A SURF-based Spatio-Temporal Feature for Feature-fusion-based Action Recognition

ECCV 2010 WS on Human Motion Understanding, Modeling, Capture and Animation

Akitsugu Noguchi

★ Keiji Yanai

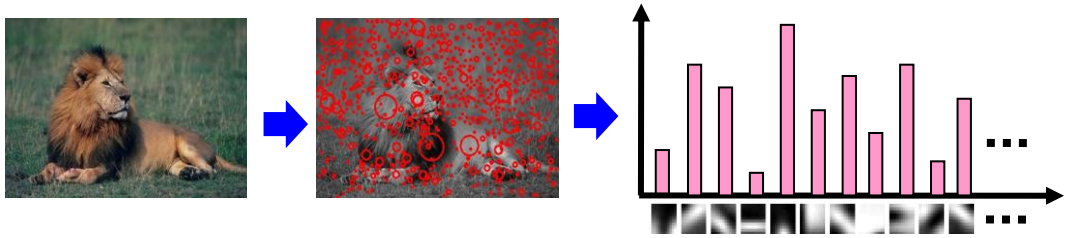
The Univ. of Electro-Communications, Tokyo, Japan

1. Background & Objective

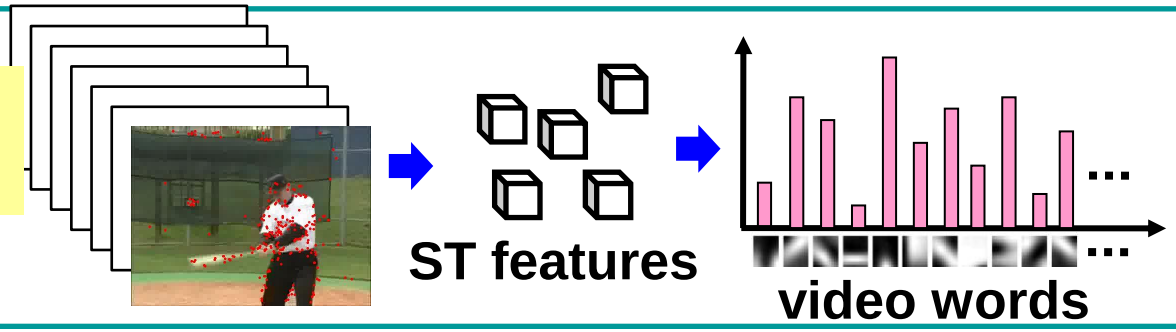
Background (1): BoF for action rec.

- **action recognition ≒ object / scene recognition**
  - Bag-of-features (BoF) of spatio-temporal features [Dollar et al. VS-PETS05]

Bag-of-words

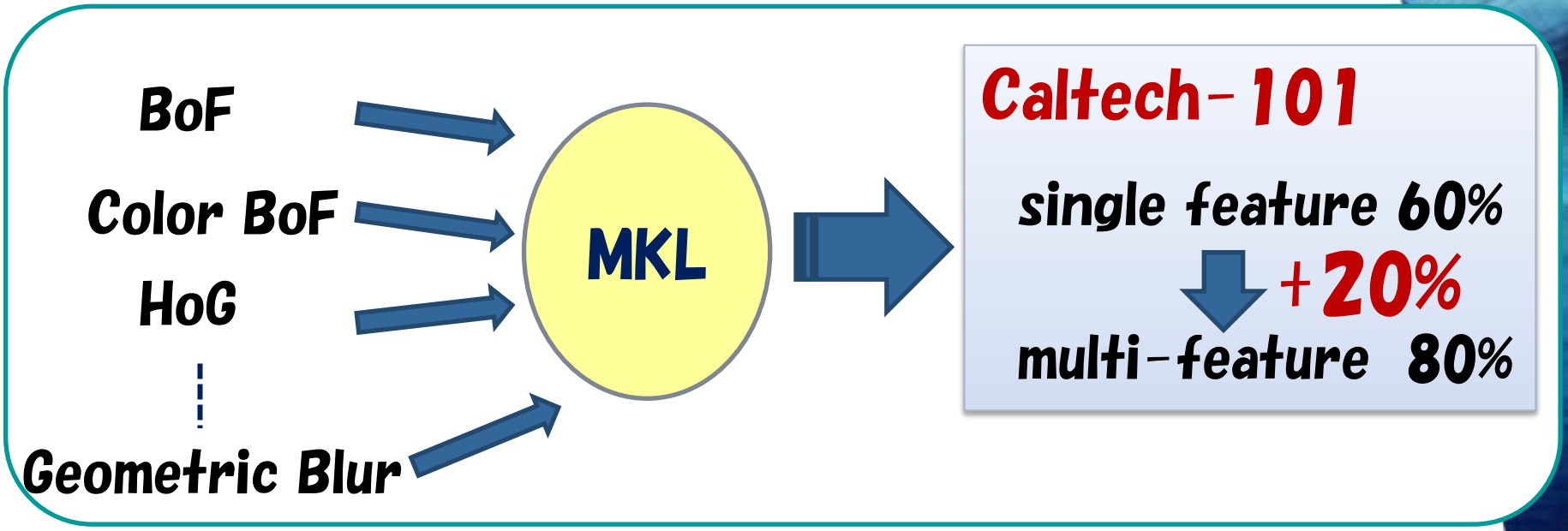


Bag-of-video-words



Background (2): Feature fusion by MKL

- **Multi-feature fusion by Multi Kernel Learning (MKL) for object recognition** [Varma el al. ICCV07]
  - MKL can estimate fusion weights adaptively.



Objective

Feature-fusion-based action recognition with MKL and heterogeneous features

Use features having different characteristics to deal with a wide range of videos from KTH to YouTube

- sparse Spatio-Temporal (ST) features
  - SURF-based new ST feature using Delauney triangulation (New!)
- temporally-dense Appearance features
- temporally-dense Motion features

Related work: MKL for action recoginition

- [Sun et al. CVPR09]
  - Select good ones from many kinds of trajectory-based features
- [Han et al. ICCV09]
  - Combine 30 part-based features

They applied MKL for the same kinds of features.

In this paper, we simply use MKL to combine (small number of) different kinds of features.

Related Work (2): fusion of heterogeneous features

- [Liu et al. CVPR09]
  - BoF of SIFT + BoF of cuboid (ST features) with Adaboost
- [Niebles et al. CVPR07][Cinbis et al. ECCV10]
- … some other papers

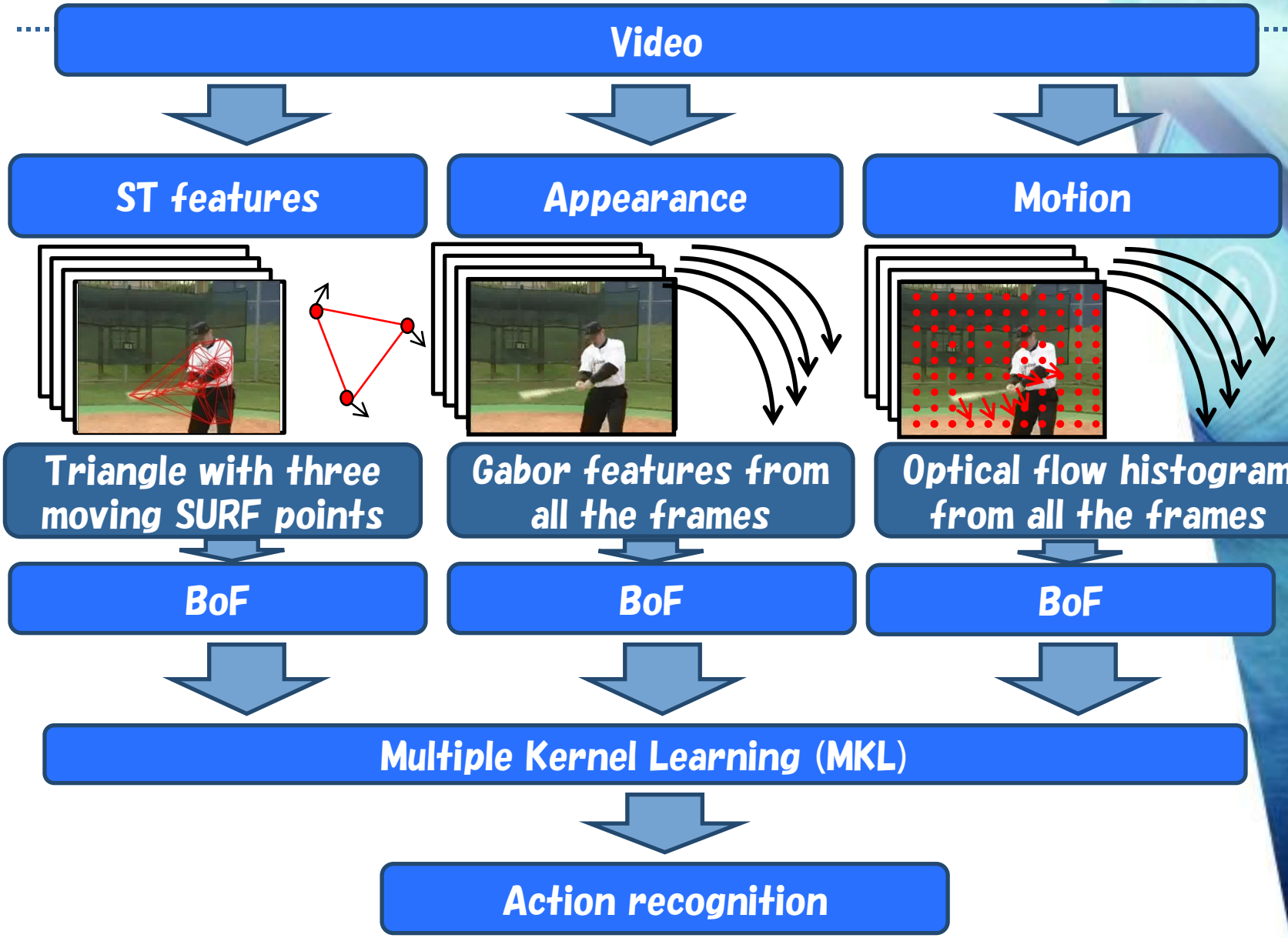
In this paper, we combine static appearance, motion, and spatio-temporal features.

Contributions

- Fusion of static appearance features, dynamic motion features and intermediate ST features
- SURF-based new ST features
- Temporally-dense sampling of appearance and motion features
  - And show their effectiveness with Both the controlled KTH dataset and uncontrolled Youtube video datasets.

2. Overview of Our Approach

Our approach

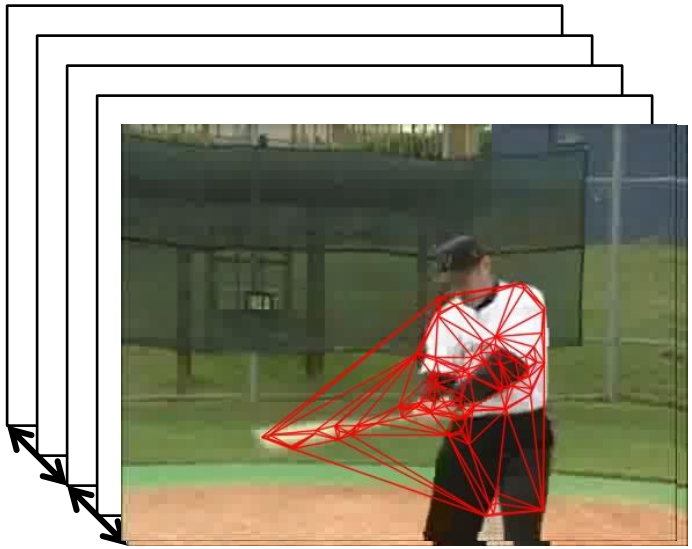


3. Proposed Method

(1) SURF-based New ST Features

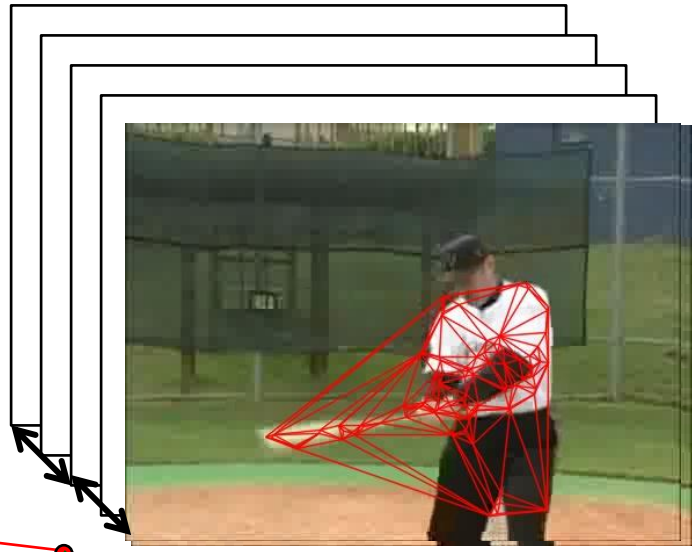
- **Extension of our previous work** [ACCV2009]
  - Combination of SURF local features, optical flows and **Delaunay triangulation**.
  - Can extract ST from not points but a set of patches ⇒ **more robust and informative**

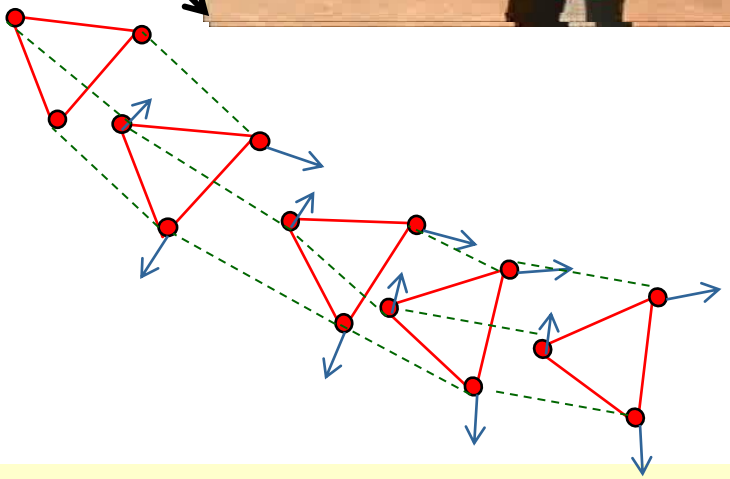
New!



① Extract SURF features
② Extract optical flows over SURF interest points by LK methods and select moving SURF points
③ Apply Delaunay triangulation over the points where flows are detected
④ Track each moving points for consecutive N frames based on optical flows (N=5)

(1) SURF-based New ST Features (cont.)





Represent ST features with a sequence of triangle patches

④ Track each moving points for consecutive N frames based on optical flows (N=5)

⑤ Convert the flow vector into a 5-dim vector regarding each moving points

⑥ Compute the difference of the size of triangles between consecutive frames.

⑦ Concatenate 3 SURF vectors, 3 5-dim flow vectors and (N-1) size difference values. (in case of N=5, totally 256dim)

over whole the video

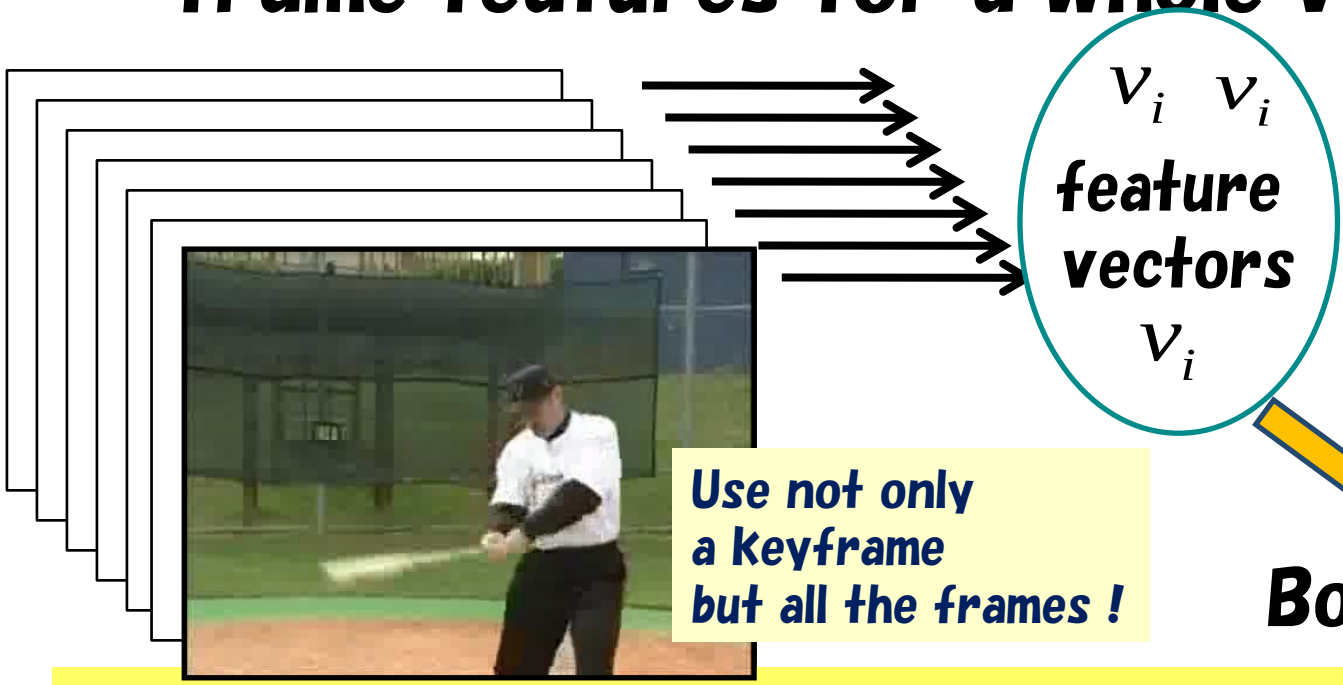
BoF vector

Codebook size is 5000.

Note that if too many flows are detected, we give up extracting ST features from the frame.

(2) Temporally-dense features & Bag-of-frames (for appearance and motion)

- Extract apperance/motion from all the frames ⇒ Temporally-dense
- Generate a BoF vector from a bag of frame features for a whole video seq.

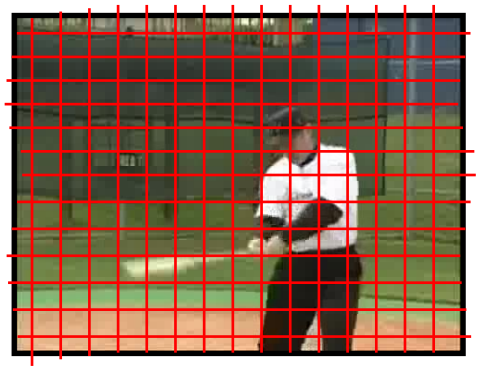


Codebook size is 3000.

"Temporally-dense" is a commom method among the TRECVID paticipants.

(2-1) Appearance features (Gabor)

- **Gabor filter response histogram:**
  - represents local textures
  - 6 directions * 4 frequencies
- **Extract from 20×20grids**
  - Obtain 400 24-dim Gabor vectors from one frame



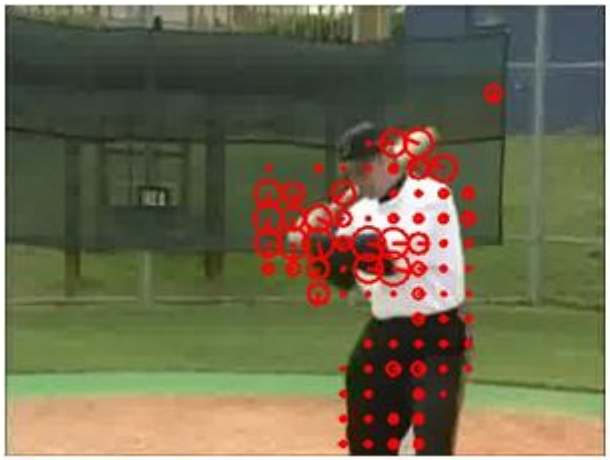
$\oplus$

$v_i \times 400$

BoF vector

(2-2) Motion features (Optical flow)

- Extract optical flows at grid points with every 8pixels by the Lucas-Kanade optical flow detector.
- Vote to a histogram consisting of 7directions and 8motion magnitude steps.



v_i

BoF vector

Multiple Kernel Learning (MKL)

[Sonnenburg et al. 2006]

- Is an extenstion of a SVM .
- Can handle "a combined kernel" which is a linear combination of kernels.
- Can estimate kernel weights and SVM model parameters simultaneously.
- Can integrate features by assigning one feature to one kernel.

Combined kernel

Visual Motion St features

$$\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{K} \beta_k \mathbf{k}_k(\mathbf{x}_i, \mathbf{x}_j)$$

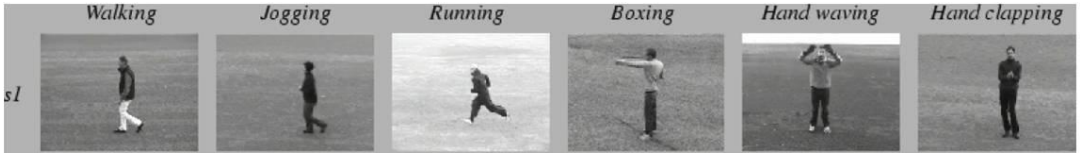
kernel weights

4. Experimental results

Experiments: Datasets

- Action classification (multiclass / 1-vs-rest)
  - (1) KTH [Schuldt et al. ICPR04]
    - 6 kinds of actions. 599 video shots.
    - Evaluated by "leave-one-out"

Controlled

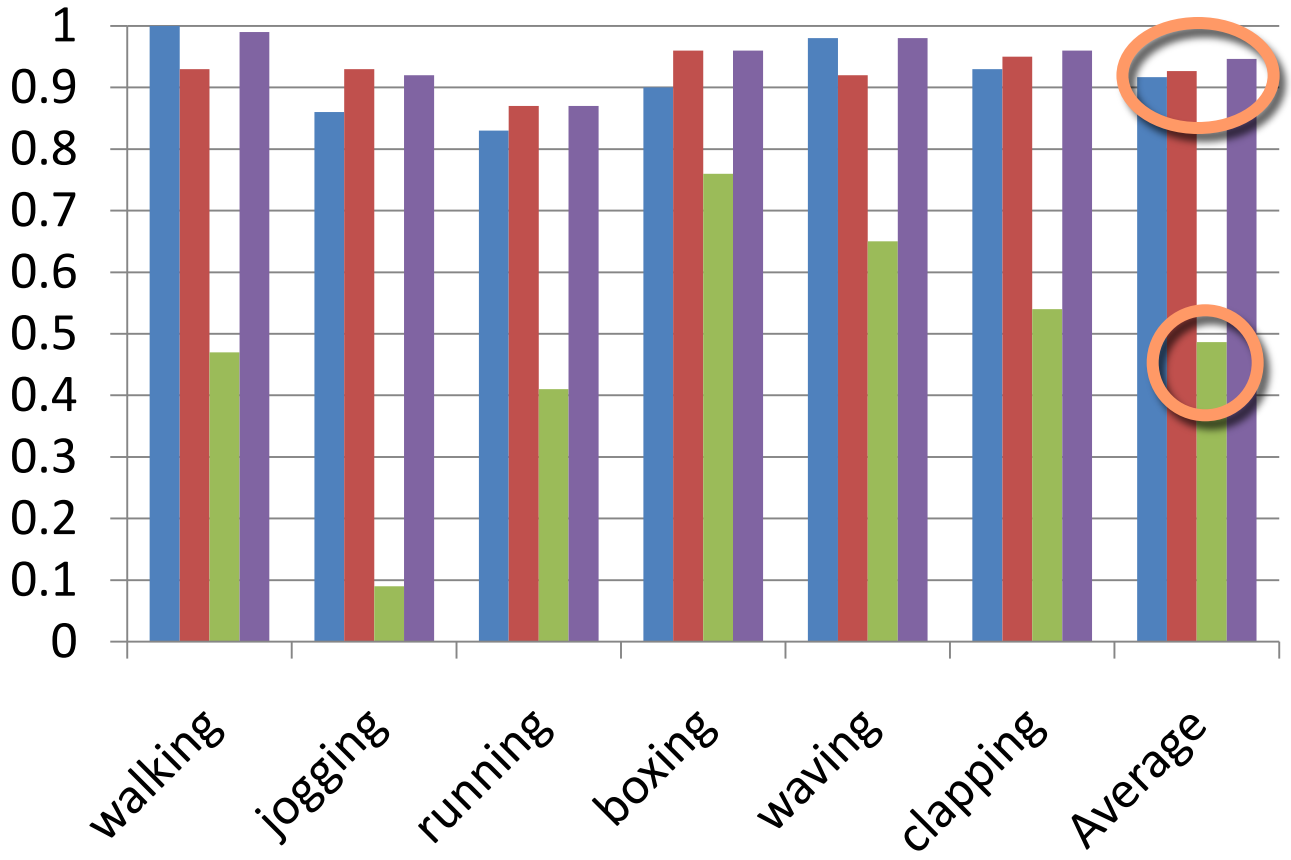


  - (2) Wild Youtube [Liu et al. CVPR09]
    - 11 kinds of actions. 1168 video shots.
    - Evaluated by "5-fold cross validation"

Uncontrolled

(UEC)

(1) KTH: Results



ST feature 91.7%
Motion 92.7%
Appearance 48.7%
MKL 94.5%

- Surprisingly, dense motion is the best single feature. 92.7%
- By combining three features, it was boosted to 94.5%.

(1) KTH: Comparison

Methods	performance
Dollar et al. [VS-PETS05]	81.2%
This paper (only Motion)	92.7%
Liu et al. [CVPR09]	93.8%
This paper (ST+App+Motion)	94.5%
Kim et al. [CVPR09]	95.3%
Gilbert et al. [ICCV09]	96.2%

- 94.5% is a good performance.
- "Motion" is the most important for KTH.

(1)KTH: proposed ST feature

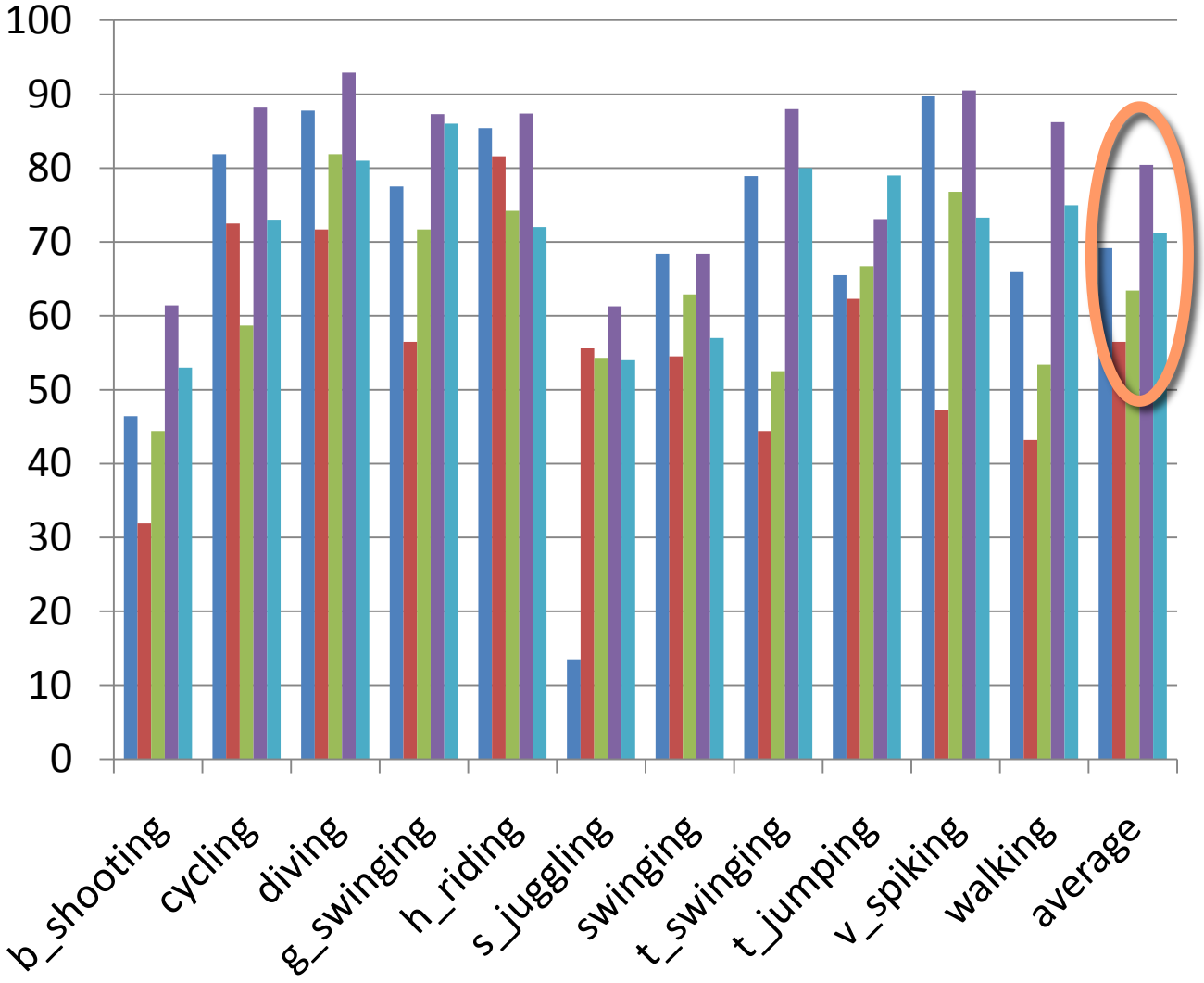
Methods	performance
No triangulation No rotation-invariance	83.3%
No triangulation With rotation-invariance	86.3%
With triangulation With rotation-invariance	91.7%

+3.0%

+5.4%

- Rotation-invariance boosted it by 3.0%.
- Triangulation boosted it by 5.4%.

(2) "Wild Youtube" : Results



Appearance 69.1%
Motion 56.5%
ST feature 63.4%
MKL(all) 80.4%
Liu et al. 71.2%
[CVPR2009]

MKL(App+Motion)73.5%

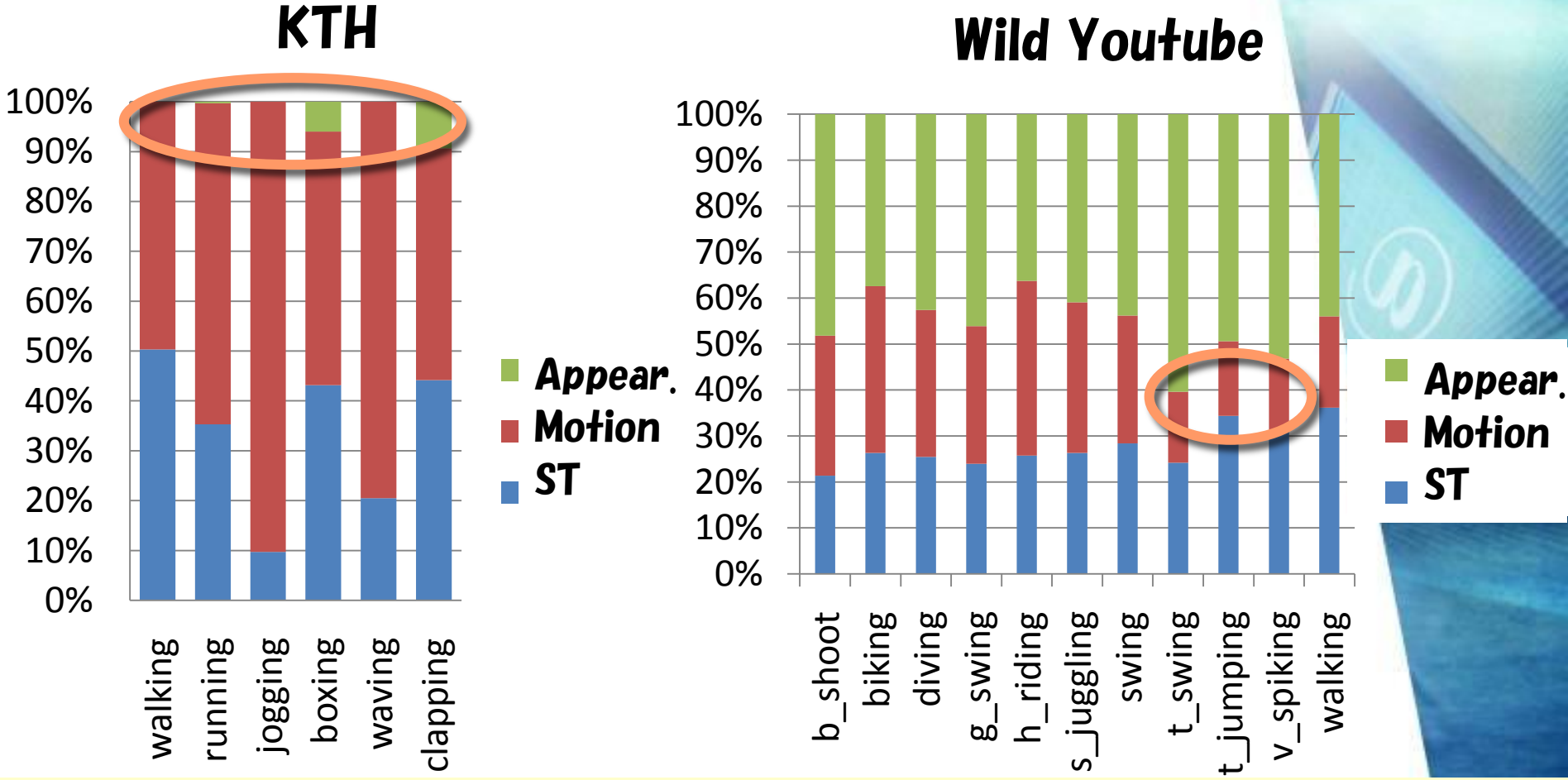
(2) Wild Youtube : Comparison

Methods	performance
This paper (only Motion)	56.5%
This paper (only ST feature)	63.4%
This paper (only Appearance)	69.1%
Liu et al. [CVPR09]	71.2% (leave-out-out)
This paper (App+Motion)	73.5% (5-fold CV)
N. I-Chinbis et al. [ECCV10]	75.2% (leave-one-out)
This paper (ST+App+Motion)	80.4% (5-fold CV)

+6.9%

- ST is not the best single feature, but it boosted MKL results.
- Leave-on-out for 1168 videos with MKL is too time-consuming !

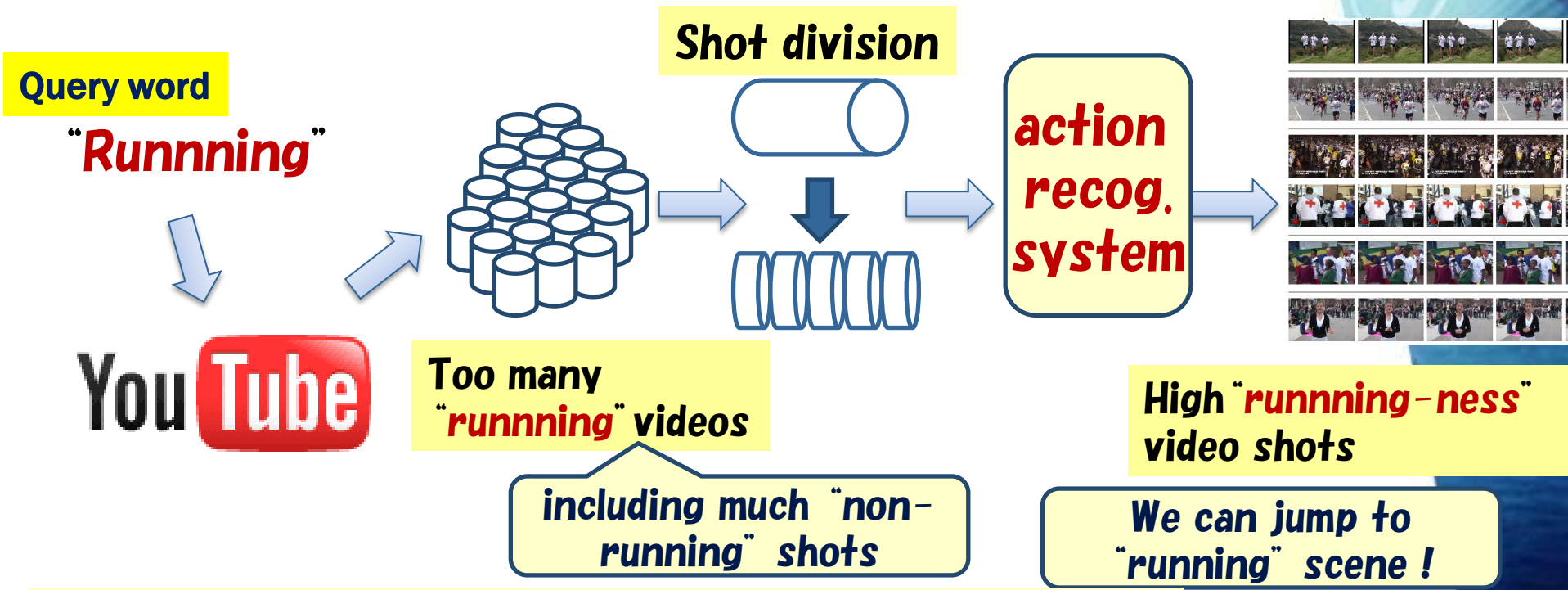
(1)&(2) Kernel weights estimated by MKL



- Motion is important for KTH, while appearance and ST are more important for Wile Youtube.
- This result shows the differences of the characteristic of two datasets.

Application : video shot ranking

- So many videos on the Web
  - Action recognition is useful to search them in addition to object recognition.



This is the same as TRECVID semantic indexing (SIN) task !

Datasets

- **Web video-shot ranking (single class)**
  - **(3) "Our Youtube" dataset [original dataset]**
    - **974 videos for 6 kinds of query words. 37,197 shots obtained by color-histogram-based shot boundary detection (33 times of Wild-Y)**
    - **Select pos./neg. training shots by hand (30/60)**

11 hours / category

batting


#videos	174
#shots	8,980

running


#videos	170
#shots	7,342

jumping


#videos	174
#shots	6,567

walking


#videos	164
#shots	7,718

shoot


#videos	142
#shots	3,442

eating

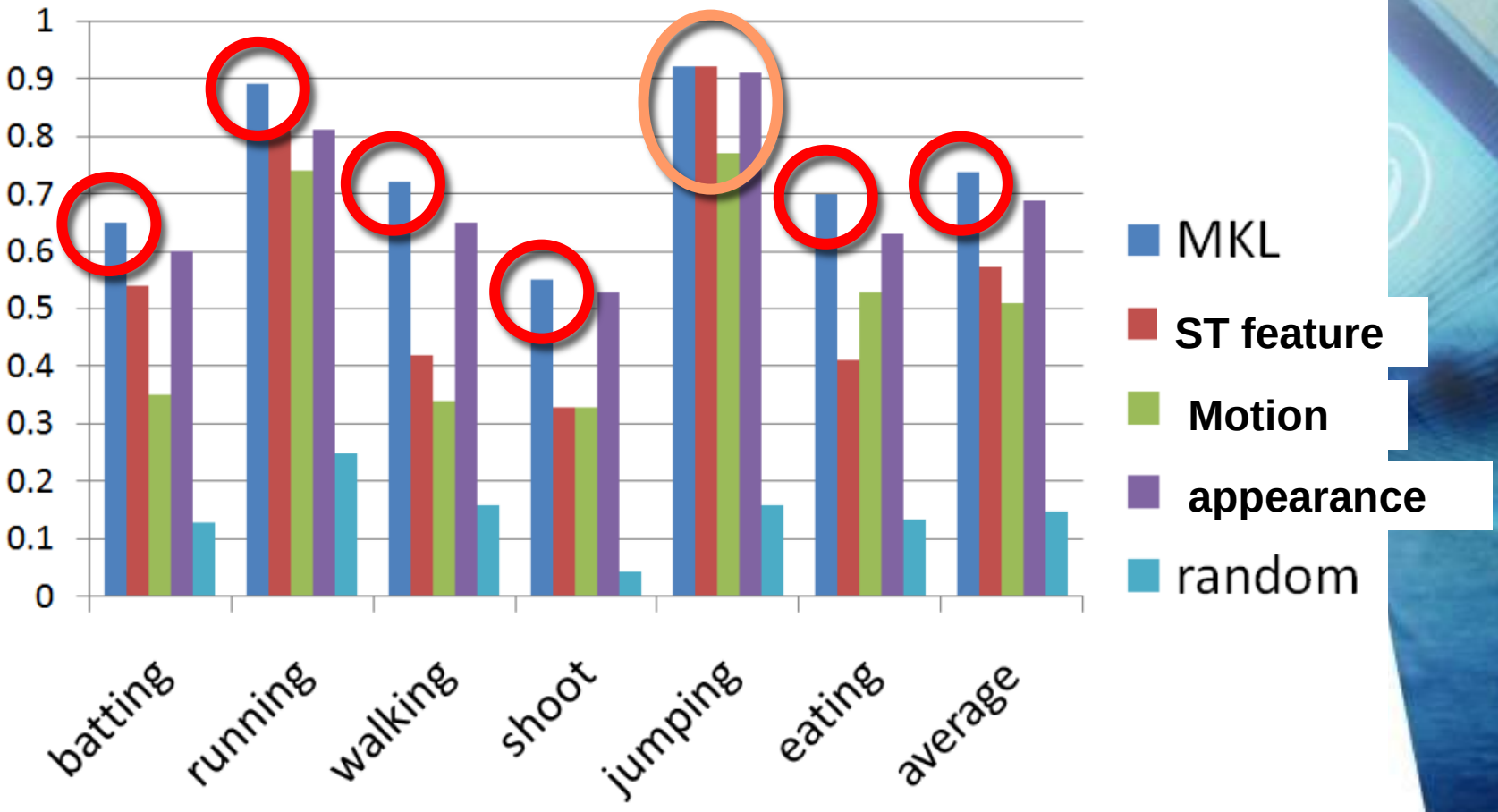

#videos	160
#shots	3,130

(3) "Youtube" shot ranking

- See the results on the browser

(3) "Youtube" shot ranking

- **Average precision until top 200**



5. Conclusions

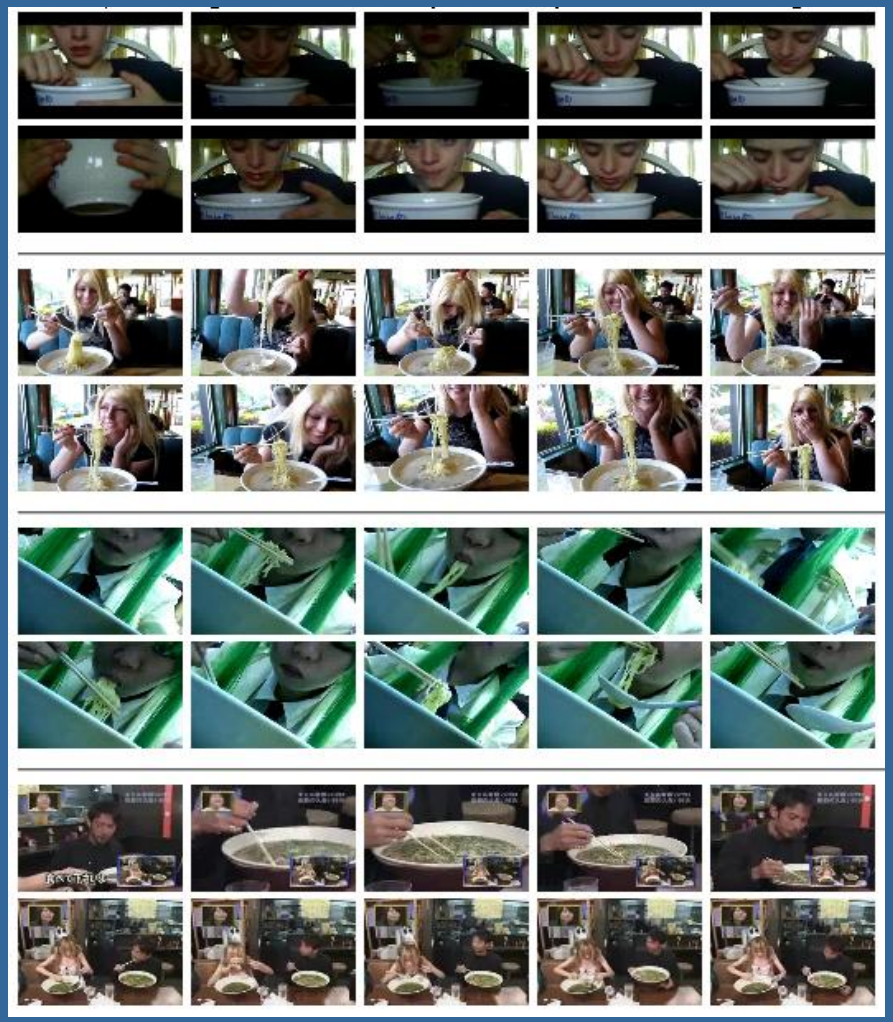
Conclusions

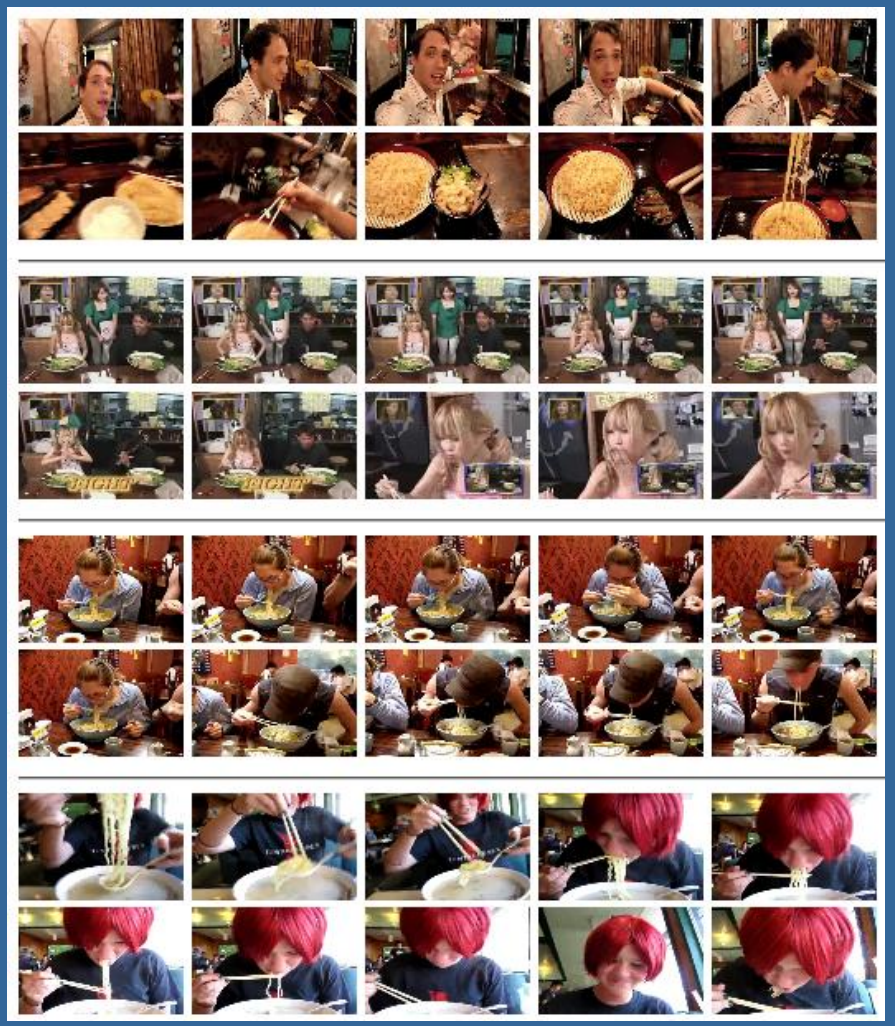
- Proposed a new ST feature extracted by SURF and L-K optical flow detectors with Delaunay triangulation.
- Combined temporally-dense appearance /motion features and the proposed sparse ST features with MKL
  - for KTH, equivalent to state-of-the-art.
  - for "wild Youtube", outperforms greatly. (71.2 ⇒ 80.4%)
  - Dense motion is strong for KTH.

Future work

- Combine other types of features such as Color and SIFT.
- Compensation of camera motion
- Multiple actions
- Apply more large-scale Youtube data
- Weakly-supervised action recognition
  - e.g. learning action from Youtube tags.

Thank you for your attention !





High "eating-Ramen-noodle-ness" video shots from Youtube !

Sorry for showing "eating scenes" just before lunch time !

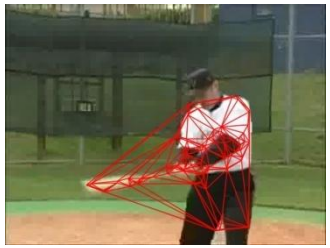
認識手法

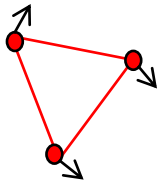
動画データセット

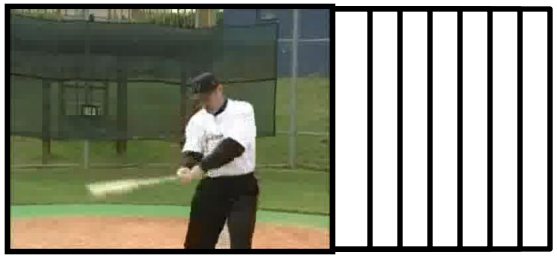
時空間特徴

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

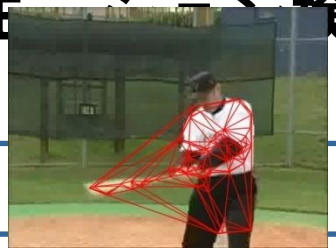
ランキング付けによるWeb動画分類

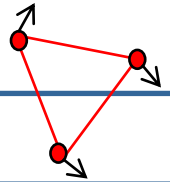
時空間特徴抽出手法(概要)

Step1

・カメラモ[illegible]検出

時空間特徴





Step2

・視覚特[illegible]

ドロネー三角形を
用いた新手法

Step3

・動き特徴抽出

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

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

  - 大規模Web動画ショット分類
    - 教師信号ありのランキング付け
    - 教師信号なしクラスタリング

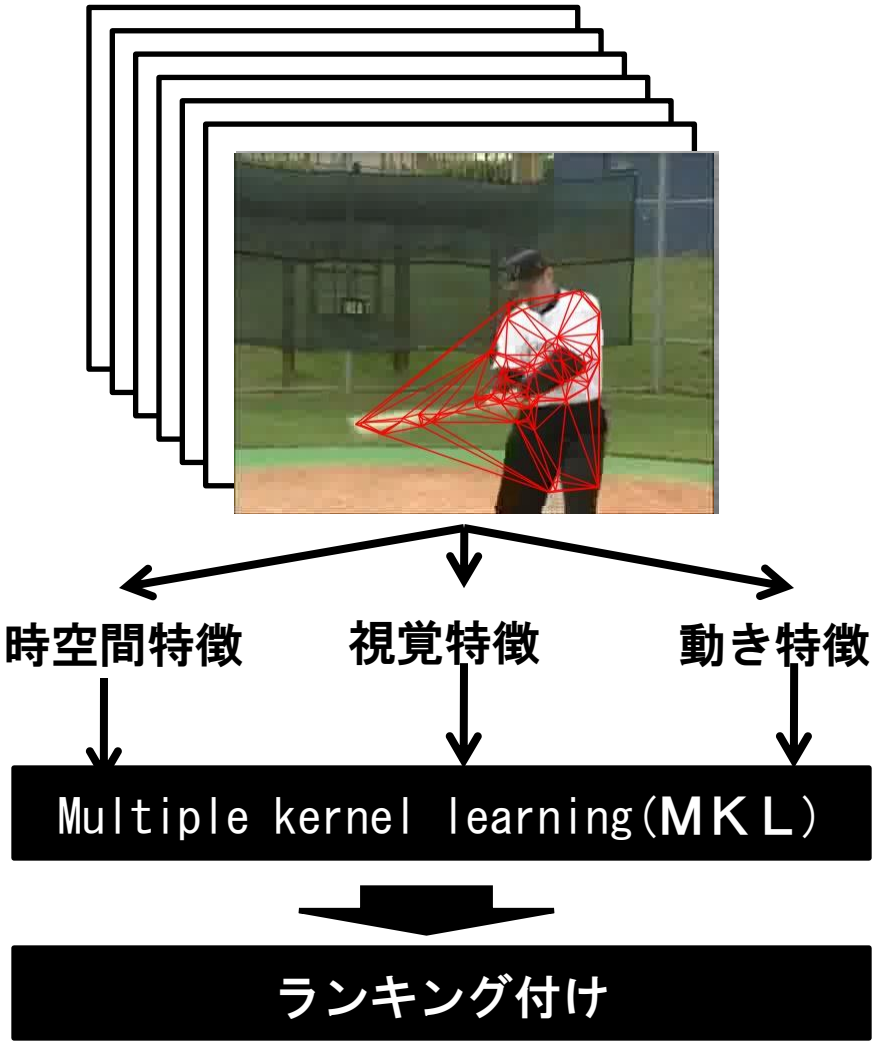
  - 特徴統合によるショット分類手法の提案

教師信号ありの大量Web動画ショット分類

大量のWebショット

Runningの学習セット → 分類器

特徴統合による動作認識



教師信号なしのWeb動画ショットクラスタリング

大量なSoccerタグのWebショット

ドリブルシーン

シュートシーン

インタビューシーン

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

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

時空間特徴抽出手法

- Web動画の特徴
  - データ量が非常に大きい
  - カメラーモーションを含む
  - 手振れなどによる動きのノイズ
  - 低い解像度
  - 雑多な背景ノイズ
  - 撮影の視点変更
- Web動画からの特徴抽出に重要なこと
  - 高速に抽出可能

時空間特徴抽出手法(概要)

カメラモーション検出
カメラモーションを検出したフレームは破棄

↓

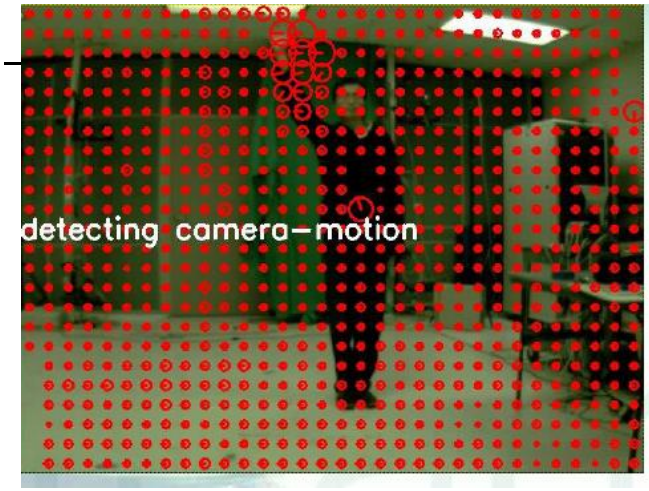
時空間特徴抽出

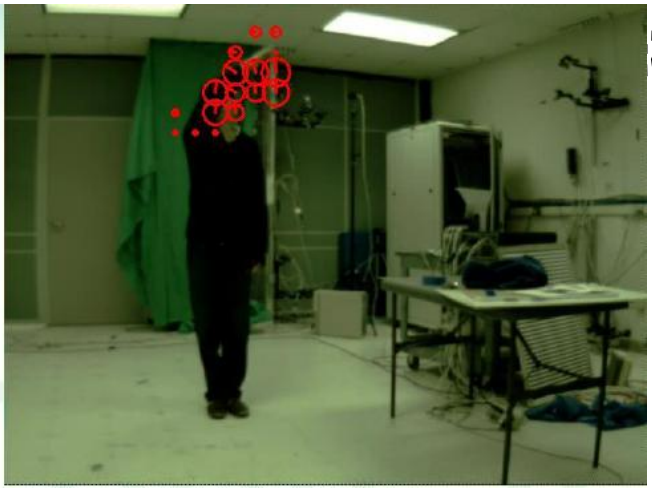
↓

特徴をBag-of-Spatio-Temporal-Features(BoSTF)で表現

カメラモーション検出

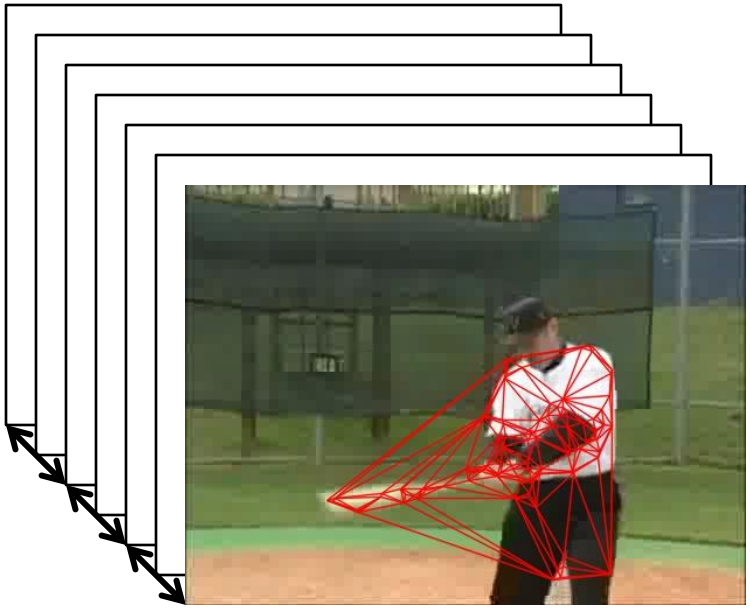
- グリッドで動きを計算
  - 動いていた領域が一定割合以上ならカメラモーション

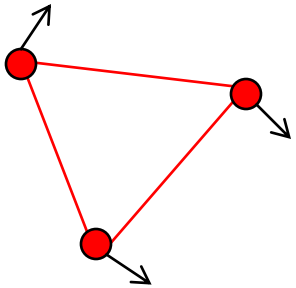




破棄

時空間特徴抽出手法

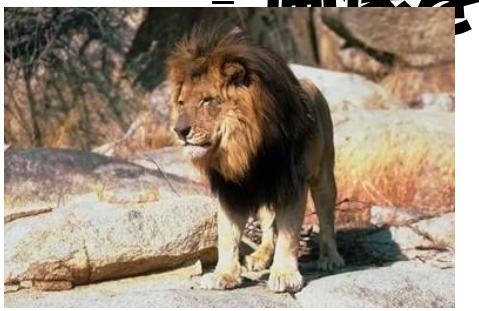




①. Nフレームを１ユニットとする
②. SURFを抽出
③. 動きがない点を削除
④. ドロネー三角形を作成
　以降三点で一組の特徴と考える
⑤. ユニットを更に区切り，それぞれの
　インターバルから動き特徴を抽出

Bag-of-Spatio-Temporal-Features

- Bag-of-Features(BoF)を動画に拡張したもの
  - 画像を特徴の出現頻度で表現したもの



特徴統合による分類手法

- 重要な特徴は異なる

MKLで自動で重みを算出

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)

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

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

  - 凸面最適化問題として解く

特徴統合による分類手法

- 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
- ◆すべてのフレームから特徴を抽出
- ◆抽出された特徴をベクトル量子化
- ◆フレームから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

時空間特徴，視覚特徴，動き特徴の３つをMKLで統合

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

評価実験

- 動作認識
  - KTHデータセット　Leave-one-outで学習

- Web動画分類
  - 教師信号ありのランキング付け
  - 教師信号なしのクラスタリング

データセット(動作認識)

- KTHデータセット
  - 6種類の動作，合計599ショット

データセット(Web動画分類)

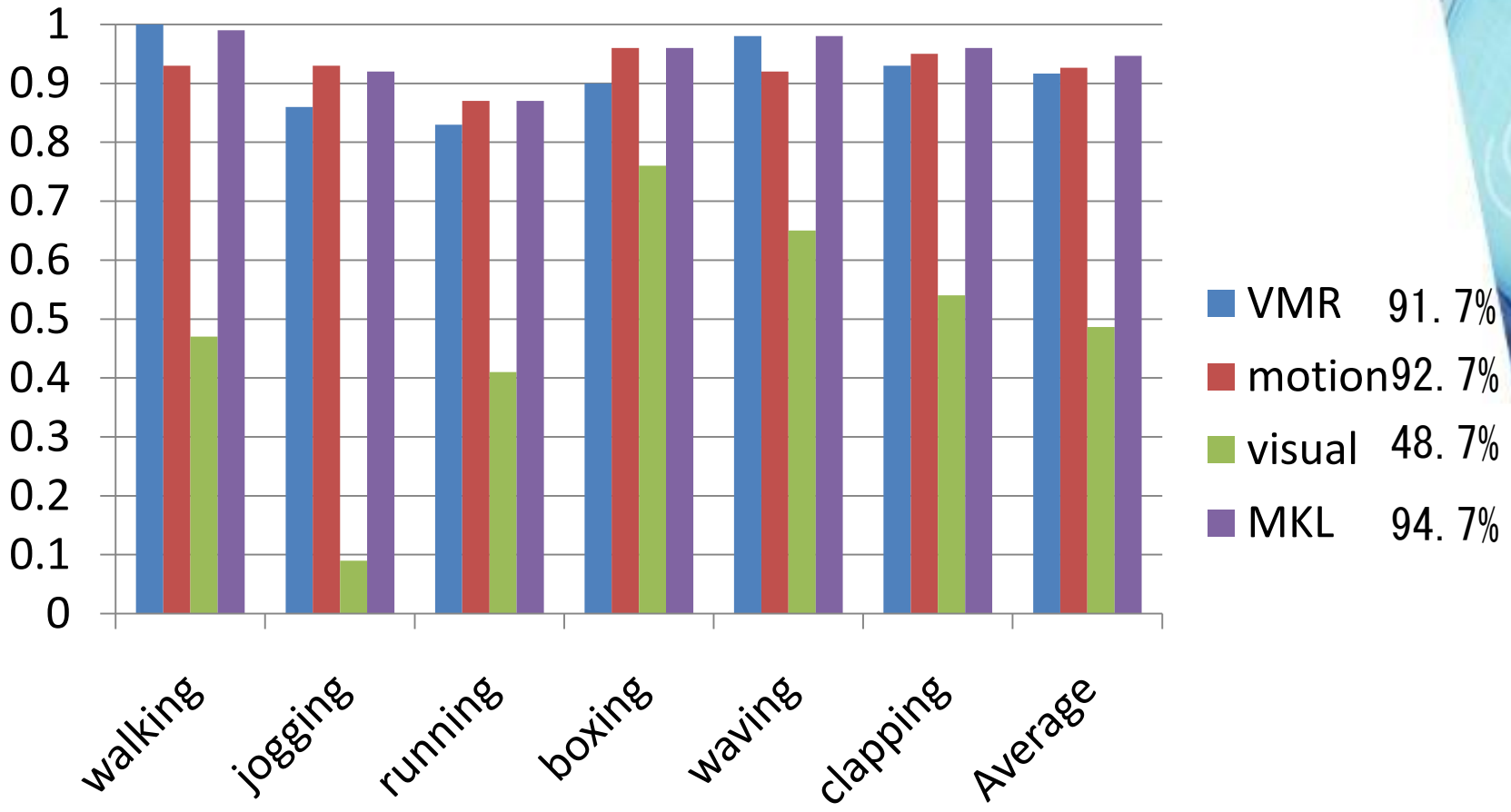
- 教師信号ありランキング付け

動作	動画数	ショット数	学習データ	
			positive	negative
batting	174	8,980	31	75
running	170	7,342	28	66
walking	174	6,567	23	63
shoot	164	7,718	14	75
eating	142	3,442	22	64
jumping	160	3,130	27	40
dancing	185	8,235		
soccer	178	10,430		
合計	948	37,179	145	385

- ラスタリング
  - クラスタ数は200に設定

実験結果

- KTHデータセット



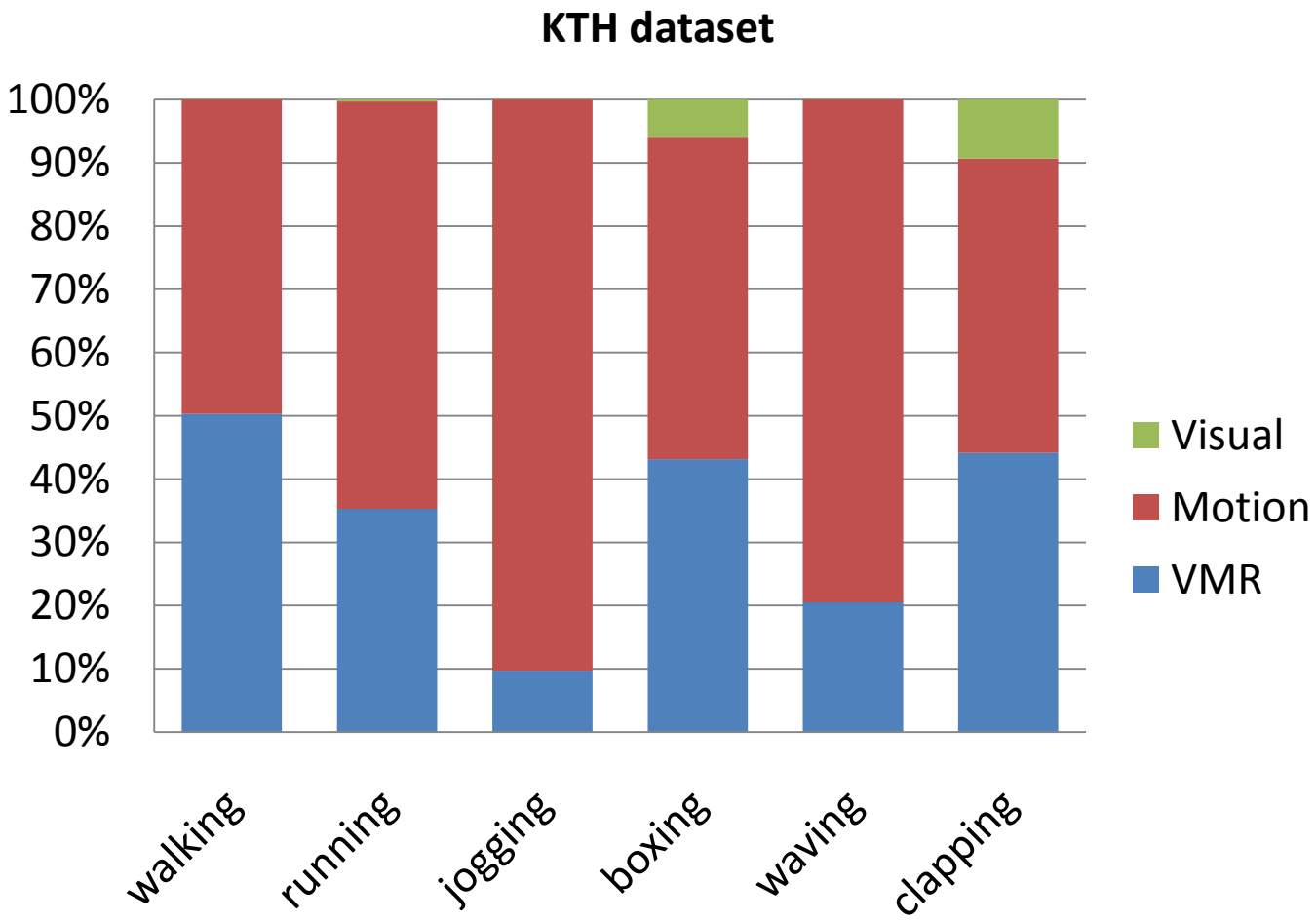
実験結果

- KTHデータセット

	walking	jogging	running	boxing	waving	clapping
walking	0.99	0.01	0	0	0	0
jogging	0.04	0.92	0.04	0	0	0
running	0	0.13	0.87	0	0	0
boxing	0.01	0	0	0.96	0	0.03
waving	0	0	0	0	0.98	0.02
clapping	0	0	0	0.04	0	0.96

実験結果

- KTHデータセット



実験結果

- KTHデータセット

分類結果（Leave-one-out）	
Ours	94.7%
Liu et al.	93.8%
Gilbert et al.	96.2%

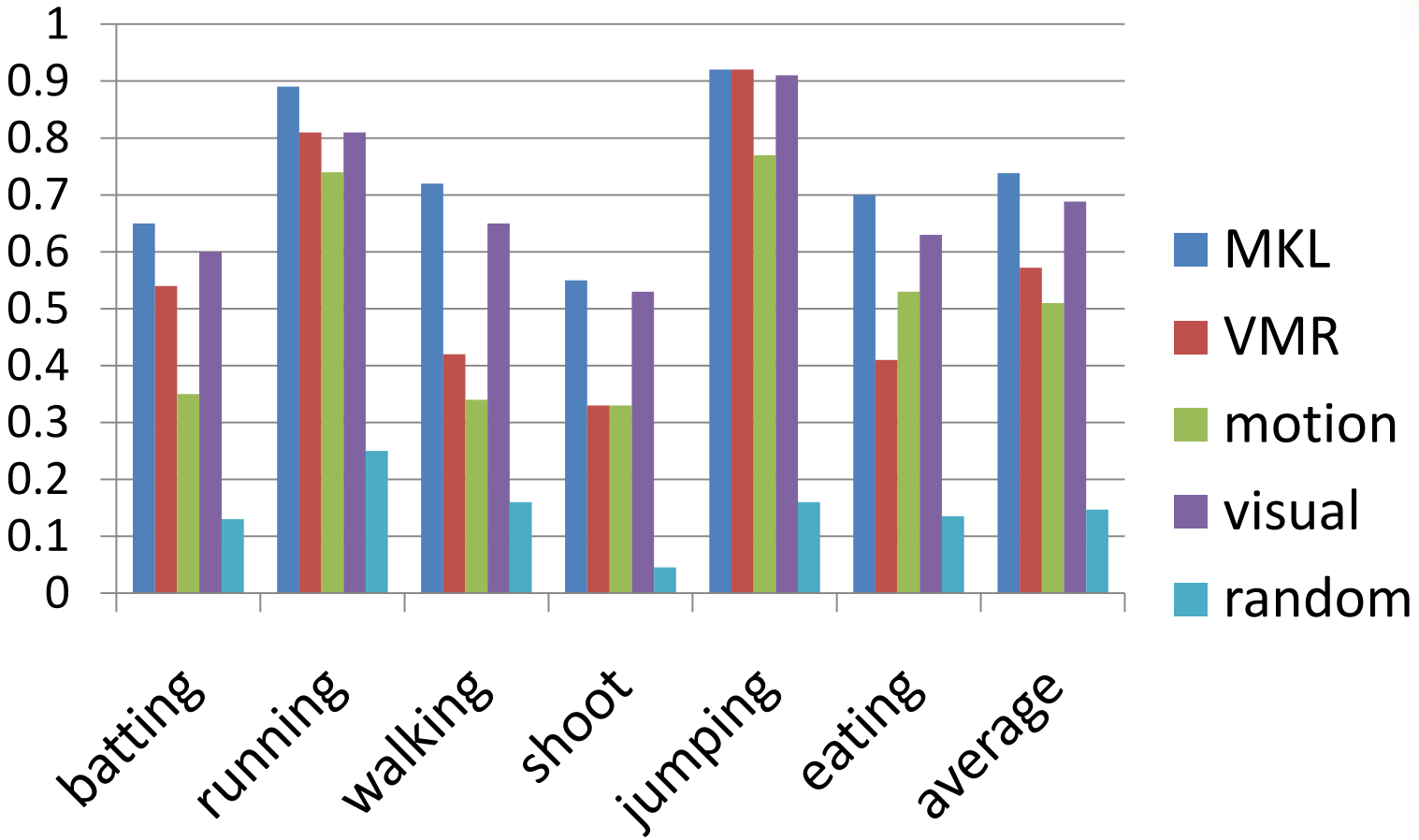
Web動画分類

- 教師信号ありランキング付け

結果のデモ

実験結果

- 教師信号ありランキング付け



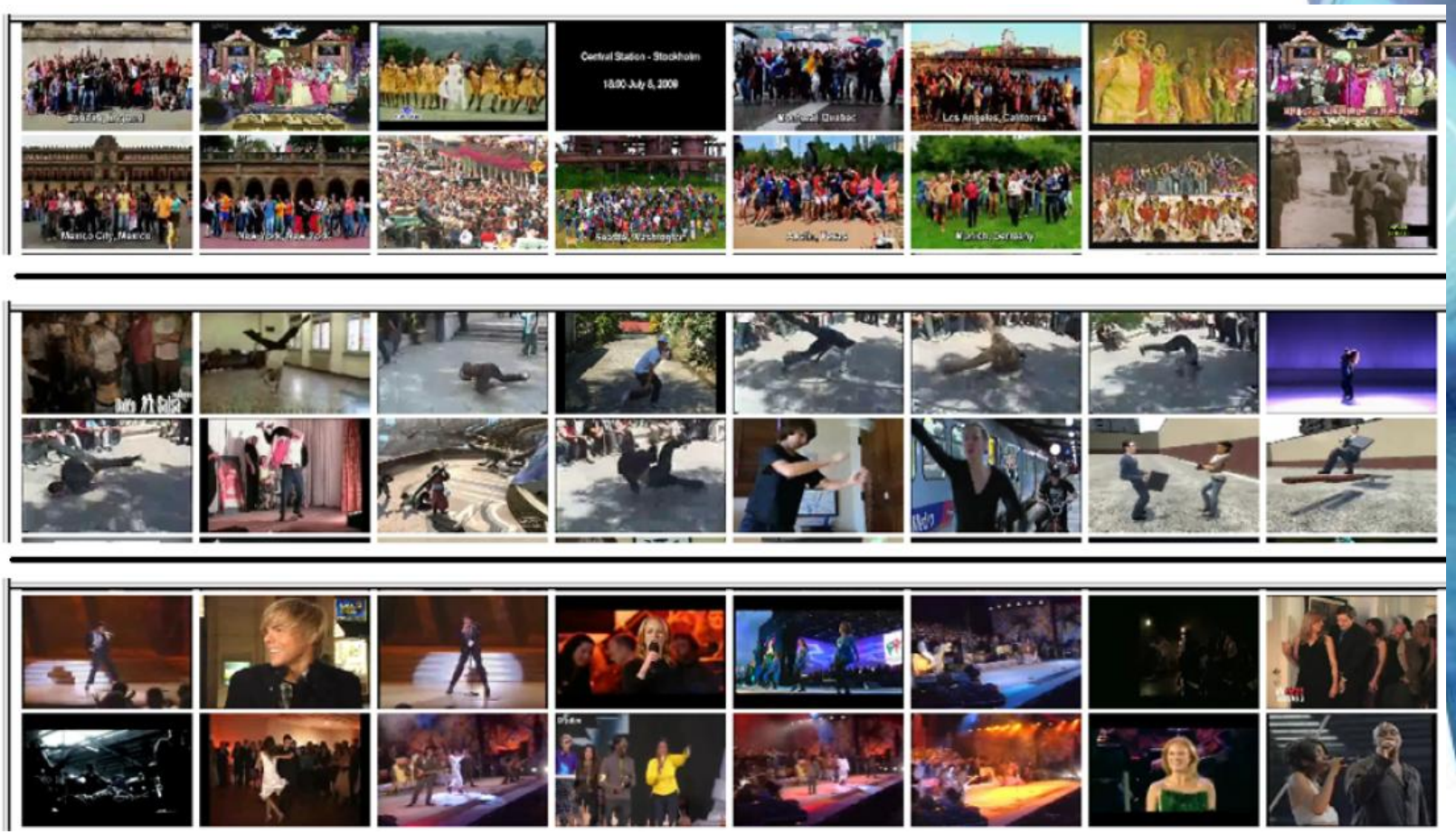
Web動画分類

- 教師信号なしクラスタリング

結果のデモ

Web動画分類

- 教師信号なしクラスタリング



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**

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
    - Web動画分類において高い精度
    - MKLによる特徴統合はWeb動画において有効
- 今後の課題
  - 時空間特徴抽出手法に関する課題
    - 複数モーションに動き補正
    - 有益な特徴の選択

ご清聴ありがとうございました