The University of Electro-Communication Yanai Lab

動作認識のための時空間特徴量と特徴統合手法の提案

MIRU 2010
電気通信大学 情報工学専攻
野口 顕嗣
下田 保志
柳井 啓司
2010年7月29日

アウトライン

- はじめに
  - 背景，研究の目的，関連研究
- 提案手法
  - 時空間特徴抽出手法の提案
  - 特徴統合による分類手法の提案
- 評価実験
  - データセット
  - 動作認識に関する実験
  - Web動画分類に関する実験
- おわりに
  - 考察，今後の課題

背景

- **Web上には大量の動画が存在**
  - Youtube
  - ニコニコ動画
- **見たい動画を探すためにはどうすれば良い？**
  - 現状ではテキストベーストな検索手法
  - 動画を完全に特定することは困難
- **コンテントベーストなアプリケーションの必要性**
  - 動画の内容によって分類することは重要
  - 大量のWeb動画の分類を行った研究は少ない
- **Web動画分類の困難さ**
  - 背景ノイズ，低解像度，カメラモーション
  - 新しい時空間特徴抽出手法を提案
  - 特徴を統合することによる精度向上

目的

- 本研究の目的は以下の二点にある
  - Web動画分類のための新しい時空間特徴の提案
    - カメラモーションに対する対応
    - 処理の高速さ
  - 特徴統合によるショット分類手法の提案
    - 時空間特徴，視覚特徴，動き特徴
    - 時空間特徴は局所的な特徴
    - 動き特徴は全体的な特徴を記述
    - 視覚特徴は背景からも特徴を抽出

本研究では動画をシーンチェンジの境界で
分割したショットを単位として利用

関連研究

- **Web動画における分類の研究**
  - CinbisらはWeb上から動作を自動学習する手法を提案[Cinbis et al 2009]
    - Web動画の動作分類
    - 学習には静的な画像からの特徴量を使用

本研究では，視覚特徴のみではなく，動き特徴も考慮

  - LiuらはPage Rankに基づいて重要な点を選択する手法[Liu et al 2009]
    - 時空間特徴と視覚特徴を統合することでWeb動画の動作認識
    - 動作認識を行うために特徴をどのように利用するかに重点

分類のための時空間特徴抽出手法の提案

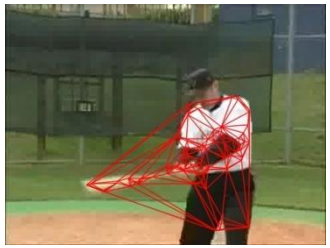
認識手法

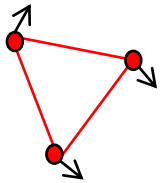
動画データセット

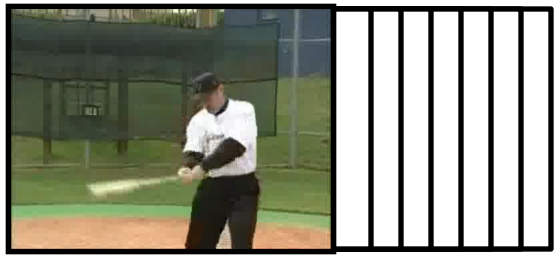
時空間特徴

視覚特徴

動き特徴







ドロネー三角形を用いた新手法

フレーム毎に取得

オプティカルフローによる表現

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

動作認識

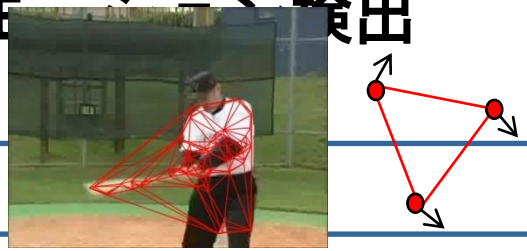
ランキング付けによるWeb動画分類

時空間特徴抽出手法(概要)

Step1

・カメラモ　　検出

時空間特徴



Step2

・視覚特

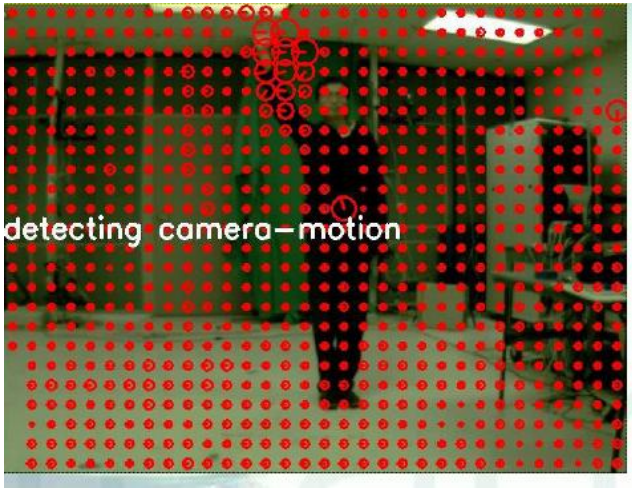
ドロネー三角形を
用いた新手法

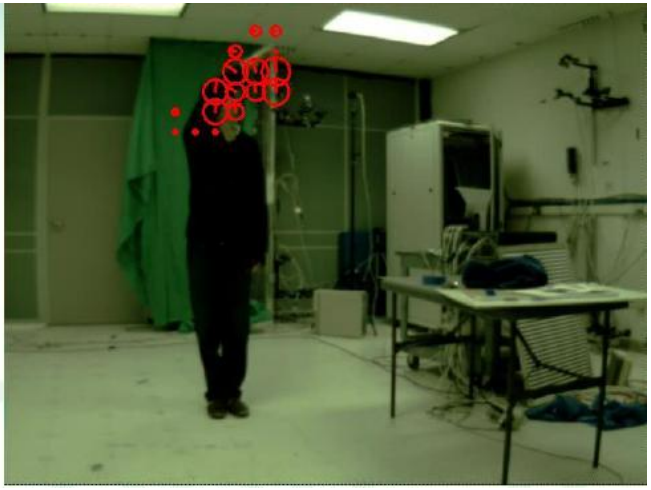
Step3

・動き特徴抽出

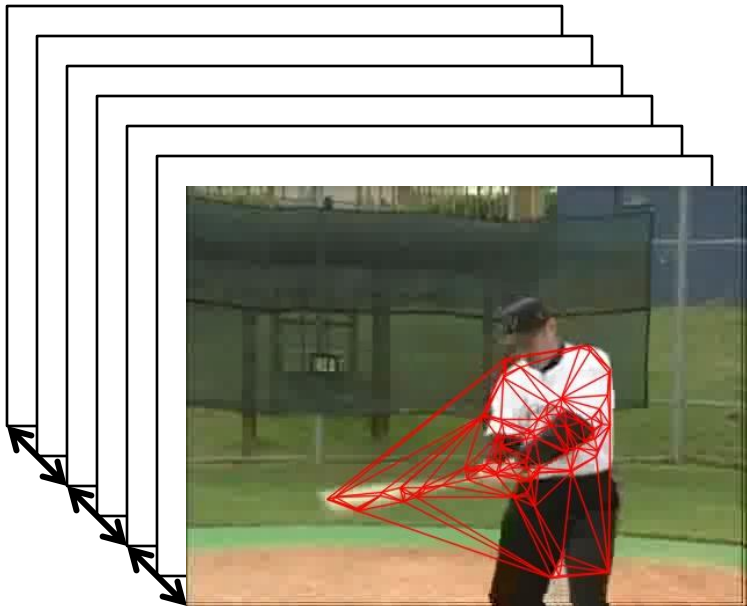
(Step 1) カメラモーション検出

- グリッドで動きを計算
  - 動いていた領域が一定割合以上ならカメラモーション
  - カメラモーションが検出されたフレームは破棄
  - Web動画における正確なカメラモーション検出は困難
    - 低い解像度
    - 手振れなどの意図しないカメラモーション





(Step 2) 視覚特徴抽出



①. Nフレームを１ユニットとする
②. SURFを抽出
③. 動きがない点を削除
④. ドロネー三角形を作成
以降三点で一組の特徴と考える

これにより隣接する特徴を考慮した特徴記述が可能に

(Step 3) 動き特徴抽出

①選ばれたNフレームを
M分割する

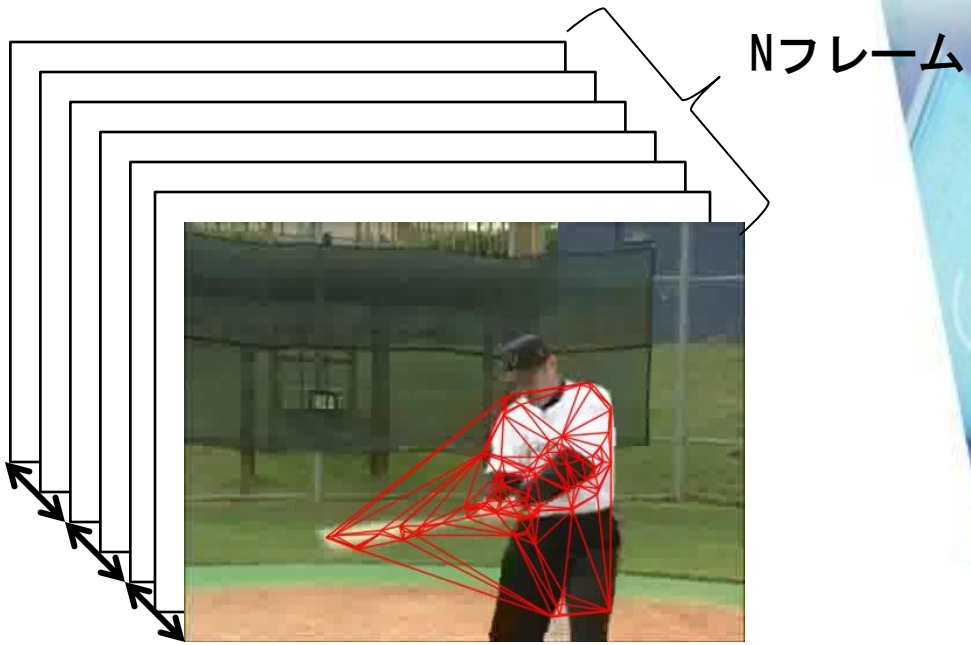
②それぞれの区間における
時空間特徴点の動き情報
を計算

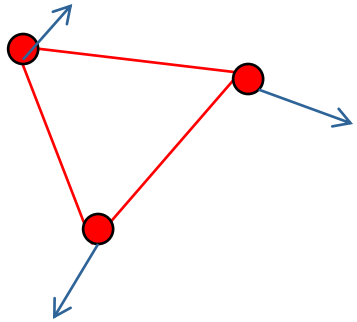
③視覚特徴と同様に動きも
三点の組として考える

④動き特徴としてそれぞれの
点の動きと三角形の面積の
変化を利用

⑤視覚特徴と動き特徴を結合

⑥抽出した特徴をベクトル量子化



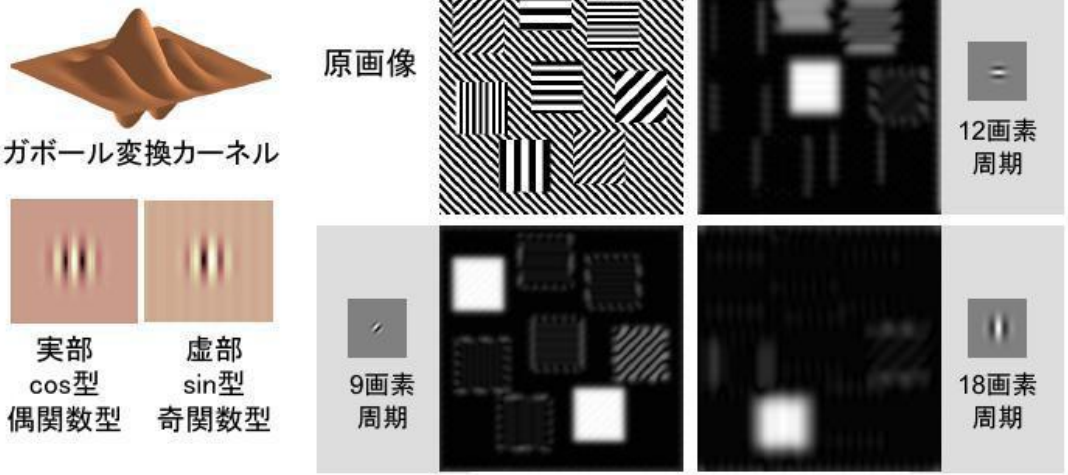


特徴統合による分類手法（視覚特徴）

- ガボールフィル の局所的な濃淡情報を表現
  - フィルタカーネ
- 画像を20×20グ
  - 一つのフレーム を抽出
- 全フレームから

視覚特徴

全フレームから取得



特徴統合による分類手法(動き特徴)

- **動き情報をグ**
  - オプティカル[illegible]き8方向/段階の大きさのヒストグラムに投票

フレーム単位から抽出する特徴について

- 多くの研究では動画を代表するキーフレームから特徴を抽出

しかし

- キーフレームの選択は重要だが難しい
  - 選ばれたフレームによって特徴は異なる

そこで

本研究ではBag-of-Framesという考えを導入

◆フレーム一枚を局所特徴と考える
◆フレームから抽出された特徴をベクトル量子化
◆抽出される特徴の出現頻度で動画を表現

キーフレームのみでなく，動画全体の特徴を考慮可能

Multiple Kernel Learning(MKL)

- **複数のサブカーネルを線形結合**
  - **最適な重みβを求める(MKL問題)**

$$K_{combined}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{K} \beta_j k_j(\mathbf{x}, \mathbf{x}')$$

$$\text{with } \beta_j \geq 0, \quad \sum_{j=1}^{K} \beta_j = 1.$$

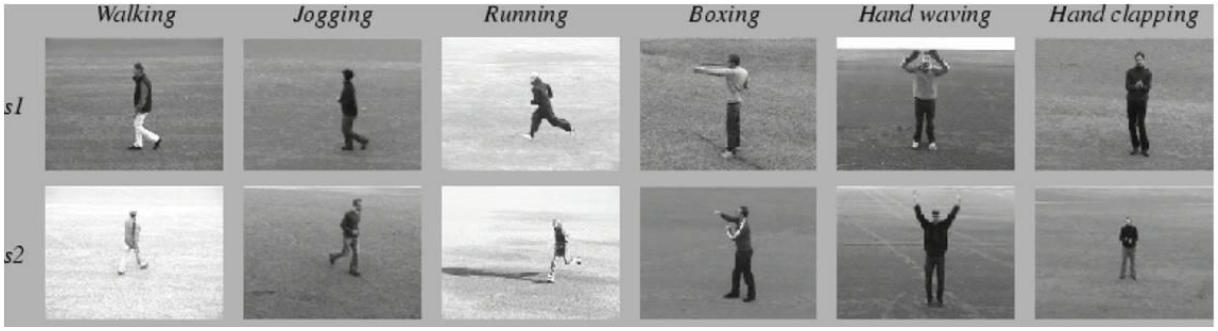
- **単一カーネルのSVM 学習を反復することによって最適なカーネルの重みβを求める手法が提案されている[12]**

評価実験

- **動作認識**
  - KTHデータセット
    - 6種類の動作、599ショット
    - Leave-one-outで学習
  - Liuらが構築したWild Youtube データセット
    - 11種類の動作、合計1595ショット
    - 5-fold cross validationで学習
- **大規模データセットによるWeb動画分類**
  - 独自に構築したOur Youtube dataset
    - 8種類の動作、92時間、合計55,845ショット
  - 教師信号ありのランキング付け

データセット(動作認識)

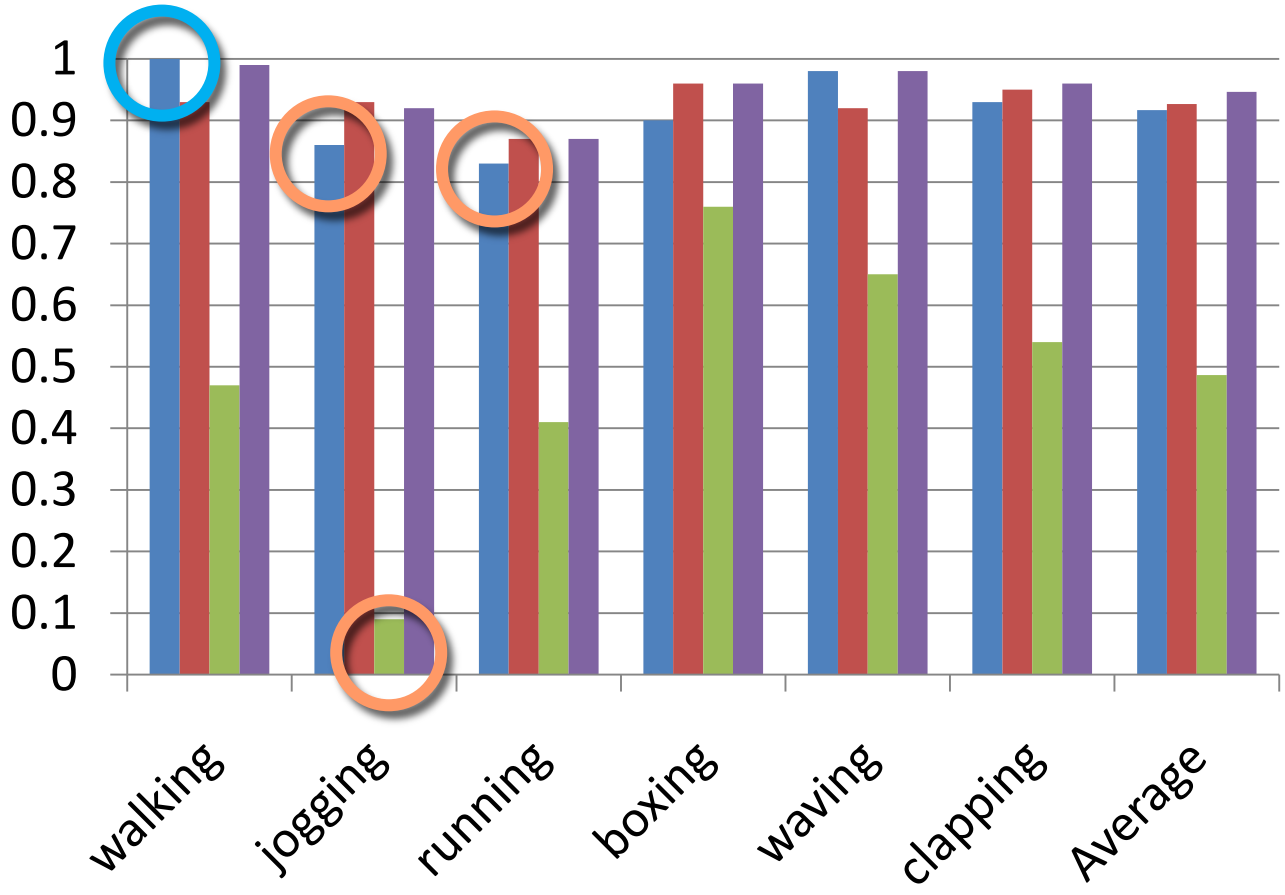
- KTHデータセット
  - 6種類の動作、599ショット



- Wild Youtubeデータセット
  - 11種類の動作、合計1595ショット

実験結果（動作認識）

- KTHデータセット



時空間特徴	91.7%
動き特徴	92.7%
視覚特徴	48.7%
MKL	94.5%

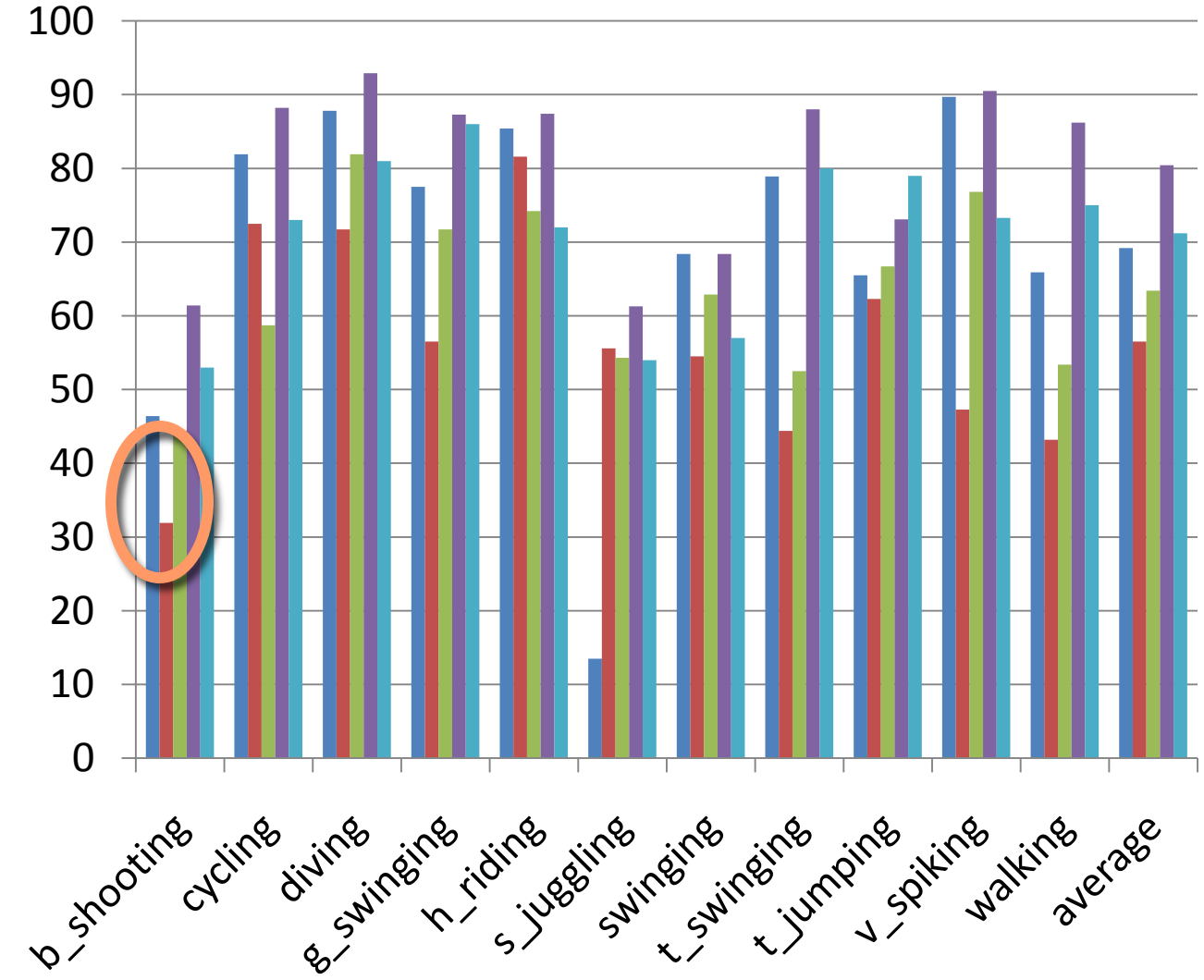
実験結果（動作認識）

- KTHデータセット

	walking	jogging	running	boxing	waving	clapping
walking	0.99	0.01	0	0	0	0
jogging	0.03	0.94	0.03	0	0	0
running	0.01	0.14	0.85	0	0	0
boxing	0.01	0	0	0.96	0.01	0.02
waving	0	0	0	0	0.98	0.02
clapping	0	0	0	0.05	0.02	0.93

実験結果（動作認識）

- 動作認識(Wild Youtube dataset)



視覚特徴
動き特徴
時空間特徴
MKL
Liu [CVPR2009]

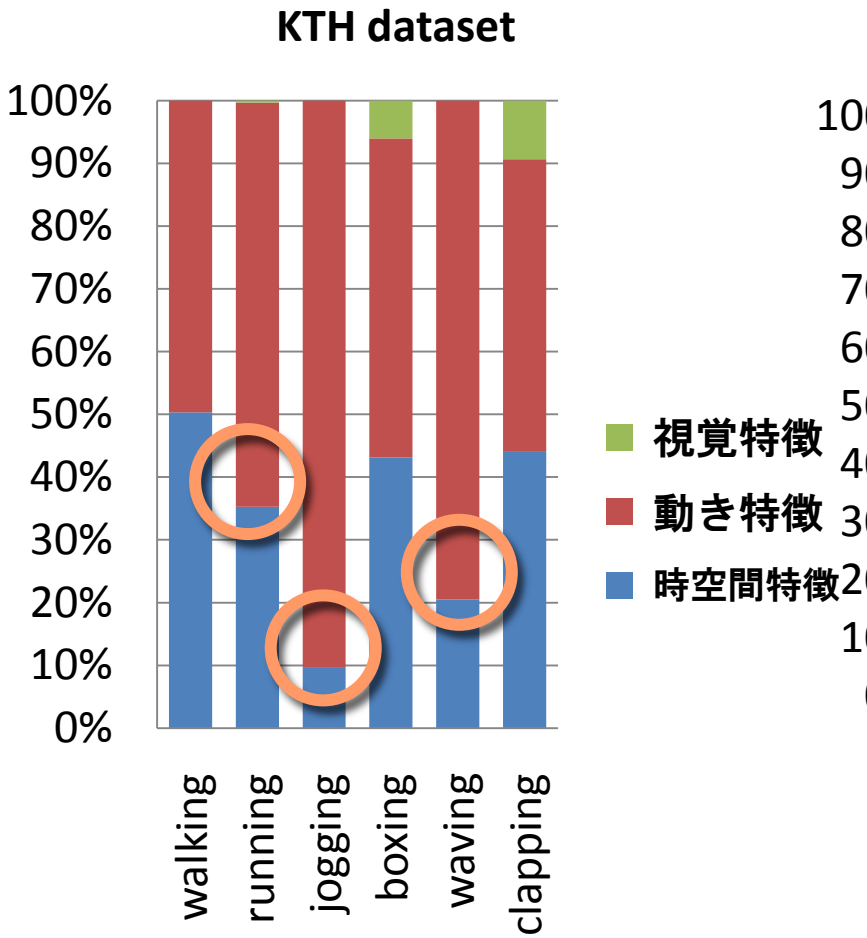
69.1%	
56.5%	
63.4%	
80.4%	
71.2%	

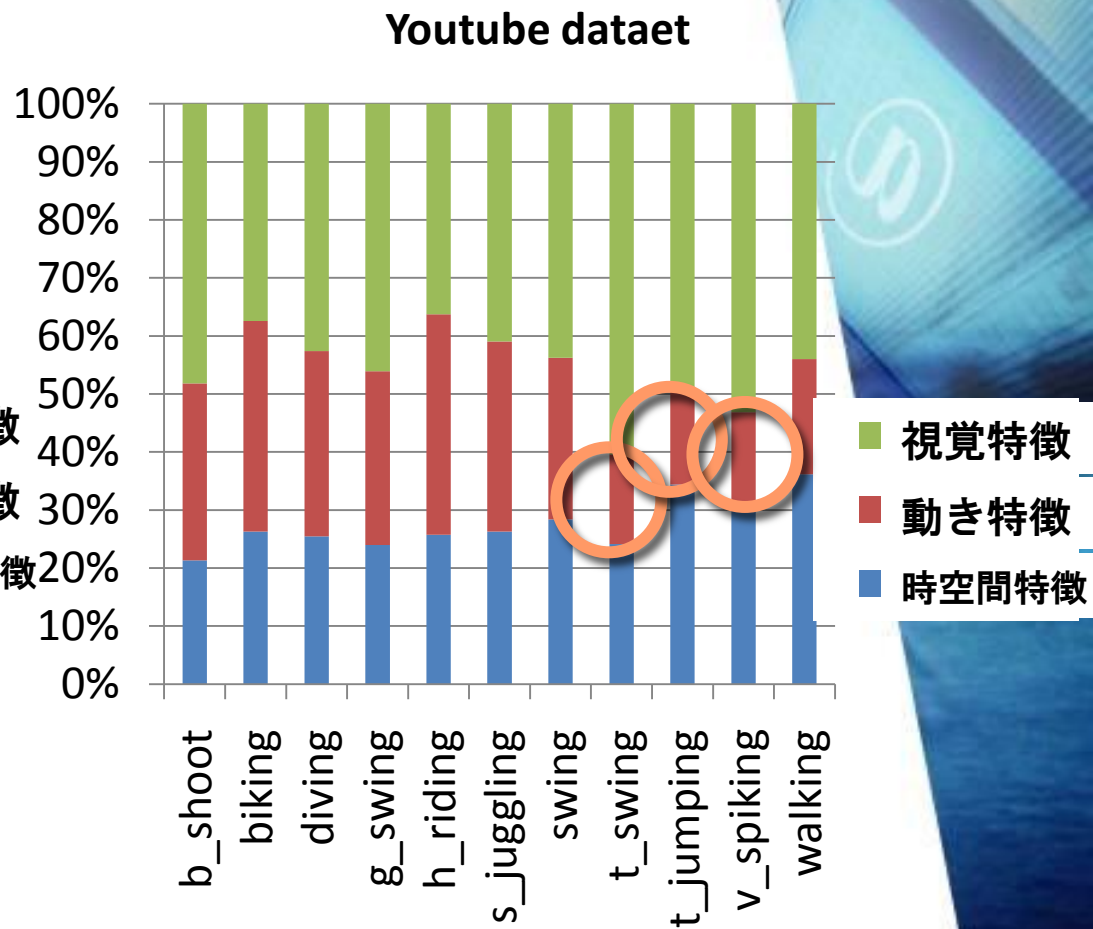
実験結果（動作認識）

	b_shooting	cycling	diving	g_swinging	h_riding	s_juggling	swinging	t_swinging	t_jumping	v_spiking	walking
b_shooting	61.4	1.4	9.3	14.3	2.1	5.7	0.7	1.4	0	3.6	0
cycling	0	88.2	1.4	0	9.7	0	0.7	0	0	0	0
diving	2.6	0	92.9	0.6	2.6	0	0.6	0	0	0.6	0
g_swinging	2.8	0	0.7	87.3	1.4	5.6	2.1	0	0	0	0
h_riding	0	6.6	2	1.5	87.4	0.5	1	0.5	0	0.5	0
s_juggling	2.6	0	3.9	14.8	2.6	61.3	9	3.2	0	2.6	0
swinging	0	9.6	0.7	5.1	0	2.2	68.4	0	9.6	0	4.4
t_swinging	0	0	4.2	0	0	0	0	88	1.8	3.6	2.4
t_jumping	0	0.8	0	1.7	0	3.4	11.8	0	73.1	0.8	8.4
v_spiking	4.3	0	0	0.9	0.9	0	0	0.9	1.7	90.5	0.9
walking	0	0.8	1.6	0.8	3.3	0	0.8	2.4	0.8	3.3	86.2

実験結果（動作認識）

- MKLのカーネル重み





実験結果（動作認識）

- KTHデータセット

分類結果（Leave-one-out）	
Ours	94.5%
D’ oller et al. [3]	81.2%
Liu et al. [6]	91.8%
Kim et al. [14]	95.33%
Gilbert et al. [13]	96.2%

- Wild Youtubeデータセット

分類結果（5-fold cross validation）	
Ours	80.4%
Liu et al. [6]	71.2%

データセット(大規模Web動画分類)

- **大規模Web動画分類**
  - Youtubeから独自に収集した8カテゴリのデータセット
    - 合計92時間、総数55,845ショット
      - Wildの33倍
    - 分類対象は6カテゴリ．学習データは528ショットを選択
    - 残りの2カテゴリはノイズデータとして利用

batting


running


jumping


動画数	174
ショット	8,980
動画数	170
ショット	7,342
動画数	174
ショット	6,567

walking


shoot


eating


動画数	164
ショット	7,718
動画数	142
ショット	3,442
動画数	160
ショット	3,130

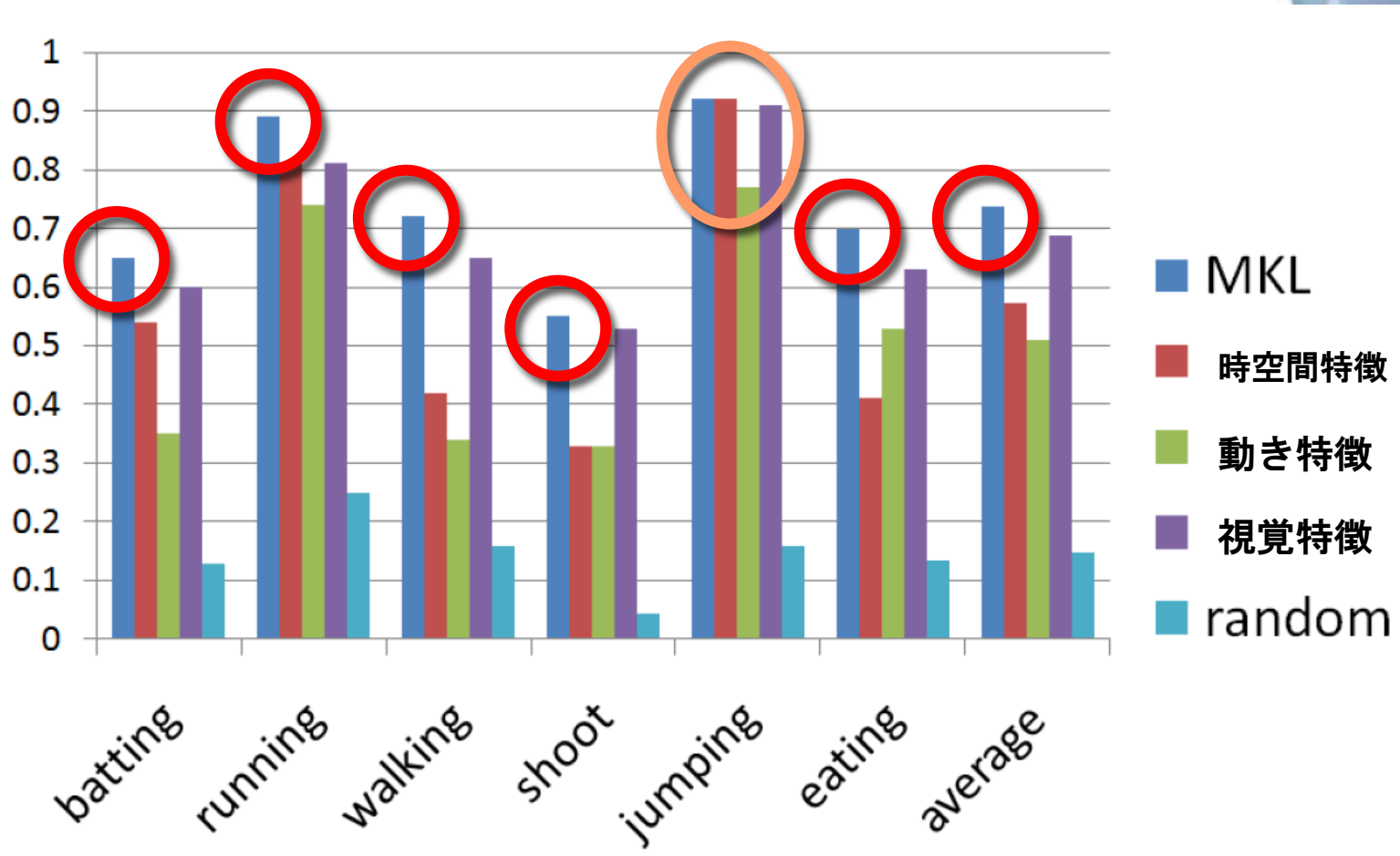
実験結果 (大規模Web動画分類)

- 大規模Web動画分類

結果のデモ

実験結果 (大規模Web動画分類)

- 上位200位までの適合率



おわりに

- まとめ
  - Web動画分類のための時空間特徴抽出手法を提案
  - 特徴統合による動作認識手法の提案
    - KTHデータセットにおいて最新手法と同等
    - MKLによる特徴統合は動作認識において有効
  - 大規模なデータセットによるWeb動画分類を実験
    - ＭＫＬの上位20位の平均適合率97.7%、200位の平均適合率73%
    - 時空間特徴の上位20位の平均適合率91.7%、200位の平均適合率57.2%
- 今後の課題
  - カメラモーションに動き補正
  - 有益な特徴の選択
  - 複数の動作を認識

ご清聴ありがとうございました