Automatic visual detection of human behavior: a review from 2000 to 2014

Palwasha Afsar[a,*], Paulo Cortez[a], Henrique Santos[a]

[a] *ALGORITMI Research Centre, Department of Information Systems, University of Minho, 4800-058 Guimarães, Portugal*

Abstract

Due to advances in information technology (e.g., digital video cameras, ubiquitous sensors), the automatic detection of human behaviors from video is a very recent research topic. In this paper, we perform a systematic and recent literature review on this topic, from 2000 to 2014, covering a selection of 193 papers that were searched from six major scientific publishers. The selected papers were classified into three main subjects: detection techniques, datasets and applications. The detection techniques were divided into four categories (initialization, tracking, pose estimation and recognition). The list of datasets includes eight examples (e.g., Hollywood action). Finally, several application areas were identified, including human detection, abnormal activity detection, action recognition, player modeling and pedestrian detection. Our analysis provides a road map to guide future research for designing automatic visual human behavior detection systems.

Keywords: Data mining, Human behavior, Literature review, Video analysis, Video databases

1. Introduction

Human detection and their corresponding behaviors have been studied under distinct perspectives in a wide variety of disciplines, such as psychology, biomechanics and computer graphics (Ko, 2008). According to (Moeslund et al., 2006; Poppe, 2010) an action primitive is an atomic movement that can be described at the limb level. An action consists of action primitives and describes a, possibly cyclic, whole-body movement. Finally, activities contain a number of subsequent actions, and give an interpretation of the movement that is being performed. The main research question underneath this work can be characterized as follows: given a succession of pictures with one or more persons performing an action, can a framework be outlined to distinguish: *who* is performing the action and *what* action was performed? In this paper, we focus on computational frameworks for the automatic detection of human behavior, which is a very recent and relevant research area due to its potential impact in a wide variety of human activities, such as gaming, visual surveillance and detection of elderly accidents.

[*] Corresponding author

Email addresses: palo_afsar77@yahoo.com (Palwasha Afsar), pcortez@dsi.uminho.pt (Paulo Cortez), hsantos@dsi.uminho.pt (Henrique Santos)

The particular interest in this area has dramatically increased with the advances of information technology. In particular, depth-imaging (3D) has substantially improved in the last few years, finally reaching an affordable consumer price (e.g., Kinect for Xbox 360). This is a multidisciplinary area, involving several fields such as artificial intelligence, data mining, psychology, biomechanics, pattern recognition and image processing. In effect, the combination of all these fields is necessary to tackle challenging tasks, such as real-time segmentation of changing scenes in natural environments, and non-rigid movement and self-occlusion.

While this is considered a trendy subject, whose interest will increase in the next few years, within our knowledge there are few recent review articles that cover well this area. Some surveys only target a portion of visual human behavior detection possibilities, such as: (Moeslund et al., 2006), which analyzed human motion capture from 2000 to 2006; (Yampolskiy and Govindaraju, 2008), which focused on behavioral biometrics from 1989 to 2007; and (Ko, 2008), which addressed video surveillance from 1985 to 2008.

Other surveys closely related to action and activity recognition has been done by (Turaga et al., 2008; Poppe, 2010). The review of Turaga et al. (2008) focus on high-level recognition of actions and activities (e.g., bending, walking, shaking hands) from 1988 to 2008, while the study of Poppe (2010) surveys low-level image representations for action recognition (e.g., optical flow, spatial-temporal volume) from 1992 to 2009. In this paper, we perform a systematic and more recent review (from 2000 to 2014) that includes both low-level and high-level methods for human behaviour recognition. Moreover, we also compare the distinct methods and highlight their advantages and limitations. Finally, we also present several updated examples of applications and the most commonly used datasets (with best performances so far achieved) in this topic.

A comprehensive search was made by analyzing recent papers, from 2000 to 2014, from high quality journals and conferences related with the addressed topic (e.g., Expert Systems with Applications, Computer Vision). The search was executed using six general scientific databases: ACM Digital Library (`dl.acm.org`), Elsevier (`www.elsevier.com`), IEEE Xplore Digital Library (`ieeexplore.ieee.org`), Springer Link (`link.springer.com`), MIT Press (`mitpress.mit.edu`) and Wiley Online Library (`onlinelibrary.wiley.com`). The list of keywords used in the search included combinations of the keywords "Human Behavior" or "Human Detection" with "Video" or "Data". The search was then filtered manually and reduced to 193 papers related to this review. The selected papers were classified into three main subjects: detection techniques, datasets and applications. Techniques that were related to detection were further divided into four categories, namely initialization, tracking, pose estimation and recognition. Moreover, a total of eight datasets were identified (e.g., Hollywood action). Finally, the applications were grouped into six main areas: human detection, abnormal activity detection, action recognition, player modeling and robotics, pedestrian detection and in-home scenarios, and person tracking. To summarize the review in terms of topics and their publishing year, Table 1 presents the evolution of the surveyed papers keywords that appear with five or more occurrences from 2000 to 2014.

This paper is organized as follows. Firstly, Section 2 introduces the visual detection of human behavior computational techniques. Then, Section 3 presents the main datasets used within the surveyed domain. Next, Section 4 lists examples of relevant human behavior detection applications. Finally, Section 5 concludes the review by performing a global analysis to the presented review and presenting future research implications.

Table 1: Automatic human behavior detection from video keywords by publication year.

Keywords	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	**Total**
Depth images	0	0	0	1	0	1	2	1	1	0	2	3	2	2	2	17
Stereo cameras	2	1	0	1	0	0	0	0	1	0	0	0	0	0	0	5
Kinect	0	0	0	0	0	0	0	0	0	0	1	3	2	2	1	9
Multiple cameras	0	0	0	1	0	1	2	1	0	2	0	1	1	1	2	12
3D	0	1	1	1	0	0	1	0	3	1	0	3	2	3	3	19
Video dataset	1	0	0	1	1	5	4	2	2	4	1	4	2	3	2	32
Background subtraction	0	0	0	1	0	3	3	2	1	1	1	1	0	1	5	19
Object tracking	0	0	2	1	1	4	4	3	2	2	3	4	1	4	5	36
Occlusion handling	0	0	1	2	0	2	2	2	4	3	3	3	3	2	4	31
Histogram Of Gradients (HOG)	0	0	0	0	0	4	5	3	0	2	2	2	0	0	3	21
Human motion-based features	1	1	0	0	1	3	3	3	1	3	3	2	2	2	3	28
Blobs	0	1	2	1	0	2	1	1	0	1	0	1	0	0	4	14
Neural Network (NN)	2	1	3	1	2	3	0	1	0	1	1	1	0	1	0	17
Support Vector Machine (SVM)	1	1	1	1	1	4	2	2	1	1	2	2	1	1	6	27
Vision	1	1	2	0	0	2	1	1	2	2	1	1	2	2	3	21
Behavior	0	0	1	2	2	4	2	0	2	0	0	1	1	2	5	22
Total	8	7	13	14	8	38	32	22	20	23	20	32	19	26	48	330

2. Techniques used for Human Behavior Detection from Video

Following the works of (Moeslund and Granum, 2001; Moeslund et al., 2006), the visual human behavior detection techniques were categorized into four groups:

- **Initialization** – in order for the system to process data, it needs to be initialized; e.g., a proper model of the system should be built;
- **Tracking** – the process of segmenting the subjects from the background and finding correspondences between segments in consecutive frames;
- **Pose** – the estimation of pose is carried out in corresponding frames (usually, a high level human model is used); and
- **Recognition** – recognizing the behavior, identity, and action of an individual or a group.

A detailed description is given in the next subsections for all these phases with a major focus towards high-level tasks. Each phase topic subsection ends with a discussion subsection that compares the different approaches.

2.1. Model Initialization

2.1.1. Main Approaches

Initialization of vision-based human motion capture frequently requires the meaning of a humanoid model approximating the appearance, shape, kinematic structure, and beginning posture of the subject to be tracked. Initialization requires prior knowledge of what constitutes an individual. Such knowledge can be separated into categories of (Moeslund et al., 2006): kinematic structure, 3D shape, color appearance and body part estimation.

The bulk of vision-based tracking frameworks acquire an initial humanoid kinematic structure incorporating a fix number of joints with specified degrees-of-flexibility. The kinematic initialization is then constrained to the estimation of limb lengths. Commercial marker-based motion capture frameworks normally oblige a fix grouping of movements which separate individual degrees-of-opportunity. Initialization of body pose and limb length from manually identified joint locations using monocular images has been addressed in several works (Taylor, 2000; Barrón and Kakadiaris, 2001; Barron and Kakadiaris, 2003; Parameswaran and Chellappa, 2004). A method for automatically initialization the kinematic structure of the upper body has been investigated by (Krahnstöver et al., 2003; Krahnstoever and Sharma, 2004) using motion segmentation of monocular video images. Song et al. (2003) presented an unsupervised learning algorithm, which uses point characteristic tracks from jumbled monocular video sequences to automate the process of developing triangulated models of entire body kinematics. Techniques that determine the kinematic structure from 3D shape sequences recreated from different perspectives have also been proposed (Cheung et al., 2003; Menier et al., 2006; Chu et al., 2003; Brostow et al., 2004). All of these approaches give a more general solution to the kinematic structure by directly deriving the structure from the scene itself.

Progressively, human motion capture sequences from commercial marker-based frameworks have been utilized to learn prior models of human kinematics and particular motions to give requirements constraints for ongoing tracking. Also, motion capture databases (Jang, 2002; Interactive, 2005) have been utilized for synthesizing picture sequences with know 3D posture correspondence to learn in advance the mapping from image to pose space for recreation.

Regarding 3D shape, representations have utilized either basic shape primitives (e.g., cylinders, cones, ellipsoids) or surfaces (e.g., polygonal mesh) to define a kinematic skeleton (Moeslund and Granum, 2001). A simultaneous capture from different aligned views has been proposed to accomplish a precise shape and appearance initialization. Plankers and Fua (2003) used an ellipsoidal ball representation for initiating the upper human body shape prior to tracking. Carranza et al. (2003) utilized a generic mesh model to multiple view silhouette image sequences of a single person in a fixed pose before doing tracking. Starck and Hilton (2003) reconstructed an entire body shape and appearance for an individual in a subjective stance by improving a generic mesh model as for stereo, silhouette, and feature correspondence obligations in various views. These model fitting methodologies give an exact parameterized rough guess of an individual assuming the expected state of the non-specific model in a sensible starting estimate. Model fitting systems usually accept short hair and fit clothing, which restrains their sweeping statement. The accessibility of sensors for entire body 3D scans also provides exact estimation of surface shape. Databases of 3D scans have additionally been utilized to learn statistical models of the inter-person variety in entire body shape (Allen et al., 2003; Magnenat-Thalmann and Seo, 2004).

Turning to color appearance initialization, it has mainly been focused around the picture set. Statistical models of color are usually utilized for tracking. Texture maps inferred from different perspective pictures were adopted in (Carranza et al., 2003; Starck and Hilton, 2003), Sidenbladh and Black (2001, 2003) modeled the probability of image perceptions for distinctive body parts. Roberts et al. (2002) used a similar approach for learning the probability of body part color appearance utilizing multimodal histograms on a 3D surface model. Results are given for 2D tracking of upper body in cluttered scenes.

Finally, regarding body part identification, a more recent trend has been body part indicators, which are then consolidated probabilistically to find individuals (Ronfard et al., 2002; Roberts et al., 2004; Micilotta et al., 2005; Ramanan et al., 2005). Machine learning methods, such as Adaboost, have been used to learn body part detectors, such as face hands, arms, legs and torso (Viola and Jones, 2001; Micilotta et al., 2005; Roberts et al., 2004). Ramanan et al. (2005) detect key-frame postures in walking arrangements and introduced a local appearance model to discover body parts at intermediate video frames. Lim et al. (2006) addressed the issue of changing appearance because of motion by demonstrating the dynamics of the appearance for walking people. The initiation of models that faultlessly show the change in appearance with respect to time because of wrinkles in clothes, hair, and subsequent change in body shape remains an open issue.

2.1.2. Discussion

Initialization of shape, appearance and motion is considered an important step in the automation of human motion. As shown in Table 2, significant advances have been made in such initialization. In particular, initialization of kinematic structure and pose for monocular sequences has been addressed by (Song et al., 2003). Also, a number of studies have proposed solution for initialization the kinematic structure from multiple-view image sequences using volumetric reconstruction (Menier et al., 2006). These techniques help in the automatic initialization of a model appearance to a specific individual.

Recent work on body part detectors has benefited from supervised learning approaches to discriminate an individual from background (Roberts et al., 2004; Ronfard et al., 2002). The issue of changes in appearance due to motion has been addressed by only a few studies. Moreover, the problem of initialization of appearance model for monocular pose and tracking is not yet solved and remains open for future research.

2.2. Tracking

The tracking in visual analysis of humans can be defined as two main processes: figure ground segmentation and temporal correspondence. The former is used to differentiate objects of interest (e.g., people) from the background image. The latter is used to detect the same objects of interest through a sequence of frames.

In this section, a higher emphasis is given towards segmentation, since it is an important step of tracking, although temporal correspondence is also covered.

2.2.1. Background Segmentation

Until the late 90s, background subtraction was adopted mainly in controlled indoor situations. Since then, interesting contributions have been proposed and are here classified into background representation, classification, updating and initialization.

Several background representation schemes have been adopted. The Mixture of Gaussians (MoG) is a standard representation that included mostly RGB color space. Still, other shade spaces, such as isolation of color and intensities (e.g., YUV) (Cucchiara et al., 2003) and

Table 2: Comparison of approaches for Model Initialization.

Reference	Approach	Advantages	Limitations
(Barrón and Kakadiaris, 2001)	Anthropometric measurements	Efficiently calculates pose	Requires human intervention
(Viola and Jones, 2001)	Cascade classifier, Adaboost	Extremely efficient, can handle complex backgrounds, high detection rate	The detectors need to be more independent for improved performance
(Allen et al., 2003)	Sparse 3D markers, template-based non-rigid registration	Robust to noise and missing data	Not efficient if poses are different
(Carranza et al., 2003)	Image-based modeling and rendering	View-invariant, runs offline, can handle broad range of motions	Cannot track all body poses
(Cheung et al., 2003)	Shape-from-silhouette (SFS)	Can handle dynamic human motion	–
(Chu et al., 2003)	Skeleton curve for kinematic postures	Can handle different human motions	Multiple cameras used
(Krahnstöver et al., 2003)	Automatic acquisition of kinematic model	Does not require prior shape and kinematic structure information	Can only handle single-view videos
(Song et al., 2003)	Decomposable triangulated graphs	Works on unlabeled data	–
(Starck and Hilton, 2003)	Model-based framework	Robust to visual ambiguities	Multiple cameras used
(Brostow et al., 2004)	*Spines* based on skeletal representation	Robust to noise	Requires multiple cameras, not view-invariant
(Krahnstoever and Sharma, 2004)	Model acquisition framework	Process human and non-human structures	Can handle occlusion only to some extent, requires input from user (not automatic)
(Roberts et al., 2004)	Learnt probabilistic model using partial configurations	Invariant to clutter, work on real world data, can handle complex backgrounds	–
(Lim et al., 2006)	Dynamic appearance modeling	Does not require prior knowledge of gender or type of activity, invariant to clutter and occlusion	Not accurate if activity changes (e.g., walking into running)
(Menier et al., 2006)	Shape-from-silhouette (SFS)/visual hull	View-invariant, robust to noise	Tracking algorithm needs to be improved

normalized RGB (McKenna et al., 2000) have also been adopted. Utilizing a MoG within a 3D color space corresponds to ellipsoids or spheres (contingent upon the presumptions on the covariance matrix) of the Gaussian representations (McKenna et al., 2000; Zhao and Nevatia, 2004). Other geometric representations are truncated cylinders (Kim et al., 2005) and truncated cones (Fihl et al., 2006). Thoughtfully diverse representations have likewise been produced. Elgammal et al. (2000) utilize a kernel-based approach that is able to handle cluttered backgrounds and non-static scenes (moving bushes). Their system adapt itself very quickly to the changing scenes which makes object detection mush faster while (Heikkila and Pietikainen, 2006) represented each background pixel by a bit sequence. Results show that the system provides many advantages when compared with the state-of-the art. Oliver et al. (2000) likewise utilize a pixel's neighbors to represent it. They apply an Eigen space representation of the background and identify new objects by contrasting the input picture with a picture recreated through the Eigen space. The most important aspect of their system is the detection of group interaction and the classification of the type of interaction. Heikkila and Pietikainen have likewise connected their texture operator for a spatio-temporal block-based (covering squares) background segmentation. The algorithm operates in real-time with a stationary camera and is able to handle multi-modal backgrounds and inherent scene changes (Heikkilä et al., 2004).

Other spatio-temporal methodologies are (Monnet et al., 2003; Zhong and Sclaroff, 2003) where an anticipated locale region found by an autoregressive methodology represents the background. The issue of background representation depends mainly on the speed of the implementation, accuracy and the application. This is true because the overall accuracy of the background subtraction is a combination of classification, representation, initialization and updating. Due to this reason, (Cucchiara et al., 2003) achieved good results on the advanced classification algorithm while he represented the background with only one pixel. MoG representation is the most commonly used method but for scenes having dynamic background, methodologies aiming at dynamic background should be used (Monnet et al., 2003; Sheikh and Shah, 2005).

Regarding background classification, false positives and negatives will regularly occur after a background subtraction (e.g., due to shadows) (Prati et al., 2003). Utilizing standard filtering methods focused on elements such as size, morphology and proximity, it is possible to improve the result (Elgammal et al., 2000; McKenna et al., 2000; Cucchiara et al., 2003; Zhao and Nevatia, 2004; Guha et al., 2005; Yang et al., 2005b). Markov Random fields have also been applied to distinguish foreground from background (Sheikh and Shah, 2005; Schindler and Wang, 2006). Recent techniques have attempted to use classifiers to separate the pixels into various sub-classes, such as unaltered background, changes because of auto iris, shadows, highlights and moving objects (Cucchiara et al., 2003; Chen et al., 2005). Classifiers have been focused on input features such as color gradients (McKenna et al., 2000), flow information (Cucchiara et al., 2003) and hysteresis thresholding (Eng et al., 2003).

A background model needs to be updated after an initialization stage. Prior methodologies assumed that no moving items are available in various consecutive frames and afterward take in the model parameters in this period. Yet, in real situations this assumption might be invalid and more recent strategies have concentrated on the neighborhood of moving objects. In MoG representation, moving object during initialization can be accepted up to some extent but this erroneous distribution will produce more false positives in classification process. An alternative methodology is to discover genuine foundation pixels, for instance by using a temporal median filter such that less than 50% of the values belong to foreground objects (Haritaoglu et al., 2000). Eng et al. (2003) join this approach with a skin detector to discover and remove people from the training images. Recent methods first divide the pixels in the initialization stage into temporal subintervals with similar values and then the "best" subinterval belonging to background is found as the subinterval with the minimum average motion (measured by optical flow) (Gutchess et al., 2001) or the subinterval with the most extreme ratio between the amount of samples in the subinterval and their variance (Wang and Suter, 2005, 2006).

Motion based figure-ground segmentation is focused around the idea that contrasts in consecutive images emerge from moving people, i.e., by discovering the movement you discover the human. The motion is measured using either flow or image differencing. Sidenbladh (2004) figures optical flow for a substantial number of picture windows, each one holding a walking human, and used Support Vector Machine (SVM) to detect strolling people in video. Given that optical flow might be noisy, an alternative is to measure image flow using higher-level entities. For instance, Gonzalez et al. (2003) track KLT-characteristics to acquire flow vectors, Sangi et al. (2001) concentrate flow vectors from displacements of pixel-blocks, and Bradski and Davis (2002) discover flow vectors as gradients in Motion History Images (MHI). Image differencing adjusts rapidly to changes in the scene, however pixels from a human that has not moved or are like their neighbors are not identified. Subsequently, an enhanced adaptation is to utilize three back-to-back pictures (Collins et al., 2000; Haritaoglu et al., 2000; Viola et al., 2003).

Segmentation focused around human appearance is based on the concept that the appearance of humans is distinct from the background. Also, distinct individuals have different appearances. The appearance-based segmentation methodologies work by building an appearance model for every human and after that either fabricating appearance models of the segmented foreground object in the current image and contrasting them with the predicted models, or by specifically segmenting the pixels in the current picture that belong to each model. Some of these systems are autonomous on the temporal context, implying that the strategies apply a general appearance model of a human, instead of techniques where the appearance model of the human is educated/overhauled focused around previous images in the current sequence. Appearance-based segmentation is divided into temporal context-free and those utilizing context.

Temporal context-free appearance segmentation routines are utilized to detect people in a still picture (Mohan et al., 2001), to detect people entering a scene (Okuma et al., 2004), or to index images in databases (Ozer and Wolf, 2002). Developments are basically on utilizing very large training datasets for adapting classifiers. For example, Okuma et al. (2004) utilize 6000 images to prepare an Adaboost-based machine learner. Some other examples are using DCT coefficients (Ozer and Wolf, 2002), utilizing partial-impediment taking care of body-part detectors(Mohan et al., 2001) or the block-based system by Utsumi and Tetsutani (2002). All these methods use a bounding box for detecting humans from the background. Yet, occlusion and drastic illumination changes will also affect the methods since they are not dynamic and do not adapt to changes in the scene.

In temporal context appearance models, the color of a human is mostly represented as either a color histogram (McKenna et al., 2000; Comaniciu et al., 2003; Hu et al., 2004; Okuma et al., 2004; Zhao and Nevatia, 2004; Xu and Puig, 2005) or a MoG (Khan and Shah, 2000; Kang et al., 2005; Roth et al., 2005; Yang et al., 2005a). Representing the entire human body with just one color model can be too coarse therefore recent developments tends to include spatial information. For instance utilizing a Correlogram, which is a co-occurrence matrix that communicates the likelihood of two distinctive colored pixels being found at a certain separation structure one another (Capellades et al., 2003). An alternate method for adding spatial information is to divide the human into various sub-regions and show each region with either a color histogram or a MoG (Mittal and Davis, 2003; Okuma et al., 2004; Roth et al., 2005; Yang et al., 2005a). Hu et al. (2004) utilized a versatile method to acquire three sub-regions representing the head, torso, and legs. A more general method is to model the human as various blobs where each one blob is a connected group of pixels having a comparable color (Khan and Shah, 2000; Park and Aggarwal, 2006).

The shape of a human is usually different from the shape of other objects in a scene. Under a temporal context-free approach,

Zhao and Thorpe (2000) used depth information to extract the silhouettes of people in the image. A Neural Network (NN) is trained on upright people and used to confirm whether the extracted silhouettes are really related to humans or not. The results of experiments performed shows that their system is able to run in real-time and can detect human in various shapes, size, clothing and also handle illumination and background changes. Leibe et al. (2005) developed a challenging system that can detect pedestrian in crowded scenes. The algorithm learns the layouts of walking people and store them as various templates. The proposed system is able to handle partial occlusion and can operate on small training sets. Wu and Yu (2006) take in a prior shape model for human edges and show it as a Boltzmann distribution in a Markov Field. Their system is also able to handle partial occlusion. Dalal and Triggs (2005) utilized SVM to distinguish people in a window of pixels. In (Dalal et al., 2006) the work is performed by including motion histograms. This takes into account people detection when the camera and/or background is moving. The algorithm produces promising results when experimented with challenging datasets. Histogram Of Gradients (HOGs) are identified with Shape Contexts (Belongie et al., 2001) and SIFT (Scale Invariant Feature Transformation) (Mittal and Davis, 2003).

When temporal context is contemplated, shape-based routines could be connected to track people with respect to time. Haritaoglu et al. (2000) developed a real-time system for detecting and tracking multiple people and monitoring their activities. It performs on a binary edge correspondence between the blueprints of the silhouettes in the last frame and the prompt surroundings in the current image. The system is able to recognize differents events like removal of an object, exchange of bags etc. Davis et al. (2000) utilized a Point Distribution Model (PDM) to show the layout of the human. The proposed framework offers robustness and can handle unrestricted camera angles. A comparable methodology is seen in (Koschan et al., 2003) where the dynamic shape model is applied to discover a fit in the current frame. Atsushi et al. (2002) used an ellipse to model the pose of human in the previous frame and foresee nine conceivable stances of the human into the current frame. Each of these is corresponded with the silhouettes in the current image so as to characterize the current stance of the human. The algorithm is able to detect multiple people in real-time. Krüger et al. (2005) correlated the extracted silhouette with a hierarchy of silhouettes of walking persons. At run-time, a Bayesian tracking structure simultaneously gauges the translation, scale, and sort of silhouette. In case of partial occlusions, the shape-based techniques fails due to lack of global shape information. Thus, recent studies include method to detect human only from a few parts of the whole shape Wu and Nevatia (2005).

Finally, turning to depth-based segmentation, systems are either built straightforwardly with respect to assessed 3D information for the scene (Ivanov et al., 2000; Haritaoglu et al., 2002; Hayashi et al., 2004; Yang et al., 2004; Lim et al., 2005) or in a roundabout way by joining distinctive camera views after features have been extracted (Mittal and Davis, 2002, 2003; Yang et al., 2003; Iwase and Saito, 2004). The improvements are essentially because of faster computers processing numerous camera inputs. Background subtraction might be affected by lighting changes. In this manner a depth-based methodology might be taken where the background is demonstrated as a depth model and contrasted with estimated depth information for each approaching frame to segment the foreground. A real-time dense stereo algorithm is still erroneous unless extraordinary hardware is used (Lim et al., 2005). A methodology to bypass this is the work by Ivanov et al. (2000) where an online depth map is not needed. Rather the mapping between pixels in two cameras is learned. Different developments in human detection focused around depth information incorporate the work by Haritaoglu et al. (2002), where depth information is obtained by roof-mounted cameras anticipated to the ground-plane. Humans are located by looking for a 3D head-shoulder profile. Mittal and Davis (2002, 2003) identify people utilizing an appearance-based system in every camera view. The center of humans found are combined with those found in another frames using region-based stereo constrained by the epipolar geometry. Yang et al. (2003) joined silhouettes from distinctive cameras into a visual structure. Size criterion and temporal history is used for correcting the incorrect interpretations. Iwase and Saito (2004) apply various cameras to detect and track numerous individuals. In every camera the feet of every individual are recognized utilizing background subtraction and information of the scene.

2.2.2. Temporal Correspondence

One of the essential roles of tracking technique is to discover the temporal correspondences. The majority of the past tracking algorithms were tested in controlled environments mainly focused on natural outdoor scenarios and challenges like presence of multiple people and occlusion. One problem is how to handle multiple people occluding one another and another one is to have a good figure-ground segmentation. Temporal correspondences before and after occlusion and during occlusion are discussed in this section.

Regarding the temporal correspondence before and after occlusion, a model of every individual must be built before any tracking can begin. Recent strategies aim to make the whole process automatic. One way is to search for (new) vast foreground objects conceivable close to the boundaries (Haritaoglu et al., 2000; McKenna et al., 2000; Antonini et al., 2006; Roth et al., 2005). Subsequently, an individual might be characterized as a foreground object detected far from any predictions (Capellades et al., 2003). At the point when the tracking has initiated the issue is to discover the temporal correspondences between predicted and measured states. Global optimization is another way that can additionally be applied. Polat et al. (2003) utilize a Multiple Hypothesis Tracker to construct diverse hypothesis each one clarifying all the predictions and estimations, and picks the hypothesis which is most likely.

Tracking during occlusion is a relatively new research issue. In recent frameworks, the first undertaking is to detect that an occlusion is present (Capellades et al., 2003). Khan and Shah (2000) recognize a non-occluded circumstance as one where the detected foreground object is a far way from another. Capellades et al. (2003) characterize a merge as a circumstance where the total number of foreground items has diminished and where two or more foreground objects from the previous frame overlap with one foreground object in the current frame. Distinctive methods for assigning pixels to people throughout impediment have also been proposed. A local approach is to assign every pixel to the most likely predicted model utilizing a probabilistic strategy (Khan and Shah, 2000; Park and Aggarwal, 2006). Local approaches are good for bypassing the occlusion problem but are also sensitive to noise hence they are combined with some post-processing to reassign the wrong pixels. Global methods attempt to arrange pixels focused around for instance the suspicion that individuals in a gathering are standing side by side regarding to the camera. This technique allows for defining vertical dividers between individuals. In the work by McKenna et al. (2000) the depth ordering is discovered explicitly. Throughout occlusion the probability of every pixel in the foreground item fitting in with an individual is computed utilizing a Bayes rule.

In (Roth et al., 2005) the depth ordering is focused around accepting a planar floor, where the closest object to the camera has the most vertical coordinate. Xu and Puig (2005) sum up this concept by utilizing projective geometry to discover the line in the picture that relates to the “horizon line” in the 3D scene. The item closest to the camera is found as the article closest to this horizontal line.

2.2.3. Discussion

The increased focus on surveillance applications has triggered the advances in figure-ground segmentation as summarized in Table 3. In order to have a fully automated system, operating in an uncontrolled environment, background segmentation plays an important role (Fihl et al., 2006). However, multiple cameras are required to cover the whole scene at a sufficient resolution.

Systems with self-calibrating cameras have also been studied by researchers (Atsushi et al., 2002), but no independent framework has been announced yet. An improvement in the field of segmentation is the use of spatial information for the color-based appearance models. Each foreground object can be divided into a number of regions, having a color representation (Hu et al., 2004; Khan and Shah, 2000; Mittal and Davis, 2003; Okuma et al., 2004; Park and Aggarwal, 2006; Roth et al., 2005; Yang et al., 2005a) or a correlograms (Capellades et al., 2003). This helps in efficient detection and tracking even in the presence of occlusion.

Significant research on natural scenes has led to improvements in temporal correspondence methods and occlusion. These advances are due to the use of probabilistic methods. For example, to segment individual pixels due to occlusion (Khan and Shah, 2000; McKenna et al., 2000; Park and Aggarwal, 2006) and also to use stochastic methods for handling multiple hypothesis and uncertainties (Hu et al., 2004; Okuma et al., 2004; Polat et al., 2003; Yang et al., 2005a).

Table 3: Comparison of approaches for Tracking

Reference	**Approach**	**Advantages**	**Limitations**
(Davis et al., 2000)	Point distribution model (PDM)	Efficient, can handle multiple camera motion and views	Does not take into account the change in shape of human model in general
(Elgammal et al., 2000)	Non-parametric background model with color information	Robust to small motion of bushes and trees, runs in real-time	Works with a stationary camera, not completely robust to motion
(McKenna et al., 2000)	Adaptive background subtraction with gradient and color information	Tracking in presence of occlusion	System will fail if two people havesame dressing
(Comaniciu et al., 2003)	Spatial masking with isotropic kernel	Invariant to camera motion, clutter and scale variation	Not robust to severe occlusion
(Zhong and Sclaroff, 2003)	Kalman filter with autoregressive moving average model (ARAM)	Can handle varying background e.g., moving clouds, trees etc.	Slow speed
(Heikkilä et al., 2004)	LBP operator	Invariant to illumination changes, preserves shape of object	Operates only with a stationary camera, cannot handle moving background
(Hu et al., 2004)	Bayesian network	Handles occlusion, efficient	–
(Okuma et al., 2004)	Mixture particle filters with Adaboost	Fully automatic, works in a cluttered environment	Fails at certain conditions
(Sidenbladh, 2004)	SVM with dense optical flow	Invariant to varying cloths, light, background, identity, weather	Only tested on outdoor video sequences
(Zhao and Nevatia, 2004)	Human shape model, stochastic sampling	Tracking in crowded environment with occlusion	–
(Chen et al., 2005)	Two cascaded classifiers with pixel classification	Works in presence of shadows and can handle complex scenes	Miss-classification of object as a background pixel
. (Guha et al., 2005)	Boolean predicates	Can handle occlusion upto an extent, ability to distinguish a variety of problem situations	Fails when complex motion or severe occlusion occurs
(Kang et al., 2005)	Spatio-temporal joint probability data association filter (JPDAF)	Robust tracking	Computationally expensive, does not use 3D information
(Kim et al., 2005)	Background subtraction with adaptive codebook	Efficient in speed and memory, handles long video sequences, invariant to moving background	–
(Krüger et al., 2005)	Model-based background with Bayes propagation	Robust to translation, scaling	Tested only on one person
(Sheikh and Shah, 2005)	Non-parametric density method	Invariant to motion	Degradation in performance due to camera motion
(Wang and Suter, 2005)	Median statistics with background initialization	Tolerate 50% noise in data	Not completely robust to noise
(Xu and Puig, 2005)	Spatial depth affinity metric (SDAM)	Works in presence of complex dynamic occlusion	May fail in dynamic occlusions which lasts for long time
(Yang et al., 2005b)	Background segmentation with merging and splitting detection	Does not require prior knowledge of shape or motion, work for long duration videos, handle complete occlusion	Two persons tracked as once in the presence of occlusion
(Fihl et al., 2006)	Multi-mode Codeword method	Can handle gradual or rapid changes	errors due to noisy foreground segmentation or insufficient tracking or appearance model
(Park and Aggarwal, 2006)	Gaussian mixture model+Attribute relational graph (ARG)+ multi association tracking (MMT)	Tolerate various views, multiple group interactions, invariant to person height and camera view	Low speed, poor tracking results for homogenous clothing

2.3. Pose Estimation

Pose Estimation alludes to the methodology of evaluating the arrangement of the underlying kinematic or skeletal articulation structure of an individual. Pose Estimation algorithms can be divided into three categories based on the human model usage (Moeslund and Granum, 2001; Moeslund et al., 2006):

- **Model free** – when no prior model is used and most techniques track body parts in 2D or map 2D sequences of picture perceptions into a 3D pose;

- **Indirect model use** – when a prior model within posture estimation is used (e.g., human body part labeling utilizing aspect ratios between limbs or posture distinguishment;

- **Direct model use** – when an explicit 3D geometric representation of human shape and kinematic structure is used to reproduce posture; most direct model approaches use an analysis-by-synthesis strategy to enhance the closeness between the model projection and observed pictures.

The three major pose estimation areas are: the incorporation of learning motion models in pose estimation to constrain the recovered 3D human motion; the introduction of probabilistic approaches to detect body parts and assemble part configurations in the model-free category; and the use of stochastic sampling techniques in model-based analysis-by-synthesis to improve robustness of 3D pose estimation.

2.3.1. Model Free

A late pattern to overcome confinements of tracking over long sequences has been the examination of direct posture detection on individual image frames. Two methods have been researched which fall into this class: probabilistic assemblies of parts, when individual body parts are initially detected and after that gathered to gauge the 2D pose; and example-based routines which specifically take in the mapping from 2D image space to 3D model space.

Probabilistic congregations of parts have been presented for direct bottom up 2D posture estimation by first locating likely areas of body parts and afterward gathering these to get the configuration which best matches the observations. A potential advantage of detection over tracking is that the stance might be assessed autonomously at each frame, permitting posture estimation for fast movements. Temporal information may be consolidated to gauge reliable stance arrangements over sequences. Using the concept of 'body plans' to represent humans, (Felzenszwalb and Huttenlocher, 2000; Ioffe and Forsyth, 2001) utilized pictorial structures to gauge 2D body part arrangements from image sequences. Probabilistic assemblies of body part detectors (face, hands, arms, legs, and torso) have been researched for bottom up estimation of entire body 2D stance in individual frames or sequences. Individual body parts are distinguished utilizing 2D shape (Roberts et al., 2004), SVM classifiers (Ronfard et al., 2002), Adaboost (Micilotta et al., 2005), and locally initialized appearance models (Ramanan et al., 2005).

Various example-based methods for human posture estimation have been proposed which contrast the observed image and a database of samples. To overcome limits of tracking, analysts have examined example-based methods which specifically lookup the mapping from silhouettes to 3D stance (Shakhnarovich et al., 2003; Agarwal and Triggs, 2004; Howe, 2004; Sminchisescu et al., 2005). Example-based methods show the mapping between images and pose space giving a compelling mechanism for straightforwardly evaluating 3D stance. Normally these methodologies endeavor rendering of motion capture data to give training examples with known 3D stance. A limitation of current example-based method is the limitation to the stances or motions utilized within training. Broadening to a more extensive vocabulary of movements may result in ambiguities in the mapping. Recent works have focused on pose estimation from a single image. Hua et al. (2005) present a method for 2D stance estimation from a single image utilizing bottom up features cues together with a Markov network to model part configurations.

2.3.2. Indirect Model Use

Several works have focused on direct reconstruction of both model shape and motion from the visual-hull without a prior model. For instance, Mikić et al. (2002, 2003) presented a coordinated framework for automatic recuperation of both a human body model and motion from numerous perspective image sequences. An extended Kalman filter is then utilized for human motion reconstruction between frames. A voxel labeling method is utilized to permit substantial information between inter-frame movements. The system performs efficiently for different motions like walking, dancing, running jumping for people having different shape sizes and belonging to various age groups. For the same purposeCheung et al. (2003) first reconstructed a model of the kinematic structure, shape, and appearance of an individual and afterward utilize this to gauge the 3D development. (Starck and Hilton, 2005) focused on another alternative method i.e., full 3D-to-3D flexible surface matching utilizing spherical mapping. The important aspect of their system as opposed to other systems is that it does not require a prior model of human shape for tracking.

2.3.3. Direct Model Use

Recreation of human posture from a solitary perspective image succession is significantly more troublesome than either the issue of 2D stance estimation or 3D posture estimation from numerous perspectives. To solve the uncertainty in monocular human movement recreation extra requirements on kinematics and movement are commonly adopted (Bregler et al., 2004). Probabilistic methods utilizing congregations of parts together with higher-level knowledge of human kinematics and shape have additionally been examined for single view 3D posture estimation. Monocular reconstruction of complex 3D human movement remains an open issue. Recent research has examined the utilization of learned motion models to give solid priors to oblige the pursuit.

There has been a growing focus towards the use of learned models of human posture and motion to compel vision-based reconstruction of human recognition from single or numerous perspectives. The accessibility of marker-based human motion capture information (Jang, 2002; Interactive, 2005) has prompted the utilization of learned models of human motion for both animation synthesis in animation illustrations and vision-based human motion synthesis. Inverse kinematics of human motion focused around learned models has already been presented in computer graphics (Grochow et al., 2004; Ong and Hilton, 2006). Consequent exploration has examined the utilization of learned movement models for 3D motion reproduction mainly from monocular image sequences to conquer the inalienable visual vagueness. Researches presenting the utilization of learned statistical models of human motion since 2000 has showed that utilizing solid motion priors encourages reconstruction of 3D posture sequences from monocular images. To date the simplification of these methods has been restricted to particular motion models with moderately little variety in movement and fixed transitions.

2.3.4. Discussion

Table 4: Comparison of approaches for Pose Estimation

Reference	**Approach**	**Advantages**	**Limitations**
(Felzenszwalb and Huttenlocher, 2000)	Pictorial structure	Efficient recognition for people and cars	–
(Ioffe and Forsyth, 2001)	Tree structured probabilistic model	Entirely automatic, handles different activities	Low configuration results for some scenarios
(Mikić et al., 2002)	Voxel data, extended kalman filter	Successfully tracks certain human poses	Multiple cameras used, tracking errors with waist rotation
(Ronfard et al., 2002)	SVM on articulated body model	Achieves good results for body parts in varying background and illumination	Worse results for torso and head, computationally expensive
(Shakhnarovich et al., 2003)	Parameter-sensitive hashing	Accurate estimation of human pose	–
(Agarwal and Triggs, 2004)	Relevance vector machine (RVM), regularized least square	Doesn't require prior labeling, robust estimation of pose	–
(Bregler et al., 2004)	Exponential maps and twist based acquisition technique	Robust to noise and occlusion, track walk and hopping cycles of human and animals	Differential method used which may fail if motion is large
(Grochow et al., 2004)	Scaled Gaussian process latent variable model	Pose recovery, real-time motion capture, automatic	–
(Roberts et al., 2004)	Learnt probabilistic model using partial configurations	Invariant to clutter, work on real world data, can handle complex backgrounds	Cannot handle high dimensional configurations with self occlusion and visual similarities
(Hua et al., 2005)	Hiddem Markov network, Monte Carlo algorithm	Robust in inferring 2D human pose from single images	Requires prior knowledge of human shape, manual labeling
(Micilotta et al., 2005)	RANSAC, Adaboost, skin color	Efficiently detects face and legs	Poor result for hand detection and background segmentation
(Ramanan et al., 2005)	Discriminative appearance model	Works in real-time, fast tracking for unusual interactions	Poor results if limbs are not located
(Sminchisescu et al., 2005)	Discriminative density propagation algorithm, Bayesian mixture of expert model (BME)	Efficiently detects human pose	Fails to track joint if they are occluded or due to folds in cloths and shadows
(Starck and Hilton, 2005)	Spherical matching, multi resolution coarse-to-fine optimization	Does not require prior shape model, can handle multiple-view videos	–
(Ong and Hilton, 2006)	Learnt inverse kinematics framework	Works well for human and animals, reconstruction of shapes from irregular terrains	Errors due to lack of diverse training data

As shown in Table 4, several methodologies have been proposed to the automatic estimation of human pose. In particular, algorithms based on 2D estimation of body parts (e.g., hands, face or limbs) in cluttered environment have also been widely studied (Felzenszwalb and Huttenlocher, 2000; Hua et al., 2005; Ioffe and Forsyth, 2001; Micilotta et al., 2005; Ramanan et al., 2005; Roberts et al., 2004; Ronfard et al., 2002). Similarly, example-based methods have been often used to learn the mapping from 2D image features such as silhouettes to 3D pose (Agarwal and Triggs, 2004; Brand, 1999; Howe, 2004; Shakhnarovich et al., 2003; Sminchisescu et al., 2005). Current example-based methods only work on a fix class of movements and range of viewpoints deployed in training. A future challenge is to make these methods viewpoint invariant and for general movements.

Model-based pose estimation focus on reliable recovery of complex movements using an analysis-by-synthesis methodology for estimation of 3D pose from multiple view-point (Carranza et al., 2003). Research on 3D pose estimation from monocular image using stochastic methods has also received significant improvements. However, monocular reconstruction of complex 3D human movement remains an open problem. The extension of learnt model for reconstruction of general human movement is also not solved. One of the limitations of the existing research that needs attention is the comparison of different techniques on a common dataset and evaluation of performance against ground-truth.

2.4. Recognition

The field of action recognition is quite old but still immature. This is an increasing area of interest in terms of research and it is related with a wide range of applications, such as: surveillance, medical studies and rehabilitation, robotics, video indexing and animations for films and games. This section is organized according to a visual abstraction hierarchy system yielding the accompanying: scene interpretation, where the whole picture is interpreted without distinguishing specific items or people; holistic recognition, where either the whole human body or individual body parts are requisitioned for recognition; and action primitives and grammars, where an action hierarchy of importance offers ascent to a semantic depiction of a scene.

2.4.1. Scene Interpretation

Numerous methods consider the camera view in general and endeavor to learn and perceive activities basically by observing the motion of objects without fundamentally knowing their identity. This is sensible in situations when the items are little enough to be shown as points on a 2D plane.

Stauffer and Grimson (2000) present a full scene interpretation framework which permits location of irregular circumstances. The framework extracts features, i.e., 2D position and speed, size and binary silhouettes. The system is able to track an object through the entire tracking sequence, however, it is unaware of the identity of the object. It is able to cope with changing lighting, repetitive motion from clutters and long-term scene changes. These can be used for scene interpretation. In Eng et al. (2003) a swimming pool surveillance framework is researched. From each of the discovered and tracked object's features, for example, posture, speed, submersion index, an activity index, and a splash index, are extracted. These features are fed into a multivariate polynomial system to recognize water crisis events. The developed algorithm achieves robust performance for different hostile environments faced by an outdoor swimming pool. Boiman and Irani (2007) approach the issue of detecting irregularities in a scene as an issue of composing newly observed data utilizing spatio-temporal patches fetched from previously seen visual samples. They extract small image and video patches which are utilized as local descriptors. They search for small patches with similar geometric configurations while allowing misalignment. This way they are able to quickly and efficiently look for different changes in behavior. Junejo et al. (2004) depict a method to focus on dynamic data for scene interpretation. Their strategy can recognize items traversing spatially different paths or items traversing spatially proximal paths but with distinctive spatio-temporal attributes. In (Chowdhury and Chellappa, 2003; Vaswani et al., 2003), action trajectories are modeled utilizing flexible shapes and a dynamic model that describes the variation in the shape structure. Chowdhury and Chellappa (2003) utilize a subspace technique to model activities as a linear blending of 3D basis shapes. Deviations from the learned activity shapes can be used to detect abnormal ones. The advantage of their proposed method is that once an activity is recognized, the deviation from it can be achieved using basis shape. A comparative complex research is done by (Xiang and Gong, 2006). They show a unified bottom-up and top-down method to model complex activities of different objects in occluded scenes. The accuracy of their system is superior in noisy and cluttered situations. Furthermore, it can operate both indoors and outdoors. Liu and Chua (2003) present a Hidden Markov Model (HMM) for perceiving multi-agent activities. Traditional HMMs fails when the number of agents increases. By using the enhanced two-layer version of HMM, the algorithm achieves robust results when agent information is partially represented.

2.4.2. Holistic Recognition approaches

Different methodologies using global body structure and dynamics are concerned with the recognition of basic actions, for example, running and walking. Very nearly all systems are silhouette or contour-based. Consequent systems are for the most part holistic, e.g., the whole silhouette or contour is, no doubt considered without recognizing individual body parts.

In Wang et al. (2003) the silhouette of a human is computed and afterward unwrapped by evenly sampling the contour. Next, the separation between each contour point and its center of gravity is processed. The unwrapped contour is then processed by Principal Components Analysis (PCA). Experiments performed on outdoor video sequences show that the system achieves overwhelming results with low computation costs. BenAbdelkader et al. (2002) utilize a variety of co-occurence procedures. After applying a suitable time-wrapping and normalization with respect to scale a self similarity plot is computed where silhouette pictures of the sequences are pairwise correlated. PCA is applied to decrease the dimensionality of these plots and a k-Nearest Neighbor (k-NN) classifier is applied in Eigen space for recognition. The proposed system is robust to tracking and segmentation errors and also invariant to variation in background and clothing. It performs reliably well when tested for both indoor and outdoor sequences at different viewpoints. Foster et al. (2003) extract, embox, and standardize silhouettes. At that point, a set of binary masks are characterized and the region of the silhouettes inside the mask is computed to give a dynamic signature of the observed individual for each mask. The system is also used for male and female discrimination, however, the accuracy is low. Kale et al. (Kale et al., 2004; Ekinci, 2006) used a HMM to model the dynamics of individual gait. A HMM is trained for every person in the database. During recognition, the HMM with largest probability identifies the individual. Yam et al. (2002) examine the relationship between walking and running. They characterize a gait signature focused on frequency examination of thigh and lower leg rotations. It seems like the signature for walking and running of an individual is related by phase modulation. The additional individual relationship between walking and running is used to derive improved gait-recognition.

While a substantial number of papers recognize people based on their dynamics, the same features can additionally be utilized to recognize what the individual is doing. An initial work in this context has been done by Efros et al. (2003). They endeavor to recognize basic actions of individuals whose pictures in the video are just 30 pixels tall and where the video quality is poor. They utilize a set of features that are focused around blurred optic flow (blurred motion channels). To start with, the individual is tracked so the image is stabilized amidst a tracking window. The blurred motion channels are computed on the residual motion derived from the motion of the body parts. Spatio-temporal cross-correlation is utilized for matching with a database.

Of further interest is the upgrade where complex actions might be alterably made out of the set of straightforward actions. Robertson and Reid (2005) endeavor to comprehend actions by building a hierarchical framework that is focused on reasoning with HMMs and belief networks on a higher level and on a lower level with features, such as, position and velocity as action descriptors. Their action descriptor is focused around the work by (Efros et al., 2003). The system can yield qualitative data, for example, walking left to right on the sidewalk.

A substantial research has focused on space-time volumes. One of the main methods is to utilize spatio-temporal XT-slices from a picture volume XYT (Ricquebourg and Bouthemy, 2000; Rittscher et al., 2002) where explained motions of a human might be connected with a regular trajectory pattern. (Zhen et al., 2014) presented spatio-temporal steerable pyramid (STSP) for holistic representation of human actions. A video sequence is viewed as a spatio-temporal volume preserving all the appearance and motion information. For invariance, spatio-temporal max pooling operation is performed between responses of filtering. Their results show that STSP achieves comparable results with the state of the art.

Bobick and Davis pioneered the introduction of temporal templates (Bobick and Davis, 2001). They propose a representation and recognition theory that is focused around motion energy images (MEI) and MHI. Bradski and Davis (2002) get the thought of MHI and create timed MHI (tMHI) for motion segmentation. A similar approach to (Bobick and Davis, 2001) was performed in Masoud and Papanikolopoulos (2003).Here motion information is presented by a feature image. However,unlike (Bobick and Davis, 2001), an action is represented by several feature images. A full and hierarchical human identification framework is presented by Ozer and Wolf (2002). They approach the tracking, posture estimation and action recognition issue in an integrated way. An alternate methodology is that of "Actions Sketches" or "Space-Time Shapes" in the 3D XYT volume. Yilmaz and Shah (2005) propose to utilize Spatio-Temporal Volumes (STV) for action recognition: the 3D contour of an individual gives rise to a 2D projection. Considering this projection over time defines STV. They approached action recognition as an object matching task by interpreting the STV as rigid 3D objects. Gao et al. (2004) consider a smart room application. A dining room action examination is performed by consolidating motion segmentation with tracking. They utilize motion segmentation focused around optical flow and RANSAC algorithm. At that point, they consolidate the motion segmentation with a tracking method which is sensitive to subtle motion. Keeping in mind the end goal to recognize activities, they recognize overwhelming directions of relative movements.

Several works are concerned with the recognition of activities focused around the dynamics and settings of individual body parts. A few methods, e.g., (Davis and Taylor, 2002), begin with silhouettes and detect the body parts utilizing a system inspired by the W4-framework (Haritaoglu et al., 2000). Others utilize 3D-model based body tracking methods where the recognition of (frequently occasional) activity is utilized as a loop-back to help posture estimation. Different methods bypass the vision issue by utilizing a motion capture framework within request to have the capacity to concentrate on the action issues (Davis and Gao, 2004; Parameswaran and Chellappa, 2006). In a work aligned with (Wang et al., 2003), Wang et al. (2004) presented a method where contours are extracted and a mean contour is computed to represent the static contour information. Dynamic data is extracted by utilizing a detailed model made out of 14 rigid body parts, every one represented by a truncated cone. Particle filtering is utilized to process the probability of a posture given an input image. For classification, a k-NN was used. The experimental results show the feasibility of the proposed system. Ren and Xu (2002) use as input a binary silhouette from which they detect the head, torso, hands, and elbow angles. At that point, a primitive-based coupled HMM is utilized to recognize natural complex and predefined actions. They extend their work in (Ren et al., 2004) by presenting primitive-based Dynamic Bayesian Networks (DBNs).

Gritai et al. (2004) address the invariant recognition of human activities, and research the utilization of anthropometry to give constraints on matching. Their work is focused around a point-light display like representation where a pose is shown through a set of points in 3D space. The proposed algorithm guarantees that both temporal and view invariance is achieved. The versatility of the developed system is demonstrated by a number of challenging datasets and applications. Davis and Gao (2003, 2004) aimed to perceive properties from visual target cues, e.g., the sex of an individual or the weight of a conveyed article is assessed from how the people move. So as to distinguish specific body parts. An importance weight is assigned to each motion trajectory and the model then automatically learns the weight from training data. The model demonstrates high accuracy for both contexts and shows greater flexibility than PCA. Fanti et al. (2005) give the structure of a human as model information. To discover the undoubtedly model arrangement with input information they exploit appearance information which remains invariant inside the same setting. Expectation maximization is utilized for unsupervised learning of the parameters and structure of the model for a specific activity and unlabeled input information. Action is then recognized by greatest probability estimation. Ning et al. (2006) used a parabola to model the shoulders of a human. Fisher discriminant analysis (FDA) on the parabola parameters are utilized to detect shrugs. Their results show that the proposed system is not only able to detect shrugs actions correctly but can also tolerate the large in-class subject variation, illumination, action speed, clutter background and partial occlusion.

(Shao et al., 2014b) recently presented a novel global descriptor for holistic human action recognition called spatio-temporal Laplacian pyramid coding (STLPC). This descriptor treats the video sequence as a whole, which prevents the information loss. 3D Gabor filter is applied to make the features invariant to noise and distortion followed by max pooling. The proposed method is tested on KTH, the multi-view IXMAS, UCF sports, HMDB51 and outperform state of the art results. Motion information is very important for describing the action of an individual. (Liu et al., 2013) used Pyramidal Motion Features (PMF) information extracted by optical flow algorithm in a subsequent frame of a video sequence. Adaboost algorithm is used to select the key frames with most discriminatory information followed by a correlogram for action representation. The system achieves high recognition accuracy for KTH, IXMAS and HMDB51.

The existing methods for realistic action recognition are based on non-probabilistic classification. They gave the predicted labels but do not outline the estimation of uncertainty. (Liu et al., 2014) proposed a probabilistic framework using Gaussian processes (GPs) that can handle the regression problems of uncertainty for action recognition. One of the major challenges of using GPs is the inversion of the large covariance matrix during inference. However, the results of their experiments show high accuracy on challenging datasets.

As the availability of digital video databases is increasing, so is increasing the need of algorithms to effectively navigate through them. (Shao et al., 2014a) developed a content-based video retrieval system based on human actions. Temporal localization based on histogram of evenly spaced time-slice is used. The proposed system is computationally efficient and simpler with comparable works for action retrieval with localization and can be used for real world scenarios.

2.4.3. Action Primitives and Grammar

There is solid neurobiological confirmation that human activities and actions are straightforwardly connected with the ”motor control” of the human body (Rizzolatti et al., 2001; Giese and Poggio, 2003).

Related ongoing research aiming imitation distinguishes a set of motor primitives that model the visually perceived activity, which allows interpreting and recognizing activities in a video scene through a chain of primitives namely, straightforward activities and actions.

Jenkins and Mataric (2002) recommend to apply a spatio-temporal non-linear dimension reduction system on manually segmented human motion capture information. Comparable segments are grouped into primitive units, which are summed up into parametrized primitives by interpolation. In the same way, they characterize action units (”behavior units”), which might be summed up into actions. Billard et al. (2004); Calinon and Billard (2004); Calinon et al. (2005) utilize a HMM-based method to learn trademark peculiarities of repetitively demonstrated movements. They recommend to utilize the HMM to combine joint trajectories of a robot. For each joint, one HMM is utilized. Various distributions endeavor to decouple actions into action primitives and to interpret actions as an arrangement on the alphabets in order of these action primitives, however, without the constraints of having to drive a motor controller with the same representation. Del Vecchio et al. (2003) utilize systems from the dynamical frameworks schema to approach segmentation and classification. System identification techniques are used to derive analytical error analysis and performance estimates. Once the primitives are detected then an iterative approach is used to find the sequence of primitives for a novel action.

Some researchers focus on the action representation by taking a vision-based method, such as performed by Rao et al. (2002) that proposes a view-invariant representation of actions focused around dynamic moments and intervals. Dynamic moments are utilized as primitives of actions, which are figured from discontinuities of 2D hand trajectories. An interval shows the time period between two element moments (key poses). Ivanov and Bobick (2000) recommend utilizing stochastic parsing for a semantic representation of an activity. They discovered that for a few actions, where it comes to semantic or temporal ambiguities or inadequate information, stochastic methods may not be efficient to model complex activities and actions. They propose decoupling actions into primitive segments and utilizing a stochastic parser for recognition. Lv and Nevatia (2006) presented an interesting method to break down the extensive joint space into a set of peculiarity spaces, where every feature corresponds to a solitary joint or combination of related joints. They utilize HMMs to recognize each activity class focused around the features and an Adaboost classifier to detect and recognize the peculiarities.

2.4.4. Discussion

As shown in Table 5, research in the area of action recognition is often linked with real-world applications, such as surveillance (Eng et al., 2003)(Boiman and Irani, 2007) and nursing care (Gao et al., 2004). Most of the publications deal only with simple human actions, such as walking, sitting, running, standing or jumping, e.g., (Chowdhury and Chellappa, 2003)(Efros et al., 2003)(Junejo et al., 2004)(Wang et al., 2004)(Kale et al., 2004). Only a small number of researchers work on complex actions and group interactions (Ivanov and Bobick, 2000; Robertson and Reid, 2005), although future research in this area is expected to highly increase, motivated by complex real-world environments. For instance, performance on the LIRIS dataset, which includes complex human-object and human-human interactions, is far from ideal (as shown in the next section).

3. Datasets Related with Human Behavior Detection

We list eight distinct public datasets that have been used in the literature (within the 2000 to 2014 range of this review) to test human behavior detection methods.

3.1. KTH Action Dataset

The KTH action video dataset (Schuldt et al., 2004) holds six sorts of distinctive human activity classes: boxing, walking, running, jogging, waving, and clapping. Each action class is performed several times by 25 human subjects, as shown in Figure 1. The sequences were recorded in four separate situations: outdoor, outdoor with scale variation, outside with distinctive attire, and indoor. The background is homogeneous and static in most arrangements. In total, the dataset comprises of 2391 video samples.

3.2. Weizmann Action Dataset

The Weizmann action dataset (Blank et al., 2005) comprises ten distinct action classes: bending downwards, running, walking, skipping, jumping-back, jumping forward, jumping in place, galloping sideways, waving with two hands, and waving with one hand. Examples of some of these actions are shown as in Figure 2. Each activity class is performed once (in some cases twice) by 9 subjects resulting in a total of 93 video sequences. Like the KTH dataset, the background is homogeneous and static.

Table 5: Comparison of approaches for Recognition

Reference	Approach	Advantages	Limitations
(Stauffer and Grimson, 2000)	Adaptive background subtraction with mixture of Gaussian	Invariant to lightening changes, clutter, dynamic scenes	Difficulty in tracking visually overlapping objects
(Bobick and Davis, 2001)	Motion energy image (MEI) and Motion history image (MHI)	Invariant to change in speed, automatic temporal segmentation	Fails when one person partially occludes another
(BenAbdelkader et al., 2002)	PCA, k-nearest neighbor (k-NN)	Invariant to dynamic background and clothing	Static camera, only tested for outdoor video, can be slow
(Chowdhury and Chellappa, 2003)	Factorization theorem with polygonal shapes	Deviation from an activity can be calculated	Static camera with low resolution, works only for known activities
(Efros et al., 2003)	Spatio-temporal cross-correlation	Recognition of actions from a distance, reliable results in low resolution and noisy data	Human operator is required
(Eng et al., 2003)	Markov random field network	Handles partial occlusion, can detect abnormal activity	Not robust for crowded scenarios
(Liu and Chua, 2003)	Observation decomposed Hidden Markov models (ODHMMs)	Less sensitive to missing data, efficient classification for three-person activities	Algorithm fails if no role parameters are assigned
(Vaswani et al., 2003)	Polygonal shapes and particle filter	Can detect abnormal activity from noisy and low resolution data	Cannot handle shape deformation, not robust
(Wang et al., 2003)	Spatial-temporal silhouettes analysis, PCA	Efficient results on low computation costs	–
(Gao et al., 2004)	RANSAC algorithm	Estimates motion direction, accurate segmentation, helpful for caregivers at nursing home	Not fully robust to motion
(Junejo et al., 2004)	Hausdorff distance metric, spatial information	Applicable for real-time pedestrian detection	–
(Kale et al., 2004)	Frame to exemplar distance (FED), Hidden Markov model (HMM)	Robust gait recognition rates	Performance drops if velocity information is used
(Wang et al., 2004)	Particle filter, k-NN, Procrustes shape analysis	Feasible results for gait	Not robust to noise, images without occlusion for testing
(Robertson and Reid, 2005)	HMM with belief network	Video annotation, Improved results, low computation cost	only simple actions are used, manual segmentation,
(Yilmaz and Shah, 2005)	Spatio-temporal volume (STV)	Invariant to camera angle	Low results for a few actions
(Xiang and Gong, 2006)	Expectation maximization (EM), Schwarz's Bayesian information criterion (BIC), DML-HMM	Works for noisy and cluttered data	Stationary camera used, require extensive training time, not reliable for complex activities
(Boiman and Irani, 2007)	Inference by composition, Patches of data	Can detect suspicious behavior, irregularities in images	Cannot handle extreme occlusion, computationally expensive in terms of memory and time
(Liu et al., 2013)	Pyramidal Motion Features (PMF), Adaboost	High recognition rates for certain databases	Computationally expensive
(Shao et al., 2014a)	Temporal localization	Computationally efficient and simple, applicable for real-world scenarios	Not appropriate for online databases
(Shao et al., 2014b)	Spatio-temporal Laplacian pyramid coding (STLPC), 3D Gabor filter	Invariant to noise and distortion	Low results with camera zooming
(Zhen et al., 2014)	spatio-temporal steerable pyramid (STSP)	Preserves shape and motion information, efficient results on three datasets	low results for complex actions and backgrounds

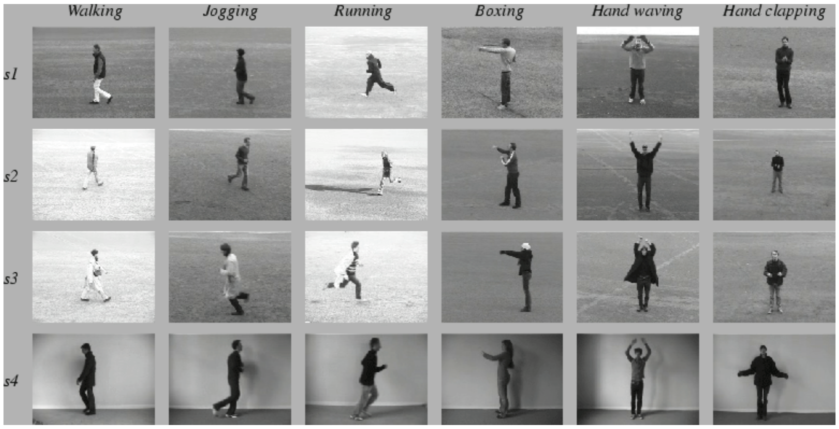

Figure 1: Examples of the KTH action dataset.

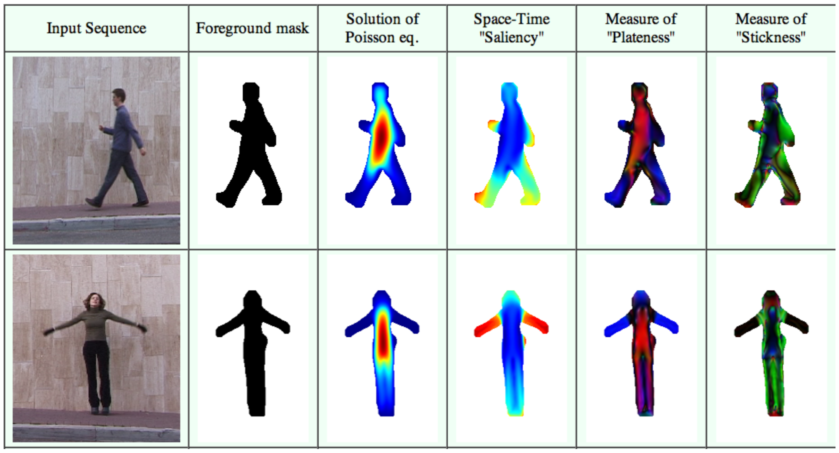

Figure 2: Examples of the Weizmann action dataset, retrieved from (Gorelick et al., 2007).

3.3. Hollywood Action Dataset

The Hollywood action dataset (Laptev et al., 2008) holds eight distinctive action classes: answering the telephone, getting out of the car, hand shaking, hugging, kissing, sitting down, sitting up, and standing up. Figure 3 shows examples of images related with some of these classes. These actions have been gathered semi-automatically from 32 diverse Hollywood films. The full dataset holds 663 video samples. It is isolated into a clean training set (219 sequences) and a clean test set (211 sequences), with training and test sequences acquired from distinct films.

3.4. YouTube Action Dataset

The YouTube action dataset (Liu et al., 2009) is a more challenging dataset when compared with the previous ones. This is because it contains a high degree variation in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc., as shown in Figure 4. This dataset contains eleven action categories: biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis

Figure 3: Examples of the Hollywood action dataset, retrieved from (Laptev et al., 2008).

swinging, trampoline jumping, volleyball spiking, and walking with a dog. For each category, the actions are divided into twenty-five groups, each with more than four action clips in it. The video clips (in MPEG4 format) in the same group may share some common characteristics (e.g., same actor, background or similar viewpoint).

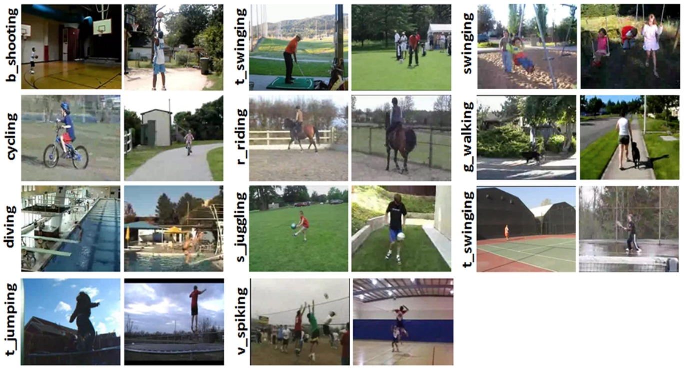

Figure 4: Examples of the YouTube action dataset, retrieved from (Liu et al., 2009).

3.5. UCF Action Dataset

The UCF dataset (Mikel D. Rodriguez, 2008) consists of a set of actions collected from various sports that were broadcasted on television (Figure 5). The video sequences are obtained from a wide range of stock footage websites, including BBC motion gallery and GettyImages. The dataset consists of 200 video sequences with a resolution of 720x480. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints. The dataset actions are classified into: diving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging and walking.

3.6. CMU Mocap Dataset

The CMU motion capture database (Guerra-Filho and Biswas, 2012) was built mainly to provide a source of motion data for animation and other applications. The database contains 2605 different motion clips of full body Mocap data, as shown in Figure 6. The actions have been performed by a total of 144 subjects (some subjects are the same person) and consist of twenty-three action categories. The database has no formal structure such

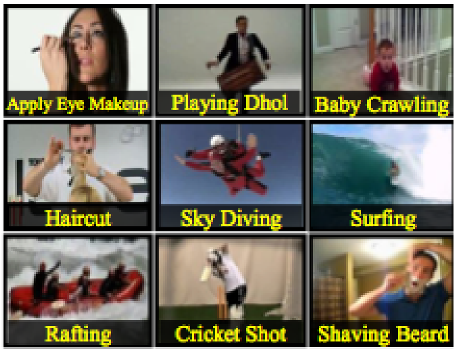

Figure 5: Examples of the UCF Action dataset, retrieved from (Soomro et al., 2012).

that most sessions have different actions. While one set may have walk actions, other set contains both walk and run actions, and another set contains basketball moves. Even the same action performed across different sets may not have been performed in the same way.

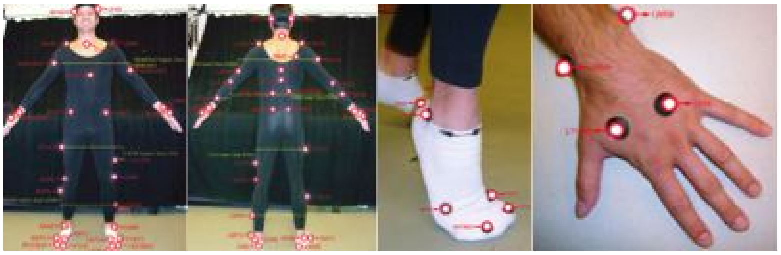

Figure 6: Examples of the CMU Mocap dataset with marker placement, retrieved from (Jang, 2002).

3.7. Caviar/Behave Dataset

Within the CAVIAR project (Brdiczka et al., 2005), a number of video clips were recorded acting out different scenarios of interest. Figure 7 shows examples of two scenarios. These scenarios include people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and leaving a package in a public place. The first section of video clips was filmed for the CAVIAR project with a wide-angle camera lens in the entrance lobby of the INRIA Labs at Grenoble, France. The resolution is half-resolution PAL standard (384 x 288 pixels, 25 frames per second) and compressed using MPEG2. The file sizes are mostly between 6 and 12 MB, a few up to 21 MB.

3.8. LIRIS Human Activities Dataset

The LIRIS human activities dataset (Wolf et al., 2012) contains videos (in gray, RGB color format and with depth data) showing people performing various activities taken from daily life (discussing, telephone calls, giving an item, etc.). Figure 8 shows examples of these activities. The dataset is fully annotated, where the annotation contains information on the action class and also about its spatial and temporal positions in the video. The dataset has been shot with two different cameras: Kinect (subset D1) and Sony consumer camcorder (subset D2).

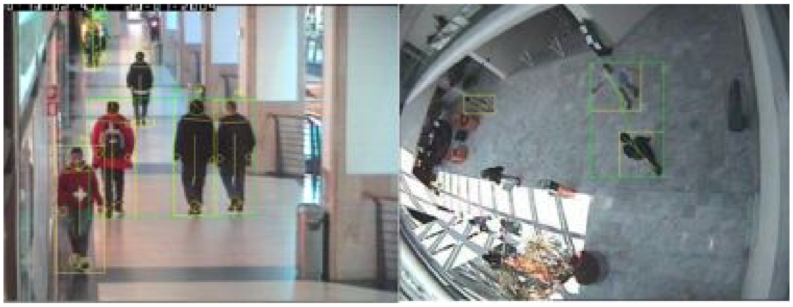

Figure 7: Examples of the Caviar/Behave dataset, retrieved from (Fisher, 2014).

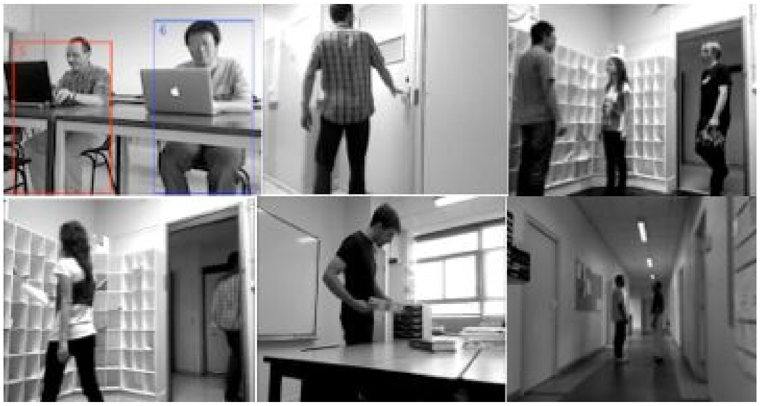

Figure 8: Examples of the LIRIS Human activity dataset, retrieved from (Wolf et al., 2012).

3.9. Discussion

With the growing interest in human behaviour recognition from video, several datasets have been created to benchmark the improvements in this field. This section lists eight of such datasets and with different characteristics. The KTH, Wiezmann and CMU motion datasets capture focus on very simple actions, such as walking, running or jogging, and were captured in a non-cluttered environment. The UCF, Hollywood and YouTube benchmarks assume a collection of broadcast videos, taken from different movies and sports. As such, they are more realistic and complex in terms of human behaviors and scenes that are approached. The Caviar dataset assumes a controlled recording of videos under different scenarios, such as people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and leaving a package in a public place. Finally, the LIRIS dataset was shot using two different cameras and it contains annotated videos of people performing several daily activities (including human–human and human-object interactions).

Table 6 presents the state-of-the art results on the eight analyzed datasets. The last column of the table (**Results**) presents the classification accuracy of the method (except for the LIRIS Dataset, where only the F-Score value is known). In the table, the datasets were sorted under a decreasing order according to the best achieved result. Table 6 shows very high classification performances (above 95%) for three datasets (CMU Mocap, Weizmann, KTH and UCF). Also, high classification accuracies (above 85%) were achieved for the YouTube

Action and Caviar benchmarks. The last two datasets (Hollywood and LIRIS) are much challenging. In effect, the state of the art methods are still far from ideal (56.4% accuracy for Hollywood and F-Score of 53% for LIRIS) and thus there is a large room for further research on methods for handling these datasets.

Table 6: State-of-the-art-results for Human Behavior Detection datasets (best values in **bold**)

Databases	**Reference**	**Approach**	**Results**
CMU Mocap Dataset	(Shotton et al., 2013)	3D joint positions	72%
	(Wang et al., 2012)	Actionlet ensemble model	**98.10**%
Weizmann Action Dataset	(Klaser et al., 2008)	Spatio-temporal Descriptors	84.30%
	(Rahman et al., 2014)	Negative space action descriptors	**95.56**%
KTH Action Dataset	(Klaser et al., 2008)	Spatio-temporal Descriptors	91.40%
	(Wang et al., 2011)	Dense trajectories	94.20%
	(Rahman et al., 2014)	Negative space action descriptors	94.49%
	(Shao et al., 2014b)	Spatio-temporal Laplacian pyramid coding	95.00%
	(Zhen et al., 2014)	Spatio-temporal steerable pyramid	94.5%
	(Liu et al., 2013)	Pyramidal motion feature + Adaboost	**95.5**%
UCF Sport Dataset	(Zhen et al., 2014)	Spatio-temporal steerable pyramid	80.7%
	(Wang et al., 2011)	Dense trajectories	88.2%
	(Shao et al., 2014b)	Spatio-temporal Laplacian pyramid coding	93.4%
	(Liu et al., 2014)	Gaussian processes (GPs)	**92.7**%
Caviar Dataset	(Kuo and Nevatia, 2011)	Appearance-based affinity models	88.1%
	(Kuo et al., 2010)	Online learned discriminative appearance models (OLDAMs)	**89.4**%
YouTube Action Dataset	(Wang et al., 2011)	Dense trajectories	84.2%
	(Liu et al., 2014)	Gaussian processes (GPs)	**85.4**%
Hollywood Action Dataset	(Klaser et al., 2008)	Spatio-temporal Descriptors	24.7%
	(Gilbert et al., 2011)	Mined hierarchical features	**56.4**%
LIRIS Dataset	(Wolf et al., 2012)	Multi-Stage Depth Contextual Information for Action Detection	F-Score=**53**%

4. Applications

In this section, we list examples of relevant applications that were grouped into 6 classes and that are detailed in the next subsections. To synthesize this section, Table 4 shows the evolution of some representative application works. The table includes some of the application key computational techniques (column **Method**), datasets analyzed (column **Database**) and mentioned limitations (column **Limitations**) and advantages (column **Advantages**).

4.1. Human Detection using 3D Depth Images

The detection of humans is a challenging problem due to variation in pose, clothing, lightning conditions and complexity of backgrounds. Several methods are proposed to target this task (Dalal and Triggs, 2005; Schwartz et al., 2009; Dalal et al., 2006; Ikemura and Fujiyoshi, 2011). Most of the research is based on images taken by visible light cameras. Some methods use statistical training based on local features, e.g., gradient-based feature such as HOG, and some involves extracting interest points in image such as scale-invariant feature transform (SIFT). High detection accuracies were reported for these method, however such performance decreases when the background is cluttered or there are complex human poses. Depth information is an important input feature because the object may not have consistent color and texture should occupy an specific region in space. Range images have several advantages over 2D intensity images, such as being more robust to change in illumination and color. Also, they are simple forms to represent 3D information. In the past, these sensors

Table 7: Chronological evolution (from 2000 to 2014) of selected application works on automatic visual detection of human behavior.

Reference	Method	Database	Limitations	Advantages
Zhao and Thorpe (2000)	Feed-forward NN	Live video in urban areas from cameras mounted on the top of a minivan	Fails with objects similar to humans, pedestrian color similar to background and two close persons	Handles occlusion and works in different weather conditions
BenAbdelkader and Davis (2002)	Non-parametric background modeling and frame to frame tracking using blob detection	Outdoor video sequence of people carrying objects	Stationary camera is used	View-invariant and robust to segmentation and tracking errors
Xu and Fujimura (2003)	Split-merge algorithm, depth information	Created their own database	Fails to identify some objects and human torso for certain poses	Insensitive to changing background, 3D sensors are used
Zhou and Hoang (2005)	Codebook, Appearance-based tracking	Created their own database	Two persons walking close classified as one, person entering the room is classified as an object	Invariant to moving background, handles occlusion
Dalal and Triggs (2005)	HOG, SVM	MIT pedestrian dataset, created their own dataset (INRIA) with 1800 persons	Not mentioned	HOG-based method outperforms wavelet, PCA-SIFT, and shape-content ones
Ferrando et al. (2006)	Hu-Moment analysis for blob detection, position and color features for tracking	Experiments performed in a laboratory and public indoor environment	Signals an alarm if an object is partly occluded by other object	Capable of real-time response
Miaou et al. (2006)	Connected component labeling	Created their own database with 20 persons	Cannot handle two persons standing side by side and shadow removal	Omni-cameras are used together with personal information
Dalal et al. (2006)	Appearance and motion descriptors, SVM	Shots from DVDs and personal camera	Appearance and motion descriptors used separately	Insensitive to moving camera and background
Klaser et al. (2008)	HOG 3D descriptor, Bag of Words (BoWs), Harris operator, SVM	KTH, Weizmann and Hollywood	Not mentioned	Competitive in comparison with the state of art
Schwartz et al. (2009)	Partial Least squares (PLS), SVM	INRIA, Daimler-Chrysler and ETHZ pedestrian datasets	Misclassification in some cases	Competitive in comparison with the state of art
Xia et al. (2011)	2D edge and 3D shape detector, and depth information	Created their own database using Kinect	Highly dependent on head detection	Performs well in indoor environment, can detect human in the presence of various objects (e.g., chairs)
Sung et al. (2011)	Maximum-entropy Markov Model (MEMM) and RGBD sensor (Kinect)	Created their own dataset for indoor environments with Kinect	Occluded data was not included	Improved results when the person was seen before and low otherwise
Wang et al. (2011)	Bag of feature, codebook for each descriptor, SVM	KTH, YouTube, UCF sport, Hollywood2	Not mentioned	Insensitive to camera and background motion
Shotton et al. (2013)	Single depth images for 3D position of body parts, Kinect	CMU mocap database	Not mentioned	Competitive in comparison with the state of art
Zhang et al. (2013)	Feed-forward NN	Facial Coding Action Units (FACU)	Cannot handle compound emotion (e.g., surprise and anger)	Efficient for developing intelligent humanoid robots
Diraco et al. (2013)	Topological and Volumetric spatial representation	Four real-home scenarios	Not mentioned	Good event and behavior detection, invariance to viewpoint
Rahman et al. (2014)	Negative-space based feature, k-NN	KTH, Weizmann, fish action dataset	Not mentioned	Handles partial occlusion, small shadows and corrupted image sequences

were expensive and difficult to use in human environment due to lasers. The scenario changed with the spread of cheap and easy to use 3D cameras, such as Microsoft Kinect.

Xu and Fujimura (2003) proposed a new method for detecting objects in a user-specified 3D zone using a single camera, adopting a new type of depth sensor together with a new depth-slicing algorithm for detecting objects in various depths. An area of interest for depth data acquisition is defined and object beyond this range does not appear in the depth image. The camera used was able to take depth information and gray (color) information simultaneously using the same optical axis, thereby achieving 3D detection by a single camera. Split-and-merge algorithm is used on range data for tracking and detecting objects at various distances in the scene. Gray images were used for extracting contextual information as to where the object of interest is, and depth images are used to extract strong evidence in discontinuity between multiple objects detected in the scene. The proposed method does not use background subtraction and hence can be applied to method where camera is non-stationary. The method was tested on various videos (e.g., walking, two people greeting, box picking).

Xia et al. (2011) proposed a new method for human detection using depth information obtained from Kinect in indoor environments. People were detected using a 2-stage head detection process, which includes a 2D edge detector and a 3D shape detector to utilize both the edge information and the relational depth change information in the depth image. From an input depth array, noise is reduced and 2D chamfer distance matching is used to locate people. Each of these regions is examined using a 3D head model, which utilizes the additional depth information for verification. A segmentation method is proposed to segment the human body from the background and the object connected to it. The overall contour of the person was extracted to recognize the activity and to perform tracking. The method is tested using Kinect for Xbox 360 in an indoor environment. Interesting results were reported, but the algorithm has high dependency on accurate head detection, which implies that if the head is occluded or if the person is wearing a strange shape, then it will not be detected.

Sung et al. (2011) performed activity recognition using a Kinect camera. Their work showed reliable performance in detection and recognition activities even in cluttered environment. The proposed method is based on machine learning techniques and useful features were extracted based on the estimated skeleton from Kinect. Five different environments were used for experimentation: office, kitchen, bedroom, bath room, and living room. Experiments showed improved results in detecting activity when the person was seen before and low otherwise. The limitation to the work is that occluded data was not included in validation.

Shotton et al. (2013) proposed a new method for predicting human pose by using 3D positions of body joints. A single depth image is used and information from preceding frames is not considered. The difficult pose estimation problem is transformed into a simpler per-pixel classification problem, for which efficient machine learning techniques exist. Computer graphics is used to synthesize a very large dataset of training images. In order to reduce the demand for training data, body is divided into parts and per-part estimates are computed to produce single pose estimate. A classifier is trained that estimate body part labels, invariant to pose, body shape, clothing and other irrelevances. Finally, re-projecting the classification result and finding local modes obtain confidence-scored 3D proposals of several body joints.

For their early experiments, they employed CMU Mocap database, which is publicly available. This dataset consist of different motions (e.g., interaction with environment, physical activities) and Xbox 360 was used.

4.2. Abnormal Activity Detection

The currently available surveillance cameras, such as Charge-Coupled Device (CCD) cameras, thermal cameras and night vision devices, are cheap and widely available these days, but it requires manpower to constantly monitor the video for suspicious event. The videos from these cameras are sparingly monitored or not at all. Thus, these cameras can be more useful if a computer constantly monitors it, in order to trigger an event when something suspicious happens in real time, in what is known as automated surveillance. Video surveillance is often based on the following stages: modeling of environment, detection of motion, and classification of moving objects, behavior understanding and description, and fusion of information from multiple cameras.

With the help of using multiple cameras, the problem of occlusion can be solved but it further introduces problems like installing multiple cameras, object matching, object calibration, automatic camera switching and data fusion. Ferrando et al. (2006) proposed a method for detecting abandoned or missing object from a scene. This is an important area and has major applications especially in crowded environments (e.g., airports, stations, sorting events and public areas). The proposed video-based surveillance aims at supporting a human operator by signaling an alarm when a critical situation occurs, such as detection of an abandoned object. Low-level module is used for updating the background if any change has occurred. Background updating is performed by information from high-level module. Blobs classified as abandoned object are excluded from updating to preserve their position in the following scenes while blobs as stolen objects are absorbed immediately in the background. In some countries, where the number of elderly people increases, the nursing demand also increases. A particular interest is the automatic detection of accidents. Indeed, Miaou et al. (2006) highlighted that around 70% of accident falls are preventable and proposed a MapCam (omni-camera) together with the personal information of an individual (e.g., age, sex, weight) to detect elderly fall. The omni camera was connected to a PC server that observed the images and sent suspicious events to PCs or smart phones. The propose system used a clean background image, when there are no object or moving people. For the objects of interest in the foreground, background subtraction is done. Morphological operators are used for the removal of noise. The system then uses connected component labeling to obtain the area, height and width of each object. A simple decision fall threshold was set to determine if a person was falling.

4.3. Action Recognition from Video

Body motion is highly relevant for action recognition but video motion can also occur due to background or camera motion. To overcome the problem of camera motion, Wang et al. (2011) introduced a local descriptor that focuses on foreground motion. Their work show that motion boundaries encoded along the trajectories significantly out perform the state of the art descriptors.

For action recognition, local descriptors based on normalized pixel values, brightness gradients and windowed optical flow were evaluated by (Dollár et al., 2005). Experiments on three datasets (e.g., KTH) have shown best results for gradient descriptors. These descriptors were computed by concatenating all gradient vectors in a region or by building histogram on gradient component. Such descriptors suffer from sensitivity to illumination changes. In (Klaser et al., 2008), descriptors on gradient orientation are used because they are robust to changes in illumination. These descriptors were based purely on spatio-temporal 3D gradients that are robust and cheap to compute. Orientation quantization is performed up to 20 bins by using polyhedrons. They also proposed integral histograms for memory-efficient computation of features at arbitrary and temporal scales. The video sequence is represented as Bag of Words (BoWs) using sparse space-time features. A sparse set of spatio-temporal interest points is obtained by a space-time extension of the Harris operator. Features are sampled at multiple spatial and temporal scales. The BoWs require creating a visual vocabulary. For representation, all features of the video sequence were assigned to the closest vocabulary work (using Euclidean distance), producing histograms of visual word occurrences that were used for classification by SVM method. They compared their work with the state of the art on three action datasets (KTH, Weizmann and Hollywood), obtaining competitive results.

Local features are popular way for representing videos. They achieve state of the art results for action classification when combined with bag-of-features representation. Wang et al. (2011) proposed an efficient way to extract dense trajectories. The trajectories are obtained by tracking densely sampled points using optical flow fields. The number of tracked points can be sealed easily, as dense flow field are already computed. Global smoothness constraints are imposed among the points in dense optical flow fields, which result in more robust trajectories than tracking or matching points separately.

Problems occur in action recognition due to motion of camera, noisy background, cluttered background, changing of viewpoint, and geometric and photometric variances of objects. Rahman et al. (2014) proposed an implicit method for recognizing action from videos that used the surrounding area to detect an object (called negative space). Features were extracted from this negative space and that work well in the presence of twist actions, such as partial occlusions, complex boundary variation and small shadows. Another system was additionally proposed to recognize cycles of diverse actions automatically. First, an input image was segmented from the background and shadows were removed from the segmented image. Then, motion based features were figured out from the sequence. Next, negative space based descriptors of each pose were collected to form an action descriptor. These actions were recognized by applying k-NN classifier and the whole system was tested on two datasets.

4.4. Player Modeling and Robotics

In the recent years, player modeling is also receiving much attention from the gaming community. Developing accurate models of player behavior can be helpful in many scenarios. These systems can help to track the player behavior along time, informing the system each time when a player change his behavior. In this way, artificial intelligence methods can better adapt itself to the changing behavior of the player. Different machine learning techniques have been used over the time to build models of human behavior. Vallim et al. (2013) proposed a method that uses incremental learning technique to track a player's behavior during his interaction with the game. A change detection technique from the area of stream mining was applied based on an incremental clustering and novelty detection method. Different simulations were performed on the data produced by Unreal tournament game, showing that the proposed method can detect changes in playing behavior.

Being able to identify what people are doing can be helpful in making personal assistant robots. These robots help in tracking daily life routines (e.g., cooking, cleaning house). Human activities are composed of sub-activities. For instance, to drink water, one needs to take a glass, pour water in it, tilt the glass and put it back. Giving robots the ability to interpret human activity is an important step towards enabling human to interact with robots in a natural way, as well as enabling a robot to be a more useful personal assistant. Work done previously on activity classification was mainly done using 2D video, which led to a low accuracy.

In order to make successful human robot interaction, it is very important that the robot understands the human expressions and responds in a similar. Zhang et al. (2013) proposed an intelligent system for humanoid robots based on an NN, for facial recognition, and Latent Semantic Analysis, for identifying topics embedded in user conversations. Their work incorporated the Facial Action Coding system for describing physical cues and anatomical knowledge of facial behavior for detecting the six basic emotions (neutral, anger, sad, disgust, happiness, fear and surprise).

4.5. Pedestrian Detection and In-Home Scenarios

Detection of humans in films and videos is a challenging problem due to the motion of subjects, the camera and the background and to variation in pose, illumination, appearance, clothing, and background clutter. Dalal et al. (2006) proposed a detector that could be used to analyze film, TV content or to detect pedestrians from a moving car. Their main aim was to develop a system in which the background and the camera move as much as the people move. The study focused on detection of people who are upright and fully visible against a background that may be stationary or moving. A linear SVM was used as a baseline classifier. The person features combined appearance descriptors extracted from a single frame of video with motion descriptors extracted either from optical flow or spatio-temporal derivatives. The classifier was trained to make person/non-person decisions using a set of manually labeled training windows.

(Ess et al., 2008) proposed an integrated system for tracking multiple people from a mobile platform. The system integrates continuous visual odometry computation with tracking-by-detection for tracking pedestrians in presence of occlusions and noise. The concept of cognitive feedback is used to derive information from one module to another. An automatic failure detection and correction mechanism has been incorporated to the system for the reduction for noise and robust performance.

(Benenson et al., 2012) developed a new pedestrian detector that is good in speed and quality in comparison with the state of the art. The novelty of the proposed work is FPDW; a detector proposed by (Dollár et al., 2010) but without resizing the original image. The idea of the system is similar to Viola and Jones believing in “scaling the features not the image”. The system is able to achieve high quality pedestrian detection at 135fps without any loss in image quality. The detector achieves competitive results on INRIA dataset and Bahnhof sequence.

Diraco et al. (2013) proposed a method for recognizing human posture in several in-home scenarios. A time-of-flight sensor was used in privacy to acquire 3D point cloud sequences while preserving modality and near real-time processed with lower power embedded PC. To cope with different application requirements in terms of discrimination capabilities, covered distance range and processing speed, a twofold discrimination methodology was used exploiting both volumetric and topological spatial representation. The approach was tested on four different real-home scenarios related with ambient assisted living and home care fields, namely dangerous event detection, anomalous behavior detection, activity recognition, natural human-ambient interaction and also invariance to viewpoint changes. The high quality results were achieved, with high classification rates around 97%.

4.6. Person Tracking and Identification

After motion detection, object tracking is performed from one frame to another in the image sequence. The tracking algorithms generally have considerable intersection with motion detection during processing. Tracking over time usually involves observing features such as points, blobs or lines in consecutive frames. Behavior understanding is the process of recognizing and analysis of motion patterns and producing high-level description of action and interaction among moving objects (Ko, 2008). To track individuals, an human model must be made. Human model incorporates human characteristics, such as color, aspect ratio, edge and velocity, to take care of the issue of occlusion. Kalman filter based strategies and manifestation based systems were proposed in the early works (Senior et al., 2006).

Zhou and Hoang (2005) developed a robust system for the detection and tracking of human beings. Their system is able to work in tough situations, such as sudden light change, heavy shadow and tree shaking. Codebook is used to model the shape of human and color histogram is used to model human features. The developed system was based on YUV color space to save the CPU usage. Background subtraction was adopted for foreground detection and then shadow detection was applied. Morphological operators were used to filter out the camera noise and irregular motions. Blobs were segmented from the foreground mask image and blob merge was used to form the whole object.

A common problem in tracking is drifting. To avoid this problem Wang et al. (2011) used some threshold on the number of frames used. Four action datasets were used in their work (KTH, YouTube, Hollywood2 and UCF). A codebook was constructed for each descriptor separately, namely trajectory, HOG, Histogram of Flow (HOF) and Motion Boundary Histogram (MBH). The number of visual words per descriptor was fixed to 4000. For limiting the complexity, a k-means cluster of a subset with 100000 randomly selected training features was adopted. For classification, a SVM with Gaussian kernel was used.

5. Conclusions and Research Implications

In this survey paper, we review the topic of automatic human behavior detection from video. A systematic search was performed using six major scientific databases and the search query “Human Behavior” or “Human Detection” and “Video” or “Data”, leading a selection of 193 journal and conference papers from 2000 to 2014. These papers were grouped into three main sections: computational detection techniques, public datasets and applications. Despite such large amount of analyzed papers, this review might have some limitations and does not claim to be exhaustive, since there might be non analyzed related articles that were not identified by the adopted query, either were available in other scientific publishing databases or were written not in English language.

As shown in the previous sections, the field of automatic human behavior detection from video has received an increasing attention and development in the last few years. Such increase is due to two main reasons. Firstly, capturing sensors (including video cameras) are becoming more affordable (including sophisticated 3D cameras), additionally allowing a more automatic and passive monitoring of human behaviors, where individuals do not need to use special hardware (e.g., magnetic trackers) and may not even be aware that they are being observed. Secondly, automatic human behavior detection can have a high impact in a wide range of human activities, such as surveillance, gaming, detection of pedestrians and preventing accidents (as shown in Section 4).

Considering the presented review on automatic human behavior detection from video, we highlight that:

- The topic is very challenging, with an ambitious goal to detect high-level features (e.g., erratic movement of an elderly person that might induce an accident) from very low level raw data (e.g., pixel) in real-time, thus requiring expertise from different disciplines (e.g., artificial intelligence and computer vision).

- Most systems adopt several assumptions and limitations (as shown in Table 4), in conjunction with straightforward computational routines, to make the topic tractable. This could be seen as a sign that this field is still in an initial stage of development.

- SVM and NN are popular machine learning methods for prediction tasks (as shown in Table 1).

- Research works adapt quickly to video technology advances (as shown in Table 1 with use of Stereo cameras in the 2000s and Kinect cameras in the 2010s).

- While a rapid progress is taking place, with recent frameworks using more advanced training methods, we are still far from a universal solution for the automatic capturing and characterization of human behavior.

- In contrast with efforts already made in other computer vision fields (e.g., face recognition), a robust evaluation needs to be developed. Based on available datasets most frameworks are tested on less than 1000 frames and several works even create their own (non disclosed) databases. Table 4 also attests wide dispersion on the datasets used.

- In the future, we expect the development of more fruitful frameworks, less dependent on non realistic assumptions, capable of fusing different image descriptors (e.g., motion and silhouettes) and with more emphasis on depth (3D) information. Also, an imminent focus should be put on robust automatic tracking and recognition results. Moreover, learnt models of human motion and pose are currently deployed only for specific movements, thus models need to be developed to cope with complex motions. Research should also focus on achieving reliable results for varying body shapes, clothing and clutter in natural scenes. While substantial work has been done on the reconstruction of human motion, the area is still immature. Major progresses are required for behavior representation in dynamic scenes, high level reasoning for interpretation and action. Furthermore, with advances in artificial intelligence, more intelligent systems will be deployed, capable of self adaptation to changes in the environment and to recover from failure. From Industrial point of view certain advances are required: human motion capture for different fields needs to be view-invariant; surveillance applications should be able to perform reliable tracking and detection with ordinary devices. As technology and computational techniques advance and prices decrease, we expect to see more commercial enthusiasm on this topic, widening even more the range of its practical applications.

Acknowledgments

This work is funded by the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia) under research grant SFRH/BD/84939/2012.

References

Agarwal, A. and Triggs, B. (2004). 3d human pose from silhouettes by relevance vector regression. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–882. IEEE.

Allen, B., Curless, B., and Popović, Z. (2003). The space of human body shapes: reconstruction and parameterization from range scans. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 587–594. ACM.

Antonini, G., Martinez, S. V., Bierlaire, M., and Thiran, J. P. (2006). Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, 69(2):159–180.

Atsushi, N., Hirokazu, K., Shinsaku, H., and Seiji, I. (2002). Tracking multiple people using distributed vision systems. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, volume 3, pages 2974–2981. IEEE.

Barrón, C. and Kakadiaris, I. A. (2001). Estimating anthropometry and pose from a single uncalibrated image. *Computer Vision and Image Understanding*, 81(3):269–284.

Barron, C. and Kakadiaris, I. A. (2003). On the improvement of anthropometry and pose estimation from a single uncalibrated image. *Machine Vision and Applications*, 14(4):229–236.

Belongie, S., Malik, J., and Puzicha, J. (2001). Matching shapes. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 454–461. IEEE.

BenAbdelkader, C., Cutler, R., and Davis, L. (2002). Motion-based recognition of people in eigengait space. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 267–272. IEEE.

BenAbdelkader, C. and Davis, L. (2002). Detection of people carrying objects: a motion-based recognition approach. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 378–383. IEEE.

Benenson, R., Mathias, M., Timofte, R., and Van Gool, L. (2012). Pedestrian detection at 100 frames per second. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2903–2910. IEEE.

Billard, A., Epars, Y., Calinon, S., Schaal, S., and Cheng, G. (2004). Discovering optimal imitation strategies. *Robotics and autonomous systems*, 47(2):69–77.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as space-time shapes. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1395–1402. IEEE.

Bobick, A. F. and Davis, J. W. (2001). The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267.

Boiman, O. and Irani, M. (2007). Detecting irregularities in images and in video. *International Journal of Computer Vision*, 74(1):17–31.

Bradski, G. R. and Davis, J. W. (2002). Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184.

Brand, M. (1999). Shadow puppetry. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1237–1244. IEEE.

Brdiczka, O., Maisonnasse, J., and Reignier, P. (2005). Automatic detection of interaction groups. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 32–36. ACM.

Bregler, C., Malik, J., and Pullen, K. (2004). Twist based acquisition and tracking of animal and human kinematics. *International Journal of Computer Vision*, 56(3):179–194.

Brostow, G. J., Essa, I., Steedly, D., and Kwatra, V. (2004). *Novel skeletal representation for articulated creatures*. Springer.

Calinon, S. and Billard, A. (2004). Stochastic gesture production and recognition model for a humanoid robot. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 3, pages 2769–2774. IEEE.

Calinon, S., Guenter, F., and Billard, A. (2005). Goal-directed imitation in a humanoid robot. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 299–304. IEEE.

Capellades, M. B., Doermann, D., DeMenthon, D., and Chellappa, R. (2003). An appearance based approach for human and object tracking. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–85. IEEE.

Carranza, J., Theobalt, C., Magnor, M. A., and Seidel, H.-P. (2003). Free-viewpoint video of human actors. In *ACM Transactions on Graphics (TOG)*, volume 22, pages 569–577. ACM.

Chen, M., Ma, G., and Kee, S. (2005). Pixels classification for moving object extraction. In *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*, volume 2, pages 44–49. IEEE.

Cheung, K., Baker, S., and Kanade, T. (2003). Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–77. IEEE.

Chowdhury, A. K. R. and Chellappa, R. (2003). A factorization approach for activity recognition. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 4, pages 41–41. IEEE.

Chu, C.-W., Jenkins, O. C., and Mataric, M. J. (2003). Markerless kinematic model and motion capture from volume sequences. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–475. IEEE.

Collins, L., Kanade, F., Duggins, T., and Tolliver, E. (2000). Hasegawa. a system for video surveillance and monitoring: Vsam final report. Technical report, Technical Report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University.

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(5):564–577.

Cucchiara, R., Grana, C., Piccardi, M., and Prati, A. (2003). Detecting moving objects, ghosts, and shadows in video streams. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(10):1337–1342.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Computer Vision-ECCV 2006*, pages 428–441. Springer.

Davis, J. W. and Gao, H. (2003). Recognizing human action efforts: An adaptive three-mode pca framework. In *ICCV*, pages 1463–1469.

Davis, J. W. and Gao, H. (2004). Gender recognition from walking movements using adaptive three-mode pca. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 9–9. IEEE.

Davis, J. W. and Taylor, S. R. (2002). Analysis and recognition of walking movements. In *Pattern Recognition, International Conference on*, volume 1, pages 10315–10315. IEEE Computer Society.

Davis, L., Philomin, V., and Duraiswami, R. (2000). Tracking humans from a moving platform. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 4, pages 171–178. IEEE.

Del Vecchio, D., Murray, R. M., and Perona, P. (2003). Decomposition of human motion into dynamics-based primitives with application to drawing tasks. *Automatica*, 39(12):2085–2098.

Diraco, G., Leone, A., and Siciliano, P. (2013). Human posture recognition with a time-of-flight 3d sensor for in-home applications. *Expert Systems with Applications*, 40(2):744–751.

Dollár, P., Belongie, S., and Perona, P. (2010). The fastest pedestrian detector in the west. In *BMVC*, volume 2, page 7. Citeseer.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE.

Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 726–733. IEEE.

Ekinci, M. (2006). Human identification using gait. *Turk J Elec Engin*, 14(2):267–291.

Elgammal, A., Harwood, D., and Davis, L. (2000). Non-parametric model for background subtraction. In *Computer Vision-ECCV 2000*, pages 751–767. Springer.

Eng, H., Toh, K.-A., Kam, A. H., Wang, J., and Yau, W.-Y. (2003). An automatic drowning detection surveillance system for challenging outdoor pool environments. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 532–539. IEEE.

Ess, A., Leibe, B., Schindler, K., and Van Gool, L. (2008). A mobile vision system for robust multi-person tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Fanti, C., Zelnik-Manor, L., and Perona, P. (2005). Hybrid models for human motion recognition. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 1166–1173. IEEE.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2000). Efficient matching of pictorial structures. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 66–73. IEEE.

Ferrando, S., Gera, G., and Regazzoni, C. (2006). Classification of unattended and stolen objects in video-surveillance system. In *Video and Signal Based Surveillance, 2006. AVSS'06. IEEE International Conference on*, pages 21–21. IEEE.

Fihl, P., Corlin, R., Park, S., Moeslund, T. B., and Trivedi, M. M. (2006). Tracking of individuals in very long video sequences. In *Advances in Visual Computing*, pages 60–69. Springer.

Fisher, R. (2014). CAVIAR: Context Aware Vision using Image-based Active Recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

Foster, J. P., Nixon, M. S., and Prügel-Bennett, A. (2003). Automatic gait recognition using area-based metrics. *Pattern Recognition Letters*, 24(14):2489–2497.

Gao, J., Hauptmann, A. G., and Wactlar, H. D. (2004). Combining motion segmentation with tracking for activity analysis. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 699–704. IEEE.

Giese, M. A. and Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, 4(3):179–192.

Gilbert, A., Illingworth, J., and Bowden, R. (2011). Action recognition using mined hierarchical compound features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):883–897.

Gonzalez, J. J., Lim, I. S., Fua, P., and Thalmann, D. (2003). Robust tracking and segmentation of human motion in an image sequence. In *International Conference on Acoustics, Speech, and Signal Processing, Hong Kong*, volume 1, pages 29–32.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253.

Gritai, A., Sheikh, Y., and Shah, M. (2004). On the use of anthropometry in the invariant analysis of human actions. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 923–926. IEEE.

Grochow, K., Martin, S. L., Hertzmann, A., and Popović, Z. (2004). Style-based inverse kinematics. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 522–531. ACM.

Guerra-Filho, G. and Biswas, A. (2012). The human motion database: A cognitive and parametric sampling of human motion. *Image and Vision Computing*, 30(3):251–261.

Guha, P., Mukerjee, A., and Venkatesh, K. (2005). Efficient occlusion handling for multiple agent tracking by reasoning with surveillance event primitives. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 49–56. IEEE.

Gutchess, D., Trajkovics, M., Cohen-Solal, E., Lyons, D., and Jain, A. K. (2001). A background model initialization algorithm for video surveillance. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 733–740. IEEE.

Haritaoglu, I., Beymer, D., and Flickner, M. (2002). Ghost 3d: detecting body posture and parts using stereo. In *Motion and Video Computing, 2002. Proceedings. Workshop on*, pages 175–180. IEEE.

Haritaoglu, I., Harwood, D., and Davis, L. S. (2000). W 4: Real-time surveillance of people and their activities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):809–830.

Hayashi, K., Hashimoto, M., Sumi, K., and Sasakawa, K. (2004). Multiple-person tracker with a fixed slanting stereo camera. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 681–686. IEEE.

Heikkila, M. and Pietikainen, M. (2006). A texture-based method for modeling the background and detecting moving objects. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):657–662.

Heikkilä, M., Pietikäinen, M., and Heikkilä, J. (2004). A texture-based method for detecting moving objects. In *BMVC*, pages 1–10.

Howe, N. R. (2004). Silhouette lookup for automatic pose tracking. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 15–22. IEEE.

Hu, M., Hu, W., and Tan, T. (2004). Tracking people through occlusions. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 724–727. IEEE.

Hua, G., Yang, M.-H., and Wu, Y. (2005). Learning to estimate human pose with data driven belief propagation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 747–754. IEEE.

Ikemura, S. and Fujiyoshi, H. (2011). Real-time human detection using relational depth similarity features. In *Computer Vision-ACCV 2010*, pages 25–38. Springer.

Interactive, C. (2005). Megamocap: Specialist in motion capture and 3D character animation. http://www.charactermotion.com/about/index.html.

Ioffe, S. and Forsyth, D. (2001). Human tracking with mixtures of trees. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 690–695. IEEE.

Ivanov, Y., Bobick, A., and Liu, J. (2000). Fast lighting independent background subtraction. *International Journal of Computer Vision*, 37(2):199–207.

Ivanov, Y. A. and Bobick, A. F. (2000). Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):852–872.

Iwase, S. and Saito, H. (2004). Parallel tracking of all soccer players by integrating detected positions in multiple view images. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 751–754. IEEE.

Jang, B. (2002). CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/faqs.php.

Jenkins, O. C. and Mataric, M. J. (2002). Deriving action and behavior primitives from human motion data. In *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, volume 3, pages 2551–2556. IEEE.

Junejo, I. N., Javed, O., and Shah, M. (2004). Multi feature path modeling for video surveillance. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 716–719. IEEE.

Kale, A., Sundaresan, A., Rajagopalan, A., Cuntoor, N. P., Roy-Chowdhury, A. K., Kruger, V., and Chellappa, R. (2004). Identification of humans using gait. *Image Processing, IEEE Transactions on*, 13(9):1163–1173.

Kang, J., Cohen, I., and Medioni, G. (2005). Persistent objects tracking across multiple non overlapping cameras. In *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*, volume 2, pages 112–119. IEEE.

Khan, S. and Shah, M. (2000). Tracking people in presence of occlusion. In *Asian Conference on Computer Vision*, volume 5. Citeseer.

Kim, K., Chalidabhongse, T. H., Harwood, D., and Davis, L. (2005). Real-time foreground–background segmentation using codebook model. *Real-time imaging*, 11(3):172–185.

Klaser, A., Marszalek, M., and Cordelia, S. (2008). A spatio-temporal descriptor based on 3d-gradients. In *19th British Machine Vision Conference (BMVC)*, pages 275–285, Leeds, UK.

Ko, T. (2008). A survey on behavior analysis in video surveillance for homeland security applications. In *Applied Imagery Pattern Recognition Workshop, 2008. AIPR'08. 37th IEEE*, pages 1–8. IEEE.

Koschan, A., Kang, S., Paik, J., Abidi, B., and Abidi, M. (2003). Color active shape models for tracking non-rigid objects. *Pattern Recognition Letters*, 24(11):1751–1765.

Krahnstoever, N. and Sharma, R. (2004). Articulated models from video. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–894. IEEE.

Krahnstöver, N., Yeasin, M., and Sharma, R. (2003). Automatic acquisition and initialization of articulated models. *Machine Vision and Applications*, 14(4):218–228.

Krüger, V., Anderson, J., and Prehn, T. (2005). Probabilistic model-based background subtraction. In *Image Analysis*, pages 567–576. Springer.

Kuo, C.-H., Huang, C., and Nevatia, R. (2010). Multi-target tracking by on-line learned discriminative appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 685–692. IEEE.

Kuo, C.-H. and Nevatia, R. (2011). How does person identity recognition help multi-person tracking? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1217–1224. IEEE.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Leibe, B., Seemann, E., and Schiele, B. (2005). Pedestrian detection in crowded scenes. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 878–885. IEEE.

Lim, H., Camps, O. I., Sznaier, M., and Morariu, V. I. (2006). Dynamic appearance modeling for human tracking. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 751–757. IEEE.

Lim, S.-N., Mittal, A., Davis, L. S., and Paragios, N. (2005). Fast illumination-invariant background subtraction using two views: Error analysis, sensor placement and applications. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 1071–1078. IEEE.

Liu, J., Luo, J., and Shah, M. (2009). Recognizing realistic actions from videos in the wild. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1996–2003. IEEE.

Liu, L., Shao, L., and Rockett, P. (2013). Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recognition*, 46(7):1810–1818.

Liu, L., Shao, L., Zheng, F., and Li, X. (2014). Realistic action recognition via sparsely-constructed gaussian processes. *Pattern Recognition*, 47(12):3819–3827.

Liu, X. and Chua, C.-S. (2003). Multi-agent activity recognition using observation decomposed hidden markov model. In *Computer Vision Systems*, pages 247–256. Springer.

Lv, F. and Nevatia, R. (2006). Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *Computer Vision–ECCV 2006*, pages 359–372. Springer.

Magnenat-Thalmann, N. and Seo, H. (2004). Data-driven approaches to digital human modeling. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 380–387. IEEE.

Masoud, O. and Papanikolopoulos, N. (2003). A method for human action recognition. *Image and Vision Computing*, 21(8):729–743.

McKenna, S. J., Jabri, S., Duric, Z., and Wechsler, H. (2000). Tracking interacting people. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 348–353. IEEE.

Menier, C., Boyer, E., Raffin, B., et al. (2006). 3d skeleton-based body pose recovery. In *3rd International Symposium on 3D Data Processing, Visualization and Transmission (DPVT'06)*, pages 389–396.

Miaou, S.-G., Sung, P.-H., and Huang, C.-Y. (2006). A customized human fall detection system using omni-camera images and personal information. In *Distributed Diagnosis and Home Healthcare, 2006. D2H2. 1st Transdisciplinary Conference on*, pages 39–42. IEEE.

Micilotta, A. S., Ong, E.-J., and Bowden, R. (2005). Detection and tracking of humans by probabilistic body part assembly. In *BMVC*.

Mikel D. Rodriguez, Javed Ahmed, M. S. (2008). Action mach: A spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, Anchorage, AK.

Mikić, I., Trivedi, M., Hunter, E., and Cosman, P. (2002). Human body model acquisition and motion capture using voxel data. In *Articulated Motion and Deformable Objects*, pages 104–118. Springer.

Mikić, I., Trivedi, M., Hunter, E., and Cosman, P. (2003). Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision*, 53(3):199–223.

Mittal, A. and Davis, L. S. (2002). M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. In *Computer Vision-ECCV 2002*, pages 18–33. Springer.

Mittal, A. and Davis, L. S. (2003). M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. *International Journal of Computer Vision*, 51(3):189–203.

Moeslund, T. B. and Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2):90–126.

Mohan, A., Papageorgiou, C., and Poggio, T. (2001). Example-based object detection in images by components. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(4):349–361.

Monnet, A., Mittal, A., Paragios, N., and Ramesh, V. (2003). Background modeling and subtraction of dynamic scenes. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1305–1312. IEEE.

Ning, H., Han, T. X., Hu, Y., Zhang, Z., Fu, Y., and Huang, T. S. (2006). A realtime shrug detector. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 505–510. IEEE.

Okuma, K., Taleghani, A., De Freitas, N., Little, J. J., and Lowe, D. G. (2004). A boosted particle filter: Multitarget detection and tracking. In *Computer Vision-ECCV 2004*, pages 28–39. Springer.

Oliver, N. M., Rosario, B., and Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):831–843.

Ong, E.-J. and Hilton, A. (2006). Learnt inverse kinematics for animation synthesis. *Graphical Models*, 68(5):472–483.

Ozer, I. B. and Wolf, W. H. (2002). A hierarchical human detection system in (un) compressed domains. *Multimedia, IEEE Transactions on*, 4(2):283–300.

Parameswaran, V. and Chellappa, R. (2004). View independent human body pose estimation from a single perspective image. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–16. IEEE.

Parameswaran, V. and Chellappa, R. (2006). View invariance for human action recognition. *International Journal of Computer Vision*, 66(1):83–101.

Park, S. and Aggarwal, J. K. (2006). Simultaneous tracking of multiple body parts of interacting persons. *Computer Vision and Image Understanding*, 102(1):1–21.

Plankers, R. and Fua, P. (2003). Articulated soft objects for multi-view shape and motion capture. *IEEE PAMI*, 25(10).

Polat, E., Yeasin, M., and Sharma, R. (2003). Robust tracking of human body parts for collaborative human computer interaction. *Computer Vision and Image Understanding*, 89(1):44–69.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990.

Prati, A., Mikic, I., Trivedi, M. M., and Cucchiara, R. (2003). Detecting moving shadows: algorithms and evaluation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(7):918–923.

Rahman, S. A., Song, I., Leung, M., Lee, I., and Lee, K. (2014). Fast action recognition using negative space features. *Expert Systems with Applications*, 41(2):574–587.

Ramanan, D., Forsyth, D. A., and Zisserman, A. (2005). Strike a pose: Tracking people by finding stylized poses. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 271–278. IEEE.

Rao, C., Yilmaz, A., and Shah, M. (2002). View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226.

Ren, H. and Xu, G. (2002). Human action recognition with primitive-based coupled-hmm. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 494–498. IEEE.

Ren, H., Xu, G., and Kee, S. (2004). Subject-independent natural action recognition. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 523–528. IEEE.

Ricquebourg, Y. and Bouthemy, P. (2000). Real-time tracking of moving persons by exploiting spatio-temporal image slices. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):797–808.

Rittscher, J., Blake, A., and Roberts, S. J. (2002). Towards the automatic analysis of complex human body motions. *Image and Vision Computing*, 20(12):905–916.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2(9):661–670.

Roberts, T. J., McKenna, S. J., and Ricketts, I. W. (2002). Adaptive learning of statistical appearance models for 3d human tracking. In *BMVC*, pages 1–10.

Roberts, T. J., McKenna, S. J., and Ricketts, I. W. (2004). Human pose estimation using learnt probabilistic region similarities and partial configurations. In *Computer Vision-ECCV 2004*, pages 291–303. Springer.

Robertson, N. and Reid, I. (2005). Behaviour understanding in video: a combined method. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 808–815. IEEE.

Ronfard, R., Schmid, C., and Triggs, B. (2002). Learning to parse pictures of people. In *Computer Vision-ECCV 2002*, pages 700–714. Springer.

Roth, D., Doubek, P., and Van Gool, L. J. (2005). Bayesian pixel classification for human tracking. In *WACV/MOTION*, pages 78–83.

Sangi, P., Heikkila, J., and Silvén, O. (2001). Extracting motion components from image sequences using particle filters. In *PROCEEDINGS OF THE SCANDINAVIAN CONFERENCE ON IMAGE ANALYSIS*, pages 508–514.

Schindler, K. and Wang, H. (2006). Smooth foreground-background segmentation for video processing. In *Computer Vision-ACCV 2006*, pages 581–590. Springer.

Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE.

Schwartz, W. R., Kembhavi, A., Harwood, D., and Davis, L. S. (2009). Human detection using partial least squares analysis. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 24–31. IEEE.

Senior, A., Hampapur, A., Tian, Y.-L., Brown, L., Pankanti, S., and Bolle, R. (2006). Appearance models for occlusion handling. *Image and Vision Computing*, 24(11):1233–1243.

Shakhnarovich, G., Viola, P., and Darrell, T. (2003). Fast pose estimation with parameter-sensitive hashing. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 750–757. IEEE.

Shao, L., Jones, S., and Li, X. (2014a). Efficient search and localization of human actions in video databases. *IEEE Trans. Circuits Syst. Video Techn.*, 24(3):504–512.

Shao, L., Zhen, X., Tao, D., and Li, X. (2014b). Spatio-temporal laplacian pyramid coding for action recognition. *Cybernetics, IEEE Transactions on*, 44(6):817–827.

Sheikh, Y. and Shah, M. (2005). Bayesian modeling of dynamic scenes for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(11):1778–1792.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124.

Sidenbladh, H. (2004). Detecting human motion with support vector machines. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 188–191. IEEE.

Sidenbladh, H. and Black, M. J. (2001). Learning image statistics for bayesian tracking. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 709–716. IEEE.

Sidenbladh, H. and Black, M. J. (2003). Learning the statistics of people in images and video. *International Journal of Computer Vision*, 54(1-3):183–209.

Sminchisescu, C., Kanaujia, A., Li, Z., and Metaxas, D. (2005). Discriminative density propagation for 3d human motion estimation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 390–397. IEEE.

Song, Y., Goncalves, L., and Perona, P. (2003). Unsupervised learning of human motion models. *Advances in Neural Information Processing Systems*, 14.

Soomro, K., Zamir, A. R., and Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Starck, J. and Hilton, A. (2003). Model-based multiple view reconstruction of people. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 915–922. IEEE.

Starck, J. and Hilton, A. (2005). Spherical matching for temporal correspondence of non-rigid surfaces. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1387–1394. IEEE.

Stauffer, C. and Grimson, W. E. L. (2000). Learning patterns of activity using real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):747–757.

Sung, J., Ponce, C., Selman, B., and Saxena, A. (2011). Human activity detection from rgbd images. In *Plan, Activity, and Intent Recognition*.

Taylor, C. J. (2000). Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 677–684. IEEE.

Turaga, P., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488.

Utsumi, A. and Tetsutani, N. (2002). Human detection using geometrical pixel value structures. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 34–39. IEEE.

Vallim, R. M., Andrade Filho, J. A., De Mello, R. F., and De Carvalho, A. C. (2013). Online behavior change detection in computer games. *Expert Systems with Applications*, 40(16):6258–6265.

Vaswani, N., Chowdhury, A. R., and Chellappa, R. (2003). Activity recognition using the dynamics of the configuration of interacting objects. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–633. IEEE.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE.

Viola, P., Jones, M. J., and Snow, D. (2003). Detecting pedestrians using patterns of motion and appearance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 734–741. IEEE.

Wang, H., Klaser, A., Schmid, C., and Liu, C.-L. (2011). Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE.

Wang, H. and Suter, D. (2005). Background initialization with a new robust statistical approach. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 153–159. IEEE.

Wang, H. and Suter, D. (2006). A novel robust statistical method for background initialization and visual surveillance. In *Computer Vision-ACCV 2006*, pages 328–337. Springer.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297. IEEE.

Wang, L., Ning, H., Tan, T., and Hu, W. (2004). Fusion of static and dynamic body biometrics for gait recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(2):149–158.

Wang, L., Tan, T., Ning, H., and Hu, W. (2003). Silhouette analysis-based gait recognition for human identification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(12):1505–1518.

Wolf, C., Mille, J., Lombardi, L., Celiktutan, O., Jiu, M., Baccouche, M., Dellandréa, E., Bichot, C.-E., Garcia, C., and Sankur, B. (2012). The liris human activities dataset and the icpr 2012 human activities recognition and localization competition. Technical report, Technical Report RR-LIRIS-2012-004, LIRIS Laboratory (March 2012).

Wu, B. and Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 90–97. IEEE.

Wu, Y. and Yu, T. (2006). A field model for human detection and tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(5):753–765.

Xia, L., Chen, C.-C., and Aggarwal, J. (2011). Human detection using depth information by kinect. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 15–22. IEEE.

Xiang, T. and Gong, S. (2006). Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67(1):21–51.

Xu, F. and Fujimura, K. (2003). Human detection using depth and gray images. In *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.*, pages 115–121. IEEE.

Xu, L.-Q. and Puig, P. (2005). A hybrid blob-and appearance-based framework for multi-object tracking through complex occlusions. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 73–80. IEEE.

Yam, C., Nixon, M. S., and Carter, J. N. (2002). On the relationship of human walking and running: automatic person identification by gait. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 287–290. IEEE.

Yampolskiy, R. V. and Govindaraju, V. (2008). Behavioural biometrics: a survey and classification. *International Journal of Biometrics*, 1(1):81–113.

Yang, C., Duraiswami, R., and Davis, L. (2005a). Fast multiple object tracking via a hierarchical particle filter. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 212–219. IEEE.

Yang, D. B., González-Baños, H. H., and Guibas, L. J. (2003). Counting people in crowds with a real-time network of simple image sensors. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 122–129. IEEE.

Yang, M.-T., Shih, Y.-C., and Wang, S.-C. (2004). People tracking by integrating multiple features. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 929–932. IEEE.

Yang, T., Pan, Q., Li, J., and Li, S. (2005b). Real-time multiple objects tracking with occlusion handling in dynamic scenes. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 970–975. IEEE.

Yilmaz, A. and Shah, M. (2005). Actions sketch: A novel action representation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 984–989. IEEE.

Zhang, L., Jiang, M., Farid, D., and Hossain, M. (2013). Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Systems with Applications*, 40(13):5160–5168.

Zhao, L. and Thorpe, C. E. (2000). Stereo-and neural network-based pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 1(3):148–154.

Zhao, T. and Nevatia, R. (2004). Tracking multiple humans in crowded environment. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–406. IEEE.

Zhen, X., Shao, L., and Li, X. (2014). Action recognition by spatio-temporal oriented energies. *Information Sciences*, 281:295–309.

Zhong, J. and Sclaroff, S. (2003). Segmenting foreground objects from a dynamic textured background via a robust kalman filter. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 44–50. IEEE.

Zhou, J. and Hoang, J. (2005). Real time robust human detection and tracking system. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 149–149. IEEE.