Chapter 20
Benchmarking Datasets for Human Activity Recognition

Haowei Liu, Rogerio Feris, and Ming-Ting Sun

Abstract Recognizing human activities has become an important topic in the past few years. A variety of techniques for representing and modeling different human activities have been proposed, achieving reasonable performances in many scenarios. On the other hand, different benchmarks have also been collected and published. Different from other chapters focusing on the algorithmic aspects, this chapter gives an overview of different benchmarking datasets, summarizes the performances of the-state-of-the-art algorithms, and analyzes these datasets.

20.1 Introduction

In the past few years, the problem of automatically recognizing human activities in videos has emerged as an important field and attracted many researchers in the vision community. The problem is challenging as in general, the videos could have been shot in an unconstrained environment where the camera could be moving, the background can be cluttered, or the camera view point can be different. All these factors already make recognition of human activities difficult, let alone possible occlusions or variations of activities different subjects perform. With that said, much progress has been made toward the automatic understanding of human activities. On one hand, many approaches (e.g. feature representations and modeling) have been proposed, which have addressed the problem to some degree. On the other hand, many benchmarks and datasets consisting of activity video sequences have been collected and published. Different from other chapters, which focus on activity representation and modeling, this chapter surveys different publicly available

H. Liu (✉) · M.-T. Sun
University of Washington, Seattle, WA 98195, USA
e-mail: hwliu@uw.edu

M.-T. Sun
e-mail: mts@uw.edu

R. Feris
IBM T.J. Watson Research Center, Hawthorn, NY 10532, USA
e-mail: rsferis@ibm.com

T.B. Moeslund et al. (eds.), *Visual Analysis of Humans*,
DOI 10.1007/978-0-85729-997-0_20,

benchmarks and summarizes the-state-of-art performances reported so far. Ideally, a good benchmarking should approximate the realistic situations as much as possible by incorporating video sequences with unrestricted camera motion, different scene contexts, different degrees of background clutter and different camera perspectives. It should also consist of video sequences with multiple subjects performing different activities in order to evaluate the robustness of activity recognition algorithms to the intra-class variations of human activities. In what follows, we will also analyze each dataset by these criteria. When summarizing the performances, we only report the best number achieved. The train/test split used in these works follow either leave-one-out or leave-one-actor-out procedure. For the former, testing is done on one sequence while training on the rest. For the latter, testing is done on sequences performed by one actor while training on the rest. The performance is reported as the average across the testing results.

20.2 Single View Activity Benchmarks with Cleaner Background

20.2.1 The KTH and the Weizmann Dataset

The KTH dataset [40] and the Weizmann dataset [10] are two widely used standard datasets, which consist of videos of different human activities performed by different subjects. The KTH dataset is published by Schuldt et al. [40] in order to benchmark their proposed motion features [40]. It contains six types of human activities (walking, jogging, running, boxing, hand waving, and hand clapping), which are performed by 25 actors in four different scenarios, resulting in 600 sequences, each with a spatial resolution of 160×120 pixels and a frame rate of 25 frames per second. The other standard benchmark, the Weizmann dataset [10], contains 10 types of activities (walking, running, jumping, gallop sideways, bending, one-hand waving, two-hand waving, jumping in place, jumping jack, and skipping), each performed by nine actors, resulting in 90 video sequences, each with a spatial resolution of 180×144 pixels and a frame rate of 50 frames per second. The background is static and clean with no camera motion. Each sequence is about three seconds long.

Since these datasets are originally published to validate proposed space–time features, they are easier compared with others as the background is cleaner and static, the camera perspective is mostly frontal, and the camera motion is mostly still, although the KTH dataset contains a certain degree of camera zooming. Therefore, they have been criticized for not being a realistic sampling of actions in the real world. With that said, many researchers use them as a validation for newly proposed algorithms. Most state-of-the-art activity recognition algorithms have already achieved higher than 90% accuracy on these two datasets. Below we summarize the published results on both datasets in Tables 20.1 and 20.2. For these two datasets, people typically use leave-one-actor-out evaluation. Hence, the training/testing split is 24:1 for the KTH dataset and 8:1 for the Weizman dataset.

Table 20.1 Performances on the KTH dataset in average accuracy

Methods	Accuracy	Methods	Accuracy	Methods	Accuracy
Gilbert et al. [9]	95.7%	Cao et al. [4]	95.02%	Raptis et al. [35]	94.8%
Kovashka et al. [21]	94.5%	Brendel et al. [3]	94.2%	Liu et al [25]	94.16%
Han et al. [12]	94.1%	Liu et al. [26]	93.8%	Liu et al. [24]	93.43%
Yuan et al. [58]	93.3%	Bregonzio et al. [2]	93.17%	Wang et al. [51]	92.51%
Liu et al. [27]	92.3%	Leptev [23]	91.8%	Jhuang et al. [17]	91.7%
Fathi et al. [8]	90.5%	Yeffet et al. [57]	90.1%	Yao et al. [56]	87.8%
Ali et al. [1]	87.7%	Wang et al. [49]	87%	Messing et al. [31]	74%

Table 20.2 Performances on the Weizmann dataset in average accuracy

Methods	Accuracy	Methods	Accuracy
Wang et al. [51]	100%	Yeffet et al. [57]	100%
Fathi et al. [8]	100%	Tran et al. [44]	100%
Brendel et al. [3]	99.7%	Wang et al. [49]	97.2%
Bregonzio et al. [2]	96.66%	Satkin et al. [39]	95.76%
Lin et al. [24]	95.48%	Ali et al. [1]	95.2%
Chaudhry et al. [5]	94.4%	Jhuang et al. [17]	92.8%
Jiang et al. [18]	90%	Nieble et al. [32]	72.8%

20.2.2 The University of Rochester Activity of Daily Living Dataset

Messing et al. [31] publish an activity of daily living dataset. The dataset is created in order to approximate daily activities people might perform. The full list of activities is: answering a phone, dialing a phone, looking up a phone number in a telephone directory, writing a phone number on a white board, drinking a glass of water, eating snack chips, peeling a banana, eating a banana, chopping a banana, and eating food with silverware, all are ordinary activities people often perform. These activities are performed three times by five different people of different shapes, sizes, genders, and ethnicities, giving large appearance variations even for the same activity. The resolution is 1280×720 at 30 frames per second. Video sequences lasted between 10 and 60 seconds, ending when the activity was completed. Table 20.3 compares the performances on the University of Rochester activity of daily living dataset using different features. The evaluation follows the leave-one-actor-out procedure.

The evaluation consisted of training on all repetitions of activities by four of the five subjects, and testing on all repetitions of the fifth subjects activities. This leave-one-out testing was averaged over the performance with each left-out subject.

Table 20.3 Performances on the UR ADL dataset in average accuracy

	Methods	Recognition Accuracy
Messing et al. [31]	Velocity History	89%
Raptis et al. [35]	Tracklet	82.67%
Satkin et al. [39]	placeCityHOF + cropping	80%
Matikainen et al. [30]	Sequential Code Map + pairwise relation	70%

20.2.3 Other Datasets

Other than the aforementioned datasets, Tran et al. [44] compose a UIUC activity dataset, consisting of 532 high resolution (1024 × 768) sequences of 14 activities performed by 8 different actors with extensive repetition. Each sequence lasts for 10∼15 seconds. They achieve an accuracy of 99.06% using the proposed metric learning method.

Another closely related source of datasets is the PETS (Performance Evaluation of Tracking and Surveillance) workshop [15], which releases high resolution surveillance footages every year. Portions of the released datasets are used as benchmarks for human activity recognition algorithms. For example, Ribeiro et al. [36] reported a 94% accuracy on the PETS04-CAVIAR dataset [13], which includes single person activities such as people fighting, walking or being immobile.

20.3 Single View Activity Benchmarks with Cluttered Background

20.3.1 The CMU Soccer Dataset and Crowded Videos Dataset

Different from the datasets introduced in previous section where the video sequences contain few or no background clutter, both the CMU soccer and CMU crowded video datasets are made to introduce cluttered background. In [7], Efros et al. record several minutes of a World Cup football game. The dataset consists of walking and running activities at different directions, giving a total of seven activities and around 5000 frames. Although the video sequences are recorded from TV programs, providing a resolution of 640 × 480, the dataset is challenging in that each human figure is only 30 pixels tall on average, and hence, fine-scale human pose estimation is not possible, making motion the only possible cue. Also, other moving humans from the background could also occlude the target subject. Table 20.4 summarizes the reported performance using leave-one-out procedure on this dataset. It suggests that putting a hierarchy or a generative model on the raw motion features could improve the performance by a 10% margin.

Ke et al. [19] collect video sequences of activities in crowded scenes to evaluate their proposed volumetric features, which are space–time templates for particular

Table 20.4 Performances on the soccer dataset in accuracy

	Methods	Accuracy
Wang et al. [50]	Motion descriptors + topic modeling	78.6%
Fathi et al. [8]	Mid-level motion descriptors	71%
Efros et al. [7]	Motion descriptors + nearest neighbor	67%

Table 20.5 Performances on the CMU crowded videos dataset in Area under ROC Curve (AUROC)

Actions/Methods	Ke et al. [19]	Brendel et al. [3]	Yao et al. [56]
Pick-up	0.47	0.60	0.58
One-hand wave	0.38	0.64	0.59
Push button	0.48	N.A.	0.74
Jumping jack	0.22	0.45	0.43
Two-hand wave	0.64	0.65	0.53

activities. These videos are recorded using a hand-held camera. Each activity is performed by three to six subjects, resulting in 110 activities of interest. The videos are downscaled to 160×120 in resolution. There is high variability in both how the subjects performed the activities and the background clutter. There are also significant spatial and temporal scale differences in the activities as well. Table 20.5 compares the performances of the state-of-the-art approaches. The performance gain of the latter two approaches comes from the incorporation of temporal features, for example, the time-series representation in [3]. Since these approaches are template-based, to test how well the templates generalize, the evaluation consists of training on sequences performed by one actor while testing on the rest.

20.3.2 The University of Maryland Gesture Dataset

Lin et al. [24] publish an UM gesture dataset consisting of 14 different gesture classes, which are a subset of the military signals. The gestures include "turn left", "turn right", "attention left", "attention right", "flap", "stop left", "stop right", "stop both", "attention both", "start", "go back", "close distance", "speed up" and "come near". The dataset is collected using a color camera with 640×480 resolution. Each activity is performed by three people for three times, giving 126 video sequences for training which are captured using a fixed camera with the person viewed against a simple, static background.

There are 168 video sequences for testing which are captured from a moving camera and in the presence of background clutter and other moving objects. Lin et al. [24] use the proposed prototype tree to achieve an accuracy of 91.07%. Brendel et al. [3] achieve 96.3% using time-series modeling while Tran et al. [44] achieve

an 100% accuracy. Note that since this dataset focuses on military signals, it might not be a suitable benchmark for generic activity recognition.

20.4 Multi-view Benchmarks

The aforementioned benchmarks only provide video sequences from a single camera perspective. In real life, it might be desirable to have a multi-camera configuration, for example, in surveillance applications. In what follows, we introduce two datasets consisting of activities from different perspectives.

20.4.1 The University of Central Florida Sports Dataset

Rodriguez et al. [37] publish a dataset consisting of a set of actions collected from various sports which are typically featured on broadcast television channels such as BBC and ESPN. It contains over 200 video sequences at a resolution of 720×480 and consists of nine sport activities including diving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging a baseball bat, and pole vaulting. These activities are featured in a wide range of scenes and viewpoints. Table 20.6 summarizes the published results using leave-one-out procedure on this dataset. Note that the space–time MACH filter [37] is a template matching approach. The low accuracy of its performance suggests that the model-based approach captures intra-class variability better when the camera view point varies.

Following the sports dataset, Yeffet et al. [57] publish a dataset of UFC videos from TV programs. UFC is a fighting sport similar to boxing. Therefore, the viewpoints and individual appearance vary differently and camera motion persists. In addition, two fighters act at the same time and could occlude each other. The dataset contains over 20 minutes of broadcast video, and two target activities are defined: the throw/take-down action and keen-kick action, two rarely occurred activities in UFC sport. Therefore, the dataset is versatile compared to other sports. One merit of this dataset is that the target activities are relevant to surveillance applications as these activities rarely occur and are similar to one person hitting another.

Table 20.6 Performances on the UCF sports dataset in average accuracy

	Methods	Recognition Accuracy
Kovashka et al. [21]	Hierarchical neighborhood feature	87.27%
Yeffet et al. [57]	Local trinity pattern feature	79.2%
Rodriguez et al. [37]	Space–time MACH filter	69%

Table 20.7 Performances on the multi-view dataset in average accuracy

Methods	Accuracy
Weinland et al. [52]	91.11%*
Lv et al. [28]	80.06%
Tran et al. [44]	80%

*Result using information from all possible views

20.4.2 The INRIA Multi-view Dataset

To the best of our knowledge, the multi-view dataset published by Weinland et al. [53] is the only known large scale multi-view dataset that provides synchronized video sequences from multiple cameras for each activity. They use multiple cameras to record 13 activities such as "walk", "sit down", "check watch", etc. Each activity is performed by multiple actors. The camera array provides five synchronized views at a resolution of 390×291 with a frame rate 23 frames per second. Each sequence lasts for a few seconds. Weinland et al. [52] demonstrate that by fusing views from multiple cameras, the accuracy can be greatly improved. Table 20.7 summarizes the performances reported so far. Note that Weinland et al. [52] use the information from all views while others, [28] and [44], use only one of the views. The evaluation follows the leave-one-actor-out procedure.

20.5 Benchmarks with Real World Footages

The datasets discussed thus far, except the UCF Sports Dataset, consist of video sequences where human actors perform different activities. Therefore, these datasets are made in a more controlled environment. In this section, we discuss datasets consisting of video sequences extracted from different real world sources, such as movies or the Internet. Since there is no limitation on how these video sequences should be made, these datasets are more difficult as the videos could contain occlusions, background clutters or could have been shot with different camera perspectives or motion.

20.5.1 The University of Central Florida Youtube Dataset

Liu et al. [26] collected video sequences from YouTube and made a dataset consisting of 11 activities, resulting in a total of 1168 sequences. These activities include basketball shooting (b_shooting), volleyball spiking (v_spiking), trampoline jumping (t_jumping), soccer juggling (s_juggling), horseback riding (h_riding), cycling, diving, swinging, golf swinging (g_swinging), tennis swinging (t_swinging), and walking (with a dog). Due to the diverse nature of video sources, these sequences contain significant camera motion, background clutters, and occlusions, variations

Table 20.8 Performances reported on the YouTube dataset in recognition accuracy

Methods/Actions	b_shooting	cycling	diving	g_winging	h_riding	s_juggling
Brendel et al. [3]	60.1%	79.3%	85.8%	89.8%	80.6%	59.3%
Liu et al. [27]	N.A.	N.A.	82%	86%	78%	60%
Ikizler-cinbis et al. [16]	48.48%	75.17%	95.0%	95.0%	73.0%	53.0%
Liu et al. [26]	53%	73%	81%	86%	72%	54%
Matikainen et al. [30]	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.

Methods/Actions	swinging	t_swinging	t_jumping	v_spiking	walking	Mean
	61.7%	87.8%	88.3%	80.5%	82.7%	77.8%
	67%	76%	80%	80.2%	N.A.	76.1%
	66.0%	77.0%	93.0%	85.0%	66.67%	75.21%
	57%	80%	79%	73.3%	75%	71.2%
	N.A.	N.A.	N.A.	N.A.	N.A.	59%

in subject appearance, illumination and view point. Also, all the sequences are low-resolution videos (240 × 320) with a frame rate of 15 frames per second. Each activity is about 3∼5 seconds long. Table 20.8 summarizes the published results using the leave-one-out procedure on the YouTube dataset.

20.5.2 The Hollywood Dataset

In order to provide a realistic benchmarking in an unconstrained environment, Laptev et al. [22] initiates an effort by creating a dataset consisting of video sequences extracted from two episodes from the movie "Coffee and Cigarettes", providing a pool of examples for atomic actions, such as "drinking" and "smoking", where each atomic event ranges from 30 to 200 frames long, with a mean of 70 frames. They show on a ∼36000 frame test set that by combining both frame-based classifier and space–time based classifier improves the precision of action detection by a 30%∼40% margin given the same recall. Similarly, Rodriguez et al. [37] published a kissing/slapping dataset consisting of ∼200 sequences from several movies. They achieved ∼66% accuracy using a template-based approach.

Laptev et al. [23] later create a Hollywood-1 dataset by extracting eight different actions (answer phone, hug person, sit up, sit down, kiss, handshake, and stand up) from various movies. The dataset consists of ∼400 video sequences. Each sequence is about 50∼200 frames long with a resolution 240 × 500 and a frame rate of 24 frames per second. Using a combination of multi-scale flow and shape features, they achieve a 30%∼50% average precision for each action class. Marszałek et al. [29] subsequently create a Hollywood-2 dataset by augmenting Hollywood-1 to include up to twelve activities with a total of 600 K frames. The scene information is also

Table 20.9 Performances on Hollywood-1 datasets in average precision

Methods/Actions	AnswerPhone	GetOutCar	HandShake	HugPerson	Kiss	SitUp
Gilbert et al. [9]	47%	47%	45.6%	42.8%	72.5%	44.0%
Han et al. [12]	43.4%	46.8%	44.1%	46.9%	57.3%	38.4%
Sun et al. [43]	40%	42%	38%	42%	55%	40%
Leptev et al. [23]	32.1%	41.5%	32.3%	40.6%	53.3%	18.2%
Raptis et al. [35]	33.0%	27%	20.1%	34.5%	53.7%	19%
Yeffet et al. [57]	35.1%	32%	33.8%	28.3%	57.6%	13.1%

Methods/Actions	SitDown	StandUp	Mean
	84.6%	70.5%	56.8%
	46.2%	57.1%	47.5%
	50%	55%	47.1%
	38.6%	50.5%	38.4%
	27.4%	60%	34.3%
	36.2%	58.3%	N.A.

Table 20.10 Performances on Hollywood-2 datasets in average precision

Methods/Actions	AnswerPhone	DriveCar	Eat	FightPerson	GetOutCar	HandShake
Gilbert et al. [9]	40.2%	75%	51.5%	77.1%	45.6%	28.9%
Satkin et al. [39]	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Han et al. [12]	15.57%	87.01%	50.93%	73.08%	27.19%	17.17%
Marszałek et al. [29]	10.7%	75%	28.6%	67.5%	19.1%	14.1%

HugPerson	Kiss	Run	SitDown	SitUp	StandUp	Mean
49.4%	56.6%	47.5%	62.0%	26.8%	50.7%	50.9%
N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	43.48%
27.22%	42.91%	66.94%	41.61%	7.19%	48.6%	42.12%
13.8%	55.6%	56.6%	31.6%	14.2%	35.0%	35.1%

annotated. They achieve an average precision of 35.5% by incorporating the context, i.e. the scene information. Both the Hollywood-1 and Hollywood-2 datasets come with a clean training set and a test set of roughly equal size (about 200 sequences).

Overall, the Hollywood datasets pose a great challenge to activity recognition as the camera views are different from sequence to sequence, the background is cluttered, multiple subjects are present, occlusions occur very often, and the intra-class variability is large, making recognition hard. Tables 20.9 and 20.10 summarize reported performance on Hollywood-1 and Hollywood-2 datasets. As we see from the tables, there is still huge room for improvement. Gilbert et al. [9] is the current state-

of-the-art by mining the spatial-temporal relationships between space–time interest points.

20.5.3 The Olympic Dataset

Recently, Niebles et al. [33] publish the Olympic Sports Dataset. The dataset contains 50 videos from each of the following 16 activities: high jump, long jump, triple jump, pole vault, discus throw, hammer throw, javelin throw, shot put, basketball layup, bowling, tennis serve, platform (diving), springboard (diving), snatch (weightlifting), clean and jerk (weightlifting) and vault (gymnastics). These sequences, obtained from YouTube, contain severe occlusions, camera movements, and compression artifacts. In contrast to other sport datasets such as the UCF Sports Dataset [37], which contains periodic or simple activities such as walking, running, golf swinging, ball kicking, the activities in the Olympic Dataset are longer and more complex. Niebles et al. [33] achieved an accuracy of 72% by modeling the temporal structure of these activities.

20.6 Benchmarks with Multiple Activities

The benchmarks introduced so far focus more on "activity recognition", i.e. video sequences in these datasets are typically pre-segmented and contain only one activity. It is desirable to have benchmarks with video sequences containing multiple activities for activity detection algorithms, i.e. finding out all possible activities in the video sequences, which is especially beneficial for surveillance applications. Uemura et al. [46] publish a Multi-KTH dataset, consisting of the same activities as the KTH dataset. The video sequences have a resolution of 640×480 and contain activities similar to the KTH dataset, except that one video sequence could contain multiple activities simultaneously and that the camera is constantly moving. By tracking space–time interest points, Uemura et al. [46] achieve an average precision of 65.4%, while Gilbert et al. [9] achieve 75.2% by data mining the space–time features. Table 20.11 summarizes the performances for each activity in terms of average precision.

In a similar setting to [19], Yuan et al. [58] publish an MSR-1 dataset containing 16 video sequences and having in total 63 actions: 14 hand clapping, 24 hand waving, and 25 boxing, performed by 10 subjects. Each sequence contains multiple

Table 20.11 Performances reported on the multi-KTH dataset in average precision

Methods/Actions	Clapping	Waving	Boxing	Jogging	Walking	Average
Gilbert et al. [9]	69%	77%	75%	85%	70%	75.2%
Uemura et al. [46]	76%	81%	58%	51%	61%	65.4%

types of actions. Some sequences contain actions performed by different people. There are both indoor and outdoor scenes. All of the video sequences are captured with cluttered and moving backgrounds. Each video is of low resolution, 320×240 and frame rate 15 frames per second. Their lengths are between $32 \sim 76$ seconds. An extended MSR-2 dataset consisting of 54 videos sequences is also available [4]. Yuan et al. [58] report a 57% recall and 87.5% precision.

Other than the aforementioned datasets, TRECVID [42] is an annual event detection challenge aiming at addressing realistic activity retrieval problems. The dataset is updated each year. It consists of videos from multiple surveillance cameras deployed at the London Gatwick airport. For example, for the 2009 challenge, the goal of the challenge was to detect several target events, including "ElevatorNoEntry", "OpposingFlow" (moving in the opposite direction), "PersonRuns", "Pointing", "CellToEar", "ObjectPut", "TakePicture", "Embrace", "PeopleMeet", and, "PeopleSplitUp". The dataset is challenging in that unlike the sequences in previous datasets where activities are repetitive, most of the target events in TRECVID are rare and subtle. For example, to detect the activity "CellToEar" or "PersonRuns" in unconstrained video sequences is extremely difficult. Also, the sequences always have cluttered background, which could also include moving people, resulting in complicated occlusion scenarios. The intra-class variations of each activity are also huge, since each person performs the same activity differently. The evaluation is done using the Detection Error Tradeoff (DET) curve, a trade off curve between miss rate and false alarm rate. The state of the art approach achieves only 90% miss rate while keeping the false alarm rate to 20 per hour. The miss rate drops to $\sim$80% while the false alarm rate is kept at 100 per hour, an indication of the difficulty of the dataset.

20.7 Other Benchmarks

Other than recognizing single subject kinematic activities, recently, researchers have tried to extend activity recognition to a broader context. For example, Prabhakar et al. [34] use temporal causality to detect activities that involve interactions among people. They evaluate their approach on a toy dataset consisting of sequences of ball playing activities ("roll-ball", "throw-ball", and "kick-ball") and a child play dataset [48] consisting of social games such as pattycake between an adult and a child, achieving 60%$\sim$70% accuracy. They also report results on the "HandShake" from the Hollywood dataset [29] for realistic evaluations. Another dataset that also involves human interactions is the PETS07-BEHAVE [14] dataset consisting of video sequences of 640×480 resolution. The activities include walking together, splitting, approaching, fighting, chasing, and so on.

Another category of activities that attracts many research works involves object manipulation. The recognition of object manipulation based activities finds its application, for example, in Programming by Demonstration in Robotics or flow optimization for factory workers. Experimental protocols for laboratory technicians and recipes for home cooks are also example tasks. Also, in object recognition, more and more context information are brought in to help recognizing the objects and

the way an object is manipulated or held significantly constrained the category of the object. On the other hand, the object class also affects how it can be grasped or manipulated and the activities that can be performed on it.

Gupta et al. [11] collect a sports image dataset consisting of five activities: "Cricket bowling", "Croquet shot", "Tennis forehand", "Tennis serve", "Volleyball smash", each with 50 images. They report a 78.9% accuracy while recently, Yao et al. [55] achieve a recognition rate of 83.3% by jointly modeling activity, body pose and manipulated object.

Similarly, Yao et al. [54] publish an instrument playing dataset consisting of seven different musical instruments: bassoon, erhu, flute, French horn, guitar, saxophone, and violin. Each class includes ∼150 people-playing-musical-instrument images. They achieve an accuracy of 65.7% using their proposed Grouplet features, an extension of local interest point features to take into account neighboring relationships.

Kjellstrom et al. [20] collect the OAC (Object–Action-Complex) dataset. The dataset consists of 50 instances, each of three different action–object combinations: "look through binoculars", "drink from cup", and "pour from pitcher". The activities are performed by 10 subjects, 5 times each. The classes are selected so that two of the activities, "look through" and "drink from" are similar, while two of the objects, "cup"and "pitch" are similar as well. They report the best performance of 6% error rate by jointly inferring the activities and the manipulated object using a CRF.

Another closely related work is the HumanEva datasets [41]. These datasets contain video sequences of six simple activities performed by four∼six subjects with motion sensors. Other than videos, the datasets also provide corresponding motion sensor values from the motion capture system in order to evaluate human pose estimation and articulated tracking algorithms.

Tables 20.12 and 20.13 summarize different properties, such as resolution, activities, degree of background clutter, of the major benchmarking datasets. We can see from the table, the numbers reported on the standard activity recognition datasets such as the KTH dataset [40] are saturated, mostly above 90%. On the other hand, there is still a huge room for improvement for realistic and multi-activity datasets, such as the Hollywood datasets [23, 29], the MSR dataset [58], or the TRECVID [42]. This suggests that more sophisticated methods are needed to address the problems of cluttered background or those of representing activities in finer scales.

20.8 Conclusions

In this chapter, we have covered the state-of-the-art benchmarking datasets for human activity recognition algorithms, ranging from standard KTH dataset [40] to realistic Hollywood dataset [23, 29] or TRECVID dataset [42]. To conclude, datasets such as the KTH dataset [40] or the Weizmann dataset [10] for which the state-of-the-art approaches have already achieved above 90% accuracy provide bench-

Table 20.12 Summary of all the datasets. "r" indicates that the dataset was made out of realistic videos. "v" indicates the dataset consists of video sequences with various perspectives. The performance is reported in average accuracy unless otherwise specified. The columns are dataset names, number of activities, number of actors, resolution of the videos (res.), and camera views

Dataset	activities	actors	res.	views
KTH [40]	6	25	160 × 120	frontal/side
Weizmann [10]	10	9	180 × 144	frontal/side
CMU Soccer [7]	7	r	30 × 30	side
CMU Crowded [19]	5	6	320 × 240	side/frontal
UCF Sports [37]	9	r	720 × 480	v
UR ADL [31]	10	5	1280 × 720	frontal
UM Gesture [24]	14	3	640 × 480	frontal
UCF Youtube [26]	11	r	240 × 320	v
Hollywood-1 [23]	8	r	240 × 500	v
Hollywood-2 [29]	12	r	240 × 500	v
MultiKTH [46]	6	5	640 × 480	side/frontal
MSR [58]	3	10	320 × 240	side/frontal
TRECVID [42]	10	r	640 × 480	v

marks in a more controlled environment, while the YouTube dataset [26], the Hollywood datasets [23, 29], and the TRECVID dataset [42] approximate realistic situations better, posing great challenges to human activity recognition algorithms. The datasets with videos containing multiple activities, such as the MSR dataset [58] provide suitable benchmarks for activity detection techniques, which are still few in its genre as most human activity recognition techniques assume pre-segmented video sequences. The properties of these major benchmarking datasets are summarized in both Tables 20.12 and 20.13. We hope that by summarizing the state-of-the-art numbers, people would be able to use them as a baseline and report improved numbers on top of them.

A dataset that is presently lacking is one that contains human actions with the information on the action context as well as on the objects that are involved in the actions. This need was also outlined in Chap. 18 where the reader may find a more detailed discussion.

20.8.1 Further Readings

We refer the interested readers to Turaga et al. [45] for generic topics about human activity recognition. For empirical methods and evaluation methodologies in Computer Vision, Henrik et al. [6] and Venkata et al. [47] both cover the design of experiments and benchmarks for various topics in Computer Vision. Interested readers could also see [38] and [59] for information about providing ground-truth labeling.

Table 20.13 Summary of all the datasets. "r" indicates that the dataset was made out of realistic videos. "v" indicates the dataset consists of video sequences with various perspectives. The performance is reported in average accuracy unless otherwise specified. The columns are dataset names, degree of background clutter (bg clutter), camera motion (c_motion), and the state-of-the-art performances

Dataset	bg clutter	c_motion	performances
KTH [40]	no	slightly	95.7% [9]
Weizmann [10]	no	no	100% [57]
CMU Soccer [7]	moderate	yes	78.6% [50]
CMU Crowded [19]	yes	yes	0.6 (AUROC) [56]
UCF Sports [37]	yes	no	87.27% [21]
UR ADL [31]	no	no	89% [31]
UM Gesture [24]	yes	yes	100% [44]
UCF Youtube [26]	yes	yes	77.8% (average precision) [3]
Hollywood-1 [23]	yes	yes	56.8% (average precision) [9]
Hollywood-2 [29]	yes	yes	50.9% (average precision) [9]
MultiKTH [46]	yes	yes	75.2% (average precision) [9]
MSR [58]	yes	yes	recall/precision: 57%/87.5% [58]
TRECVID [42]	yes	no	90% miss rate 20 false positives/hr

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **32**(2), 288–303 (2010) [413]
2. Bregonzio, M., Gong, S., Xiang, T.: Recognising action as clouds of space–time interest points. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2009) [413]
3. Brendel, W., Todorovic, S.: Activities as time series of human postures. In: IEEE European Conference on Computer Vision (ECCV) (2010) [413,415,418,424]
4. Cao, L., Liu, Z., Huang, T.: Cross-dataset action detection. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2010) [413,421]
5. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2009) [413]
6. Christensen, H., Phillips, J.: Empirical Evaluation Methods in Computer Vision. World Scientific, Singapore (2002) [423]
7. Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision (ICCV) (2003) [414,415,423,424]
8. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2008) [413, 415]
9. Gilbert, A., Illingworth, J., Bowden, R.: Action recognition using mined hierarchical compound features. IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI) (2010) [413,419,420,424]
10. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space–time shapes. In: IEEE International Conference on Computer Vision (ICCV) (2005) [412,423,424]

11. Gupta, A., Kembhavi, A., Davis, L.: Observing human–object interactions: Using spatial and functional compatibility for recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(10), 1775–1789 (2009) [422]
12. Han, D., Bo, L., Sminchisescu, C.: Selection and context for action recognition. In: IEEE International Conference on Computer Vision (ICCV) (2009) [413,419]
13. IEEE: Performance Evaluation of Tracking and Surveillance (2004) [414]
14. IEEE: Performance Evaluation of Tracking and Surveillance (2007) [421]
15. IEEE: Performance Evaluation of Tracking and Surveillance (2009) [414]
16. Ikizler-Cinbis, N., Sclaroff, S.: Object, scene and actions: Combining multiple features for human action recognition. In: IEEE European Conference on Computer Vision (ECCV) (2010) [418]
17. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: IEEE International Conference on Computer Vision (ICCV) (2007) [413]
18. Jiang, H., Martin, D.: Finding actions using shape flows. In: IEEE European Conference on Computer Vision (ECCV) (2008) [413]
19. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in cluttered videos. In: IEEE International Conference on Computer Vision (ICCV) (2007) [414,415,420,423,424]
20. Kjellström, H., Romero, J., Martínez, D., Kragić, D.: Simultaneous visual recognition of manipulation actions and manipulated objects. In: IEEE European Conference on Computer Vision (ECCV) (2008) [422]
21. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space–time neighborhood features for human action recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2010) [413,416,424]
22. Laptev, I., Perez, P.: Retrieving actions in movies. In: IEEE International Conference on Computer Vision (ICCV), pp. 1–8 (2007) [418]
23. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2008) [413,418,419,422-424]
24. Lin, Z., Jiang, Z., Davis, L.: Recognizing actions by shape-motion prototype trees. In: IEEE International Conference on Computer Vision (ICCV), pp. 444–451 (2009) [413,415,423]
25. Liu, J., Shah, M.: Learning human action via information maximization. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2008) [413]
26. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2009) [413, 418,423,424]
27. Liu, J., Yang, Y., Shah, M.: Learning semantic visual vocabularies using diffusion distance. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2009) [413,418]
28. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and Viterbi path searching. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2007) [417]
29. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2009) [418,421-423]
30. Matikainen, P., Hebert, M., Sukthankar, R.: Representing pairwise spatial and temporal relations for action recognition. In: IEEE European Conference on Computer Vision (ECCV) (2010) [414,418]
31. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. In: IEEE International Conference on Computer Vision (ICCV) (2009) [413,414, 423,424]
32. Niebles, J., Li, F.-F.: A hierarchical model of shape and appearance for human action classification. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2007) [413]
33. Niebles, J., Chen, C.-W., Li, F.-F.: Modeling temporal structure of decomposable motion segments for activity classification. In: IEEE European Conference on Computer Vision (ECCV) (2010) [420]

34. Prabhakar, K., Oh, S., Wang, P., Abowd, G., Rehg, J.: Temporal causality for the analysis of visual events. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2010) [421]
35. Raptis, M., Soatto, S.: Tracklet descriptors for action modeling and video analysis. In: IEEE European Conference on Computer Vision (ECCV) (2010) [413,414,419]
36. Ribeiro, P., Santos-Victor, J.: Human activity recognition from video: modeling, feature selection and classification architecture. In: International Workshop on Human Activity Recognition and Modelling (2005) [414]
37. Rodriguez, M., Ahmed, J., Shah, M.: Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2008) [416,418,420,423,424]
38. Russell, B., Torralba, A., Murphy, K.: Labelme: A database and web-based tool for image annotation. Int. J. Comput. Vis. **77**(1–3), 157–173 (2008) [423]
39. Satkin, S., Hebert, M.: Modeling the temporal extent of actions. In: IEEE European Conference on Computer Vision (ECCV) (2010) [413,414]
40. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: International Conference on Pattern Recognition (ICPR) (2004) [412,422-424]
41. Sigal, L., Balan, A., Black, M.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal on Computer Vision (IJCV) **87**(1–2) (2010) [422]
42. Smeaton, A., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: ACM International Conference on Multimedia Information Retrieval (MIR) (2006) [421-424]
43. Sun, J., Wu, X., Yan, S., Cheong, L.-F., Chua, T.-S., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2009) [419]
44. Tran, D., Sorokin, A.: Human activity recognition with metric learning. In: IEEE European Conference on Computer Vision (ECCV) (2008) [413,414,417,424]
45. Turaga, P., Chellappa, R.: Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Technol. **18**(11), 1473–1488 (2008) [423]
46. Uemura, H., Ishikawa, S., Mikolajczyk, K.: Feature tracking and motion compensation for action recognition. In: British Machine Vision Conference (BMVC) (2008) [420,423]
47. Venkata, S., Ahn, I., Jeon, D., Gupta, A., Louie, C., Garcia, S., Belongie, S., Taylor, M.: Sd-vbs: The San Diego Vision Benchmark Suite (2009) [423]
48. Wang, P., Abowd, G., Rehg, J.: Quasi-periodic event analysis for social game retrieval. In: IEEE International Conference on Computer Vision (ICCV) (2009) [421]
49. Wang, Y., Mori, G.: Learning a discriminative hidden part model for human action recognitio. In: Advances in Neural Information Processing Systems (NIPS) (2008) [413]
50. Wang, Y., Mori, G.: Human action recognition by semilatent topic models. IEEE Trans. Pattern Anal. Mach. Intell. **31**(10), 1762–1774 (2009) [415,424]
51. Wang, Y., Mori, G.: Max-margin hidden conditional random fields for human action recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2009) [413]
52. Weinland, D., Boyer, E., Ronfard, R.: Action recognition from arbitrary views using 3d exemplars. In: IEEE International Conference on Computer Vision (ICCV) (2007) [417]
53. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding (2006) [417]
54. Yao, B., Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interactions. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2010) [422]
55. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human–object interaction activities. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2010) [422]
56. Yao, B., Zhu, S.-C.: Learning deformable action templates from cluttered videos. In: IEEE International Conference on Computer Vision (ICCV) (2009) [413,415,424]

57. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: IEEE International Conference on Computer Vision (ICCV) (2009) [413,416,419,424]
58. Yuan, J., Liu, Z., Wu, Y.: Discriminative subvolume search for efficient action detection. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2009) [413,420-424]
59. Yuen, J., Russell, B., Liu, C., Torralba, A.: Labelme video: Building a video database with human annotations. In: IEEE International Conference on Computer Vision (ICCV) (2009) [423]