論文 / 著書情報
Article / Book Information

論題	映像検索技術の最前線
著者	篠田浩一
出典	第18回画像センシングシンポジウム講演論文集, Vol. , No. , OS3-02-1-4
発行日 / Issue date	2012, 6
Note	第18回画像センシングシンポジウム講演論文集より転載

映像検索技術の最前線

篠田浩一
（東京工業大学）

参考文献

1. http://trecvid.nist.gov/
2. 佐藤真一, ``映像内容解析におけるTRECVIDの取組み,''
   電子情報通信学会誌, vol. 91, no. 1, 55-59, 2008.
3. 井上中順, 篠田浩一, ``映像の高性能なセマンティックインデクシング
   を目指して,'' 信学技報, vol. 111, no. 353, pp. 89-94, 2011.

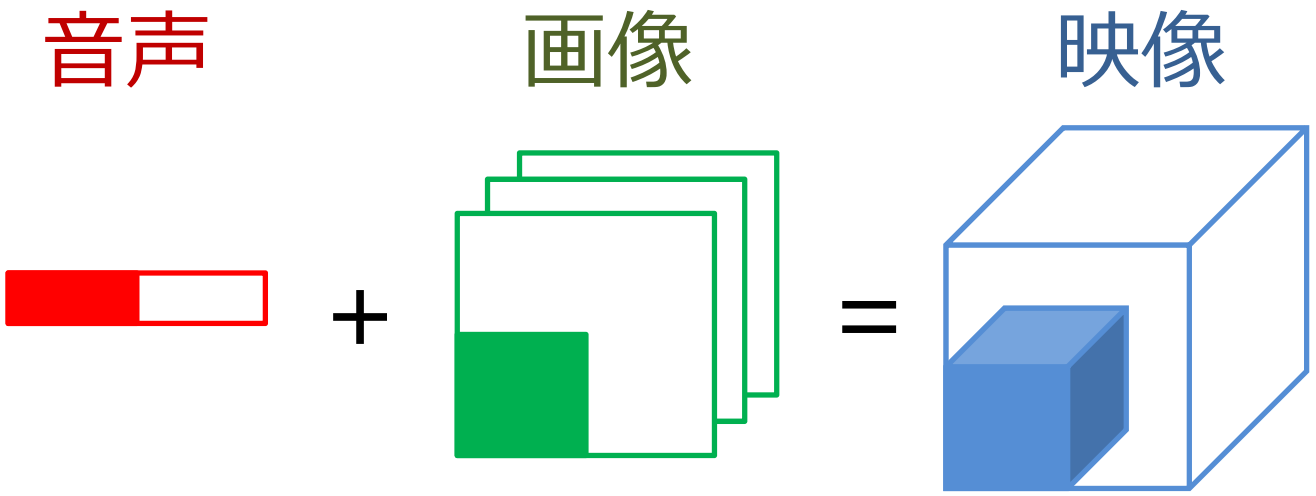
映像検索

タグのない映像コンテンツから情報を抽出

Content-Based Information Retrieval; CBIR

- 意味インデクシング
  - 目的はテキスト検索と同じ
- 類似映像検索
  - コピー検出(知的財産権管理)
- 異常検知
  - 監視カメラ映像、セキュリティ目的

機械学習によるアプローチ



データ収集、計算のコスト、セマンティックギャップ

- 特定用途に限定
- 複数の研究機関の連携

TRECVID
(TREC Video Retrieval Evaluation)

Text REtrieval Conference (TREC)より2001年に派生

主催： 米国標準技術局 NIST
(National Institute of Standard and Technology)

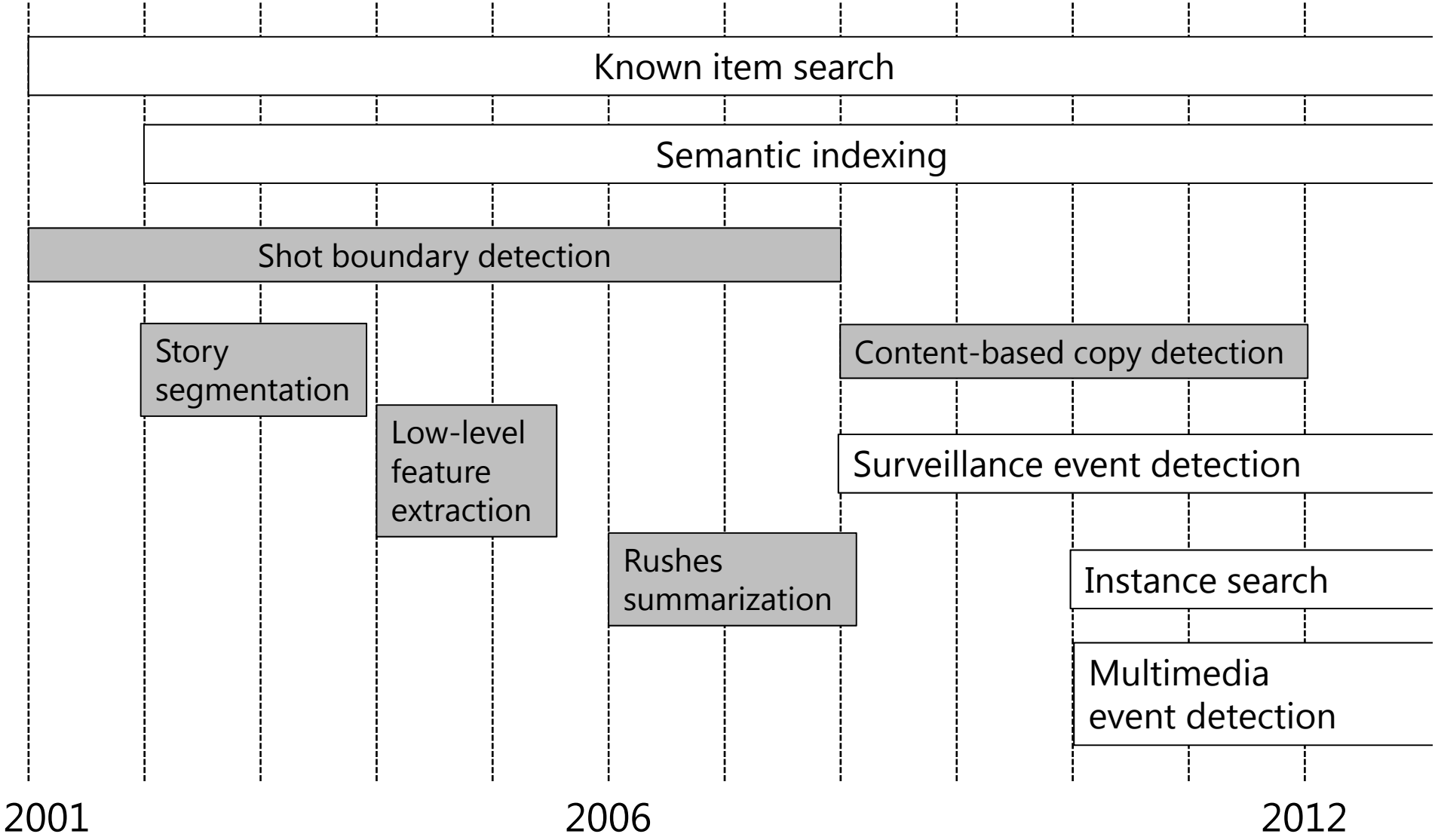
目的：映像コンテンツの解析・検索技術の高度化

形態：競争型国際プロジェクト, 世界中の研究機関が参加
クローズド(参加者のみ出席可能)

ホームページ： http://trecvid.nist.gov

- 大規模データベースを利用可能(権利問題を回避)
- 技術の比較が容易、全体の進歩がスピードアップ
- データベース作成を分担
- (参加者には)勝ち負けがはっきりわかる

TRECVIDのタスク



TRECVID 2011

世界中から66チームが参加。日本からは12チーム。
世界的に有名な企業、大学が参加。

年間スケジュール　※タスク・年度により変動あり

2月：参加者募集
～4月：タスクの詳細決定
～6月：開発データの構築・配布
～8月：Dry Run (結果提出のシミュレーション)
～9月：評価データの配布
～10月：各機関が結果を提出
～11月：NISTから各機関に評価結果通知
12月：ワークショップ(今年度結果と次年度タスクについて議論)

Semantic Indexing (SIN)

目的

映像のショットから意味をもつ事象(高次特徴)を検出

オブジェクト、シーン

例: Flower, Bus, Dancing,
Speaking, Car Racing


Human


Boat_Ship


Cityscape


Night_Time

タスクの特徴と位置づけ

入門的かつ中心的タスク

一般物体認識(静止画)の拡張

データベース

IACC

(Internet Archive videos with Creative Commons licenses)

600時間のビデオデータ

IACCデータベース:19,772本のwebビデオクリップ計600時間

ショット数：開発用 264,673，テスト用137,327

正解ラベル: 高次特徴346種類

選択基準：汎用性, 有用性, 過去に使用したもの

参加者が分担して開発データにラベル付け

テスト用ショットの順位付きリスト(上位2,000個)を提出

1チーム最大４つのRunを提出可能

評価基準: Average Precision (AP)

$$\frac{1}{K}\sum_{k=1}^{K} p(k)$$

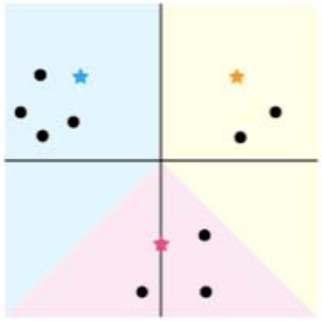
k: リスト中の順位, $p(k)$: 1位からk位までのprecision

※ショットの任意の位置に１回でも出現していれば正例

基本：Bag of Words (BoW)

Shot

Local features



Histogram

SVM

※各ショットの
キーフレーム
のみを利用

新しい動き(1)：頑健性

少ないデータをいかに有効に使うか

(＝いかに頑健にするか)

- もっと特徴を！

  SIFT, Color SIFT, SURF, HOG, GIST, Dense特徴
- マルチモーダル

  画像だけでなく、音響特徴も ⇨ 歌、ダンス、自動車など
- マルチフレーム

  キーフレームだけではなく、複数フレーム
- ソフトクラスタリング

  量子化誤差の影響を軽減

新しい動き(2)：高速化

かといって、計算には時間がかかる

- 近似アルゴリズム
- 複数CPUによる並列計算
- Graphical Processing Unit (GPU)の利用

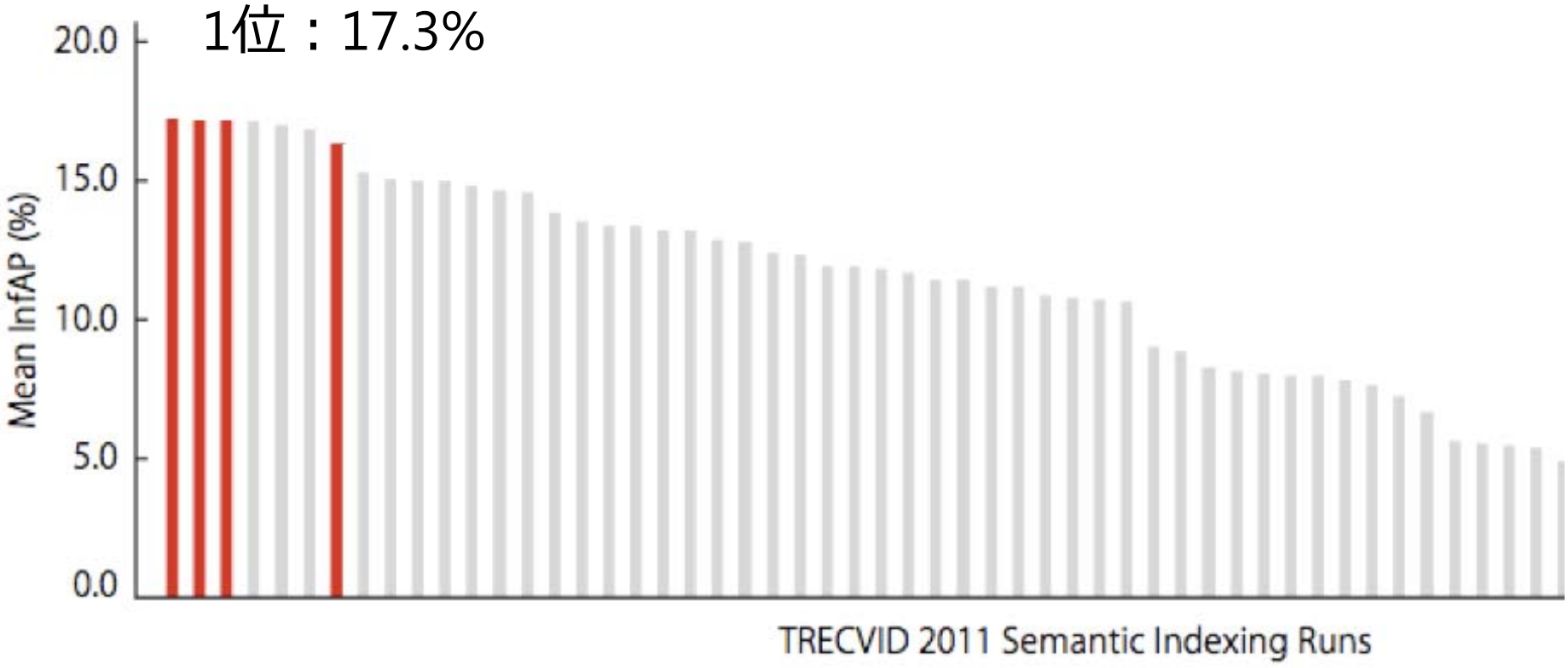
計算資源の確保、計画性が重要

エントリ56チーム中、結果提出は28チーム

期待したほど効果がないもの

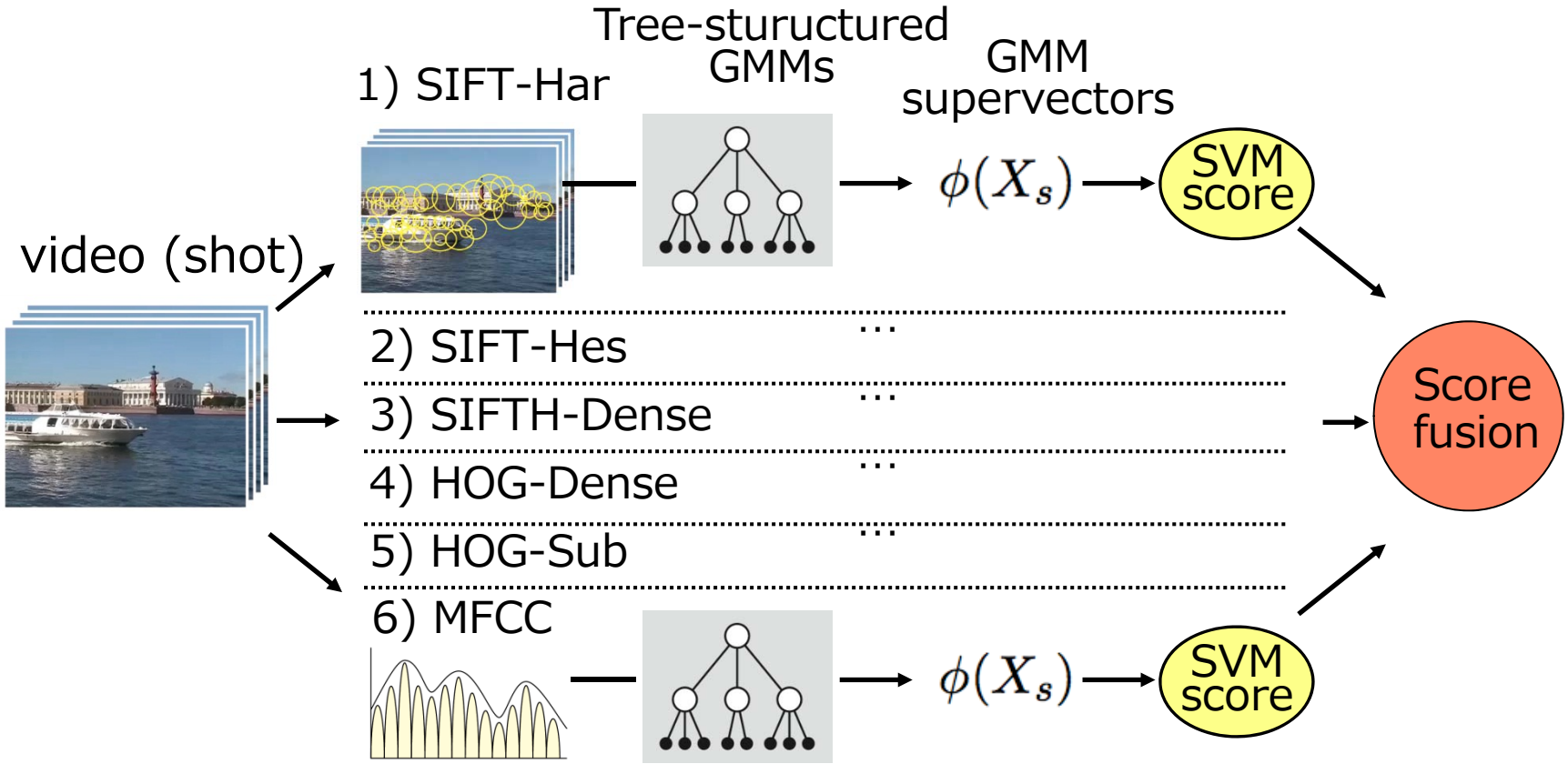
- 色ヒストグラムなどの大局的特徴
  局所的特徴のみで十分（相補的でない）
- 音声認識結果、OCR結果
  役立つケースが少ない／満足な性能に達しない
- オブジェクトの位置
  位置検出に失敗／位置をもたない特徴が多い
- 高次特徴の共起関係などのコンテキスト
  学習データの不足

結果



※ Mean InfAP: 全高次特徴のAPの平均(推定量)

1位チームのフレームワーク



※上位チームの間はさほど違いがない

1位チームの特色

- 6種類の相補的な画像特徴量

  SIFT-Har, SIFT-Hes, SIFTH-Dense, HOG-Dense, HOG-Sub
- マルチモーダル

  音響特徴: Mel-Frequency Cepstral Coefficient (MFCC)
- マルチフレーム

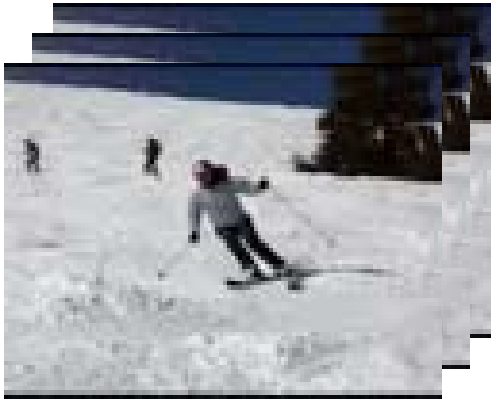
  毎フレーム、2フレーム毎、2秒に1フレーム
- ガウス混合モデル(Gaussian Mixture Model; GMM)
  - 確率論に基づくソフトクラスタリング(BoWの拡張)
  - 少量データに頑健なパラメータ推定: MAP適応
  - GMMとSVMの組み合わせ: GMM-Supervector SVM
- 近似アルゴリズムによる高速化

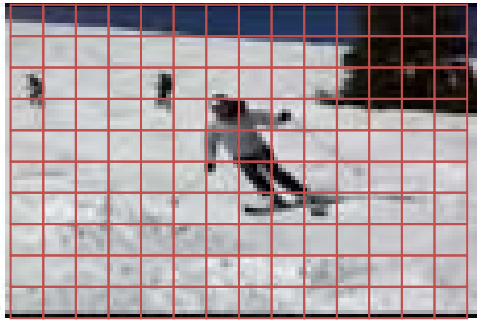
  GMMにおける効率的な分布サーチ
  (BoWにおけるコードワードのサーチに相当)

複数の画像局所特徴

1) SIFT-Har
  - Harris-affine検出器 (Harrisのコーナー検出の拡張)
  - マルチフレーム (1フレームおき)
2) SIFT-Hes
  - Hessian-affine検出器
  - マルチフレーム (1フレームおき)
3) SIFTH-Dense
 - SIFT + Hue histogram
 - 30,000サンプル(キーフレーム)
4) HOG-Dense
 - 32次元のHOG特徴量
 - 10,000サンプル(キーフレーム)
5) HOG-Sub
 - 時間的差分画像から特徴抽出
 - 動きを検出

※ 画像特徴はそれぞれ主成分分析(PCA)で32次元に圧縮

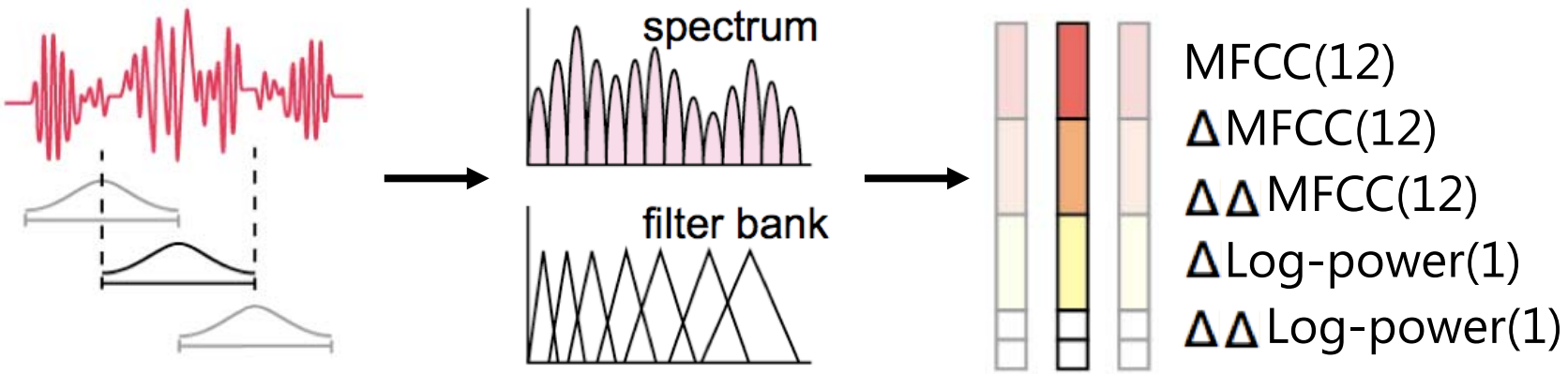




音響特徴: MFCC

Mel-frequency cepstrum coefficients

音声認識、音響イベント検出で良く用いられる



Gaussian Mixture Model (GMM)

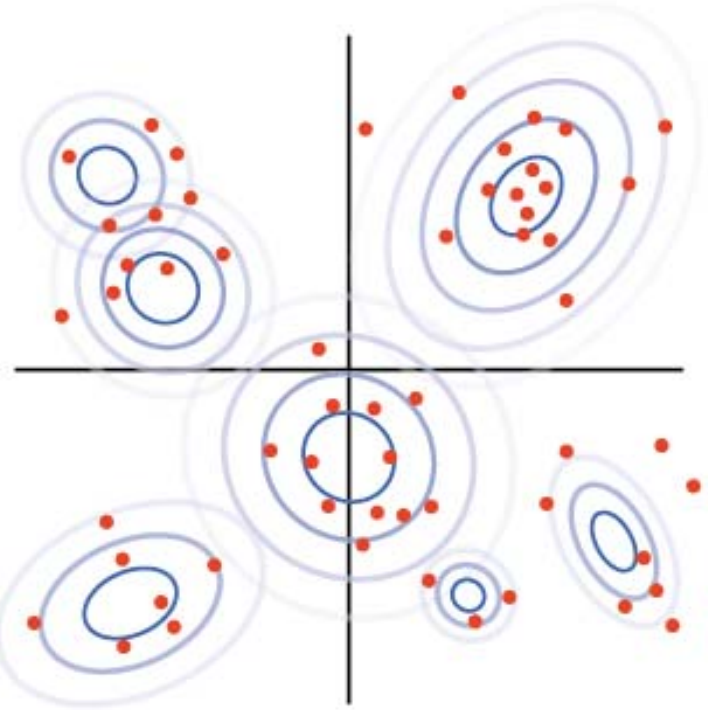
正規分布の重み付け和

$X_{\mathrm{F}} = \{x_i\}_{i=1}^{n}$ ：入力特徴

$$p(x|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

w_k: 混合成分kの重み係数 ($\sum w_k = 1$)

μ_k, Σ_k:混合成分kの平均ベクトル、
共分散行列



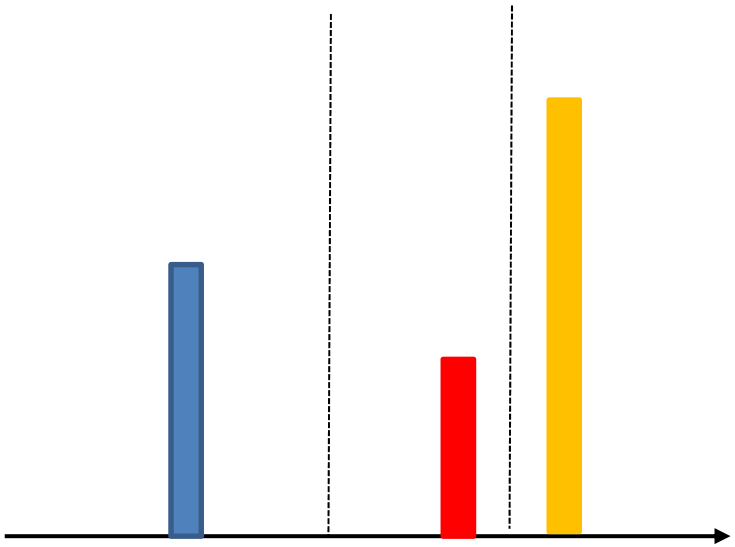
学習時：各高次特徴毎にGMMを作成

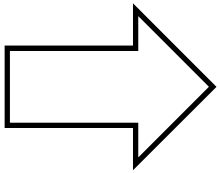
認識時：入力ショットのGMMを作成し、
各高次特徴GMMとの「距離」を用いて検出

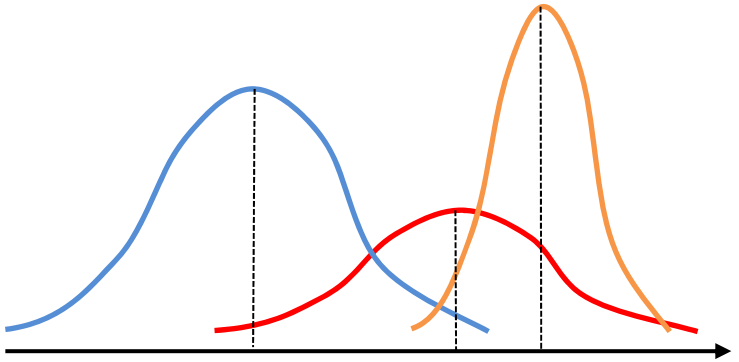
GMMはBoWの拡張

BoW

GMM







コードベクトル ⇒ 正規分布, 中心(平均)も更新

ヒストグラム ⇒ 混合重み係数の分布

Maximum A Posteriori (MAP)適応

- GMMの平均ベクトルの推定手法
- 事前知識を確率分布（事前分布）で表現

1. すべての学習データを用いたEMアルゴリズムで
   全高次特徴に共通のGMMを作成
   Universal background model (UBM)

2. UBMを初期モデルとし各GMMの平均ベクトルをMAP推定
   平均ベクトルの事前分布：UBMの対応する分布

最尤推定に比べ、少量のデータでも高い性能

GMM Supervector

平均ベクトルを結合→GMM supervector

$$\phi(X_{\mathrm{F}}) = \begin{pmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \\ \vdots \\ \tilde{\mu}_K \end{pmatrix}$$

where

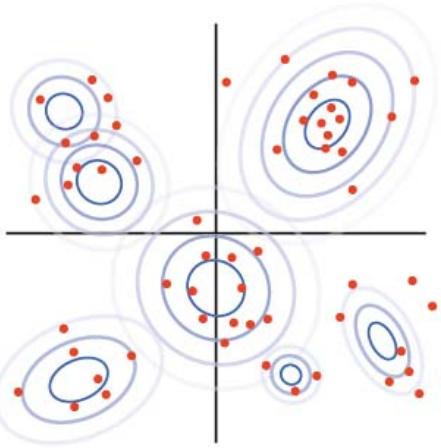
$$\tilde{\mu}_k = \sqrt{w_k^{(U)}}(\Sigma_k^{(U)})^{-\frac{1}{2}}\hat{\mu}_k$$

normalized mean

$X_s = \{x_i\}_{i=1}^{n_s}$

UBM

MAP adaptation



$\phi(X_{\mathrm{F}})$

GMM supervector

スコア融合

1. 個々の特徴量を用いた識別

RBF-kernel のサポートベクターマシン(SVM)

$$k(X_{\mathrm{F}}, X'_{\mathrm{F}}) = \exp\left(-\gamma \|\phi(X_{\mathrm{F}}) - \phi(X'_{\mathrm{F}})\|_2^2\right),$$

2. SVMスコアの重み付き和を用いて検出

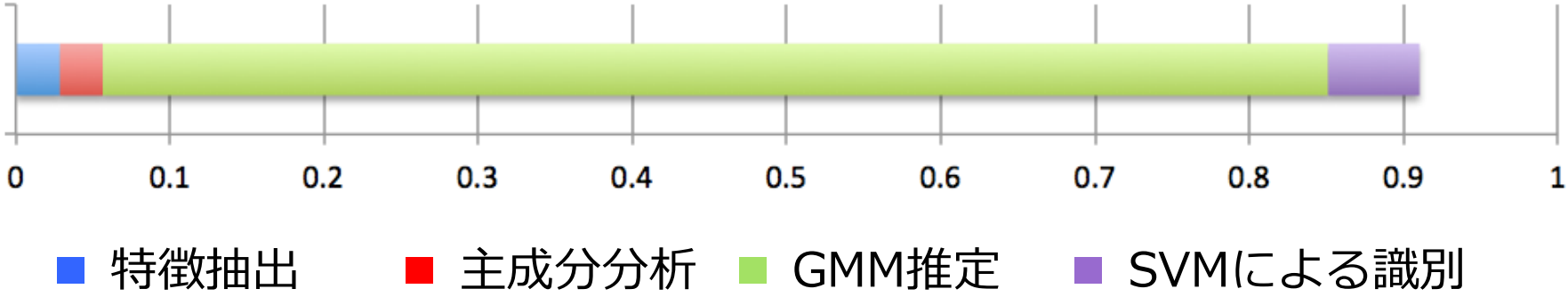
$$f(X) = \sum_{\mathrm{F}\in\mathcal{F}} \alpha_{\mathrm{F}} f_{\mathrm{F}}(X_{\mathrm{F}}), \quad 0 \le \alpha_{\mathrm{F}} \le 1, \quad \sum_{\mathrm{F}} \alpha_{\mathrm{F}} = 1$$

where $\mathcal{F} = \{\text{SIFT-Har}, \text{SIFT-Hes}, \text{SIFTH-Dense}, \text{HOG-Dense}, \text{HOG-Sub}, \text{MFCC}\}$

※重み係数は高次特徴ごとに開発セットで最適化

計算コスト

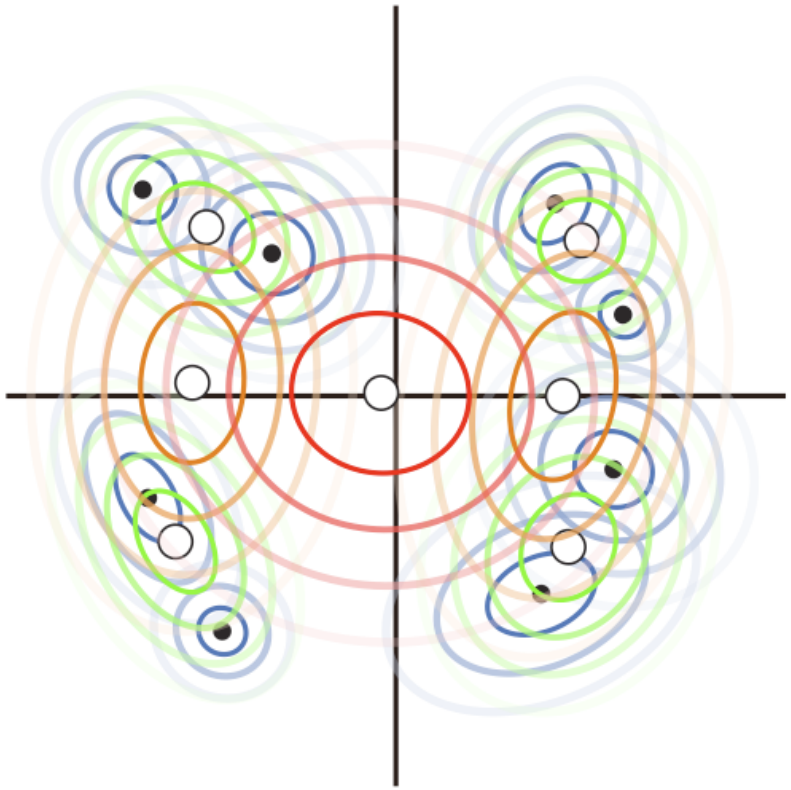
HOG-Dense特徴を用いた検出における計算時間 (sec)

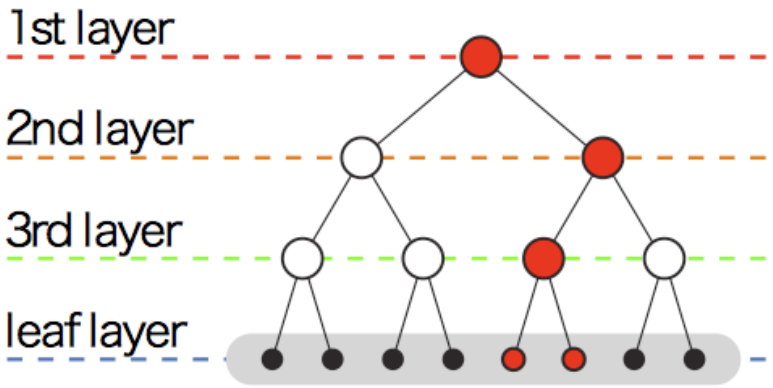


GMM推定(MAP適応)の高速化が必要

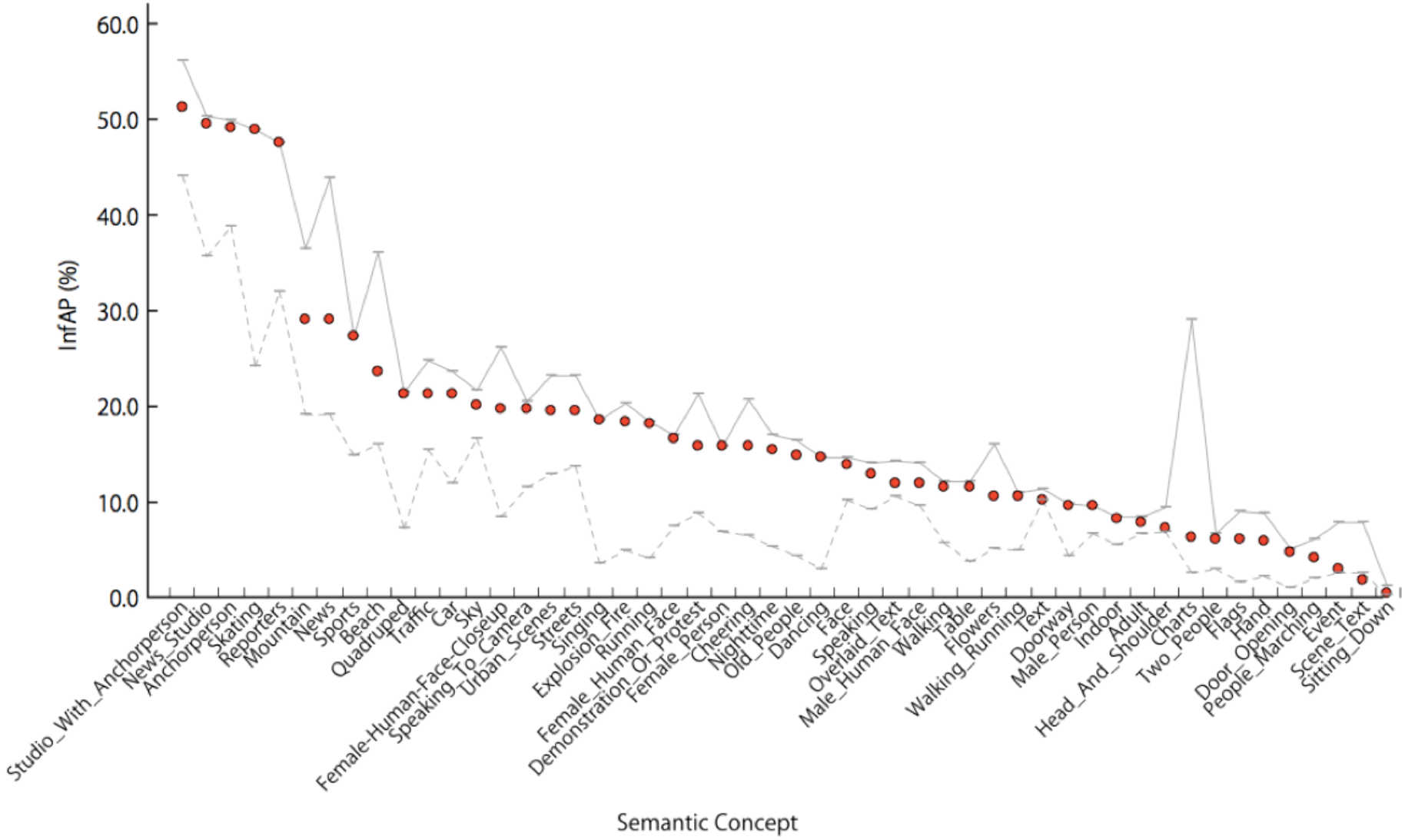
木構造GMMによる高速化

入力特徴ベクトルがどの分布にどの程度属するか
(BoW: どのコードに割り当てるか)

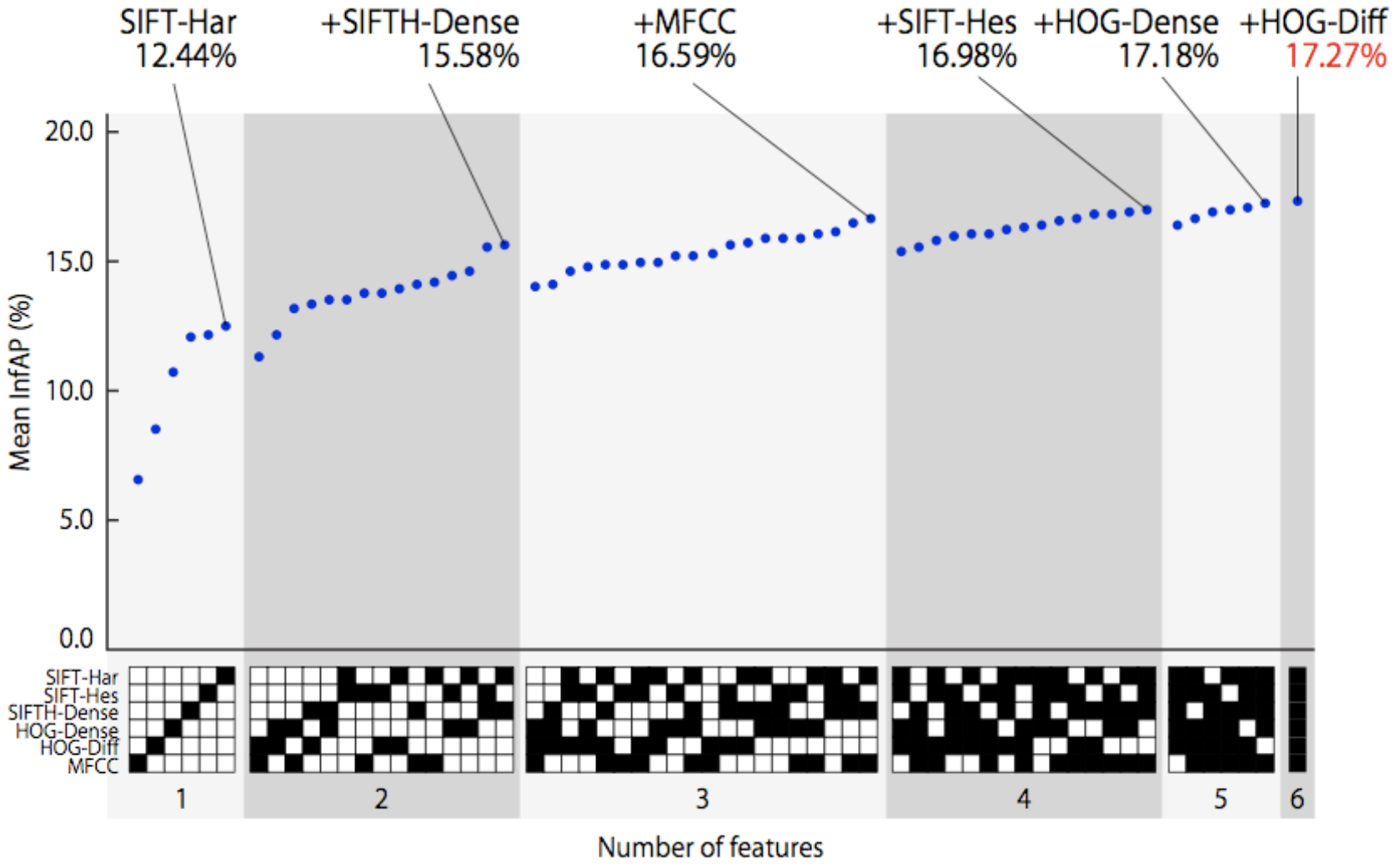




高次特徴ごとのAP


60.0
50.0
40.0
30.0
20.0
10.0
0.0
InfAP (%)
Studio_With_Anchorperson
News_Studio
Anchorperson
Skating
Reporters
Mountain
News
Sports
Beach
Quadruped
Traffic
Car
Sky
Female-Human-Face-Closeup
Speaking_To_Camera
Urban_Scenes
Streets
Singing
Explosion_Fire
Running
Female_Human_Face
Demonstration_Or_Protest
Female_Person
Cheering
Nighttime
Old_People
Dancing
Face
Speaking
Overlaid_Text
Male_Human_Face
Walking
Table
Flowers
Walking_Running
Text
Doorway
Male_Person
Indoor
Adult
Head_And_Shoulder
Charts
Two_People
Flags
Hand
Door_Opening
People_Marching
Event
Scene_Text
Sitting_Down
Semantic Concept

特徴量


SIFT-Har
12.44%
+SIFTH-Dense
15.58%
+MFCC
16.59%
+SIFT-Hes
16.98%
+HOG-Dense
17.18%
+HOG-Diff
17.27%
20.0
15.0
10.0
5.0
0.0
Mean InfAP (%)
SIFT-Har
SIFT-Hes
SIFTH-Dense
HOG-Dense
HOG-Diff
MFCC
1
2
3
4
5
6
Number of features

Multimedia Event Detection (MED)

目的

映像から一般的なイベントを検出

例: Batting a run in

Making a cake

タスクの特徴と位置づけ

SINよりも長く複雑な事象

スポーツ動画からのハイライト 検出の発展

データベース

HAVIC：ホームビデオ約2000時間

Linguistic data consortium (LDC)が提供

Multimedia Event Detection (MED) (2)

- 2010年に開始(まだ始まったばかり)
- 2011年は18チームが参加(うち日本から5チーム)
- 米国防総省 The Intelligence Advanced Research Projects Activity(IARPA)のプロジェクト、Automated Low-Level Analysis and Description of Diverse Intelligence Video (ALADDIN) と連携

HAVICデータベース

- 平均長2分程度のビデオクリップ3488個
- 各イベントにつき100個(学習と認識で半々に)

2010 (3イベント)	2011 (10イベント)	
Assembling a shelter	Birthday party	Making a sandwich
Batting a run in	Changing a vehicle tire	Parade
Making a cake	Flash mob gathering	Parkour
	Getting a vehicle unstuck	Repairing an appliance
	Grooming an animal	Working on a sewing project

タスク：

各イベントの検出を試み、任意個数の候補を提出

評価基準：

- Missed Detection Probability P_{miss}

  1 – Recall

- False Alarm Probability P_{FA}

  誤報数 / イベントのないクリップ数

- Normalized Detection Cost (NDC)

  上２つを組み合わせたスコア

$$NDC = \frac{Cost_{miss}\, P_{miss} P_{target} + Cost_{FA}\, P_{FA}\left(1 - P_{target}\right)}{MIN\left(Cost_{miss} P_{target} + Cost_{FA}\left(1 - P_{target}\right)\right)}$$

$Cost_{Miss} = 80$
$Cost_{FA} = 1$
$P_{target} = 0.001$

主な取り組み

- SIN手法がベース
  多種類の特徴量＋BoW+SVM
- 時空間の局所特徴
  STIP (Space-time interest point)など
- 高次特徴量間の相関をモデル化(セマンティックモデル)
  効果はあまりない←データ量が少なすぎる？
- 音声認識、OCRの利用
  効果はあまりない←SINと同様の理由

今後の展開

- 映像検索における有望なタスク・有効な方法論がはっきりしてきた
- データ量がまだまだ足りない。
  - 協調的アノテーション
- 単語レベル(SIN)から文レベル(MED)へ
  - 映像のコミュニケーションモデル
  - コンテキストの利用(位置、音声、共起、etc)
- より効果的な特徴量は？

TRECVID 以外の取り組み

LSCOM, MediaEval