Computer Vision and Image Understanding 117 (2013) 633–659

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A survey of video datasets for human action and activity recognition ☆

Jose M. Chaquet [a],*, Enrique J. Carmona [a], Antonio Fernández-Caballero [b]

[a] Dpto. de Inteligencia Artificial, Escuela Técnica Superior de Ingeniería Informática, Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain
[b] Instituto de Investigación en Informática de Albacete, Universidad de Castilla-La Mancha, 02071 Albacete, Spain

ARTICLE INFO

Article history:
Received 2 March 2012
Accepted 30 January 2013
Available online 13 February 2013

Keywords:
Human action recognition
Human activity recognition
Database
Dataset
Review
Survey

ABSTRACT

Vision-based human action and activity recognition has an increasing importance among the computer vision community with applications to visual surveillance, video retrieval and human–computer interaction. In recent years, more and more datasets dedicated to human action and activity recognition have been created. The use of these datasets allows us to compare different recognition systems with the same input data. The survey introduced in this paper tries to cover the lack of a complete description of the most important public datasets for video-based human activity and action recognition and to guide researchers in the election of the most suitable dataset for benchmarking their algorithms.

1. Introduction

Human activity and action recognition systems aim to identify the actions and goals of one or more agents from a series of observations on each agent and a given context. An increasing interest in this type of systems has been reported so far [9,67,99,171,198,199,227]. Action recognition is one of the keys of several applications such as visual surveillance [76,80,93,97,169], video retrieval [63] and human–computer interaction [88], among others. Recognition of human activities can be considered as the last step of a set of previous tasks, such as image capture, segmentation, tracking, identification, and classification. Other surveys closely related to the action and activity recognition, such as motion analysis [8,10,36,89,170,224], understanding dynamic scene activity [32], understanding human behaviour [162], classifying human actions [176] or human motion capture [132,133] are also available.

Although, in recent years, more and more video datasets dedicated to human action and activity recognition have been created, currently there is not a survey in this field. In fact, to our knowledge, there is only a short paper in the literature devoted to this subject [11]. Therefore, this survey tries to cover the lack of a complete description of the most important public datasets suited for video-based human action and activity recognition. The use of publicly available datasets has two main advantages. On the one hand, they save time and resources, that is, there is no need to record new video-sequences or pay for them, so researchers can focus on their particular algorithms and implementations. On the other hand, and this is even more important, the use of the same datasets facilitates the comparison of different approaches and gives insight into the abilities of the different methods. This survey is mainly focused on the video datasets that are composed by heterogeneous action sets, i.e., typical actions that can appear in a variety of situations or scenarios and are recorded by visible spectrum cameras. Nonetheless, there are some databases created for very specific action recognition, such as detection of abandoned objects, recognition of activities of daily living (ADL), crowd behaviour, detection of human falls, gait analysis, or pose and gesture recognition. These datasets will be also described here, but in a very short fashion, due to space limitations. In Fig. 1, the taxonomy adopted in this work is shown. There are three categories. Two of them are related to the type of actions provided by the dataset: *Heterogeneous* and *Specific Actions*. A third category, called *Others*, is defined according to the specific techniques to capture the actions: Infrared and thermal, and Motion Capture (MOCAP).

An action can be considered like a sequence of primitive actions that fulfil a function or simple purpose, such as jumping, walking, or kicking a ball. On the other hand, an activity is composed of sequences of actions over space and time, such as a person preparing a dish following the steps from a recipe, or people playing football. One additional feature of activities is that they are normally related to the concept of interaction: between a person with one or more people, or between one or more people with objects of the surrounding environment. However, differences between actions and activities are not always clear. For instance, the running of a person from one place to another can be considered an action or,

☆ This paper has been recommended for acceptance by J.K. Aggarwal.
* Corresponding author.
E-mail addresses: jose.chaquet@gmail.com (J.M. Chaquet), ecarmona@dia.uned.es (E.J. Carmona), antonio.fdez@uclm.es (A. Fernández-Caballero).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.01.013

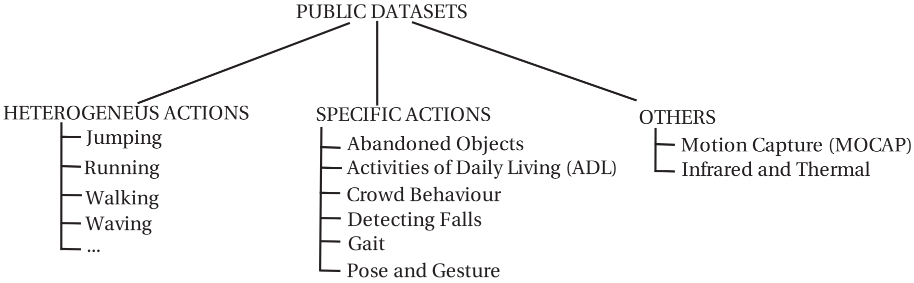


Fig. 1. A possible taxonomy of the datasets according to the type of actions.

by the contrary, could be considered an activity if the action is seen in a particular context, such as a person who is running away from a possible risk. That is why a lot of the datasets described along this paper do not make distinction between actions and activities. In a generic way, it is said that these datasets store action sets.

The rest of the paper is organized as follows. In Section 2, the main characteristics of 28 public video datasets for heterogeneous action or activity recognition are provided. In Section 3, the datasets described in the previous section are compared from different points of view. In Section 4 some brief comments about specific action datasets and other datasets related to action recognition are provided. Finally, Section 5 presents the conclusion of this survey.

2. Video datasets for action recognition

In this section, the main public video datasets for human action and activity recognition are described. The common feature of each of the datasets included here is that all of them are composed by a large and varied repertoire of different actions or activities that can be applied to different contexts or situations.

The chronologic appearance order of the different human action video datasets runs parallel to the challenges that the scientific community has been considering to face the problem of automatic and visual recognition of human activities and actions in video images. Thus the first challenge was to analyse single-human and single-action and to this end, the first action video datasets were created: Weizmann (2001& 2005) and KTH (2004). They are the datasets more known and used, but however the main inconvenience of them is that they were recorded in controlled conditions, that is, neither of these datasets are representative of human actions in a real world. Here an individual actor performs an action for video clip and each action is performed in a near identical fashion, from a fixed point of view and against a simple and static background. The KTH adds more complexity by varying clothing and lighting, but it is still unrealistic.

On the other hand, in real problems, more complex situations are managed. That is why soon appeared new datasets with video clips recorded in more realistic conditions or gathered directly from web. Datasets like CAVIAR (2004), ETISEO (2005), CASIA Action (2007), MSR Action (2009) and UT-Tower (2010) belong to the first group. Here, illumination conditions are not controlled (outdoors) and backgrounds are complex and not static. Representative examples of the second group are HOLLYWOOD (2008), UCF Sports (2008), UCF *YouTube* (2009), UCF50 (2010), Olympic Sports (2010) and HMDB51 (2011), where the most of their videos were compiled from *YouTube*. Another typical feature of realistic situations is to consider human–human or object–human interactions. Although several of the datasets above mentioned (CAVIAR, ETISEO, CASIA Action, HOLLYWOOD and HMDB51) already contain actions with person-to-person interactions, other datasets have been created specifically to study this issue: BEHAVE (2004), TV Human Interaction (2010) and UT-Interaction (2010).

So far, we have focused on datasets created to visual analysis of behaviour from a single observational viewpoint. However, recently, the scientific community has also been interested in to address the problem of understanding and recognising human behaviours in realistic conditions too, but analyzed from multiples viewpoints. Thus, for example, it has become increasingly necessary to use networks of multiple cameras for monitoring large public spaces such as shopping malls, airports, and train or subway stations. Several video datasets have been created specifically for studying the problems related to this context: IXMAS (2006), i3DPost Multi-view (2009), MuHAVi (2010), VideoWeb (2010) and CASIA Action. The first three were recorded in controlled conditions (indoor) while the last two in real conditions (outdoor). In addition, some of the aforementioned datasets also contain some multi-view sub-datasets. This is the case of BEHAVE, CAVIAR and ETISEO.

Beside of the above mentioned datasets, it has been created other type of datasets that are authentic repositories of large quantities of video, containing thousands of hours of footage. We are speaking of VISOR (2005) and VIRAT (2011), the latter of very recent creation. In this type of datasets, we can find a wide repertory of actions performed by single person or interactions person-to-person, person-to-vehicle, person-to-object, or person-to-facility. The above chronological description is summarized in Table 1.

In the following sections the description of each dataset is made in a systematic way, attending to a collection of common features, namely, dataset identification, original goal, example video frames, context, ground truth, and reference papers. The dataset identification includes the following information: information related to the institution (congress, university, research center, etc.) which built the dataset, the country, the year of creation, the web site for downloading and the descriptive paper (if any). This information is followed by the description of goals (tasks, subtasks, application domains) for which the dataset was built. A figure showing several example video frames is also included. The context makes reference, for example, to the information related to the set of actions compiled, the type of sceneries, the number of actors or the number of stored videos. The ground truth includes information about what type of knowledge on each dataset video is available, such as segmented silhouettes, bounding boxes, annotations about physical objects or list of events that appears or happens in the scene. Finally, a list of example papers using the dataset is also included. As can be seen in Section 3, this characterization is extended with other more technical or specific types of features, and lastly the complete information is summarized and compiled in a table. This allows a faster and direct comparison of all the datasets mentioned in this paper. Notice that a great number of references cited in this work are URLs. Because web addresses are not always permanent and can be modified, all the cited links have been trusted by the

Table 1
Historical development of the most important public action datasets.

COMPLEXITY	TYPE OF PROBLEM	SOURCE	DATASET	AGE
—	UNREALISTIC ACTION ANALISYS (Simple and static background)	Recorded videos (indoor/outdoor)	Weizmann (2001&2005) KTH (2004)	+
	REALISTIC ACTION ANALISYS (Complex and not static background and illumination conditions not controlled)	Recorded videos (indoor/outdoor)	CAVIAR (2004) ETISEO (2005) CASIA Action (2007) MSR Action (2009) UT Tower (2010)	
⇓		Videos from web (indoor/outdoor)	HOLLYWOOD (2008&2009) UCF Sports (2008) UCF *YouTube* (2009) UCF 50 (2010) Olympic Sports (2010) HMDB51 (2011)	⇑
	INTERACTION ANALYSIS	Videos from recording&TV shows	BEHAVE (2004) TV Human Interaction (2010) UT Interaction (2010)	
+	MULTIVIEW ANALYSIS	Recorded video (indoor)	IXMAS (2006) i3DPost Multi-view (2009) MuHAVi (2010)	—
		Recorded videos (outdoor)	CASIA Action (2007) VideoWeb (2010)	
	REPOSITORIES	Videos from different sources	ViSOR (2005) VIRAT (2011)	

authors of this work and the date of the last access is provided. The datasets will be presented in chronological order.

2.1. WEIZMANN datasets

The Weizmann Institute of Science (Faculty of Mathematics and Computer Science, Israel) provides two datasets. They are described below.

2.1.1. Weizmann event-based analysis

The Weizmman Event-Based Analysis dataset [236], recorded in 2001, was one of the first datsets created, and was recorded for studying algorithms for clustering and temporal segmentation of videos using some statistical measures, in contrast to other common approaches at that time where parametrical models were used. The main intention of the authors using non-parametric measures was to handle a wide range of dynamic events without prior knowledge of the types of events, their models, or their temporal extent. The dataset is formed by a unique long sequence of around 6000 frames, displaying different people, wearing different clothes, and performing four activities: running in place, waving, running, and walking. The only ground truth provided is the action annotation for each frame. In Fig. 2, some frames are shown. One example work using this dataset applied to automatic temporal segmentation can be consulted in [235].

2.1.2. Weizmann actions as space-time shapes

The Weizmann Actions as Space-Time Shapes dataset [69] was recorded in 2005 with the intention of studying new algorithms that improved the human action recognition systems at that time: optical flow was difficult to apply, other approaches were based on feature tracking which could not deal properly with self-occlusions, other works could only use periodic actions, etc. On the contrary, this dataset was applied to algorithms based on space-time shape volumes. Therefore, the background is relatively simple and only one person is acting in each frame. It contains 10 human actions (walking, running, jumping, galloping sideways, bending, one-hand waving, two-hands waving, jumping in place, jumping jack, and skipping), each performed by nine people (see Fig. 3). A detailed description of the dataset can be found in [68]. The backgrounds are static and the foreground silhouettes (in MATLAB format) of each moving person and the background sequences used for background subtraction are included in the dataset as ground truth. The view-point is static. In addition to this dataset, two separate sets of sequences are recorded for robustness evaluation. One set shows walking movement viewed from different angles. The second set shows front-parallel walking actions with slight variations (carrying objects, with different clothing, or with different styles). Examples of works using this dataset applied to action recognition are [30,68,174,182,220,230].

2.2. BEHAVE: computer-assisted prescreening of video streams for unusual activities

The BEHAVE project [54] started in 2004 under the coordination of the School of Informatics of Edinburgh University. The project investigated two novel computer-based image analysis processes to prescreen video sequences for abnormal or crime-oriented behavior. The objectives of the project were: (1) to investigate and extend methods for classifying the interaction among multiple individuals, being capable of discriminating between subtly different behaviors; (2) to develop methods for flow-based analysis of the behavior of many interacting individuals; (3) to apply the results of these two approaches to the detection of criminal or dangerous situations in interactions between small groups and crowd situations; (4) to filter out image sequences where uninteresting normal activity is occurring.

The public downloaded dataset is composed of two sets: *Optical Flow Data* and *Multiagent Interaction Data*. The first one is composed of optical flow sequences from the Waverly train station.

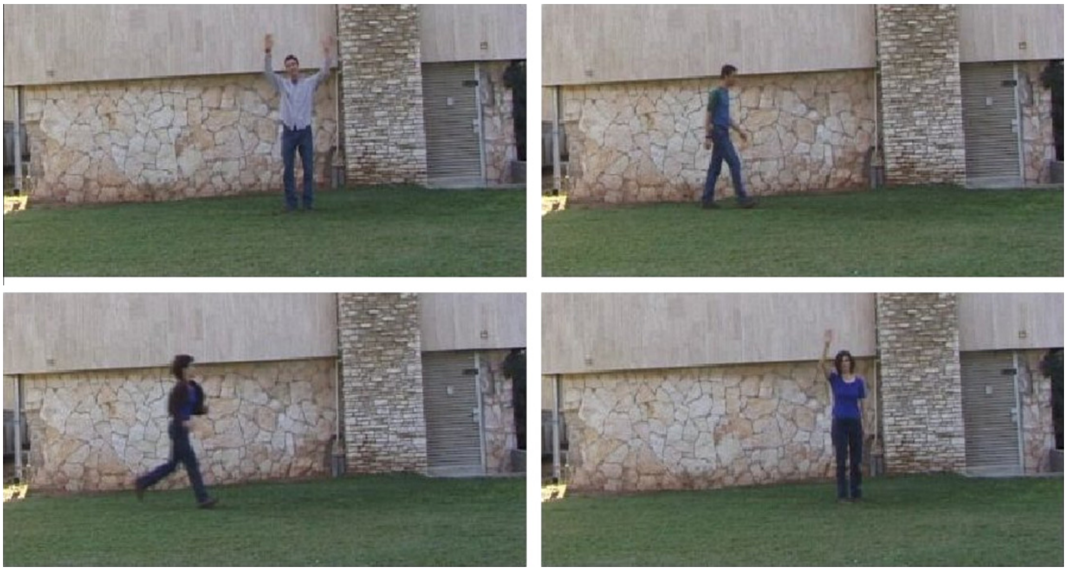


Fig. 2. Four example frames from Weizmann event dataset.

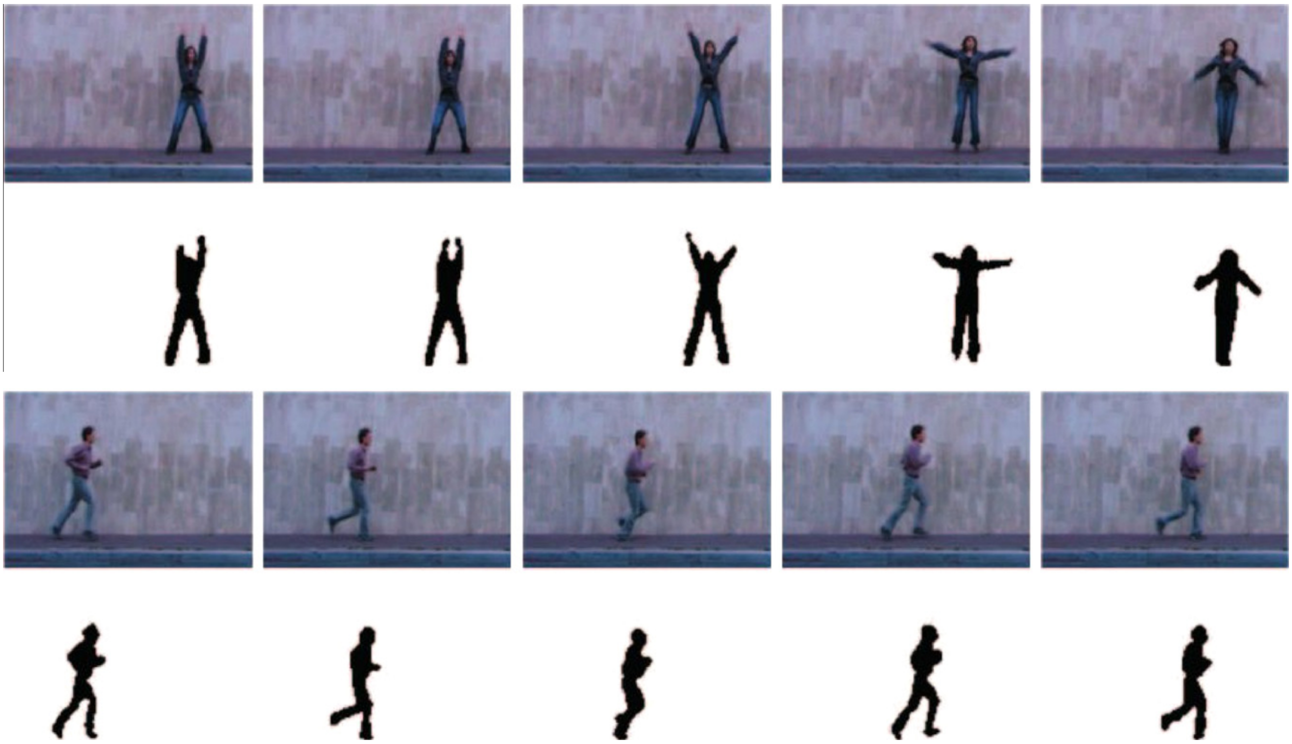


Fig. 3. Examples of video sequences and extracted silhouettes from Weizmann action database.

There are 30 files divided into the following groups: Group 1 (normal-training), Group 2 (normal-testing) and Group 3 (emergency-blocked exit at the bottom of the scene). There is also a text file containing labels and marks that characterize the interactions.

The second dataset comprises two views of various scenarios where people interact. The ground truth contains the coordinates of the bounding boxes of the pedestrians for almost all video sequences and it is described in VIPER XML [104], a video description format. Ten basic scenarios are used: *InGroup, Approach, WalkTogether, Split, Ignore, Following, Chase, Fight, RunTogether* and *Meet* (see Fig. 4). The ground plane homography can be computed using the available data. The BEHAVE project has published several works using this dataset with a focus on crowd analysis, such as event detection [16,17,20], simulation and modelling of crowd problems [15,20], optical flow anomalies [18,19] and multiagent activities recognition [27].

2.3. CAVIAR: Context Aware Vision using Image-based Active Recognition

The main objective of the CAVIAR project [1], (2002–2005), was to address the following question: Can rich local image descriptions and other image sensors, selected by a hierarchal visual attention process and guided and processed using task, scene,

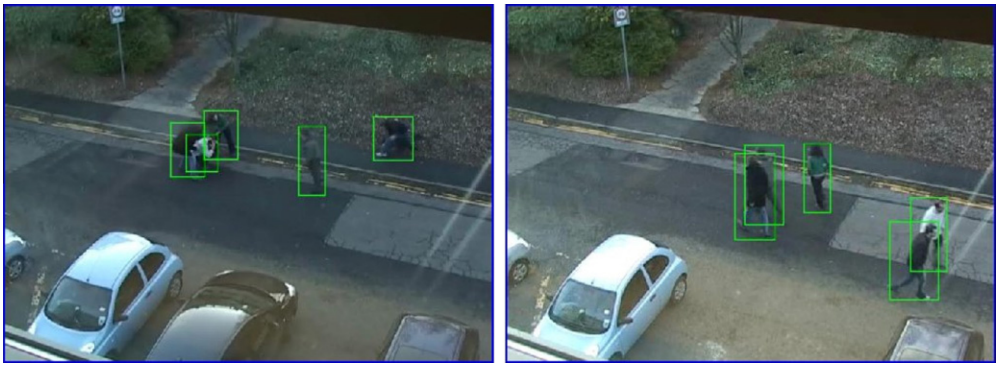


Fig. 4. An example of the BEHAVE dataset (Multiagent Interaction Data) showing two frames with the bounding boxes of several people.

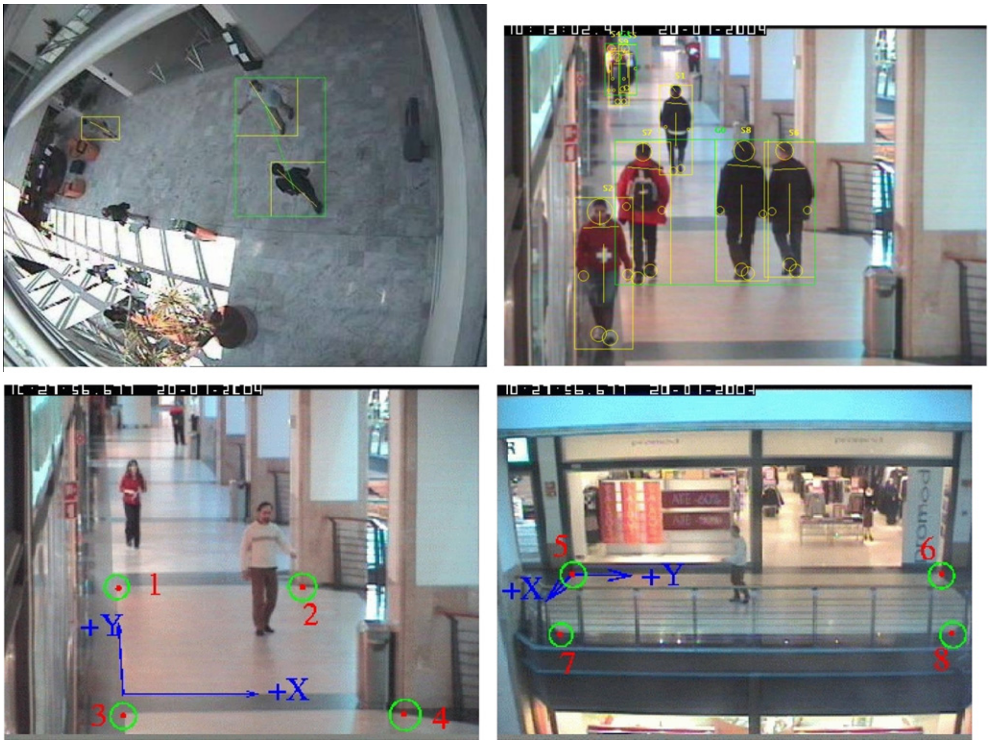


Fig. 5. Examples of typical frames in CAVIAR dataset: frame recorded at INRIA Labs (left top), example of a marked up frame with identification heads, gaze, hands, feet and shoulders recorded at the hallway in a shopping center in Lisbon (right top) and examples of two frames synchronized along (left bottom) and across (left right) the same corridor, showing ground plane homography data.

function and object contextual knowledge, improve image-based recognition processes?

Among other activities of that project, a video dataset was created. The CAVIAR dataset includes people performing 9 activities: walking, browsing, slump, left object, meeting, fighting, window shop, shop entering, and shop exiting. The videos are recorded in two different places. The first section of video clips is filmed with a wide angle camera lens in the entrance lobby of the INRIA Labs at Grenoble, France (see Fig. 5, left top). The second set also uses a wide angle lens along and across the hallway in a shopping center in Lisbon. In this case, there are two timely synchronized videos for each sequence, one with the view across (see Fig. 5, right bottom) and the other along the hallway (as shown in Fig. 5, left bottom). For each scenario, a ground truth file expressed in CVML [117] (an XML-based computer vision markup language) was constructed. The file contains the coordinates of a bounding box, an activity label (appear, disappear, occluded, inactive, active, walking, running), and a scenario label (fighter role, browser role, left victim role, leaving group role, walker role, left object role) for each individual (see Fig. 5, right top). Also, for each frame, a situation label (moving, inactive, browsing) and a scenario label (browsing, immobile, walking, drop down) are provided. Further information about the ground truth can be obtained in [55,120]. The CAVIAR project has produced a big amount of publications dedicated to different applications, such as target detectors [78], activity recognition [26,118,119,139,168,173,200], human activity monitoring [33], clustering of trajectories [7], human activities identification [49], motion segmentation [140,141], tracking [77,91] or multi-agent activity recognition [28]. At the PETS-ECCV-04 workshop web page [5], some example papers using CAVIAR dataset, and dedicated to activity recognition, can be also consulted.

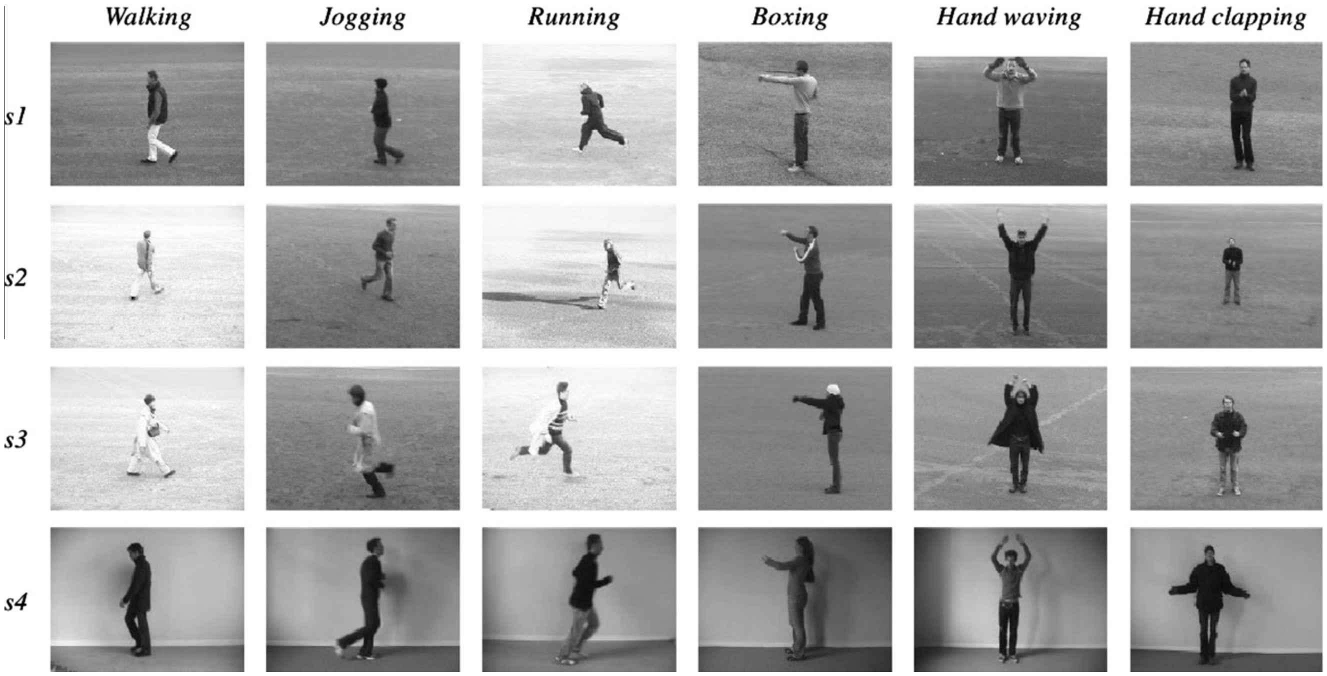


Fig. 6. Examples of sequences corresponding to different types of actions and scenarios from KTH database. The four different scenarios are outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4).

2.4. KTH recognition of human actions

The KTH Royal Institute of Technology created this dataset [108] in 2004 achieving an important milestone in the computer vision community. At that time it became the largest video database with sequences of human actions taken over different scenarios. It allowed a systematic comparison of different algorithms using the same input data. Nevertheless, all the sequences were taken over homogeneous background with a static camera.

This dataset contains six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed several times by 25 people in four different scenarios. A useful description of this dataset is performed in [180]. Fig. 6 shows several examples frames corresponding to different types of actions and scenarios. There are a total of 25 × 6 × 4 = 600 video files for each combination of 25 individuals, 6 actions, and 4 scenarios. The ground truth data is provided in an ASCII file giving for each frame the action performed. Each action is characterized by the actor, action type and context: static homogenous background (SHB), SHB with scale variations, SHB with different clothes, and SHB with lighting variations. Some applications of this dataset for action recognition are [30,174,182,195,230] and for feature extraction [105,109–111].

2.5. ETISEO: Evaluation du Traitement et de l'Interpretation de Sequences Video

ETISEO [85] was a two-year project started in 2005 and developed in the INRIA institute. It was created to improve video surveillance algorithm robustness and to reduce dependencies between algorithms and their conditions of use. For that, the project tried to acquire precise knowledge of vision algorithms and to create a space for discussion among participants. One of the main objectives was to create two ontologies: the first one to describe technical concepts used in the whole video interpretation chain (e.g., a blob, a mobile object, an individual trajectory) as well as concepts associated to evaluation criteria; and the second one to describe concepts of the application domain (e.g., a bank attack event). Other project goal was to conceive automatic evaluation tools for vision algorithms to allow a fair and quantitative comparison between algorithm results and reference data.

In the ETISEO dataset, the videos are grouped into five topics: *apron*, *building corridor*, *building entrance* (see Fig. 7), *metro* and *road*. Focusing in human activity recognition, the available actions provided in this dataset are: walking, running, sitting, lying, crouching, holding, pushing, jumping, pick up, puts down, fighting, queueing, tailgating, meeting and exchanging an object. Further details of this dataset can be found in [142,143].

As ground truth information, a database describes the contents of each video clip with the following attributes: those related to general information such as sequence identification, name of the dataset, name of the sequence, and difficulty level of the video; attributes related to the scene information such as indoor/outdoor scene and sequence type (road, apron, metro, etc.); other attributes related to video acquisition information such as date-hour, duration, number of frames, camera type, and camera calibration; attributes with information about light conditions such as light source (natural or artificial), weather condition (sunny, cloudy, foggy, rainy or snowy), illumination condition, and illumination variation; context attributes such as background update (empty scene available or not), object number (few or crowed), occlusions, and artifacts (shadows, reflections, noise); and, finally, attributes related to the context such as the nature of physical objects (physical object of interest, contextual object), physical object (sub)-type, and events or states. At the web page [85], there are a lot of papers (describing some applications oriented to object detection, tracking and event recognition) using this dataset.

2.6. ViSOR: Video Surveillance On-line Repository for Annotation Retrieval

The Imagelab Laboratory of the University of Modena and Reggio Emilia started the ViSOR project [159] in 2005, a video surveillance online repository for annotation retrieval. The first aim of

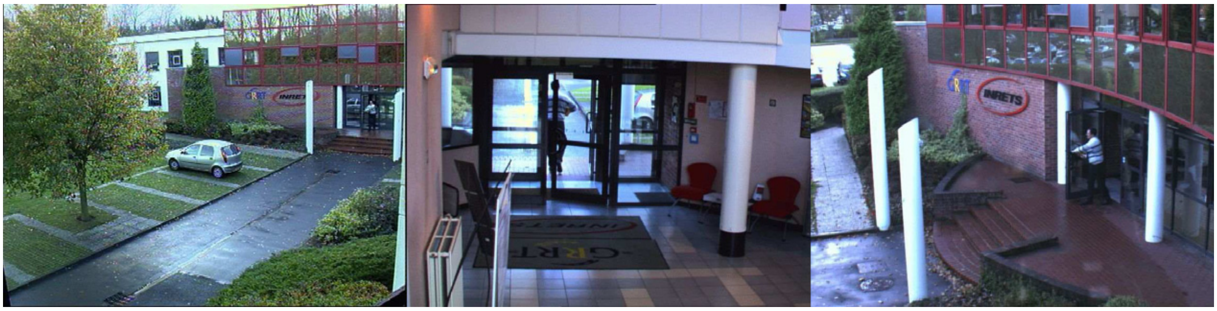
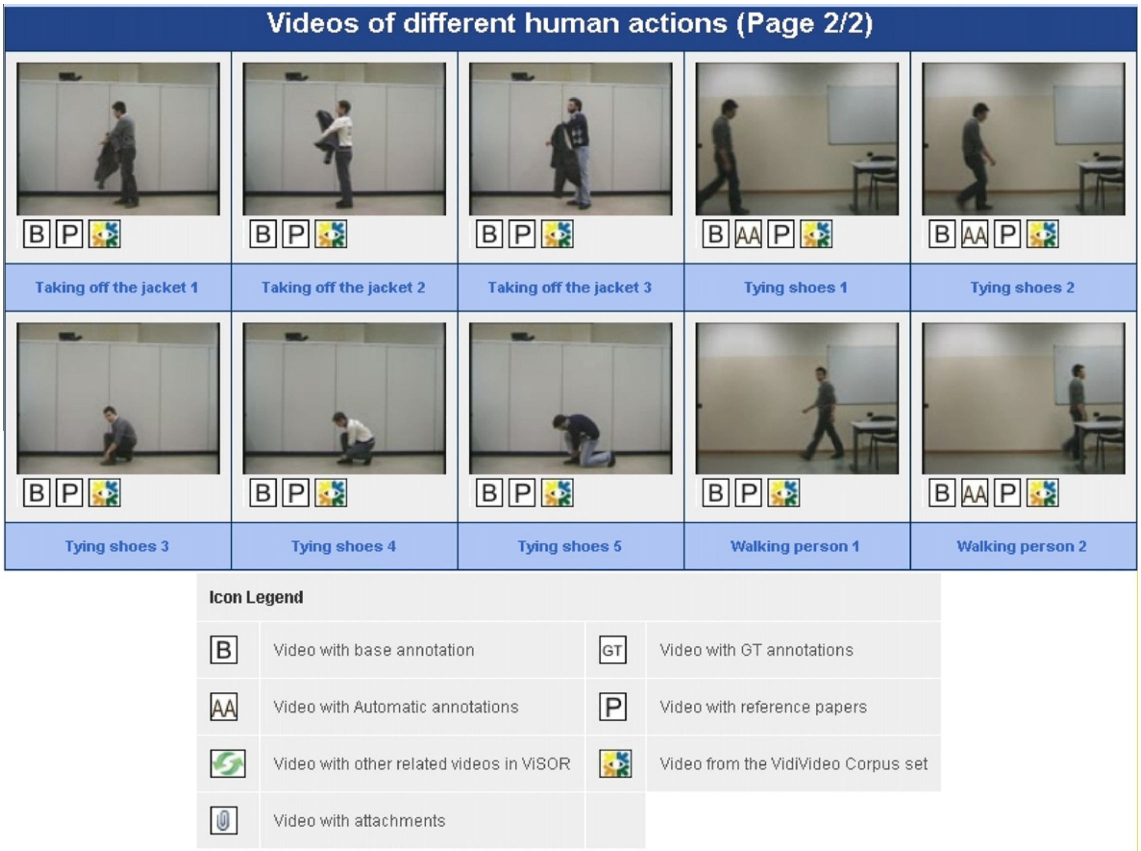


Fig. 7. Three views of the ETISEO building entrance dataset.

ViSOR was to gather and make freely available surveillance video footages for the research community on pattern recognition and multimedia retrieval. Together with the videos, ViSOR defines an ontology for metadata annotations, both manually provided as ground truth and automatically obtained by video surveillance systems. Annotation refers to a large ontology of concepts on surveillance and security related to objects and events. Vezzani and Cucchiara, the direct authors involved in the ViSOR project, have written several publications describing this general repository [212–214]. The native annotation format, supported by ViSOR, is ViPER [104].

There are several types of videos sorted in different categories. Some categories can be used directly for human action and activity recognition, as those found inside the category "Videos for human action recognition in video surveillance" (by MICC-University of Florence) with 130 video sequences. Because ViSOR has been designed as a big repository of video sequences, the contents are updated continuously. Fig. 8 shows some screenshots of videos belonging to the indoor category.

The available ground truth depends on each specific video. ViSOR has created a general framework for doing the annotations. To this end, there are three directions based on which an annotation can be differently detailed: the temporal level, the spatial level, and the domain level [214]. The temporal level can define three temporal description sub-levels: none or video-level (no temporal information is given), clip (the video is partitioned into clips and each of them is described by a set of descriptor instances) and frame (the annotation is provided frame by frame). Similar considerations can be made for the spatial level: none or image level (no spatial information is given and the concept refers to the whole frame), position (the location of the concept is specified by an individual point), ROI (the region of the frame containing the concept is reported, for example, using the bounding box) and mask (a pixel level mask is reported for each concept instance).



Icon Legend

B	Video with base annotation	GT	Video with GT annotations
AA	Video with Automatic annotations	P	Video with reference papers
⟳	Video with other related videos in ViSOR	✱	Video from the VidiVideo Corpus set
📎	Video with attachments		

Fig. 8. Thumbnails of videos from the ViSOR dataset belonging to the Indoor category.

Considering the domain level, four sub-levels can be used: none (no semantic information is provided, although free-text keywords and title can be provided), one concept (only one particular concept is considered and annotated), subset (only a subset of the ViSOR surveillance concepts is considered and the subset adopted should be indicated) and whole ontology (all the ViSOR surveillance concepts are considered).

The number of publications that have used this dataset is big. In the ViSOR web-page [159], more than 50 papers are reported, with applications to human behavior analysis, human tracking, event analysis, people counting, pedestrian crossing, human identification, smoke detection and human action recognition.

2.7. IXMAS: INRIA Xmas Motion Acquisition Sequences

IXMAS dataset [86] aims to disseminate the data acquired in 2006 with the GRImage multi-camera platform, hosted at INRIA Grenoble, France. The dataset was created to investigate how to build spatio-temporal models of human actions that could support categorization and recognition of simple action classes, independently of viewpoint, actor gender and body sizes. It is a multi-view dataset (5 cameras) for view-invariant human action recognition, where 13 daily-live motions are performed each 3 times by 11 actors. The actors freely choose position and orientation. As ground truth, silhouettes in BMP format (390×291) and reconstructed volumes in MATLAB format ($64 \times 64 \times 64$) are provided. The ground truth labelling used is the following: nothing, checking watch, crossing arms, scratching head, sitting down, getting up, turning around, walking, waving, punching, kicking, pointing, picking up, throwing (over head), and throwing (from bottom up). Two examples of publications that use this dataset for action recognition are [225,226]. Fig. 9 shows several frames from the 5 view angles.

2.8. CASIA action database

The Center for Biometrics and Security Research (CBSR) was founded by the Institute of Automation, Chinese Academy of Sciences (CASIA). CBSR aims to research and develop cutting-edge biometrics and intelligent surveillance technologies and applications, and to develop biometric standards, databases and protocols performing biometric product testing.

Specifically, the CASIA action database [58], created in 2007, is a sequence collection of human activities captured by outdoor video cameras from different angles of view. The intention of this dataset was to study how to deal with big changes in the view angle whitout using complex 3D models.

It contains eight types of human actions (walking, running, bending, jumping, crouching, fainting, wandering and punching a car) each one performed by 24 individuals, as well as seven types of interactions involving two persons (robbing, fighting, following, following and gathering, meeting and parting, meeting and gathering, overtaking). All video sequences are captured simultaneously with static three non-calibrated cameras from different angles of view (horizontal, angle and top down views).

A simple ground truth is provided using the names of the files: [view angle]_[person label]_[action]_[action_count].AVI, where [view angle] codes the angle of view (angle, horizontal or top down view), [person label] encodes the individual ID, [action] is one of the actions mentioned above, and [action count] represents the action number. Fig. 10 shows some frame examples. This dataset has been employed for humam behavior analysis in [81,82].

2.9. UCF datasets

The Department of Electrical Engineering and Computer Science at University of Central Florida (UCF, USA) has developed several useful human action datasets.

To develop algorithms for wide-area surveillance systems and high-resolution imagery acquisition, it is necessary to simulate and get representative datasets from aerial platforms. The combined need for metadata and high-resolution imagery also requires sufficient disk-storage space, which adds to the aerial platform's payload and eliminates the possibility of using small-scale, fixed-wing aircraft or rotorcraft for data collection of the images. The University of Florida's Compute Vision Lab developed a platform which was used for generating images to both UCF Aerial and UCF ARG datasets. The first one is mono-view, while the second one provides as well two other views.

UCF Sports Action dataset was devoted to benchmarking of algorithms based on temporal template matching.

UCF *YouTube* action dataset was created to recognize action from videos captured under uncontrolled conditions, such as videos recorded by an amateur using a hand-held camera. This type of video generally contains significant camera motion, background clutter, and changes in object appearance, scale, illumination conditions, and viewpoint. Finally UCF50 is an extension of *YouTube* action dataset providing more action examples.

2.9.1. UCF aerial action dataset

This dataset [148] was created in 2007 using an R/C-controlled blimp equipped with an HD camera mounted on a gimbal. The dataset comprises a diverse pool of actions featured at different heights and aerial viewpoints. Therefore it is an example of mobile camera dataset. Multiple instances of each action are recorded at different flying altitudes which range from 400 to 450 feet and are performed by different actors (see Fig. 11). The actions collected in this dataset include: walking, running, digging, picking up an object, kicking, opening a car door, closing a car door, opening a car trunk, and closing a car trunk. As ground truth, all actions were annotated using VIPER [104] including bounding boxes, action labels and descriptors (car or man). This dataset has been used, for instance, in action recognition [124].

2.9.2. UCF-ARG

The UCF-ARG (University of Central Florida-Aerial camera, Rooftop camera and Ground camera) dataset [147] was recorded in 2008 and is a multi-view human action dataset. It consists of 10 actions, performed by 12 actors, and recorded from different points of view: a ground camera, a rooftop camera at a height of 100 feet,

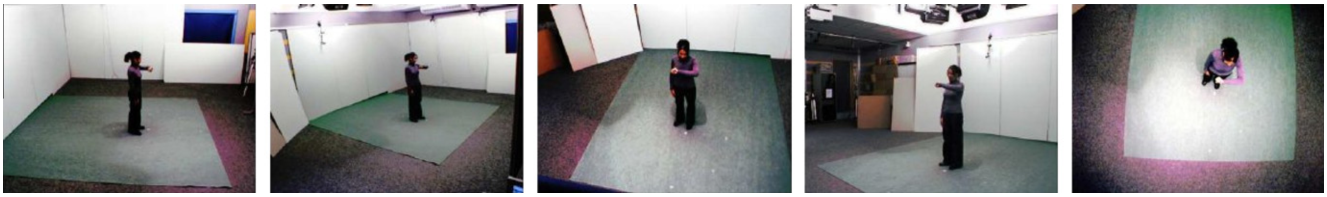


Fig. 9. Example views of 5 cameras used during acquisition in IXMAS dataset.

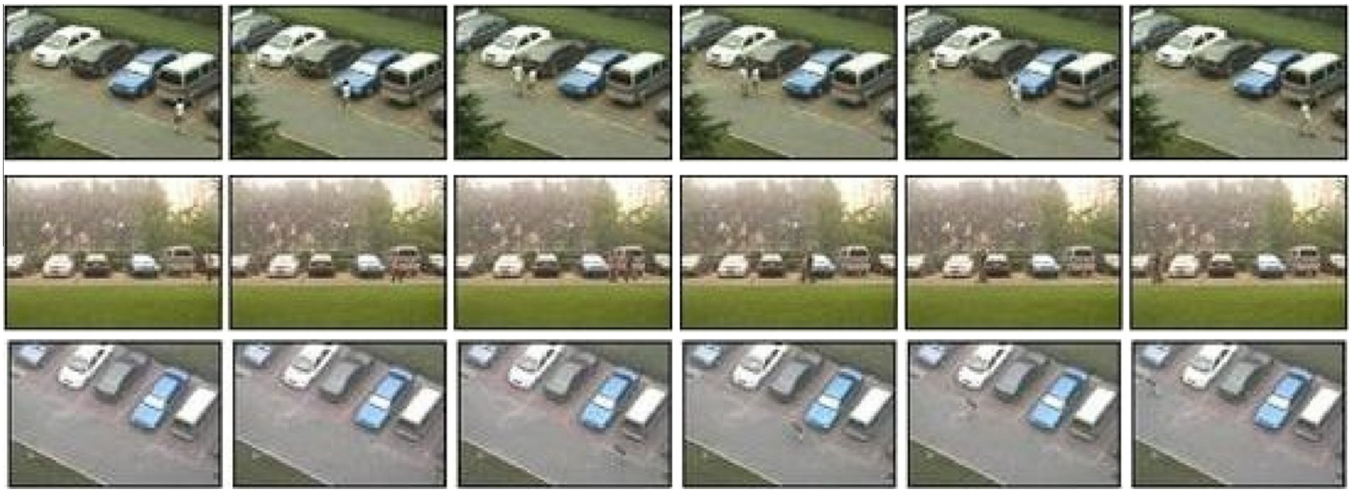


Fig. 10. Some frames from angle (top row), horizontal (middle row) and top down (bottom row) views from CASIA action dataset.

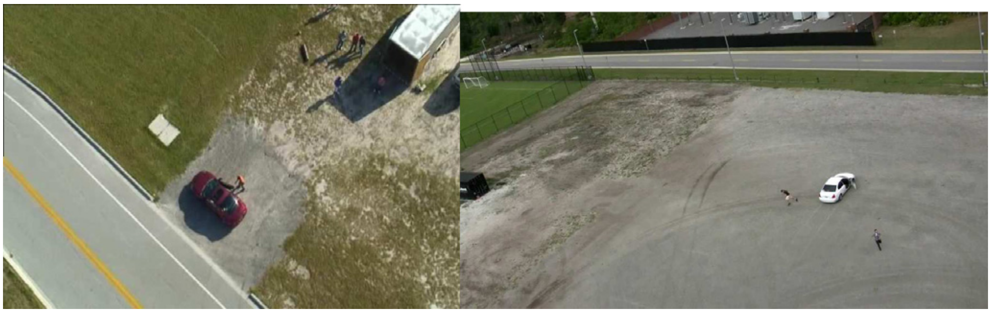


Fig. 11. Two frame examples from UCF aerial action dataset.

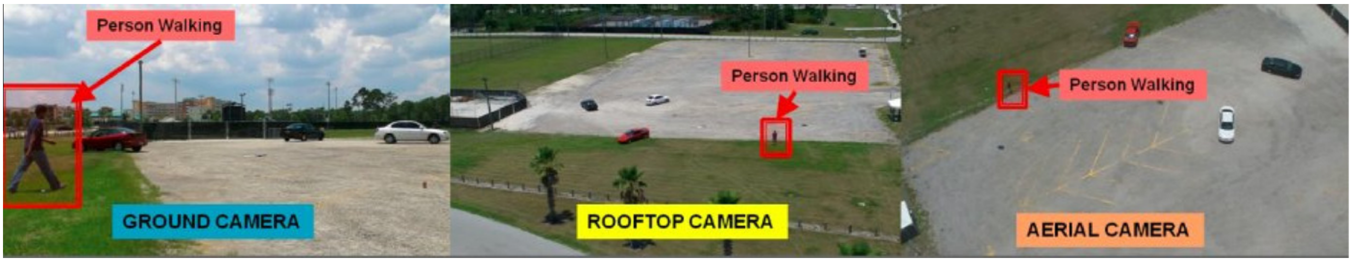


Fig. 12. Screen-shots of walking action from the three cameras of UCF-ARG dataset.

and an aerial camera mounted onto the payload platform of a helium balloon. The 10 actions are: boxing, carrying, clapping, digging, jogging, opening-closing trunk, running, throwing, walking, and waving. Except for opening-closing trunk, all the other actions are performed 4 times by each actor in different directions. Opening-closing trunk is performed only 3 times, i.e. on 3 cars parked in different directions. As ground truth, all actions are annotated using VIPER [104], including bounding boxes, action labels and descriptors (car or man). As an example, this dataset has been used for action recognition [228]. Fig. 12 shows a frame example from the three views.

2.9.3. UCF sports action dataset

This dataset [150] consists of several actions collected in 2008 from various sporting events which are typically featured on broadcast television channels such as the BBC and ESPN. The video sequences were obtained from a wide range of stock footage web sites including BBC *Motion gallery*, and *GettyImages* (see Fig. 13). The dataset comprises a natural pool of actions featured in a wide range of scenes and viewpoints.

The actions in this dataset include: diving (16 videos), golf swinging (25 videos), kicking (25 videos), lifting (15 videos), horseback riding (14 videos), running (15 videos), skating (15 videos), swinging (35 videos) and walking (22 videos). A very simple ground truth (action annotations) is provided. In [174] this dataset is used for action recognition.

2.9.4. UCF YouTube action dataset

This dataset [149] was created in 2009 with videos from *YouTube*. It contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This dataset is very challenging due to large variations in camera motion, object appearance

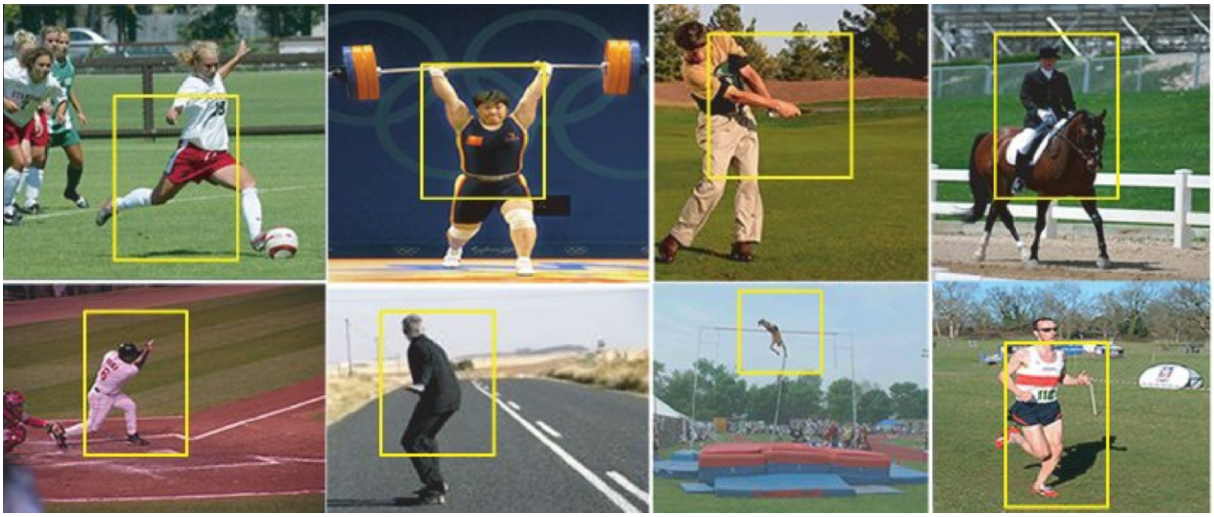


Fig. 13. Screenshots of UCF *Sports Action Dataset*.

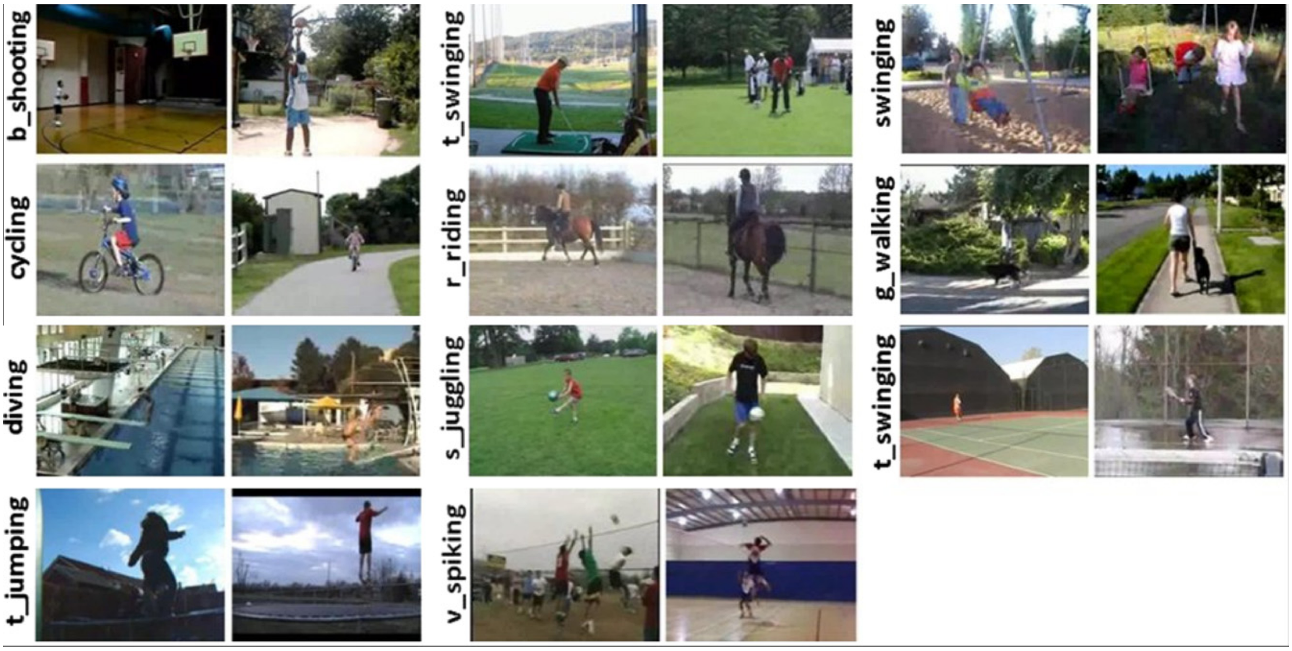


Fig. 14. Screenshots of the 11 actions of UCF *YouTube* action dataset.

and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. For each category, the videos are grouped into 25 groups with more than four action clips in it. The video clips in the same group may share some common features, such as the same actor, similar background, similar viewpoint, and so on. The ground truth is provided in VIPER [104] format giving bounding boxes and action annotation. Two examples of papers using this dataset for action recognition are [122,121]. Fig. 14 shows two example frames for each of the 11 actions.

2.9.5. UCF50

UCF50 is an action recognition dataset [150] with 50 action categories, consisting of realistic videos taken from *YouTube* in 2010. This dataset is an extension of the UCF *YouTube* Action dataset. The dataset's 50 action categories are: baseball pitch, basketball shooting, bench press, biking, billiards shot, breaststroke, clean and jerk, diving, drumming, fencing, golf swing, playing guitar, high jumping, horse racing, horse riding, hula hoop, javelin throwing, juggling balls, jumping rope, jumping jack, kayaking, lunges, military parade, mixing batter, nun chucks, playing piano, pizza tossing, pole vault, pommel horse, pull ups, punching, pushing ups, rock climbing indoor, rope climbing, rowing, salsa spins, skate boarding, skiing, skijet, soccer juggling, swing, playing tabla, taichi, tennis swing, trampoline jumping, playing violin, volleyball spiking, walking with a dog, and yo-yo. Ground truth is provided in VIPER [104] format giving bounding boxes and action annotation. For example, this dataset has been used as benchmarking of event recognition algorithms [130].

2.10. HOLLYWOOD & HOLLYWOOD-2: human actions datasets

These two datasets were created at the IRISA institute (France). They provide video more challenging that the previous datasets where only a few action classes recorded in controlled and simpli-

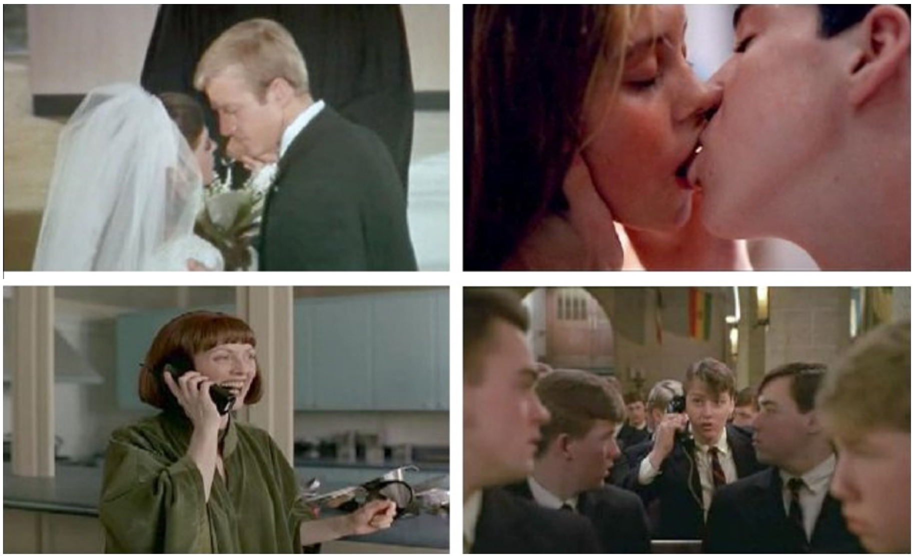


Fig. 15. Frame examples for two classes of human actions (kissing and answering a phone) from HOLLYWOOD dataset.

fied settings were available. Both of them provide individual variations of people in expression, posture, motion and clothing; perspective effects and camera motions; illumination variations; occlusions and variation in scene surroundings.

On the one hand, the HOLLYWOOD dataset [107], 2008, contains video samples with human actions from 32 movies. Each sample is labelled according to one or more of eight action classes: *AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp,* and *StandUp* (see Fig. 15 for examples). The dataset is divided into a test set obtained from 20 movies and two training sets obtained from 12 movies different from the test set. Specifically, the so-called *automatic training set* is obtained using automatic script-based action annotation and contains 233 video samples with approximately 60% correct labels. The so-called *clean training set* contains 219 video samples with manually verified labels. Finally, the *test set* contains 211 samples with manually verified labels. The ground truth information is made simply by the frame ranges and the corresponding actions. The dataset was originally used for action recognition [113].

On the other hand, HOLLYWOOD-2 [106] is an extension of the earlier HOLLYWOOD dataset and was created in 2009. The dataset intends to provide a comprehensive benchmark for human action recognition in realistic and challenging settings. HOLLYWOOD-2 dataset provides 12 classes of human actions and 10 classes of scenes distributed over 3669 video clips extracted from 69 movies and approximately 20 hours of video in total. Specifically, the 12 human actions are: *AnswerPhone, DrivingCar, Eat, Fight, GetOutCar, HandShake, HugPerson, Kiss, Run, SitDown, SitUp,* and *StandUp.*

Action samples are collected by means of automatic script-to-video alignment in combination with text-based script classification [113]. Video samples generated from training movies correspond to the automatic training subset with noisy action labels. Based on this subset, a clean training subset with action labels manually verified to be correct is also constructed. A test subset, different from the training set and with manually checked action labels, is also provided. Scene classes are selected automatically from scripts in order to maximize co-occurrence with the given action classes and to capture action context. Scene video samples are then generated using script-to-video alignment. Fig. 16 shows some frame examples. This dataset has been used, for instance, for action recognition [126].

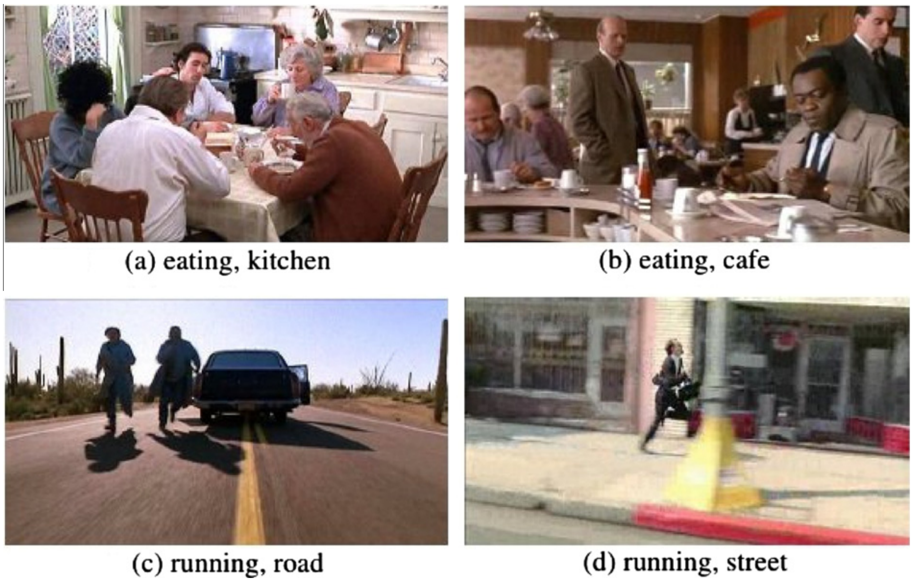
2.11. UIUC action dataset

The University of Illinois at Urbana-Champaign (UIUC) created the UIUC Action Dataset [197] in 2008 for human activity recognition with two main objectives: to reject unfamiliar activities and to learn with few examples. More details are explained in [196]. The dataset is composed of two sets. The first one (see Fig. 17, top) consists of 532 high resolution sequences of 14 activities performed by eight actors. The activities are: walking, running, jumping, waving, jumping jacks, clapping, jumping from sit-up, raise one hand, stretching out, turning, sitting to standing, crawling, pushing up and standing to sitting. Foreground masks, bounding boxes and action annotation are provided as ground truth. The second database (Fig. 17, bottom) consists of 3 badminton sequences downloaded from *YouTube*. The sequences are 1 single and 2 double matches at the Badminton World Cup 2006. As ground truth, manual annotation of type of motion (run, walk, hop, jump, unknown), type of shot (forehand, backhand, smash, unknown) and shot detection (shot, non-shot) are provided. One published work that uses this dataset as benchmarking for recognizing human actions is [123].

2.12. i3DPost multi-view dataset

i3DPost is a multi-view/3D human action/interaction database [156] created in 2009 as a cooperation between University of Surrey and CERTH-ITI (Centre of Research and Technology Hellas Informatics and Telematics Institute) within the i3DPost project. According to the authors of this dataset, it was expected that full view invariant action recognition, robust to occlusion, would be much more feasible through algorithms based on multi-view videos or 3D posture model sequences. On the contrary, the vast majority of early human action recognition methods used single-view video sources and posed the requirement that the human was captured from the same viewing angle during both the testing and training stage.

The database was recorded using a convergent eight camera setup to produce high definition multi-view videos, where each video depicts one of eight people performing one of twelve different human motions. Various types of motions were 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 individuals interact with each other. There are 8 people



(a) eating, kitchen

(b) eating, cafe

(c) running, road

(d) running, street

Fig. 16. Video samples from HOLLYWOOD-2 dataset.

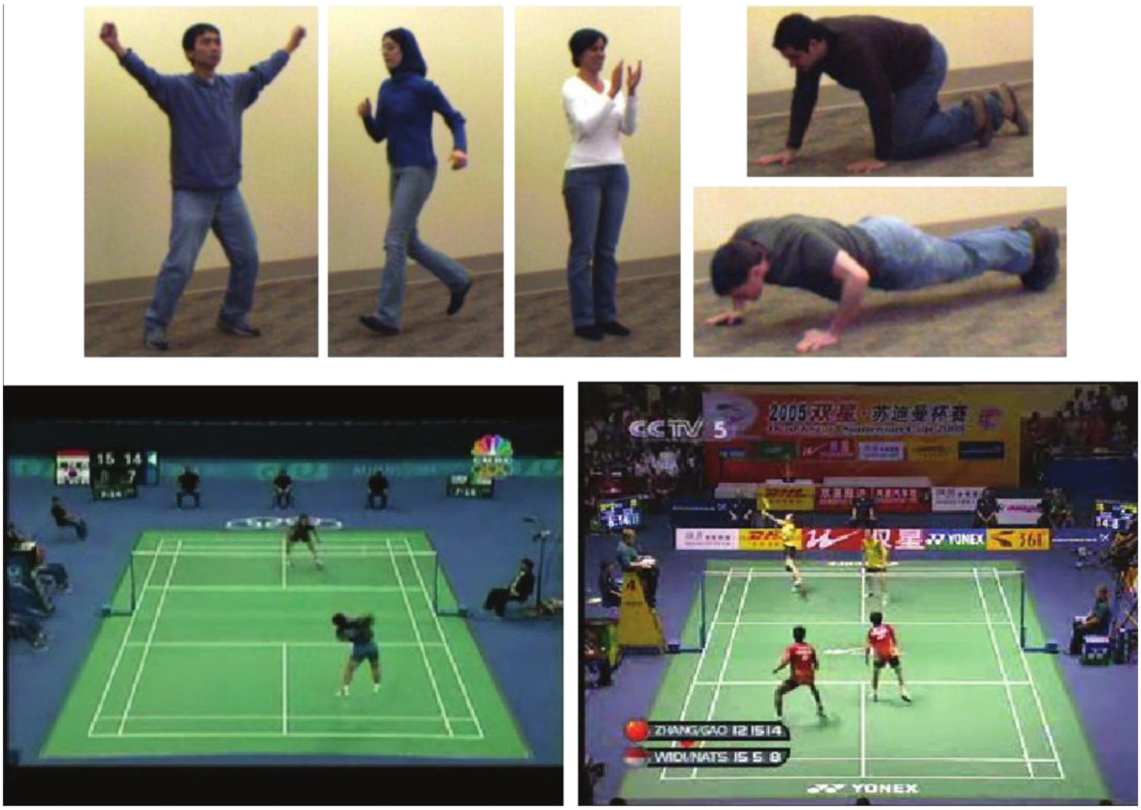


Fig. 17. Example frames of UIUC Action Dataset: first subset (top) and second one (bottom).

performing 13 actions (walking, running, jumping, bending, hand-waving, jumping in place, sitting-stand up, running-falling, walking-sitting, running-jumping-walking, handshaking, pulling, and facial-expressions) each one. Therefore a total of 104 video-sequences are recorded. The actors have different body sizes, clothing and are of different sex, nationality, etc. The multi-view videos have been processed to produce a 3D mesh at each frame describing the respective 3D human body surface. Details about the dataset are described in [65]. See Fig. 18 for some example frames. Background images are provided, and camera calibration parameters (intrinsic and extrinsic) for 3D reconstruction mesh models were computed using a global optimization method [190]. Some examples of works using this dataset are human movement recognition [66,87] and 3D human action recognition [79].

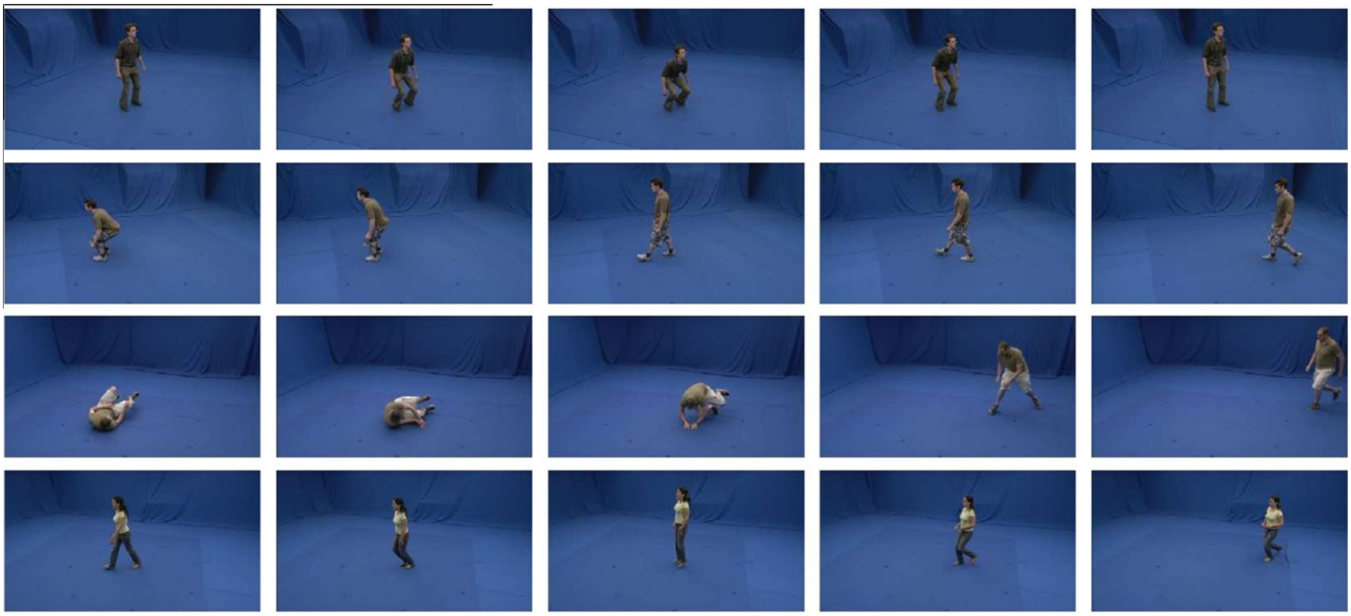


Fig. 18. From top to bottom row, frames of the following actions of i3DPost multi-view dataset are shown: sitting down-standing up, walking-sitting down, running-falling and running-jumping-walking.

2.13. MSR action dataset

The MSR Action Dataset [234], created in 2009 to study the behavior of recognition algorithms in presence of clutter and dynamic backgrounds and other types of action variations. A detailled description can be seen in [233]. The dataset contains 16 video sequences and includes three types of actions: hand clapping (14 examples), hand waving (24 examples), and boxing (25 examples), performed by 10 people. Each sequence contains multiple types of actions. Some sequences contain actions performed by different people. There are both indoor and outdoor scenes. All the video sequences are captured with clutter and moving backgrounds with lengths ranging from 32 to 76 s. As ground truth, there are manually labelled spatio-temporal bounding boxes for each action (see Fig. 19 for some frame examples). Application examples of this dataset are: adaptive action recognition [34] and action recognition [29].

2.14. MuHAVi: Multicamera Human Action Video Data

The Faculty of Science, Engineering and Computing of Kingston University collected in 2010 a large body of human action video data named MuHAVi (Multicamera Human Action Video dataset) [204] for the purpose of evaluating silhouette-based human action recognition methods. It provides a realistic challenge to both the segmentation and human action recognition communities and can act as a benchmark to objectively compare proposed algorithms. There are 17 action classes performed by 14 actors. Each actor performs each action several times in the action zone highlighted using white tapes on the scene floor. A total of eight non-synchronized cameras located at four sides and four corners of a rectangular platform are used. The patterns on the scene floor can be used to calibrate the cameras of interest. Further information can be seen in [187].

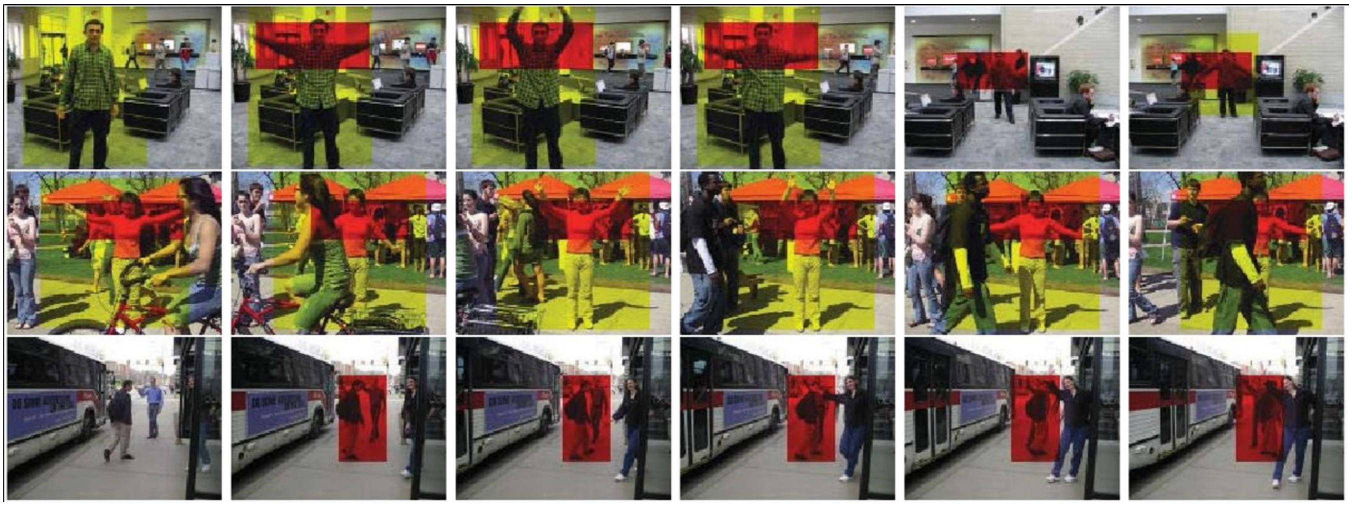


Fig. 19. Detection result example frames related to two-hand waving in MSR action dataset. The yellow bounding box is the ground truth label of the whole human body action and the red bounding box is the detection of two-hand waving described in [233].

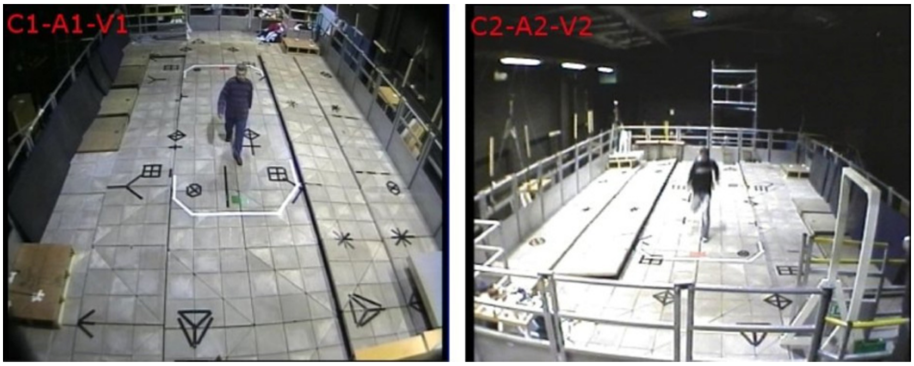


Fig. 20. Sample image frames from the MuHAVi dataset.

The action types stored in this dataset are: *WalkTurnBack, RunStop, Punch, Kick, ShotGunCollapse, PullHeavyObject, PickupThrowObject, WalkFall, LookInCar, CrawlOnKnees, WaveArms, DrawGraffiti, JumpOverFence, DrunkWalk, ClimbLadder, SmashObject,* and *JumpOverGap*. Each action class may be broken into at least two primitive actions. For instance, the action *WalkTurnBack* consists of two primitive actions: walk and turn back. Further, although it is not quite natural to have a collapse action due to shotgun followed by standing up action, one can simply split them into two separate action classes. Masks corresponding to several sequences are provided as ground truth. Fig. 20 shows two frames from this dataset. An application example of this dataset for action recognition can be consulted in [127].

2.15. Olympic sports dataset

The Olympic Sports Dataset [210] was created in 2010 and contains videos of athletes practicing different sports. The motivation of this dataset was to study the approach of modeling motion by exploiting the temporal structure of the human activities. For that, activities were represented as temporal compositions of motion segments. Therefore the dataset should provided complex human activities.

All video sequences were obtained from *YouTube* and their class labels were annotated with the help of *Amazon Mechanical Turk* [163]. The current release contains 50 videos from 16 different sports: high jump, long jump, triple jump, pole vault, discus throw, hammer throw, javelin throw, shot put, basketball lay-up, bowling, tennis serve, platform diving, springboard diving, snatch (weightlifting), clean and jerk (weightlifting), and gymnastic vault. The sequences were stored in the SEQ video format. There are MATLAB routines for reading/writing/modifying SEQ files in Piotr's MATLAB Toolbox [46]. In [144], further details of this dataset can be consulted. No ground truth is provided except simple action labels. Fig. 21 shows one frame from each class. In paper [144] a method for activity classification using Olympic Sports dataset is presented.

2.16. TV human interactions dataset

TV Human Interactions Dataset [74] was created by the Visual Geometry Group in 2010. The aim of this dataset is the recognition of interactions between two people in videos, in the context of video retrieval. In this context, several challenges must be addressed as background clutter, a varying number of people in the scene, camera motion and changes of camera viewpoints, to name a few.

This dataset consists of 300 video clips collected from over 20 different TV shows and containing four types of interactions such as handshakes, high fives, hugs and kisses, as well as clips that do not contain any of the previous interactions. The ground truth is composed by the upper body of people (with a bounding box), the discrete head orientation (profile-left, profile-right, frontal-left, frontal-right and backwards) and the interaction label of each person. In [165] a full description of this dataset is provided. Some example frames are shown in Fig. 22. The work [166] is an example paper which uses this dataset for recognizing different types of interactions between two people.

2.17. UTexas databases

The University of Texas has also developed two useful human action datasets. They were created in the frame of the Contest on Semantic Description of Human Activities (SDHA) [158], a research competition to recognize human activities in realistic scenarios, which was held in conjunction with the 20th International Conference on Pattern Recognition (ICPR 2010) [4].

The general idea behind the challenges was to test methodologies with realistic surveillance-type videos having multiple actors and pedestrians. The objective of the first challenge (using UT_Interaction dataset) was to recognize high-level interactions between two humans, such as hand-shake and push. The goal of the second challenge (UT-Tower dataset) was to recognize relatively simple one-person actions (e.g. bend and dig) taken from a low-resolution far-away camera.

2.17.1. UT-interaction dataset

The UT-Interaction dataset [179] contains videos of continuous executions of 6 classes of human–human interactions: shaking hands, pointing, hugging, pushing, kicking and punching. Ground truth labels for these interactions are provided, including time intervals and bounding boxes. There are a total of 20 video sequences whose lengths are around 1 minute. Each video contains at least one execution per interaction, providing an average of eight executions of human activities per video. Several participants with more than 15 different clothing conditions appear in the videos. The first version of this dataset was presented in [178]. Fig. 23 shows some frames of the interactions. For example, this dataset has been used for recognition of complex activities [61].

2.17.2. UT-tower dataset

The UT-Tower dataset [39] consists of 108 low-resolution video sequences from 9 types of actions. Each action was performed 12 times by 6 individuals. The dataset is composed of two types of scenes: concrete square and lawn. The possible actions are *pointing, standing, digging, walking, carrying, running, waving_1, waving_2*, and *jumping*. The camera is stationary with jitter. Ground truth labels for all actions videos are provided and consists of

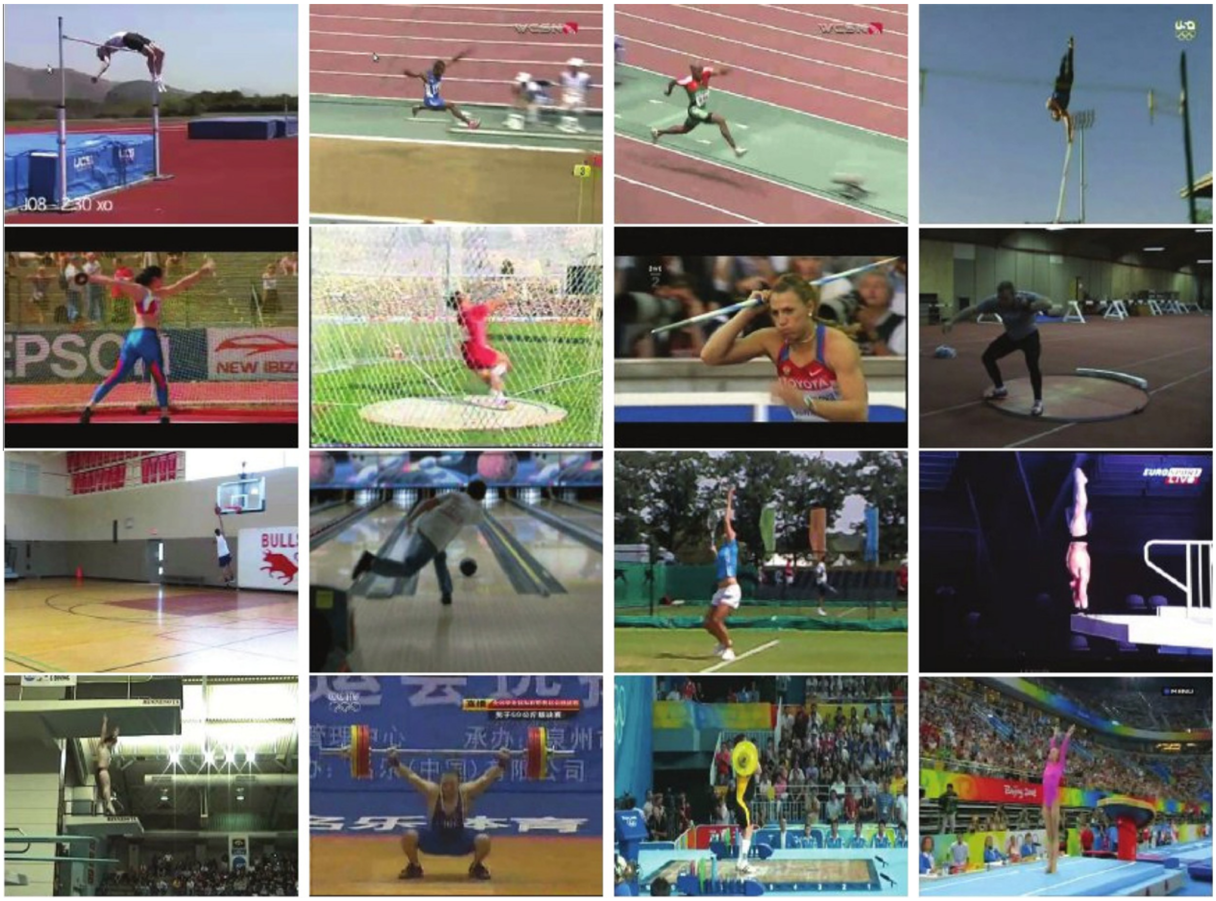


Fig. 21. One frame example of each of the 16 classes from olympic sports dataset.

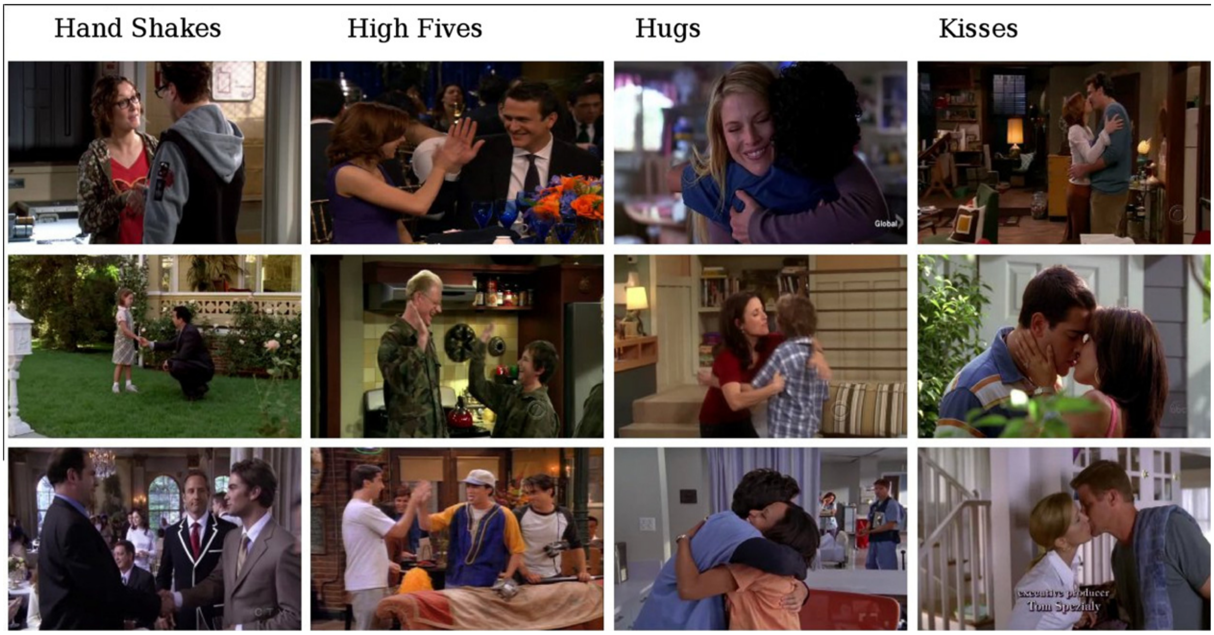


Fig. 22. Same frame examples from TV human interactions dataset.

bounding boxes and foreground masks. More details of this dataset are given in [38]. Fig. 24 shows example frames of *carrying* and *waving_1* actions. In the aforementioned work [38], this dataset is used for benchmarking an action recognition algorithm. At the ICPR 2010 web page [4] other works related with this dataset can be consulted.

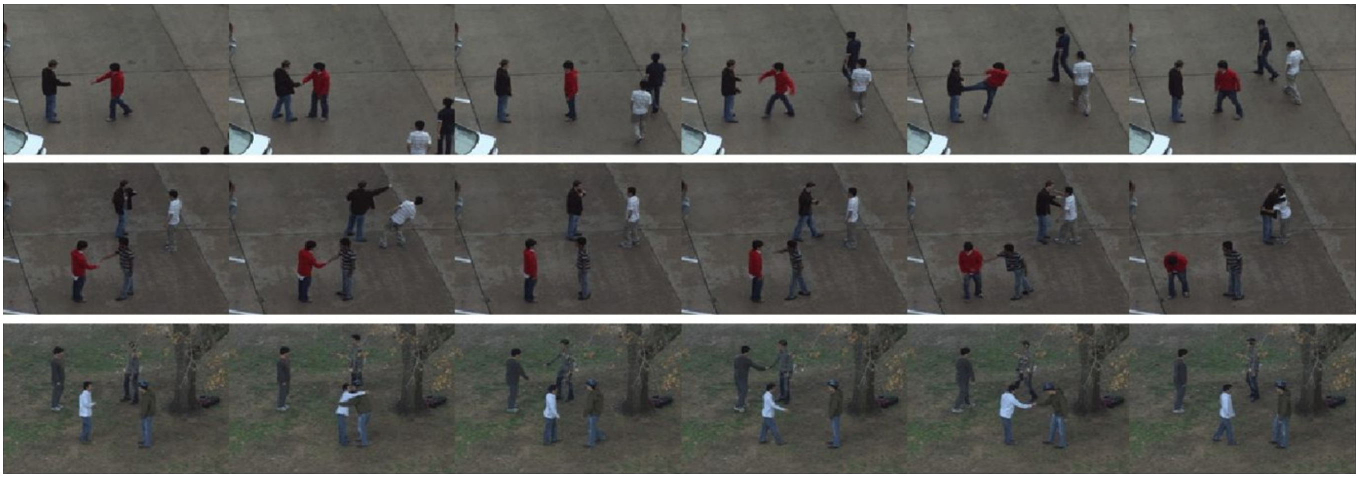


Fig. 23. Example video sequences of the UT-interaction dataset.

Fig. 24. Examples of *carrying* (top row) and *waving_1* (bottom row) actions in the UT-tower dataset.

2.18. VideoWeb dataset

The VideoWeb dataset [71] was created in 2010 by the Video Computing Group, belonging to the Department of Electrical Engineering at University of California Riverside (UCR). It tried to cover the lack of a public dataset suitable for recognizing non-verbal communication (NVC) among multiple persons. It consists of about 2.5 hours of video recorded from a minimum of 4 and a maximum of 8 cameras. The data are divided into a number of scenes that were collected over many days. Each video is recorded by a camera network whose number of cameras depends on the type of scene. The dataset involves up to 10 actors interacting in various ways (with each other, with vehicles or with facilities). The activities are: people meeting, people following, vehicles turning, people dispersing, shaking hands, gesturing, waving, hugging, and pointing. Annotations are available for each scene (frame numbers and camera ID for each activity). The videos from the different cameras are approximately synchronized. Further details of this dataset can be consulted in [62]. Fig. 25 shows a snapshot obtained from one of the eight different cameras at a particular time frame. For example, this dataset has been used for recognition of complex activities [61]. At the web page of ICPR 2010 [4] other works related with this dataset are provided.

2.19. HMDB51: a large video database for human motion recognition

The Serre lab at Brown University, USA, built the HMDB dataset [103] in 2011. Their videos were collected from various sources which were mostly from movies and, a small proportion, from public databases, such as the *Prelinger* archive, *YouTube* and *Google* videos. The dataset contains 6849 clips divided into 51 action categories, each containing a minimum of 101 clips. In paper [100], a full description is given.

As the HMDB's authors point out, the amount of daily new videos generated on the iternet is huge. There is an immediate need for robust algorithms that can help organize, summarize and retrieve this massive amount of data. HMDB dataset is an effort to advance in that direction. It tries to cover as well the reduce number of actions in previous datasets (such as KTH and Weizmann).

The actions categories can be grouped in five types: *general facial actions* (smile, laugh, chew, talk), *facial actions with object manipulation* (smoke, eat, drink), *general body movements* (cart-

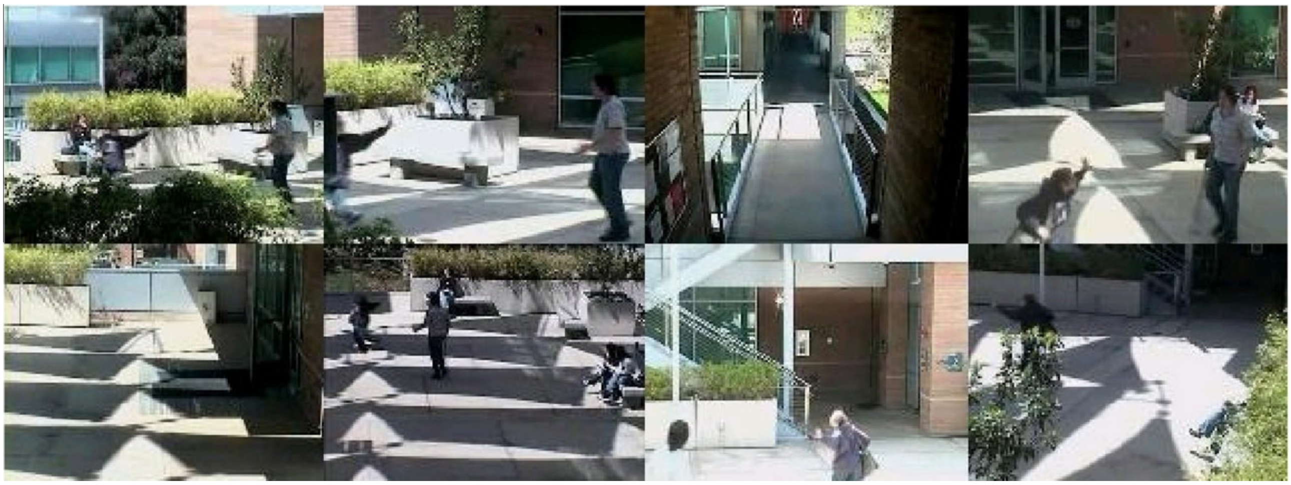


Fig. 25. Frame examples from VideoWeb dataset. Each image shows a snapshot obtained from one of 8 different cameras at a particular time frame.

wheel, clap hands, climb, climb stairs, dive, fall on the floor, backhand flip, handstand, jump, pull up, push up, run, sit down, sit up, somersault, stand up, turn, walk, wave), *body movements with object interaction* (brush hair, catch, draw sword, dribble, golf, hit something, kick ball, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw) and *body movements for human interaction* (fencing, hug, kick someone, kiss, punch, shake hands, sword fight).

Annotations are made using *space-time interest points* (STIP) [219]. In addition to the label of the action category, metadata contain the following fields: visible body parts/occlusions indicating whether the head, upper body, lower body or the full body is visible; camera motion indicating whether the camera is moving or is static; camera view point relative to the actor (labelled front, back, left or right); and the number of people involved in the action (one, two or multiple people).

The clips are also annotated according to their video quality: High (detailed visual elements, such as the fingers and eyes of the main actor, are identifiable through most of the clip), medium (large body parts, like the upper and lower arms and legs, identifiable through most of the clip) and low (large body parts are not identifiable due in part to the presence of motion blur and compression artifacts). Fig. 26 exemplifies each quality grade to show the differences.

One major challenge associated with the use of video clips extracted from real-world videos is the potential presence of significant camera/background motion, which is assumed to interfere with the local motion computation. To remove the camera motion, the authors use standard image stitching techniques to align the frames and stabilize clips. The aforementioned work [100] describes an implementation of a human recognition algorithm using this dataset.

2.20. VIRAT video dataset

The VIRAT Video Dataset [96] was designed by Kitware company [95] to be more realistic, natural and challenging for video surveillance domains than existing action recognition datasets in terms of its resolution, background clutter, diversity in scenes, and human activity/event categories. The created dataset was supported by the Defence Advanced Research Projects Agency (DARPA). The first released was published in 2011. A detailed description of this dataset can be found in [160]. Data are collected in natural scenes showing people performing normal actions in standard contexts, with uncontrolled, cluttered backgrounds. There are frequent incidental movers and background activities. Actions performed by directed actors are minimized, most are actions performed by the general population. Data are collected at multiple sites distributed throughout the USA. A variety of camera viewpoints and resolutions are used. Diverse types of human actions and human–vehicle interactions are included, with a large number of examples (more than 30) per action class. Several releases are available, being 2.0 the current one.

As ground truth, there are 12 different types of annotated events. In this context, annotated objects could be: people, vehicle or arbitrary objects, such as bags, being loaded into vehicles. Every annotated object has a duration information which consists of a starting frame number and the duration. Bounding boxes are provided as well. Static objects, such as parked vehicles, which are not involved in any activities, are also annotated. The events are annotated and represented as the set of objects being involved and the temporal interval of interest. There is a total of 12 different types of events in release 2.0: person loading an object to a vehicle, person unloading an object from a vehicle, person opening a vehicle trunk, person closing a vehicle trunk, person getting into a vehicle, person

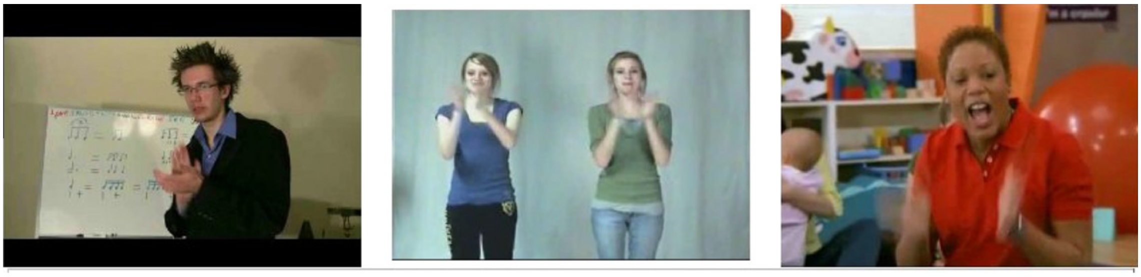


Fig. 26. Frames examples for each of the three quality grades of recording used in HMDB51 dataset.

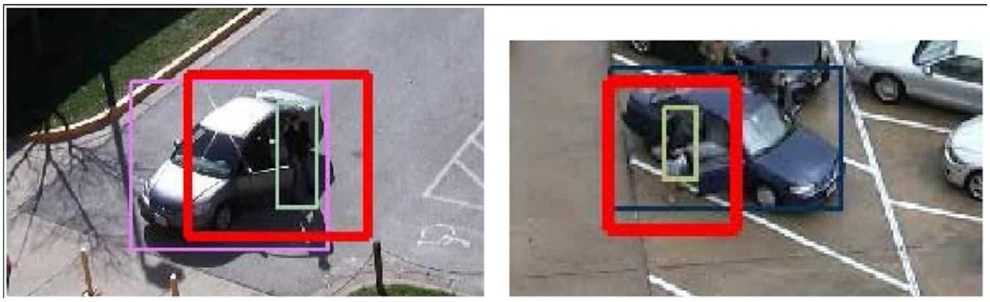


Fig. 27. Examples of ground truth from VIRAT database where both the person and vehicle bounding boxes are plot in different colurs, and the event boxes are marked by thick red box.

getting out of a vehicle, person gesturing, person digging, person carrying an object, person running, person entering a facility, and person exiting a facility. This dataset offers 3 × 3 homographies for different scenes. Each homograph provides a mapping image coordinate to scene-dependent worl coordinate. Fig. 27 shows examples of the frames annotated with bounding boxes. A license agreement is needed for downloading the dataset, and it is also mandatory to install a proprietary server to use it. An application using this dataset as benchmarking for video annotation and tracking can be found in [217] and for human–vehicle interaction recognition in [114].

3. Discussion

In this section, a comparison between the different datasets described in this survey will be accomplished.

For the sake of clarity, the main characteristics of the 28 public video datasets for human action and activity recognition described in Section 2 are shown divided in three tables (see Tables 2–4). Each table contains a subgroup of these characteristics. For an easy search, the dataset names are sorted alphabetically. Table 2 shows the following features: the year of creation or first publication that describes the dataset, the web page from which the dataset can be downloaded, the reference to the paper (if any) which describes in full detail the dataset, and areas of application where the dataset has been used for benchmarking. Normally, the year of creation (2nd column) is directly related to the number of publications which reference the dataset, being the older ones the more referenced. On the contrary, modern datasets are less cited because they are *younger*, but normally provides better ground truth information. The web page addresses (3rd column) are not only useful for downloading the datasets, but other information, such as ground truth, project, or workshop related to, can be consulted or downloaded. It is common that the dataset is presented to the scientific community by means of one or several specific papers (4th column), where detailed information about the dataset is described directly by the dataset's authors. The last column shows

Table 2
Summary of characteristics that describe datasets dedicated to the recognition of sets of heterogeneous human actions (I).

Name	Year	Web	Description	Area of application
BEHAVE	2004	[54]	–	Event detection [16,17,20], modeling crowd problems [15,20], optical flow anomalies [18,19], multiagent activities recognition [27], visual tracking [47,92]
CASIA *Action*	2007	[58]	–	Humam behavior analysis [81,82]
CAVIAR	2004	[1]	[55,120]	Target detectors [78], activity monitoring [33], activity recognition [26,118,119,139,168,173,200], clustering of trajectories [7], activities identification [49], motion segmentation [140,141], tracking [77,91], multiagent activity recognition [28]
ETISEO	2005	[85]	[142,143]	Object detection, tracking and event recognition [85]
HMDB51	2011	[103]	[100]	Action recognition [100]
HOLLYWOOD	2008	[107]	[113]	Action recognition [113]
HOLLYWOOD-2	2009	[106]	[126]	Action recognition [126]
IXMAS	2006	[86]	–	Action recognition [225,226]
i3DPost Multi-view	2009	[156]	[65]	Human movement recognition [66,87], 3D human action recognition [79]
KTH	2004	[108]	[180]	Action recognition [30,174,180,182,195,230], feature extraction [105,109–111]
MSR Action	2009	[234]	[233]	Action recognition [29,34]
MuHAVi	2010	[204]	[187]	Action recognition [127]
Olympic Sports	2010	[210]	[144]	Action recognition [40,144], complex event detection [192]
TV Human Interaction	2010	[74]	[165]	Recognizing interactions between two people [166], activity recognition [41,115]
UCF *Aerial*	2007	[148]	–	Action recognition [124]
UCF-ARG	2008	[147]	–	Action recognition [228]
UCF *Sports*	2008	[150]	[174]	Action recognition [174], action retrieval [90], object detection [60]
UCF *YouTube*	2009	[149]	[122]	Action recognition [122,121], modeling actions [35]
UCF50	2010	[150]	–	Event recognition [130], action recognition [222] representation of activities [43]
UIUC Action	2008	[197]	[196]	Action recognition [123,196]
URADL	2009	[153]	[131]	Action recognition [131,232]
UT-Interaction	2010	[179]	[178]	Complex activities recognition [61,238], recognizing interactions [135], activity prediction [177], localizing participants of group activities [14]
UT-Tower	2010	[39]	[38]	Action recognition [38], recognizing action at a distance [135]
VideoWeb	2010	[71]	[62]	Complex activities recognition [61], modeling multi-object activities [183]
VIRAT	2011	[96]	[160]	Annotation and tracking [217], human-vehicle interaction recognition in [114], action recognition [25,124]
ViSOR	2005	[159]	[212–214]	Behavior analysis [44], tracking [215], identification [24], action recognition [23]
WEIZMANN *Actions*	2005	[69]	[68]	Action recognition [30,68,174,182,195,220,230], motion recognition [231], pose estimation [50,184], gesture recognition [138], person recognition [181].
WEIZMANN *Event*	2001	[236]	–	Temporal segmentation [235]

Table 3
Summary of characteristics that describe datasets dedicated to the recognition of sets of heterogeneous human actions (continuation II).

Name	No. of actions: for 1 person/multiple people	No. of actors	Scenes	No. of views	Camera movement
BEHAVE	0/10	–	Outdoors	2	Static
CASIA *Action*	8/7	24	Outdoors	3	Static
CAVIAR	7/2	–	In/Outdoors	1, 2	Static
ETISEO	10/5	–	In/Outdoors	1, 3, 4	Static
HMDB51	44/7	–	In/Outdoors	1	Several
HOLLYWOOD	5/3	–	In/Outdoors	1	Several
HOLLYWOOD-2	8/4	–	In/Outdoors	1	Several
IXMAS	13/0	11	Indoors	5	Static
i3DPost Multi-view	11/2	8	Indoors	8	Static
KTH	6/0	25	In/Outdoors	1	Static
MSR Action	3/0	10	In/Outdoors	1	Static
MuHAVi	17/0	14	Indoors	8	Static
Olympic Sports	16/0	–	In/Outdoors	1	Mobile
TV Human Interaction	0/4	–	In/Outdoors	1	Several
UCF *Aerial*	9/0	–	Outdoors	1	Mobile
UCF-ARG	10/0	12	Outdoors	3	Mobile
UCF *Sports*	9/0	–	In/Outdoors	1	Several
UCF *YouTube*	11/0	–	In/Outdoors	1	Several
UCF50	50/0	–	In/Outdoors	1	Several
UIUC Action	14/0	8	Indoors	1	Static
URADL	10/0	5	Indoors	1	Static
UT-Interaction	0/6	–	Outdoors	1	Static
UT-Tower	9/0	6	Outdoors	1	Static
VideoWeb	2/6	–	Outdoors	4–8	Static
VIRAT	12/0	–	Outdoors	–	Static
ViSOR	–	–	In/Outdoors	–	Several
WEIZMANN *Actions*	10/0	9	Outdoors	1	Static
WEIZMANN *Event*	4/0	–	Outdoors	1	Static

areas of application in which each dataset has been used and also provides references to some published works in each of them. This reference list has no aspirations to be exhaustive, that is, it only contains a sample of papers that use each dataset.

Table 3 summarizes the following information: the total number of actions registered, the total number of actors, the type of scenes, the number of views of a scene, and if the camera used is mobile or not. Because the problem related to action recognition involving interactions is reaching more and more interest, the total number of actions (2nd column) is presented distinguishing between those carried out by a single actor and those made by two or more people interacting. When it is known, the total number

Table 4
Summary of characteristics that describe datasets dedicated to the recognition of sets of heterogeneous human actions (continuation III).

Name	Ground truth	Calibration data?	Agreement required?
BEHAVE	Bounding boxes in VIPER	Yes	No
CASIA *Action*	Action annotation		Yes
CAVIAR	Bounding boxes and action annotation in CVML	Yes	No
ETISEO	General attributes (sequence id, difficulty level, etc.), scene attributes (In/Outdoors, sequence type), video acquisition (duration, camera calibration, etc.), light conditions (light source, weather conditions, etc.), illumination conditions. VIPER format.	Yes	Yes
HMDB51	Actions, body parts, camera motion, number of actors, video quality	No	No
HOLLYWOOD	Action annotation	No	No
HOLLYWOOD-2	Frame ranges, action annotation	No	No
IXMAS	Silhouettes, action annotation, 3D mesh models	Yes	No
i3DPost Multi-view	Action annotation, background, 3D volumes	Yes	Yes
KTH	Simple action annotation	No	No
MSR Action	Action labels, bounding boxes	No	Yes
MuHAVi	Silhouettes bounding boxes	Yes	Yes
Olympic Sports	Simple action annotation	No	No
TV Human Interaction	Upper body, head orientation, interaction level	No	No
UCF *Aerial*	Bounding boxes, object and action labels, in VIPER	Yes	No
UCF-ARG	Bounding boxes, object and action labels, in VIPER	No	No
UCF *Sports*	Simple action annotation	No	No
UCF *YouTube*	Frame ranges, bounding boxes, action annotation, in VIPER	No	No
UCF50	Frame ranges, bounding boxes, action annotation, in VIPER	No	No
UIUC Action	Bounding boxes, foreground masks	No	No
URADL	Temporal segmentation, name activities	No	No
UT-Interaction	Time intervals, bounding boxes	No	No
UT-Tower	Action labels, foreground masks, bounding boxes	No	No
VideoWeb	Frame number and camera ID for each activity	No	Yes
VIRAT	Time annotation, bounding boxes	Yes	Yes
ViSOR	Several levels, in VIPER	No	No
WEIZMANN *Actions*	Silhouettes	No	No
WEIZMANN *Event*	Temporal annotation	No	No

Table 5
Rough dataset ranking according to an estimate of the number of papers which uses directly them for benchmarking.[a] Those datasets with the same number of references are sorted according to the publication date.

Rank	Name	Year	Number of references
1	CAVIAR	2007	>100
2	KTH	2004	>100
3	WEIZMANN (Action&Event)	2001/2005	>100
4	ViSOR	2005	65
5	IXMAS	2006	59
6	UCF *Sports*	2008	21
7	UT-Interaction	2010	15
8	HOLLYWOOD	2008	15
9	ETISEO	2005	15
10	MSR Action	2009	12
11	BEHAVE	2004	12
12	UT-Tower	2010	11
13	VideoWeb	2010	11
14	MuHAVi	2010	11
15	Olympic Sports	2010	11
16	i3DPost Multi-view	2009	10
17	HOLLYWOOD-2	2009	9
18	UCF *YouTube*	2009	9
19	VIRAT	2011	7
20	CASIA *Action*	2007	7
21	URADL	2009	5
22	TV Human Interaction	2010	4
23	UCF50	2010	3
24	UIUC *Action*	2008	3
25	HMDB51	2011	2
26	UCF-ARG	2008	1
27	UCF *Aerial*	2007	1

[a] Due to space limitations, not all the references used to prepare this table are included in Reference section. However, some of them appear at column 5 of Table 2 to illustrate the different uses of each dataset. To elaborate the ranking, the following search-engines have been employed: IEEE Xplore, Springer Link, ACM Digital Library and Science Direct.

of actors or ordinary people engaged in all the videos stored in a dataset is provided (3rd column). The datasets recorded in controlled conditions normally use indoor scenarios. However, like can be seen in the fourth column, there are also datasets obtained in outdoor scenarios or in both. Obviously, outdoor scenario datasets are more challenging than indoor ones. The fifth column shows the number of views used for recording each scene. Multiple views of a scene could be interesting when, for example, the hidden information in a view, due to occlusion, can be obtained from the other views. Other important information is whether the videos are recorded with mobile or static cameras (6th column). This will strongly influence the choice of algorithm to be used for working with the dataset.

Finally, Table 4 presents the following information: ground truth, calibration data, and whether an agreement is required or not to download each dataset. The ground truth (2nd column) is valuable information because it aids in the process of interpretation and analysis of the actions and activities recorded in the videos of each dataset. The more ground truth information you have, the more versatile the dataset will be. Related to the ground truth, calibration data (3rd column) allow us to compute the real physical coordinates of the moving objects and actors on the image plane. The last column is devoted to the license agreement requirements. The majority of the datasets are directly downloadable. However others need a license agreement. In both cases, the dataset is free of charge and the only condition to be used is to reference the dataset source.

The datasets described in Section 2 can be classified in different clusters depending on different features considered. For example, some datasets do not have a controlled background because the original sequences were recorded for purposes other than action recognition. This is the case of HMDB51, HOLLYWOOD, HOLLYWOOD-2, Olympic Sport, TV Human Interactions, UCF *Sports*, UCF *YouTube*, UCF50 and ViSOR. Nevertheless, in some applications, such as modern surveillance systems, the background characteristics are known in advance.

A big dispersion in the number of actions considered by each dataset is observed. Specifically, this number ranges going from 3 to 51. The datasets with the greater number of actions (over 10) are: UCF *YouTube* (11), HOLLYWOOD2 (12), VideoWeb (12), IXMAS (13), i3DPost (13), ETISEO (14), CASIA Action (15), Olympic (16), MuHAVi (17), UCF50 (50) and HMDB51 (51).

Some human actions require several actors in the same scene and are suitable for studding social interactions. This type of actions appears in the following datasets: BEHAVE, CASIA Action, CAVIAR, HMDB51, HOLLYWOOD, HOLLYWOOD2, ETISEO, TV Human Interaction, UT-Interaction and VideoWeb. Equally, other datasets related to sports, such as Olympic Sports, UCF Sports, UCF *YouTube*, UVF50, and second set of UIUC Action dataset, can be suitable for studying interactions with objects or other people (sportsmen).

In order to evaluate the effect of each dataset in scientific community, Table 5 shows a simple metric based on a rough estimate of the number of papers which use each dataset for benchmarking. Other papers which do not use directly the datasets, as surveys, are not included in the list. In any case, the numbers of works has been taken as an approximation. Note that there is a certain bias in the figures because older datasets have more chances to be used. Therefore, although the datasets are sorted according the number of publications, the year of creation is provided as well. As we can see, the most prolific datasets dealing with the number of publications are KTH, WEIZMANN and CAVIAR with more than 100 publications each one. Specially successful is CAVIAR dataset, which is more recent than KTH and WEIZMANN and has roughly the same number of references. Other datasets are still not referenced too much because they were created recently. This is the case of VIRAT and HMDB51 datasets, from 2011. In particular, 86% of the datasets described here were created from 2005 onward.

Another important feature to take into account is the number of available views of each scene. All datasets described are mono-view, except BEHAVE, CASIA Action, CAVIAR, ETISEO, IXMAS, i3DPost Multi-view, MuHAVi, UCF-ARG and VideoWeb that are multi-view. Other datasets are recorded from an aerial view, such as UCF Aerial, UCF-ARG and UT-Tower.

Attending to the type of people who is present in the scene, the datasets can be also grouped in two categories. The first one is composed by those datasets where specific actors (normally amateurs) are recruited for the performance: BEHAVE, CASIA Action, IXMAS, i3DPost Multi-view, KTH, MSR Action, MuHAVi, UCF Aerial, UCF-ARG, first dataset of UIUC, UT-Interaction, UT-Tower, VideoWeb, ViSOR, WEIZMANN Event and WEIZMANN Action. The other group is composed by those datasets whose individuals are ordinary people just passing in front of the camera, or people and actors that originally were recorded for other purposes (*YouTube* videos, movies, etc.): CAVIAR, HMDB51, ETISEO, HOLLYWOOD, HOLLYWOOD-2, Olympic Sports, TV Human Interactions, UCF Sports, UCF *YouTube*, UCF50, second dataset of UIUC and VIRAT.

The ground truth annotation data is one of the most relevant features of a dataset. Some datasets only provide a very simple ground truth (name of the actions for each frame): CASIA *Action*, HOLLYWOOD, HOLLYWOOD-2, KTH, Olympic Sports, UCF Sports and WEIZMANN Event. Other datasets provide silhouettes of each person appearing in the scene: IXMAS, MuHAVi and WEIZMANN Actions. On the contrary, other datasets give a high quality ground truth using XML derived languages as VIPER [104] or CVML [117]. This is the case of BEHAVE, CAVIAR, ETISEO, UCF Aerial, UCF-ARG, UCF *YouTube*, UCF50 and ViSOR.

Table 6
Heterogeneous action dataset classification according to different features.

	Datasets
View type	
Mono-view	HMDB51, HOLLYWOOD, KTH, MSR, Olympic, TV Human Interaction, UCF Sports, UCF *YouTube*, UCF50, UIUC Action, URADL, UT-Interaction, UT-Tower, WEIZMANN
Multi-view	BEHAVE, CASIA Action, CAVIAR, ETISEO, IXMAS, i3DPost, MuHAVi, UCF-ARG, VideoWeb
Aerial-view	UCF Aerial, UCF-ARG
Ground truth	
Bounding Boxes	BEHAVE, CAVIAR, MSR, MuHAVi, UCF Aerial, UCF-ARG, UCF *YouTube*, UCF50, UIUC Action, UT-Interaction, UT-Tower, VIRAT
Scene Attributes	ETISEO
Video Quality	HMDB51
Silhouettes	IXMAS, MuHAVi, WEIZMANN
3D models/ volumes	IXMAS, i3DPost
Foreground masks	UIUC Action, UT-Tower
Actions number	
1-5	MSR (3), TV Human Intereaction (4), WEIZMANN Event (4)
6-10	BEHAVE (10), CAVIAR (9), HOLLYWOOD (8), KTH (6), UCF Aerial (9) UCF-ARG (10), UCF Sports (9), URADL (10), UT-Interaction (6), UT-Tower (9), VideoWeb (8), WEIZMANN Action (10)
11-15	CASIA Action (15), ETISEO (15), HOLLYWOOD-2 (12), IXMAS (13), i3DPost (13), UCF *YouTube* (11), UIUC Action (14), VIRAT (12)
16-20	MuHAVi (17), Olympic Sports (16)
>20	HMDB51 (51), UCF50 (50)
Interaction	
None	CASIA Action, CAVIAR, ETISEO, HMDB51, HOLLYWOOD, IXMAS, i3DPost, KTH, MSR, MuHAVi, Olympic Sports, UCF Aerial, UCF-ARG, UCF Sports, UCF *YouTube*, UCF50, UIUC Action, URADL, UT-Tower, VideoWeb, VIRAT, WEIZMANN
Person-Person	BEHAVE, CASIA Action, CAVIAR, ETISEO, HMDB51, HOLLYWOOD, i3DPost, TV Human Interaction, UT-Interaction, VideoWeb
Person-Object	CASIA Action, HMDB51, HOLLYWOOD
Movil camera	
Yes	HMDB51, HOLLYWOOD, Olympic Sports, TV Human Interaction, UCF Aerial, UCF-ARG, UCF Sports, UCF *YouTube*, UCF50, ViSOR
No	BEHAVE, CASIA Action, CAVIAR, ETISEO, HMDB51, HOLLYWOOD, IXMAS, i3DPost, KTH, MSR, MuHAVi, TV Human Interaction, UCF Sports, UCF *YouTube*, UCF50, UIUC Action, URADL, UT-Interaction, UT-Tower, VideoWeb, VIRAT, ViSOR, WEIZMANN

The above information has been summarized in Table 6, which facilitates the search of a specific dataset fulfilling several requirements. Note that, in that table, the same dataset can appear in several groups and sub-groups. For instance, CASIA Action provides actions of one person, interactions between several people and one interaction between one person and a car. Therefore, this dataset is in all the sub-groups inside the classification according to the type of interactions.

4. Specific action datasets

There are other datasets in the literature which focus on specific actions. This type of datasets is included here for the sake of completeness but, due to space constraints, they are only described in a summarized way (see Table 7). Now, instead of storing sets of heterogeneous human actions, like do the datasets described in Section 2, this new type of datasets are dedicated to study specific type of actions, such as detecting falls, detecting abandoned objects, crowd behaviour, and pose or gesture recognition. Thus, fall detection is a task where computer vision provides new and promising solutions such as the development of new health-care systems to help elderly people staying at home in a secure environment [37]. The abandoned object detection is also an important issue in video surveillance systems for public facilities. As it was described in Section 3, the study of different social interactions between two or more people is an area of growing interest. In the limit, there are some studies devoted to the analysis of the interaction of multiple people (crowd behaviour) [237]. Other datasets, thought to recognise specific human actions, can be classified in the ADL (Activities of Daily Living) group. ADL is a term used in healthcare to refer to daily self-care activities within an individual's place of residence, in outdoor environments, or both. Finally, the pose or gesture recognition is dedicated to actions performed by a specific part of the human body, such as hands, arms or upper body part [48,98,167]. There exists other special types of datasets that are also considered in this summary: the gait analysis datasets and the so-called motion capture (MOCAP) datasets (see [75] for a recently survey). The former are dedicated to the task of studying systematically the human motion, augmented by instrumentation for measuring body movements and body mechanics. The latter were originally created for purposes such as the study of the human movement, medical applications, or 3D computer animation. However both of them are mentioned here because, nowadays, they are also being used for action recognition. Finally, we also dedicate a place to mention datasets whose videos were recorded by thermal and infrared cameras. They can be useful to benchmark those action recognition algorithms that work with videos recorded in the infrared part of the electromagnetic spectrum. Table 7 summarizes the following information related to each dataset: the year of creation, the web page for downloading, the reference to the paper (if any) which describes

Table 7
Summary of characteristics that describe datasets dedicated to the recognition of specific human actions and other related. The datasets are listed in alphabetical order.

Name	Year	Web/description paper	Category
Biological Motion library	2006	[161]/[125]	MOCAP
Buffy pose classes package	2009	[73]/[51]	Pose and Gesture Recognition
Buffy Stickmen	2008	[72]/–	Pose and Gesture Recognition
Cambridge Hand Gesture	2007	[84]/[94]	Pose and Gesture Recognition
CANDELA	2004	[21]/–	Detecting abandoned objects
CASIA Gait Recognition A	2001	[57]/[221]	Gait analysis
CASIA Gait Recognition B	2005	[57]/[223]	Gait analysis
CASIA Gait Recognition C	2005	[57]/[191]	Infrared and Thermal
CMU-MMAC	2008	[102]/[101]	MOCAP
CMU MoBo	2001	[203]/[70]	MOCAP
CMU Graphics Lab Motion Capture	2004	[202]/–	MOCAP
DRINKING&SMOKING	2007	[107]/[112]	Pose and Gesture Recognition
Gait Based Human ID Challenge Problem	2001	[154]/–	Gait analysis
HDM05	2007	[145]/[136]	MOCAP
HMD	2011	[157]/[75]	MOCAP
HUMAN-EVA	2006	[201]/[185,186]	MOCAP
Human Identification at a Distance	2001	[64]/–	MOCAP
ICS Action	2003	[134]/–	MOCAP
IEMOCAP	2008	[155]/[31]	MOCAP
i-Lids (AVSS-2007)	2007	[3]/–	Detecting abandoned objects
Keck Gesture	2009	[116]/[122]	Pose and Gesture Recognition
Korea University Gesture	2006	[206]/[83]	MOCAP
LDB	2002	[154]/[211]	Gait analysis
MPII Cooking Activities	2012	[59]/[175]	ADL
Multiple cameras fall dataset	2010	[207]/[22]	Detecting falls, ADL
NATOPS	2011	[189]/[188]	Pose and Gesture Recognition
OTCBVS	2007	[208]/[45]	Infrared and Thermal
PETS_ICVS 2006	2003	[2]/–	Pose and Gesture Recognition
PETS 2006	2006	[209]/–	Detecting abandoned objects
PETS 2007	2007	[6]/–	Detecting abandoned objects
PETS 2009, Winter-PETS 09, PETS 2010	2009	[42,193]/[52,53]	Crowd Behavior
POETICON	2010	[56]/[164,218]	MOCAP
RVL-SLLL ASL	2006	[216]/–	Pose and Gesture Recognition
TUM Kitchen	2009	[137]/[194]	MOCAP, ADL
UCF Crowd Segmentation	2007	[151]/[12]	Crowd Behavior
UCF Tracking in High Density Crowds	2008	[152]/[13]	Crowd Behavior
UCF Web Dataset: Abnormal/Normal Crowds	2009	[128]/[129]	Crowd Behavior
URADL	2009	[153]/[131]	ADL
ViHASI	2008	[205]/[172]	MOCAP
WAR	2008	[146]/[229]	MOCAP

the dataset in full detail, and the classification category of each one. Evidence that these datasets have gained recent interest is the fact that all of them were created not before 2007. Specifically, a big number of MOCAP, pose and gesture datasets have been published (15 and 8 respectively).

5. Conclusions

There is a great number of datasets available for human action and activity recognition. Concretely, a total of 68 datasets are reported in this survey. Although all of them are included in a time window ranging from 2001 to 2012, around 80% of the datasets described here were created from 2005 onward. A coarse first classification can be done attending to the variety of the repertoire of actions stored: those datasets devoted to sets of heterogeneous human actions (28 datasets belonging to this class are described in this survey), those ones focused on specific actions (23 datasets are mentioned) and others (17). Among the former, other classifications are proposed, considering different types of features, such as type of background (controlled or not), type of ground truth, type of interaction (none, person-to-person and person-to-object), type of actor engaged in the action (professional-amateur or ordinary people), number of views (mono or multi-view), and if the camera is moving or not.

Although the ground truth provided in older datasets is limited to simple manual annotation, the majority of modern datasets give high quality ground truth. The popularization of standard description languages based on XML such as ViPER or CVML have simplified the annotation process and led to richer descriptions of what is happening in each frame (see BEHAVE, CAVIAR, ETISEO and ViRAT datasets). There are also repositories, such as ViSOR, which contain a collection of different action datasets and is continuously updating. Other important subject to take into account is the proliferation of MOCAP databases. They are very attractive for action recognition because provide parametric data which allow us to build full 3D models.

Although new datasets devoted to sets of heterogeneous action have been created in recent years (among the datasets reported in this survey, nine new datasets were created from 2010 to 2012), it is also observed a great proliferation of new datasets dedicated to sets of more specific actions.

In relation to the impact in the scientific community of the different datasets associated to the recognition of heterogeneous actions, we can cite the most used dataset in each one of the categories described in column 2 of Table 1. In absolute terms, Weizmann, KTH and CAVIAR are the most popular. The former two belong to the category of “unrealistic action analysis” and their great success could be explained by two reasons: they are one the first databases to appear in the history of action datasets and, moreover, the challenge that they pose is not high. On the other hand, the success of CAVIAR, which belongs to the category of “realistic action analysis”, could be explained by the fact that it was the first dataset recorded in environments more complex, where background and illumination are not controlled. In the same category as CAVIAR, but in a slightly more complex context (videos are collected directly from the web), one of the most cited datasets is UCF-Sports. In the following category, “interaction analysis”, UT-Interaction is the most used dataset in tasks related to the automatic recognition of different types of interactions. In the category of “multi-view analysis”, IXMAS and VideoWeb datasets are the most cited in its respective subcategories, “indoor” and “outdoor”. Finally, in the category of “repositories”, ViSOR is far more popular than VIRAT but, in this case, it is necessary to take in count that the first dataset is much older than the second one.

Finally, this survey tries to cover the lack of a complete description of the most important public datasets for video-based human activity recognition and to guide the researchers in the election of the most suitable dataset for benchmarking their algorithms. A comparison of all datasets, focusing on different practical issues such as ground truth data, type of scenes, number of actions and actors, or references of published papers using these datasets, is provided.

Acknowledgements

This work was supported by the Spanish Ministerio de Economía y Competitividad/FEDER under Projects TIN2010-20845-C03-01 and TIN2010-20845-C03-02.

References

[1] Caviar: context aware vision using image-based active recognition, November 2011. <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm>.

[2] PETS-ICVS 2003 datasets, December 2011. <http://www.cvg.cs.rdg.ac.uk/PETS-ICVS/pets-icvs-db.html>.

[3] Advanced video and signal based surveillance 2007, datasets, December 2011. <http://www.eecs.qmul.ac.uk/ andrea/avss2007d.html>.

[4] 20th international conference on pattern recognition, February 2012. <http://www.icpr2010.org/ICPR10ProgramAtAGlanceWeb.html>.

[5] PETS-ECCV 2004, sixth IEEE international workshop on performance evaluation of tracking and surveillance, January 2012. <http://www-prima.inrialpes.fr/PETS04/index.html>.

[6] Tenth IEEE international workshop on performance evaluation of tracking and surveillance, January 2012. <http://pets2007.net/>.

[7] J. Acevedo-Rodríguez, S. Maldonado-Bascón, R. López-Sastre, P. Gil-Jiménez, A. Fernández-Caballero, Clustering of trajectories in video surveillance using growing neural gas, Lecture Notes in Computer Science 6686 (2011) 461–470.

[8] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Computer Vision and Image Understanding 73 (1999) 428–440.

[9] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: a review, ACM Computing Surveys (CSUR) 43 (3) (2011).

[10] J.K. Aggarwal, Q. Cai, W. Liao, B. Sabata, Nonrigid motion analysis: articulated and elastic motion, Computer Vision and Image Understanding 70 (1998) 142–156.

[11] M.A.R. Ahad, J. Tan, H. Kim, S. Ishikawa. Action dataset – a survey, in: 2011 Proceedings of SICE Annual Conference (SICE), September 2011, pp. 1650–1655.

[12] S. Ali, M. Shah, A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[13] S. Ali, M. Shah, Floor fields for tracking in high density crowd scenes, in: The 10th European Conference on Computer Vision (EECV), 2008.

[14] M.R. Amer, S. Todorovic, A chains model for localizing participants of group activities in videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 786–793. <www.scopus.com>.

[15] E. Andrade, S. Blunsden, R. Fisher, Simulation of crowd problems for computer vision, in: First International Workshop on Crowd Simulation (V-CROWDS '05), 2005.

[16] E. Andrade, S. Blunsden, R. Fisher, Performance analysis of event detection models in crowded scenes, in: Workshop on Towards Robust Visual Surveillance Techniques and Systems at Visual Information Engineering, 2006.

[17] E. Andrade, S. Blunsden, R. Fisher, Detection of emergency events in crowded scenes, in: IEE Int. Symp. on Imaging for Crime Detection and Prevention (ICDP 2006), 2006.

[18] E.L. Andrade, S. Blunsden, R.B. Fisher, Characterisation of optical flow anomalies in pedestrian traffic, in: IEE Int. Symp. on Imaging for Crime Detection and Prevention (ICDP 2005), 2005.

[19] E.L. Andrade, S. Blunsden, R.B. Fisher, Hidden markov models for optical flow analysis in crowds, in: Int. Conf. on Pat. Recog., 2006.

[20] E.L. Andrade, S. Blunsden, R.B. Fisher, Modeling crowd scenes for event detection, in: Int. Conf. on Pat. Recog., 2006.

[21] Multitel A.S.B.L. Abandoned object dataset, December 2011. <http://www.multitel.be/ va/candela/>.

[22] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, J. Rousseau. Multiple cameras fall dataset, Technical report, Universite de Montreal, 2010.

[23] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, G. Serra, Recognizing human actions by fusing spatio-temporal appearance and motion descriptors, in: Proc. of IEEE International Conference on Image Processing (ICIP), IEEE Computer Society, Cairo, Egypt, 2009. <http://www.micc.unifi.it/serra/wp-content/uploads/2009/07/serra-icip09.pdf>.

[24] D. Baltieri, R. Vezzani, R. Cucchiara. 3D body model construction and matching for real time people re-identification, in: Proceedings of Eurographics Italian Chapter Conference 2010 (EG-IT 2010), Genova, Italy, November 2010.

[25] S. Bhattacharya, R. Sukthankar, R. Jin, M. Shah, A probabilistic representation for efficient large scale visual recognition tasks, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 2593–2600.

[26] J. Bins, T. List, R.B. Fisher, D. Tweed, An intelligent and task-independent controller for video sequence analysis, in: IEEE Int. Workshop on Computer Architecture for Machine, Perception (CAMP'05), 2005.

[27] S. Blunsden, R. Fisher, Recognition of coordinated multi agent activities: the individual vs. the group, in: Workshop on Computer Vision Based Analysis in Sport Environments (CVBASE), 2006.

[28] S. Blunsden, E. Andrade, R. Fisher, Non parametric classification of human interaction, in: Proc. 3rd Iberian Conf. on Pattern Recog. and Image, Analysis, 2007.

[29] H. Boyraz, M.F. Tappen, R. Sukthankar, Localizing actions through sequential 2d video projections, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2011.

[30] M. Bregonzio, S. Gong, T. Xiang, Recognising action as clouds of space-time interest points, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, 2009.

[31] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, S. Narayanan, Iemocap: interactive emotional dyadic motion capture database, Journal of Language Resources and Evaluation 42 (4) (2008) 335–359.

[32] H. Buxton, Learning and understanding dynamic scene activity: a review, Image and Vision Computing 21 (2003) 125–136. <http://www.sciencedirect.com/science/article/pii/S0262885602001270>.

[33] A. Fernández-Caballero, J.C. Castillo, J.M. Rodríguez-Sánchez, Human activity monitoring by local and global finite state machines, Expert Systems with Applications 39 (8) (2012) 6982–6993.

[34] L. Cao, Z. Liu, T.S. Huang, Cross-dataset action detection, in: CVPR, 2010.

[35] A. Castrodad, G. Sapiro, Sparse modeling of human actions from motion imagery, International Journal of Computer Vision, 0920-5691 100 (2012) 1–15. http://dx.doi.org/10.1007/s11263-012-0534-7. 10.1007/s11263-012-0534-7.

[36] C. Cedras, M. Shah, Motion-based recognition: a survey, Image and Vision Computing 13 (1995) 129–155.

[37] A.A. Chaaraoui, P. Climent-Perez, F. Florez-Revuelta, A review on vision techniques applied to human behaviour analysis for ambient-assisted living, Expert Systems with Applications, (0), 2012. ISSN: 0957–4174. <http://www.sciencedirect.com/science/article/pii/S0957417412004757>.

[38] Chia-Chih Chen, J.K. Aggarwal, Recognizing human action from a far field of view, in: IEEE Workshop on Motion and Video Computing (WMVC), 2009.

[39] Chia-Chih Chen, M.S. Ryoo, J.K. Aggarwal, UT-Tower Dataset: Aerial View Activity Classification Challenge, January 2012. <http://cvrc.ece.utexas.edu/SDHA2010/AerialViewActivity.html>.

[40] J. Chen, G. Zhao, V. Kellokumpu, M. Pietikainen, Combining sparse and dense descriptors with temporal semantic structures for robust human action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1524–1531. <www.scopus.com>.

[41] S.Y. Cho, H.R. Byun, Human activity recognition using overlapping multi-feature descriptor, Electronics Letters 47 (23) (2011) 1275–1277. <www.scopus.com>.

[42] Reading University Computational Vision Group, School of Systems Engineering, Pets 2009 benchmark data, December 2011. <http://www.cvg.rdg.ac.uk/PETS2009/a.html>.

[43] J.J. Corso, S. Sadanand, Action bank: a high-level representation of activity in video, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1234–1241. ISSN: 1063-6919.

[44] R. Cucchiara, C. Grana, A. Prati, R. Vezzani, Probabilistic posture classification for human behaviour analysis, IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans 35 (1) (January 2005) 42–54.

[45] J. Davis, M. Keck, A two-stage approach to person detection in thermal imagery, in: Workshop on Applications of Computer Vision, 2005.

[46] P. Dollár, Piotr's image and video matlab toolbox, January 2012. <http://vision.ucsd.edu/ pdollar/toolbox/doc/>.

[47] E. Erdem, S. Dubuisson, I. Bloch, Visual tracking by fusing multiple cues with context-sensitive reliabilities, Pattern Recognition 45 (5) (May 2012) 1948–1959.

[48] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognition, 0031-3203 36 (1) (2003) 259–275. <http://www.sciencedirect.com/science/article/pii/S0031320302000523>.

[49] A. Fernández-Caballero, J.C. Castillo, J.M. Rodríguez-Sánchez, A proposal for local and global human activities identification, Lecture Notes in Computer Science 6169 (2010) 78–87.

[50] V. Ferrari, M. Marin-Jimenez, A. Zisserman, Progressive search space reduction for human pose estimation, in: 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008. <www.scopus.com> Cited By (since 1996): 63.

[51] V. Ferrari, M. Marin-Jimenez, A. Zisserman, Pose search: retrieving people using their pose, in: Proceedings of the IEEE Conference on Computer Vision and, Pattern Recognition, 2009.

[52] J. Ferryman, A. Ellis, Pets 2010: Dataset and challenge, in: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2010.

[53] J. Ferryman, A. Shahrokni, Pets 2009: Dataset and challenge, in: 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009.

[54] R. Fisher, Behave: Computer-assisted prescreening of video streams for unusual activities, November 2011. <http://homepages.inf.ed.ac.uk/rbf/BEHAVE/>.

[55] R.B. Fisher, Pets04 surveillance ground truth data set, in: Proc. Sixth IEEE Int. Work. on Performance Evaluation of Tracking and Surveillance (PETS04), 2004.

[56] Max Planck Institute for Biological Cybernetics. Popeticon corpus, Janaury 2012. <http://poeticoncorpus.kyb.mpg.de/>

[57] Center for Biometrics and Security Research, Casia gait database, 2011. <http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp>.

[58] Center for Biometrics and Security Research, Casia action database for recognition, November 2011. <http://www.cbsr.ia.ac.cn/english/Action%20Databases%20EN.asp>.

[59] Max Planck Institute for Informatics, Mpii cooking activities dataset, September 2012. <https://www.d2.mpi-inf.mpg.de/mpii-cooking>.

[60] J. Gall, A. Yao, N. Razavi, L. Van Gool, V. Lempitsky, Hough forests for object detection, tracking, and action recognition, Transactions on Pattern Analysis and Machine Intelligence, 0162-8828 33 (11) (2011) 2188–2202. http://dx.doi.org/10.1109/TPAMI.2011.70.

[61] U. Gaur, Y. Zhu, B. Song, A. Roy-Chowdhury, A string of feature graphs model for recognition of complex activities in natural videos, in: IEEE Conf. on Computer Vision, 2011.

[62] G. Denina, B. Bhanu, H. Nguyen, C. Ding, A. Kamal, C. Ravishankar, A. Roy-Chowdhury, A. Ivers, B. Varda, Distributed Video Sensor Networks, Springer, 2011 (Chapter: VideoWeb Dataset for Multi-camera Activities and Non-verbal Communication).

[63] P. Geetha, V. Narayanan, A survey of content-based video retrieval, Computer Science 4 (6) (2008) 474–486. <http://www.thescipub.com/abstract/>.

[64] GVU Center/College of Computing Georgia Tech, Human identification at a distance, January 2012. <http://www.cc.gatech.edu/cpl/projects/hid/>.

[65] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, I. Pitas, The i3dpost multi-view and 3d human action/interaction, in: CVMP, 2009, pp. 159–168.

[66] N. Gkalelis, N. Nikolaidis, I. Pitas, View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation, in: ICME, 2009.

[67] J. Gonzalez, T.B. Moeslund, L. Wang, Semantic understanding of human behaviors in image sequences: from video-surveillance to video-hermeneutics, Computer Vision and Image Understanding, 1077-3142 116 (3) (2012) 305–306 (Special issue on Semantic Understanding of Human Behaviors in Image Sequences) <http://www.sciencedirect.com/science/article/pii/S1077314212000100>.

[68] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, Transactions on Pattern Analysis and Machine Intelligence 29 (12) (2007) 2247–2253. <http://www.wisdom.weizmann.ac.il/ vision/SpaceTimeActions.html>.

[69] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Weizmman actions as space-time shapes, November 2011. <http://www.wisdom.weizmann.ac.il/vision/SpaceTimeActions.html>.

[70] R. Gross, J. Shi, The cmu motion of body (mobo) database, Technical report, Robotics Institute, 2001.

[71] Video Computing Group. Videoweb dataset, January 2012. <http://www.ee.ucr.edu/ amitrc/vwdata.php>.

[72] Visual Geometry Group, Buffy stickmen v3.0: Annotated data and evaluation routines for 2d human pose estimation, January 2012. <http://www.robots.ox.ac.uk/ vgg/data/stickmen/index.html>.

[73] Visual Geometry Group. Buffy pose classes, January 2012. <http://www.robots.ox.ac.uk/ vgg/data/buffyposeclasses/index.html>.

[74] Visual Geometry Group. Tv human interactions dataset, Janury 2012. <http://www.robots.ox.ac.uk/ vgg/data/tvhumaninteractions/index.html>.

[75] G. Guerra-Filho, A. Biswas, The human motion database: a cognitive and parametric sampling of human motion, in: 9th IEEE Conference on Automatic Face and Gesture Recognition (FG), 2011.

[76] N. Haering, P.L. Venetianer, A. Lipton, The evolution of video surveillance: an overview, Machine Vision and Applications 19 (2008) 279–290.

[77] D. Hall, Automatic parameter regulation for a tracking system with an auto-critical function, in: IEEE Int. Workshop on Computer Architecture for Machine, Perception (CAMP'05), 2005.

[78] D. Hall, J. Nascimento, P. Ribeiro, E. Andrade, P. Moreno, S. Pesnel, T. List, R. Emonet, R.B. Fisher, J. Santos-Victor, J.L. Crowley, Comparison of target detection algorithms using adaptive background models, in: Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), 2005.

[79] M. Holte, T. Moeslund, N. Nikolaidis, I. Pitas, 3D human action recognition for multi-view camera systems, in: 3DIMPVT, 2011.

[80] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 34 (2004) 334–352.

[81] K. Huang, D. Tao, Y. Yuan, X. Li, T. Tan, View-independent behavior analysis, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 1083-4419 39 (4) (2009) 1028–1035.

[82] K. Huang, S. Wang, T. Tan, S.J. Maybank, Human behavior analysis based on a new motion descriptor, IEEE Transactions on Circuits and Systems for Video Technology, 1051-8215 19 (12) (2009) 1830–1840.

[83] B.-W. Hwang, S. Kim, S.-W. Lee, A full-body gesture database for automatic gesture recognition, in: 7th IEEE International Conference on Automatic Face and Gesture Recognition, 2006, pp. 243–248.

[84] London Imperial Collage, Cambridge-gesture data base, January 2012. <http://www.iis.ee.ic.ac.uk/ tkkim/gesdb.htm>.

[85] INRIA, Etiseo video understanding evaluation, December 2011. <http://www-sop.inria.fr/orion/ETISEO/index.htm>.

[86] INRIA, Inria xmas motion acquisition sequences (ixmas), November 2011. <http://4drepository.inrialpes.fr/public/viewgroup/6>.

[87] A. Iosifidis, A. Tefas, N. Nikolaidis, I. Pitas, Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis, Computer Vision and Image Understanding, 1077-3142 116 (3) (2012) 347–360 (Special issue on Semantic Understanding of Human Behaviors in Image Sequences) <http://www.sciencedirect.com/science/article/pii/S1077314211002074>.

[88] A. Jaimes, N. Sebe, Multimodal human-computer interaction: a survey, Computer Vision and Image Understanding 108 (2007) 116–134.

[89] X. Ji, H. Liu, Advances in view-invariant human motion analysis: a review, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 1094-6977 40 (1) (2010) 13–24.

[90] S. Jones, L. Shao, J. Zhang, Y. Liu, Relevance feedback for real-world human action retrieval, Pattern Recognition Letters, 0167-8655 33 (4) (2012) 446–452. <http://www.sciencedirect.com/science/article/pii/S016786551100136X>.

[91] P.M. Jorge, A.J. Abrantes, J.S. Marques, On-line tracking groups of pedestrians with bayesian networks, in: International Workshop on Performance Evaluation for tracking and Surveillance (PETS, ECCV), 2004.

[92] V. Karavasilis, C. Nikou, A. Likas, Visual tracking using the earth mover's distance between gaussian mixtures and kalman filtering, Image and Vision Computing 29 (5) (April 2011) 295–305.

[93] I.S. Kim, H.S. Choi, K.M. Yi, J.Y. Choi, S.G. Kong, Intelligent visual surveillance – a survey, International Journal of Control, Automation, and Systems, 1598-6446 8 (2010) 926–939. <http://dx.doi.org/10.1007/s12555-010-0501-4>.

[94] T.-K. Kim, S.-F. Wong, R. Cipolla, Tensor canonical correlation analysis for action classification, in: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[95] Kitware, Kitware, inc., January 2012. <http://www.kitware.com/>.

[96] Kitware, Virat video dataset, January 2012. <http://www.viratdata.org/>.

[97] T. Ko, A survey on behavior analysis in video surveillance for homeland security applications, in: 37th IEEE Applied Imagery Pattern Recognition Workshop, 2008 (AIPR '08), October 2008, pp. 1–8.

[98] S.G. Kong, J. Heo, B.R. Abidi, J. Paik, M.A. Abidi, Recent advances in visual and infrared face recognition – a review, Computer Vision and Image Understanding, 1077-3142 97 (1) (2005) 103–135. <http://www.sciencedirect.com/science/article/pii/S1077314204000451>.

[99] V. Kruger, D. Kragic, A. Ude, C. Geib, The meaning of action: a review on action recognition and mapping, Advanced Robotics 21 (13) (2007) 1473–1501. <http://www.ingentaconnect.com/content/vsp/arb/2007/00000021/00000013/ar t00002>.

[100] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, Hmdb: a large video database for human motion recognition, in: ICCV, 2011.

[101] F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, J. Macey, Guide to the carnegie mellon university multimodal activity (cmu-mmac) database, Technical Report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University, 2009.

[102] Carnegie Mellon's Motion Capture Lab, Cmu multi-modal activity database, 2012 January. <http://kitchen.cs.cmu.edu/>.

[103] Serre lab, Hmdb: a large video database for human motion recognition, November 2011. <http://serre-lab.clps.brown.edu/resources/HMDB/index.htm>.

[104] Language and Media Processing Laboratory, Viper: The video performance evaluation resource, November 2011. <http://viper-toolkit.sourceforge.net/>.

[105] I. Laptev, Local Spatio-Temporal Image Features for Motion Interpretation, PhD thesis, Computational Vision and Active Perception Laboratory (CVAP), NADA, KTH, Stockholm, 2004.

[106] I. Laptev, Hollywood2: human actions and scenes dataset, November 2011. <http://www.irisa.fr/vista/actions/hollywood2/>.

[107] I. Laptev, Irisa download data/software, December 2011. <http://www.irisa.fr/vista/Equipe/People/Laptev/download.html>.

[108] I. Laptev, B. Caputo, Recognition of human actions, November 2011. <http://www.nada.kth.se/cvap/actions/>.

[109] I. Laptev, T. Lindeberg, Space-time interest points, in: ICCV'03, 2003.

[110] I. Laptev, T. Lindeberg, Local descriptors for spatio-temporal recognition, in: ECCV Workshop Spatial Coherence for Visual Motion, Analysis, 2004.

[111] I. Laptev, T. Lindeberg, Velocity adaptation of space-time interest points, in: ICPR'04, 2004.

[112] I. Laptev, P. Perez, Retrieving actions in movies, in: ICCV, 2007.

[113] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[114] J.T. Lee, C.-C. Chen, J.K. Aggarwal, Recognizing human–vehicle interactions from aerial video without training, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2011, pp. 53–60.

[115] B. Li, M. Ayazoglu, T. Mao, O.I. Camps, M. Sznaier, Activity recognition using dynamic subspace angles, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011, pp. 3193–3200. <www.scopus.com>.

[116] Z. Lin, Z. Jiang, L.S. Davis, Recognizing actions by shape-motion prototype trees, January 2012. <http://www.umiacs.umd.edu/ zhuolin/Keckgesturedataset.html>.

[117] T. List, R.B. Fisher, Cvml – an xml-based computer vision markup language, in: Proceedings of the 17th International Conference on, Pattern Recognition, 2004.

[118] T. List, J. Bins, R.B. Fisher, D. Tweed, A plug-and-play architecture for cognitive video stream analysis, in: IEEE Int. Workshop on Computer Architecture for Machine, Perception (CAMP'05), 2005.

[119] T. List, J. Bins, R.B. Fisher, D. Tweed, K.R. Thorisson, Two approaches to a plug-and-play vision architecture – caviar and psyclone, in: AAAI05 workshop on Modular Construction of Human-like, Intelligence, 2005.

[120] T. List, J. Bins, J. Vazquez, R.B. Fisher, Performance evaluating the evaluator, in: Proc. 2nd Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), 2005.

[121] J. Liu, Y. Yang, M. Shah, Learning semantic visual vocabularies using diffusion distance, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[122] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos in the wild, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[123] J. Liu, B. Kuipers, S. Savarese, Recognizing human actions by attributes, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3337–3344.

[124] J. Liu, Y. Yang, I. Saleemi, M. Shah, Learning semantic features for action recognition via diffusion maps, Computer Vision and Image Understanding 116 (3) (2012) 361–377.

[125] Y. Ma, H. Paterson, F. Pollick, A motion capture library for the study of identity, gender, and emotion perception from biological motion, Behavior Research Methods 38 (2006) 134–141.

[126] M. Marszaek, I. Laptev, C. Schmid, Actions in contex, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[127] F. Martinez-Contreras, C. Orrite-Urunuela, E. Herrero-Jaraba, H. Ragheb, S.A. Velastin, Recognizing human actions using silhouette-based hmm, in: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '09), 2009, pp. 43–48.

[128] R. Mehran, Abnormal crowd behavior detection using social force model, January 2012. <http://www.cs.ucf.edu/ ramin/?pageid=24>.

[129] R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[130] M. Merler, B. Huang, L. Xie, G. Hua, A. NAstev, Semantic model vectors for complex video event recognition, IEEE Transactions on Multimedia, 2011, pp. 1–14.

[131] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: Proceedings of the Twelfth IEEE International Conference on Computer Vision (ICCV), 2009.

[132] T.B. Moeslund, A. Hilton, V. Kruger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding 104 (2006) 90–126.

[133] T.B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Computer Vision and Image Understanding 81 (2001) 231–268.

[134] T. Mori, Ics action database, January 2012. <http://www.ics.t.u-tokyo.ac.jp/action/>.

[135] S. Mukherjee, S.K. Biswas, D.P. Mukherjee, Recognizing interaction between human performers using'key pose doublet', in: MM'11 – Proceedings of the 2011 ACM Multimedia Conference and Co-Located Workshops, 2011, pp. 1329–1332. <www.scopus.com>.

[136] M. Muller, T. Roder, M. Clausen, B. Eberhardt, B. Kruger, A. Weber, Documentation: Mocap database hdm05, Technical Report CG-2007-2, Universität Bonn, 2007.

[137] Technische Universitat Munchen, Tum kitchen data set, January 2012. <http://ias.in.tum.de/software/kitchen-activity-data>.

[138] B.M. Nair, V.K. Asari, Time invariant gesture recognition by modeling body posture space, LNAI of Lecture Notes in Computer Science, vol. 7345 (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012. <www.scopus.com>.

[139] J.C. Nascimento, M.A.T. Figueiredo, J.S. Marques, Recognizing human activities using space dependent switched dynamical models, in: IEEE Int. Conf. on Image Processing, 2005.

[140] J.C. Nascimento, M.A.T. Figueiredo, J.S. Marques, Motion segmentation for activity surveillance, in: ISR Workshop on Systems, Decision and Control Robotic Monitoring and Surveillance, 2005.

[141] J.C. Nascimento, M.A.T. Figueiredo, J.S. Marques, Segmentation and classification of human activities, in: Workshop on Human Activity Recognition and Modeling (HAREM 2005 – in conjunction with BMVC 2005), 2005.

[142] A.T. Nghiem, F. Bremond, M. Thonnat, R. Ma, New evaluation approach for video processing algorithms, in: WMVC 2007 IEEE Workshop on Motion and Video, Computing, 2007.

[143] A.T. Nghiem, F. Bremond, M. Thonnat, V. Valentin, Etiseo, performance evaluation for video surveillance systems, in: Proceedings of AVSS 2007, 2007.

[144] J.C. Niebles, C.-W. Chen, L. Fei-Fei, Modeling temporal structure of decomposable motion segments for activity classification, in: 11th European Conference on Computer Vision (ECCV), 2010.

[145] University of Bonn, Motion capture database hdm05, January 2012. <http://www.mpi-inf.mpg.de/resources/HDM05/>.

[146] University of California, Distributed human action recognition via wearable motion sensor networks, January 2012. <http://www.eecs.berkeley.edu/yang/software/WAR/>.

[147] University of Central Florida, UCF aerial camera, rooftop camera and ground camera dataset, November 2011. <http://vision.eecs.ucf.edu/data/UCF-ARG.html>.

[148] University of Central Florida, UCF aerial action dataset, November 2011. <http://server.cs.ucf.edu/ vision/aerial/index.html>.

[149] University of Central Florida, UCF youtube action dataset, November 2011. <http://www.cs.ucf.edu/ liujg/YouTubeActiondataset.html>.

[150] University of Central Florida, UCF sports action dataset, February 2012. <http://vision.eecs.ucf.edu/datasetsActions.html>.

[151] University of Central Florida, Crowd segmentation data set, February 2012. <http://vision.eecs.ucf.edu/datasetsCrowd.html>.

[152] University of Central Florida, Tracking in high density crowds data set, February 2012. <http://vision.eecs.ucf.edu/datasetsTracking.html>.

[153] University of Rochester, Activities of daily living dataset, January 2012. <http://www.cs.rochester.edu/ rmessing/uradl/>.

[154] University of South Florida, National Institute of Standards, Technology, and University of Notre Dame, Human id gait challenge problem, January 2012. <http://marathon.csee.usf.edu/GaitBaseline/>.

[155] University of Southern California, The interactive emotional dyadic motion capture (iemocap) database, January 2012. <http://sail.usc.edu/iemocap/index.html>.

[156] University of Surrey and CERTH-ITI, i3dpost multi-view human action datasets, January 2012. <http://kahlan.eps.surrey.ac.uk/i3dpostaction/>.

[157] University of Texas at Arlington, Human motion database, January 2012. <http://smile.uta.edu/hmd/>.

[158] The University of Texas at Austin, Icpr 2010 contest on semantic description of human activities, January 2012. <http://cvrc.ece.utexas.edu/SDHA2010/>.

[159] The Imagelab Laboratory of the University of Modena and Reggio Emilia, Visor (video surveillance online repository), November 2011. <http://www.openvisor.org/index.asp>.

[160] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J.T. Lee, S. Mukherjee, J.K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, M. Desai, A large-scale benchmark dataset for event recognition in surveillance video, in: Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR), 2011.

[161] PACO, Perception action and cognition, January 2012. <http://paco.psy.gla.ac.uk/index.php?option=comjdownloads&view=viewcategories&Itemid=62>.

[162] M. Pantic, A. Pentland, A. Nijholt, T. Huang, Human computing and machine understanding of human behavior: a survey, in: Proceedings of the 8th International Conference on Multimodal interfaces, ICMI '06, ACM, New York, NY, USA, 2006, ISBN 1-59593-541-X, pp. 239–248. http://doi.acm.org/10.1145/1180995.1181044.

[163] G. Paolacci, J. Chandler, P.G. Ipeirotis, Running experiments on amazon mechanical turk, Judgment and Decision Making 5 (5) (2010) 411–419.

[164] K. Pastra, C. Wallraven, M. Schultze, A. Vatakis, K. Kaulard, The poeticon corpus: capturing language use and sensorimotor experience in everyday interaction, in: Seventh conference on International, Language Resources and Evaluation (LREC'10), 2010.

[165] A. Patron-Perez, M. Marszalek, A. Zisserman, I. Reid, High five: recognising human interactions in tv shows, in: Proceedings of the British Machine Vision Conference, 2010.

[166] A. Patron-Perez, M. Marszalek, I. Reid, A. Zisserman, Structured learning of human interactions in tv shows, IEEE Transactions on Pattern Analysis and Machine Intelligence 99 (2012).

[167] V.I. Pavlovic, R. Sharma, T.S. Huang, Visual interpretation of hand gestures for human-computer interaction: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence, 0162-8828 19 (7) (1997) 677–695.

[168] F. Pla, P. Ribeiro, J. Santos-Victor, A. Bernardino, Extracting motion features for visual human activity representation, in: Proc. IBPRIA – 2nd Iberian Conference on, Pattern Recognition and Image Analysis, 2005.

[169] O.P. Popoola, K. Wang, Video-based abnormal human behavior recognition review, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 1094-6977 PP (99) (2012) 1–14.

[170] R. Poppe, Vision-based human motion analysis: an overview, Computer Vision and Image Understanding 108 (2007) 4–18.

[171] R. Poppe, A survey on vision-based human action recognition, Image and Vision Computing 28 (2010) 976–990.

[172] H. Ragheb, S. Velastin, P. Remagnino, T. Ellis, Vihasi: virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods, in: Workshop on Activity Monitoring by Multi-Camera Surveillance Systems, 2008.

[173] P. Ribeiro, J. Santos-Victor, Human activities recognition from video: modeling, feature selection and classification architecture, in: Workshop on Human Activity Recognition and Modeling (HAREM 2005 – in conjunction with BMVC 2005), 2005.

[174] M.D. Rodriguez, J. Ahmed, M. Shah, Action mach: a spatio-temporal maximum average correlation height filter for action recognition, in: Proceedings of IEEE International Conference on Computer Vision and, Pattern Recognition, 2008.

[175] M. Rohrbach, S. Amin, M. Andriluka, B. Schiele, A database for fine grained activity detection of cooking activities, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2012.

[176] M. Del Rose, C. Wagner, Survey on classifying human actions through visual sensors, Artificial Intelligence Review, 0269-2821 37 (2012) 301–311. http://dx.doi.org/10.1007/s10462-011-9232-z. 10.1007/s10462-011-9232-z.

[177] M.S. Ryoo, Human activity prediction: early recognition of ongoing activities from streaming videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1036–1043. <www.scopus.com>.

[178] M.S. Ryoo, J.K. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, in: IEEE International Conference on Computer Vision (ICCV), 2009.

[179] M.S. Ryoo, J.K. Aggarwal, UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA), January 2012. <http://cvrc.ece.utexas.edu/SDHA2010/HumanInteraction.html>.

[180] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: International Conference on Pattern Recognition, 2004, pp. 32–36. <http://www.nada.kth.se/cvap/actions/>

[181] M. Selmi, M. El Yacoubi, B. Dorizzi, On the sensitivity of spatio-temporal interest points to person identity, in: Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation, 2012, pp. 69–72. <www.scopus.com>.

[182] H.J. Seo, P. Milanfar, Action recognition from one example, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (2011) 867–882.

[183] R.J. Sethi, A.K. Roy-Chowdhury, Individuals, groups, and crowds: modeling complex, multi-object behaviour in phase space, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1502–1509. <www.scopus.com>.

[184] A.M. Sharma, K.S. Venkatesh, A. Mukerjee, Human pose estimation in surveillance videos using temporal continuity on static pose, in: ICIIP 2011 – Proceedings: 2011 International Conference on Image Information Processing, 2011. <www.scopus.com>.

[185] L. Sigal, M.J. Black, Humaneva: synchronized video and motion capture dataset for evaluation of articulated human motion, Technical report, Brown University, 2006.

[186] L. Sigal, A. Balan, M.J. Black, Humaneva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, International Journal of Computer Vision 87 (2010).

[187] S. Singh, S.A. Velastin, H. Ragheb, Muhavi: a multicamera human action video dataset for the evaluation of action recognition methods, in: 2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems (AMMCSS), 2010.

[188] Y. Song, D. Demirdjian, R. Davis, Tracking body and hands for gesture recognition: Natops aircraft handling signals database, in: Proceedings of the 9th IEEE International Conference on Automatic Face and Gesture Recognition, 2011.

[189] Y. Song, D. Demirdjian, R. Davis, Natops aircraft handling signals database, February 2012. <http://groups.csail.mit.edu/mug/natops/>.

[190] J. Starck, A. Hilton, Surface capture for performance based animation, IEEE Computer Graphics and Applications 27 (2007) 21–31.

[191] D. Tan, K. Huang, S. Yu, T. Tan, Efficient night gait recognition based on template matching, in: International Conference on Pattern Recognition (ICPR06), 2006.

[192] K. Tang, L. Fei-Fei, D. Koller, Learning latent temporal structure for complex event detection, in: Computer Vision and Pattern Recognition (CVPR), June 2012, pp. 1250–1257.

[193] IEEE Computer Society (PAMI TC) and IEEE Signal Processing Society (IVMSP TC), in: 7th IEEE International Conference on Advanced Video and Signal-based Surveillance, February 2012. <http://www.avss2010.org/>.

[194] M. Tenorth, J. Bandouch, M. Beetz, The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition, in: IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), in conjunction with ICCV2009, 2009.

[195] T.H. Thi, J. Zhang, L. Cheng, L. Wang, S. Satoh, Human action recognition and localization in video using structured learning of local space-time features, in: Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2010.

[196] D. Tran, A. Sorokin, Human activity recognition with metric learning, in: ECCV08, 2008.

[197] D. Tran, A. Sorokin, D. Forsyth, Human activity recognition with metric learning, January 2012. <http://vision.cs.uiuc.edu/projects/activity/>.

[198] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Transactions on Circuits and Systems for Video Technology, 1051-8215 18 (11) (2008) 1473–1488.

[199] P. Turaga, R. Chellappa, A. Veeraraghavan, Advances in video-based human activity analysis: challenges and approaches, in: Marvin V. Zelkowitz (Ed.), Advances in Computers, Advances in Computers, vol. 80, Elsevier, 2010, pp. 237–290.

[200] D. Tweed, W. Fang, R. Fisher, J. Bins, T. List, Exploring techniques for behaviour recognition via the caviar modular vision framework, in: Workshop on Human Activity Recognition and Modeling, 2005.

[201] Brown University, Humaneva: synchronized video and motion capture dataset for evaluation of articulated human motion, November 2011. <http://vision.cs.brown.edu/humaneva/index.html>.

[202] Carnegie Mellon University, Cmu graphics lab motion capture database, December 2011. <http://mocap.cs.cmu.edu/>.

[203] Carnegie Mellon University, The cmu motion of body (mobo) database, January 2012. <http://www.ri.cmu.edu/publicationview.html?pubid=3904>.

[204] Kingston University, Muhavi: Multicamera human action video data, November 2011. <http://dipersec.king.ac.uk/MuHAVi-MAS/>.

[205] Kingston University, Vihasi: virtual human action silhouette data, November 2011. <http://dipersec.king.ac.uk/VIHASI/>.

[206] Korea University, Full-body gesture database, January 2012. <http://gesturedb.korea.ac.kr/index.html>.

[207] Montreal University, Multiple cameras fall dataset, December 2011. <http://www.iro.umontreal.ca/ labimage/Dataset/>.

[208] Oklahoma State University, Otcbvs benchmark dataset collection, November 2011. <http://www.cse.ohio-state.edu/otcbvs-bench/>.

[209] Reading University, Ninth IEEE International workshop on performance evaluation of tracking and surveillance, November 2011. <http://www.cvg.rdg.ac.uk/PETS2006/data.html>.

[210] Standford University, Olympic sports dataset, January 2012. <http://vision.stanford.edu/Datasets/OlympicSports/>.

[211] G.V. Veres, L. Gordon, J.N. Carter, M.S. Nixon, What image information is important in silhouette-based gait recognition?, CVPR 2 (2004) 776–782

[212] R. Vezzani, R. Cucchiara, Visor: video surveillance on-line repository for annotation retrieval, in: 2008 IEEE International Conference on Multimedia and Expo, April 23–26, 2008 pp. 1281–1284.

[213] R. Vezzani, R. Cucchiara, Annotation collection and online performance evaluation for video surveillance: the visor project, in: IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance, 2008 (AVSS '08), September 2008, pp. 227–234.

[214] R. Vezzani, R. Cucchiara, Visor: video surveillance online repository, Annals of the BMVA 2 (2010) 1–13.

[215] R. Vezzani, C. Grana, R. Cucchiara, Probabilistic people tracking with appearance models and occlusion classification: the ad-hoc system, Pattern Recognition Letters, November 2010.

[216] Robot Vision and the Sign Language Linguistics Labs at Purdue University, Rvl-slll american sign language database, January 2012. <https://engineering.purdue.edu/RVL/Database/ASL/asl-database-front.htm>.

[217] C. Vondrick, D. Ramanan, Video annotation and tracking with active learning, in: Neural Information Processing Systems (NIPS), 2011.

[218] 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, in: 9th IEEE conference on Authomatic Face and Gesture Recognition (FG'11), 2011.

[219] H. Wang, M.M. Ullah, A. Kläser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: BMVC'09, 2009.

[220] L. Wang, C. Leckie, Encoding actions via quantized vocabulary of averaged silhouettes, in: 20th International Conference on Pattern Recognition (ICPR), 2010.

[221] L. Wang, H. Ning, W. Hu, T. Tan, Gait recognition based on procrustes shape analysis, In International Conference on Image Processing 3 (2002) 433–436.

[222] S. Wang, Y. Yang, Z. Ma, X. Li, C. Pang, A.G. Hauptmann, Action recognition by exploring data distribution and feature correlation, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2012, pp. 1370–1377.

[223] Y. Wang, S. Yu, Y. Wang, T. Tan, Gait recognition based on fusion of multi-view gait sequences, Advances in Biometrics, Lecture Notes in Computer Science, vol. 3832/2005, 2005, pp. 605–611.

[224] W. Wei, A. Yunxiao, Vision-based human motion recognition: a survey, in: Second International Conference on Intelligent Networks and Intelligent Systems, 2009 (ICINIS '09), November 2009, pp. 386–389.

[225] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding, 2006.

[226] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3d exemplars, in: International Conference on Computer Vision, 2007.

[227] D. Weinland, R. Ronfard, E. Boyer, A survey of vision-based methods for action representation, segmentation and recognition, Computer Vision and Image Understanding 115 (2011) 224–241.

[228] S. Wu, O. Oreifej, M. Shah, Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories, in: ICC11, 2011.

[229] A. Yang, P. Kuryloski, R. Bajcsy, Ward: a wearable action recognition database, in: CHI, Workshop, 2009.

[230] W. Yang, Y. Wang, G. Mori, Human action recognition from a single clip per action, in: IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), 2009, pp. 482–489.

[231] F. Ye, L. Xu, J. Du, J. Yang, Head-reference human contour model, Zhejiang Daxue Xuebao (Gongxue Ban)/Journal of Zhejiang University (Engineering Science) 45 (7) (2011) 1175–1180. <www.scopus.com>.

[232] F. Yuan, G.G. Xia, H. Sahbi, V. Prinet, Mid-level features and spatio-temporal context for activity recognition, Pattern Recognition 45 (12) (2012) 4182–4191. <www.scopus.com>.

[233] J. Yuan, Z. Liu, Y. Wu, Discriminative subvolume search for efficient action detection, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2009.

[234] J. Yuan, Z. Liu, Y. Wu, Discriminative video pattern search for efficient action detection, January 2012. <http://users.eecs.northwestern.edu/ jyu410/ indexfiles/actiondetection. html>.

[235] L. Zelnik-Manor, M. Irani, Event-based analysis of video, in: IEEE Computer Society Conference on Computer Vision and, Pattern Recognition, 2001.

[236] L. Zelnik-Manor, M. Irani, Weizmann event-based analysis of video, November 2011. <http://www.wisdom.weizmann.ac.il/ vision/VideoAnalysis/ Demos/EventDetec tion/EventDetection.html>.

[237] B. Zhan, D. Monekosso, P. Remagnino, S. Velastin, L.-Q. Xu, Crowd analysis: a survey, Machine Vision and Applications, 0932-8092 19 (2008) 345–357. http://dx.doi.org/10.1007/s00138-008-0132-4. 10.1007/s00138-008-0132-4.

[238] X. Zhang, J. Cui, L. Tian, H. Zha, Local spatio-temporal feature based voting framework for complex human activity detection and localization, in: 1st Asian Conference on Pattern Recognition, ACPR 2011, 2011, pp. 12–16. <www.scopus.com>.