A Multi-viewpoint Outdoor Dataset for Human Action Recognition

Asanka G. Perera, Yee Wei Law, Titilayo T. Ogunwa, and Javaan Chahl, *Member, IEEE*

Abstract—Advancements in deep neural networks have contributed to near perfect results for many computer vision problems such as object recognition, face recognition and pose estimation. However, human action recognition is still far from human-level performance. Owing to the articulated nature of the human body, it is challenging to detect an action from multiple viewpoints, particularly from an aerial viewpoint. This is further compounded by a scarcity of datasets that cover multiple viewpoints of actions. To fill this gap and enable research in wider application areas, we present a multi-viewpoint outdoor action recognition dataset collected from YouTube and our own drone. The dataset consists of 20 dynamic human action classes, 2324 video clips and 503086 frames. All videos are cropped and resized to 720×720 without distorting the original aspect ratio of the human subjects in videos. This dataset should be useful to many research areas including action recognition, surveillance and situational awareness. We evaluated the dataset with a two-stream CNN architecture coupled with a recently proposed temporal pooling scheme called kernelized rank pooling that produces nonlinear feature subspace representations. The overall baseline action recognition accuracy is 74.0%.

Index Terms—Video dataset, human action recognition, multi-viewpoint, kernelized rank pooling.

I. INTRODUCTION

INTERPRETTING the dynamics of the human body in videos and inferring the action the human subject is performing is a challenging research problem. Sometimes, even for a human observer, it is difficult to accurately recognize actions in many scenarios. Such scenarios include scenes with occlusions, motion blur, low resolution, perspective distortion, etc. However, popular video datasets such as [1–8] have provided a stimulus to research in human action recognition, with many successes reported over the last few years based on deep learning techniques.

Owing to the articulated nature and infinite poses the human body can make, human action recognition is a challenging research problem, and is far from resolved. Recent advances in autonomous driving have created a lot of interest among researchers attempting to understand human actions using different data capturing methods. This includes analyzing challenging videos recorded from moving cameras [9–11]. There are a number of datasets available to study road environments especially focusing on the behavior of pedestrians and vehicles [9, 10, 12–14]. Research related to autonomous vehicles rely heavily on such datasets.

In addition to autonomous ground vehicles such as self-driving cars and buses, there is increasing commercial and research interest in application areas such as search and rescue [15, 16], surveillance, crowd monitoring [17, 18], sports

activity recording [18] and situational awareness. In oder to achieve human-level performance in such challenging missions a significant amount of training data and affective algorithms are needed.

To meet these demands, datasets that cover common outdoor actions from multiple viewpoints are necessary. Currently most available video datasets are limited to ground-based action recognition *or* aerial action recognition. It is difficult to find a single dataset that contains clips from both aerial and ground viewpoints covering realistic outdoor actions. Our work presented here is specifically focused on providing a *multi-viewpoint* and single/multi-person video dataset to the research community to advance research in outdoor human action recognition.

The dataset presented in this article consists of 20 action classes collected from YouTube and our own drone. Most of the actions with an aerial viewpoint were recorded from an aerial platform (UAV or helicopter) or a rooftop camera while the actions with ground-level viewpoints were recorded from a fixed or moving camera. This dataset is suitable for use on research involving action recognition, situational awareness and surveillance. All the videos were cropped around the action and resized to a resolution of 720×720 without distorting the original aspect ratio of the action.

As the baseline action recognition algorithm, we used a two-stream CNN architecture coupled with a recently proposed temporal pooling scheme called *kernelized rank pooling based on feature subspaces* (KRP-FS) [19] that produces nonlinear spatiotemporal feature subspace representations. KRP-FS transforms nonlinear action dynamics to a low-rank subspace in the (potentially) infinite-dimensional reproducing kernelized Hilbert space (RKHS) [20], and creates linear and efficient feature descriptor for action classification. The baseline accuracy of KPR-FS for our dataset was found to be comparable to those for recently published human action datasets.

The rest of this article is organized as follows. Section II discusses closely related work on popular action recognition datasets and their limitations in terms of suitability for multi-viewpoint human action recognition. Section III describes the steps involved in preparing the dataset, and compares the dataset with other recently published video datasets. Section IV reports experimental results. A discussion of issues and potential improvements is presented in Section V. Section VI concludes.

II. RELATED WORK

The surveys [21] and [22] compare action recognition datasets that were published up to the years 2013 and 2016

respectively. Here, we discuss datasets related to our work.

Early datasets: The problem of recognizing human pose in statically captured videos has been studied extensively in recent literature [22]. Early datasets such as KTH [23] and Weizmann [24] have helped progress action recognition research. However, the videos were captured indoor using a static camera. The static backgrounds used restrict their applicability in a variety of studies. One of the largest and most diverse datasets in this category is UCF101 [1], with 13320 selected instances from YouTube videos. The videos belong to 101 action classes with dynamic camera movements and different backgrounds. However, the actions are mainly related to indoor and sports events with a low resolution of QVGA (320×240).

Recent small datasets: Recent datasets created from YouTube and movie clips include JHMDB [3] and Penn action [4]. Videos in these datasets have been annotated for both pose and action. JHMDB has 21 action classes across 928 videos (320x240 resolution) and Penn action consists of 15 actions across 2326 video sequences with a maximum dimension of 480 pixels (width or height). These datasets contain both indoor and outdoor scenes and static and dynamic camera movement. However, the types of action and limited outdoor scenes from a moving camera make them inappropriate for multi-viewpoint action recognition.

Recent large datasets: The datasets discussed so far are limited in size. Large action datasets such as Sports-1M [5] and ActivityNet [6] were introduced in the recent past. They mainly focus on actions recorded from ground cameras and do not provide spatial localization details of the human subjects. For example, Sports-1M videos are annotated automatically using YouTube topics and are therefore, weakly labeled [7].

There are large action recognition datasets that were prepared from realistic YouTube videos. ActivityNet [6], YouTube-8M [25] Kinetics [8] and HACS [26] are such popular datasets. The first release of Kinetics dataset [8] consists of 400 action classes with 400 minutes of video. It covers a diverse range of actions collected from YouTube videos. The successive Kinetics releases have 600 and 700 classes respectively. The HACS [26] dataset consists of 1.55 M annotated clips across 200 classes. The diversity of these large datasets helps to pre-train action recognition network on them and enhance the network performance on small datasets.

Aerial action recognition datasets: Of increasing interest to the action recognition community is aerial videos because drones are an increasingly popular means of capturing videos. Human action detection from an aerial view proves more challenging than from a fronto-parallel view. Oh et al. [27] presented a large-scale VIRAT dataset with about 550 videos. The 29 hours of video in the dataset cover 23 types of events, recorded from both static and moving cameras. Using resolutions of 1920×1080 and 1280×720, the VIRAT ground dataset was recorded using stationary overhead cameras placed in multiple locations. However, in the case of the VIRAT aerial dataset, using a low resolution of 480×720 increased the difficulty in retrieving rich activity information from human subjects that cover a relatively small part of the frame. Both the ground and aerial datasets were recorded against cluttered and uncontrolled backgrounds.

The Okutama-Action 4k resolution video dataset introduced in [28] focused on concurrent detection of actions by multiple subjects. A clutter-free baseball field was where the videos were recorded using 2 UAVs. The abrupt camera movements captured 12 actions, taken from various view angles and at altitudes from 10 to 45 meters. The 90 degree elevation angle of the camera resulted in self-occlusions and severe perspective distortion in the videos.

Other important aerial action datasets are Mini-drone [29], UCFARG [30] and UCF aerial action [31]. An R/C-controlled blimp and a helium balloon were used to record videos for UCF aerial action and UCF ARG respectively with similar action classes. While the UCF ARG dataset is multi-viewpoint (recorded from aerial, rooftop and ground cameras), the UCF aerial action is a single-viewpoint dataset. The Mini-drone dataset was developed as a surveillance dataset in order to evaluate various aspects and definitions of privacy. The recording of the dataset was done in a car park with a low-altitude flying drone to detect actions categorized under illicit, suspicious and normal behaviors.

Research on outdoor action recognition would benefit from more datasets, such as the dataset presented here which covers 20 common outdoor actions captured from multiple viewpoints.

III. PREPARING THE DATASET

This section explains the data collection process, the selected action classes and variations in the collected data. The dataset is compared with recently published aerial and ground video datasets.

A. Data collection

We collected a total of 2324 video clips: 6 of them captured using our own quadrotor UAV, and 2318 from YouTube published under the Creative Commons license.

All clips were resized to 720×720 pixels. We cropped and resized the clips from their original videos without distorting their original aspect ratios. All of our clips have an undistorted 1 : 1 aspect ratios making them easy to use as inputs to CNN architectures. All videos were re-sampled at 29.97 fps.

When collecting the data we tried to cover multiple viewpoints of the same actions while prioritizing aerial viewpoints. We discretized the viewing sphere of the action as shown in Fig. 1(a). The viewpoint distribution across the dataset is shown in Fig. 1(b). Among the viewpoints, 54% are aerial viewpoints (viewpoints A and B) while 46% are ground viewpoints (viewpoints C and D). Videos with aerial viewpoints were mostly recorded by an aerial platform (a UAV or helicopter) or rooftop cameras. The dataset contains videos recorded from both fixed and moving cameras.

The collected videos cover realistic scenarios across 20 selected classes (see Figure 2). The 20 selected actions were backpacking, chainsawing_trees, cliff_jumping, cutting_wood, cycling, dancing, fighting, figure_skating, fire_fighting, jetskiing, kayaking, motorbiking, nfl_catches, rock_climbing, running,
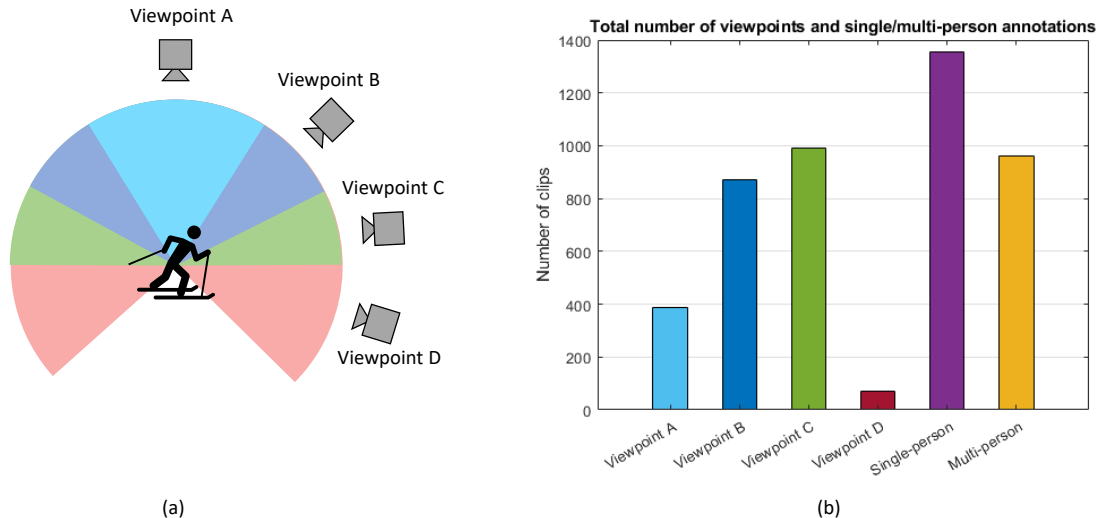
Fig. 1. Each clip was annotated with its viewpoint and number of people (single-person or multi-person). (a) The upper viewing hemisphere is discretized into three sections (viewpoints A, B and C) and any viewpoint falls into the lower viewing hemisphere is considered as viewpoint D. (b) Number of clips per viewpoint type and people count.

stakeboarding, skiing, standup_paddling, surfing and windsurfing. Since our focus was to create an outdoor action recognition dataset we selected 18 outdoor actions consisting of popular outdoor sports (e.g., jetskiing, surfing) and general outdoor events (e.g., chainsawing_trees, fire_fighting). The remaining two actions, namely dancing and fighting, are applicable to both indoor and outdoor settings.

There are substantial variations in the actions in terms of the phase, orientation, camera movement, viewpoint, number of humans in an action, human sizes and the backgrounds. These variations create a challenging dataset for action recognition, and also makes it more representative of real-world situations.

B. Data annotation

All clips were annotated for the action class, viewpoint and the number of people involved in the action. The number of people was annotated as "single-person" or "multi-person".

C. Dataset summary

The dataset contains a total of 2324 clips with 503086 frames, consistently cropped to a resolution of 720×720 and re-sampled at 29.97 fps. The dataset is 280 minutes (i.e., 4 hours and 40 minutes) long, with an average clip length of 7.4 seconds. For every action class, there are a minimum of 105 clips, a maximum of 153 clips, and an average of 116.2 clips. Table I provides a summary of the dataset.

Figure 3(a) shows the total clip length (blue) and mean clip length (amber) for each action class. Actions cliff_jumping and nfl_catches have a relatively low clip durations since both are very short actions. The clip durations are grouped and shown in Figure 3(b). The longest clip duration is 21.3 seconds and the shortest is 1.84 seconds. The majority of the clips lie in the 2-10 seconds range.

Fig. 3(c) shows the viewpoint breakdown across the classes. There is a low number of clips with viewpoint D. This is due

TABLE I
A SUMMARY OF THE DATASET.

Feature	Value
# Actions	20
# People	One to many
# Clips	2324
# Clips per class	116.2
Mean clip length	7.4 sec
Total duration	280 mins
# Frames	503086
Frame rate	29.97 fps
Resolution	720×720
Viewpoints	Multi-viewpoints
Annotations	Action, single/multi-person, viewpoint
Sources	YouTube and own drone

to the nature of actions selected here. The rock_climbing action has the largest number of viewpoint D clips, because most of the rock_climbing actions were recorded by a person on the ground. Actions chansawing_trees and cutting_wood were mostly recorded from the ground-level viewpoint C. In contrast, the windsurfing action has the largest number of aerial viewpoints owing to the aerial camera platforms used to record them. The other 16 actions have a relatively balanced aerial and ground viewpoint distribution.

Figure 3(d) shows the distribution of single-person versus multi-person clips per action class. There is a mix of single-person and multi-person clips for all classes excepting cutting_wood, fighting and nfl_catches. cutting_wood is normally a single-person action, and fighting and nfl_catches are always multi-person actions.

D. Comparison with other datasets

In Table II, we compare our dataset with some recently published video datasets. These datasets have helped to advance research in action recognition and event recognition. Datasets such as UCF101, HMDB51 and JHMDB have been used as the standards in recent action recognition studies. The
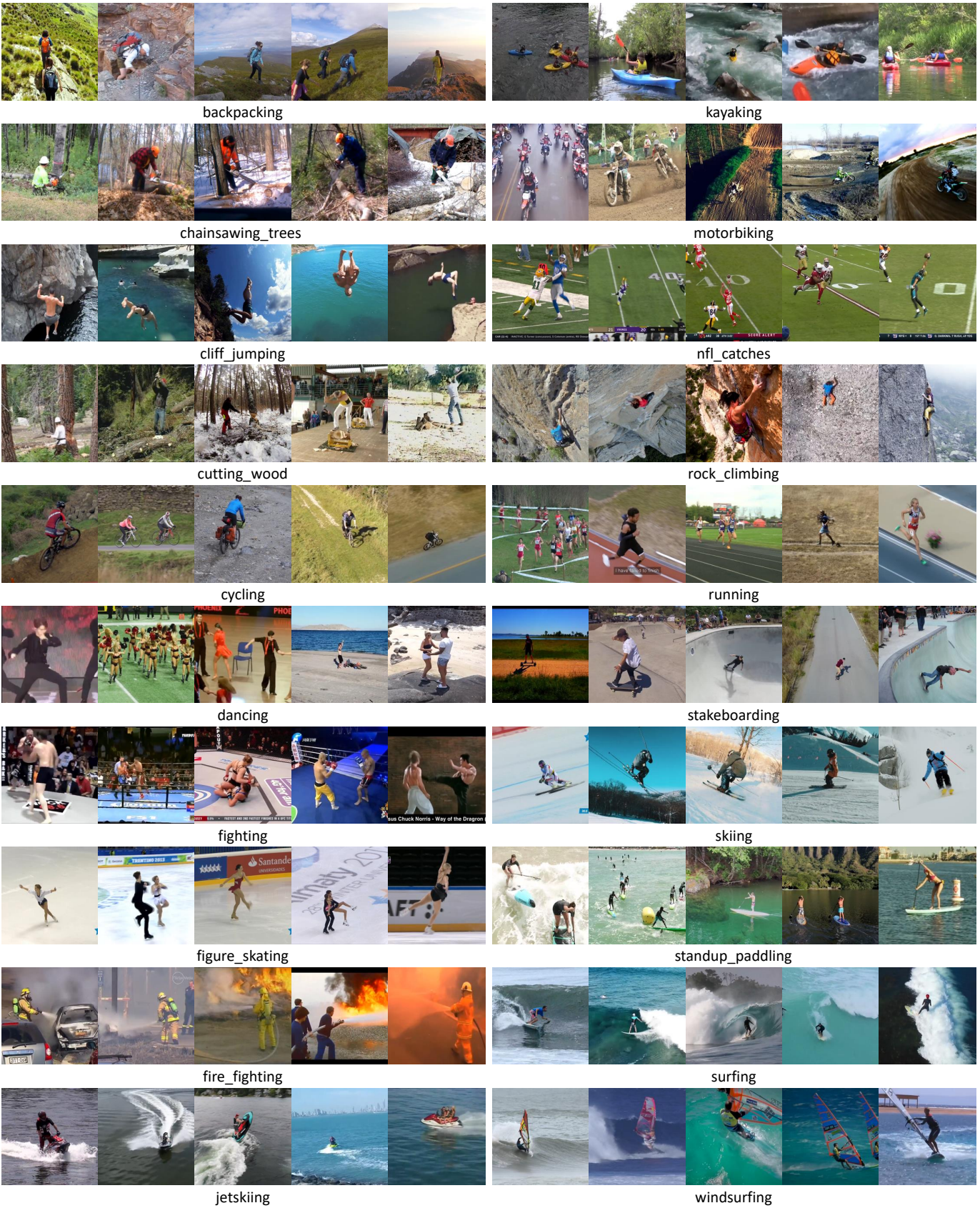
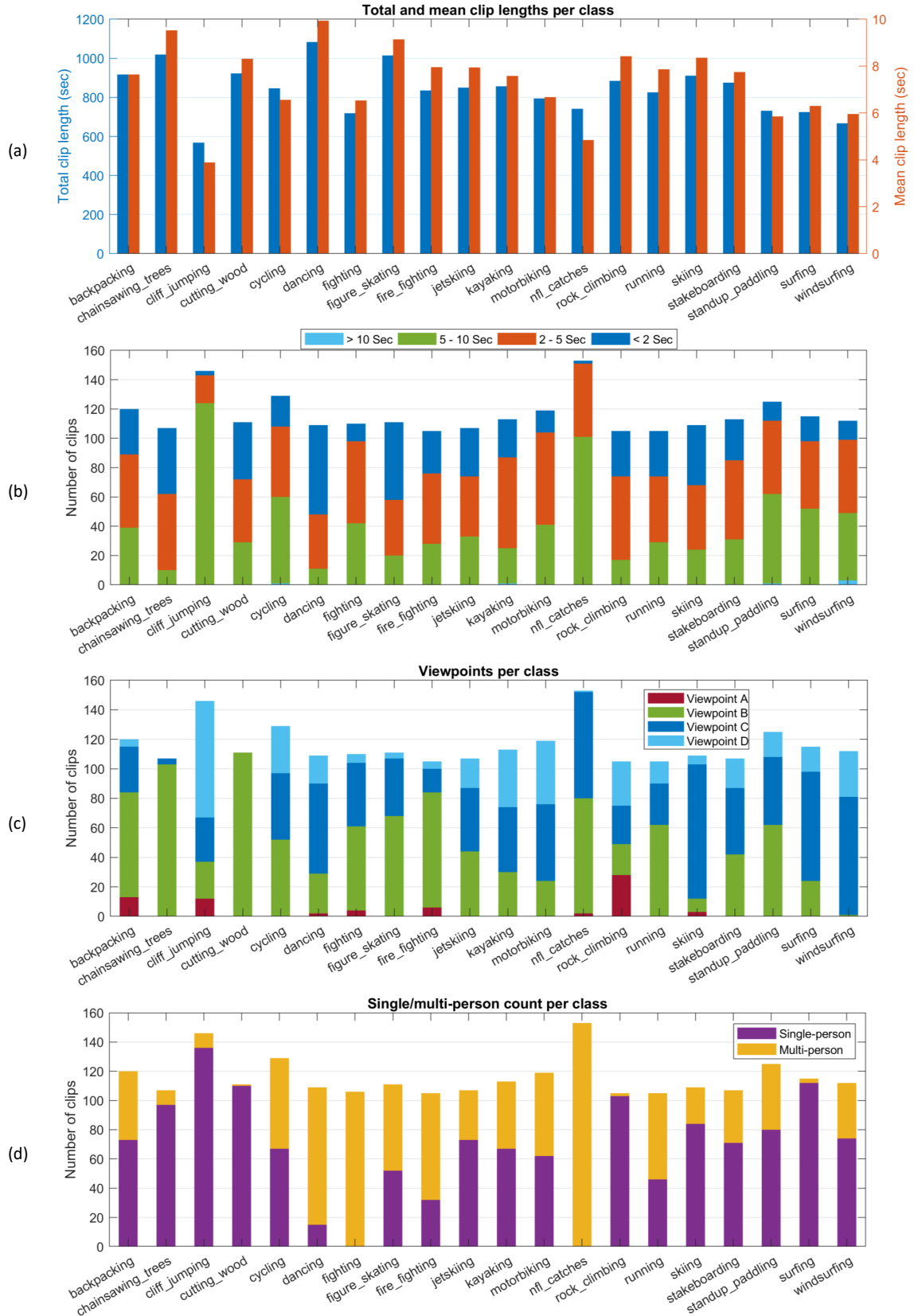Fig. 2. Randomly selected images from the 20 classes.

Fig. 3. (a) Total and mean clip lengths per action class. (b) Clip lengths grouped per action class. (c) Viewpoint distribution per action class. (d) Distribution of single/multi-person clips per action class.

design of our dataset is similar to these datasets, but there are some unique differences in our dataset. We selected a range of mostly outdoor actions and collected videos to cover multiple viewpoints. Further, the resolution of our dataset is 720×720 compared to the 360×240 resolution of UCF101, HMDB51 and JHMDB.

Drone-recorded datasets like VIRAT, Mini-drone and Okutama-Action are focused on various degrees of outdoor action recognition. Compared to these datasets, our dataset covers a more diverse range of outdoor actions. Recently released Kinetics[1] and HACS[2] datasets support large numbers of classes, and provide download scripts and annotations for a large number of YouTube videos, but since these videos are online, they may disappear over time. In comparison, ours is an offline dataset[3] and are nonvolatile.

E. Generation of training and test sets

We experimented with the dataset by generating 3 split sets, with a $70 : 30$ train-to-test ratio for each split set. After separating the clips into classes, the first, mid and last 30% of clips for each action class were assigned to test sets 1, 2 and 3 respectively, while the remainder were assigned to the corresponding training sets. This split generation method avoided or minimized the use of clips generated from the same original video in both train and test splits.

IV. EXPERIMENTAL RESULTS

We evaluated our dataset using the metric of classification accuracy, calculated using the scores returned by the action classifiers.

For action classification, we used a two-stream CNN architecture coupled with a recently proposed temporal pooling scheme called *kernelized rank pooling based on feature subspaces* (KRP-FS) [19] that produces nonlinear spatiotemporal feature subspace representations, as our baseline algorithm. KRP-FS was selected because it has been shown to produce state-of-the-art results for several action recognition datasets such as JHMDB, HMDB51, MPII Cooking Activities and UT-Kinect actions. We used the publicly available KRP-FS code[4] and fine-tuned it for our dataset.

Our evaluation results include results for KRP-FS as well as for *basic kernelized rank pooling* (BKRP), as a means of sanity check. Valid results should show that KRP-FS provides better accuracy than BKRP.

A. Introduction to BKRP and KRP-FS

The feature representation employed here is based on *rank pooling*. Rank pooling refers to the temporal pooling strategy where a learning-to-rank algorithm [33] is trained on different videos of the same action, and the parameters of the resultant "ranking machine" are used as a video representation that captures the video-wide temporal evolution of the video [34].

The first ever rank pooling scheme to be proposed is *VideoDarwin* [34]. Denote an n-frame video by

$$X_n = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n], \tag{1}$$

where $\boldsymbol{x}_t \in \mathbb{R}^d$ is the d-dimensional feature vector of the frame at time t. VideoDarwin

- defines a vector-valued function $V : \{X_t, t\} \to \boldsymbol{v}_t \in \mathbb{R}^d$;
- imposes the order constrains $\boldsymbol{v}_n \succ \cdots \succ \boldsymbol{v}_t \succ \cdots \succ \boldsymbol{v}_1$;
- trains a ranking machine to learn a *linear* function characterized by parameters $\boldsymbol{u} \in \mathbb{R}^D$, namely $\phi(\boldsymbol{v}; \boldsymbol{u}) = \boldsymbol{u}^\top \boldsymbol{v}$, such that $\boldsymbol{v}_t \succ \boldsymbol{v}_s \iff \boldsymbol{u}^\top \boldsymbol{v}_t > \boldsymbol{u}^\top \boldsymbol{v}_s$.

The parameter vector $\boldsymbol{u}$ is essentially a compact representation of X that summarizes the sequence dynamics. Defining $\boldsymbol{v}_t \triangleq \boldsymbol{m}_t/\|\boldsymbol{m}_t\|$, where $\boldsymbol{m}_t \triangleq (\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_t)/t$, gives us the variant of VideoDarwin called the *Forward VideoDarwin*.

Observing that VideoDarwin uses linear projection — a poor fit for nonlinear geometries — Cherian et al. [19] proposed applying the kernel trick [20] to rank pooling, enabling X to be mapped to a potentially infinite-dimensional *reproducing kernel Hilbert space* (RKHS), in which data is linear. Suppose for kernel $K(\boldsymbol{x}_t, \boldsymbol{z})$, there exists a feature map $\Phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space for which $\langle \Phi(\boldsymbol{x}_t), \Phi(\boldsymbol{z}) \rangle_{\mathcal{H}} = K(\boldsymbol{x}_t, \boldsymbol{z})$. The most basic form of kernelized rank pooling is the *basic kernelized rank pooling* (BKRP) formulation, which finds the pre-image $\boldsymbol{z}$ whose embedding $\Phi(\boldsymbol{z})$ in the feature space defines a direction/line, onto which the embeddings $\Phi(\boldsymbol{x}_1), \ldots, \Phi(\boldsymbol{x}_n)$ are projected in a way that preserves the temporal order. The BKRP scheme uses the pre-image $\boldsymbol{z}$ as the action description.

BKRP assumes that the pre-image $\boldsymbol{z}$ always exists. However, given that pre-images are finite-dimensional representatives of infinite-dimensional Hilbert space points, they may not be unique or may not even exist [35]. As a solution, instead of estimating a single hyperplane in feature space, the *kernelized rank pooling based on feature subspaces* (KRP-FS) formulation estimates a low-rank kernelized subspace subject to the constraint that projections of the embeddings $\Phi(\boldsymbol{x}_1), \ldots, \Phi(\boldsymbol{x}_n)$ onto this subspace should preserve the temporal order (as increasing distances from the subspace origin). Both BKRP and KRP-FS assume nonlinear input data and allow compact linear order-preserving projections.

B. Feature extraction using KRP-FS

A two-stream CNN architecture — the same as that of P-CNN [36] — was used, where one stream extracts RGB features from RGB images, and the other optical flow features from optical flow images. For extracting RGB features, the ImageNet-trained "VGG-f" network provided by Chatfield et al. [37] was used. For extracting (optical) flow features, the motion network that Gkioxari and Malik [38] trained on the UCF101 split 1 dataset was used. Both networks have the same architecture consisting of 5 convolutional layers and 3 fully-connected layers, as shown in Table III. The CNN features were obtained from the 7th fully connected layer of both networks. The loss functions of RGB and flow networks are *dropout* and *cross entropy* respectively. RGB and flow data were used for kernelized rank pooling computations.

---

[1]https://deepmind.com/research/open-source/open-source-datasets/kinetics/
[2]http://hacs.csail.mit.edu/
[3]https://asankagp.github.io/mod20/
[4]https://github.com/anoopcherian/kernelized_rank_pool/

TABLE II
COMPARISON OF RECENTLY PUBLISHED VIDEO DATASETS.

Dataset	Scenario	Purpose	Environment	Frames	Classes	Resolution	Year
UT Interaction [32]	Surveillance	Action recognition	Outdoor	36k	6	360×240	2010
VIRAT [27]	Drone, surveillance	Event recognition	Outdoor	Many	23	Varying	2011
HMDB51 [2]	Movies, YouTube	Action recognition	Varying	655k	51	320×240	2011
UCF101 [1]	YouTube	Action recognition	Varying	558k	24	320×240	2012
J-HMDB [3]	Movies, YouTube	Action recognition	Varying	32k	21	320×240	2013
Mini-drone [29]	Drone	Privacy protection	Outdoor	23.3k	3	1920×1080	2015
Okutama-Action [28]	Drone	Action recognition	Outdoor	70k	13	3840×2160	2017
Kinetics [8]	YouTube	Action recognition	Varying	Many	400	Varying	2017
HACS [26]	YouTube	Action recognition	Varying	Many	200	Varying	2019
MOD20 (this dataset)	YouTube, own drone	Action recognition	Varying	503k	20	720×720	2019

TABLE III
THE TWO-STREAM CNN ARCHITECTURE EMPLOYED: EACH CELL REPRESENTS A CONVOLUTIONAL LAYER ($C(n \times k \times k)$) OR FULLY CONNECTED LAYER ($FC(n)$) WITH n FILTERS AND KERNEL SIZE $k \times k$.

Network	C1	C2	C3	C4	C5	FC6	FC7	FC8
RGB [37] (input size: 224×224)	$C(64 \times 11 \times 11)$	$C(256 \times 5 \times 5)$	$C(256 \times 3 \times 3)$	$C(256 \times 3 \times 3)$	$C(256 \times 3 \times 3)$	$FC(4096)$ (dropout)	$FC(4096)$ (dropout)	$FC(1000)$ (softmax)
Flow [38] (input size: 227×227)	$C(96 \times 7 \times 7)$	$C(384 \times 5 \times 5)$	$C(512 \times 3 \times 3)$	$C(512 \times 3 \times 3)$	$C(384 \times 3 \times 3)$	$FC(4096)$ (cross entropy)	$FC(4096)$ (cross entropy)	$FC(1000)$ (softmax)
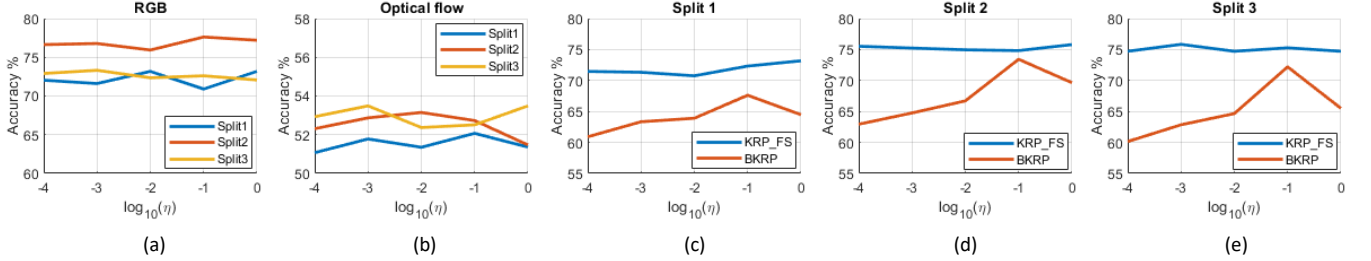
Fig. 4. (a) Average accuracies of KRP-FS using RGB features for different ranking threshold values. (b) Average accuracies of KRP-FS using flow features for different rank threshold values. (c)-(e) Average accuracies of KRP-FS and BKRP for the split sets 1, 2 and 3.

Here, we briefly discuss two important parameters for feature extraction using KRP-FS: the ranking threshold η and the Nyström (approximation) ratio.

The ranking threshold η is the parameter that enforces the temporal order of rank pooling calculation. η can be used to verify if an algorithm reliably detects temporal order in the extracted features. In Fig. 4, we analyzed the influence of η on the classification accuracy by varying it from 0.0001 to 1 at a multiplicative step size of 10. Fig. 4(a) and (b) show the accuracy variation with different thresholds for RGB and flow data respectively. Fig. 4(c)-(e) show the average accuracies of KRP-FS and BKRP for the 3 splits of our dataset. KRP-FS achieves the highest average accuracy when $\eta = 1$.

Nyström kernel approximation [39] is a technique for reducing the computational cost of kernel matrix calculation. It computes only a few columns of the kernel matrix and approximates the full kernel by a low-rank outer product. Table IV shows the classification accuracies for five Nyström ratios calculated on split 1 of our dataset. The data was uniformly sampled in the order of $(1/2^j)^{th}$ the original data size, while j is varied from 5 to 2 at a step size of -1. Consistent with [19], $(1/8)^{th}$ was found to be the optimal Nyström ratio for the HMDB51 dataset. We experimented with different Nyström values and found that $(1/8)^{th}$ was the

TABLE IV
ANALYZING NYSTRÖM APPROXIMATION RATIOS.

Nyström approx. ratio	Accuracy (%)
1/32	67.33
1/16	68.61
1/8	71.47
1/4	71.04

optimal Nyström ratio for our dataset too.

C. Performance evaluation

Over 3 split sets, the average classification accuracy is 74.05%.

We used a 3.6 GHz four-core computer for the experiments. The average execution time (total time/frames) to achieve the baseline results was 1.43 milliseconds for KRP-FS (excluding the CNN forward pass time of 27.8 milliseconds per frame), consistent with Cherian et al.'s timing [19].

Table V compares the accuracy of KRP-FS to that of BKRP for different action recognition datasets. The average classification accuracies of KRP-FS for our dataset, HMDB51, MPII Cooking and JHMDB lie in a close range (between 69% and 74%). The results can potentially be improved by using Fisher vector (IDT-FV) features [40] with KRP-FS [19].

suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 3354–3361.

[10] C. G. Keller, M. Enzweiler, and D. M. Gavrila, "A new benchmark for stereo-based pedestrian detection," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, June 2011, pp. 691–696.

[11] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4457–4465.

[12] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, 2012.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving video database with scalable annotation tooling," *CoRR*, vol. abs/1805.04687, 2018.

[15] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, "Help from the sky: Leveraging uavs for disaster management," *IEEE Pervasive Computing*, vol. 16, no. 1, pp. 24–32, Jan.-Mar. 2017. [Online]. Available: doi.ieeecomputersociety.org/10.1109/MPRV.2017.11

[16] J. M. Peschel and R. R. Murphy, "On the human–machine interaction of unmanned aerial system mission specialists," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 1, pp. 53–62, Jan 2013.

[17] R. L. Finn and D. Wright, "Unmanned aircraft systems: Surveillance, ethics and privacy in civil applications," *Computer Law and Security Review*, vol. 28, no. 2, pp. 184 – 194, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0267364912000234

[18] A. Al-Kaff, F. M. Moreno, L. J. S. José, F. García, D. Martín, A. de la Escalera, A. Nieva, and J. L. M. Garcéa, "VBII-UAV: Vision-based infrastructure inspection-UAV," in *Recent Advances in Information Systems and Technologies*, Á. Rocha, A. M. Correia, H. Adeli, L. P. Reis, and S. Costanzo, Eds. Cham: Springer International Publishing, 2017, pp. 221–231.

[19] A. Cherian, S. Sra, S. Gould, and R. Hartley, "Non-linear temporal subspace representations for activity recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[20] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[21] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633 – 659, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314213000295

[22] S. Kang and R. P. Wildes, "Review of action recognition and detection methods," *CoRR*, vol. abs/1610.06906, 2016. [Online]. Available: http://arxiv.org/abs/1610.06906

[23] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3, Aug 2004, pp. 32–36 Vol.3.

[24] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2, Oct 2005, pp. 1395–1402 Vol. 2.

[25] S. Abu-El-Haija, N. Kothari, J. Lee, A. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A large-scale video classification benchmark," *CoRR*, vol. abs/1609.08675, 2016.

[26] H. Zhao, Z. Yan, L. Torresani, and A. Torralba, "HACS: Human action clips and segments dataset for recognition and temporal localization," *arXiv preprint arXiv:1712.09374*, 2019.

[27] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR 2011*, June 2011, pp. 3153–3160.

[28] M. Barekatain, M. Martí, H. F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-Action: An aerial view video dataset for concurrent human action detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2153–2160.

[29] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi, "Privacy in mini-drone based video surveillance," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 04, May 2015, pp. 1–6.

[30] University of Central Florida, "UCF-ARG data set," http://crcv.ucf.edu/data/UCF-ARG.php, November 2011.

[31] ——, "UCF aerial action dataset," http://crcv.ucf.edu/data/UCF_Aerial_Action.php, November 2011.

[32] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *2009 IEEE 12th International Conference on Computer Vision*, Sept 2009, pp. 1593–1600.

[33] T.-Y. Liu, "Learning to rank for information retrieval," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[34] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[35] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "Kernel PCA and de-noising in feature spaces," in *Proceedings of the 1998 Conference*

on Advances in Neural Information Processing Systems II. Cambridge, MA, USA: MIT Press, 1999, pp. 536–542. [Online]. Available: http://dl.acm.org/citation.cfm?id=340534.340729

[36] G. Cheron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[37] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *CoRR*, vol. abs/1405.3531, 2014. [Online]. Available: http://arxiv.org/abs/1405.3531

[38] G. Gkioxari and J. Malik, "Finding action tubes," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[39] P. Drineas and M. W. Mahoney, "On the nyström method for approximating a gram matrix for improved kernel-based learning," *journal of machine learning research*, vol. 6, no. Dec, pp. 2153–2175, 2005.

[40] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 581–595.

[41] L. Xia, C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2012, pp. 20–27.

[42] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1194–1201.