Sports Videos in the Wild (SVW): A Video Dataset for Sports Analysis

Seyed Morteza Safdarnejad[1], Xiaoming Liu[1], Lalita Udpa[1], Brooks Andrus[2], John Wood[2], Dean Craven[2]
[1] Michigan State University, East Lansing, MI, USA
[2] TechSmith Corporation, Okemos, MI, USA

***Abstract*— Considering the enormous creation rate of user-generated videos on websites like YouTube, there is an immediate need for automatic categorization, recognition and analysis of videos. To develop algorithms for analyzing user-generated videos, unconstrained and representative datasets are of great significance. For this purpose, we collected a dataset of *Sports Videos in the Wild (SVW)*, consisting of videos captured by users of the leading sports training mobile app (Coach's Eye) while practicing a sport or watching a game. The dataset contains 4100 videos selected by reviewing ∼85,000 videos and consists of 30 sports categories and 44 actions. Videos of sports practice, which frequently happens outside the typical sports field, have huge intra-class variations due to background clutter, unrepresentative environment, existence of different training equipment and most importantly, imperfect actions. On the other hand, using smartphones for video capturing by ordinary people, in comparison to videos captured by professional crew for broadcasting, leads to challenges due to camera vibration and motion, occlusion, view point variation, and poor illumination. Given various manual labels, this dataset can be used for a wide range of computer vision applications, such as action recognition, action detection, genre categorization, and spatio-temporal alignment. On the sport genre categorization problem, we design the evaluation protocol and evaluate three different methods to provide baselines for future works.**

I. Introduction

The amount of digital videos being created is increasing exponentially, e.g., YouTube has reached the upload rate of 100 hours of video per minute. A great deal of this growth is due to the tremendous popularity of smartphones and ubiquitous Internet access. This means that *amateur-user-generated videos* form the new trend in content generation. Thus, there is an immediate need for robust algorithms to automatically analyze and retrieve videos.

Many computer vision approaches are data-driven and the existence of representative and realistic datasets is crucial for developing robust approaches. Therefore, there has been a trend from research on *controlled* datasets toward *unconstrained* datasets. For instance, recent face recognition research focuses on datasets like LFW [11] and YouTube faces database [12] rather than controlled datasets like FERET [13] or FIA [14]. Similarly, for human action recognition, datasets with less controlled videos, e.g., Hollywood2 [6], HMDB [8] and UCF101 [9], are gaining popularity, compared with staged datasets like KTH [1] or Weizmann [2]. While these datasets ([6], [8], [9]) are from YouTube videos and movies and thus have unconstrained environment and actions relative to staged datasets, many of the videos are captured *professionally*. Therefore, in aspects like camera vibration, view angle variation, and illumination, they are bound to common practices of filmmaking. On the other hand, specifically for sports videos, most videos in public datasets are representative of successful completion of the actions that may not truly reflect the highly complex and diverse real-world sports activities. Finally, for sports videos, due to strong correlation of background and the actions in existing datasets, the state-of-the-art performance on genre categorization is very high.

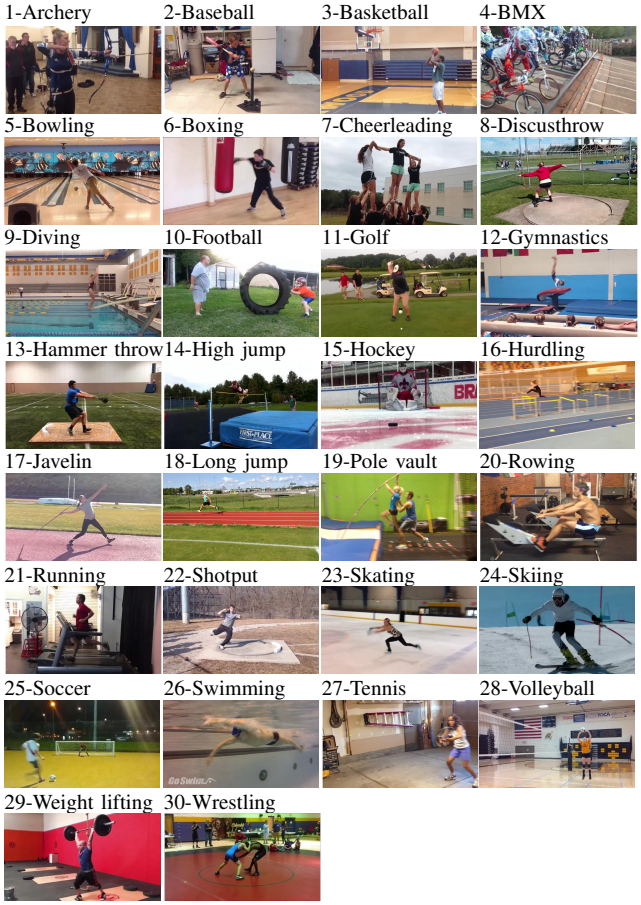


Fig. 1. Sample frames from all 30 sports categories of SVW.

Given the explosion of user-generated videos and the lack of real-world datasets for the research community, we present a highly unconstrained dataset of sports videos, called *Sport Videos in the Wild (SVW)*. SVW is comprised of videos captured *solely* with mobile devices by users of Coach's Eye® mobile app, a leading app for sports training

This work was partially supported by TechSmith Corporation. Contact author: Xiaoming Liu, liuxm@cse.msu.edu

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

TABLE I

DATASET COMPARISON FOR ACTION RECOGNITION (AR), ACTION DETECTION (AD), SCENE UNDERSTANDING (SU), AND GENRE CATEGORIZATION (GC).

Dataset	Purpose	Categ. #	Clip #	Avg. length	Unconst. actions	Unconst. capturing	Camera vibration	Orientation	Sources
KTH [1]	AR	6	100	NA	No	No	No	Landscape	Staged
Weizmann [2]	AR	9	9	NA	No	No	No	Landscape	Staged
IXMAS [3]	AR	11	30	NA	No	No	No	Landscape	Staged
UCF Sports [4]	AR	9	14+	NA	Yes	No	No	Landscape	Broadcast TV
Olympic [5]	AR	16	50	NA	Yes	No	No	Landscape	YouTube
Hollywood2 [6]	AR SU	A:12 S:10	61+ 62+	NA	Yes	No	No	Landscape	Movies
UCF50 [7]	AR	50	100+	NA	Yes	No	Slight	Landscape	YouTube
HMDB [8]	AR	51	101+	NA	Yes	No	Slight	Landscape	Movies & Internet
UCF101 [9]	AR	101	100+	7.2	Yes	No	Slight	Landscape	YouTube
THUMOS [10]	AR/AD	101	100+	NA	Yes	No	Slight	Landscape	YouTube
SVW	AR/AD GC	A:44 G:30	50+ 110+	11.6	Yes	Yes	Yes	Landscape & Protrait	Smartphone & Tablet

developed by TechSmith Corporation. The app allows users to conveniently capture videos whenever they practice a sport or watch a game. Fig. 1 shows sample frames from different categories of SVW. Being captured by mobile devices and by ordinary people, along with the fact that many videos are of practices of amateurs, not professional athletes, makes SVW the most unconstrained dataset of sports and action videos.

SVW is annotated to serve for multiple purposes. For *action recognition*, videos are labeled with 44 different actions and timespan of each action. For *action detection*, we provide time stamps of the actions, rather than trimming around each action. In addition, more than 50% of the videos are annotated with spatio-temporal bounding boxes around each and every action in the video. For *sports genre categorization*, each video is labeled with generic name of the sport being practiced, resulting in 30 sports categories.

For the sport genre categorization problem, we design the evaluation protocol and compare the performance of three algorithms on the proposed dataset as baselines for future research. First, the performance of the state-of-the-art motion-based dense trajectories algorithm [15] is reported. Second, purely context-based algorithm of describing videos with SIFT features [16] is presented. Finally, experiments using a motion-assisted context-based algorithm are conducted. All data, including the dataset, labels, evaluation protocol, and experimental results, are *publicly available* to the research community for future research[1].

II. RELATED WORK

Table I compares different aspects of most popular action recognition (AR) datasets, among which HMDB [8] and UCF101 [9] are the most challenging ones in terms of having unconstrained videos. To the best of our knowledge, there is no publicly available dataset for sports genre categorization.

[1]www.cse.msu.edu/~liuxm/sportsVideo

KTH [1] and Weizmann [2] datasets contain simple actions and their AR accuracies are reported to be above 90% [8]. IXMAS [3] contains staged actions captured by 5 calibrated cameras, where an AR accuracy of 93.5% is reported in [17].

UCF Sports [4] and Olympic [5] are the only datasets that cover just sports activities. While the environment is not controlled, the videos are *captured* by professional crew, the actions are *performed* by professional athletes, and the background is restricted to standard sports fields. As noted in [8], the actions in these datasets are highly distinguishable from shape cues alone. For Olympic, an accuracy of 91.1% is reported in [18]. Having limited number of categories and distinct activities in each category, a recognition rate of 98% is reported in [8] for UCF Sports using the information from static joint locations alone.

Gathered from 69 movies, Hollywood2 [6] is labeled for both action recognition and scene understanding. Being selected from movies, it contains unconstrained environment and actions while benefiting from professional capturing. Its main restrictions are the small number of actions and the fact that clips extracted from the same movie share similar scenes. A 64.3% AR accuracy is reported in [18].

For UCF50 [7], Kuehne et al. [8] suggest that low-level features are as predictive as mid-level features and Wang et al. [18] report a 91.2% AR accuracy. As an extension of UCF50, UCF101 has 101 categories and is the largest AR dataset available [9]. Being collected from YouTube, the actions are fairly unconstrained, but no comment can be made about the capturing process. Karpathy et al. [19] report a 66% AR accuracy for UCF101 (80% for the sports group). Probably due to the low resolution of source videos, all clips are normalized to the relatively low resolution of 320×240. At the mean clip length of 7.2 second, UCF101 is fairly short compared to SVW, making it less suitable for action detection problems. Recently, in THUMOS challenge [10],

three sets of temporally *untrimmed* validation, background, and test videos are used along with UCF101 videos as training, to push the action recognition and detection tasks toward real-word scenarios.

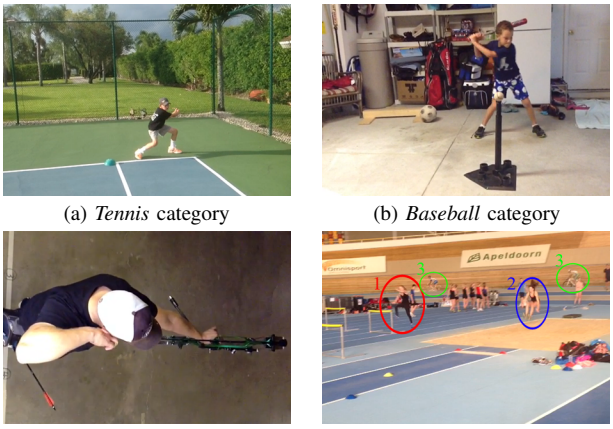
HMDB [8] is collected by looking for non-ambiguous human actions in Internet videos and movies. As a quality standard, selection of videos has been constrained to having a single action per clip and 40% of the clips are not affected by camera motion. The dataset is prepared in two versions of original videos and stabilized videos and good performance is reported for stabilization. In [18], an accuracy of 57.2% is reported for HMDB. HMDB is very challenging due to not only the unconstrainedness of the dataset, but also having multiple shots in a single clip, where both factors contribute to the low AR accuracy.

Although existing datasets have some levels of unconstrained actions and environment, there is still more complexity in real-world videos that need to be represented in research datasets. Specifically for sports videos, current datasets do not provide highly unconstrained conditions. For UCF101, one of the most challenging datasets, sport videos achieve the highest recognition rate ([9], [19]) among different types of videos. This is claimed to be due to distinctiveness of sports motions and less cluttered background in official sports field than other types of actions, which does not hold for sports in the wild. Considering high performances reported for UCF Sports, Olympic, and sports groups of UCF101, SVW specifically fills the research gap for analyzing *challenging* sports videos. On the other hand, uploading a video to YouTube implies that the action of desire has been successfully performed and completed in the video. But for a completely unconstrained video, there might be failure cases (e.g., batting practice). In addition, unlike UCF101 or Hollywood2, in SVW no two videos are trimmed from a single footage captured by users, which keeps the variance of the actions, environment, and shooting conditions in SVW as high as possible. Furthermore, due to unconstrained environment and illumination as well as a high rate of scene occlusion by people, video stabilization of SVW is very challenging and our experiments show a high failure rate of stabilization using the common RANSAC algorithm. Finally, the video resolution and length of SVW are larger than all the current datasets, and SVW includes both landscape and portrait orientation of videos.

III. Sports Videos in the Wild (SVW) Dataset

A. Dataset details and statistics

Dataset collection: SVW is selected from the videos captured by ordinary users of Coach's Eye mobile app developed by TechSmith Corporation, when users practice a sport or watch a game. The users can review the videos and compare them with those of coaches or professional athletes side by side. A user may also upload the videos to the app server for other users to review and comment on his sports training progress. At the time of writing this paper, an average of 4 videos per minute are being uploaded to the app server by users, and among 700,000 uploaded videos, users have marked ∼418,000 as publicly usable. Due to the highly non-uniform distribution of sports categories, 85,000 videos from the public set have been reviewed and labeled to collect enough videos for 30 sports category and 44 action categories with at least 110 and 50 videos per category, respectively.



(a) *Tennis* category (b) *Baseball* category

(c) *Archery* category (d) *Hurdling* category

Fig. 2. SVW challenges: (a) Related equipment does not exist, (b) Background is cluttered and uncorrelated with the sport, (c) Uncommon camera angles increase the intra-class variations, (d) Multiple sports co-exist (1: Hurdling, 2: Long jump, 3: Cycling).

Challenges of SVW: Compared to broadcasting videos, sports videos in the wild have many unique challenges for visual analysis, due to both the imperfect practices of *amateur players* and unprofessional capturing by *amateur users*. Firstly, the static image context is less discriminative for categorization. For example, in a video of tennis forehand drill (Fig. 2 (a)), no assumption can be made about existence of the racquet (and in some cases the tennis court). The only reliable clue may be the unique motion characteristics of the hands. Secondly, in these videos, existence of training equipment is more likely than the broadcasting videos (Fig. 2 (b)). On the other hand, cluttered backgrounds as well as common environments also cause difficulties in unconstrained sports videos. There are many SVW videos that the sport is practiced inside the house, in the garage, or in the backyard (Fig. 2 (b)). Thirdly, unprofessional capturing by amateur users introduces additional challenges like extreme camera vibration, improper camera movement, occlusion from audience, judges and fences due to improper camera location, and uncommon view angles (Fig. 2 (c)). Finally, for amateur videos, it is more probable to have multiple activities in a single video (Fig. 2 (d)).

It is important to note that unlike other action recognition datasets that are recently widely used, multiple actions defined in SVW may come from a *single* sport (see Fig. 3). In other words, while the environment is quite similar for these subsets of actions, movements are completely different. This introduces further challenges in visual analysis of sports videos in the wild for the purpose of action recognition. On the other hand, this arises difficulties for genre categorization. Each sport category has huge intra-class variations due to containing multiple actions that can appear at any timespan of the entire video length.

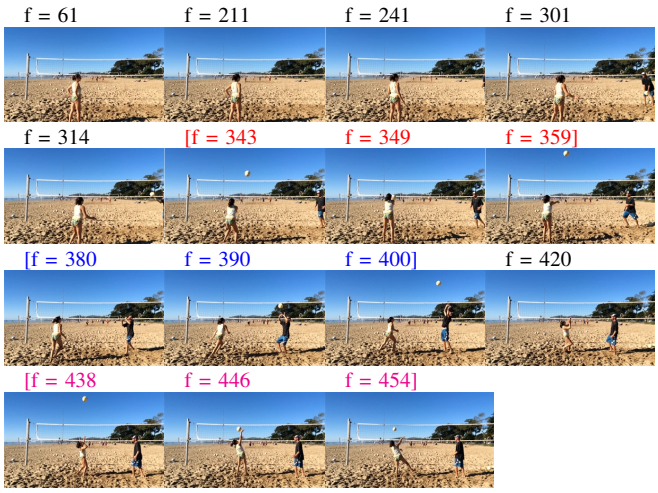


Fig. 3. Annotated actions categories ([343, 359, Forearm], [380, 400, Set], [438, 454, Spike]) within a video from *Volleyball* category. Since distinct actions from the same sport genre may share a common field, visual appearance alone is not enough for action recognition in SVW.

Dataset labeling: Videos are manually labeled in a two-round scheme. First, for each clip, 6 frames uniformly sampled across the video length constitute a montage, which is saved as an image. A GUI equipped with a button for each category shows the saved montage and records the pressed button from the labeler. In the next round, all labeled clips are reviewed one by one. Clips over 1-minute long are trimmed to loosely cover representative motions, but not precisely around the action of interest so that the dataset is also suitable for action detection. To prepare SVW for action recognition, at least 50% of the videos are reviewed closely to annotate *all* pre-defined actions within a clip and their corresponding timespans. The same videos are also annotated with bounding boxes around each action in the video for action detection. Fig. 3 shows how different actions within a clip are annotated with the label and time stamps.

For each video clip, we also label various meta tags. Fig. 4 represents the distribution of the number of participants in videos, commonality/uniqueness of the action environment, and the camera view angles for 30 sports categories. Meta tags reveal that 19% of SVW videos are affected by considerable camera vibration and the videos of three categories have the highest rates of training equipment usage, Running (9%), Weight lifting (9%), and Boxing (4%). Multiple activities in a single video are more common for categories such as Hurdling, High jump, Running, Weight lifting, and Diving.

Spatial resolution normalization: The resolution of the original videos varies from 480×272 to 1280×720 (irrespective of video orientation) with 640×360 being the most common size. Since for some analysis algorithms variation of video sizes might result in the confusion of scene scales, a normalized version of the dataset is provided along with the original one. Having both landscape and portrait orientations in the dataset, normalized clips have the maximum size (width or height) of 480 pixels.

Evaluation protocol: In line with UCF101 and HMDB, three splits of 70% training and 30% testing are generated for the genre categorization application of SVW. We designate the splits by aiming to evenly distribute different actions, camera view angles, and field characteristics over the splits. The genre categorization accuracy is used as the performance metric and is defined as the fraction of testing videos whose genres are correctly classified.

B. Potential applications of SVW

Action recognition: Due to the huge number of video content available online and the desire to content understanding, a great deal of effort has been focused on action recognition from videos [6], [20], [21], [22]. Inherently, for sports videos, action recognition is a subset of the genre categorization problem, i.e., for the former, labels for a single action are available but for the latter, a group of different actions within each sport are all labeled with the genre of the sport, resulting in higher intra-class variations.

Action detection: Although there has been great emphasis on action recognition, the action detection problem has not been extensively studied. Action detection by itself and as part of *recognition by detection* systems [23] is an important problem to be tackled. Especially, in real-world videos, actions of interest may cover a relatively short period of a video and it is important to be able to detect these actions. Existing approaches use rather simple datasets with short videos [24], [25] or proprietary datasets [26]. SVW enables researchers to push the limit of action detection toward more realistic videos.

Genre categorization: Sports genre categorization is vastly studied for broadcasting TV videos [27], [28], [29], [30], [31], [32]. In these works, it is assumed that sports occur in sports arena (implicitly assuming the existence of specific equipment and field lining) and are captured by professional TV broadcasting crew. Low-level features like color, motion, and histogram of edge directions are used for categorization. Our experiments show that this type of approaches does not perform well on sports in the wild. On the other hand, in [33], authors report superior performance of the dense trajectories method [15] for genre categorization of unconstrained proprietary videos. This paper aims to provide a dataset of such videos. Well-known sports-only datasets of UCF Sports [4] and Olympic [5], include specific actions not generic sports categories, and have been reported to achieve $\sim$90% accuracy [17], [18] (the method in [17] achieves $\sim$62% accuracy on SVW). Thus, a challenging video dataset for this application is highly desirable.

Spatio-temporal alignment: Given two video sequences of the same action, spatio-temporal alignment estimates the spatial and temporal coordinate transformation that maps the actions of interest in one video to those of the other [34], [35], which recovers the body pose information [36]. For the case of sports videos, spatio-temporal alignment of actions enables effective comparison of actions performed by different people. This is specifically useful for the purposes of sports training and grading. Furthermore, the *joint alignment*

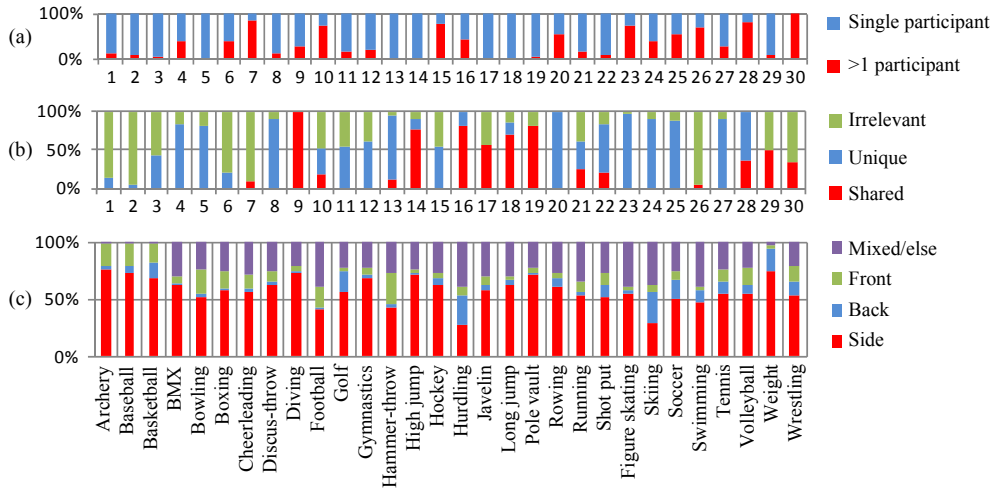


Fig. 4. Distribution of (a) number of participants in videos, (b) aspects of the action field and (c) camera views angles, in 30 categories. *Irrelevant field* is a field that from its appearance, the sports category cannot be deduced (e.g., practicing in the backyard). *Shared field* refers to the condition in which from just field appearance, more than one sports category might be inferred (e.g., track and field sports). *Unique field* is the one that just from field context, the corresponding sports category can be conjectured (e.g., Bowling tracks).

of multiple videos, from either one user over time or a diverse set of users, allows us to study the temporal evolving and inter-subject variations of a particular action, which are novel research problems by themselves. Similar problems of joint alignment of images have been studied for faces [37] and general objects [38]. Having unconstrained videos where action of interest may happen at any temporal segment of the video, SVW serves as a realistic and challenging dataset for the alignment problem.

IV. Baseline Experiments

In this section, we present the performances of three different algorithms for the genre categorization problem on SVW. The first algorithm summarizes features extracted from dense trajectories [15] using the widely used Bag of Words (BoW) approach [39]. The second algorithm analyzes the context of video frames using the BoW on the SIFT features [16]. The third one, a motion-assisted context-based algorithm, segments the moving and stationary pixels using trajectory information and then analyzes the appearance of these two groups of pixels separately. To the interest of computational cost and memory, for all methods, a two-level bottom-up codebook generation scheme is used [32]. At the first layer, for each class, a set of codewords are generated using K-means clustering. At the second layer, codewords of all classes are aggregated and by another round of clustering, the final codewords are obtained. We use Support Vector Machine (SVM) as the classifier for all the algorithms. Table II summarizes the genre categorization accuracies of the baseline algorithms. Unlike UCF Sports, which is conjectured in [8] to be equally predictable using contextual or motion information due to the fact that many sports in UCF Sports are location-specific, and similarly Olympic dataset, sports videos in the wild are better recognized using motion features due to existence of many practice videos in environments uncorrelated with the activities. However, the accuracy achieved by motion-based algorithm is relatively low due to miscellaneous aforementioned challenges of SVW. Fig. 5 represents confusion matrices of context-based and motion-based algorithms. The contrast of off-diagonal elements indicates the potential benefits of fusing these two algorithms. More detail on all three algorithms follows.

A. Motion-based algorithm

For motion-based algorithm, the state-of-the-art approach of dense trajectories is used [15]. The BoW approach on top of dense trajectory based features has been reported to outperform those of space-time interest points on various datasets [17]. This approach consists of three main steps: video stabilization, trajectory extraction and description, and BoW representation of trajectory information. We use implementations in [18] for the second step.

a) Video stabilization: Frame by frame motion stabilization is achieved by matching interest points on consecutive frames and applying RANSAC to obtain the affine transformation between frames. Due to issues such as poor illumination, moving subjects and audience, and uniform or non-rigid backgrounds (e.g., water), the failure rate of video stabilization is quite high, which deteriorates the overall performance of the motion-based algorithm.

b) Dense trajectories: As proposed in [15], dense trajectories are extracted at multiple spatial scales. Each point $p_t = (x_t, y_t)$ at frame t is tracked to the next frame $t+1$ by performing the median filtering in a dense optical flow field $\mathbf{W} = (u_t, v_t)$, $p_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (K * \mathbf{W})|_{(\overline{x_t}, \overline{y_t})}$, where K is the median filtering kernel and $(\overline{x_t}, \overline{y_t})$ is the rounded position of (x_t, y_t). Trajectories are started from the sample points on a grid spaced by W pixels (set to 5). The

TABLE II
Genre categorization accuracy on SVW.

Method	Motion-based	Context-based	Motion-assisted context
Performance	61.53%	37.08%	39.13%

TABLE III
PERFORMANCE OF DIFFERENT COMBINATIONS OF TRAJECTORY DESCRIPTORS ON SVW.

Descriptors	**s**	HOG	HOG + **s**	MBH	MBH + HOG	MBH + HOG + **s**
Performance	44.65%	56.45%	60.43%	58.12%	60.69%	61.53%

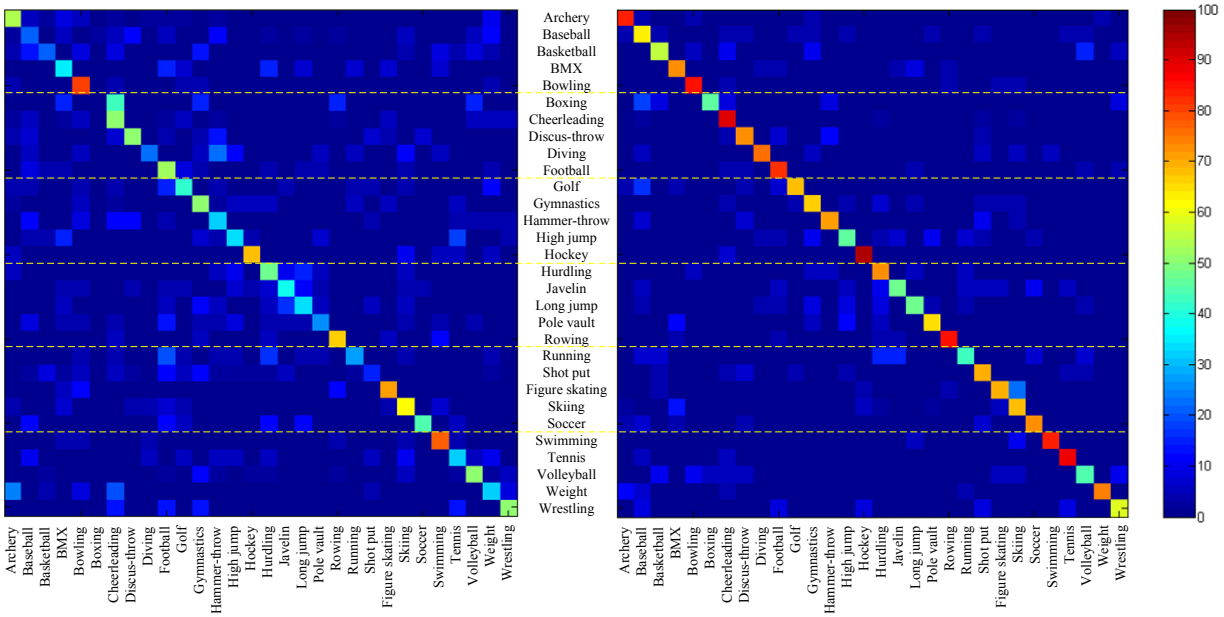


Fig. 5. Confusion matrices of context-based (left) and motion-based (right) categorization algorithms.

length of each trajectory is limited to L (set to 15), and after reaching this length, the trajectory is removed from the tracking process and new sample points are tracked.

c) Trajectory descriptors: The shape of the trajectories can be used as a representative feature, especially for sports analysis. In [15], the displacements of trajectory, $\Delta p_t = (x_{t+1} - x_t, y_{t+1} - y_t)$, over L consecutive frames are concatenated to be a vector, $\hat{\mathbf{s}} = (\Delta p_t, ..., \Delta p_{t+L-1})$, which is further normalized to be a trajectory descriptor $\mathbf{s} = \hat{\mathbf{s}} / \sum_{j=t}^{t+L-1} \|\Delta p_j\|$. Similar to [15], the video volume of a neighborhood of each trajectory is aligned and the resultant volume is described by using the Motion Boundary Histogram (MBH) [40] and the Histogram of Oriented Gradients (HOG) [41]. Table III shows the performance of different combinations of descriptors. The highest accuracy of 61.53% is achieved by combining MBH, HOG, and **s** descriptors.

B. Context-based algorithm

We follow the algorithm in [32] for analyzing the videos using only the static contextual information. In this algorithm, we sample one frame per second of the video and use the BoW representation of SIFT descriptors for categorization. For this algorithm, no inter-frame information is utilized, thus video stabilization is not required. For dictionary learning, we use 10 videos per category, and the codebook size of 4000. As shown in Table II, this method achieves the categorization accuracy of 37.08%.

C. Motion-assisted context algorithm

Along with the idea in [7], we augment the context-based method with the information of moving and stationary pixels. This can be loosely considered as foreground-background segmentation using motion information. For this purpose, the mean position of trajectories of the stabilized videos, for which the standard deviation of the trajectory points is beyond a threshold, is considered as a moving point. The decision about a moving point at a certain frame is propagated to 15 frames before and after the frame on which the trajectory ends. Having groups of moving and stationary pixels ready, SIFT descriptors and BoW representation are calculated for them separately and the resulting histograms are concatenated to represent the video. This algorithm achieves an accuracy of 39.13% (Table II), which is slightly better than the algorithm using context information only.

D. Discussion

Comparing ∼62% accuracy achieved by motion-based algorithm on SVW with ∼91% accuracy obtained by applying the same method to both UCF50 and Olympic Sports datasets [18], demonstrates that SVW is a very challenging sports video dataset. In addition, comparing accuracy obtained by applying motion-based and context-based algorithms (∼62% vs ∼39%) reveals that in SVW, motion is the main cue for categorization and action recognition. While the motion-assisted context based algorithm results in ∼39% accuracy for SVW, as reported in [7], a similar method achieves an accuracy of ∼67% for UCF50. This essentially suggests that background and equipment appearance in SVW are not as informative as in UCF50. In [19], a 80% accuracy is reported for Sports group of UCF101. Considering all these results, we may conclude that although sports videos

feature unique movements, analysis of truly unconstrained videos is still challenging and needs further research.

V. Conclusions

To advance computer vision research, and to push the limits of various video analysis problems toward more realistic and unconstrained scenarios, representative and unconstrained datasets are essential. In this regard, we introduced Sports Videos in the Wild (SVW), as a very challenging real-world dataset of sports videos available for genre categorization, action detection, action recognition, and spatio-temporal alignment. We evaluated three different baseline algorithms for sports genre categorization. Experimental results suggest that due to weak correlation between environment and actions in SVW, as well as amateur capturing of the videos, the presented SVW dataset is indeed the most challenging sports and action dataset available.

References

[1] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local SVM approach," in *ICPR*. IEEE, 2004, vol. 3, pp. 32–36.

[2] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," in *ICCV*. IEEE, 2005, vol. 2, pp. 1395–1402.

[3] Daniel Weinland, Edmond Boyer, and Remi Ronfard, "Action recognition from arbitrary views using 3D exemplars," in *ICCV*. IEEE, 2007, pp. 1–7.

[4] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah, "Action mach: a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, 2008.

[5] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *ECCV*, pp. 392–405. Springer, 2010.

[6] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid, "Actions in context," in *CVPR*. IEEE, 2009, pp. 2929–2936.

[7] Kishore K Reddy and Mubarak Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[8] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, "HMDB: a large video database for human motion recognition," in *ICCV*. IEEE, 2011, pp. 2556–2563.

[9] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[10] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," http://crcv.ucf.edu/THUMOS14/, 2014.

[11] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[12] Lior Wolf, Tal Hassner, and Itay Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR*. IEEE, 2011, pp. 529–534.

[13] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms," *IEEE T-PAMI*, vol. 22, no. 10, pp. 1090–1104, October 2000.

[14] Rodney Goh, Lihao Liu, Xiaoming Liu, and Tsuhan Chen, "The CMU Face In Action (FIA) database," in *Proc. IEEE Intl. Workshop on Anal. and Modeling of Faces and Gestures*, 2005, pp. 255–263.

[15] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *CVPR*. IEEE, 2011, pp. 3169–3176.

[16] David G Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[17] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, 2013.

[18] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.

[19] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[20] Ivan Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2-3, pp. 107–123, 2005.

[21] Yimeng Zhang, Xiaoming Liu, Ming-Ching Chang, Weina Ge, and Tsuhan Chen, "Spatio-temporal phrases for activity recognition," in *ECCV*. 2012, pp. 707–721, Springer.

[22] Cen Rao and Mubarak Shah, "View-invariance in action recognition," in *CVPR*. IEEE, 2001, vol. 2, pp. II–316.

[23] Sreemanananth Sadanand and Jason J Corso, "Action bank: A high-level representation of activity in video," in *CVPR*. IEEE, 2012, pp. 1234–1241.

[24] Junsong Yuan, Zicheng Liu, and Ying Wu, "Discriminative video pattern search for efficient action detection," *IEEE T-PAMI*, vol. 33, no. 9, pp. 1728–1743, 2011.

[25] Tae-Kyun Kim and Roberto Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE T-PAMI*, vol. 31, no. 8, pp. 1415–1428, 2009.

[26] Konstantinos G Derpanis, Mikhail Sizintsev, Kevin Cannons, and Richard P Wildes, "Efficient action spotting based on a spacetime oriented structure representation," in *CVPR*. IEEE, 2010, pp. 1990–1997.

[27] Jinjun Wang, Changsheng Xu, and Engsiong Chng, "Automatic sports video genre classification using pseudo-2D-HMM," in *ICPR*. IEEE, 2006, vol. 4, pp. 778–781.

[28] Ling-Yu Duan, Min Xu, Qi Tian, Chang-Sheng Xu, and Jesse S Jin, "A unified framework for semantic shot classification in sports video," *IEEE T-MI*, vol. 7, no. 6, pp. 1066–1083, 2005.

[29] Ning Zhang, Ling-Yu Duan, Qingming Huang, Lingfang Li, Wen Gao, and Ling Guan, "Automatic video genre categorization and event detection techniques on large-scale sports data," in *Proceedings of the 2010 Conference of the Center for Advanced Studies on Collaborative Research*. IBM Corp., 2010, pp. 283–297.

[30] Xun Yuan, Wei Lai, Tao Mei, Xian-Sheng Hua, Xiu-Qing Wu, and Shipeng Li, "Automatic video genre categorization using hierarchical SVM," in *ICIP*. IEEE, 2006, pp. 2905–2908.

[31] C Krishna Mohan and B Yegnanarayana, "Classification of sport videos using edge-based features and autoassociative neural network models," *Signal, Image and Video Processing*, vol. 4, no. 1, pp. 61–73, 2010.

[32] Ning Zhang, Ling-Yu Duan, Lingfang Li, Qingming Huang, Jun Du, Wen Gao, and Ling Guan, "A generic approach for systematic analysis of sports videos," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 46, 2012.

[33] S. Morteza Safdarnejad, Xiaoming Liu, and Lalita Udpa, "Genre categorization of amateur sports videos in the wild," in *ICIP*. IEEE, 2014.

[34] Yaron Ukrainitz and Michal Irani, "Aligning sequences and actions by maximizing space-time correlations," in *ECCV*. 2006, pp. 538–550, Springer.

[35] Serge Ayer and Martin Vetterli, "Method and system for combining video sequences with spatio-temporal alignment," 2001, US Patent 6,320,624.

[36] Xiaoming Liu, Ting Yu, Thomas Sebastian, and Peter Tu, "Boosted deformable model for human body alignment," in *CVPR*, 2008, pp. 1–8.

[37] Yan Tong, Xiaoming Liu, Frederick W. Wheeler, and Peter Tu, "Automatic facial landmark labeling with minimal supervision," in *CVPR*, 2009.

[38] Xiaoming Liu, Yan Tong, and Frederick W. Wheeler, "Simultaneous alignment and clustering for an image ensemble," in *ICCV*, 2009.

[39] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, 2004, pp. 1–22.

[40] Navneet Dalal, Bill Triggs, and Cordelia Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*, pp. 428–441. Springer, 2006.

[41] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005, vol. 1, pp. 886–893.