RESEARCH ARTICLE OPEN ACCESS

Recent Progress in Dataset Development for Video Classification

Sasithradevi. A [1], S. Mohamed Mansoor Roomi [2], P. Kalaivani [3], navarasi [4]
Electronics and Communication Engineering
Thiagarajar College of Engineering, Madurai
Tamil Nadu, India

ABSTRACT
Video classification has been a hot research topic in computer vision owing to its application in domains like video analytics, video retrieval, human machine interaction and video surveillance. The success of any video classification algorithm relies on the efficient representation of the video content and it's performance has long been demonstrated on complex datasets. The evolution of such datasets, challenges the growing computerized algorithms for video classification. Hence, we provide a detailed survey of existing datasets that has been developed in last decades to assess the quality of computer understanding in video classification. The available datasets for video classification is categorized based on application as: Activity classification and object classification. The objective of this paper is to provide knowledge on available datasets in these categories and to highlight the existing techniques capabilities on these datasets.
Keywords :— Video classification, Constrained datasets, Unconstrained datasets, Human activity recognition, Generic video classification.

I. INTRODUCTION

. Development in video capturing devices has major impact on the overflow of personal digital videos in Google, Flickr and You tube. Similarly, evolution of low cost storage devices has made the transfer of high-quality videos through less expensive hardware. Hence digital videos are surplus and had been seeking the need for effective video understanding methodologies and in particular, human activity and object classification. Over the recent years, this has made action recognition research curious, providing a wide range of effective techniques and systems in video research community. The benchmarks for action recognition have evolved along-side the growing capabilities of efficient machine vision techniques. As these methods perform well, Benchmarks have become more challenging like acquiring videos under uncontrolled/unconstrained environment. These datasets comprise of thousands of videos obtained outside the lab without constraints. In this article, a survey on these datasets is detailed. On one hand, we intend to review the growth of datasets in video classification and on the other hand we highlight

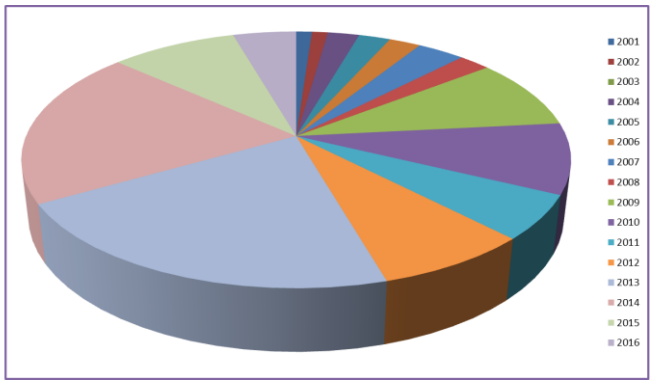

Fig. 1. Quantum datasets Vs year.

the capabilities of existing techniques on these datasets. This paper is not the predecessor to review the available datasets for video classification. The more recent articles include [1], [2], [3] provided an exhaustive research on simply activity recognition/classification. As shown in figure 1, majority of datasets had been evolved during 2013 and least during 2003. This figure gives an overall idea of evolution of datasets emerged over the

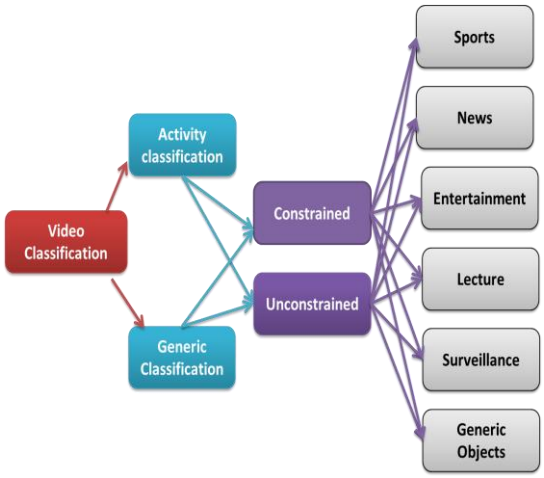

Fig. 2. Proposed Taxonomy.

decade. Together with this growth, efficient algorithms had also been proposed in the research community to solve the challenges in the dataset thereby reducing the semantic gap. As shown in figure 2, this paper aims to provide wide knowledge to the reader on available video datasets in various categories like Sports, Entertainment, Lecture, News, Surveillance and Generic objects videos. Some of these datasets include various challenges like moving camera, moving and cluttered foreground, varying illumination, moving background, different environments like winter and fog.

Section II briefs about the recently available datasets for activity recognition/classification in fields like sports, media, Entertainment, etc. Section III details the dataset available for generic classification of videos. Section IV tabulates the prevailing datasets in activity and generic video classification. Section V concludes this study.

II. ACTIVITY CLASSIFICATION DATASET

KTH [4] and Weizmann [5] are the two benchmarks available in literature, which has long been used in activity recognition/classification/retrieval. Among these, KTH considers different scenarios like outdoor, outdoor scenes at different scales and clothing and indoor scaling as shown in figure 3. Dataset was created using low resolution videos of activities such as walking, jogging, and running. This constrained video dataset was developed in the lab and person under observation act the scripted behavior. The videos were captured with static camera and un-cluttered background and actors too appear without occlusions. Even though techniques proposed in [6], [7] reported good accuracy, these datasets had also been frequently used today in [8] and [9]. Other constrained datasets like IXMAS proposed in [10] include videos captured at different view points. This Dataset also provide depth information for recognition purposes and 13 daily life actions: checking watch, crossing arms, scratching head, sitting down, getting up, turning around, walking, waving, punching, kicking, pointing, picking, overhead throwing and bottom up throwing performed three times by 11 subjects. In order to include pose variation and viewing conditions, this research community have turned to create dataset with TV, sports broadcasts and motion pictures as sources.

Fig. 3. Samples of KTH video dataset.

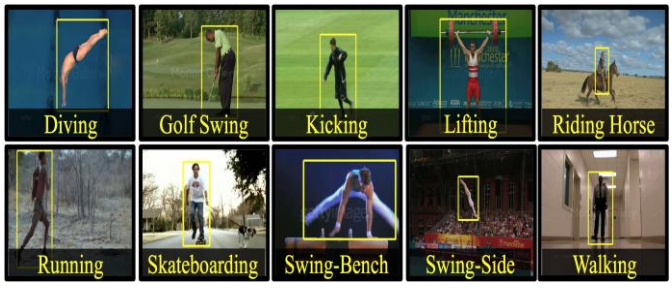

Figure 4. Samples of UCF video dataset.

These benchmarks include illuminations and occlusions thereby significantly raising the bar for action recognition systems. HOHA-1 dataset consists of eight actions collected 32 from movies: answering phone, getting out a car, hand shaking, hugging, kissing, sitting down, sitting up, and standing up [11]. HOHA-2 [12] is an extension of HOHA-1, comprising of 12 actions including HOHA-1 actions, collected from 69 movies: driving car, eating, fighting, and running. High-Five collection [13] is the recently proposed dataset which focuses on interactions, apart from single person activities. Similar to UCF Feature Films Dataset, this dataset include TV and motion pictures. The "Kissing-Slapping" benchmark, provides 90 videos of kissing and 110 of slapping scenes obtained from a range of classic movies. The popular dataset from UCF community is the UCF-Sports benchmark dataset, shown in figure 4 with its 200 videos of nine different sports events: Diving, Golf Swing, Kicking, Lifting, Riding Horse, Running, Skate Boarding, Swing-Bench, Swing-Side and Walking has been collected from TV broadcasts. The Imagelab Laboratory of University of Modena resumed the project of Video Surveillance On-line Repository for Annotation Retrieval (ViSOR). An ontology for annotations of metadata is defined by ViSOR [22].The Human Eva-I [23] dataset covers four gray scale video sequences and three color video sequences from a motion capture system which are calibrated and synchronized with 3D body poses. The database contains 4 subjects covering 6 actions { walking, jogging, gesturing, catching, boxing and combination of walking and jogging. The sequences are with resolution of 640x480 pixels captured at 60 Hz. The Institute of Automations, Chinese Academy of Sciences founded Center for Biometrics and Security Research (CBSR). It was created in 2007. It has eight kinds of human actions with ground truth of AVI file. It can be used for human behavior analysis. i 3D-Post Multi-view dataset [24] was developed in 2009. It contains 13 types of actions performed by 8 people each, hence there are 104 video sequences. Multi Camera Human Action Video data (MuHAVi) [25] was created in 2010. It contains 17 kinds of actions done

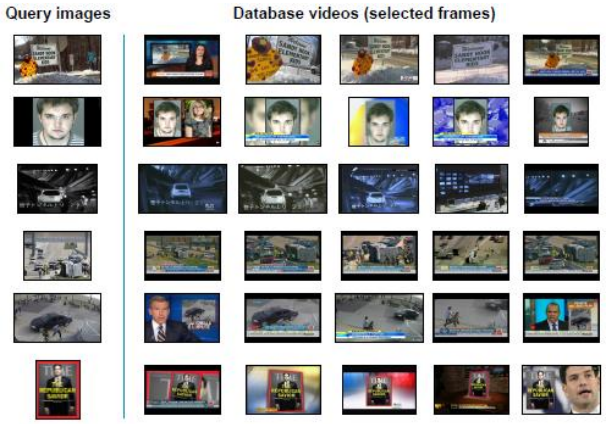

Figure 5. Samples of Stanford I2V video dataset.

Strange action Wrong direction

Abnormal object

Figure 6. Samples of Avenue video dataset.

by 14 persons. It addresses the challenges of camera motion, changing number of persons in scene, cluttered background, variations in view points. It contains 300 video clippings from 20 TV shows handling 4 kinds of interactions. Video Web Dataset [26] was developed in 2010. It contains 2.5 hours video recorded with 4-8 cameras. BEHAVE [27] was created in 2004. It prescreens video sequences to detect unusual or suspected behavior. It has two sets: Optical flow data, Multi agent interaction data.

Evaluation du Traitement et de l'Interpretation de Sequences Video (ETISEO). It was resumed in 2005 and created by INRIA institute [28].University of Illinois at Urbana-Champaign (UIUC) [29] dataset was developed in 2008. It has two sets (i) the first set consists of 14 types of activities done by 8 persons.(ii)the second set has 3 badminton youtube videos.

MSR was built in 2009. It contains 16 video streams with 3 kinds of actions accomplished by 10 people. It addresses the challenges of dynamic and cluttered background .Olympic dataset [29] has variety of sport videos created in 2010. It represents activities as motion segment temporal compositions. It uses Amazon Mechanical Turk for annotation .UTexas released two different datasets (i) UT-Interaction dataset having 6 kinds of 2 person interaction (ii) UT-Tower dataset having 9 classes of actions done by individuals.

III. GENERIC CLASSIFICATION DATASETS

As shown in figure 5, Stanford I2V dataset has been developed based on the idea of searching the video by image query which represent content in video. A single news story is segmented from full length news broadcast and grouped into a video clip in this dataset. This story clip is the collection of successive shots which cover a single event. Hence, each story clip usually contains tens of shots [14]. Similarly, CNN2h dataset [15] can also be used for querying a database by images. The total length of video is 2 hours and 139 image queries with annotated ground truth can be used for query. The annotations also include: i) 2,951 pairs of matching image queries and video frames, and ii) 21,412 pairs of non-matching image queries and video frames. Apart from news and entertainment videos, a large scale dataset namely classX [16] is developed for searching lecture video database. Lecture segments were collected from 21 popular Stanford University courses, to form video clips in this dataset. The courses were offered by 6 different departments:

- Computer Science
- Electrical Engineering
- Management Sciences & Engineering
- Material Sciences,
- Civil & Environmental Engineering
- Computational & Mathematical Engineering

The EVVE video event detection dataset contains 2375 620 videos which were returned to 13 different queries on You tube [17]. The avenue dataset [18] has been developed for detecting abnormality behavior in a scenario like strange action, wrong direction and abnormal object a shown in figure 6. Apart from surveillance videos, generic classes like Basket Ball, Base Ball, Soccer, ice skating, skiing, swimming, biking, cat, dog, bird, graduation, birthday, wedding ceremony, wedding reception, wedding dance, musical performance, non-musical performance, parade, beach, playground has been collected from you tube keyword searches and accumulated into Columbia Consumer Video dataset (CCV) [19]. EDds dataset [20] is focused on two types of human-related events: interactions and activities. In particular, two activities (Hand Up and Walking) and three human-object interactions (Leave, Get and Use object) have been annotated. VIRAT [21] dataset includes large numbers of instances for 23 event types distributed throughout 29 hours of video. Of these datasets TRECVID dataset is the most challenging in the sense it is the unconstrained video dataset taking moving objects, background and varying illumination conditions.

IV. GIST OF BENCHMARKS

This section provides information about datasets like number of classes, details of the classes and the type of videos like sports, news, movie, surveillance, etc. Table 1 illustrates the details of activity classification dataset and Table 2 gives the needed information about generic video classification dataset. Figure 7 shows the performance on recently developed algorithms on UCF101 and HMDFB dataset.

F

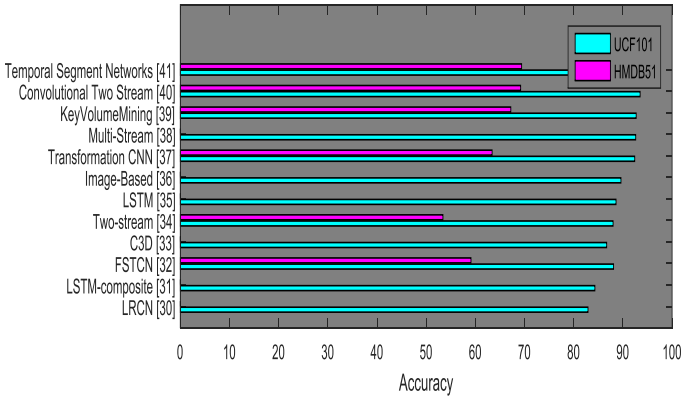

Fig. 7. State of the arts in UCF101 and HMDB1

TABLE I. GIST OF ACTIVITY CLASSIFICATION DATASETS.

S. No	Dataset	Year	Classes	Description	Website
1	KTH	2004	walking, jogging, running, boxing, hand waving, and hand clapping.	600 videos (192 training, 192validation, 216 testing). Resolution = 160x120. Black and white videos. Static camera.	http://www.nada.kth.se/cvap/actions/
2	WEIZMANN	2005	walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack, skip.	90 videos(evaluate by leave one out cross validation). Resolution = 180x144. Static camera.	http://www.wisdom.weizmann.ac.il/%7Evision/SpaceTimeActions.html
3	UCF-Sports	2008	diving, golf swinging, kicking, lifting, horse back riding, running, skating, baseball swinging, walking.	182 videos (evaluate by leave one out cross validation). Resolution = 720x480. Static camera.	crcv.ucf.edu/data/UCF_Sports_Action.php
4	UCF You tube action	2008	101 classes	13320 videos from 101 action categories, UCF101 gives the largest diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions,	crcv.ucf.edu/data/UCF101.php
5	HOHA 1	2008	answer phone, get out car, hand shake, hug person, kiss, sit down, sit up, stand up.	30 videos (219 training, 211 testing). Resolution = 400-300x300-200 (varies between videos). Non-static camera. Shots may happen within clips.	http://www.di.ens.fr/%7Elaptev/download.html
6	HOHA 2	2009	HOHA 1 + driving car, eating, fighting, and running	The training set - 33 movies with 823 samples. The test set - 36 movies other than those used in training with 884 samples	https://www.di.ens.fr/~laptev/actions/
7	IXMAS	2006	checking watch, crossing arms, scratching head, sitting down, getting up, turning around, walking, waving, punching, kicking, pointing, picking, overhead throwing and bottom up throwing	total = 2145 sequences. All sequences captured by 5 calibrated and synchronized cameras	http://4drepository.inrialpes.fr/public/viewgroup/6

TABLE II. GIST OF GENERIC VIDEO CLASSIFICATION DATASETS

S. No	Dataset Name	Description of dataset	Applications	Website
1	Stanford I2V-A News Video dataset	3,800 hours of newscast videos 200 ground-truth queries	video search, image-based retrieval, visual search, news videos	https://purl.stanford.edu/zx935qw7203
2	CNN2h dataset	2 hours of video 139 image queries	Video search by image queries	https://purl.stanford.edu/pj408hq3574
3	ClassX -- A Lecture Video Dataset	21 popular Stanford University courses The queries are 258 clean images of slide	Lecture video retrieval	https://purl.stanford.edu/sf888mq5505
4	The EVVE video Dataset	2375 + 620 videos which were returned to 13 different queries on You tube. Total length=166Hrs	Surveillance	http://pascal.inrialpes.fr/data/evve/
5	Avenue Dataset for Abnormal Event Detection	16 training videos, 21 testing videos	Surveillance	http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html
6	CCV database	9,317 YouTube videos over 20 semantic categories.	Generic	www.ee.columbia.edu/dvmm/CCV/
7	EDds dataset	17 sequences taken using a stationary camera at resolution of 320x240 at 12 fps.	Surveillance	http://www-pu.eps.uam.es/EDds/content.htm
8	VIRAT Dataset	23 event types distributed throughout 29 hours of video.	Surveillance	www.viratdata.org

V. CONCLUSION

A complete review of all datasets available for video classification is beyond reach. This article reviews the benchmarks frequently used by the researchers for evaluating their algorithm for video classification. Based on the application, the available datasets are grouped into activity classification dataset and Generic video classification dataset. A detailed discussion on these categories has also been provided to show pathway for budding researchers.

REFERENCES

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

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

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

[4] 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.

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

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

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

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

[10] 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.

[11] 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.

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

[13] 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.

[14] purl.stanford.edu/zx935qw7203

[15] purl.stanford.edu/pj408hq3574

[16] purl.stanford.edu/sf888mq5505

[17] pascal.inrialpes.fr/data/evve/

[18]http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html

[19] www.ee.columbia.edu/dvmm/CCV/

[20] www-pu.eps.uam.es/EDds/content.htm

[21] www.viratdata.org

[22] www.openvisor.org

[23] http://humaneva.is.tue.mpg.de/

[24] kahlan.eps.surrey.ac.uk/i3dpost_action/

[25] Sanchit Singh, Sergio A Velastin and Hossein Ragheb : A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods, Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance,2 010

[26] www.ee.ucr.edu/~amitrc/datasets.php

[27] groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/

[28] www-sop.inria.fr/orion/ETISEO/

[29] cogcomp.cs.illinois.edu/Data/Car/

[30] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., Long-term recurrent convolutional networks for visual recognition and description. In: CVPR., 2015.

[31] Srivastava, N., Mansimov, E., Salakhutdinov, R., Unsupervised learning of video representations using LSTMs. In: ICML, 2015.

[32] Sun, L., Jia, K., Yeung, D.-Y., Shi, B. E.,. Human action recognition using factorized spatio-temporal convolutional networks. In: CVPR., 2015.

[33] Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., Paluri, M., C3d: Generic features for video analysis. In: ICCV., 2015.

[34] Simonyan, K., Zisserman, A., Two-stream convolutional networks for action recognition in videos. In: NIPS, 2014.

[35] Ng, J. Y.-H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O.,Monga, R., Toderici, G., Beyond short snippets: Deep networks for video classication. In: CVPR., 2015

[36] Zha, S., Luisier, F., Andrews, W., Srivastava, N., Salakhutdinov, R., Exploiting image-trained cnn architectures for unconstrained video classication. *In: BMVC.*, 2015

[37] Wang, X., Farhadi, A., Gupta, A., 2016. Actions ~ transformations. In:CVPR.

[38] Wu, Z., Jiang, Y.-G., Wang, X., Ye, H., Xue, X., 2016. Multi-stream multi-class fusion of deep networks for video classication. In: ACM Multimedia.

[39] Zhu, W., Hu, J., Sun, G., Cao, X., Qiao, Y., 2016. A key volume mining deep framework for action recognition. In: CVPR.

[40] Feichtenhofer, C., Pinz, A., Zisserman, A., 2016. Convolutional twostream network fusion for video action recognition. In: CVPR.

[41] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool,L., 2016. Temporal segment networks: Towards good practices for deep action recognition. In: ECCV.