Video-Based Human Behavior Understanding: A Survey

Paulo Vinicius Koerich Borges, Nicola Conci, and 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 3-D 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 unambiguous 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.**

Index Terms—**Behavior analysis, computer vision, human detection, video analysis.**

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 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 of 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 intrabody 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].

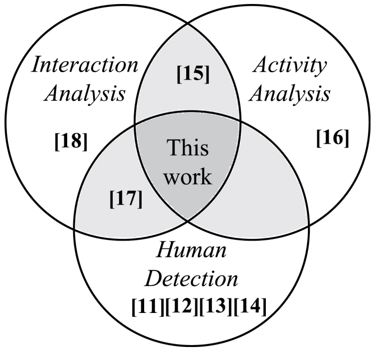


Fig. 1. Visual description of the contents of this paper in relation to previous surveys (indicated by the references).

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] surveyed features and model motion trajectories focusing on trajectory analysis. Moeslund *et al.* [16] covered 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] reviewed 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

Manuscript received February 18, 2013; revised April 28, 2013; accepted May 23, 2013. Date of publication June 20, 2013; date of current version November 1, 2013. This paper was recommended by Associate Editor S. Yan.

P. V. K. Borges is with the Autonomous System Laboratory, ICT Centre, Commonwealth Scientific and Industrial Research Organisation, Pullenvale, QLD 4069, Australia (e-mail: paulo.borges@csiro.au).

N. Conci is with the Department of Information Engineering and Computer Science, University of Trento, Trento 38123, Italy (conci@disi.unitn.it).

A. Cavallaro is with the Centre for Intelligent Sensing, Queen Mary University of London, London, U.K. (andrea.cavallaro@eecs.qmul.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2013.2270402

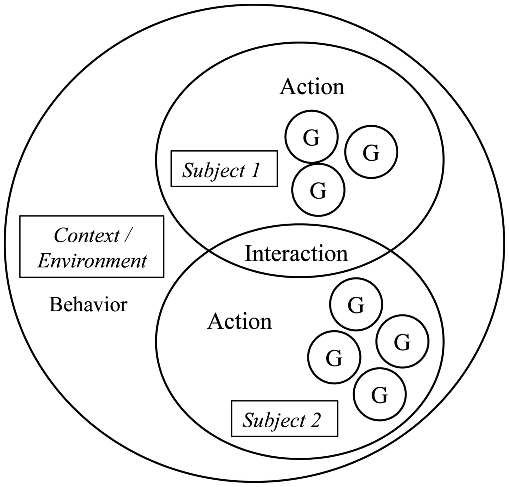


Fig. 2. Visual representation of the definitions for human behavior understanding as a combination of gestures (G), actions and interactions in a two-person scenario.

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 (Fig. 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 challenges.

The paper is organized as follows. In Section II, we introduce 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. Fig. 2 illustrates the relationship between these groups in a two-person scenario.

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. Gestures 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 decodable 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 characterizing visual elements, the recognition of behaviors requires a joint analysis of the (image) content and the (scene) context [22]. While actions can be analyzed in terms of motion and appearance features, the analysis of behaviors also requires information about the context and other factors influencing behaviors. Examples of contextual information include place and presence of other objects. Fig. 3 illustrates how the same spatio-temporal features corresponding to the action of running lead to different behaviors depending on the context.

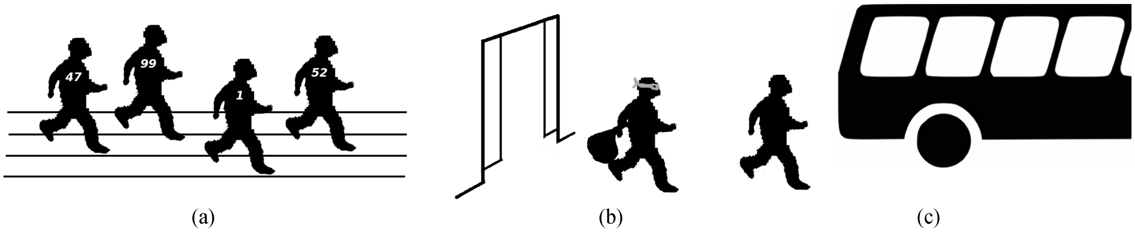
An interaction happens when other subjects or objects become a distinctive element to interpret the behavior of a person. We can distinguish between interactions with objects and social interactions. Interactions with objects happen when for example people use kitchen appliances, manipulate doors, or operate ATMs. Social interactions can be divided into two-person and group interactions (Fig. 4). A social interaction implies a feedback mechanism that contributes to establishing a two-way communication [23], [24]. Examples of interactions are shaking hands and fighting, but also a person stealing a bag from another, since stealing serves as a stimulus to establish a relationship between the two people. Note that this definition differs from that of [20], which made no distinctions between activities and interactions. Group interactions extend the concept of one-to-one interactions to multiple people sharing a common objective, such as a group marching and a group meeting. The common objective is in general regulated by a set of implicit or explicit rules, which are leveraged to alternate the action-reaction loop. As the number of people increases, it becomes more difficult to isolate them and to understand what they are doing, mainly due to occlusions and the difficulty in interpreting the role of each person in the group.

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 specific case of object detection in still images and are used 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.

A. Appearance-Based Methods

Algorithms based on appearance features are specialized detectors trained on large pedestrian databases [11], [12]. These algorithms generally scan each frame, searching for patterns that match predefined models. Appearance-based methods can be directly applied to nonstatic cameras (e.g., for automatic driver-assistance) [25], [26]. Effective representations include the histogram of oriented gradients (HOG) [27], the Viola and Jones algorithm [28], and Haar-like features [29].

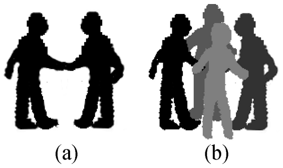


(a) (b) (c)

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 (FRAMES/S); HOG: HISTOGRAM OF ORIENTED GRADIENTS; SVM: SUPPORT VECTOR MACHINE; MPL: MULTIPOSE LEARNING)

Ref.	Features	Classifier	DR	DT	Comments
[25]	Appearance, HOG	linear and kernel SVM	89	-	Landmark algorithm, extended in [26],[27]
[28]	Appearance, shape (local and global)	Probabilistic top-down segmentation	71	-	Good results for partially occluded humans
[26]	HOG and pictorial structures [29]	SVM	82	-	Uses rendered synthetic pedestrian models for training
[27]	Edge orientation histograms, a variation of HOG	SVM	80	5	Part-based: lower dimensionality than traditional HOG; better performance on INRIA dataset
[30]	Motion History Image	Metric thresholding	-	30-50	Uses motion characteristics of legs and vehicles
[31]	Motion power spectral analysis	Spectral similarity comparison	91	1.1	Not only detection but also activity classification
[32]	Contour motion and Haar-like features	Boosting	90+	24	Combined motion and static image analysis. Very good results on the INRIA dataset
[33]	Skeleton extremities	Metric based	98	25	Good results but skeleton analysis requires high quality segmentation
[34]	Maximal principal gait angle	cyclic pattern in human movement	90	10	Use phase-locked-loop
[35]	Optical flow and HOG	Linear Kernel SVM and MPL Boost [36]	95+	-	Very good results on the Caltech and TUD-Brussels datasets
[37]	Motion point trajectories	Random Decision Forest	-	2-5	Moving camera and near range
[38]	Appearance, overall blob motion	Kernel SVM	88+	15	Blob motion statistics combined with appearance methods
[39]	Infrared HOG and Haar	Boosting	93	20	Tuned HOG detector, optimized to the infrared domain
[40]	HOG adapted to 3D	Linear SVM	85	33	Combination of 2D and 3D data, adapting HOG to RGB-D
[41]	3D surface model	Model fitting metric	98.4	-	3D surface model for the head using RGB-D combined with 2D contour model



(a) (b)

Fig. 4. Examples of (a) one-to-one interaction and (b) group interaction.

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 [30], [31]. Another popular approach is the combination of local and global cues via a probabilistic top-down segmentation [32]. A combination of multiple features, such as silhouette, appearance, holistic, and part-based, can be used as input to a SVM classifier [33], [34]. The Viola and Jones algorithm, used initially in face detection, can be trained for pedestrian detection [35]. Haar-like features have also been used in combination with covariance features [29], leading to computationally efficient algorithms with performance comparable to HOGs in the INRIA dataset [36].

Benchmark papers on appearance-based people detection introduce a database with richly annotated videos and discuss evaluation measures for performance comparison [13], [14]. Pishchulin *et al.* [37] 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 [38]. Although joint segmentation and detection processes have been proposed [32], [39], 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 undersegmentation 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 [40].

When observing cyclic motion, gait is analyzed based on pixel or region-wise oscillations such that the general statistical

TABLE II

DATASETS FOR HUMAN DETECTION. (KEY. A: ANNOTATION; Y: YES; N: NO; P: PARTIAL)

Ref.	Name	Features	A	Comments
[71]	MIT Pedestrian Database	923 images, with windows around the subjects	N	Pose of people limited to frontal and rear views. Images scaled to 64 × 128
[73], [72]	USC Pedestrian Detection Testset	369 images, with 816 humans	Y	Images collected from the Internet and from the CAVIAR video database [76], with frontal/rear view walking/standing human
[49]	INRIA Dataset	Approx. 2500 images, 128 × 94	Y	Refined combination of images from different datasets (including Internet), with pedestrians with bounding box area > 100 pixels
[74]	ETH Dataset	12.298 pedestrians in approx. 2.000 frames	Y	Two-camera dataset. The unbayered images, the camera calibration, and annotation are provided for both cameras. Annotated pedestrians are taller than 50 pixels
[13], [14]	Caltech Pedestrian Detection Benchmark	10 hours of 640 × 480 video	Y	Recorded from moving vehicle in urban area. 350,000 bounding boxes with 2300 unique pedestrians annotated

periodic behavior is used for classification. Early works performed a discrete Fourier transform (DFT) to quantify pixel oscillations [41], [42]. Variations of the method analyze the power spectral similarity in the walking pattern [43] or the amount of change in a motion history image [44]. An alternative to analyzing the full pixel intensity information is to high-pass filter the image, observing only a contour motion feature [45]. Cutler and Davis [46] look for the gait period by calculating a similarity matrix for every image pair in a sequence. He *et al.* [47] 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 [48] and autoregressive moving average models (ARMA) for estimating frequencies of sinusoidal signals. For example, Quinn and Hannan [49] 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 [50]. Ran *et al.* [51] discuss the methods above addressing the specific problem of periodicity estimation for pedestrian detection. Other works analyzed the periodic change in the optical flow [52], [53]. Perbet *et al.* [54] use a number of point trajectories and identify those that 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 [55], [56]. The adaptation of a generic pedestrian detector to a specific traffic scene can be performed [56], adding for example information about common pedestrian 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 [57]. Liu *et al.* [58] 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 [59]. Moreover, infrared imaging presents particular contrast characteristics that can be exploited [60], [61]. Although simple brightness analysis can be inefficient for characterizing people due to the polarity switch phenomenon [61], employing appropriate statistical models, such as the Infinite Gaussian mixture model [60] for the contrast around potential areas yields effective classification.

C. Hybrid Methods

Efficient people detection solutions combine appearance and motion [35], [62]. 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 (frames/s) 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 [28] operating on differences between pairs of images [35] and training using AdaBoost [63]. 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 [35] can be improved by one order of magnitude at the expense of using a higher number of frames [64]. A related approach is also used by Dalal [62] and Bouchirika *et al.* [65], combining HOGs with motion and gait patterns. The concept can be extended to also include behavior as a cue in the detection process [66], 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.* [67] 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 [68], where the direction of depth changes are locally encoded. An accuracy of 85% in ranges up to 8 m is achieved.

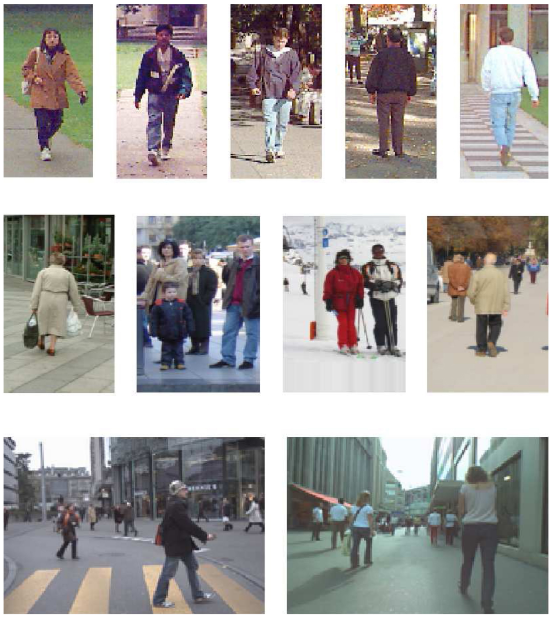


Fig. 5. Samples of different datasets used for training and evaluation of human detection algorithms. Scenarios considered range from simple cropped images (row A) to wide views with multiple people and complex background (row C).

D. Datasets

Datasets for testing human detection algorithms include the MIT pedestrian database [69], the USC Pedestrian Detection Test Sets [70], [71], the CALTECH database [13], the INRIA person dataset [36], and the ETH dataset [72]. 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 × 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 [13], [70] contain lower resolution pedestrians in larger images, recorded in cluttered urban areas. These datasets contain rich annotation, including occlusion information.

The PETS database [73] contains single and multiview footages including outdoor and indoor scenarios. The ETH dataset [72] 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 Fig. 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 [74], [75], silhouettes [76], and long-term analysis of motion [77]. Semantics and context-based techniques can be used to support and increase the system reliability [78], [79]. 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 [80], [81] and spatio-temporal interest points (STIPs) [75], [82]–[87]. Laptev and Lindeberg [88] originally presented STIPs for action recognition as a space-time variation of the Harris corner detector [89]. 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 [83], [90]. 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 Schuldt *et al.* [75] considered individual feature detections to be independent and built a single histogram for the full video sequence. Ullah *et al.* [91] performed video segmentation prior to STIP detection to eliminate misplaced interest points. Alternative approaches use several binning designs [92], [93] 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.* [94] introduced 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 [80], [81], [95]. Wang *et al.* [80], [95] used 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 showed that motion boundaries encoded along the trajectories can outperform state-of-the-art trajectory descriptors, on the KTH [75], Youtube [96], Hollywood2 [97], and UCF Sport datasets [98].

An action spectrogram for interest points in 3-D volumes is used by Chen and Aggarwal [99], 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 [100], with 70% accuracy in UT Dataset [101].

Given the redundancy of information collected through spatio-temporal features, Castrodad and Sapiro [102] 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.* [77] 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

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)

Ref.	Representation	Classifier	Key Aspects
[83]	MEI and MHI	"Pooled" Mahalanobis distance	Temporal templates and dynamic matching. Real time and multi-view
[84]	Fourier Descriptors	SVM and HMM	90% DR on ad-hoc dataset with 101 activities (5 subjects)
[85]	GMM to model salient postures	EM-based learning	Tolerance to noise, robustness across subjects and datasets
[86]	Spatio-temporal template (star-figure)	GMM framework	GMM features arranged in a two dimensional map, transformed into a gray level image
[87]	Hierarchical	χ^2 test	Automatic model adaptation; unsupervised behavior analysis and abnormality detection
[88]	Action graphs and 3D points	Bi-gram maximum likelihood decoding	Sparse sampling of 3D points to characterize the 3D shape of each salient postures
[89]	HMM with multiple independent observations (HMM-MIO)	HMM	Robustness to outliers, dimensionality reduction, handles sparse observations; performance comparable to discriminative classifiers
[90]	Self similarity matrix	Nearest neighbor	Self similarity matrices combined over time. 100% accuracy on KTH and Weizmann datasets
[91]	Motion trajectory	SCFG, Allen's temporal logic, Multi-Thread Parsing	Error detection and recovery, automatic rules induction
[80]	3D trajectories and pose	Random Forest and 3D-MRF	Group analysis as the spatial distribution of people and its temporal evolution
[92]	Motion trajectories	SVM, MRF	Classification based on trajectory grouping
[93]	Low or mid-level features	SVM, Dynamic Programming	Joint segmentation and classification of actions
[82]	Template-based action detector	SVM	High-level representation. Invariance to scale, viewpoint and tempo
[94]	Motion trajectories	Spectral analysis of graphs	Trajectories as transitions of graph nodes
[95]	Dense optical flow	Motion boundary histogram	Dense optical flow trajectories and an efficient tracker and descriptor, segmenting foreground motion from background motion
[96]	3D image patches	Sparse coding	Dictionary learning to extract action primitives and used as basis vectors. Classification obtained by reconstruction

dataset [103] with actions, such as standing, queuing, walking, and talking, they obtain an average accuracy of 82%. Raptis *et al.* [104] 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.* [105] present a framework for joint segmentation and recognition of actions based on multiclass 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 the Weizmann dataset and 42.4% on the Hollywood dataset. Daldoss *et al.* [106] 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.* [107] transformed the motion trajectories in a set of basic motion patterns and used 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 [108] 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 [78], [79], [109]. Gupta *et al.* [109] used 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.* [78] 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 [110] 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 [111]. 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 [112], Sadanand and Corso [79] proposed 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 [113], where the semantics are incorporated to his previous work [114] 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.* [115] used 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 [116], [117]. Alternatively a dynamic model of the action of interest can be generated using characteristics from each silhouette [76], [118].

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) [116]. Hu moments [117] 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.* [82] 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.* [119] used a hierarchical approach based on silhouettes to detect falls, whereas Li *et al.* [120] extended the silhouette approach to 3-D sensors.

Dynamic models of actions generated by extracting characteristics from each silhouette generally employ statistics-based techniques, such as conditional random fields (CRF) [121], [122] and hidden Markov models (HMM) [118], [123]. The features extracted from each silhouette capture the shape of the body and possible local motion. Kellokumpu *et al.* [124] 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 behavior of whole bodies, an alternative is to perform parts-based analysis [125], [126], where template matching is combined with a pictorial structure model to detect and localize actions.

Sun *et al.* [127] 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 [16], [20], [128], [129], 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 five-point star [128]. Most statistics are derived from the absolute and relative motion of those points with respect to the star centroid. The statistics can be analyzed directly or can serve as input to a classifier.

When subjects are close to the camera, using RGB-D, Shotton *et al.* [130] predict the 3-D 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.* [131], [132] 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 subactivities 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 [75] and the Weizmann dataset [133] 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% [126], [134] in the KTH and Weizmann datasets, which provide an initial basis for comparison among algorithms [123]. More realistic datasets include UCF50 [135] and Hollywood [134]. Other staged databases include INRIA Xmas [136] and MuHAVi [137], which contain multicamera views. The UCF YouTube database [96] has large variations in camera motion, object appearance, and illumination conditions, as illustrated in Fig. 6. Large variations are also found in the Hollywood Human Action datasets [97], [134], which contain thousands of video movies clips corresponding to more than 20 hours of data representing 12 human actions classes, and ten 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 3-D People Surveillance Dataset (3DPES) [138], the MIT trajectory dataset [139], and the Edinburgh Informatics Forum Pedestrian Database (EIFPD) [140]. 3DPES [138] 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 [139] contains more than 40 000 trajectories collected over five days from a single camera monitoring a parking lot. EIFPD [140] covers several months of observation resulting in about 1000 trajectories/day. Although data do not include the ground truth, information

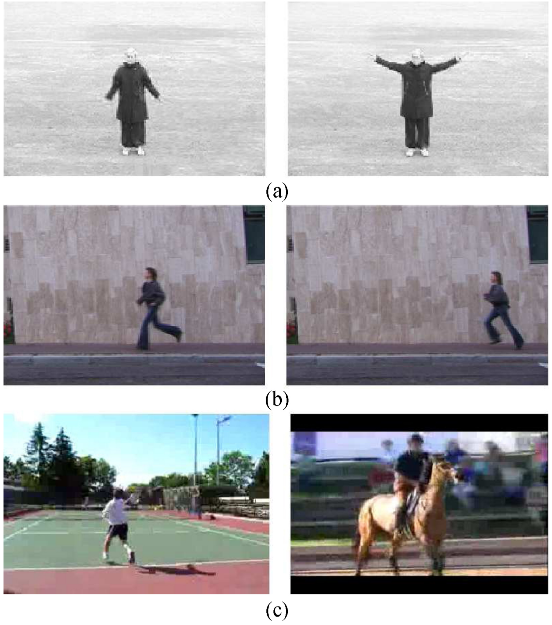


Fig. 6. Samples of different datasets used for training and evaluation of human activity classification algorithms. (a) and (b) Staged scenarios. (c) Real sequences obtained from the Internet. Dataset Key: (a) KTH [75]. (b) WEIZMANN [133]. (c) YouTube [96].

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 Australian Sign Language (ASL) [141], the Face and Gesture Recognition Working Group (FGnet) [142]–[144], Pointing’04 [145], Cambridge Gesture Database [146], and the ChaLearn Gesture Challenge dataset [147]. The ASL [141] 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 [142] provided by FGnet. The videos contain seven two-handed gestures, collected using seven different subjects, and recording ten videos per gesture per subject. The same working group proposed in [148] additional datasets for hand posture recognition, as well as a specific dataset for hand gesture recognition [144]. This dataset includes deictic and symbolic gestures for 400 samples. The benchmark data of Pointing’04 workshop [145] include ground-truth information indicating where users are pointing. Nine static and four dynamic gestures are presented and annotated in [143], including besides the identifiers for event and user, the gesture, action start, stroke start, stroke end, and gesture end. 900 image sequences of nine gesture classes defined by three primitive hand shapes and three primitive motions are presented in [146]. Each class contains 100 image sequences (five different illuminations, 10ten arbitrary motions, two subjects). The target task for the data set is to classify different shapes as well as different motions at a time.

Recent RGB-D datasets include the RGBD-HuDaAct [149] and the ChaLearn Gesture Challenge [147] 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 s, totaling approximately 46 hours of video at 640 × 480 resolution. Each subject performs the same activity 2–4 times. In the ChaLearn Gesture Challenge dataset [147], 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 [150], [151]. Proxemics exploits the so-called social space (the space between subjects) 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.* [152] assimilate the human behavior in a social context as a summation of forces that 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.* [153] used 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 [154]. Pellegrini *et al.* [155] applied this model to crowded scenes, considering single moving entities as agents. The motion of each agent is driven by its destination, and planned to prevent collisions with other moving objects. Every agent is associated with a set of features, comprising position, speed, and direction of motion. Path

TABLE IV

DATASETS FOR GESTURE RECOGNITION. (KEY. A: ANNOTATION; Y: YES; N: NO; P: PARTIAL; CGD: CAMBRIDGE GESTURE DATABASE; CGC: CHALEARN GESTURE CHALLENGE)

Ref.	Name	Features	A	Comments
[144]	ASL	95 signs, 27 examples per sign	Y	Samples of Auslan (Australian Sign Language) signs captured from a native signer
[145]	FGNet	10 postures, 24 people, 3 backgrounds for static; 1 sequence, 4 gestures for dynamic	Y	Static and dynamic hand posture database
[146]	FGNet	7 gestures, 7 users, 10 videos per gesture	Y	Two-handed gestures
[147]	FGNet	about 400 videos	Y	Hand posture recognition, including deictic and symbolic gestures
[148]	FGNet	8 videos, each with user pointing in 8 directions	Y	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.
[149]	FGNet	13 gestures, 9 static, 4 dynamic, 16 users	Y	Static and dynamic gestures performed under different illumination conditions by different subjects.
[150]	CGD	900 sequences, 9 gestures, defined by 3 primitive hand shapes and 3 primitive motions.	Y	Sequences recorded in front of a fixed camera having roughly isolated gestures in space and time. Videos are uniformly resized into $20 \times 20 \times 20$
[151]	CGC	about 50.000 images	P	The dataset includes high quality RGB videos complemented by depth information of single users.

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 [156] describe the coupling between multiple object states using Coupled HMMs on relative spatio-temporal features among the people under analysis. Rota *et al.* [157] recognized normal and abnormal two-people interactions exploiting proxemics cues combined with motion information, also revealing the intentionality of a specific interaction. Zen *et al.* [158] identified proxemics cues in a close-range analysis to discriminate personality traits, such as neuroticism and extraversion, and used 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.* [159] 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 [160], where features characterizing each category are identified based on the trajectories of the subjects.

The social interaction model proposed in [155] and [161] 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 CRF. 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 [162]. 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 [163], 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 3-D 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 [164] 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 3-D spatial dimensions and to the temporal dimension to detect a feature point. A feature descriptor then clusters the intensity and depth gradients within a 4-D 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 [165] 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.* [166] use a mobile robot and consider the problem of labeling subactivities 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 [167], the UT interaction dataset [101], the TV human interactions Dataset [168], and the CMU MOCAP [169] (Fig. 7). The BEHAVE dataset [167] 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 (FI), RunTogether (RT), and Meet (M). A smaller dataset is the UT interaction dataset for the SDHA contest [101] that covers continuous

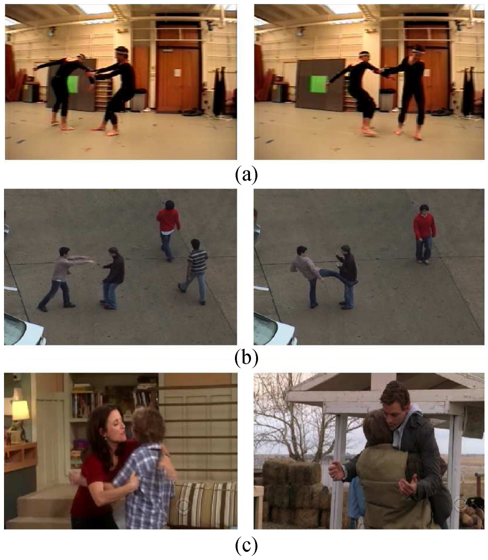
TABLE V

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)

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

executions of six 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) [168] is used to analyze interactions occurring in TV footages [170]. The dataset is annotated. Finally, the CMU MOCAP [169] is an annotated 3-D 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 [131] and CAD-120 [166], and the LIRIS [171]. 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 s 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 subactivities 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 subactivity labels, tracked human skeleton joints, and tracked object bounding boxes. This dataset combines actions, gestures, and interactions in its annotations.



(a)

(b)

(c)

Fig. 7. Samples from different datasets used for training and evaluation of interaction classification algorithms. Examples include pulling, fighting, and hugging. Dataset key: (a) MOCAP [169]. (b) UT [101]. (c) TV HID [170].

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

VI. Conclusion

This survey presented an extensive overview of video-based behavior understanding, its related definitions, and an analysis of datasets used for benchmarking. The goal was 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 3-D observations, datasets and annotation.

From a representation viewpoint, and in spite of their simplicity, low-level features (such as pixels), but also

TABLE VI

DATASETS FOR ACTION AND INTERACTION RECOGNITION. (KEY. A: ANNOTATION; Y: YES; N: NO; P: PARTIAL)

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

spatio-temporal features as STIPs, revealed excellent performances (e.g., [99], [102]), achieving 100% accuracy on some of the common benchmark datasets. Low-level features benefited from the fact that they were generally easy to extract. However, they were 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 [98] or objects level [107], [111]. Also in this case, efficient performances can be achieved, as demonstrated by Sadanand and Corso [79], who reached 98.2% on the KTH dataset and 95% on the UCF Sports dataset. Enriching the descriptors with contextual information can also provide significant improvements in the classification performances. This can be achieved by complementing the features with semantics [109], 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 [172] have demonstrated their robustness in dealing with noise both in data acquisition and behavior interpretation [173]. Examples include action [77], [105], [174], and interaction analysis [161], [175].

The problem of reconstructing the motion of a 3-D articulated body from 2-D point correspondences from a single camera for behavior understanding has gained increased attention [176], [177]. However, methods still require capturing the object at a high resolution and proximity to the camera. New opportunities are offered by devices, such as RGB-Depth (RGB-D) sensors (e.g., Asus Xtion [178] and Microsoft Kinect [179]), that simplify tasks at close range for human behavior analysis [67], [131], [132]. Another natural extension is the use of multiple cameras [72], [136], [137]. Moreover, to increase the performance achieved by pure vision systems in real scenarios, multimodal sensing should be considered [180], [181]. Examples include the integration of 2-D laser scanners with cameras [180] or cameras and microphones [182].

Datasets and their fields of applicability were 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 [183]. Interesting quality assessment for crowdsourced object annotations have been proposed [184], illustrating the need for large and reliably annotated datasets to help increase the performance of algorithms.

References

[1] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 893–908, May 2008.

[2] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Human computing and machine understanding of human behavior: A survey," in *Proc. Artif. Intell. Human Comput.*, pp. 47–71, 2007.

[3] A. Shirai, E. Geslin, and S. Richir, "WiiMedia: Motion analysis methods and applications using a consumer video game controller," in *Proc. ACM SIGGRAPH Symp. Video Games*, 2007, pp. 133–140.

[4] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3-D facial expression database for facial behavior research," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recogn.*, Apr. 2006, pp. 211–216.

[5] C. Kidd, R. Orr, G. Abowd, C. Atkeson, I. Essa, B. MacIntyre, E. Mynatt, T. Starner, and W. Newstetter, "The aware home: A living laboratory for ubiquitous computing research," in *Proc. Cooperative Build. Integr. Inform., Org., Arch.*, Oct. 1999, pp. 191–198.

[6] Z. Niu, X. Gao, and Q. Tian, "Tactic analysis based on real-world ball trajectory in soccer video," *Pattern Recogn.*, vol. 45, no. 5, pp. 1937–1947, May 2012.

[7] C. G. M. Snoek, K. E. A. van de Sande, O. de Rooij, B. Huurnink, J. C. van Gemert, J. R. R. Uijlings, J. He, X. Li, I. Everts, V. Nedovic, M. van Liempt, R. van Balen, F. Yan, M. A. Tahir, K. Mikolajczyk, J. Kittler, M. de Rijke, J. M. Geusebroek, T. Gevers, M. Worring, A. W. M. Smeulders, and D. C. Koelma, "The mediamill TRECVID 2008 semantic video search engine," 2009.

[8] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 436–450, Mar. 2012.

[9] A. Yao, J. Gall, and L. Van Gool, "Coupled action recognition and pose estimation from multiple views," *Int. J. Comput. Vision*, vol. 100, no. 1, pp. 16–37, 2012.

[10] S. Ba and J. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 101–116, Jan. 2011.

[11] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detecion: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[12] D. Gerónimo, A. M. López, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[13] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recogn.*, Jun. 2009, pp. 304–311.

[14] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state-of-the-art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[15] B. Morris and M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1114–1127, Aug. 2008.

[16] T. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vision Image Understand.*, vol. 104, nos. 2–3, pp. 90–126, 2006.

[17] R. Poppe, "A survey on vision-based human action recognition," *Image Vision Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[18] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surveys*, vol. 43, no. 3, p. 16, 2011.

[19] Y. Wu and T. Huang, "Vision-based gesture recognition: A review," in *Proc. Gesture-Based Commun. Human Comput. Interaction*, 1999, pp. 103–115.

[20] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[21] A. Maslow, R. Frager, and J. Fadiman, *Motivation and Personality*, vol. 2. New York, NY, USA: Harper and Row, 1970.

[22] J. Baird and D. Baldwin, "Making sense of human behavior: Action parsing and intentional inference," in *Proc. Intentions Intentionality: Foundations Soc. Cogn.*, 2001, pp. 193–206.

[23] J. Thibaut and H. Kelley, *The Social Psychology of Groups*. New York, NY, USA: Wiley, 1959.

[24] W. Doise and A. Palmonari, *Social Interaction in Individual Development*, vol. 3. Cambridge, U.K.: Cambridge Univ., 1984.

[25] D. Gavrila and S. Munder, "Multicue pedestrian detection and tracking from a moving vehicle," *Int. J. Comput. Vision*, vol. 73, no. 1, pp. 41–59, 2007.

[26] A. Howard, L. Matthies, A. Huertas, M. Bajracharya, and A. Rankin, "Detecting pedestrians with stereo vision: Safe operation of autonomous ground vehicles in dynamic environments," in *Proc. 13th Int. Symp. Robotics Res.*, 2007, pp. 26–29.

[27] N. Dalal, B. Triggs, I. Rhone-Alps, and F. Montbonnot, "Histogram of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recogn.*, vol. 1. Jun. 2005, pp. 886–893.

[28] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, vol. 1. 2001, pp. I-511–I-518.

[29] S. Paisitkriangkrai, C. Shen, and J. Zhang, "Fast pedestrian detection using a cascade of boosted covariance features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1140–1151, Aug. 2008.

[30] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2009.

[31] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2009, pp. 1014–1021.

[32] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recogn.*, vol. 1. Jun. 2005, pp. 878–885.

[33] Z. Lin and L. Davis, "A pose-invariant descriptor for human detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 423–436.

[34] M. Enzweiler and D. M. Gavrila, "Combination of feature extraction method for SVM pedestrian detection," *IEEE Trans. Intell. Transport. Syst.*, vol. 8, no. 2, pp. 292–307, Jun. 2007.

[35] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *Int. J. Comput. Vis.*, vol. 63, no. 2, pp. 153–161, 2005.

[36] INRIA. (2012) *Person Dataset* [Online]. Available: http://pascal.inrialpes.fr/data/human/

[37] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormahlen, and B. Schiele, "Learning people detection models from few training samples," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 1473–1480.

[38] T. Bouwmans, F. E. Baf, and B. Vachon, *Statistical Background Modeling for Foreground Detection: A Survey*, vol. 4. Singapore: World Scientific Publishing, 2010, ch. 3, pp. 181–199.

[39] B. Wu and R. Nevatia, "Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses," *Int. J. Comput. Vision*, vol. 82, no. 2, pp. 185–204, Apr. 2009.

[40] P. Borges, "Pedestrian detection based on blob motion statistics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 2, pp. 224–235, Feb. 2013.

[41] J. J. Little and J. E. Boyd, "Recognizing people by their gait: The shape of motion," *Videre J. Comput. Vision Res.*, vol. 1, no. 2, pp. 1–32, 1998.

[42] P. S. Tsai, M. Shah, K. Keiter, and T. Kasparis, "Cyclic motion detection for motion based recognition," *Pattern Recogn.*, vol. 27, no. 12, pp. 1591–1603, Dec. 1994.

[43] Q. Meng, B. Li, and H. Holstein, "Recognition of human periodic movements from unstructured information using a motion-based frequency domain approach," *Image Vision Comput.*, vol. 24, no. 8, pp. 795–809, Aug. 2006.

[44] S. Johnsen and A. Tews, "Real-time object tracking and classification using a static camera," in *Proc. IEEE Int. Conf. Robotics Autom. Workshop People Detect. Track.*, May 2009.

[45] Y. Liu, X. Chen, H. Yao, X. Cui, C. Liu, and W. Gao, "Contour-motion feature (CMF): A space-time approach for robust pedestrian detection," *Pattern Recogn. Lett.*, vol. 30, no. 2, pp. 148–156, 2009.

[46] R. Cutler and L. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 781–796, Aug. 2000.

[47] Y. Chen, Q. Wu, and X. He, "Motion based pedestrian recognition," in *Proc. Int. Congr. Image Signal Process.*, 2008, pp. 376–380.

[48] J. E. Boyd, "Synchronization of oscillations for machine perception of gaits," *Comput. Vision Image Understand.*, vol. 96, no. 1, pp. 35–59, Oct. 2004.

[49] B. G. Quinn and E. J. Hannan, *The Estimation and Tracking of Frequency*. Cambridge, U.K.: Cambridge Univ. Press, 2001, pp. 143–162.

[50] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing*. New York, NY, USA: McGraw-Hill, 2000.

[51] Y. Ran, I. Weiss, Q. Zheng, and L. S. Davis, "Pedestrian detection via periodic motion analysis," *Int. J. Comput. Vision*, vol. 71, no. 2, pp. 143–160, Feb. 2007.

[52] S. Walk, N. Majer, K. Schindler, and B. Schiele, "New features and insights for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recogn.*, Jun. 2010, pp. 1030–1037.

[53] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2003, pp. 726–733.

[54] F. Perbet, A. Maki, and B. Stenger, "Correlated probabilistic trajectories for pedestrian motion detection," in *Proc. IEEE Int. Conf. Comput. Vision*, Sep.–Oct. 2009, pp. 1647–1654.

[55] X. Wang, X. Ma, and W. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 539–555, Mar. 2009.

[56] M. Wang and X. Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 3401–3408.

[57] J. Rosell, G. Andreu, A. Rodas, V. Atienza, and J. Valiente, "Feature sets for people and luggage recognition in airport surveillance under real-time constraints," in *Proc. VISIGRAPP*, 2008, pp. 662–665.

[58] X. Liu, L. Lin, S. Yan, H. Jin, and W. Tao, "Integrating spatio-temporal context with multiview representation for object recognition in visual surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 393–407, Apr. 2011.

[59] J. Ge, Y. Luo, and G. Tei, "Real-time pedestrian detection and tracking at nighttime for driver-assistance systems," *IEEE Trans. Intell. Transport. Syst.*, vol. 10, no. 2, pp. 283–298, Jun. 2009.

[60] T. Elguebaly and N. Bouguila, "A nonparametric Bayesian approach for enhanced pedestrian detection and foreground segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 21–26.

[61] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Comput. Vision Image Understand.*, vol. 106, nos. 2–3, pp. 288–299, 2007.

[62] N. Dalal, "Finding people in images and videos," Ph.D. thesis, Inst. Nat. Polytechnique de Grenoble, Isère, France, 2006.

[63] Y. Freund and R. Schapire, "A desicion-theoretic generalization of online learning and an application to boosting," in *Computational Learning Theory*. Berlin, Germany: Springer, 1995, pp. 23–37.

[64] M. Jones and D. Snow, "Pedestrian detection using boosted features over many frames," in *Proc. IEEE Int. Conf. Pattern Recogn.*, Dec. 2008, pp. 1–4.

[65] I. Bouchrika, J. Carter, M. Nixon, R. Morzinger, and G. Thallinger, "Using gait features for improving walking people detection," in *Proc. IEEE Int. Conf. Pattern Recogn.*, Aug. 2010, pp. 3097–3100.

[66] W. Ge, R. Collins, and R. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, May 2012.

[67] L. Xia, C. Chen, and J. Aggarwal, "Human detection using depth information by Kinect," in *Proc. Workshop Human Activity Understand. 3-D Data Conjunction IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 15–22.

[68] L. Spinello and K. O. Arras, "People detection in RGB-D data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 3838–3843.

[69] MIT. (2000). *CBCL Pedestrian Database* [Online]. Available: http://cbcl.mit.edu/cbcl/software-datasets/PedestrianData.html

[70] B. Wu and R. Nevatia, "Cluster boosted tree classifier for multiview, multipose object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2007, pp. 1–8.

[71] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. IEEE Int. Conf. Comput. Vision*, vol. 1. Oct. 2005, pp. 90–97.

[72] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2007, pp. 1–8.

[73] PETS. (2005). *Computational Vision Group—University of Reading, U.K.* [Online]. Available: http://www.cvg.cs.rdg.ac.uk

[74] I. Laptev, "On space-time interest points," *Int. J. Comput. Vision*, vol. 64, no. 2, pp. 107–123, 2005.

[75] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. IEEE Int. Conf. Pattern Recogn.*, vol. 3. Aug. 2004, pp. 32–36.

[76] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 1992, pp. 379–385.

[77] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 3273–3280.

[78] L. Lin, H. Gong, L. Li, and L. Wang, "Semantic event representation and recognition using syntactic attribute graph grammar," *Pattern Recogn. Lett.*, vol. 30, no. 2, pp. 180–186, 2009.

[79] S. Sadanand and J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2012, pp. 1234–1241.

[80] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 3169–3176.

[81] S. Wu, O. Oreifej, and M. Shah, "Action recognition in videos acquired by a moving camera using motion decomposition of Lagrangian particle trajectories," in *Proc. IEEE Int. Conf. Comput. Vision*, Nov. 2011, pp. 1419–1426.

[82] D. Chen, S. Shih, and H. Liao, "Human action recognition using 2-D spatio-temporal templates," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2007, pp. 667–670.

[83] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. 2nd Joint IEEE Int. Workshop Visual Surveillance Performance Eval. Tracking Surveillance.*, Oct. 2005, pp. 65–72.

[84] J. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[85] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proc. IEEE 11th Int. Conf. Comput. Vision*, Oct. 2007, pp. 1–8.

[86] J. Liu and M. Shah, "Learning human actions via information maximization," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2008, pp. 1–8.

[87] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Proc. IEEE Int. Conf. Comput. Vision*, Sep.–Oct. 2009, pp. 1593–1600.

[88] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2003, pp. 432–439.

[89] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vision Conf.*, Manchester, U.K., vol. 15. 1988, p. 50.

[90] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2009, pp. 1948–1955.

[91] M. Ullah, S. Parizi, and I. Laptev, "Improving bag-of-features action recognition with non-local cues," in *Proc. Brit. Mach. Vision Conf.*, vol. 2. 2010, p. 7.

[92] D. Han, L. Bo, and C. Sminchisescu, "Selection and context for action recognition," in *Proc. IEEE Int. Conf. Comput. Vision*, Sep.–Oct. 2009, pp. 1933–1940.

[93] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2009, pp. 2004–2011.

[94] B. Chakraborty, M. Holte, T. Moeslund, J. Gonzalez, and F. Roca, "A selective spatio-temporal interest point detector for human action recognition in complex scenes," in *Proc. IEEE Int. Conf. Comput. Vision*, Nov. 2011, pp. 1776–1783.

[95] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[96] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2009, pp. 1996–2003.

[97] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2009, pp. 2929–2936.

[98] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: A spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2008, pp. 1–8.

[99] C. Chen and J. Aggarwal, "Modeling human activities as speech," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 3425–3432.

[100] M. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *Proc. IEEE Int. Conf. Comput. Vision*, Nov. 2011, pp. 1036–1043.

[101] M. S. Ryoo and J. K. Aggarwal. (2010) *UT-Interaction Dataset, ICPR Contest on Semantic Description of Human Activities (SDHA)* [Online]. Available: http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html

[102] A. Castrodad and G. Sapiro, "Sparse modeling of human actions from motion imagery," *Int. J. Comput. Vision*, vol. 100, no. 1, pp. 1–15, 2012.

[103] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, Sep.–Oct. 2009, pp. 1282–1289.

[104] M. Raptis, I. Kokkinos, and S. Soatto, "Discovering discriminative action parts from mid-level video representations," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2012, pp. 1242–1249.

[105] M. Hoai, Z. Lan, and F. De la Torre, "Joint segmentation and classification of human actions in video," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 3265–3272.

[106] M. Daldoss, N. Piotto, N. Conci, and F. G. B. De Natale, "Learning and matching human activities using regular expressions," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 4681–4684.

[107] Z. Zhang, T. Tan, and K. Huang, "An extended grammar system for learning and recognizing complex visual events," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 240–255, Feb. 2011.

[108] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, and N. Tishby, "Detecting anomalies in people's trajectories using spectral graph analysis," *Comput. Vision Image Understand.*, vol. 115, no. 8, pp. 1099–1111, 2011.

[109] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis, "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2009, pp. 2012–2019.

[110] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, to be published, 2013.

[111] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3-D human pose annotations," in *Proc. IEEE Int. Conf. Comput. Vision*, Sep.–Oct. 2009, pp. 1365–1372.

[112] L. Li, H. Su, E. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, Dec. 2010, pp. 1378–1386.

[113] R. Messing, "Human activity recognition in video: Extending statistical features across time, space and semantic context," Ph.D. dissertation, University of Rochester, Rochester, NY, USA, 2011.

[114] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Proc. IEEE Int. Conf. Comput.Vision*, Sep.–Oct. 2009, pp. 104–111.

[115] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *Proc. Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 3281–3288.

[116] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[117] M. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[118] J. Davis and A. Tyagi, "Minimal-latency human action recognition using reliable-inference," *Image Vision Comput.*, vol. 24, no. 5, pp. 455–472, 2006.

[119] F. Nater, H. Grabner, and L. Van Gool, "Exploiting simple hierarchies for unsupervised human behavior analysis," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2010, pp. 2014–2021.

[120] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3-D points," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2010, pp. 9–14.

[121] L. Wang and D. Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2007, pp. 1–8.

[122] X. Ji, H. Liu, and Y. Li, "Viewpoint insensitive actions recognition using hidden conditional random fields," in *Proc. Knowl. Based Intell. Inf. Eng. Syst.*, Sep. 2010, pp. 369–378.

[123] O. Concha, D. Xu, R. Yi, Z. Moghaddam, and M. Piccardi, "HMM-MIO: An enhanced hidden Markov model for action recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 62–69.

[124] V. Kellokumpu, M. Pietikäinen, and J. Heikkilä, "Human activity recognition using sequences of postures," in *Proc. IAPR Conf. Mach. Vision Applicat.*, 2005, pp. 570–573.

[125] Y. Ke, R. Sukthankar, and M. Hebert, "Event detection in crowded videos," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2007, pp. 1–8.

[126] Y. Wang and G. Mori, "Hidden part models for human action recognition: Probabilistic vs. max-margin," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1310–1323, Jul. 2011.

[127] C. Sun, I. Junejo, and H. Foroosh, "Action recognition using rank-1 approximation of joint self-similarity volume," in *Proc. IEEE Int. Conf. Comput. Vision*, Nov. 2011, pp. 1007–1012.

[128] H. Fujiyoshi and A. Lipton, "Real-time human motion analysis by image skeletonization," in *Proc. IEEE Workshop Appl. Comput. Vision*, Oct. 1998, pp. 15–21.

[129] W. Li, Z. Zhang, and Z. Liu, "Expandable data-driven graphical modeling of human actions based on salient postures," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1499–1510, Nov. 2008.

[130] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 1297–1304.

[131] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," in *Proc. AAAI Workshop Pattern, Activity Intent Recogn.*, 2011, pp. 47–55.

[132] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *Proc. IEEE Int. Conf. Robotics Autom*, May 2012, pp. 842–849.

[133] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. IEEE Int. Conf. Comput. Vision*, vol. 2. Oct. 2005, pp. 1395–1402.

[134] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2008, pp. 1–8.

[135] K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," in *Proc. Mach. Vision Applicat.*, 2012, pp. 1–11.

[136] D. Weinland, R. Ronfard, and E. Boyer, “Free viewpoint action recognition using motion history volumes,” *Comput. Vision Image Understand.*, vol. 104, no. 2, pp. 249–257, 2006.

[137] S. Singh, S. Velastin, and H. Ragheb, “MuHAVI: A multicamera human action video dataset for the evaluation of action recognition methods,” in *Proc. IEEE Int. Comput. Soc. Conf. Adv. Video Signal Based Surveillance*, 2010, pp. 48–55.

[138] D. Baltieri, R. Vezzani, and R. Cucchiara, “3DPEs: 3-D people dataset for surveillance and forensics,” in *Proc. 1st Int. ACM Workshop Multimedia Access 3-D Human Objects*, Nov. 2011, pp. 59–64.

[139] X. Wang, K. T. Ma, G.-W. Ng, and E. Grimson, “Trajectory analysis and semantic region modeling using a nonparametric Bayesian model,” in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2008, pp. 1–8.

[140] EIFPD. *Edinburgh Informatics Forum Pedestrian Database* [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING/

[141] M. W. Kadous, “Temporal classification: Extending the classification paradigm to multivariate time series,” Ph.D. dissertation, University of New South Wales, Sydney, New South Wales, Australia, 2002.

[142] FGnet. *Two Handed Gesture Dataset* [Online]. Available: http://www-prima.inrialpes.fr/FGnet/data/04-TwoHand/main.html

[143] M. Holte and M. Stoerring. (2004). *Pointing and Command Gestures Under Mixed Illumination Conditions: Video Sequence Dataset* [Online]. Available: http://www-prima.inrialpes.fr/FGnet/data/03-Pointing/index.html

[144] S. Marcel, O. Bernier, J. Viallet, and D. Collobert, “Hand gesture recognition using input-output hidden Markov models,” in *Proc. IEEE Int. Conf. Autom. Face Gesture Recogn.*, Mar. 2000, pp. 456–461.

[145] FGnet. *FGnet Pointing’04 Benchmark Data* [Online]. Available: http://www-prima.inrialpes.fr/Pointing04/data-hand.html

[146] T. Kim, S. Wong, and R. Cipolla, “Tensor canonical correlation analysis for action classification,” in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2007, pp. 1–8.

[147] ChaLearn. (2011). *Gesture Dataset* [Online]. Available: http://www-prima.inrialpes.fr/FGnet/data/03-Pointing/index.html

[148] S. Marcel. (2000). *Hand Posture and Gesture Datasets* [Online]. Available: http://www-prima.inrialpes.fr/FGnet/data/10-Gesture/gestures/main.html

[149] B. Ni, G. Wang, and P. Moulin, “RGBD-HUDAACT: A color-depth video database for human daily activity recognition,” in *Proc. Consumer Depth Cameras Computer Vision Workshop—Int. Conf. Comput. Vision*. Berlin, Germany: Springer, 2013, pp. 193–208.

[150] E. Hall, *The Hidden Dimension*, vol. 6. New York, NY, USA: Doubleday, 1966.

[151] E. Hall, *The Silent Language*. Harpswell, ME, USA: Anchor, 1973.

[152] D. Helbing and P. Molnar, “Social force model for pedestrian dynamics,” *Phys. Rev. E*, vol. 51, no. 5, p. 4282, 1995.

[153] R. Mehran, A. Oyama, and M. Shah, “Abnormal crowd behavior detection using social force model,” in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2009, pp. 935–942.

[154] P. Scovanner and M. Tappen, “Learning pedestrian dynamics from the real world,” in *Proc. IEEE Int. Conf. Comput. Vision*, Sep.–Oct. 2009, pp. 381–388.

[155] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, “You’ll never walk alone: Modeling social behavior for multitarget tracking,” in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2009, pp. 261–268.

[156] M. Taj and A. Cavallaro, “Recognizing interactions in video,” in *Proc. Intell. Multimedia Anal. Security Applicat.*, 2010, pp. 29–57.

[157] P. Rota, N. Conci, and N. Sebe, “Real time detection of social interactions in surveillance video,” in *Proc. 3rd Int. Workshop Anal. Retrieval Tracked Events Motion Imagery Streams, Eur. Conf. Comput. Vision*, 2012, pp. 111–120.

[158] G. Zen, B. Lepri, E. Ricci, and O. Lanz, “Space speaks: Towards socially and personality aware visual surveillance,” in *Proc. ACM MPVA*, 2010, pp. 37–42.

[159] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, and V. Murino, “Towards computational proxemics: Inferring social relations from interpersonal distances,” in *Proc. SocialCom*, 2011, pp. 290–297.

[160] B. Ni, S. Yan, and A. Kassim, “Recognizing human group activities with localized causalities,” in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2009, pp. 1470–1477.

[161] S. Pellegrini, A. Ess, and L. Van Gool, “Improving data association by joint modeling of pedestrian trajectories and groupings,” in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 452–465.

[162] X. Cui, Q. Liu, M. Gao, and D. Metaxas, “Abnormal detection using interaction energy potentials,” in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, Jun. 2011, pp. 3161–3167.

[163] J. Lei, X. Ren, and D. Fox, “Fine-grained kitchen activity recognition using RGB-D,” in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 208–211.

[164] H. Zhang and L. E. Parker, “4-dimensional local spatio-temporal features for human activity recognition,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 2044–2049.

[165] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[166] H. Swetha Koppula, R. Gupta, and A. Saxena, “Learning human activities and object affordances from RGB-D videos,” 2013.

[167] BEHAVE. (2007). *The Behave Dataset* [Online]. Available: http://homepages.inf.ed.ac.uk/rbf/BEHAVE

[168] A. Patron-Perez. (2010). *Tv Interactions Dataset* [Online]. Available: http://www.robots.ox.ac.uk/~vgg/data/tv_human_interactions/index.html

[169] CMU. (2013). *CMU Graphics Lab Motion Capture Database* [Online]. Available: http://mocap.cs.cmu.edu/

[170] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. Reid, “High five: Recognising human interactions in TV shows,” in *Proc. British Mach. Vision Conf.*, Aug. 2010, pp. 1–11.

[171] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandréa, C.-E. Bichot, C. Garcia, and B. Sankur, “The Liris human activities dataset and the ICPR 2012 human activities recognition and localization competition,” Tech. Rep. LIRIS RR-2012-004, Laboratoire Informatique en Images et Systmes Information, INSA de Lyon, France, Tech. Rep., 2012.

[172] M. J. Wainwright and M. I. Jordan, “Graphical models, exponential families, and variational inference,” *Found. Trends Mach. Learn.*, vol. 1, no. 1-2, pp. 1–305, Jan. 2008 [Online]. Available: http://dx.doi.org/10.1561/2200000001

[173] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert, “Activity forecasting,” in *Proc. Eur. Conf. Comput. Vision*. Springer, 2012.

[174] M. Amer, D. Xie, M. Zhao, S. Todorovic, and S. Zhu, “Cost-sensitive top-down/bottom-up inference for multiscale activity recognition,” in *Proc. Eur. Conf. Comput. Vision*. Springer, 2012.

[175] W. Choi and S. Savarese, “A unified framework for multitarget tracking and collective activity recognition,” in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 215–230.

[176] J. Valmadre, Y. Zhu, S. Sridharan, and S. Lucey, “Efficient articulated trajectory reconstruction using dynamic programming and filters,” in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 72–85.

[177] X. Wei and J. Chai, “Modeling 3-D human poses from uncalibrated monocular images,” in *Proc. IEEE Int. Conf. Comput. Vision*, Sep.–Oct. 2009, pp. 1873–1880.

[178] Asus. (2012). *Xtion Pro* [Online]. Available: http://www.asus.com/Multimedia/Motion_Sensor/Xtion_PRO/

[179] Microsoft. (2013). *Microsoft Kinect* [Online]. Available: http://www.xbox.com/en-GB/Kinect

[180] L. Oliveira, U. Nunes, P. Peixoto, M. Silva, and F. Moita, “Semantic fusion of laser and vision in pedestrian detection,” *Pattern Recogn.*, vol. 43, no. 10, pp. 3648–3659, 2010.

[181] A. Broggi, P. Cerri, S. Ghidoni, P. Grisleri, and H. Jung, “A new approach to urban pedestrian detection for automatic braking,” *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 594–605, Dec. 2009.

[182] M. Taj and A. Cavallaro, “Interaction recognition in wide areas using audiovisual sensors,” in *Proc. IEEE Int. Conf. Image Process.*, Sep.–Oct. 2012, pp. 1113–1116.

[183] C. Vondrick, D. Ramanan, and D. Patterson, “Efficiently scaling up video annotation with crowdsourced marketplaces,” in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 610–623.

[184] S. Vittayakorn and J. Hays, “Quality assessment for crowdsourced object annotations,” in *Proc. Brit. Machine Vision Conf.*, 2011, pp. 109–111.

Paulo Vinicius Koerich Borges received the B.E. and M.Sc. degrees in electrical engineering from Federal University of Santa Catarina, Brazil, in 2002 and 2004, respectively. In 2007, he received the Ph.D. degree from the Queen Mary University of London (QMUL), London, U.K.

From 2007 to 2008, he was a Post-Doctoral Researcher at QMUL, involved in video event detection. In 2001, he was a Visiting Research Student in the subject of image restoration with the University of Manchester, Manchester, U.K. Since 2009, he has been a Research Scientist with the Autonomous System Laboratory at Commonwealth Scientific and Industrial Research Organisation, Brisbane, Australia. He is also an Adjunct Lecturer with the School of Information Technology and Electrical Engineering, University of Queensland, Queensland, Australia. His research interests include visual-based robot localization, pedestrian tracking, video event detection/classification, video tracking, and digital watermarking, besides general image processing and pattern recognition.

Nicola Conci received the Ph.D. degree from University of Trento, Trento, Italy, in 2007.

In 2007, he was a Visiting Student with the Image Processing Laboratory, University of California, Santa Barbara, CA, USA. From 2008 to 2009, he was a Post-Doctoral Researcher with the Multimedia and Vision Research Group, Queen Mary University of London, London, U.K. Since 2009, he has been an Assistant Professor with the University of Trento. His research interests include machine vision for behavior understanding and human computer interfaces.

Dr. Conci received the Best Student Paper Award at the International ACM Conference Mobimedia held in Alghero, Italy, in 2006.

Andrea Cavallaro received the Laurea degree (*summa cum laude*) from University of Trieste, Trieste, Italy, in 1996, and the Ph.D. degree from Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2002, both in electrical engineering.

He was a Research Consultant with the Image Processing Laboratory, University of Trieste, in 1996 and 1998, focusing on compression algorithms for very low bitrate video coding. From 1998 to 2003, he was a Research Assistant with the Signal Processing Laboratory, EPFL. Since 2003, he has been with Queen Mary University of London, London, U.K., where he is a Professor of multimedia signal processing. He has authored over 130 papers and published two books entitled *Multicamera Networks* (Elsevier, 2009) and *Video Tracking* (Wiley, 2011).

Dr. Cavallaro was the recipient of a Research Fellowship from British Telecommunications (BT), London, U.K., in 2004 and 2005, three Student Paper Awards from IEEE International Conference on Acoustic, Speech, and Signal Processing, in 2005, 2007, and 2009, and the Best Paper Award from the IEEE Advanced Video and Signal-based Surveillance (AVSS) in 2009. He is an Elected Member of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee. He served as a Technical Chair for IEEE AVSS 2011, the Workshop on Image Analysis for Multimedia Interactive Services in 2010, and the European Signal Processing Conference in 2008, and as General Chair for the IEEE/ACM International Conference on Distributed Smart Cameras in 2009, the British Machine Vision Conference in 2009, and the IEEE AVSS in 2007. He is an Area Editor for the IEEE *Signal Processing Magazine* and an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and the IEEE TRANSACTIONS ON SIGNAL PROCESSING.