Bag-of-Frames と時空間特徴量を用いた Semantic Indexing Taskへの取り組み

PRMU 2011
電気通信大学 総合情報学専攻
下田 保志
野口 顕嗣
柳井 啓司
2011年2月18日

アウトライン

- はじめに
  - 関連研究
- 手法
  - 時空間特徴抽出手法
  - Bag of Framesによる全フレーム認識
  - キーフレームの利用
- 実行環境
- 結果
  - 反省点
- おわりに

はじめに

- 国際映像処理ワークショップTRECVID
  - 膨大な動画データを利用
  - 課題の提示、結果の公表
- Semantic Indexing Task
  - 動画中の概念を認識し、概念に沿った動画から順に索引付けを行うタスク
  - 2010年から概念カテゴリが20種類から130種類に増加
    - 公式の平均適合率の算出は30種類の概念

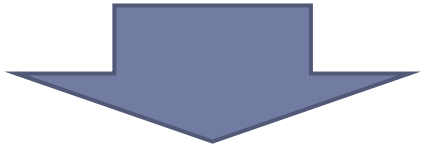
関連研究

- TRECVIDにおけるマルチフレーム認識
  - MediaMillチーム[2]
    - 複数のフレームから取得した特徴量を統合
  - 東京工業大学チーム[3]

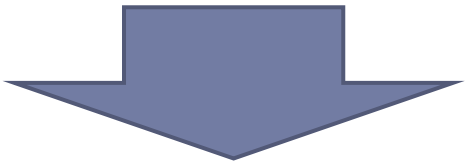
- 時空間特徴と視覚特徴の統合
  - Liu[4]らの研究
    - Adaboostによる統合

方針

- 2010年からTRECVIDのデータセットが更新
  - Webから収集した動画に

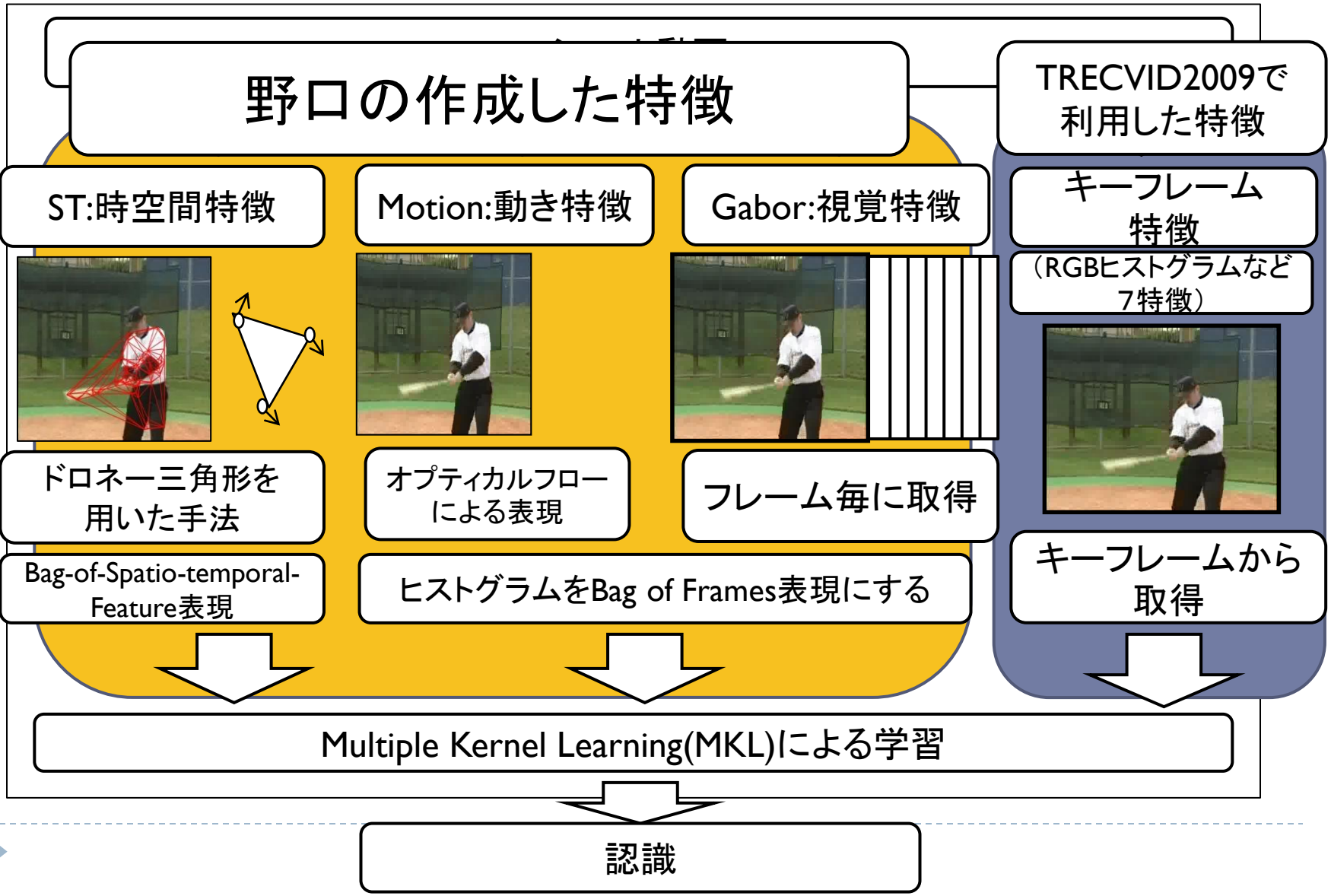


- 野口[1]によるWeb動画認識で利用した特徴
  - ドロネー三角形を利用した時空間特徴など
- 本チームがTRECVID　2009で利用した特徴
  - キーフレームから取得したカラーヒストグラムなど



Multiple Kernel Learningで統合

動画認識の概要


野口の作成した特徴
TRECVID2009で利用した特徴
ST:時空間特徴
Motion:動き特徴
Gabor:視覚特徴
キーフレーム特徴
（RGBヒストグラムなど7特徴）
ドロネー三角形を用いた手法
オプティカルフローによる表現
フレーム毎に取得
キーフレームから取得
Bag-of-Spatio-temporal-Feature表現
ヒストグラムをBag of Frames表現にする
Multiple Kernel Learning(MKL)による学習
認識

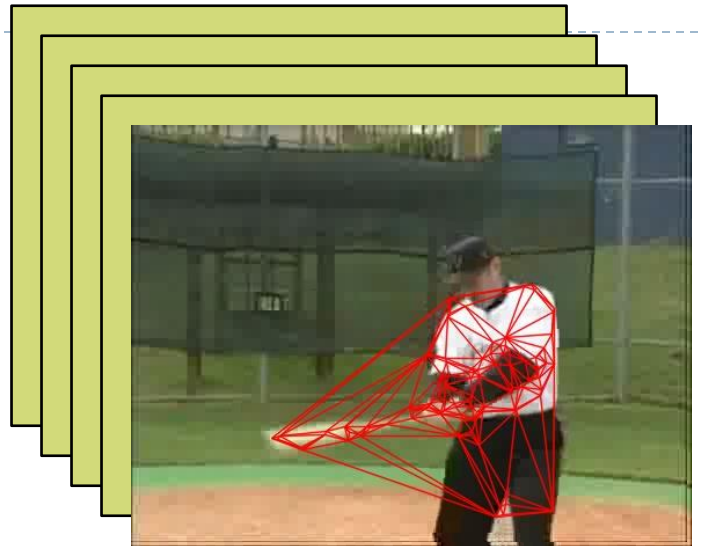
ST:時空間特徴

- １．SURFによる特徴点抽出
- ２．オプティカルフローの小さい

特徴点を除去(Lucas-Kanade法)

- ３．ドロネー三角分割法により

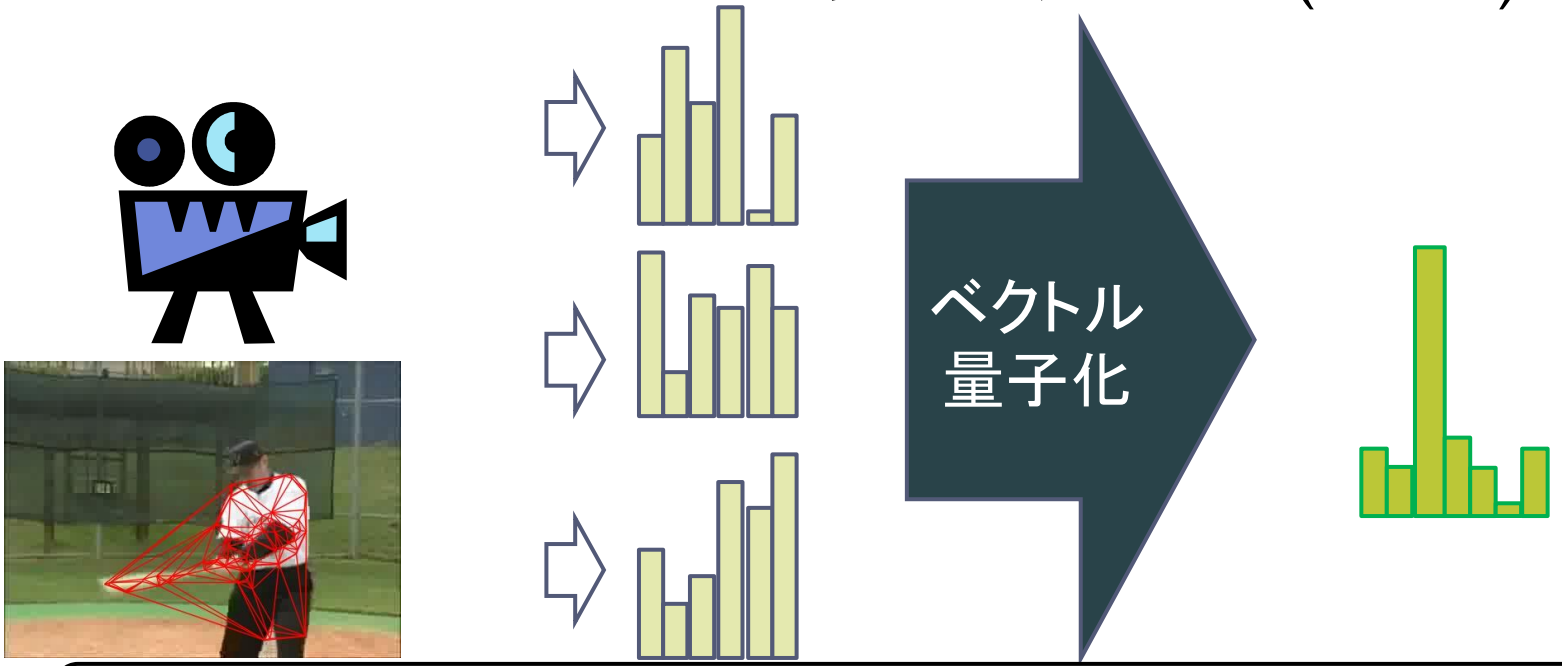
三座標による特徴を取得



- ４．各特徴の3頂点のオプティカルフローと面積変化を計算
- ５．視覚特徴と動き特徴を結合し257次元の特徴抽出
  - SURF視覚特徴64次元×3
  - 動き特徴ヒストグラム(5次元*4フレーム*3座標)
  - 三角面積(1次元*5フレーム)

Bag-of-Spatio Temporal Feature(BoSTF)

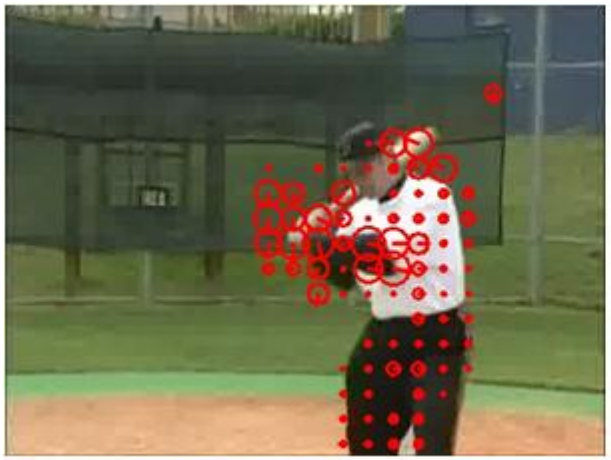
- Bag-of-Features(BoF)を動画に拡張したもの
  - 画像を局所特徴の出現頻度で表現する(BoF)
- 動画を時空間特徴の出現頻度で表現する(BoSTF)

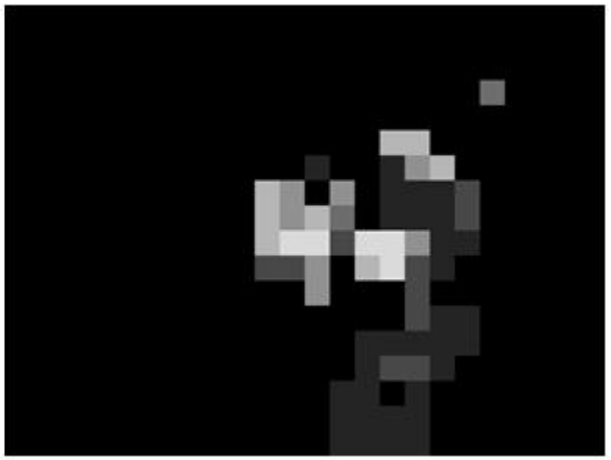


ST:時空間特徴は5000次元のBoSTF表現として利用

Motion:動き特徴

- オプティカルフローを用いてフレーム全体の動き情報を表現
- ８方向、７段階に分類し５６次元のヒストグラムを抽出
- カメラモーションが生じたフレームからは取得しない

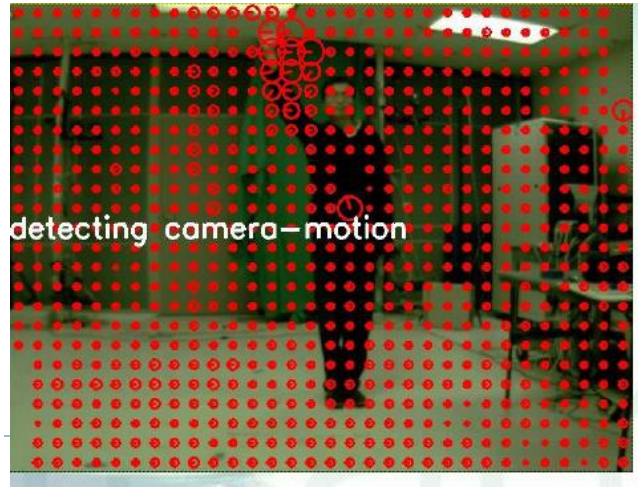


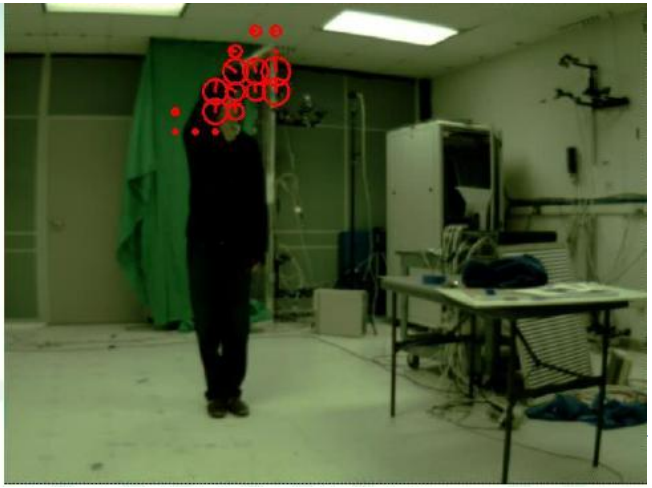


カメラモーション除去

- 動画特有の特徴利用の際、カメラモーション除去が重要

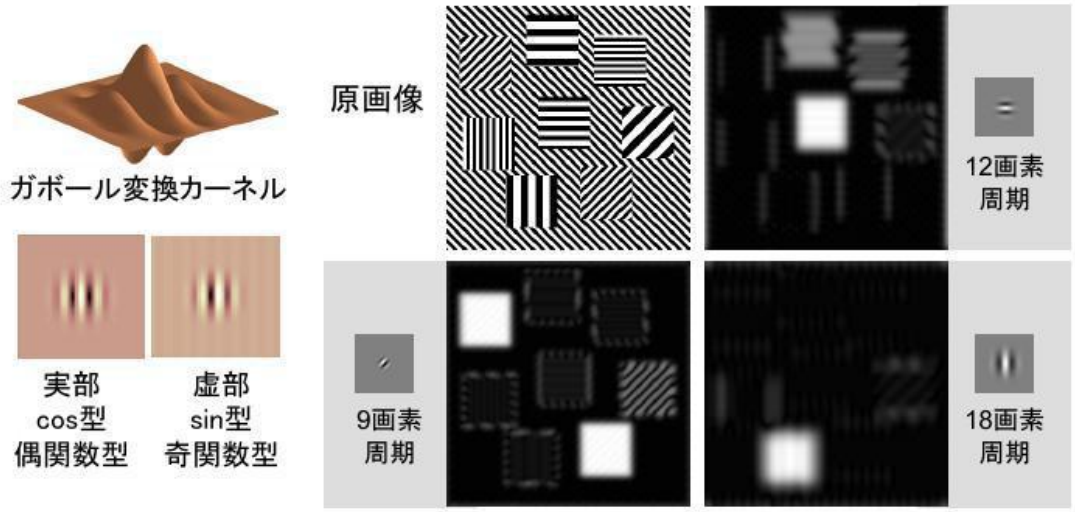
- グリッドで動きを計算
  - 動いた領域の割合が一定以上ならカメラモーションと見なす
  - カメラモーションが検出されたフレームは時空間特徴、動き特徴に利用しない





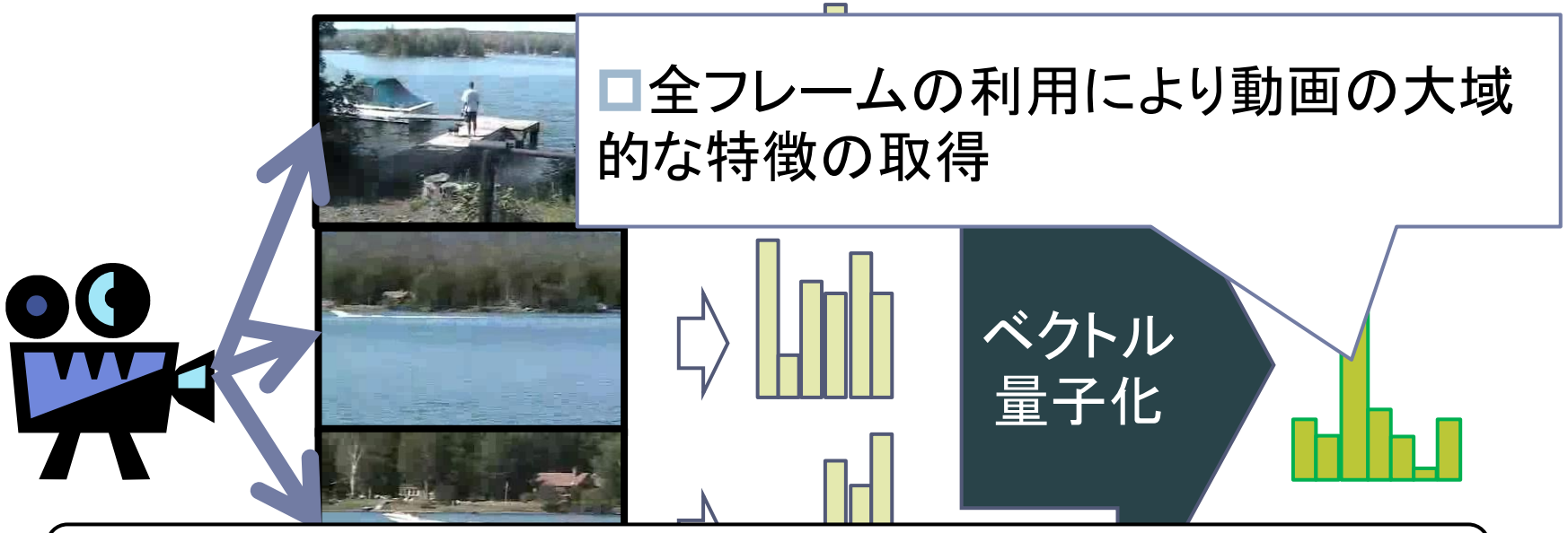
Gabor:ガボール特徴

- ガボールフィルターを用い，画像の局所的な濃淡情報を表現
  - フィルタカーネル：6方向，4周期
- 画像を20×20グリッドに分割
  - 一つのフレームから400個の24次元の特徴を抽出
  - 合計9600次元の特徴として利用



Bag of Frames(BoFr)による全フレーム認識

- Gabor、Motion特徴はBag of Frames表現で利用する
- フレーム一枚から得られた特徴全体を、一つの局所特徴と見なし、1動画中の出現頻度で動画を表現



Motion:動き特徴は3000次元のBoFr表現として利用

Gabor:ガボール特徴は5000次元のBoFr表現として利用

キーフレームの利用

- 2009年のTRECVIDで用いたキーフレームから取得した視覚特徴も一部利用した

- カラーヒストグラム
  - 画像全体からHSV、RGB、Luvそれぞれの色空間ヒストグラム64次元
  - １画像4*3分割しHSV、RGB、Luvそれぞれの色空間ヒストグラム768次元

- 顔特徴
  - Haar-Likeによる顔検出を利用
  - 検出した顔個数１次元のみ取得

Multiple Kernel Learning(MKL)

- カテゴリごとに認識で重要な特徴は異なるはず
- 重要な特徴に適切な重みづけを行うことで実現

- 複数のサブカーネルの線形結合
  - 最適な重みβを求める(MKL問題)
  - 凸面最適化問題として解く

$$K_{combined}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{K} \beta_j k_j(\mathbf{x}, \mathbf{x}') \qquad with \;\; \beta_j \geq 0, \;\; \sum_{j=1}^{K} \beta_j = 1.$$

実行環境

- 主に利用した計算クラスタマシン
  - Phenom II X4 3.0 quad core メモリ4GB
  - OS:Fedora 11(x64)

- 処理時間(130カテゴリ合計)
  - 特徴抽出:3.5週間程度
    - 学習データ:2週間程度（40コア程度利用）
    - テストデータ:1.5週間程度(80コア利用)
  - MKLの学習:2日前後(100コア利用)
  - テスト分類:2日前後(100コア利用)

実験

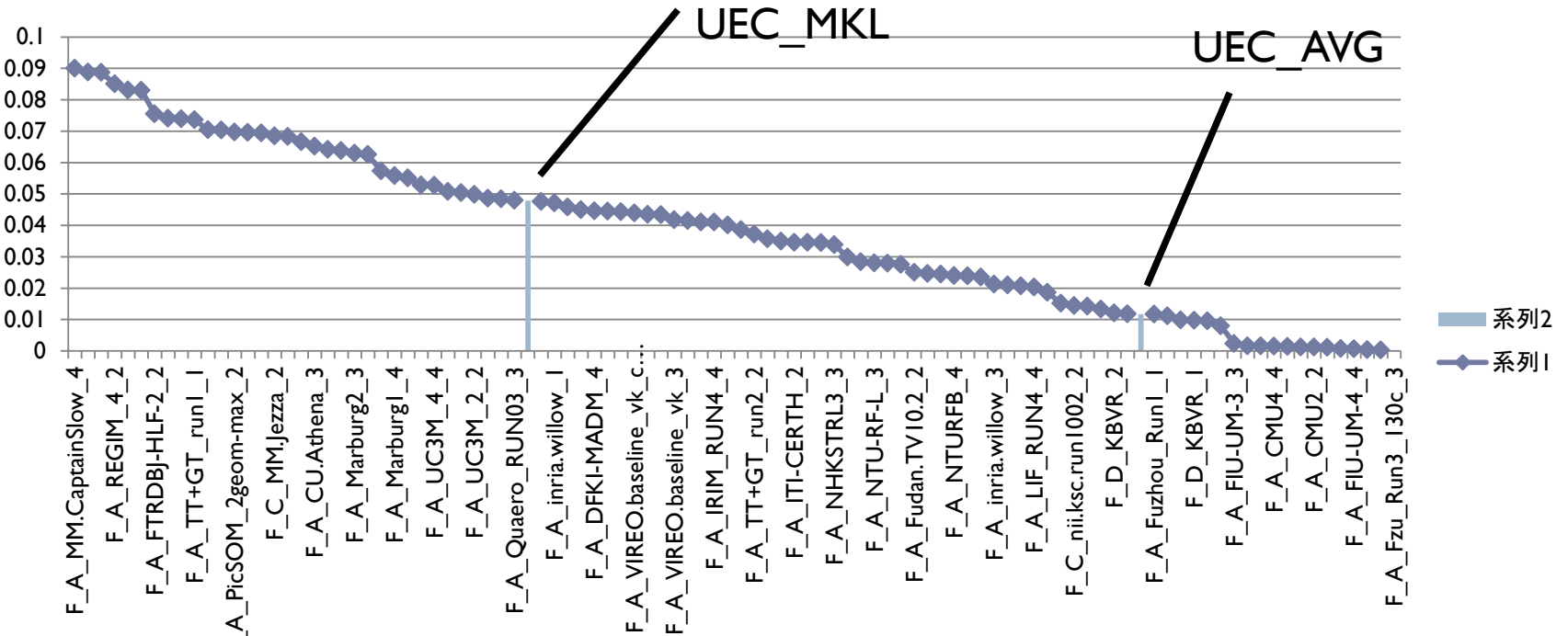
- TRECVIDデータセット
  - 学習動画118305ショット
  - テスト動画144591ショット

- 2000位までの推定平均適合率(infAP)で評価
- Full Category
  - 30カテゴリ認識
- Light Category
  - 10カテゴリ認識

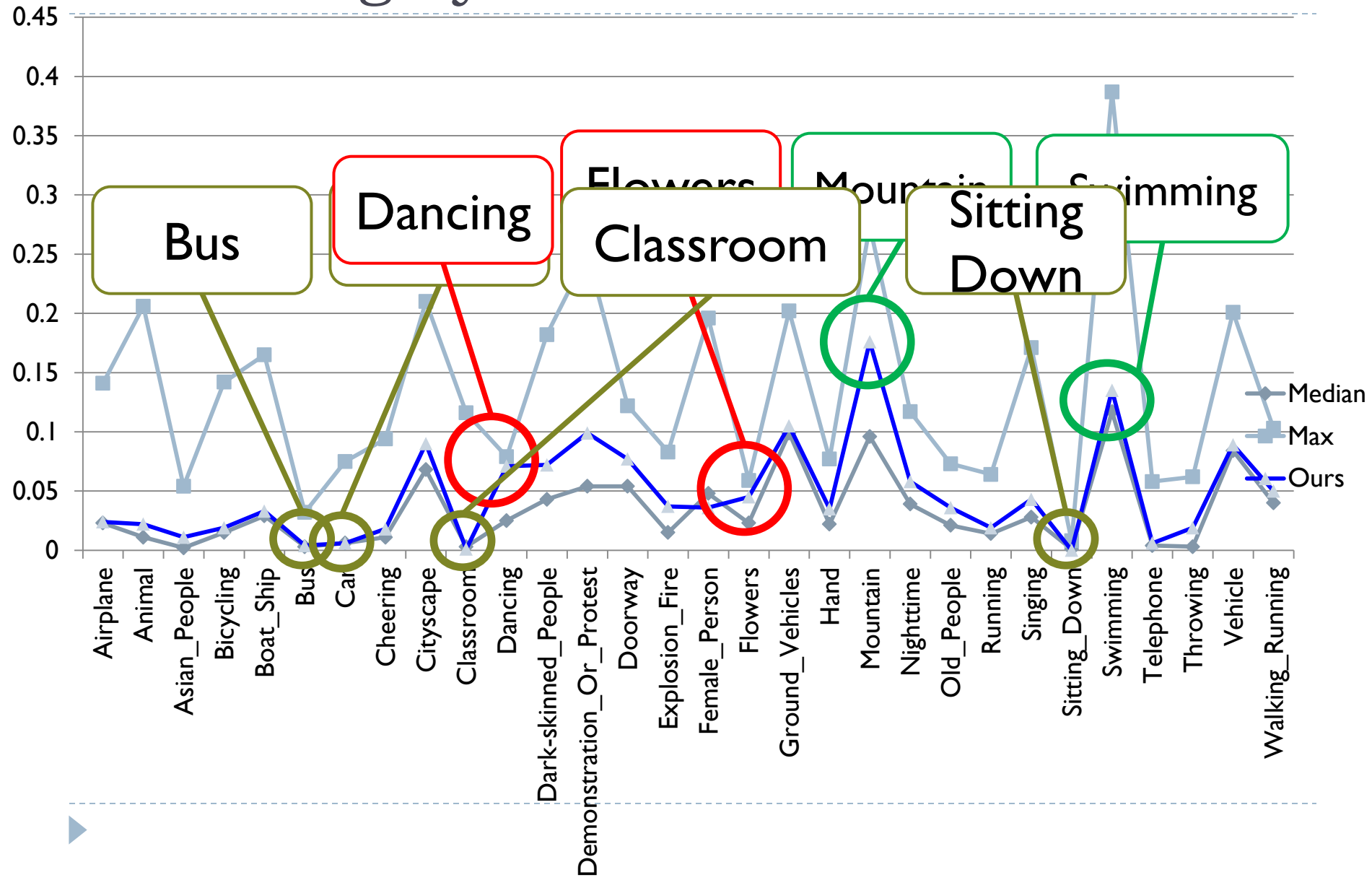
- 手法
  - 統合手法にMKLを利用(UEC_MKL)
  - 特徴を単純結合しSVMを利用(UEC_AVG)

結果(Full Category)

- 全30チーム中14位(UEC_MKL)
- infAP=0.0478(UEC_MKL)86手法中32位
- infAP=0.0117(UEC_AVG) 86手法中70位

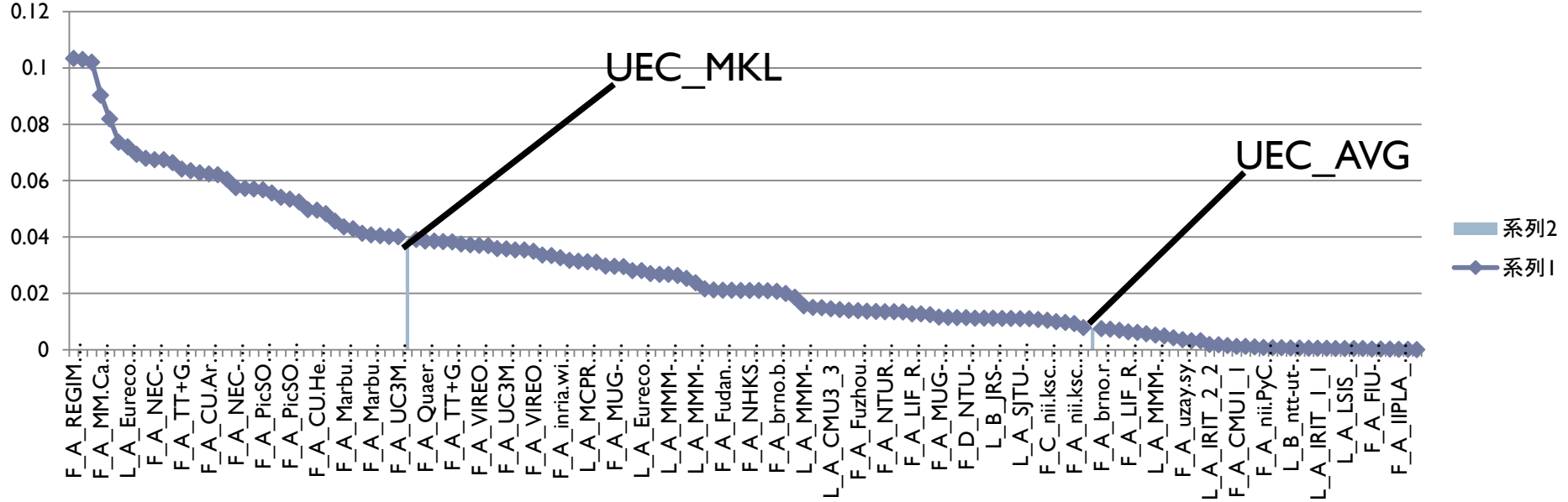


Full Categoryのカテゴリ別認識精度

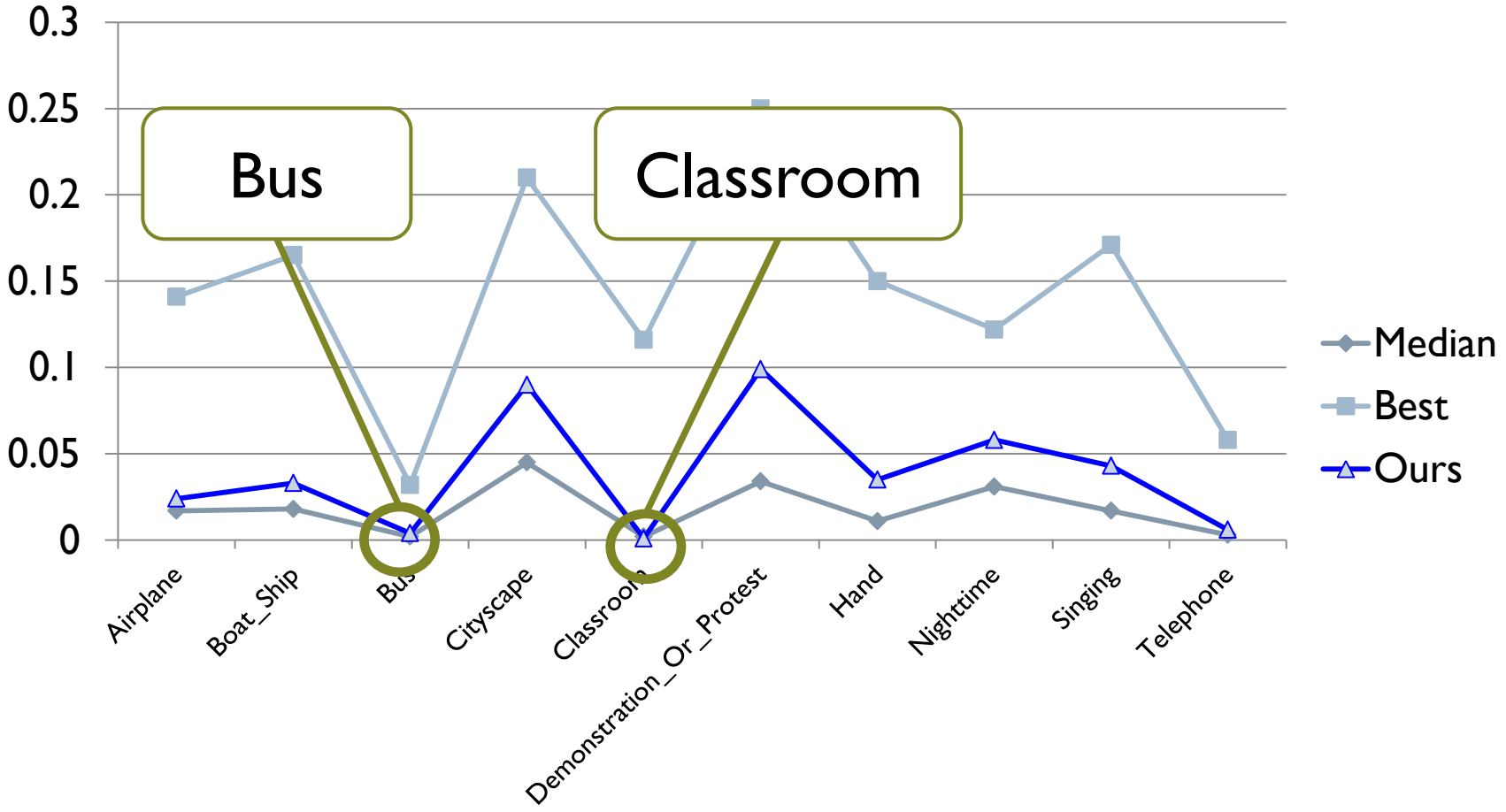


結果(Light Category)

- 全37チーム中12位(UEC_MKL)
- infAP=0.0393(UEC_MKL) 128手法中31位
- infAP=0.0077(UEC_AVG)128手法中94位



Light Categoryのカテゴリ別認識精度



結果紹介

- 成功例：Dancing
  - 466枚：1つの動画から複数選択されていることが多い
- 失敗例：Bus
  - 31枚：わずかな学習データ
- 失敗例2：Walking_Running
  - 2379枚：多様な学習データ

反省点

- メモリ容量による学習データ数の限度あり
  - 15000個利用した場合2週間計算しても終わらないことも
  - たくさんのネガティブデータをどう利用するか
  - 今回の場合ネガティブデータ5000個程度

- 段取りが重要
  - 130種類のカテゴリの認識なので、計画的に行う必要
  - MKLのパラメータ調整を行う時間が作れなかった

おわりに

- まとめ
  - 時空間特徴、bag-of-Framesなどの特徴の利用、MKLによる統合を行った結果、
  - 30チーム中14位、86手法中32位の結果を得た

- TRECVID 2011に向けて
  - 学習データの少ないカテゴリが存在
  - 認識精度向上のためWeb上の情報を利用する(Image-net)