SICE Annual Conference 2011
September 13-18, 2011, Waseda University, Tokyo, Japan

Action Dataset – A Survey

Md. Atiqur Rahman Ahad, J. Tan, H. Kim, S. Ishikawa

Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Japan
E-mail: atiqahad@yahoo.com

Abstract: Human action understanding and recognition have various demands for different applications in the field of computer vision and human-machine interaction. Due to these issues, more than a decade, extensive researches are going on in this arena – to recognize various actions and activities. Researchers have been exploiting various action datasets and some of them become prominent. Though there are some good datasets, unfortunately, to have a strong survey on these datasets has been a long due. This paper attempts this and presents the key datasets and analyzes them in different perspectives.

Keywords: action, activity, recognition, database, dataset, gesture.

1. INTRODUCTION

Though there are many researches on human actions representation and understanding for more than the last decade, till-to-date, the research progress is not up-to the mark to deal with multi-dimensional actions, social interactions, outdoor regular daily activities, etc. Most of the cases, we see that action recognition and understanding are done for single person and with simple activities. On the other hand, though there are various approaches for action representation, most of these methods are limited to various conditions and constraints [1-2]. Social behavior analysis and understanding are very crucial part for broader real-life applications. However, to this extend, the research progress is not enough. Recent survey papers on action understanding and recognition clarify this statement [1-2]. In this paper, we address various dataset issues and analyze them, which was a long due in the research community. We strongly feel that this paper will be beneficial for the action/gesture recognition community. Note that we have not cover gait analysis [13,18] as gait recognition is considered different than action/activity.

2. ACTION DATASETS

In this section, we present major databases employed and developed in important works. Based on our knowledge, till-to-date, no other work have done similar works on action database survey.

In this Section, we present key and prominent datasets on action, activity and gesture. Most of these are widely exploited in various experimentations by researchers due to their availabilities online as free of cost (except a few cases where one must buy the dataset, e.g., Korea Gesture Dataset; in few cases, one has to submit some documentation for licensing and then have those datasets). Even though some of the papers present good datasets while presenting their methods and performances, due to their unavailability for common use as free materials, those datasets have not been exploited by others and hence remain unknown to the vision community. However, few datasets, though these are simple and have various constraints to analyze, turn into very well-known datasets due to their presence in their websites and research communities are exploiting those to compare their results with prior results published elsewhere. Therefore, to mention a dataset as the top-rated one is highly subjective and it depends on various aspects on what to recognize, how many actions, presence of other subjects in the view, cluttered-ness in the background, movement of camera or background, single- or multi-view, resolution, various articulations in the video clips, interactions, etc.

Another point to mention is here is the presence and usage of standard nomenclature of the terminologies like – gesture, gait, action, activity, interactions, etc. Unfortunately, many papers published in reputed journals and conference proceedings have overlapped these terminologies and used these interchangeably [1]. This issue is a problem for a new or inexperienced researcher in this arena – as which one to call basic or atomic action, how to define activity or action; where the basic distinguishing factor is to name one hand gesture or head gesture to an action or basic action. In some papers, gait is reported as action, head gestures are confused with actions, and vice versa. Gaidon et al. [50] defined a sequence of atomic action units as *actoms*, which are characteristic for the action. Fig. 1 shows one of the taxonomies, where interactions are also considered. Note that still action and activity are covered in one block and level.

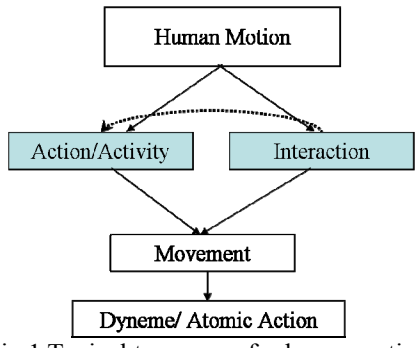

Fig.1 Typical taxonomy for human action.

Most of the cases, mainly optical sensors (e.g., digital video camera) are employed to capture videos. Motion capture, or mocap, is an important new technique for capturing and analyzing human articulations. At present, mocap is widely used to animate computer graphics figures in motion pictures and video games [66]. In some cases, various other sensors, audio data, acceleration, etc. are considered to analyze actions or activities.

KTH Human Action Dataset: The KTH human action dataset [5] has six types of human actions including walking, running, jogging, boxing, hand-waving, and handclapping; each type of action is performed by 25 actors in indoor and outdoor settings. There are 600 video sequences in the dataset. For the common KTH dataset, results are often non-comparable due to the different experimental settings used [8].

Weizmann human action dataset: It [21] consists of 90 low-resolution (180x144, deinterlaced 50 fps) videos of 9 different subjects, each performing 10 natural actions: bend, jumping jack, jump forward, jump in place, run, gallop sideways, skip, walk, wave one hand and wave both hands. This dataset uses a fixed camera setting and a simple background.

ICS Action Database: The ICS Action Database [16] is used for human action recognition and segmentation. The ICS database contains 25 different actions (e.g., getup, lying, lookup, turn, walk, run, standup, sit-down, sitting, etc.) with each action having five trials.

Korea University Gesture Database: The Korea University Gesture (KUG) database [14] is created to assist human tracking and gesture recognition problems, by 20 subjects. It includes 14 normal gestures (such as sitting, walking), 10 abnormal gestures (different forms of falling such as forward, backward, or from a chair), and 30 command gestures (well-defined and commonly used in gesture based studies such as yes, no, pointing, and drawing numbers). However, it is not free like other datasets.

IXMAS dataset: It records in a controlled environment [23] by 11 actors, each performing 3 times for 13 actions: check watch, cross arms, scratch-head, sit-down, get-up, turn-around, walk, wave, punch, kick, point, pick up, and throw. Multiple views of 5 synchronized and calibrated cameras are provided.

WARD (*Wearable Action Recognition Database*): WARD [44] consists of continuous sequences of human actions measured by a network of wearable motion sensors. The wireless sensors are instrumented at five body locations: two wrists, the waist, and two ankles. 13 male and 7 female produce a rich set of 13 action categories that covers some of the most common actions in a human's daily activities, such as standing, sitting, walking, and jumping [45].

CASIA Action Database: It is a collection of sequences of human activities captured by video cameras outdoors from different angle of view [38]. There are 1446 sequences in all containing eight types of actions of single person (walk, run, bend, jump, crouch, faint, wander and punching a car) performed each by 24 subjects and seven types of two person interactions (rob, fight, follow, follow and gather, meet and part, meet and gather, overtake) performed by every 2 subjects [46]. They post the dataset [47] if one wants to use this.

Biological Motion Library: The Biological Motion library [17] is built to analyze and identify features such as gender, identity, and affects. This dataset is structured and evenly distributed across these parameters. There are 15 male and 15 female non-professional actors who perform: walk, arm motions (knocking, throwing, and lifting), and sequences of a walk; as well as their affective styles.

HDM05 Motion Capture Database: The Hochschule der Medien (HDM05) database [15] has more than 70 motion classes in 10~50 realizations executed by five actors. Most of the motion sequences are performed several times as per fixed guidelines.

Cambridge Gesture Dataset: The Cambridge-Gesture dataset [3] contains nine classes of gestures. In total, there are 900 video sequences which are partitioned into five different illumination subsets (Set1, Set2, Set3, Set4, and Set5).

NATOPS Dataset: The Naval Air Training and Operating Procedures Standardization (NATOPS) aircraft handling signals database [26] is a body-and-hand gesture dataset containing an official gesture vocabulary used for communication between carrier deck personnel and Navy pilots (e.g., yes/no signs, taxing signs, fueling signs, etc.). The dataset contains 24 gestures, with each gesture performed by 20 subjects 20 times, resulting in 400 samples per gesture. Each sample has a unique duration. Unlike previous gesture databases, this data requires knowledge about both body and hand in order to distinguish gestures.

Keck Gesture Dataset: The gesture dataset [32] consisting of 14 different gesture classes, is performed by three people, repeated by 3 times, which are a subset of military signals. Hence there are $3 \times 3 \times 14 = 126$ video sequences for training which are captured using a fixed camera with the person viewed against a simple, static background. There are $4 \times 3 \times 14 = 168$ video sequences for testing which are captured from a moving camera and in the presence of background clutter and other moving objects. Another dataset consists of 14 army signaling gestures is developed by [33] where each gesture is performed 5 times by 5 subjects.

UCF Sports Dataset: The UCF sport action dataset [4] consists of ten categories of human actions including swinging on the pommel horse, driving, kicking, lifting weights, running, skateboarding, swinging at the high bar, swinging golf clubs, and walking. The number of videos for each action varies from 6 to 22 and there are 150 video sequences in total. Furthermore, the videos presented in this dataset have non-uniform backgrounds and both the camera and the subject are moving in some actions.

Hollywood2 human action (HOHA) datasets: The Hollywood2 actions dataset [7] has been collected from 69 different Hollywood movies. There are 12 action

classes: answering the phone, driving car, eating, fighting, getting out of the car, hand shaking, hugging, kissing, running, sitting down, sitting up, and standing up. Since action samples in Hollywood2 are collected from movies, they contain many shot boundaries, which cause many artificial interest points.

YouTube dataset: The YouTube dataset [19] contains actions obtained from YouTube, TV broadcast, and personal video collections and are captured under uncontrolled conditions. The videos are of varying resolution, and contain significant variation. There are 11 action categories.

VIRAT Video Dataset (http://www.viratdata.org/): The dataset is designed to be realistic, natural and challenging for video surveillance domains in terms of its resolution, background clutter, diversity in scenes, and human activity/event categories than existing action recognition datasets, with a large number of examples (>30) per action class. Here, both ground camera videos and aerial videos are collected.

MSR Action Dataset [31]*:* The test dataset contains 16 video sequences and has in total 63 actions: 14 hand clapping, 24 hand waving, and 25 boxing, performed by 10 subjects. Each sequence contains multiple 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 clutter and moving backgrounds.

UTexas Database: Ref. [34] started three different types of activity recognition challenges: i. high-level human interaction recognition challenge; ii. aerial view activity classification challenge; and iii. wide-area activity search and recognition challenge, based on their datasets. The general idea behind these three challenges is to test methodologies with realistic surveillance-type videos having multiple actors and pedestrians.

HumanEva-I/ -II Database: The HumanEva database [12] provides ground-truth data to assist in the evaluation of algorithms for pose estimation and tracking of human motion. For every video of a human figure performing some action, there is the corresponding motion capture data of the same performance. The dataset also provides a standard evaluation metric which can be used to evaluate algorithms using the data. There are six actions performed by four subjects in the HumanEva dataset.

CMU Mocap Database: The CMU motion capture database [10] is built mainly to provide a source of motion data for animation and other applications. The database contains 2605 different motion clips of full body mocap data. The actions are performed by a total of 144 subjects (some subjects are the same person). The database has no formal structure such that most sessions have different actions. While one set may have walk actions, other set contains both walk and run actions, and another set contains basketball moves. Even the same action performed across different sets may not have been performed in the same way.

Human Motion Database (HMD): Human Motion Database [9] is developed by controlled sampling methods (parametric and cognitive sampling) to obtain the structure necessary for the quantitative evaluation of several motion-based research problems. It is organized into several components: the *praxicon* dataset, the cross-validation dataset, the generalization dataset, the compositionality dataset, and the interaction dataset. The *praxicon* dataset, a corpus of human motion from a single subject with a wide range of more than 350 commonly performed actions.

Interactive Emotional Dyadic MoCo DB (*IEMOCAP*) *Database:* The SAIL at University of Southern California collected the IEMOCAP database [11] from ten actors in dyadic sessions, for the study of "expressive human communication". The database contains audio and motion capture data of the human face, head, and hands by many markers. The used VICON motion capture system with 8 cameras.

MuHAVi: Multi-camera Human Action Video Data: This dataset has a large body of human action video (MuHAVi) [36] data using 8 cameras, located at 4 sides and 4 corners of a rectangular platform. There are 17 action classes (e.g., walk-turn-back, run-stop, punch, kick, walk-fall, etc.) performed by 14 actors. Note that to prepare training data for action recognition methods, each of the action classes may be broken into at least two primitive actions. For instance, the action 'walk-turn-back' consists of walk and turn back primitive actions.

MAS: Manually Annotated Silhouette Data: From the MuHAVi dataset, selected 5 action classes (i.e., walk-turn-back, run-stop, punch, kick, shot-gun-collapse) are manually annotated into the corresponding image frames to generate the corresponding silhouettes of the actors [36]. This is done to facilitate for the evaluation of action recognition methods.

Virtual Human Action Silhouette (ViHASi) Data: The ViHASi dataset is developed for the evaluation of silhouette-based action recognition methods and the evaluation of silhouette-based pose recovery methods [41-43]. The dataset has 20 different actions by 9 actors and up to 40 synchronized perspective camera views. The perspective of having silhouette-based dataset is that at many processing in lower-level while action recognition, we develop silhouette image or similar. Hence, having this dataset eases the process of recognition and the silhouettes are of good quality [41].

POETICON Enacted Scenario Corpus: Wallraven et al. [22] introduces such a corpus for (inter)action understanding that contains six everyday scenarios taking place in a kitchen / living-room setting. Each scenario was acted out several times by different pairs of actors and contains simple object interactions as well as spoken dialogue.

TMU Kitchen Dataset: It [25] contains recordings of a few (up to four) subjects, who are setting a table according to a pre-defined layout. It contains multi-view camera data, RFID data from a few objects, as well as kinematic data from a marker-less body tracking software.

Kitchen Capture Dataset: It is part of the Carnegie Mellon University Multimodal Activity (CMUMMAC) Database [24]. The dataset is still under development and currently contains data from 25 subjects each of whom was asked to prepare five different recipes in the kitchen. It contains video data from five stationary and one mobile camera, audio data, some inertial motion data from the subjects hand, as well as motion capture data of the full body including the hand.

Assisted Daily Living (ADL) dataset: The ADL dataset [20] consists of high resolution videos of activities performed in daily living. Actions include: answer phone, chop banana, dial phone, look up in directory, write on whiteboard, drink water, eat snack, peel banana, eat banana, and eat with silverware.

i3DPost Multi-view Dataset: In [37], a multi-view/3D human action/interaction database is presented. The database has been created using a convergent eight camera setup to produce high definition multi-view videos, where each video depicts one of eight persons performing one of twelve different human motions. Various types of motions have been recorded, i.e., scenes where one person performs a specific movement, scenes where a person executes different movements in a succession and scenes where two persons interact with each other. Moreover, the subjects have different body sizes, clothing and are of different sex, nationalities, etc. The multi-view videos have been further processed to produce a 3D mesh at each frame describing the respective 3D human body surface. The database is freely available for research purposes. Here, actions are captured in HD-SDI 20-bit 4:2:2 format with 1920 × 1080 resolution at 25Hz progressive scan. Several few other 3D datasets are referred in [23, 38-40].

CHIL 2007 Evaluation Dataset: It provides for the Rich Transcription 2007 Meeting Recognition Evaluation (RT07) in terms of recording setup, scenario, speaker demagogic and transcription process [11]. The corpus consists of 25 interactive seminars recorded at five different recording sites in Europe and USA in multi-sensory smart rooms [11].

DLSBP Dataset: It is introduced by Duchenne et al. [48], consists of two action categories 'OpenDoor' and 'SitDown'. The training set includes 38 'OpenDoor' and 51 'SitDown' actions sequences obtained from 15 movies.

PETS has sequences of datasets, e.g., PETS2006 benchmark data-sets are multi-sensor sequences containing left-luggage scenarios with increasing scene complexity; PETS2007 contains the following 3 scenarios, with increasing scene complexity: loitering, attended luggage removal (theft), unattended luggage. Based on these datasets, workshops were organized [49] like HumanEva.

Another recent benchmark for temporal action detection is called 'Coffee and Cigarettes' [50]. It consists of a single movie composed of 11 short stories, each with different scenes and actors [51].

Visual Geometry Gr. opens up several datasets in [35], e.g., TV human interactions dataset. Yilmaz and Shah [27] develop a dataset having 18 Sequences, 8 Actions, i.e., 3 x Running, 3 x Bicycling, 3 x Sitting-down, 2 x Walking, 2 x Picking-up, 1 x Waving hands, 1 x Forehand stroke, 1 x Backhand stroke.

Another DB from M. Shah et al. [28] has 6 Actions: Sitting, Standing, Falling, Walking, Dancing, Running. Ref. [29] develops dataset on Walking, Diving, Jumping, Waving arms, Waving hands, Ballet figure, Water fountain. Ref. [30] has 25 x Walk, 6 x Run, 18 x Sit-down.

CAVIAR project develops a number of different sequences with ground truth which can be employed for recognition as well as tracking [53]. Similar to this project, ViSOR (Video Surveillance Online Repository) produces different sets of videos for various applications, along with a dataset called 'videos of different human actions' [54]. This action dataset contains 40 items.

Though for gait analysis and recognition, a number of good datasets are already developed (e.g., [55-57]), we are not covering gait datasets here. The RVL-SLLL American Sign Language (ASL) Database is developed for sign language of various single signs, alphabets, numbers and examples of short discourses [58]. Similar to the gait dataset, sign language datasets are not covered here.

There are several synthetically-generated rendered computer graphics human surface models or datasets as well [59-62] by employing POSER or MAYA. From the captured video, motion capture data are extracted to produce re-rendered poses for recognition, e.g., in Ref. [60]. Another database [61] consists of about 3000 samples that involve a variety of human activities including walking, running, turns, gestures in conversations, quarreling and pantomime.

3. DISCUSSIONS

A systematically constructed gesture database in carefully controlled environment is essential due to the fact that it can aid to observe or analyze the characteristics of human motion and verify or evaluate the developed algorithm and its application system.

Few datasets consider mainly upper body gestures, whereas, in some case, lower-body movements are systematically collected, mainly for gait recognition and analysis. Despite clear advances in the field of action recognition and understand, evaluation of these methods remains mostly heuristic and qualitative [14], therefore, making it difficult to evaluate the current state-of-the-art with any certainty or even to compare different methods with any rigor. For example, it has been found that local descriptors more suitable for KTH, while the holistic features are more suitable for Weizmann [45] due to the different characteristics of these two databases. Sun et al. [63] summarizes various recognition results for KTH and Weizmann datasets. Similarly, many of the datasets do not include ground-truth, detailed pose information so that researchers can think and consider beyond the conventional approaches in analyzing actions (e.g., developing mathematical foundation for computer

vision based on new approaches or existing physics-based approaches). Sigal et al. [12] define a standard set of error measures to evaluate their action datasets. More efforts on this similar track are required.

As covered above, datasets of human motion have been widely used for recognizing human motion and synthesizing humanoid motions. It is also necessary to develop smart data structure approaches for storing and extracting human motion data and demonstrate that the database can be applied to the recognition and motion synthesis problems in robotics, humanoid robots and other applications [64]. The last decade brought major advances in our understanding of the cerebral structures involved in processing the how, what, and why of other people's actions [65]. Datasets that can somehow relate to research on neuron system is a due now. It is necessary to investigate various scaling aspects for large motion capture databases as well as reconstructing multiple people and handling occlusions and different observation representations. Apart from these issues, developing more robust action datasets at night or dark (by employing IR imaging), from far distance by manipulating PNZ cameras, abnormal activities understanding in cluttered outdoor scene, datasets with multi-modes and interactions are essential to address. It is important to develop datasets that can blend the photographic analysis of human movements (as above) and neural recording of human brains – so that we can understand and define human activity in robust manners. We also need to develop datasets with people having various disabilities of movements due to partial paralyzed body-parts or other problems, so that by analyzing these datasets, we can work and understand rehabilitation and medical treatment issues.

4. CONCLUSIONS

In this paper, we survey action/activity datasets for the computer vision community. Due to space constraints, we could not categorize these datasets based on various parameters; employing Tables and Figures to make the paper more analytical. We hope that this paper will be helpful for the researchers who are or will work in human action, activity, gesture recognition and understanding. Future work will concentrate on further development and detailed analysis of the survey.

REFERENCES

[1] Md. Atiqur Rahman Ahad, J. Tan, H. Kim, and S. Ishikawa, "Motion history image: its variants and applications", Machine Vision and Applications, pp. 1-27, 2010. DOI: 10.1007/s00138-010-0298-4

[2] Y Lui and J. Beveridge, "Tangent bundle for human action recognition", IEE AFGR, 2011.

[3] T.-K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection", IEEE Trans. PAMI, Vol. 31, No. 8, pp. 1415–1428, 2009. The database is publicly available at http://mi.eng.cam.ac.uk/~tkk22

[4] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach: A spatiotemporal maximum average correlation height filter for action recognition", IEEE CVPR, 2008. Available at http://www.cs.ucf.edu/vision/public_html/

[5] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach", ICPR, 2004. Available at http://www.nada.kth.se/cvap/actions/

[6] T.-K. Kim and R. Cipolla, "Gesture recognition under small sample size", ACCV, 2007.

[7] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context", IEEE CVPR, 2009. Available at http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

[8] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition", BMVC, 2009.

[9] G. G.-Filho and A Biswas, "The human motion database: A cognitive and parametric sampling of human motion", IEEE AFGR 2011.

[10] CMU Mocap Database, 2001. http://mocap.cs.cmu.edu

[11] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database", Language Resources and Evaluation, Vol. 42, No. 4, pp. 335-359, 2008.

[12] L. Sigal, A. Balan and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion", International Journal of Computer Vision, Vol. 87, No. 1-2, 2010.

[13] R. Gross and J. Shi, "The CMU motion of body MOBO database", Technical report, Carnegie Mellon Univ., 2001.

[14] B.-W. Hwang, S. Kim, and S.-W. Lee, "A full-body gesture database for automatic gesture recognition," IEEE AFGR, pp. 243-248, 2006.

[15] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation: Mocap database HDM05", Technical report CG-2007-2, Universität Bonn, 2007.

[16] ICS Action Database, The University of Tokyo, 2003-2009. http://www.ics.t.u-tokyo.ac.jp/action/

[17] Y. Ma, H. Paterson, and F. Pollick, "A motion capture library for the study of identity, gender, and emotion perception from biological motion", Behavior Research Methods, Vol. 38, pp. 134-141, 2006.

[18] Georgia Tech Human Identification at Distance Database. http://www.cc.gatech.edu/cpl/projects/hid/

[19] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," IEEE CVPR, 2009.

[20] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints", ICCV, 2009.

[21] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes", IEEE Trans. PAMI, Vol. 29, No. 12, pp. 2247-2253, 2007.

[22] C. Wallraven, M. Schultze, B. Mohler, A. Vatakis, K. Pastra, "The POETICON enacted scenario corpus - a tool for human and computational experiments on action understanding", IEEE AFGR 2011. Available in: http://poeticoncorpus.kyb.mpg.de

[23] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes", Computer Vision and Image Understanding, Vol. 104, No. 2-3, pp. 249-257, 2006.

[24] F. Hodgins and J. Macey, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database", CMU-RI-TR-08-22, Jan. 2009.

[25] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition", *IEEE Int. Workshop on*

Tracking Humans for the Evaluation of their Motion in Image Sequences with ICCV, 2009.
[26] Y. Song, D. Demirdjian, and R. Davis, "Tracking Body and Hands For Gesture Recognition: NATOPS Aircraft Handling Signals Database", *IEEE AFGR*, 2011.
[27] A. Yilmaz and M. Shah, "Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras", *ICCV* 2005.
[28] Y. Sheikh and M. Shah, "Exploring the Space of an Action for Human Action Recognition", *ICCV*, 2005.
[29] E. Shechtman and M. Irani, "Space-Time behavior-based correlation - OR - how to tell if two underlying motion fields are similar without computing them?", *IEEE Trans. PAMI*, Vol. 29, No. 11, 2007.
[30] V. Parameswaran and R. Chellappa, "View invariants for human action recognition", *IEEE CVPR*, 2003.
[31] J. Yuan, Zicheng Liu and Ying Wu, "Discriminative subvolume search for efficient action detection", *IEEE CVPR*, 2009.
[32] Z. Lin, Z. Jiang, and L. Davis, "Recognizing actions by shape-motion prototype trees," *ICCV*, pp. 444-451, 2009.
[33] "http://www.umiacs.umd.edu/~shivnaga/supplmat_ActionRecBallisticDyn_CVPR08/action_rec_using_ballistic_dynamics.html#gesture_rec"
[34] "http://cvrc.ece.utexas.edu/SDHA2010/"
[35] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid, "High five: Recognising human interactions in TV shows", *BMVC*, 2010. "http://www.robots.ox.ac.uk/~vgg/data/"
[36] MuHAVi and MAS: "http://dipersec.king.ac.uk/MuHAVi-MAS/"
[37] N. Gkalelis, H. Kim, A. Hilton, N Nikolaidis and I. Pitas, "The i3DPost multi-view and 3D human action/interaction database", *Conf. on Visual Media Production*, 2009.
[38] Y. Wang, K. Huang, and T. Tan, "Human activity recognition based on R transform", *IEEE CVPR*, 2007.
[39] A. Ogale, A. Karapurkar, and Y. Aloimonos, "Viewinvariant modeling and recognition of human actions using grammars", *Workshop on Dynamical Vision with ICCV*, 2005.
[40] F. Huang and G. Xu, "Viewpoint insensitive action recognition using envelop shape", *ACCV*, pp. 477-486, 2007.
[41] Virtual Human Action Silhouette (ViHASi) Database. "http://dipersec.king.ac.uk/VIHASI/"
[42] C. Orrite, F. Martınez, E. Herrero, H. Ragheb, S. Velastin, "Independent viewpoint silhouette-based human action modelling and recognition", *Int. Workshop on Machine Learning for Vision-based Motion Analysis* (MLVMA'08) with ECCV, 2008.
[43] H. Ragheb, S. Velastin, P. Remagnino and T. Ellis, "ViHASi: Virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods", *Workshop on Activity Monitoring by Multi-Camera Surveillance Systems*, 2008.
[44] A. Yang, R. Jarafi, P. Kuryloski, S. Iyengar, S. Sastry, and R. Bajcsy, "Distributed segmentation and classification of human actions using a wearable motion sensor network", *Workshop on Human Communicative Behavior Analysis* with *CVPR*, 2008.
[45] "http://www.eecs.berkeley.edu/~yang/software/WAR/"
[46] Z. Zhang, K. Huang, and T. Tan, "Multi-thread parsing for recognizing complex events in videos", *ECCV*, 2008.
[47] CASIA Action DB. Available in: http://www.cbsr.ia.ac.cn/english/Action%20Databases%20EN.asp
[48] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce, "Automatic annotation of human actions in video", *ICCV*, 2009.
[49] *Workshops on Performance Evaluation of Tracking & Surveillance* (PETS)
"http://www.cvg.rdg.ac.uk/PETS2007/data.html"
[50] A. Gaidon, Z. Harchaoui, and C. Schmid, "Actom sequence models for efficient action detection", *IEEE CVPR*, 2011.
[51] I. Laptev and P. Perez, "Retrieving actions in movies", *ICCV*, 2007.
[52] S. Burger, "The CHIL RT07 Evaluation Data", *Multimodal Technologies for Perception of Humans*, 2008.
[53] CAVIAR test case scenarios: "http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/"
[54] "http://imagelab.ing.unimore.it/visor/video_categories.asp"
[55] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition", *ICPR*, pp. 441-444, 2006.
[56] S. Sarkar, P. Phillips, Z. Liu, I. Vega, P. Grother, and K. Bowyer, "The humanid gait challenge problem: data sets, performance, and analysis", *IEEE Trans. PAMI*, Vol. 27, No. 2, pp. 162-177, 2005.
[57] "http://www.cc.gatech.edu/cpl/projects/hid/Description.html"
[58] R. Wilbur and A. Kak, "Purdue RVL-SLLL american sign language database", School of Electrical and Computer Engineering Technical Report, TR-06-12, Purdue University, 2006.
[59] A. Agarwal and B. Triggs, "3D human pose from silhouettes by relevance vector regression", *IEEE CVPR*, pp. 882-888, 2004.
[60] K. Grauman, G. Shakhnarovich and T. Darrell, "Inferring 3D structure with a statistical image-based shape model", *ICCV*, pp. 641-648, 2003.
[61] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Discriminative density propagation for 3D human motion estimation", *IEEE CVPR*, pp. 390-397, 2005.
[62] C. Tomasi, S. Petrov, and A. Sastry, "3D tracking = classification + interpolation", *ICCV*, 2003.
[63] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features", *CVPR4HB*, 2009.
[64] K. Yamane and Y. Nakamura, "Human motion database with a binary tree and node transition graphs", *J. of Autonomous Robots*, Vol. 29, No. 2, 2010.
[65] M. Thioux, V. Gazzola, and C. Keyesers, "Action understanding: how, what and why", *Current Biology*, Vol. 18, No. 10, pp. R431-R434, 2008.
[66] G. Liu, J. Zhang, W. Wang, and L. McMillan, "A system for analyzing and indexing human-motion databases", *SIGMOD* 2005.