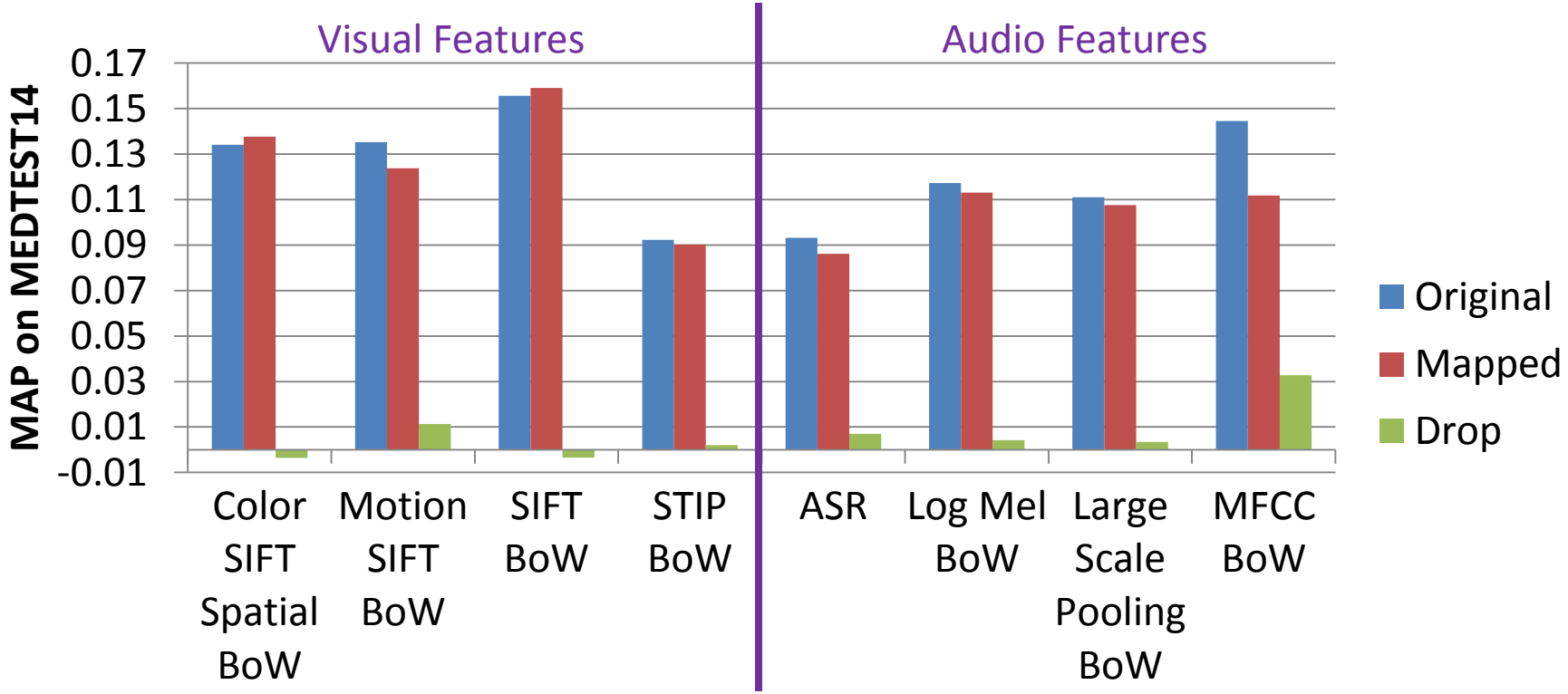
Carnegie Mellon University

Fast Training and Prediction

- Goal: create interactive MED system
  - Requires fast training and prediction
- Methodologies
  - Replaced **kernel classifiers** with **linear classifiers**
    - For χ^2 kernel-based features, use explicit feature map.
  - **Product Quantization (PQ)** for **compressing evaluation data** (I/O is the bottleneck in prediction!)
  - Created **hybrid CPU/GPU pipeline** to fully utilize CPUs and GPUs on COTs workstation

Explicit Feature Maps to "linearize" kernel-based features

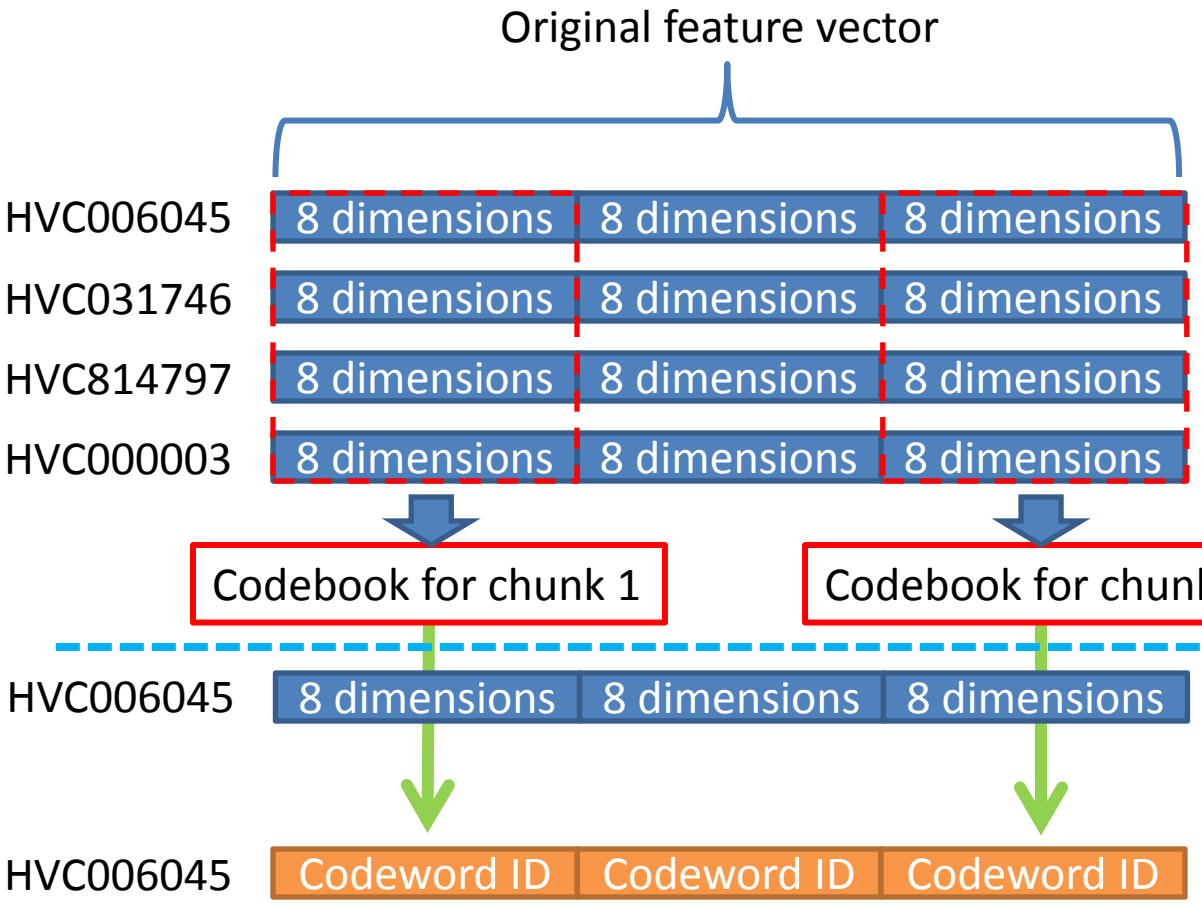
- Use Explicit Feature Maps (EFM, [1]) to map χ^2 kernel-based features to linear features.



GMM-based features given up because cannot apply EFM to RBF kernel-based features.

[1] Vedaldi, Andrea, and Andrew Zisserman. "Efficient additive kernels via explicit feature maps." *PAMI 2012.*

Product Quantization (PQ) for Compressing Evaluation Features



Step 0: chunk the original feature by every 8 (4x) dimensions for 32x (16x) compression.

Step 1: learn codebook with K-means for each chunk. K=256.

Step 2: map each chunk of a vector to a code word. Store only the code ID for each chunk.

8 dim. floating point: **32 bytes**, codeword ID: **1 byte**. **32x compression**!

Jegou, Herve, Matthijs Douze, and Cordelia Schmid. "Product quantization for nearest neighbor search." *PAMI 2011.*

Comparison in Performance and Timing before and after EFM & PQ

- Fusion and speed experiments on EFM, PQ features.

Runs (MEDTEST14)	MAP Performance		Timing (s) for 100Ex	
	100Ex	010Ex	EQG	ES
Original (no EFM, no PQ, with GMM features)	0.405	0.266	12150[1]	5430[1]
With EFM, PQ 32X, no GMM features	0.394	0.270	926	142
Improvement	-2.7%	1.5%	1940%	3823%

- No significant drop in performance after EFM and PQ, but speed increases significantly. **EQG: 19x, ES: 38x**.
- Event search for 1 event (**100 classifiers**): **142 seconds**.

[1]. Extrapolated timing from MED13 pipeline.

Summary of CMU's MED 2014 System

Carnegie Mellon University

Better semantics

- Semantic concepts have significantly improved
  - Fast to train, fast to predict and performs better than the low-level feature it was trained on
  - Vast improvement in 0Ex system

Better features & fusion

- Low-level feature and good fusion still key for 010Ex and 100Ex system
  - CMU improved trajectories, better DCNNs
  - Multistage Hybrid Late Fusion

Good reranking

- Self-Paced Reranking

Training done in < 16 minutes. Prediction done in < 5 minutes.

- Towards interactive MED systems
  - Explicit Feature Map & Product Quantization
  - Hybrid CPU, GPU pipeline