Automatic Construction of Action Datasets using Web videos with Density-based Cluster Analysis and Outlier Detection

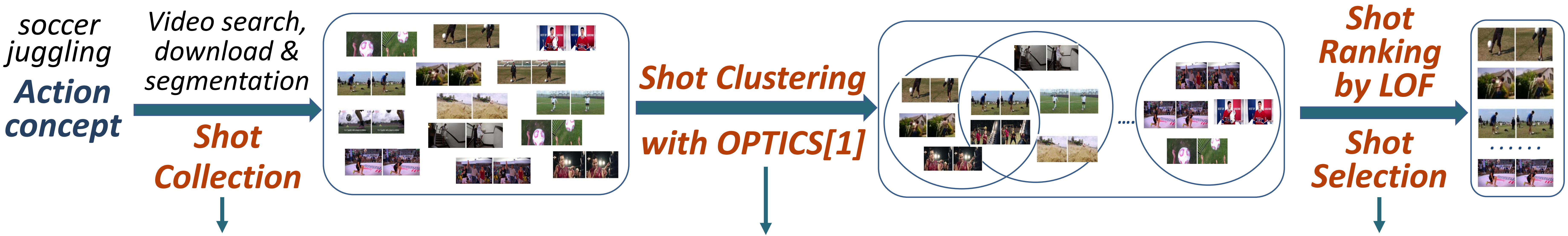
Nga Do and Keiji Yanai (The University of Electro-Communications, Tokyo)

Introduction

Web videos
Action Concept
Automatic Approach
Relevant Shots
Action Dataset

- Previous work: require additional data (e.g.: tags[3]), ignore concept diversity problem
- This work: can exploit Web videos without tags, copes with concept diversity

Proposed Approach



i. Word preparation

- "verb" (dive), "verb+non-verb" (throw hammer), "non-verb" (vault)

ii. Video search

- "verb" & "verb-ing" (dive & diving)

iii. Video filtering

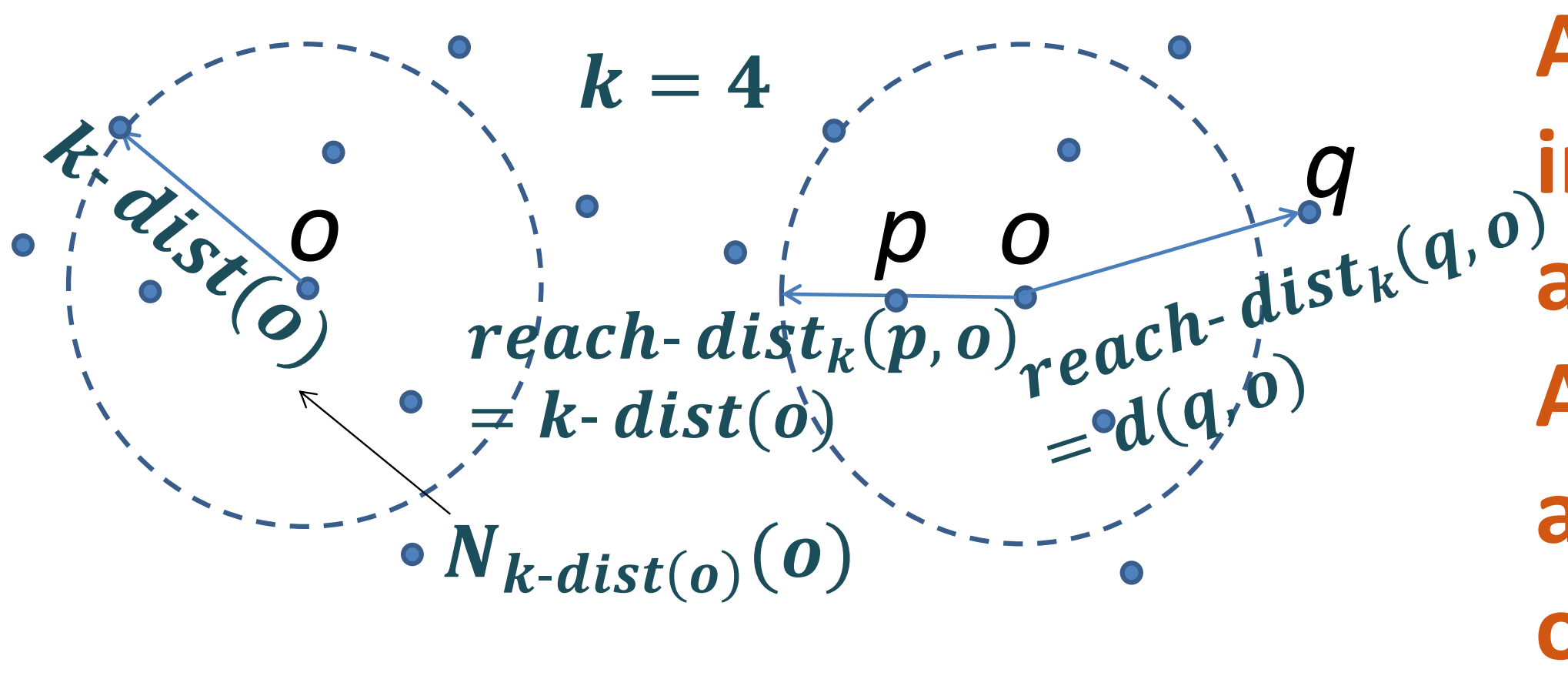
- No videos of "Entertainment"

iv. Video downloading

- Web API (e.g. Youtube API)

v. Shot segmentation

- Color histogram


$k = 4$
$k\text{-}dist(o)$
o
p
o
q
$reach\text{-}dist_k(p,o) = k\text{-}dist(o)$
$reach\text{-}dist_k(q,o) = d(q,o)$
$N_{k\text{-}dist(o)}(o)$

A low reachability distance indicates an object within a cluster.

A high reach-dist indicates a noise or a jump from one cluster to another.

$$k\text{-}dist(o) = d(o,p): \begin{cases} 1.\ at\ least\ k\ objects\ q:\ d(o,q) \le d(o,p) \\ 2.\ at\ most\ k-1\ objects\ q:\ d(o,q) < d(o,p) \end{cases}$$

$$reach - dist(p,o) = max\big(k\text{-}dist(o), d(p,o)\big)$$

As visual features, we extract motion features using ConvNet models trained on UCF-101 dataset (split 1) with multi-frame stacking optical flows[4].

LOF (Local Outlier Factor) [5]

$$LOF_{MinPts}(p) = \frac{\sum_{o \epsilon N_{MinPts-dist(p)}(p)} \frac{MinPts - dist(p)}{MinPts - dist(o)}}{\left|N_{MinPts-dist(p)}(p)\right|}$$

Small $MinPts - dist$ corresponds to a region with high density.

Shots with low LOF are considered as relevant shots and ranked to the top.

Shots are selected from all clusters to guarantee diversity of selection results.

Experiments and Results

Experiment 1: Dataset Construction

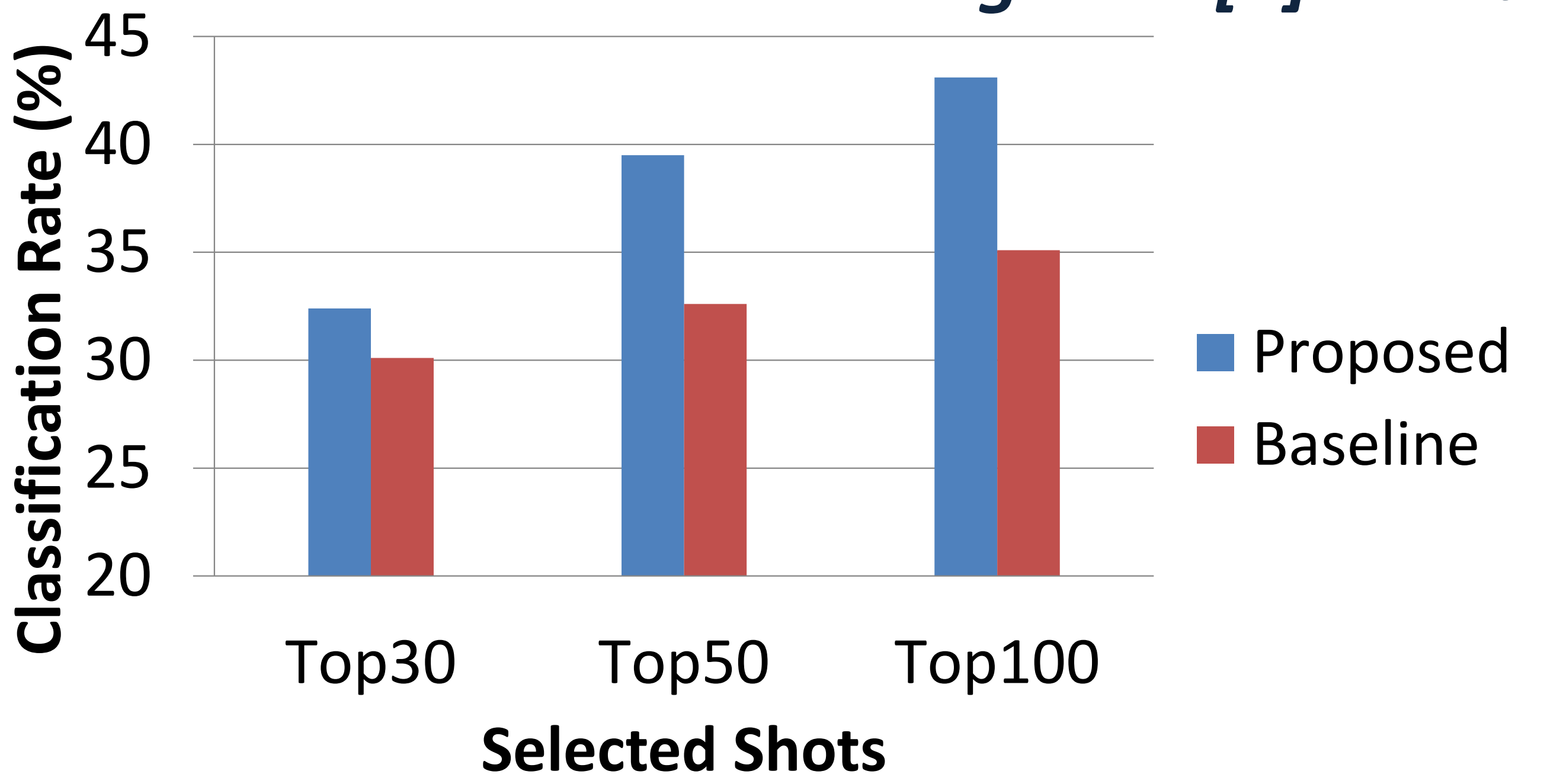
- Data: Web videos (YouTube)
- Actions: 11 actions in UCF11[2]
- Precision rate = percentage of relevant shots among top 100 shots [3]
- Baseline[3]: VisualRank based method

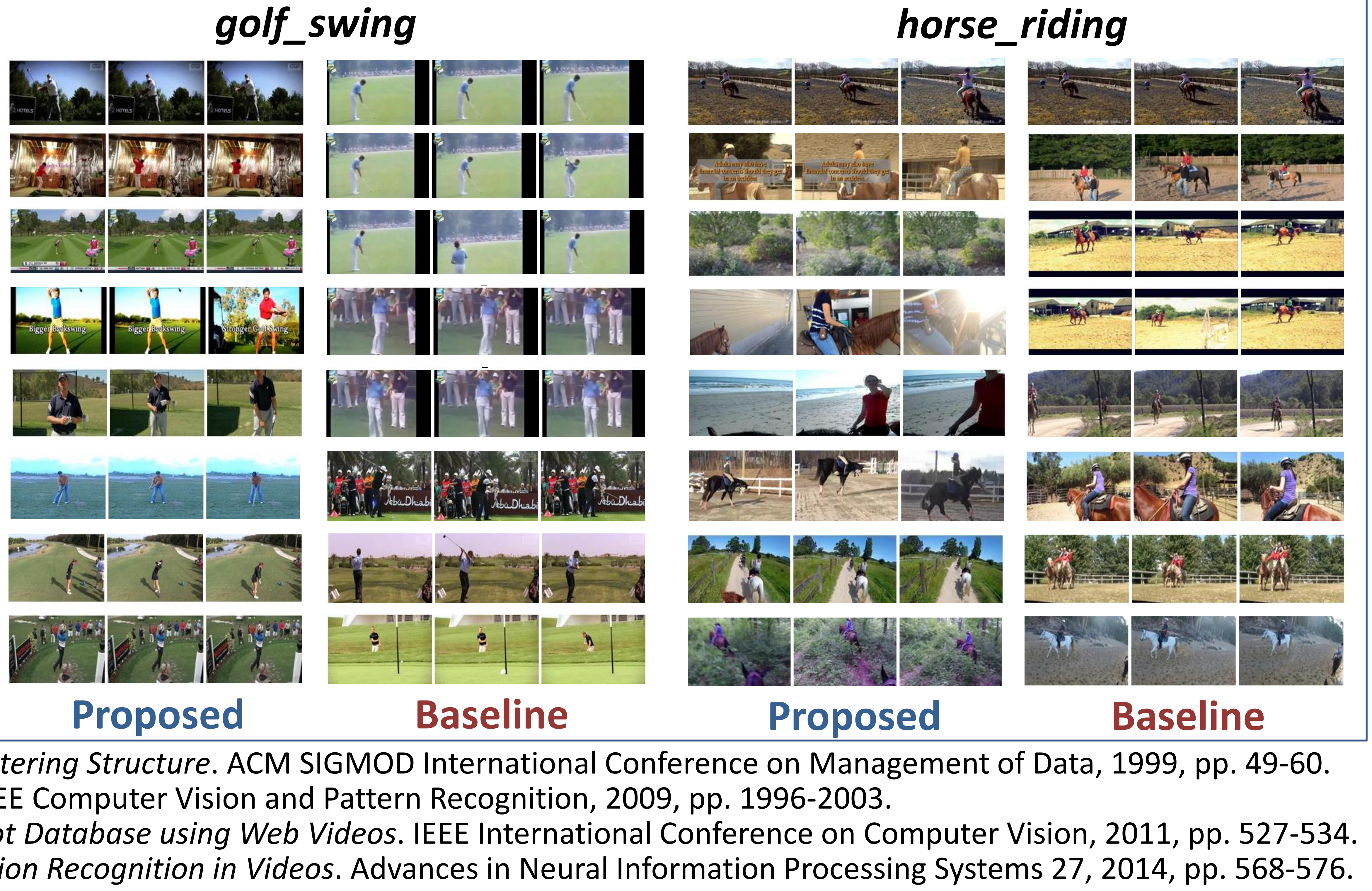
Action	Proposed	Baseline	Action	Proposed	Baseline
basketball	**50**	35	swing	**36**	31
biking	**23**	17	tennis_swing	47	**51**
diving	**35**	28	trampoline_jumping	**54**	**54**
golf_swing	52	**54**	volleyball_spiking	58	**69**
horse_riding	**50**	42	walking	**14**	9
soccer_juggling	**68**	63	Average	**44.3**	41.1

Experiment 2: Action Classification

- Dataset: UCF11[2]
- Precision = average of 25-fold validation
- Training data: standard data[2] & shots automatically obtained in Experiment 1

With standard training data [2]: 81.5%





[1] Mihael et al. *OPTICS: Ordering Points To Identify the Clustering Structure*. ACM SIGMOD International Conference on Management of Data, 1999, pp. 49-60.
[2] Jingen et al. *Recognizing realistic actions from videos*. IEEE Computer Vision and Pattern Recognition, 2009, pp. 1996-2003.
[3] Nga et al. *Automatic Construction of an Action Video Shot Database using Web Videos*. IEEE International Conference on Computer Vision, 2011, pp. 527-534.
[4] Karen et al. *Two-Stream Convolutional Networks for Action Recognition in Videos*. Advances in Neural Information Processing Systems 27, 2014, pp. 568-576.
[5] Chiu et al. *Enhancements on local outlier detection*. IEEE Database Engineering and Applications Symposium, 2003, pp. 298 – 307.