CMU-informedia @ TRECVID 2011 Semantic Indexing

Lei Bao[1,2], Shoou-I Yu[1], Alexander Hauptmann[1]

[1]Language Technologies Institute, Carnegie Mellon University

[2]Advanced Computing Research Laboratory, Beijing Key Laboratory of Mobile Computing and Pervasive Device, ICT, CAS

Outline

Feature Extraction

- Image-based feature: SIFT and CSIFT
- Video-based feature: MoSIFT
- Representation: Spatial bag-of-word

Training Classifier

- Kernel matrix pre-computation
- Sequential Boosting SVM

Fusing

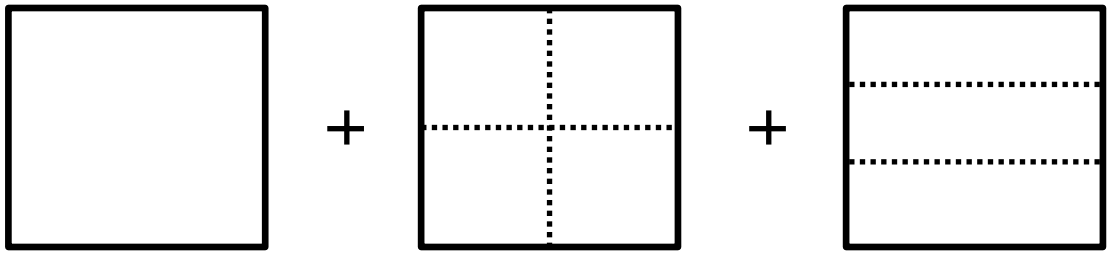
- Early fusion
- Multi-modal Sequential Boosting SVM

Feature Extraction

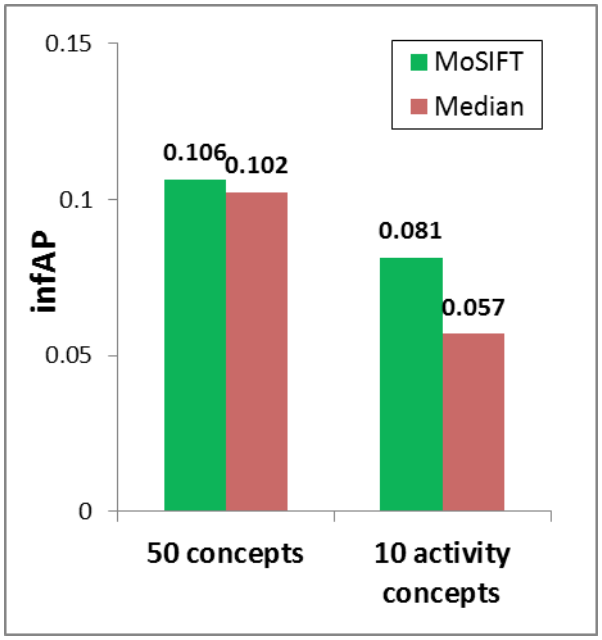
Table 1. SIFT[1], Color SIFT[1], and MoSIFT[2] raw features

	Image-based features				Video-based feature
	SIFT-HL	**CSIFT-HL**	**SIFT-DS**	**CSIFT-DS**	**MoSIFT**
Detector	Harris-Laplace	Harris-Laplace	Dense-Sampling	Dense-Sampling	Difference of Gaussian with optical flow filter
Descriptor	• SIFT • 128-d	• CSIFT • 384-d	• SIFT • 128-d	• CSIFT • 384-d	• SIFT plus optical flow • 256-d

- Generate codebook by K-Means
  - Size of codebook: 4096
- Spatial bag-of-word feature representation:
  - Soft voting: 10-nearest
  - Dimension: 4096*(1+2*2+1*3)=32,768



Performance of MoSIFT



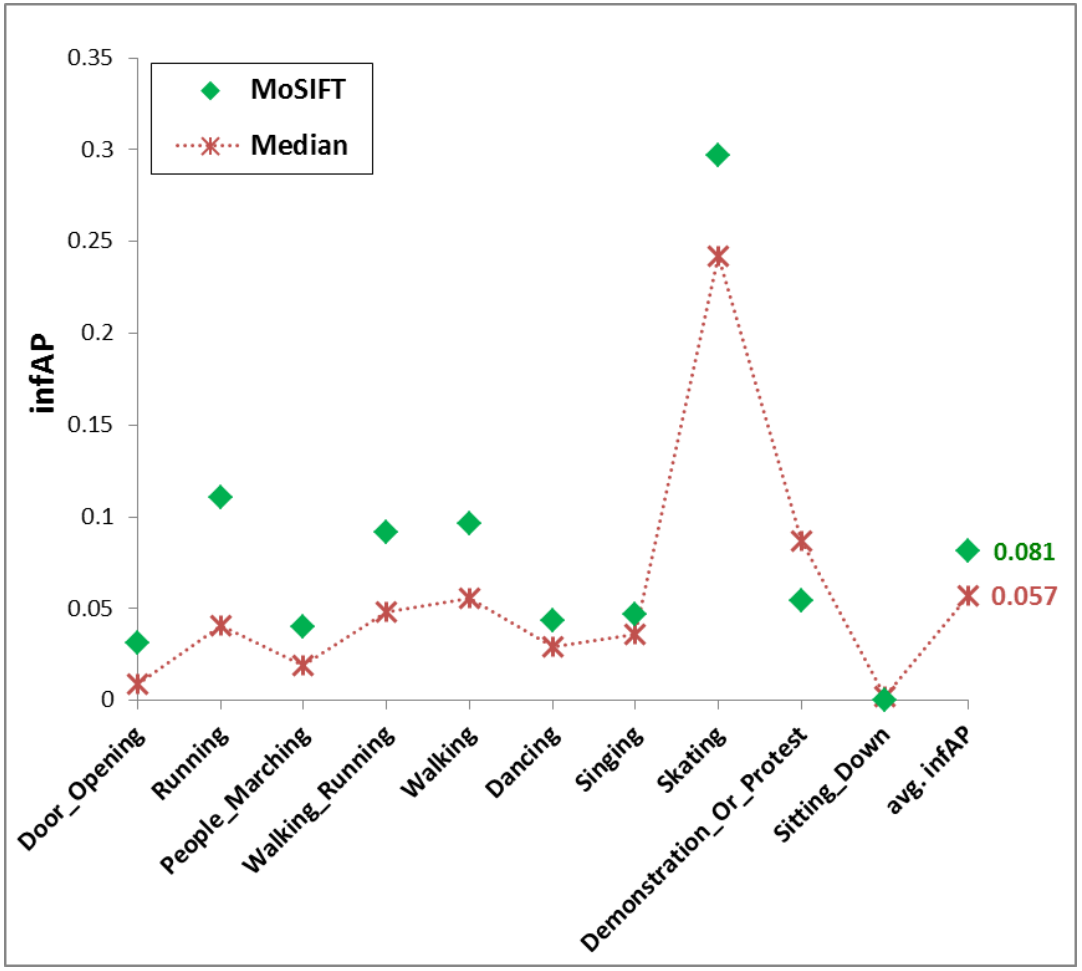


Fig. 1. Performance of MoSIFT feature

Performance of Early Fusion Feature

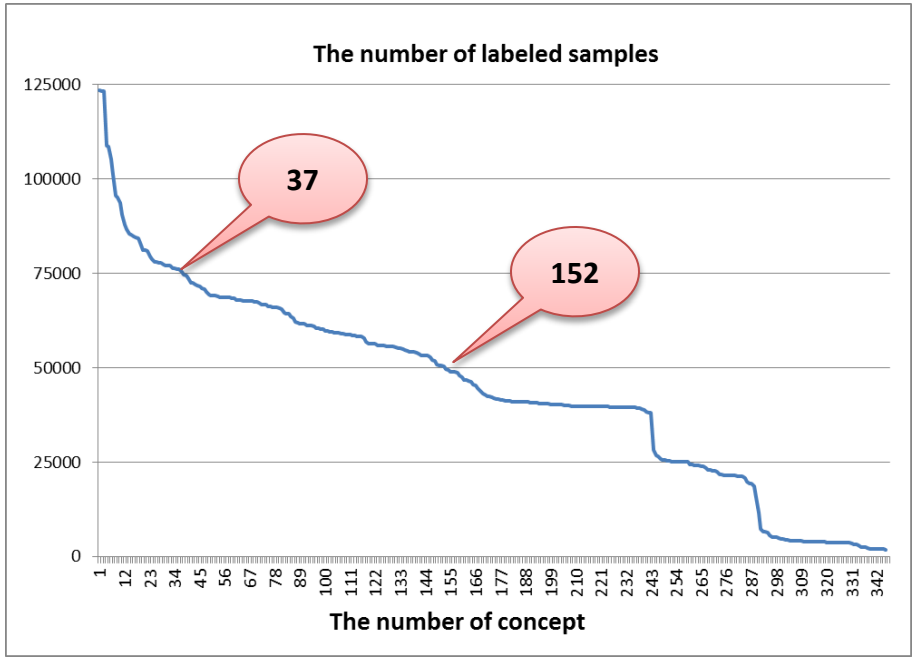
- **MoSIFT vs. SIFT and CSIFT:**
  - MoSIFT: describes the gradient and motion information of a video clip;
  - SIFT and CSIFT: describes the gradient and color information of a static image.
- **Harris-Laplace vs. Dense-Sampling**
  - Harris-Laplace provides meaning feature points but sometime it only can detect a few of feature points when the scene is simple;
  - Dense-Sampling provides enough points but it also involves a lot of noise.

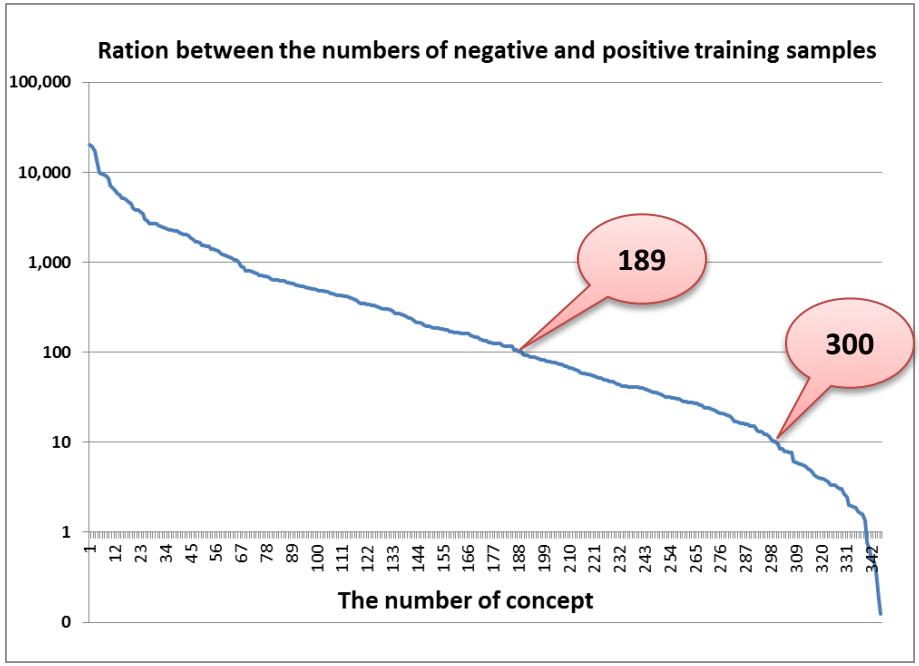
Table 2. Performance of early fusion features

	MoSIFT	SIFT_HL	CSIFT_HL	SIFT_DS	CSIFT_DS	Avg. infAP	Improvement
MoSIFT	√					0.106	0.0%
MoSIFT-SIFT-CSIFT	√	√	√			0.134	**26.4%**
MoSIFT-SIFT2-CSIFT2	√	√	√	√	√	0.141	**33.0%**

Training Classifier

- ***Task:*** train 346 concept detectors on annotated development set (over 260,000 shots), and predict them on evaluation set (over 130,000 shots).





***Challenge*:** Large-scale unbalanced classification problem!

Kernel Distance Pre-computation

- **Distance:**
  - Train: Chi-square distances between training examples;
  - Prediction: Chi-square distances between training and testing examples.
- **Reasons:** reduce computation cost
  - Train: Distances are repeatedly computed during cross-validation;
  - Predict: All of the 346 concepts share the same training and testing set;
  - Early fusion = weighted combination of distance matrix.
- **Tip:** Save the distance matrix as binary files to speed up the write/read process.

Sequential Boosting SVM

- **Sequential Boosting SVM**: train a sequence of SVM classifiers, and a limited and balanced training examples are boosted sampled for each classifier.
  - ***Large-scale***: divide a large-scale classification problem to several much smaller classification problems; loading the distance matrix to memory is durable;
  - ***Unbalance***: keep the balance of training examples in each small classification problem;
  - ***Performance***: boosted sampling enforces the further classifier to focus on the easily misclassified samples and boost the performance.

Sequential Boosting SVM

Bagging[3]

- Training examples for each classifier are generated by uniformly sampling with replacement.

Asymmetric Bagging[4]

- Sample all of the positive examples;
- Uniformly sample the same number of negative examples from all of the negative examples set to keep the balance of training examples.

Sequential Boosting

- Sample the most “important” examples for each small classifier;
- The examples that can be easily misclassified get high possibility to be sampled while examples that can be easily classified get low possibility.

Sequential Boosting SVM

Algorithm 1: Algorithm of Sequential Boosting SVM.

Input: positive example set $\mathbf{S}^+ = (x_1^+, y_1^+), \ldots, (x_{N^+}^+, y_{N^+}^+)$, where $y_i^+ = 1$; negative example set $\mathbf{S}^- = (x_1^-, y_1^-), \ldots, (x_{N^-}^-, y_{N^-}^-)$, where $y_i^- = 0$; SVM classifier $\mathbf{I}$; number of generated classifiers: $\mathbf{T}$; sample $\mathbf{K}^+$ positive examples and $\mathbf{K}^-$ negative examples in each iteration.

begin

$D_1^+(i) = 1/N^+$;

$D_1^-(i) = 1/N^-$;

for $t \leftarrow 1$ **to** T **do**

Sample:

- Sample positive example set $\mathbf{S}_t^+$ from $\mathbf{S}^+$ via distribution D_t^+, $|\mathbf{S}_t^+| = \mathbf{K}^+$;
- Sample negative example set $\mathbf{S}_t^-$ from $\mathbf{S}^-$ via distribution D_t^-, $|\mathbf{S}_t^-| = \mathbf{K}^-$;

Train SVM classifier: $C_t = \mathbf{I}(\mathbf{S}_t^+, \mathbf{S}_t^-)$;

Predict: $C^*(x_i) = \frac{1}{t}\sum_{p=1}^{t} C_p(x_i)$;

Update:

- $D_{t+1}^+(i) = \frac{D_t^+(i)}{Z_t^+} \times (1 - C^*(x_i^+))$, where Z_t^+ is a normalization factor (chosen so that D_{t+1}^+ will be a distribution);
- $D_{t+1}^-(i) = \frac{D_t^-(i)}{Z_t^-} \times (C^*(x_i^-))$, where Z_t^- is a normalization factor (chosen so that D_{t+1}^- will be a distribution);

end

Output: classifier $C^*(x_i) = \frac{1}{\mathbf{T}}\sum_{p=1}^{\mathbf{T}} C_p(x_i)$

Sequential Boosting vs. Asymmetric Bagging

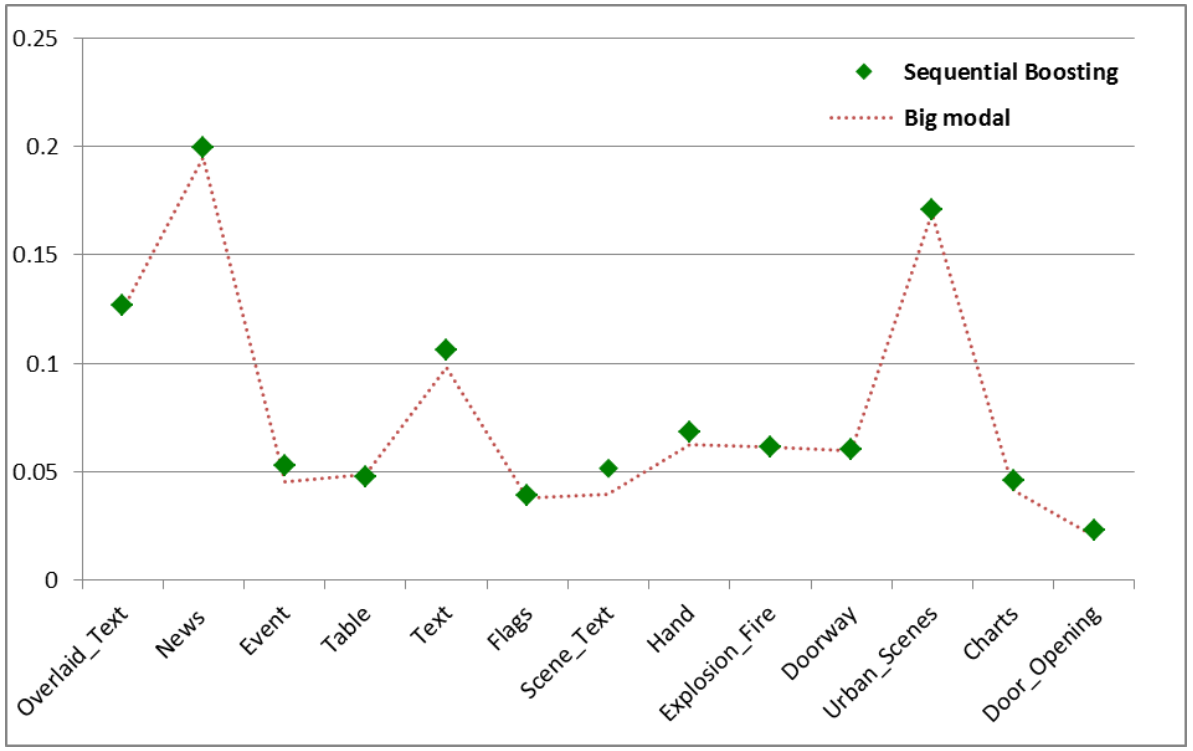
- MoSIFT-SIFT-CSIFT: early fusion of MoSIFT, SIFT-HL and CSIFT-HL
- # Bagging: 10
- # Sampled positive examples in each iteration: [0, 1000]
- # Sampled negative examples for each iteration = # Sampled positive examples
- Evaluation metric: avg. infAP

Table 3. Sequential Boosting vs. Asymmetric Bagging

# Negative examples	# Concepts	Asymmetric Bagging	Sequential Boosting	Improvement
[0, +∞)	50	0.125	0.132	6.05%
[0, 25,000]	13	0.078	0.080	2.88%
(25,000, 50,000]	18	0.132	0.137	4.13%
[50,000, +∞)	19	0.150	0.163	8.76%

Sequential Boosting vs. Big SVM modal

- Choose 13 concepts which the numbers of positive and negative samples are both less than 25,000;
- Avg. infAP of Big SVM: 0.077
- Avg. infAP of Sequential Boosting SVM: 0.081 (+4.89%)



Fusion

- **Early fusion:** weighted fuse kernel distance matrixes of different features.
  - **SIFT-HL-DS**: averagely fuse distance matrixes of SIFT-HL and SIFT-DS;
  - **CSIFT-HL-DS**: averagely fuse distance matrixes of CSIFT-HL and CSIFT-DS;
  - **MoSIFT-SIFT-CSIFT**: averagely fuse distance matrixes of MoSIFT, SIFT-HL and CSIFT-HL;
  - **MoSIFT-SIFT2-CSIFT2:** averagely fuse distance matrix of MoSIFT, SIFT-HL-DS and CSIFT-HL-DS.
- **Multi-modal Sequential Boosting SVM:**
  - The examples which are misclassified in current layer have high probabilities to be correctly classified in next layer if a different feature is used.

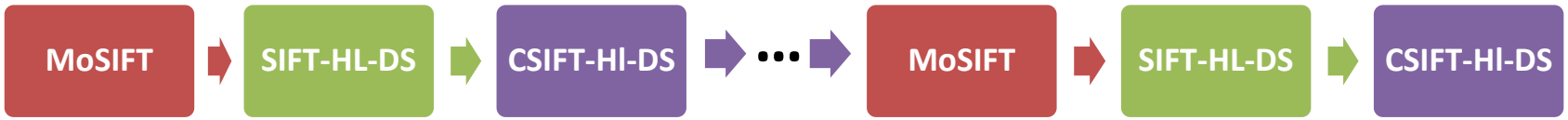


Fig. 2. Multi-modal Sequential Boosting SVM

Submissions

Run_ID	Avg. infAP	Name	Description
CMU_1	0.1064	MoSIFT	• MoSIFT • 10-layer Sequential Boosting SVM
CMU_2	0.1337	MoSIFT-SIFT-CSIFT	• MoSIFT-SIFT-CSIFT • 10-layer Sequential Boosting SVM
	0.1407	MoSIFT-SIFT2-CSIFT2	• MoSIFT-SIFT2-CSIFT2 • 10-layer Sequential Boosting SVM
CMU_3	0.1458	MoSIFT-SIFT2-CSIFT2 multimodal	• MoSIFT, SIFT-HL-DS, CSIFT-HL-DS • 20-layer Multi-modal Sequential Boosting SVM
CMU_4	0.1464	MoSIFT-SIFT2-CSIFT2 latefusion	• Averagely fused the prediction scores from MoSIFT-SIFT2-CSIFT2 and MoSIFT-SIFT2-CSIFT2_multimodal

Lessons Learned

- **Features:**
  - MoSIFT feature works well for activity concepts;
  - MoSIFT, SIFT and Color SIFT features are complementary visual features.
- **Classification**:
  - Pre-computing kernel distance matrix reduced computation time a lot;
  - Sequential Boosting SVM is a good solution to deal with the large-scale unbalanced classification problem.
- **Fusion**:
  - Sequential Boosting SVM can be successfully extended to handle multi-modal problem.

Future work

- **Features:**
  - Video-based feature: STIP
  - Audio feature: MFCC
- **Classification:**
  - Optimize the number of classifiers in Sequential Boosting SVM
- **Fusion:**
  - Optimize the feature path in Multi-modal Sequential Boosting SVM
- Others**:**
  - Explore the relationship of concepts.

References

[1] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1582–1596, 2010.

[2] M.-Y. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos, 2009.

[3] L. Breiman and L. Breiman. Bagging predictors. In Machine Learning, pages 123–140, 1996

[4] D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans. Pattern Anal. Mach. Intell., 28:1088–1099, July 2006.


Q&A?