A Critical Review of Action Recognition Benchmarks

Tal Hassner
The Open University of Israel
hassner@openu.ac.il

Abstract

Understanding human actions in videos has been a central research theme in Computer Vision for decades, and much progress has been achieved over the years. Much of this progress was demonstrated on standard benchmarks used to evaluate novel techniques. These benchmarks and their evolution, provide a unique perspective on the growing capabilities of computerized action recognition systems. They demonstrate just how far machine vision systems have come while also underscore the gap that still remains between existing state-of-the-art performance and the needs of real-world applications. In this paper we provide a comprehensive survey of these benchmarks: from early examples, such as the Weizmann set [1], to recently presented, contemporary benchmarks. This paper further provides a summary of the results obtained in the last couple of years on the recent ASLAN benchmark [12], which was designed to reflect the many challenges modern Action Recognition systems are expected to overcome.

1. Introduction

Capturing digital videos has long ceased to be restricted to experts with high-end equipment. Similarly, storage and transfer of high-quality videos now no longer requires expensive hardware. Not surprising, digital videos are now abundant. With this abundance comes a growing need for effective video understanding techniques, and in particular, human action recognition. Over the last decade, this has led to a growing interest in action recognition research, yielding a wide range of techniques and systems being proposed.

Similarly to other Computer Vision problems, interest in action recognition has led many to assemble and put forth benchmarks for action recognition. These benchmarks have evolved along-side the growing capabilities of machine vision methods. As methods improved in performance, so did the benchmarks become more challenging: From the early benchmarks, which included few videos of atomic actions acquired under controlled conditions, to the more recent benchmarks offering thousands of videos obtained outside the lab, "in the wild", and representing a wide range of complex actions and behaviors.

In this paper we survey these benchmarks. On one hand we intend to review the significant progress made by action recognition systems over a relatively short period of time. On the other hand, we highlight the wide gap that remains between existing capabilities and the needs of real-world applications. This gap is particularly meaningful when comparing action recognition in videos to other Computer Vision tasks, particularly in image classification: The most recent and challenging benchmarks for action recognition fall far short, in size and class variety, from those assembled for image understanding (e.g., the "80 million tiny images" [35] and "ImageNet" [5] collections).

We conclude by reviewing the existing performance on the Action Similarity LAbeliNg benchmark (ASLAN) [12], one of the recent benchmarks proposed for this problem and the focus of the ACTion Similarity (ACTS) in unconstrained videos workshop, at the Computer Vision and Pattern Recognition (CVPR) conference, 2013, further illustrating the distance that still remains to be crossed remaining by action recognition systems.

1.1. Related work

This paper is by no means the first to review the various benchmarks proposed for action recognition. Two very recent examples include [3, 12] and [19]. It is therefore reasonable to ask what is the need for another survey? Rather than providing a taxonomy of the benchmarks according to their design and intended purposes, our goal here is to learn what Computer Vision systems can and cannot do, by considering the benchmarks used to test them; the evolution of these benchmarks testifies to the growing capabilities of the systems developed over the years. Consequently, reports of perfect scores on some of the older benchmarks, suggest that the settings they present, no longer offer significant challenges to modern systems. The design of the newer benchmarks, however, provides an important picture of the challenges that still remain.

2. A survey of action recognition benchmarks

We survey many of the benchmarks commonly used in action recognition research. These are summarized in Ta-

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.43

ble 1. We note that our survey focuses exclusively on action recognition from visual data, excluding benchmarks which use additional channels of information (e.g., depth, as in [24]). We further limit the discussion to benchmarks designed to measure performance on general action recognition tasks, rather than specific applications (e.g., violence detection in [8] or gait recognition [31]).

Whenever possible, we report existing state-of-the-art results. We note that different benchmarks define different testing protocols and different performance measures. Due to space considerations, we report only the best scores for each set without elaborating on the different measures. Care must therefore be taken when interpreting these values, as different benchmarks use different evaluation protocols and have different levels of chance recognition.

2.1. The early years: Action recognition "in the lab"

Two early benchmarks are the **KTH** [32] and **Weizmann** [1] sets, both used extensively over the years. These sets provide low resolution videos of a few, "atomic" action categories, such as walking, jogging, and running. These videos were produced "in the lab", and present actors that perform scripted behavior. The videos they provide were acquired under controlled conditions, with static cameras and static, un-clutered backgrounds. In addition, actors appear without occlusions, thus allowing action recognition to be performed by considering silhouettes alone [1].

Performances on these databases has saturated over the years, with perfect accuracy reported on the Weizmann set [40] and near-perfect accuracy on KTH (e.g., $\sim 95\%$ in [7]), which has been reported saturated as far back as [23]. Curiously, despite being saturated and despite their simplistic settings, these benchmarks are still frequently used today, with recent examples including [13] on the Weizmann set and [34] on KTH.

Over the years other benchmarks have been proposed, providing videos obtained under laboratory conditions, with different emphasis placed on the data set design and the setting these benchmarks attempted to reflect. These sets include the **IXMAS** benchmark, proposed in [41], designed to study action recognition under varying viewpoints. Although IXMAS provides synchronized, multi-view footage of each action, and this has been used to include depth information for recognition, tests on this benchmark often include videos only. Additional sets are The **UMD Common Activities Dataset** [37], which records activities from a synchronized pair of viewpoints, the University of Illinois at Urbana-Champaign **UIUC1** benchmark [36], and finally, the **University of Rochester Activities of Daily Living (ADL)** [22] benchmark, released in 2009 and the latest such benchmark proposed, to our knowledge.

As we report in Table 1, perfect, or near perfect results have been reported on all these benchmarks. We believe this is a strong testament on the capabilities of modern action recognition systems: controlled setting such as those offered by these early sets no longer seem to pose significant challenges to modern computer vision systems.

2.2. Interim years: TV, sports, and motion pictures

In an effort to increase the diversity of appearances and viewing conditions, benchmark designers have turned to TV, sports broadcasts and motion pictures as sources for challenging videos of human actions. These benchmarks no longer represent controlled conditions; viewpoints, illuminations, occlusions are all arbitrary, thereby significantly raising the bar for action recognition systems.

One early example is the **UIUC2** benchmark [36], which provided unconstrained sports videos of badminton matches downloaded from YouTube. To our knowledge, this is the first benchmark to provide such unconstrained data. Another early, popular example is the **UCF-Sports** benchmark [28] with its 200 videos of nine different sports events, collected from TV broadcasts. Sports videos were also considered in the **Olympic-Games benchmark** of [23].

Feature films were used by some as a source for challenging action recognition videos. These included the **UCF Feature Films Dataset** [28], sometimes referred to as the "Kissing-Slapping" benchmark, which provides 90 videos of kissing and 110 of slapping scenes obtained from a range of classic movies, and the challenging **Hollywood-2 (HOHA2)** [21] from 2009, an extension to the **Hollywood-1 (HOHA)** benchmark [16] released a year earlier. Finally, the most recent benchmark to use TV and motion picture data, included in this survey, is the **High-Five** collection of [26], which, like the UCF Feature Films Dataset, focuses on interactions, rather than single person activities.

Of these benchmarks, the two Hollywood sets seems to remain challenging even today. This, despite the relatively small number of action categories they include. One possible reason for this is that the action samples in these sets are not well localized in time; action recognition systems tested on these sets must include temporal action *detection*, localizing the action in time, not just recognizing it, thereby making it more challenging than other sets.

2.3. Recent years: Action recognition "in the wild"

TV and motion picture videos introduced unconstrained viewing conditions, yet they were in some sense still limited in the range of challenges they represented. Specifically, although unconstrained, the videos in these sets were all typically produced under favorable conditions, were of high quality, and were shot from carefully selected viewpoints. Not surprising, high performances were already reported for most of these sets. All this cannot be said of the many videos accumulating in online repositories such as YouTube. In recent years, several have offered bench-

Table 1. **Action Recognition Databases.** ♯acn. is the number of action classes. ♯clips are the numbers of clips in the collection. SotA reflects the best performances obtained on each benchmark in recent years. It is important to note that these measures are *not* comparable, as the different benchmarks use different performance measures. Moreover, these scores do not reflect the improvement over chance, which varies from one benchmark to the other (e.g., 2% for the UCF50 benchmark [27] and 50% for ASLAN [12]).

Database	Year	♯acn.	♯clips	Setting	Technical details	SotA
KTH [32]	2004	6	600	Laboratory: 25 actors, 4 conditions, 4 repetitions = 2391 sub-sequences	Homogeneous background, static camera, 25fps, 160x120px, 4s duration, AVI DVIX-compressed	$\sim 95\%$ [7]
WEIZMANN [1]	2005	9	81	Laboratory: 9 actors	Static background, low resolution, 25fps	100% [40]
UMD [37]	2006	10	100	Laboratory: 1 actor, many repetitions	Two synchronized views, resolution 300px	100% [37]
IXMAS [41]	2006	11	110	Laboratory: 10 actors arbitrary orientation, 5 view points with multiple cameras, 30 sub-sequences	Resolution 100-200px, very short sequences	93.6% [38]
UIUC1 [36]	2008	14	532	Laboratory: 8 actors, single view, extensive repetition	Resolution 400px	N/A
UIUC2 [36]	2008	2-4-5	3	Youtube videos: 3 badminton games	Resolution 80px	98.3% [38]
UCF-sports [28]	2008	9	200	Real sports broadcasts	Unconstrained: wide range of scenes and viewpoints, simple background, resolution 720x480px	95% [29]
Olympic-games [23]	2008	17	166	Video footage from Olympic games	5065 manually annotated frames: high intra class variability, background clutter, large camera motion, motion blur, occlusions and appearance variations	69.6% [23]
Kissing-Slapping [28]	2008	2	200	Feature Films	Large variability in genres, scenes and views, actors	96.75% [2]
Hollywood1 (HOHA1) [16]	2008	8	430	Short sequences from 32 movies	See below.	62% [33]
Hollywood2 (HOHA2) [21]	2009	12	3669	69 movies 20.1 hours of video in total.	Large intra-class variability, label ambiguity, multiple persons, challenging camera motion, rapid scene changes, unconstrained and cluttered background, high quality, 240x450px, 24fps.	59.9% [38]
UCF-YouTube [20]	2009	11	1168	Youtube and personal videos	25 sub-groups: different environments and sources. Mix of steady and shaky cameras, cluttered background, variation in object scale, views points and illumination, low resolution (mpeg4-codec)	86.1% [30]
ADL [22]	2009	10	150	Laboratory: 5 actors, 3 repetition	Complex activities in living environment, static background, res. 1280x720px, 30fps, duration 10-60s	96% [39]
High-Five [26]	2010	4	300	23 different TV shows	30-600 frames, realistic human interactions: varying number of actors, scale, and views	68% [18]
YouTube Olympic Sports [25]	2010	16	800	YouTube videos	Complex activities, 50 sequences per class, labeled by Mechanical Turk	82.7% [6]
HMDB51 [14]	2011	51	6,766	Motion pictures, YouTube	Over 100 clips per class. Stabilized videos.	46.6% [38]
UCF50 [27]	2012	50	6,676	YouTube videos	100 videos per action category. Each category organized into 25 groups with clips sharing common features	72.68% [10]
ASLAN [12]	2012	432	3,697	YouTube videos	Binary, "same / not-same" classification of never-before-seen action pairs (training on action-category-exclusive data)	66.1% [17]

marks assembled from such "in-the-wild" sources.

One of the first to offer such videos was the **UCF-YouTube** set [20], also referred to as the YouTube Actions benchmark. Taking advantage of the availability of huge numbers of videos, it provides 1,168 videos in eleven categories, obtained from both YouTube and personal video collections. More recently, and returning to Olympic sports events, the **Olympic Sports Dataset** [25] was released, containing 16 complex actions represented by 50 videos each, all downloaded from YouTube. Here, an attempt was made to go beyond atomic actions, by considering events involving several different stages or different actions performed at once. An example being the long-jump, which combines standing, running, jumping, landing and standing up again.

2.4. Emerging trends

Two conclusions are clearly evident from the survey above and further illustrated in Figure 1. The first is that many of these benchmarks have been saturated. This testifies to the high performance and capabilities of modern action recognition techniques, as well as suggests that new benchmarks must be designed to reflect more challenging problem settings. The second conclusion is evident when comparing these benchmarks to those used in photo classification applications: action recognition benchmarks currently provide a few dozen categories to the thousands of-

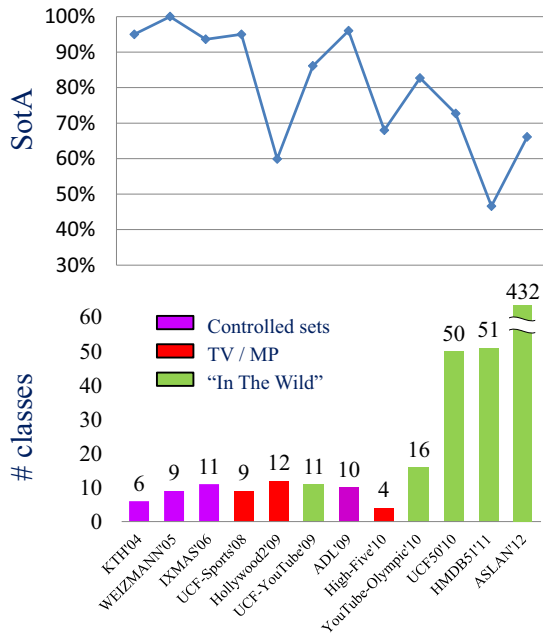

Figure 1. **Visual summary benchmarks.** Some benchmarks from Table 1, ordered from left to right by year of publication. Bottom: number of action categories in each benchmark, color coded by the video sources (controlled sets in purple, TV and motion pictures in red, and "in the wild", web-videos in green). Top: state-of-the-art performances for each benchmark. Note that these performances are not directly comparable as different performance measures are used with each benchmark, and in particular, chance values vary according to the number of categories.

fered by image classification sets, and thousands of videos to the millions of images (e.g., [5, 35]).

In an effort to address these issues, three recent benchmarks were proposed, offering significantly more videos and more action categories than those listed above. The first is the **HMDB51** set [14]. Released in 2011, it provided $6,766$ videos in 51 categories, significantly more than previous benchmarks. These were collected from various web repositories and motion pictures. The challenge this set poses is evident in the state-of-the-art results published on this set, reaching only 46.6% in [38].

A similar number of classes and videos was made available by the **UCF50** benchmark [27], which extends the UCF-YouTube discussed above. Here, all videos were harvested from YouTube, and focus on articulated actions – excluding actions restricted to facial expressions, etc. of which four are included in the HMDB set. The evaluation protocol recommended by [27] is a Leave-One-group-Out (LOgO) cross validation scheme for which the best reported scores are 72.68% in [10] with chance being 2%.

3. The ASLAN benchmark

A different benchmarking approach was taken when assembling the **Action Similarity LAbeiNg** (ASLAN) challenge [12], which put an emphasis on providing substantially more action categories. ASLAN contains video samples downloaded from YouTube, of 432 action categories. Labeling videos into separate action groups can be costly and particularly difficult for videos where action definitions may be ambiguous. To this end, and following the example of [9], rather than designing a multi-class categorization benchmark, the ASLAN benchmark is a binary, "same"/"not-same" classification task. That is, given two videos of never-before-seen actions, the task is to determine if the same action is being performed in both videos. In so doing, it encourages the development of action *similarity* measures and techniques, rather than the design of methods for learning the properties of particular actions.

Table 2. **Results on the ASLAN benchmark.** See text for details.

	Method	Acc. $\pm$ SE	AUC
1	HOG, $\sum(x_1.*x_2)$ [12]	56.58 $\pm$ .74	61.6
2	HOF, $\sqrt{\sum(x_1.*x_2)}$ [12]	56.82 $\pm$ .57	58.5
3	HNF, $\sqrt{\sum(x_1.*x_2)}$ [12]	58.87 $\pm$ .89	62.1
4	STIP, 12 sim. [12]	60.88 $\pm$ .77	65.3
5	OSSML [11]	64.25 $\pm$ .70	69.1
6	MIP [10]	64.62 $\pm$ .80	70.4
7	MIP+STIP [10]	65.45 $\pm$ .80	71.9
8	Traj. [17]	62.02 $\pm$ 1.1	66.9
9	MBH [17]	64.25 $\pm$ .90	69.9
10	Multi-rep. [17]	**66.13** $\pm$ **1.0**	**73.2**
11	Human [12]	94.19	97.8

We next briefly review the performances reported on the ASLAN benchmark by different methods. Results are reported for View-2, on a ten-fold cross validation task, where the *estimated mean accuracy* (Acc.) $\hat{\mu}$ is given by

$$\hat{\mu} = \frac{\sum_{i=1}^{10} Pi}{10}$$

Pi being the percentage of correct classifications on View-2, using subset i for testing. In addition, the *standard error of the mean* (SE) is reported, given by,

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}},$$

where,

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{10}(Pi - \hat{\mu})^2}{9}}.$$

Finally, we report the area under the ROC curve (AUC) for each method. Performances are summarized in Table 2.

The original results on the ASLAN benchmark, reported in [12], evaluated performance using Bags-of-features (BoF) produced using 5,000 word vocabularies and using different Space Time Interest Points (STIP) [15], namely HOG, HOF and HNF. Two videos were compared using different (dis-)similarities and a threshold was automatically determined over these values using linear support vector machines trained on separate training data (Table 2, rows

1-3). Twelve different similarities were computed for all three STIP representations and combined into 26D vectors representing each video pair. These were again classified as same/not-same using linear-SVM in Table 2 row 4. In [11] a metric learning approach was designed for the One Shot Similarity of [42]. Their OSSML approach improved performance but at a significant computational cost (row 5).

Motion Interchange Patterns (MIP), presented in [10], are low-level features, encoding at each space-time pixel a measure of the likelihood for a change in motion occurring at that pixel. A video is represented by building BoF using these codes. Mechanisms are described for both motion stabilization and screening of irrelevant codes. Their system is more efficient than OSSML, yet achieves better performance rates (row 6). By combining MIP with the three STIP (HOG, HOF, and HNF) results are slightly improved, suggesting that the MIP descriptors already capture much of the information available from the STIP descriptors.

The Dense Motion Trajectories representation (Traj.), originally described in [38], encode point trajectories over time, on a dense grid in space and scale, using optical flow computed between subsequent frames. Each trajectory is encoded as a vector of (normalized) 2D displacements of points. ASLAN results obtained with this representation were reported in [17] using the code made available by the authors of Traj. [38] and appear here in row 7 of Table 2.

In [17] the Motion Boundary Histogram (MBH) representation of [4] were also evaluated. MBH represents motion by considering the horizontal and vertical derivatives of motion in each pixel, separately, in order to compare the relative motion of pixels to their neighbors. In [38] this representation was used to encode actions by quantizing these derivatives into 8-bin, weighted histograms. Results from [17] using MBH are reported in row 8.

Finally, [17] propose a number of extensions to the MIP descriptor [10]. For the HistMIP descriptor, MIP codes are computed separately over different channels of each frame. Each such channel captures spatial gradient orientations, quantized into local spatial histograms and weighted by their magnitude. Thus, HistMIP is designed to reflect the changes in motion of local *textures*, rather than intensities of the original MIP. The DoGMIP, on the other hand, first processes each frame to produce multiple Difference of Gaussians layers. These are then processed, producing MIP codes for each layer. The results reported in [17] show that none of these representations, on their own, provides a significant performance boost over existing methods. Their combination, along with the other descriptors described above, however, does improve performance, as evident in row 10 of Table 2, which is the best performance reported on the ASLAN benchmark to date.

Table 2 also provides human performances on the ASLAN benchmark, originally published in [12]. These results demonstrate the remaining gap between man and machine on the task represented by the ASLAN benchmark.

4. Conclusions

The survey presented here demonstrate that contemporary action recognition systems can now perform very well in controlled settings, with few atomic actions and clear, known viewing positions, and even on uncontrolled videos of high quality, when actions are well localized in time. Unconstrained action recognition from "real-world" videos, however, remains a challenging problem. State-of-the-art methods are currently very far from the perfect scores obtained by systems developed for image classification. This fact is underscored by the properties of the benchmarks used to develop, train, and test systems for action recognition in videos, compared to systems for image classification: where the latter include hundreds of action categories, peaking at thousands of video samples, the former provide thousands of categories and millions of photo samples. This is particularly troubling when considering that actions can vary greatly in how they appear in videos, possibly far more than static items vary in appearance in photos, implying that data sets for action recognition must be designed to offer *more* examples, not less.

In an effort to bridge this gap, the ASLAN benchmark has recently been proposed in [12]. With 432 classes, it offers a far greater variability of categories than other existing benchmarks. In this paper we review the performance reported on this set and show the wide gap that remains between the best performing methods and human capabilities. We do so in an effort to motivate subsequent research into action recognition on larger and more challenging video collections, reflecting realistic, "in the wild" conditions.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2005.

[2] M. Bregonzio, J. Li, S. Gong, and T. Xiang. Discriminative topics modelling for action feature selection and recognition. In *Proc. British Mach. Vision Conf.*, volume 6, 2010.

[3] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Comput. Vision Image Understanding*, 2013.

[4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conf. Comput. Vision*, pages 428–441. Springer, 2006.

[5] C. Fellbaum. Wordnet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243, 2010.

[6] A. Gaidon, Z. Harchaoui, C. Schmid, et al. Recognizing activities with cluster-trees of tracklets. In *Proc. British Mach. Vision Conf.*, 2012.

[7] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *Proc. IEEE Int. Conf. Comput. Vision*, 2009.

[8] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *Int. Workshop on Socially Intelligent Surveillance and Monitoring at the IEEE Conf. Comput. Vision Pattern Recognition*, 2012.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. UMASS, TR 07-49, 2007.

[10] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *European Conf. Comput. Vision*, 2012.

[11] O. Kliper-Gross, T. Hassner, and L. Wolf. One shot similarity metric learning for action recognition. *Proc. of the Workshop on Similarity-Based Pattern Recognition (SISM)*, 2011.

[12] O. Kliper-Gross, T. Hassner, and L. Wolf. The action similarity labeling challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3):615–621, 2012.

[13] T. Kobayashi and N. Otsu. Motion recognition using local auto-correlation of space–time gradients. *Pattern Recognition Letters*, 33(9):1188–1195, 2012.

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proc. IEEE Int. Conf. Comput. Vision*, 2011.

[15] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2):107–123, 2005.

[16] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2008.

[17] N. Levy, Y. Hanani, and L. Wolf. Evaluating new variants of motion interchange patterns. In *CVPR Workshop on Action Similarity in Unconstrained Videos*, 2013.

[18] B. Li, M. Ayazoglu, T. Mao, O. I. Camps, and M. Sznaier. Activity recognition using dynamic subspace angles. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 3193–3200. IEEE, 2011.

[19] H. Liu, R. Feris, and M.-T. Sun. Benchmarking datasets for human activity recognition. In *Visual Analysis of Humans*, pages 411–427. Springer, 2011.

[20] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 1996–2003, 2009.

[21] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 2929–2936, 2009.

[22] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Proc. IEEE Int. Conf. Comput. Vision*, pages 104–111, 2009.

[23] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2008.

[24] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Comput. Vision*. 2013.

[25] J. C. Nieblesand, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European Conf. Comput. Vision*, 2010.

[26] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid. High five: Recognising human interactions in tv shows. In *Proc. British Mach. Vision Conf.*, 2010.

[27] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Mach. Vision and Applications*, 2012.

[28] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 1–8, 2008.

[29] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 1234–1241. IEEE, 2012.

[30] M. Sapienza, F. Cuzzolin, and P. Torr. Learning discriminative space-time actions from weakly labelled videos. In *Proc. British Mach. Vision Conf.*, 2012.

[31] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(2):162–177, 2005.

[32] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *Proc. 17th Int. Conf. Pattern Recognition*, volume 3, pages 32–36, 2004.

[33] A. H. Shabani, D. A. Clausi, and J. S. Zelek. Evaluation of local spatio-temporal salient feature detectors for human action recognition. In *Comput. and Robot Vision*, 2012.

[34] F. Shi, E. Petriu, and R. Laganiére. Sampling strategies for real-time action recognition. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*. IEEE, 2013.

[35] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, 2008.

[36] D. Tran and A. Sorokin. Human activity recognition with metric learning. In *European Conf. Comput. Vision*, 2008.

[37] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury. The function space of an activity. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 959–968, 2006.

[38] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vision*, pages 1–20, 2012. Available: lear.inrialpes.fr/people/wang/dense_trajectories.

[39] J. Wang, Z. Chen, and Y. Wu. Action recognition with multiscale spatio-temporal contexts. In *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, pages 3185–3192. IEEE, 2011.

[40] Y. Wang and G. Mori. Human action recognition by semilatent topic models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10):1762–1774, 2009.

[41] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Comput. Vision Image Understanding*, 104(2-3):249–257, 2006.

[42] L. Wolf, T. Hassner, and Y. Taigman. The one-shot similarity kernel. In *ICCV*, 2009.