Video-based human behavior understanding:
a survey

Paulo Vinicius Koerich Borges, Nicola Conci, Andrea Cavallaro

Abstract—Understanding human behaviors is a challenging problem in computer vision that has recently seen important advances. Human behavior understanding combines image and signal processing, feature extraction, machine learning and 3D geometry. Application scenarios range from surveillance to indexing and retrieval, from patient care to industrial safety and sports analysis. Given the broad set of techniques used in video-based behavior understanding and the fast progress in this area, in this paper we organize and survey the corresponding literature, define unambiguously key terms and discuss links among fundamental building blocks ranging from human detection to action and interaction recognition. The advantages and the drawbacks of the methods are critically discussed, providing a comprehensive coverage of key aspects of video-based human behavior understanding, available datasets for experimentation and comparisons, and important open research issues.

I. INTRODUCTION

The capability of automatically detecting people and understanding their behaviors is a key functionality of intelligent video systems. The interest in behavior understanding has dramatically increased in recent years, motivated by a societal needs that include security [1], natural interfaces [2], gaming [3], affective computing [4] and assisted living [5]. Significant technological advances in hardware and communication protocols are also facilitating new services such as real-time collection of statistics on group sports [6] and annotation of videos for event detection and retrieval [7]. A number of processing steps are necessary to analyze the scene at different levels of abstraction, starting from the behaviors of objects of interest. The first step consists in detecting and tracking subject(s) of interest to generate motion descriptions (e.g. motion trajectory or combination of local motions), which are then processed to identify actions or interactions. When considering local motions, the analysis generally deals with a fine-grained level of understanding to recognize gestures and motion patterns at intra-body level [8]. Depending on the quality of the camera view, position information can be complemented by other descriptors, such as the trajectories of body joints [9] or head pose changes [10].

Recognizing specific behaviors requires for example the definition of a set of templates that represent different classes of behaviors. However, in many scenarios not all behaviors can be characterized by a predefined number of classes nor can be known (and therefore represented) a priori. In such cases, it is common to use the concept of anomaly, namely a deviation from the learned behaviors [1]. In fact, although anomalous behaviors are generally difficult to model a priori, they can be detected as dissimilar from patterns acquired during regular behaviors.

Recent reviews on human behavior understanding focused on specific aspects of the overall problem, such as monocular pedestrian detection [11], human detection for driver-assistance systems [12] and pedestrian detection benchmarks [13], [14]. Morris and Trivedi [15] survey features and model motion trajectories focusing on trajectory analysis. Moeslund et al. [16] cover methods focusing on motion capture and analysis for human understanding. Comprehensive reviews on algorithms for detecting high-level behaviors compare activity recognition and interaction approaches [17]. Aggarwal and Ryoo [18] review action recognition approaches, extending actions and gestures to human-object interactions and group activities. However, the literature still lacks a comprehensive and up-to-date survey that discusses different abstraction levels at which behavior analysis can be developed depending on the application requirements and the granularity at which the captured information is represented. For this reason, in this paper we link fragmented aspects of this field in a coherent context for human behavior understanding from video (Figure 1). We present a critical analysis and a comparisons of relevant approaches, and we identify common features used by the methods and standard datasets to assess the performances of human behavior analysis algorithms. Moreover, because the
use of terminology is at times ambiguous in the literature, we
provide a consistent definition of the terms used in human
behavior analysis. Finally, we highlight open research chal-

The paper is organized as follows. In Section II we intro-
duce the definitions that will be used throughout the paper
to describe human behaviors. Section III presents a review
of human detection methods, whereas Section IV discusses
actions and gestures. Next, Section V discusses how contextual
information can be exploited to infer interactions. Finally,
Section VI concludes the paper with a discussion and presents
an overview on future trends and open research issues.

II. DEFINITIONS

Human activities can be categorized into four main groups,
namely gestures, actions, behaviors, and interactions. Figure 2
illustrates the relationship between these groups in a two-

Gestures are movements of body parts that can be used to
control and to manipulate, or to communicate [19]. Examples
of gestures include stretching an arm and raising a leg. Ges-
tures are the atomic components characterizing (describing)
the motion of a person. From these atomic elements it is
possible to compose actions, namely temporal concatenations
of gestures [20]. Actions represent voluntary body movements
of arbitrary complexity. An action implies a detailed sequence
of elementary movements that make them univocally decod-
able by an observer and can be combined to compose single
motion patterns or periodic motion patterns. Examples of
single motion patterns are bending, jumping and grabbing an
object; whereas examples of periodic motion patterns include
walking, running and swimming.

A behavior is the response of a person to internal, external,
conscious, or unconscious stimuli [21]. Unlike the recognition
of actions that are represented by a sequence of characteriz-
ing visual elements, the recognition of behaviors requires a

III. HUMAN DETECTION

Video-based people detection methods can be divided into
three main classes, namely appearance-based, motion-based
and hybrid methods. Appearance-based techniques are a spe-
cific case of object detection in still images and use for
example edge information. Motion-based methods consider
temporal information for the definition of the features defining
a human, in particular the movements of the legs. Hybrid
methods use combinations of the previous two classes. Table I
compares the characteristics of human detection methods,
which will be discussed next together with available datasets
to test their performance.
Fig. 3: Different behaviors generated by similar spatio-temporal features corresponding to the action of running: (a) sports; (b) robbery; (c) catching a bus.

TABLE I: Characteristics of human detection approaches (Key. DR: Detection Rate (%); DT: Detection Time (fps); HOG: Histogram of Oriented Gradients; SVM: Support Vector Machine; MPL: Multi-Pose Learning).

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Features</th>
<th>Classifier</th>
<th>DR</th>
<th>DT</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>[28]</td>
<td>Appearance, shape (local and global)</td>
<td>Probabilistic top-down segmentation</td>
<td>71</td>
<td>-</td>
<td>Good results for partially occluded humans</td>
</tr>
<tr>
<td>[26]</td>
<td>HOG and pictorial structures [29]</td>
<td>SVM</td>
<td>82</td>
<td>-</td>
<td>Uses rendered synthetic pedestrian models for training</td>
</tr>
<tr>
<td>[27]</td>
<td>Edge orientation histograms, a variation of HOG</td>
<td>SVM</td>
<td>80</td>
<td>5</td>
<td>Part-based: lower dimensionality than traditional HOG; better performance on INRIA dataset</td>
</tr>
<tr>
<td>[30]</td>
<td>Motion History Image</td>
<td>Metric thresholding</td>
<td>-</td>
<td>30-50</td>
<td>Uses motion characteristics of legs and vehicles</td>
</tr>
<tr>
<td>[31]</td>
<td>Motion power spectral analysis</td>
<td>Spectral similarity comparison</td>
<td>91</td>
<td>1.1</td>
<td>Not only detection but also activity classification</td>
</tr>
<tr>
<td>[32]</td>
<td>Contour motion and Haar-like features</td>
<td>Boosting</td>
<td>90+</td>
<td>24</td>
<td>Combined motion and static image analysis. Very good results on the INRIA dataset</td>
</tr>
<tr>
<td>[33]</td>
<td>Skeleton extremities</td>
<td>Metric based</td>
<td>98</td>
<td>25</td>
<td>Good results but skeleton analysis requires high quality segmentation</td>
</tr>
<tr>
<td>[34]</td>
<td>Maximal principal gait angle</td>
<td>cyclic pattern in human movement</td>
<td>90</td>
<td>10</td>
<td>Use phase-locked-loop</td>
</tr>
<tr>
<td>[37]</td>
<td>Motion point trajectories</td>
<td>Random Decision Forest</td>
<td>-</td>
<td>2-5</td>
<td>Moving camera and near range</td>
</tr>
<tr>
<td>[38]</td>
<td>Appearance, overall blob motion</td>
<td>Kernel SVM</td>
<td>88+</td>
<td>15</td>
<td>Blob motion statistics combined with appearance methods</td>
</tr>
<tr>
<td>[39]</td>
<td>Infrared HOG and Haar</td>
<td>Boosting</td>
<td>93</td>
<td>20</td>
<td>Tuned HOG detector, optimized to the infrared domain</td>
</tr>
<tr>
<td>[40]</td>
<td>HOG adapted to 3D</td>
<td>Linear SVM</td>
<td>85</td>
<td>33</td>
<td>Combination of 2D and 3D data, adapting HOG to RGB-D</td>
</tr>
<tr>
<td>[41]</td>
<td>3D surface model</td>
<td>Model fitting metric</td>
<td>98.4</td>
<td>-</td>
<td>3D surface model for the head using RGB-D combined with 2D contour model</td>
</tr>
</tbody>
</table>

A. Appearance-based methods

Algorithms based on appearance features are specialized detectors trained on large pedestrian databases [12],[11]. These algorithms generally scan each frame, searching for patterns that match pre-defined models. Appearance-based methods can be directly applied to non-static cameras (e.g., for automatic driver-assistance) [42],[43]. Effective representations include the histogram of oriented gradients (HOG) [25], the Viola and Jones algorithm [44] and Haar-like features [45].

Using HOG, the local shape of a human can be represented by the distribution of edge directions, given by intensity gradients. HOGs are usually combined with a support vector machine (SVM) classifier. However, despite being significantly invariant to scale and illumination transformations, HOGs change with object orientation. Several extensions of this representation have been proposed using the concept of HOGs for different body parts that are then combined [46],[29]. Another popular approach is the combination of local and global cues via a probabilistic top-down segmentation [28]. A combination of multiple features such as silhouette, appearance, holistic, and part-based can be used as input to a SVM classifier [27],[47]. The Viola and Jones algorithm, used initially in face detection, can be trained for pedestrian detection [48]. Haar-like features have also been used in combination with covariance features [45], leading to computationally efficient algorithms with performance comparable to HOGs in the INRIA dataset [49].

Benchmark papers on appearance-based people detection introduce a database with richly annotated videos and discuss evaluation measure for performance comparison [13], [14]. Pishchulin et al. [26] perform training on synthetically generated models, by employing a rendering-based reshaping technique to create thousands of images from only a small number of real images.

B. Motion-based methods

Motion-based methods generally detect the cyclic motion of the legs and assume a static camera to identify the moving foreground [50]. Although joint segmentation and detection processes have been proposed [28],[51], person detection often considers a blob segmented from the background, over which further analysis is performed to verify whether the blob is a
pedestrian. However, when under-segmentation occurs, these methods become unreliable. An alternative is to analyze the motion statistics of the tracked blobs as a whole. Useful cues include the cyclic pattern in blob trajectory and an in-phase relationship between change in blob size and position [52].

When observing cyclic motion, gait is analyzed based on pixel or region-wise oscillations, such that the general statistical periodic behavior is used for classification. Early works performed a Discrete Fourier Transform (DFT) to quantify pixel oscillations [53] [54]. Variations of the method analyze the power spectral similarity in the walking pattern [31] or the amount of change in a motion history image [30]. An alternative to analyzing the full pixel intensity information is to high-pass filter the image, observing only a contour motion feature [32]. Cutler and Davis [55] look for the gait period by calculating a similarity matrix for every image pair in a sequence. He et al. [33] determine the angle formed by the centroid point and the two bottom end points of the object skeleton. The histogram of this angle over time is used for detection.

Other periodic detection algorithms use the Phase-Locked Loop [56] and autoregressive moving average models (ARMA) for estimating frequencies of sinusoidal signals. For example, Quinn and Hannan [57] use a second order ARMA model and derive theoretical performance bounds. Another well-known frequency estimation method is the multiple signal classification (MUSIC) algorithm, which estimates the frequency content of a signal using an eigenspace method [58]. Ran et al. [34] discuss the methods above addressing the specific problem of periodicity estimation for pedestrian detection. Other works analyzed the periodic change in the optical flow [35] [59]. Perbet et al. [37] use a number of point trajectories and identify those which are spatio-temporally correlated as arising from feet in walking motion.

People detection performance can be improved by considering scene modeling that helps reducing the search space. Knowledge of the scene can be used to train for a specific area [60] [61]. The adaptation of a generic pedestrian detector to a specific traffic scene can be performed automatically [61], adding for example information about common pedestrian appearance in that scene, and the most likely paths. Moreover, when the homography between the ground and the camera is known, size features can be used for the detection [62]. Liu et al. [63] manually annotate the main surfaces in the scene providing spatial contextual knowledge. The density distribution over surfaces is learned from the scene and it is conditional on the surfaces semantic properties. As an example, on a zebra crossing pedestrians occur with a higher probability than cars, for a given direction of movement. All these cues improve the pedestrian detection reliability.

People detection in the infrared domain can be achieved with slight modifications to visible-spectrum detectors [39]. Moreover, infrared imaging presents particular contrast characteristics that can be exploited [64] [65]. Although simple brightness analysis can be inefficient for characterizing people due to the polarity switch phenomenon [65], employing appropriate statistical models such as the Infinite Gaussian Mixture Model [64] for the contrast around potential areas yields effective classification.

C. Hybrid methods

Efficient people detection solutions combine appearance and motion [48], [66]. These two approaches can separately analyze the data, merging the final result according to a given decision function, or apply a still-image detector to regions potentially containing pedestrians, as indicated by a blob tracker. The advantage of the former solution is that the detection is not restricted to moving blobs. However, the computational complexity is largely increased, as the appearance-based detector performs a search over the full frame, reducing the number of frames per second (fps) in video analysis. The latter solution reduces computational complexity and the false-positive rate.

Methods that consider a background model to segment potential pedestrians generally present a better detection rate, both in terms of false positives and false negatives.

One of the earlier works jointly exploiting both types of features combines a Viola and Jones detector [44] operating on differences between pairs of images [48] and training using AdaBoost [67]. Each step of the AdaBoost selects from the several motion and appearance features the feature with lowest weighted error on the training samples. The final classifier weighs pixel intensity and motion information such that the detection rates are maximized. The accuracy of this technique [48] can be improved by one order of magnitude at the expense of using a higher number of frames [68]. A related approach is also used by Dalal [66] and Bouchirika et al. [69], combining HOGs with motion and gait patterns. The concept can be extended to also include behavior as a cue in the detection process [70], where blobs with affine movements have a higher likelihood of representing a group of pedestrians.

Alternative sensing modalities using depth information are becoming increasingly popular. Xia et al. [41] perform detection by using a 2-D head contour model and a 3-D head surface model, in a constrained office area. An adaptation of HOGs to RGB-D data is proposed by Spinello and Arras [40], where the direction of depth changes are locally encoded. An accuracy of 85% in ranges up to 8 meters is achieved.

D. Datasets

Datasets for testing human detection algorithms include the MIT Pedestrian Database [71], the USC Pedestrian Detection Testsets [72], [73], the CALTECH Database [13], the INRIA Person Dataset [49] and the ETH Dataset [74]. The MIT Pedestrian Database was one of the earliest datasets, with the people poses limited to frontal and rear views. The images are scaled to the size $64 \times 128$ and aligned such that the person’s body is centered in the image. The INRIA Person Dataset presents similar scenarios but with a larger variety in poses and comprehensive annotation. Later datasets [72], [13] contain lower resolution pedestrians in larger images, recorded in cluttered urban areas. These datasets contain rich annotation, including occlusion information.

The PETS Database [75] contains single and multi-view footages including outdoor and indoor scenarios. The ETH
Dataset [74] is collected using two cameras, with annotation and camera calibration information. A comprehensive discussion on datasets for pedestrian detection is presented in [14]. Sample images from detection datasets are shown in Figure 5, illustrating from top to bottom how the complexity of the scenarios has evolved over time.

IV. ACTIONS

Action recognition approaches commonly use as descriptors low-level features such as optical flow (both dense and sparse) with gradient information [77], [78], silhouettes [79], and long-term analysis of motion [80]. Semantics and context-based techniques can be used to support and increase the system reliability [81], [82]. A selection of relevant techniques for action recognition using these techniques are described below and compared in Table III.

A. Low-level features and spatio-temporal interest points

Common approaches for action recognition include dense optical flow [95], [97] and spatio-temporal interest points (STIPs) [86], [78], [98], [99], [100], [101], [102]. Laptev and Lindeberg originally presented STIPs for action recognition [103] as a space-time variation of the Harris corner detector [104]. Regions with high intensity changes in both space and time are defined as STIPs. Alternatively temporal Gabor filters can be employed to increase the robustness of the method [98], [105]. STIPs do not necessarily rely on a person detector, they are robust to clutter and are relatively invariant to temporal and appearance changes. However, they are generally combined with bag-of-features models and therefore do not account for temporal and spatial probabilistic feature distributions.

Early work by Schuld et al. [78] considered individual feature detections to be independent and built a single histogram for the full video sequence. Ullah et al. [106] perform video segmentation prior to STIP detection to eliminate misplaced interest points. Alternative approaches use several binning designs [107], [108] to obtain the relative structure of STIPs across a video. Although results are generally better than the single-bin (histogram) approach, rigid bin partitions can affect the spatial or temporal shift in the video volume. Chakraborty et al. [109] introduce a bag of visual words vocabulary approach by combining spatial pyramid and vocabulary compression techniques.

Dense optical flow has also been successfully used for action recognition [95], [97], [110]. Wang et al. [95], [110] use dense optical flow trajectories and segment the foreground from background motion for moving cameras. Dense trajectories are more robust to irregular abrupt motions and are more accurate in capturing complex motion patterns. The authors show that motion boundaries encoded along the trajectories can outperform state-of-the-art trajectory descriptors, on the KTH [78], Youtube [111], Hollywood2 [112] and UCF Sport datasets [113].

An action spectrogram for interest points in 3D volumes is used by Chen and Aggarwal [114], with successful results on the Weizmann dataset and 91% accuracy on the KTH dataset. A variation of the problem is to consider not only action recognition, but also action prediction [115], with 70% accuracy in UT Dataset [116].

Given the redundancy of information collected through spatio-temporal features, Castrodad and Sapiro [96] propose a sparse coding pipeline, by classifying the incoming videos using a dictionary of learned primitives. In spite of the simplicity of the method, the achieved results reach very high performances, scoring 100% on the KTH and UT datasets, and 96% on the UCF-Sports dataset.

B. Mid and high-level representation

At a higher level, action recognition can be performed by exploiting mid and high-level features, such as long-term tracked trajectories and semantics. Choi et al. [80] present a framework for collective activity recognition exploiting motion trajectories and poses. By analyzing coherent activity, better action recognition performance is obtained. Using their own dataset [117] with actions such as standing, queuing, walking and talking, they obtain an average accuracy of 82%. Raptis et al. [92] use motion trajectories to spatially locate where an event occurs. Trajectories are clustered, and the clusters’ properties are analyzed to classify the action. They achieve a 79.4% accuracy on the UCF-Sports Dataset. Hoai et al. [93] present a framework for joint segmentation and recognition of actions based on multi-class SVM for action classification, while segmentation inference is achieved by dynamic programming. Regular expressions can be exploited to automatically associate an observed activity pattern to a template of activities learned a priori, achieving a 87.7% accuracy on...
TABLE II: Datasets for human detection. (Key. A: Annotation; Y: yes; N: no; P: partial)

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Name</th>
<th>Features</th>
<th>A</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>[71]</td>
<td>MIT Pedestrian Database</td>
<td>923 images, with windows around the subjects</td>
<td>N</td>
<td>Pose of people limited to frontal and rear views. Images scaled to 64 × 128</td>
</tr>
<tr>
<td>[73], [72]</td>
<td>USC Pedestrian Detection Testset</td>
<td>369 images, with 816 humans</td>
<td>Y</td>
<td>Images collected from the Internet and from the CAVIAR video database [76], with frontal/rear view walking/standing human</td>
</tr>
<tr>
<td>[49]</td>
<td>INRIA Dataset</td>
<td>Approx. 2500 images, 128 × 94</td>
<td>Y</td>
<td>Refined combination of images from different datasets (including Internet), with pedestrians with bounding box area &gt; 100 pixels</td>
</tr>
<tr>
<td>[74]</td>
<td>ETH Dataset</td>
<td>12,298 pedestrians in approx. 2.000 frames</td>
<td>Y</td>
<td>Two-camera dataset. The unbayered images, the camera calibration, and annotation are provided for both cameras. Annotated pedestrians are taller than 50 pixels</td>
</tr>
<tr>
<td>[13], [14]</td>
<td>Caltech Pedestrian Detection Benchmark</td>
<td>10 hours of 640 × 480 video</td>
<td>Y</td>
<td>Recorded from moving vehicle in urban area. 350,000 bounding boxes with 2300 unique pedestrians annotated</td>
</tr>
</tbody>
</table>

TABLE III: Comparative analysis of action recognition algorithms (Key. Ref: Reference; DR: Detection Rate; MEI: Motion-Energy Image; MHI: Motion-History Image; GMM: Gaussian Mixture Model; HMM: Hidden Markov Model; HMM-MIO: HMM with Multiple Independent Observations; SVM: Support Vector Machine; EM: Expectation Maximization; MRF: Markov Random Field)

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Representation</th>
<th>Classifier</th>
<th>Key Aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td>[83]</td>
<td>MEI and MHI</td>
<td>“Pooled” Mahalanobis distance</td>
<td>Temporal templates and dynamic matching. Real time and multi-view</td>
</tr>
<tr>
<td>[84]</td>
<td>Fourier Descriptors</td>
<td>SVM and HMM</td>
<td>90% DR on ad-hoc dataset with 101 activities (5 subjects)</td>
</tr>
<tr>
<td>[85]</td>
<td>GMM to model salient postures</td>
<td>EM-based learning</td>
<td>Tolerance to noise, robustness across subjects and datasets</td>
</tr>
<tr>
<td>[86]</td>
<td>Spatio-temporal template (star-figure)</td>
<td>GMM framework</td>
<td>GMM features arranged in a two dimensional map, transformed into a gray level image</td>
</tr>
<tr>
<td>[87]</td>
<td>Hierarchical</td>
<td>χ² test</td>
<td>Automatic model adaptation; unsupervised behavior analysis and abnormality detection</td>
</tr>
<tr>
<td>[88]</td>
<td>Action graphs and 3D points</td>
<td>Bi-gram maximum likelihood decoding</td>
<td>Sparse sampling of 3D points to characterize the 3D shape of each salient postures</td>
</tr>
<tr>
<td>[89]</td>
<td>HMM with multiple independent observations (HMM-MIO)</td>
<td>HMM</td>
<td>Robustness to outliers, dimensionality reduction, handles sparse observations; performance comparable to discriminative classifiers</td>
</tr>
<tr>
<td>[90]</td>
<td>Self similarity matrix</td>
<td>Nearest neighbor</td>
<td>Self similarity matrices combined over time. 100% accuracy on KTH and Weizmann datasets</td>
</tr>
<tr>
<td>[91]</td>
<td>Motion trajectory</td>
<td>SCFG, Allen’s temporal logic, Multi-Thread Parsing</td>
<td>Error detection and recovery, automatic rules induction</td>
</tr>
<tr>
<td>[80]</td>
<td>3D trajectories and pose</td>
<td>Random Forest and 3D-MRF</td>
<td>Group analysis as the spatial distribution of people and its temporal evolution</td>
</tr>
<tr>
<td>[92]</td>
<td>Motion trajectories</td>
<td>SVM, MRF</td>
<td>Classification based on trajectory grouping</td>
</tr>
<tr>
<td>[93]</td>
<td>Low or mid-level features</td>
<td>SVM, Dynamic Programming</td>
<td>Joint segmentation and classification of actions</td>
</tr>
<tr>
<td>[82]</td>
<td>Template-based action detector</td>
<td>SVM</td>
<td>High-level representation. Invariance to scale, viewpoint and tempo</td>
</tr>
<tr>
<td>[94]</td>
<td>Motion trajectories</td>
<td>Spectral analysis of graphs</td>
<td>Trajectories as transitions of graph nodes</td>
</tr>
<tr>
<td>[95]</td>
<td>Dense optical flow</td>
<td>Motion boundary histogram</td>
<td>Dense optical flow trajectories and an efficient tracker and descriptor, segmenting foreground motion from background motion</td>
</tr>
<tr>
<td>[96]</td>
<td>3D image patches</td>
<td>Sparse coding</td>
<td>Dictionary learning to extract action primitives and used as basis vectors. Classification obtained by reconstruction</td>
</tr>
</tbody>
</table>

the Weizmann dataset and 42.4% on the Hollywood dataset. Duldoss et al. [118] model actions through an abstraction of the top-view trajectory. The expressions corresponding to the activity models are learned as separate Context-Free Grammars (CFGs) using a set of training sequences, parsed using the Earley-Stolcke algorithm. Zhang et al. [91] transform the motion trajectories in a set of basic motion patterns and use a rule induction algorithm based on the Minimum Description Length (MDL) to derive the spatio-temporal structure of the event from the primitives stream. Anomalous actions can be detected based on spectral analysis of graphs [94] by representing human motion trajectories as transitions between nodes. Individuals span only a limited portion of all possible trajectories on the graph and this subspace is characterized by large connected components of the graph, on which it is possible to implement invariant metrics for anomaly detection.

Most of the works discussed above define human actions by their motion and appearance features. However, other contextual and semantic information can also be employed [119], [81], [82]. Gupta et al. [119], use a storyline to describe causal relationships between actions (e.g. baseball swinging and running are typically related actions). AND-OR graphs are used as a mechanism for representing storyline models. The parameters and structure of the graph are learned from weakly labeled videos, using linguistic annotations and visual data, leading to joint learning of both storyline and appearance models. Lin et al. [81] use attribute graphs for activity recognition. They model the variability of semantic events by a set “event primitives”, which are learned as a object-trajectory mapping that describes mobile object attributes (location, velocity, and visibility). With this representation, one observed event is parsed into an event parse graph, and variabilities of one event are modeled into an AND-OR graph.

Raptis and Segal [120] model an action as a very sparse sequence of temporally local discriminative keyframes. Keyframes are learned as collection of poselets, a description
recently used in person recognition [121]. Poselets capture discriminative action parts that carry partial pose information, thus reducing the impact of occlusions. The model semantically summarizes actions in a storyline, which represents a contextual temporal orderings of discriminant partial poses.

Inspired by the object bank approach for image segmentation [122], Sadanand and Corso [82] propose a high-level representation of actions. The main idea is that a large set of action detectors can be seen as the bases of a high-dimensional “action-space”. This characteristic, combined with a simple linear classifier, can form the basis of a semantically-rich representation for activity recognition in video. The authors report accuracy of 98.2% on the KTH, 95.0% on the UCF Sports, and 57.9% on the UCF50 dataset. Promising results are also given by Messing [123], where the semantics are incorporated to his previous work [124] on activity recognition using a model for the velocity history of tracked points. Higher level information, the relative position of a person, their body parts and objects in known scenarios are incorporated in the activity recognition system. As the knowledge of objects aid activity recognition, Fathi et al. [125] use an egocentric approach, where video captured from a wearable camera is used to view the scene in front of the user at all times. They perform the analysis of hand movements and interaction with objects, using a weakly supervised technique. The method automatically segments the active object areas, assigning regions to each object and propagating their information using semi-supervised learning.

C. Silhouettes

The use of silhouettes to classify actions assumes that human movements can be represented as a continuous progression of the body posture. These approaches are mainly based on background segmentation [11]. Action descriptors can be extracted from a sequence of silhouettes in consecutive frames and traditional classifiers can be employed for recognition [83], [126]. Alternatively a dynamic model of the action of interest can be generated using characteristics from each silhouette [79], [127].

Action descriptors extracted from a sequence of silhouettes capture and combine spatio-temporal characteristics of the activity. A common technique is to accumulate silhouettes to generate motion energy images (MEI) as well as motion history images (MHI) [83]. Hu moments [126] can be extracted from both MEI and MHI as action descriptors, and action classification is based on the Mahalanobis distance between each moment descriptor of the known actions and the one under analysis. Chen et al. [86] fit star figure models to human silhouettes to identify the five extremities of the shape corresponding to head, arms and legs. To model the spatial distribution of the five points over time, Gaussian Mixture Models are used, ignoring the temporal order of the silhouettes in the action sequence. Nater et al. [87] use a hierarchical approach based on silhouettes to detect falls, whereas Li et al. [88] extend the silhouette approach to 3D sensors.

Dynamic models of actions generated by extracting characteristics from each silhouette generally employ statistics-based techniques such as Conditional Random Fields (CRF) [128], [129] and Hidden Markov Models (HMM) [127], [89]. The features extracted from each silhouette capture the shape of the body and possible local motion. Kellokumpu et al. [84] apply Fourier-based shape descriptors to cluster a number of possible postures. HMMs are employed to model the activity dynamics, such that each cluster is assumed as a discrete symbol from the hidden states in the HMM. Instead of analyzing the behaviour of whole bodies, an alternative is to perform parts-based analysis [130], [131], where template matching is combined with a pictorial structure model to detect and localize actions.

Sun et al. [90] were the first to obtain 100% accuracy on both the KTH and Weizmann datasets by generating a self-similarity matrix (SSM) for each frame in the video using a feature vector. The SSMS from all frames are combined and the result is decomposed into its rank-1 approximation, yielding a set of compact vectors that efficiently discriminate different actions. Alternative methods analyze skeletons [132], [85], [20], [16], which represent the human body in terms of articulated joints, thus extending the information provided by silhouette and shape to analyze motion and pose at a finer level of detail. In a simple form, skeleton-based articulation considers a small number of points: the head, hands, and feet, for example, forming a 5 point star [132]. Most statistics are derived from the absolute and relative motion of those points with respect to the star centroid. The statistics can be analysed directly or can serve as input to a classifier.

When subjects are close to the camera, using RGB-D, Shotton et al. [133] predict the 3D positions of skeletal body joints from a single image. Using a highly diversified training dataset, they generate an intermediate body parts representation that maps the complex pose estimation problem into a simpler per-pixel classification task. Extending the poses to activities, Sung et al. [134], [135] present learning algorithms to infer the activities, using a hierarchical maximum entropy Markov model on RGB-D data. Their method considers an activity as composed of a set of sub-activities of body parts, inferring a two-layered graph structure using dynamic programming. They achieve an average precision rate of 84.3% in house and office environments.

D. Datasets

Activity recognition datasets are either staged or natural, and contain images or trajectory data. The KTH dataset [78] and the Weizmann dataset [136] contain instances of walking, jogging, running, boxing, hand waving, and hand clapping, performed several times by different subjects. State-of-the-art methods achieve recognitions rates of more than 90% [131], [137] in the KTH and Weizmann datasets, which provide an initial basis for comparison among algorithms [89]. More realistic datasets include UCF50 [138] and Hollywood [137]. Other staged databases include INRIA Xmas [139] and MuHAvi [140], which contain multi-camera views. The UCF YouTube database [111] has large variations in camera motion, object appearance and illumination conditions, as illustrated in Figure 6. Large variations are also found in the Hollywood Human Action datasets [137], [112], which contain thousands
of video movies clips corresponding to more than 20 hours of data representing 12 human actions classes and 10 scene classes (e.g. kitchen, car, hotel).

Other datasets only consider the spatio-temporal motion information provided by the trajectories of moving people. These datasets include the 3D People Surveillance Dataset 3DPES [141], the MIT trajectory dataset [142] and the Edinburgh Informatics Forum Pedestrian Database (EIFPD) [143]. The 3D People Surveillance Dataset 3DPES [141] contains paths of about 200 people, acquired over several days and complemented with ground-truth information in the form of bounding boxes and direction of motion. The MIT trajectory dataset [142] contains more than 40,000 trajectories collected over five days from a single camera monitoring a parking lot. The Edinburgh Informatics Forum Pedestrian Database (EIFPD) [143] covers several months of observation resulting in about 1000 trajectories/day. Although data do not include the ground truth, information about the detected blobs, the raw samples, and the interpolated trajectory using splines is provided.

As subset of activity recognition, datasets for gesture recognition include ASL (Australian Sign Language) [144], FGnet [146], [149], [147], Pointing’04 [148], Cambridge Gesture Database [150], and the ChaLearn Gesture Challenge dataset [151]. The ASL [144] dataset consists of a wide set of samples of Auslan (Australian Sign Language) signs: 27 examples of each of 95 Auslan signs were captured from a native signer using high-quality position trackers. A small two-handed dataset is presented in [146] provided by the Face and Gesture Recognition Working Group (FGnet). The videos contain 7 two-handed gestures, collected using 7 different subjects and recording 10 videos per gesture per subject. The same working group proposed in [145] additional datasets for hand posture recognition, as well as a specific dataset for hand gesture recognition [147]. This dataset includes deictic and symbolic gestures for 400 samples. The benchmark data of Pointing’04 workshop [148] include ground-truth information indicating where users are pointing. Nine static and four dynamic gestures are presented and annotated in [149], including besides the identifiers for event and user, the gesture, action start, stroke start, stroke end and gesture end. 900 image sequences of 9 gesture classes defined by 3 primitive hand shapes and 3 primitive motions are presented in [150]. Each class contains 100 image sequences (5 different illuminations, 10 arbitrary motions, 2 subjects). The target task for the dataset is to classify different shapes as well as different motions at a time.

Recent RGB-D datasets include the RGBD-HuDaAct [152] and the ChaLearn Gesture Challenge [151] datasets. The RGBD-HuDaAct database contains videos showing subjects performing various common daily life activities, such as talking on the phone, getting up, and mopping the floor. The dataset is annotated and one single subject is present in each action. Each sequence spans about 30 to 50 seconds, totaling approximately 46 hours of video at $640 \times 480$ resolution. Each subject performs the same activity 2-4 times. In the ChaLearn Gesture Challenge dataset [151] ground-truth annotation is partially available, including temporal segmentation into isolated gestures and body part annotations (head, shoulders, elbows, and hands).

A summary of the characteristics of the datasets discussed above is given in Table IV.

V. INTERACTIONS

We can distinguish between one-to-one and group interactions. One-to-one interactions can be seen as extensions of action for single subjects complemented by contextual and social information. Group interactions require the detection of a group entity in terms of social aggregation. In both cases motion features need to be combined with sociological and psychological information that rule interpersonal relations. However, the literature usually addresses them as two separate problems. Table V compares the interaction recognition techniques described below.

Recognizing the social component of behaviors and social activities that involve multiple people requires observing the activity of a person in reference of his/her neighbors. Considering position as the main source of information, interactions can be modeled by the permanence for a specific amount of time in the neighborhood of one or more meaningful spots in the scene. Models from psychology and sociology that can be used include the theory on proxemics [153][154]. Proxemics exploits the so-called social space (the space between subjects) in order to infer interpersonal relationships. These relationships can be seen as a summation of attractive and repulsive forces that drive human behaviors, linking them with other people in the surroundings. Helbing et al. [155] assimilate the human behavior in a social context as a summation of forces that
TABLE IV: Datasets for gesture recognition. (Key. A: Annotation; Y: yes; N: no; P: partial; CGD: Cambridge Gesture Database; CGC: ChaLearn Gesture Challenge)

<table>
<thead>
<tr>
<th>Ref</th>
<th>Name</th>
<th>Features</th>
<th>A</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>[144]</td>
<td>ASL</td>
<td>95 signs, 27 examples per sign</td>
<td>Y</td>
<td>Samples of Auslan (Australian Sign Language) signs captured from a native signer</td>
</tr>
<tr>
<td>[145]</td>
<td>FGNet</td>
<td>10 postures, 24 people, 3 backgrounds for static; 1 sequence, 4 gestures for dynamic</td>
<td>Y</td>
<td>Static and dynamic hand posture database</td>
</tr>
<tr>
<td>[146]</td>
<td>FGNet</td>
<td>7 gestures, 7 users, 10 videos per gesture</td>
<td>Y</td>
<td>Two-handed gestures</td>
</tr>
<tr>
<td>[147]</td>
<td>FGNet</td>
<td>about 400 videos</td>
<td>Y</td>
<td>Hand posture recognition, including deictic and symbolic gestures</td>
</tr>
<tr>
<td>[148]</td>
<td>FGNet</td>
<td>8 videos, each with user pointing in 8 directions</td>
<td>Y</td>
<td>Includes a video set for pointing on a whiteboard in a multi-camera setup and a set of videos recorded using a head-mounted camera.</td>
</tr>
<tr>
<td>[149]</td>
<td>FGNet</td>
<td>13 gestures, 9 static, 4 dynamic, 16 users</td>
<td>Y</td>
<td>Static and dynamic gestures performed under different illumination conditions by different subjects.</td>
</tr>
<tr>
<td>[150]</td>
<td>CGD</td>
<td>900 sequences, 9 gestures, defined by 3 primitive hand shapes and 3 primitive motions.</td>
<td>Y</td>
<td>Sequences recorded in front of a fixed camera having roughly isolated gestures in space and time. Videos are uniformly resized into $20 \times 20 \times 20$.</td>
</tr>
<tr>
<td>[151]</td>
<td>CGC</td>
<td>about 50,000 images</td>
<td>P</td>
<td>The dataset includes high quality RGB videos complemented by depth information of single users.</td>
</tr>
</tbody>
</table>

lead a subject to its target (Social Force Model). The idea of a social force that binds and regulates the relationships between subjects has been widely adopted to model both one-to-one interactions and group interactions, as a generalization of the one-to-one case. Mehran et al. [156] use grid of particles over the image that are moved by the forces created by the space-time optical flow as they were individuals. The moving forces are then estimated using the social force model to identify normal or abnormal behaviors.

A. One-to-one interactions

The interaction level between two subjects can be measured as an energy function (or potential) computed along the axis connecting them [157]. Pellegrini et al. [158] applied this model to crowded scenes, considering single moving entities as agents. The motion of each agent is driven by its destination, and planned so as to prevent collisions with other moving objects. Every agent is associated with a set of features, comprising position, speed and direction of motion. Path planning is achieved by implementing a social force model to evaluate the interaction level between two subjects. From the analysis of the energy function, the authors propose a Linear Trajectory Avoidance (LTA) model, based on a short time prediction, to model the avoidance path. Longer predictions would lose effectiveness, because of the casualness of the human movement. Taj and Cavallaro [159] describe the coupling between multiple object states using Coupled HMMs on relative spatio-temporal features among the people under analysis. Rota et al. [160] recognize normal and abnormal two-people interactions exploiting proxemics cues combined with motion information, also revealing the intentionality of a specific interaction. Zen et al. [161] identify proxemics cues in a close-range analysis to discriminate personality traits such as neuroticism and extraversion, and use the collected data to construct a correspondent behavioral model. The social interaction cues are then used to complement a people tracker to improve its accuracy.

B. Group interactions

Extending the concept to a group of several subjects implies analyzing the scene from the social viewpoint and the dynamics ruling these interactions are significantly different. Cristani et al. [162] detect the so-called F-Formations in the scene to infer whether an interaction between two or more persons is occurring. The interactions in a small groups of people can be encoded in three categories, namely self-causality, pair-causality and group-causality [163], where features characterizing each category are identified based on the trajectories of the subjects.

The social interaction model proposed in [158] [164] assumes that people behave consistently when walking in groups. Trajectory hypotheses are generated for each person within a time window, and the best hypothesis is selected taking into account social factors, while estimating group membership using Conditional Random Fields. The group membership allows optimizing motion prediction also considering constraints imposed by the group. The algorithm improves tracking performance by introducing additional constraints for the tracker. Finally, an interaction energy potential can be extracted to model the relationships among groups of people [165]. The relationship between the current state of the subject and the corresponding reaction is then used to model normal and abnormal behaviors.

RGB-D data can be used to detect fine-grained human-object interaction [166], identifying hand movements and other cooking activities, considering the interactions between hands and objects. The system combines global and local features, also considering the duration of the events. The global feature uses the principal component analysis (PCA) on the gradients of 3D hand trajectories. The local feature employs bag-of-words of trajectory gradients snippets, which are useful for distinctive isolated actions such as chopping.

Zhang and Parker [167] employ RGB-D sensors on a mobile robotic platform, aiming at efficient interaction between the robot and humans. The feature detector applies separated filters on the 3D spatial dimensions and to the temporal dimension to detect a feature point. A feature descriptor then clusters the intensity and depth gradients within a 4D cuboid, which is
centered at the detected feature point as a feature. As a classifier for the activity recognition, the authors employ the Latent Dirichlet Allocation [168] combined with Gibbs sampling. Detection rates are relatively high in a staged scenario, for basic activities such as waving, walking, and signaling. Koppula et al. [169] use a mobile robot and consider the problem of labeling sub-activities performed by a human, but also use interaction with objects as cues (associated affordances). The human activities and object affordances are jointly modeled in a Markov Random Field (MRF) framework. The learning problem is formulated using a structural SVM approach, with 75.8% precision and 74.2% recall in a challenging dataset.

C. Datasets

Interaction detection datasets include the BEHAVE dataset [170], the UT Interaction Dataset [116], the TV Human Interactions Dataset [171] and the CMU MOCAP [172] (Figure 7). The BEHAVE dataset [170] covers interactions between multiple persons with subtly different behaviors such as InGroup (IG), Approach (A), WalkTogether (WT), Split (S), Ignore (I), Following (FO), Chase (C), Fight (F), RunTogether (RT) and Meet (M). A smaller dataset is the UT Interaction Dataset for the SDHA Contest [116] that covers continuous executions of 6 classes of human-human interactions: shake-hands, point, hug, push, kick and punch. Ground-truth labels for these interactions are provided, including time intervals and bounding boxes. There are 20 video sequences of around 1-minute length. For close-range interactions, the TV Human Interactions Dataset (TV HID) [171] is used to analyze interactions occurring in TV footages [173]. The dataset is annotated. Finally, the CMU MOCAP [172] is an annotated 3D dataset that also includes a section for two-person interactions. The position and motion information of body parts are captured using markers that create the skeleton of the person. The dataset includes several types of behaviors, ranging from normal interactions such as handshaking and passing an object to quarrels and fights.

In the RGB-D domain, datasets include the Cornell CAD-60 [134] and CAD-120 [169], and the LIRIS [174]. The LIRIS dataset contains video of multiple subjects, with partial focus on interactions, such as discussing, handling items, handshaking, apart from some single subject actions present in the RGBD-HuDaAct dataset. Full annotation is provided, with information on the type of action and related spatial and temporal position in the video. This dataset contains fully annotated standard 720 × 576 video of the same scenes, recorded with a consumer camcorder, providing a useful basis for comparisons between RGB and RGB-D techniques. The CAD-60 and CAD-120 datasets consider most environments of a regular household (office, kitchen, bedroom, bathroom), with three to four common activities identified in each location with about 45 seconds of data for each activity and each person. The CAD-120 extends the CAD-60 by separating high-level human-object interactions (e.g. taking medicine, cleaning objects, microwaving food) and sub-activities like punching, reaching, and drinking. Object characteristics labels such as movable, reachable and drinkable are also included.

This dataset contains a comprehensive annotation and each video is annotated with sub-activity labels, tracked human skeleton joints and tracked object bounding boxes. This dataset combines actions, gestures and interactions in its annotations.

A summary of the characteristics of the datasets discussed above is given in Table VI.

VI. DISCUSSION AND CONCLUSIONS

This survey presented an extensive overview of video-based behavior understanding, its related definitions and an analysis of datasets used for benchmarking. The goal is to provide the reader with a critical analysis of the major steps, from detection to high-level interpretation, contrasting key elements of the different approaches using comparative tables. Based on this survey, we can highlight successful approaches as well as several important future directions, which include promising methods, the generation and reconstruction of 3D observations, datasets and annotation.

From a representation viewpoint, and in spite of their simplicity, low-level features (such as pixels), but also spatio-temporal features as STIPs, have revealed excellent performances (e.g., [114][96]), achieving 100% accuracy on some of the common benchmark datasets. Low-level features benefit from the fact that they are generally easy to extract. However, they are unable to handle the temporal structure of the action/behavior, implying the need for a higher level analysis to construct a suitable temporal model. Mid and high-level representations can bridge this gap, providing also the spatio-temporal model (motion trajectories) associated to each descriptor, both at a feature level [113] or objects level [91][121]. Also in this case, efficient performances can be achieved, as demonstrated by Sadanand and Corso [82], who reached 98.2% on the KTH dataset and 95% on the UCF Sports dataset. Enriching the descriptors with contextual information
TABLE V: Comparison of social interaction modeling approaches (Key. SFM: Social Force Model; BoW: Bag of Words; GA: Genetic Algorithm; SVM: Support Vector Machine; EM: Expectation Maximization; CRF: Conditional Random Field; STIP: Space Time Interest Points).

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Feature</th>
<th>Matching</th>
<th>Objective</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>[156]</td>
<td>Optical flow of particles</td>
<td>SFM + BoW</td>
<td>Anomaly detection</td>
<td>Far range, particles are assimilated to individuals, model regular and abnormal flow</td>
</tr>
<tr>
<td>[157]</td>
<td>Energy potentials</td>
<td>Energy minimization</td>
<td>Learn pedestrian dynamics for trajectory estimation</td>
<td>Three energy potentials to be minimized: destination, constant velocity and avoidance</td>
</tr>
<tr>
<td>[158]</td>
<td>Energy potentials</td>
<td>GAs.</td>
<td>Improve multi-target tracking exploiting social behavioral models</td>
<td>Energy potentials for position, speed and direction of motion</td>
</tr>
<tr>
<td>[160]</td>
<td>Potentials for distance and velocity</td>
<td>SVM + Proxemics</td>
<td>Classification of two-persons interactions: intentional, normal, abnormal</td>
<td>Analysis carried out with a temporal sliding window</td>
</tr>
<tr>
<td>[161]</td>
<td>Distance</td>
<td>MRF + Proxemics</td>
<td>Improve visual tracking</td>
<td>Joint analysis of social space, visual attention and personality traits</td>
</tr>
<tr>
<td>[162]</td>
<td>Distance</td>
<td>Clustering + EM + Proxemics</td>
<td>Infer social relations from interpersonal distances</td>
<td>Detection of F-Formations. Social distances used to infer roles of people in the party</td>
</tr>
<tr>
<td>[163]</td>
<td>Visual Codebooks on motion trajectories</td>
<td>SVM</td>
<td>Recognition of self, pair, and group-causality</td>
<td>A dataset is also proposed. Activities include walking and running together, gathering, standing, fighting</td>
</tr>
<tr>
<td>[164]</td>
<td>Energy potentials</td>
<td>CRF</td>
<td>Infer social interactions to improve tracking in groups</td>
<td>Third-order graphical model to jointly estimate trajectories and group memberships over a time window</td>
</tr>
<tr>
<td>[165]</td>
<td>Energy potentials on STIPs</td>
<td>BoW + SVM</td>
<td>Detection of abnormal events</td>
<td>Human detection and segmentation are not necessary. Events modeled through STIPs</td>
</tr>
</tbody>
</table>

can also provide significant improvements in the classification performances. This can be achieved by complementing the features with semantics [119], defining (or learning) a description of the video in terms of causal relationships, as it is common, for example, in sport events. These approaches are particularly efficient when observing long-term actions, in which atomic elements can be combined together to compose a more complex and structured behaviors.

Although the appropriateness of the feature representation is a key element to create a correct behavior model, the choice of the classifier may strongly influence the quality of the results. Deterministic classifiers (such as neural networks, SVM), often combined with some efficient representation of the input data as bag-of-words or dimensionality reduction, are commonly used. However, a deterministic model is strongly dependent on the set of training data. For this reason, deterministic models are well suited to detect behaviors that show a certain level of regularity and for which the salient details that characterize them are persistent across the samples of the dataset. A different category of methods model behaviors as stochastic processes. A stochastic, or probabilistic process assumes that the variables to be analyzed behave as probability distributions, instead of single values, thus incorporating in the model a certain degree of randomness. Some of the most widely used include Markovian models and their variations (e.g., HMM, MRF), as well as CRF and stochastic context-free grammars (SCFG). These solutions often match very well with the nature of behaviors that tend to evolve or change, especially when observed over long temporal intervals, due to the high variability with which humans accomplish the same tasks over time.

Graphical models [175] have demonstrated their robustness in dealing with noise both in data acquisition and behavior interpretation [176]. Examples include action [177][80][93] and interaction analysis [178][164].

The problem of reconstructing the motion of a 3D articulated body from 2D point correspondences from a single camera for behavior understanding has been gaining increased attention [179], [180]. However, methods still require capturing the object at a high resolution and proximity to the camera. New opportunities are offered by devices such as RGBDepth (RGB-D) sensors (e.g. Asus Xtion [181] and Microsoft Kinect [182]) that simplify tasks at close range for human behavior analysis [41][134][135]. Another natural extension is the use of multiple cameras [74] [139] [140]. Moreover, in order to increase the performance achieved by pure vision systems in real scenarios, multi-modal sensing should be considered [183], [184]. Examples include the integration of 2D laser scanners with cameras [183] or cameras and microphones [185].

Datasets and their field of applicability have been discussed throughout the survey and are summarized in Tables II, IV, and VI. Each dataset generally comprises a specific facet of human behavior understanding. Annotated datasets that combine detection, activity and interaction analysis are still missing and would greatly facilitate progress in this field whose building components have been traditionally analyzed separately. Moreover, especially in the area of action and interaction recognition, datasets are generally small, making it difficult to use a relevant number of examples for learning and testing. As human analysis datasets grow larger and more complex, it becomes harder to perform accurate annotation, which is paramount for training. For this reason, crowdsourcing annotation of datasets is growing in popularity [186]. Interesting quality assessment for crowdsourced object annotations have been proposed [187], illustrating the need for large and reliably annotated datasets to help increasing the performance of algorithms.

REFERENCES


<table>
<thead>
<tr>
<th>Ref.</th>
<th>Name</th>
<th>Features</th>
<th>A</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>[78]</td>
<td>KTH dataset</td>
<td>2391 sequences at 25 fps, downsammped to $160 \times 120$ pixels.</td>
<td>Y</td>
<td>25 subjects repeating six actions (walking, jogging, running, boxing, hand waving and hand clapping) in four scenarios</td>
</tr>
<tr>
<td>[136]</td>
<td>Weizmann dataset</td>
<td>90 sequences, of resolution $80 \times 144$ at 50 fps</td>
<td>Y</td>
<td>9 people performing 10 natural actions: run, walk, skip, jump-in-place, jumping-jack, jump-forward-on-two-legs, wave-one-hand, wave-two-hands, bend, gallop-sideways.</td>
</tr>
<tr>
<td>[139]</td>
<td>INRIA Xmas</td>
<td>429 sequences of resolution $390 \times 291$ at 25 fps</td>
<td>Y</td>
<td>5-view recording of 13 daily-live motions performed each 3 times by 11 actors. The actors choose freely position and orientation</td>
</tr>
<tr>
<td>[140]</td>
<td>MuhAVi</td>
<td>238 sequences of resolution $704 \times 576$</td>
<td>Y</td>
<td>8-views recording of 17 action classes performed by 14 actors</td>
</tr>
<tr>
<td>[111]</td>
<td>UCF YouTube database</td>
<td>1100 sequences of resolution $320 \times 240$ at 30 fps</td>
<td>Y</td>
<td>Non-staged, from YouTube, activities include cycling, diving, jumping and horse riding</td>
</tr>
<tr>
<td>[138]</td>
<td>UCF50</td>
<td>50 actions</td>
<td>Y</td>
<td>Extension of UCF Youtube database</td>
</tr>
<tr>
<td>[137], [112]</td>
<td>Hollywood Human Action datasets</td>
<td>Samples from 12 movies, at 25 fps.</td>
<td>Y</td>
<td>Non-staged human actions from feature films, sitcoms and news segments</td>
</tr>
<tr>
<td>[141]</td>
<td>3DPES</td>
<td>Recording over several days of more than 200 people</td>
<td>Y</td>
<td>Multi-camera videos with non-overlapped field of view, designed for re-identification</td>
</tr>
<tr>
<td>[142]</td>
<td>MIT trajectory dataset</td>
<td>Contains more than 40000 trajectories</td>
<td>–</td>
<td>Videos recorded from a camera monitoring a parking lot.</td>
</tr>
<tr>
<td>[143]</td>
<td>EIPFP</td>
<td>about 92000 trajectories, 8 millions of targets</td>
<td>N</td>
<td>Data recorded over months of working days.</td>
</tr>
<tr>
<td>[170]</td>
<td>BEHAVE</td>
<td>10 types of interactions</td>
<td>P</td>
<td>Types of interactions: in group, approach, walk together, split, ignore, following, chase, fight, run together, and meet</td>
</tr>
<tr>
<td>[116]</td>
<td>UT Interaction</td>
<td>20 videos, 6 types of interactions</td>
<td>Y</td>
<td>Small dataset for simple human interactions (hug, hand-shake, push, kick, punch, point)</td>
</tr>
<tr>
<td>[172]</td>
<td>MOCAP</td>
<td>about 110 videos</td>
<td>Y</td>
<td>3D dataset including two-person interactions</td>
</tr>
<tr>
<td>[171]</td>
<td>TV Human Interactions Dataset</td>
<td>300 videos, 4 interactions</td>
<td>Y</td>
<td>Short clips from TV-Shows including kiss, hug, high-five, hand-shake</td>
</tr>
<tr>
<td>[152]</td>
<td>RGBD-HuDaAct</td>
<td>RGB-D data in 1189 videos (46 hours) at 30 fps and resolution $640 \times 480$</td>
<td>Y</td>
<td>One single subject is present in each action. Each sequence spans about 30 to 50 seconds</td>
</tr>
<tr>
<td>[134]</td>
<td>CAD-60</td>
<td>RGB-D data in 60 videos at 30 fps and resolution $640 \times 480$</td>
<td>Y</td>
<td>3 to 4 common household activities identified in each location about 45 seconds of data for each activity and each person</td>
</tr>
<tr>
<td>[169]</td>
<td>CAD-120</td>
<td>RGB-D data in 120 activities, at 30 fps and resolution $640 \times 480$</td>
<td>Y</td>
<td>tracked human skeleton joints and tracked object bounding boxes are annotated</td>
</tr>
<tr>
<td>[174]</td>
<td>LIRIS</td>
<td>RGB-D data in 828 actions at 25 fps and $640 \times 480$ (RGB-D) + $720 \times 576$ (RGB)</td>
<td>Y</td>
<td>In addition to RGB-D standard video of the same scenes, recorded with a consumer camcorder, for comparisons between higher quality RGB and RGB-D techniques</td>
</tr>
</tbody>
</table>


