Computers **2013**, *2*, 88-131; doi:10.3390/computers2020088

OPEN ACCESS

computers

ISSN 2073-431X

www.mdpi.com/journal/computers

Review

A Review on Video-Based Human Activity Recognition

Shian-Ru Ke [1,*], Hoang Le Uyen Thuc [2], Yong-Jin Lee [1], Jenq-Neng Hwang [1], Jang-Hee Yoo [3], Kyoung-Ho Choi [4]

[1] Department of Electrical Engineering, University of Washington, Seattle, WA 98195-2500, USA; E-Mails: zeroth@uw.edu (Y.-J.L.); hwang@uw.edu (J.-N.H.)

[2] Department of ETE, Danang University of Technology, Danang, Vietnam; E-Mail: hluthuc@dut.udn.vn

[3] Video Surveillance Research Section, ETRI, 305-700 Daejeon , Korea; E-Mail: jhy@etri.re.kr

[4] Department of Information & Electronics Engineering, Mokpo National University, Jeollanam-do 534-729, Korea; E-Mail: khchoi@mokpo.ac.kr

***** Author to whom correspondence should be addressed; E-Mail: srke@uw.edu; Tel.: +1-206-708-3446; Fax: +1-206-543-3842.

Received: 29 November 2012; in revised form: 21 February 2013 / Accepted: 30 April 2013 / Published: 5 June 2013

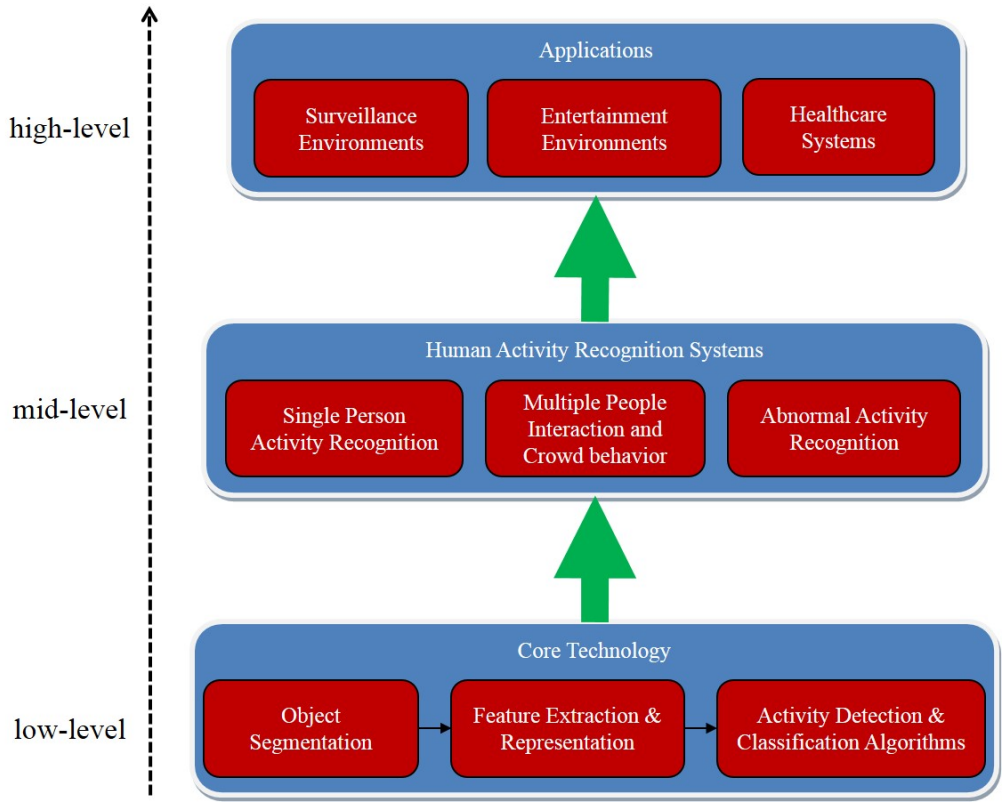
Abstract: This review article surveys extensively the current progresses made toward video-based human activity recognition. Three aspects for human activity recognition are addressed including core technology, human activity recognition systems, and applications from low-level to high-level representation. In the core technology, three critical processing stages are thoroughly discussed mainly: human object segmentation, feature extraction and representation, activity detection and classification algorithms. In the human activity recognition systems, three main types are mentioned, including single person activity recognition, multiple people interaction and crowd behavior, and abnormal activity recognition. Finally the domains of applications are discussed in detail, specifically, on surveillance environments, entertainment environments and healthcare systems. Our survey, which aims to provide a comprehensive state-of-the-art review of the field, also addresses several challenges associated with these systems and applications. Moreover, in this survey, various applications are discussed in great detail, specifically, a survey on the applications in healthcare monitoring systems.

Keywords: human activity recognition; segmentation; feature representation; security surveillance; healthcare monitoring; human computer interface

1. Introduction

In recent years, automatic human activity recognition has drawn much attention in the field of video analysis technology due to the growing demands from many applications, such as surveillance environments, entertainment environments and healthcare systems. In a surveillance environment, the automatic detection of abnormal activities can be used to alert the related authority of potential criminal or dangerous behaviors, such as automatic reporting of a person with a bag loitering at an airport or station. Similarly, in an entertainment environment, the activity recognition can improve the human computer interaction (HCI), such as the automatic recognition of different player's actions during a tennis game so as to create an avatar in the computer to play tennis for the player. Furthermore, in a healthcare system, the activity recognition can help the rehabilitation of patients, such as the automatic recognition of patient's action to facilitate the rehabilitation processes. There have been numerous research efforts reported for various applications based on human activity recognition, more specifically, home abnormal activity [1], ballet activity [2], tennis activity [3,4], soccer activity [5], human gestures [6], sport activity [7,8], human interaction [9], pedestrian traffic [10] and simple actions [11–22], and healthcare applications [1,23–38]. In this paper, the video based technologies for human activity recognition will be extensively reviewed and discussed.

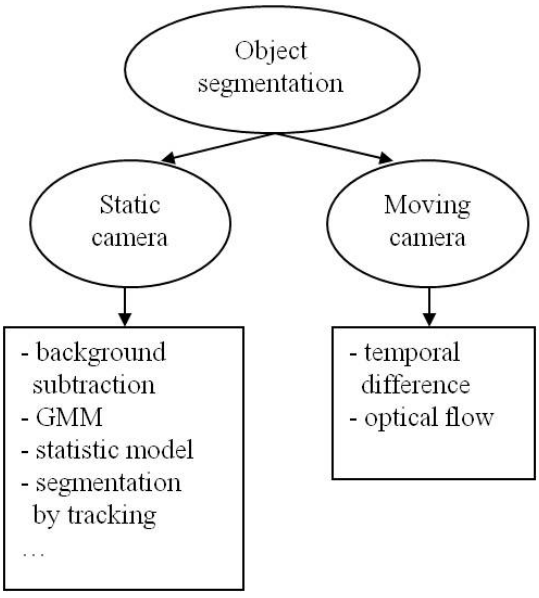
Figure 1. The overview of a general system for human activity recognition.



Generally speaking, human activity recognition can be separated into three levels of representations, individually the low-level core technology, the mid-level human activity recognition systems and the high-level applications as shown in Figure 1. In the first level of core technology, three main processing stages are considered, i.e., object segmentation, feature extraction and representation, and activity detection and classification algorithms. The human object is first segmented out from the video sequence. The characteristics of the human object such as shape, silhouette, colors, poses, and body motions are then properly extracted and represented by a set of features. Subsequently, an activity detection or classification algorithm is applied on the extracted features to recognize the various human activities. Moreover, in the second level of human activity recognition systems, three important recognition systems are discussed including single person activity recognition, multiple people interaction and crowd behavior, and abnormal activity recognition. Finally, the third level of applications discusses the recognized results applied in surveillance environments, entertainment environments or healthcare systems.

In the first stage of the core technology, the object segmentation is performed on each frame in the video sequence to extract the target object. Depending on the mobility of cameras, the object segmentation can be categorized as two types of segmentation, the static camera segmentation and moving camera segmentation as shown in Figure 2.

Figure 2. The categories for object segmentation.



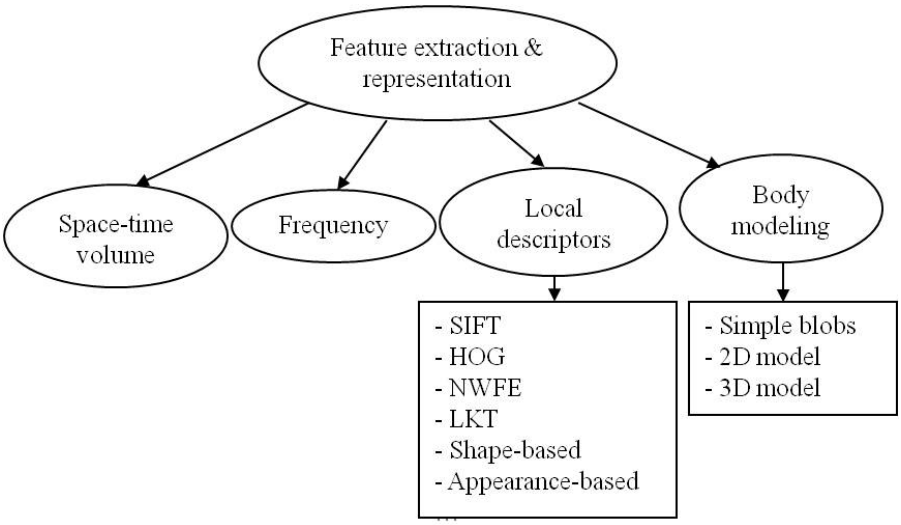
For the static camera segmentation, the camera is fixed in a specific position with a fixed angle. Hence, the viewpoint of the object and the background are fixed. The most popular method for static camera segmentation is background subtraction [39–41] due to its simplicity and efficiency. The background image without any foreground object is first established. After that, the current image in the analyzed video sequence is subtracted from the background image to obtain the foreground objects. However, the background subtraction is very sensitive to illumination changes. More complex methods to form the background model from a bunch of background images are proposed. The background model can be built by a single Gaussian or a mixture of Gaussians for each pixel [42,43], or by statistical parameters including the intensity change and chromaticity change for each pixel [44].

In addition, static camera segmentation by tracking [45,46] has also been proposed, i.e., unlike the point-based segmentation methods to form a background model in advance, based on a dynamic time warping like algorithm, the object can also be segmented by tracking regions which are spatially cohesive with locally smooth motion.

On the other hand, for moving camera segmentation, the camera is moving with a photographer or a robot. The moving camera segmentation is more challenging than a static one because, in addition to the motion of the target object, it also needs to consider the motion of the camera and the change of background. Because of the dynamic background, the background model is not appropriate for moving camera segmentation. The most common method for moving camera segmentation is the temporal difference [47,48], i.e., the difference between consecutive image frames. Also, the motion of moving camera can be inferred by optical flow [49,50], which estimates the pixel-level motion between two images. The features of images are then tracked and the coordinate between consecutive images is also transformed. Hence the moving object can be segmented based on the transformed coordinate of moving camera.

In the second stage of the core technology, characteristics of the segmented objects such as shape, silhouette, colors and motions are extracted and represented in some form of features. Generally speaking, the features can be categorized as four groups, space-time information, frequency transform, local descriptors and body modeling, as shown in Figure 3.

Figure 3. The categories for feature extraction and representation.

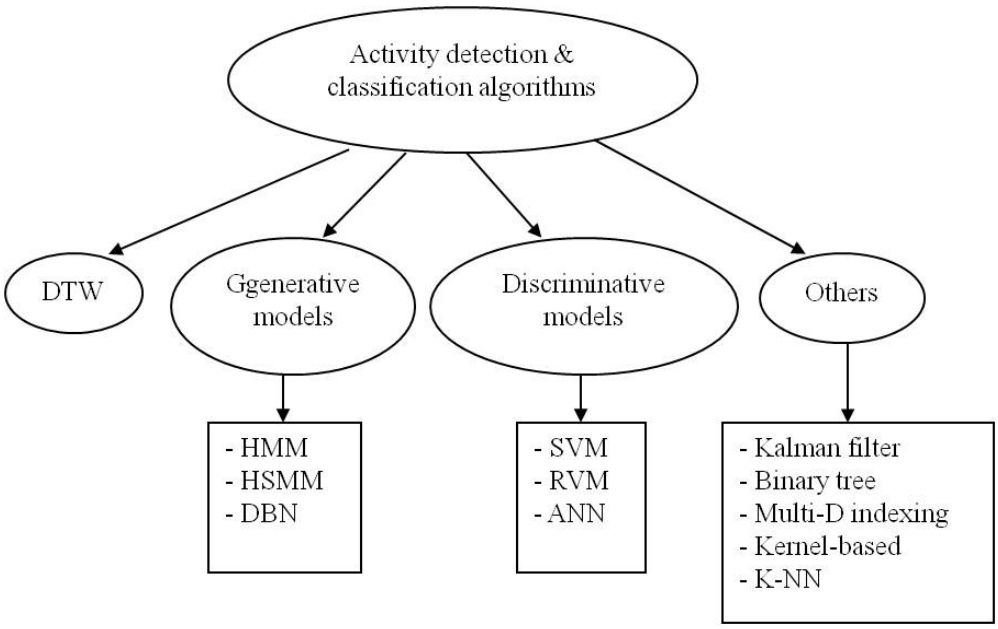


The space-time information is first considered. The space-time volume (STV) [2,3,11,51] is built as the image features by concatenating the consecutive silhouette of objects along the time axis. The extracted 3D XYT volume (along *x*-*y* spatial coordinates and time) can capture the continuity of human action. But the STV is limited on non-periodic activities. Compared to spatial-temporal domain image, the frequency domain information can also be exploited. More specifically, the discrete Fourier transform (DFT) [22], which has been widely used to represent information about the geometric structure of the object. The STV and DFT features belong to global features which consider the whole image so that they are limited on viewpoint changes and occlusion. Hence, some local descriptors are considered. The local descriptors [5,8,12–14,52–56], such as scale-invariant feature transform (SIFT)

[57,58] and histogram of oriented gradient (HOG) [59] capture the characteristics of an image patch. They are ideally invariant to background clutters, appearance and occlusions, and also invariant to rotation and scale in some cases. However, the above mentioned feature representations do not fully capture the whole body actions. Therefore, some human modeling methods [15,39,52,60–70] are also proposed to model the human body including simple blobs, 2D body modeling and 3D body modeling. Generally, the body modeling requires the 2D/3D pose estimation problem. Usually, after the pose estimation, the 2D/3D coordinates of the human body are further converted into other dimension-reduced or more discriminative feature representations, such as polar coordinate representation [71], Boolean
features [72] and geometric relational features (GRF) [73,74].

In the third stage of the core technology, the activity detection and classification algorithms are used to recognize various human activities based on the represented features. They can generally be categorized as dynamic time warping (DTW), generative models, discriminative models and others as shown in Figure 4.

Figure 4. The categories for activity detection and classification algorithms.



The dynamic time warping (DTW) [14,16], a method for measuring similarity between two temporal sequences, which may vary in time or speed, is one of the most common temporal classification algorithms due to its simplicity; however, DTW is not appropriate for a large number of classes with many variations. Some probability-based methods by generative models (dynamic classifiers) are proposed such as Hidden Markov Models (HMM) [1,4,6,17,75–77] and Dynamic Bayesian Networks (DBN) [7,9,78]. On the other hand, discriminative models (static classifiers) such as Support Vector Machine (SVM) [18,19,79,80], Relevant Vector Machine (RVM) [54,81,82] and Artificial Neural Network (ANN) [29,83,84], can also be used in this stage. In addition to the dynamic and static classifier difference nature, another main difference between generative models and discriminative models [85] is that the generative classifiers commonly learn a model of the joint probability, $p(x,y)$, of the input x and the label y, or equivalently the likelihood $p(x|y)$ according to

Bayes' rule; while the discriminative classifiers model the posterior $p(y|x)$ directly. Therefore, the generative models can be used to simulate values of any variables in the models, while the discriminative models allow only sampling of the target variables conditional on the observed variables. For both of the probability model-based algorithms, including generative models and discriminative models, their performance relies on extensive training dataset. Therefore, other methods are proposed, such as Kalman filter [10,86], binary tree [20,87], multidimensional indexing [21], and K nearest neighbor (K-NN) [22]. Different classification algorithms usually require different sets of suitable feature representations.

After object segmentation, feature representation, classification stages in the low-level core technology, the three main aspects of the mid-level human activity recognition systems are discussed, including single person activity recognition, multiple people interaction and crowd behavior, and abnormal activity recognition. Finally, the recognized human activities have been applied mainly in three fields, i.e., surveillance environments, entertainment environments, and healthcare systems.

Several video-based activity recognition survey papers have been published in the last two decades. The survey paper by Aggarwal and Shangho [88] focuses on modeling of motion and recognition of actions and interactions. Valera and Velastin [89] survey the automated visual surveillance systems. Moeslund *et al.* [90] emphasize the human motion capture and analysis, including human model initialization, tracking, pose estimation and action recognition. Krüger *et al.* [91] pay more attention to the recognition of actions at different levels of complexity. Turaga *et al.* [92] focus on the recognition of actions and activities and not on the lower-level processing modules such as detection and tracking. Enzweiler and Gavrila [93] concentrate on the detection of pedestrians. Candamo *et al.* [94] focus on the recognition of human behaviors in transit scenes. The most recent survey by Aggarwal and Ryoo [95] emphasizes the recognition of human actions, interactions and group activities. Jiang *et al.* [96] focus on event recognition, rather than on human activity or crowd behavior. Enzweiler and Gavrila [97] emphasize the pedestrian detection. Our survey, which aims to provide the comprehensive state-of-the-art review of the field, covers all processing stages of the activity recognition system in a wide scope, from the low-level processing stages to the mid-level human activity recognition systems , and even to the high-level applications, especially, on the applications to healthcare monitoring systems.

In this paper, we will conduct a thorough review of all the three representation levels, i.e., core technology, the human activity recognition systems and the relevant applications. The paper is organized as follows. In Section 2, we review objection segmentation methods for static cameras and moving cameras. Section 3 addresses the feature extraction and representations. Section 4 discusses the activity detection and classification algorithms. Three main aspects of human activity systems are described in Section 5. Three applications are introduced in Section 6. The conclusions and future direction are drawn in Section 7.

2. Object Segmentation

For human activity recognition, the first stage is to do the object segmentation, i.e., the human objects are segmented from the background image. Based on the mobility of the camera, the object

segmentation task can be divided into two categories, the static camera segmentation and the moving camera segmentation.

2.1. Static Camera

In static camera segmentation, the camera is fixed in a specific position and angle. Since the background never moves, it is natural to build a background model in advance, so that the foreground object can be segmented from the image of the background model.

2.1.1. Background Subtraction

The most common method for static camera segmentation is background subtraction due to its simplicity and efficiency [98]. The background model contains only the stationary background scene without any foreground object, and any image change is assumed to be caused only by moving objects. Hence the foreground object can be obtained by subtracting the current image of the background image, followed by a magnitude thresholding to obtain the segmentation mask. The segmentation mask often contains rough and fractional foreground object(s) and usually requires some post-processing, such as closing and opening morphological operations. The background subtraction has been extensively applied in all kinds of scenarios with various improved modifications. For example, for real-time human body tracking [39], the color distribution of each pixel in the background is first modeled with a Gaussian with a full covariance matrix. This background scene texture map is considered to be class zero. The foreground textures in different classes are grouped by the mean of a point and the covariance associated with that point. Another improvement is to discriminate moving objects, ghosts and shadow [40], based on statistical assumptions, with object-level knowledge, of moving objects, apparent objects (ghosts) and shadows. Besides, in order to overcome the limitation of the background subtraction on stationary background, Seki *et al.* [41] proposes a method to handle the dynamic (waving) background. The method learns the chronological changes in the observed scene's background in terms of distribution of image vectors. Generally speaking, the background subtraction is simple and efficient, but the simplicity of the background model sometimes causes the inaccurate classification of the pixels. Continuous and effective updating of the background in response to gradual changes of background also poses some challenges.

2.1.2. Gaussian Mixture Model (GMM)

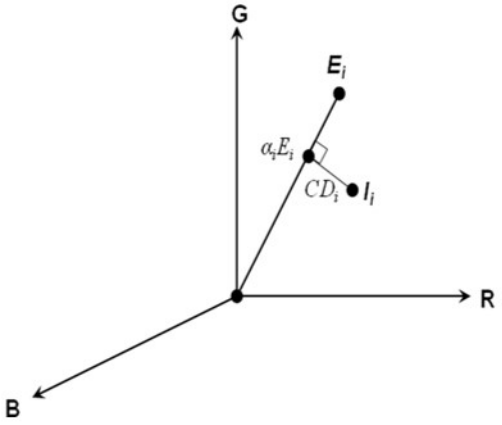
Instead of the simple one-Gaussian per pixel background modeling, the pixel values at location (x,y) can be modeled as a more complicated form, such as a mixture of Gaussians, to accommodate different background scenarios. The Gaussian mixture model (GMM) has been extensively applied in many fields to allow the adaptation to the multi-modal environments. Generally, GMM is learned by the expectation maximization (EM) algorithm. The higher the probability of a pixel value in the GMM, the more likely the pixel belongs to the background. Therefore, for the image sequence, a pixel of the image is classified to belong to the foreground object if the probability of the pixel value is less than a predefined threshold. In [42], the GMMs are constructed over a variety of different color and texture feature spaces. Instead of using EM, Permuter *et al.* apply the *k-means* clustering algorithm to reduce the high computation incurred by the EM algorithm. Although the likelihood value attained by the k-means is slightly lower than that attained by the EM algorithm, with the same amount of data involved in training the background model, the difference of performance is insignificant. It is sometimes

necessary to reduce the dimensions of the features to avoid the frequently encountered singular covariance problems when the training data are not sufficient [43]. In short, the GMMs can be effectively modified to describe more complex background, but the high computation cost involved in the EM training imposes a trade-off.

2.1.3. Statistical Models

Besides GMM, the methods to build more complicated background models are also proposed. One of the typical methods is to build the background model as a statistical model. In [44], each pixel is modeled by four parameters, individually brightness distortion, chromaticity distortion, the variation of the brightness distortion, and the variation of the chromaticity distortion. As shown in Figure 5 [44], E_i is the expected color value (say, as recorded in the background image) and I_i is the color value of the *i*th pixel. The *brightness distortion* (CD_i) is defined as the shortest distance between the *i*th pixel I_i and the line $\overline{E_iO}$. The *chromaticity distortion* is defined as the coefficient α_i which minimizes the brightness distortion distance, CD_i. Based on the thresholding on brightness distortion and chromaticity distortion, every pixel in the current image can be segmented (classified) into one of the following group: the original background, shadow, highlighted background, and the moving foreground. More specifically, the background pixels are classified if brightness and chromaticity distortion are small. The shadow pixels are classified if the chromaticity distortion is small but with lower brightness. And the foreground pixels are classified if the chromaticity distortion is high. Compared to GMM, the statistical models are generally more efficient in building the background model, and can be used to segment out not only the foreground objects but also the shadows.

Figure 5. Color value I_i of the *i*th pixel for the statistic model in RGB color space [44].



2.1.4. Segmentation by Tracking

The above point-based methods, including background subtraction, GMM and statistical models, describe the low-level segmentation, which can capture the pixel-level details of the objects but is lacking the global information of objects. Hence, the region-based methods for segmentation by tracking are proposed. Brendel *et al.* [45] propose a segmentation method by tracking regions across frames with a new circular dynamic-time warping (CDTW) algorithm, which generalizes the

conventional DTW algorithm to match closed boundaries of two regions, for region matching to identify the longest, best matching boundary portion of two regions. Moreover, Yu *et al.* [46] propose another segmentation method by tracking the spatial-color Gaussian mixture models (SCGMM). The SCGMMs are built in five dimensions, *X*, *Y*, *R*, *G*, *B*, with *X*, *Y* for spatial information and *R*, *G*, *B* for color information. Yu *et al.* also propose a tracking algorithm to iteratively update SCGMMs by fixing the color Gaussian models while updating spatial Gaussian models, and vice versa, with a constrained expectation maximization (EM) algorithm. Both [45] and [46] are region-based segmentation by tracking the objects' regions, whose performance highly depends on the performance of the tracking algorithms. It can avoid the pepper-and-salt noise frequently encountered in the pixel-based segmentation methods by tracking the target regions. However, the methods of segmentation by tracking need to have a robust scheme to segment out the target objects at the very first frame. Several methods have been proposed for the very first frame object segmentation, e.g., [45] applies mean-shift segmentation, on the other hand [46] adopts a boosting face detector [99] and a graph-cut segmentation.

2.2. Moving Camera

Unlike static camera with fixed location and angle, moving camera (e.g., the camera installed on cars, moving robots, flying vehicles, *etc.*), including the use of active camera (e.g., pan-tilt camera), is with dynamic location and angle. Moving camera segmentation is much more challenging than static camera segmentation because two questions are needed to be considered simultaneously, i.e., the motion of the background and the motion of each foreground moving object. Generally, camera motion decomposition is needed to separate the motion of the camera and the motion of the objects.

2.2.1. Temporal Difference

The most common method for moving camera segmentation is the temporal difference between consecutive frames. Unlike static camera segmentation, where the background is comparably stable, the background is changing along time for moving camera; therefore, it is not appropriate to build a background model in advance. Instead, the moving object is detected by taking the difference of consecutive image frames *t-1* and *t*. However, the motion of the camera and the motion of the object are mixed in the moving camera. Hence, the motion of camera is estimated first. In [47], Murray *et al.* propose a temporal difference method for the segmentation on pan-tilt active camera. The background compensation is first applied for apparent motion of the background caused by the camera motion, and for finding a relationship between pixels representing the same 3-D point in images taken from different camera orientations. Morphological operations, including erosion and dilation, are then applied to smooth the absolute difference between current frame and previous frame to obtain the motion edges, resulting in the detection the moving objects when combined with the object edge information. Moreover, Kim *et al.* [48] propose another temporal difference method to estimate the pan and tilt movements of the camera by edge features in consecutive frames. The estimated pan-tilt motion parameters are then used to transform the image coordinate. Consequently, a motion image is created by using the estimated difference value among three consecutive frames by Equation (1), in

which a pixel is considered as a motion pixel if the difference with previous frame and the difference with next frame are larger than a predefined threshold.

$$\left|f_t(x,y)-f_{t-1}(x,y)\right|>T_1 \;\&\; \left|f_{t+1}(x,y)-f_t(x,y)\right|>T_1 . \quad (1)$$

The advantage of the temporal difference methods [47,48] is efficient due to computational simplicity. However, they need to first perform the camera motion compensation, which is generally sensitive to noise due to the consecutive image difference.

2.2.2. Optical Flow

Another category for segmentation on moving camera is optical flow, which denotes a displacement of the same scene in the image sequence at different time instant. The pixel-based local optical flow in image sequence can be robustly evaluated by the Lucas-Kanade-Tomasi (LKT) feature tracker [100,101], which effectively selects corner feature points of the reference image patch. In [49], Daniilidis *et al.* apply an FIR-kernel based LKT feature tracker to estimate the optical flow and to infer the motion of objects. The spatial FIR-kernels are binomial approximations to the first derivatives of the Gaussian function. Moreover, Huang *et al.* [50] also apply the LKT feature tracker to obtain the optical flow. Those features points with similar optical flows (similar magnitude and orientations) are then grouped together. Finally, the detected moving object patch is validated by target's color histogram as well as contour outlier removing. Even though the optical flow can be estimated by the LKT feature tracker, which robustly captures the local descriptor, it will perform poorly when the reference image patch is occluded by the moving target, the feature points originally located at the background will be moved with the target and result in inaccurate estimation of the optical flow. To overcome this issue, the LKT feature points need to be updated every few frames.

3. Feature Extraction and Representation

The second stage for human activity recognition is feature extraction and representation, where the important characteristics of image frames are extracted and represented in a systematical way as features. Feature extraction and representation have crucial influence in the performance of recognition, therefore it is essential to select or represent features of image frames in a proper way. In a video sequence, the features that capture the space and time relationship are known as space-time volumes (STV) [2,3,11,51]. In addition to spatial and temporal information, discrete Fourier transform (DFT) [22] of image frames mainly captures the image intensity variation spatially. The STV and DFT are global features which are extracted by globally considering the whole image. However, the global features are sensitive to noise, occlusion and variation of viewpoint. Instead of using global features, some methods are proposed to consider the local image patches as local features. Ideally, the local features are designed to be more robust to noise and occlusion, and possibly to rotation and scale. Besides global and local features, other methods are also proposed to directly or indirectly model human body, to which the pose estimation and body part tracking techniques can be applied. Moreover, the coordinates of the body modeling can be further converted into lower-dimensional or more discriminative features, such as polar coordinate representation [71], Boolean features [72] and geometric relational features (GRF) [73,74], for effective recognition purpose.

3.1. Space-Time Volumes (STV)

The space-time volume (STV) is formed by temporally stacking frames over a video sequence as a 3D cuboid of spatial-temporal shape. Blank *et al.* [2] propose a method, by stacking segmented silhouette frame-by-frame, to form a 3D spatial-temporal shape, from which the space-time features such as local space-time saliency, action dynamics, shape structure and orientation can be extracted [2]. Ke *et al.* [3] further uses the spatial-temporal shapes for shaped-based matching, including spatial-temporal region extraction and region matching. For region matching, an unsupervised clustering technique is applied to group the video into classes of 3D volumes of consistent appearance. In order to overcome the limitation of shape-based approaches, such as changes in camera view and variability in the speed of actions, Ke *et al.* [3] also incorporate Shechtman and Irani's flow-based features [51] into the classifier to improve the performance. Moreover, Dollar *et al.* [11] applies a spatio-temporal interest point detector to find local region of interest in the cuboids of space and time for activity recognition. First, cuboids of spatio-temporally windowed data surrounding a feature point extracted from sample behaviors are clustered to form a dictionary of cuboid prototypes. The histogram of the cuboid types is then used as an activity descriptor for object recognition. Generally, the STV features provide a proper way to combine spatial and temporal information; however, STV features normally require good segmented silhouette and are sensitive to viewpoint and occlusion.

3.2. Discrete Fourier Transform (DFT)

Besides spatial and temporal information, the frequency domain information, that is, the intensity variation of an image can also be taken advantage of. Kumari and Mitra [22] use discrete Fourier transforms (DFTs) of small image blocks as the selected features for activity recognition. It is generally assumed that the gray-level intensity of foreground object is different from that of background object; therefore, DFT of an image can be used to obtain information of the geometric structure (shape) in the spatial domain. The K-nearest neighbor (K-NN) algorithm can then be applied to the DFT features for human activity recognition. Generally, the DFT features can capture the shape energy of image frames, but are also sensitive to noise and occlusion.

3.3. Local Descriptors

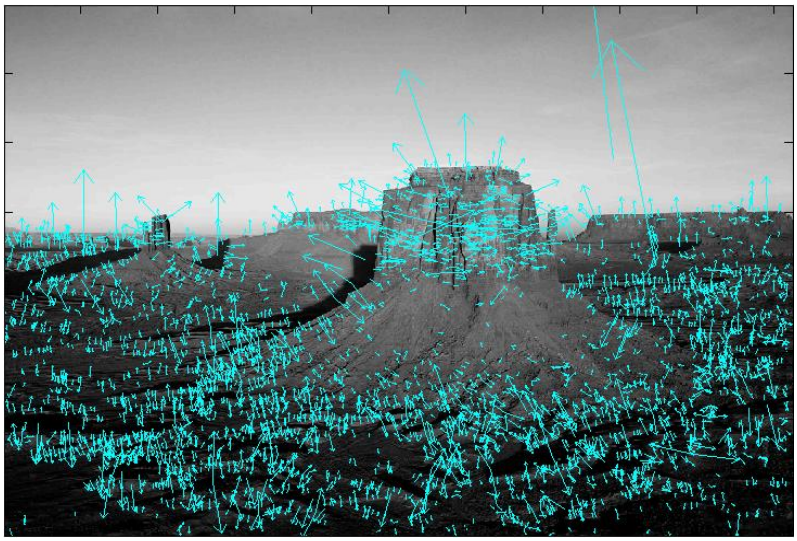
As discussed previously, local descriptors are designed to be more robust to noise and occlusion, and possibly invariant to rotation and scale, such as scale-invariant feature transform (SIFT) features [12,57,58], histogram of oriented gradient (HOG) features [5,59], nonparametric weighted feature extraction (NWFE) features [13], Lucas-Kanade-Tomasi (LKT) features [8,100,101], shape-based features [14,54–56] and appearance-based features [52,53].

3.3.1. SIFT Features

Lowe [57,58] first proposes a class of local image features, which are invariant to image scaling, translation, and rotation, and partially invariant to illumination change and affine or 3D projection, known as scale invariant feature transform (SIFT). The scale-invariant features are enabled by using a

staged filtering approach. As proposed in [58], the four major steps applied on an image frame to generate SIFT features are summarized as follows. The first step is the scale-space extrema detection, which applies a difference-of-Gaussian (DoG) function to search over all scales and locations of the image, and to identify interest points that are invariant to scale and orientation. The second step is the keypoint localization. Based on the identified interest points in the first step, a detailed model is used to determine the scale and location, and then the more stable points are selected as keypoints. The third step is the orientation assignment based on local image gradient directions. The last step is the keypoint descriptor. At each keypoint, the gradients of the image are measured at the selected scale, and then the measured image gradients are transformed into a representation to allow for local shape distortion and illumination changes. One example of the selected keypoints for SIFT features is shown in Figure 6 [58]. Scovanner *et al.* [12] further introduce a 3D SIFT descriptor, which can reliably capture the spatio-temporal nature of the video data as well as 3D imagery such as MRI data. The SIFT descriptor is a popular descriptor due to its invariance to image rotation and scale and robust to affine distortion, noise corruption and illumination change. However, the high dimensionality is a drawback of SIFT in the feature matching step, because the distinctiveness of SIFT descriptor is achieved by assembling a high-dimensional vector representing the image gradients within a local region of the image. Another significant drawback is the SIFT features are not sufficiently discriminative. For example, the SIFT features are equally likely to be found on human objects and on the background. Moreover, the SIFT uses only grayscale information and misses important appearance information, such as color.

Figure 6. The 2D scale-invariant feature transform (SIFT) descriptor, where the arrows indicate the local image gradient directions with information of scale, orientation and location [58].

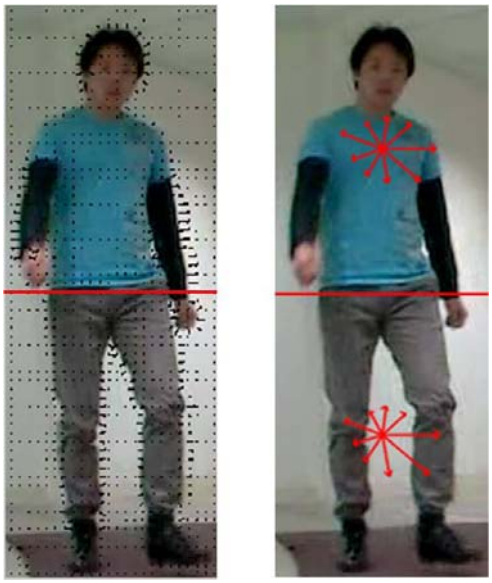


3.3.2. HOG Features

Dalal and Triggs [59] were the first to propose histogram of oriented gradient (HOG) descriptors for human detection. The HOG is derived based on evaluating normalized local histograms of image gradient orientations in a dense grid by considering fine-scale gradient, fine orientation binning,

relatively coarse spatial binning and high-quality local contrast normalization in overlapping descriptor blocks. An example of HOG descriptors is shown in Figure 7. Lu and Little [5] further propose a template-based algorithm to track and recognize athlete's actions by exploitation of the PCA-HOG descriptor. First, an HOG descriptor is applied to images in a video, and then is projected to a linear subspace by the principal component analysis (PCA). The proposed PCA-HOG descriptor in [5] is invariant to the variation of illuminations, poses, and viewpoints. At each time instance *t*, three procedures are performed, which are individually sequentially tracking, action recognition and template updating [5]. The major drawback of HOG features is that the local descriptors are extracted at a fixed scale; therefore, the size of the human in the image can have great influence on the performance.

Figure 7. An example of histogram of oriented gradient (HOG) descriptor: (**a**) the image gradients of local grids; (**b**) the HOG descriptor with nine bins for the upper body and the lower body [59].



(**a**) (**b**)

3.3.3. NWFE Features

Taking into account the distance information and the width feature of a silhouette, Lin *et al.* [13] propose a new feature, called nonparametric weighted feature extraction (NWFE), to build histogram vectors for human activity recognition by using the nearest neighbor classifiers. NWFE features are extracted from the pose contour by combining the distance [102] and width features [103], which are then projected from the original high-dimensionality feature space to a low-dimensional subspace by PCA transformation and K-means clustering. The NWFE features reduce the computational complexity and still achieve high-recognition rate. However, the main drawbacks of NWFE features are that it relies on accurate human body silhouette and contour, and ignore the color appearance information of the image.

3.3.4. LKT (Lucas-Kanade-Tomasi) Features

Lucas-Kanade [100] and Tomasi [101] propose a point tracking method based on the sum of squared intensity differences, named LKT (Lucas-Kanade-Tomasi) feature tracker. An example of fist tracking using LKT feature trackers is shown in Figure 8. Lu *et al.* [8] use an LKT feature tracker to track human body joints in key frames and actual frames, where key frames contain the prototype information to represent defined posture (one step of action). And a proposed factorized sampling algorithm is used to replace one tracker with a set of LKT trackers to enhance the joints tracking accuracy and recognize non-rigid human actions. Since the LKT feature tracker assumes that neighboring pixels in a small window have the same flow vector, resulting in the main limitation of being difficult to deal with large motion between frames.

Figure 8. An example of fist tracking using one Lucas-Kanade-Tomasi (LKT) tracker. The red dots denote the LKT feature points, and the yellow line denotes the tracking by one LKT tracker [100,101].

3.3.5. Shape-Based Features

Shape analysis is essential for object tracking and activity recognition. Generally, the shape extraction of human objects should be robust to errors in the silhouette extraction process and likely designed to be invariant to translation, scale and rotation. Several features describing a shape have been developed in the literature [14,55,56,104,105]. Veeraraghavan *et al.* [14] propose a method, which exploits the shape deformations of the human silhouette, to extract the shape-based features to be used for human gaits recognition. The similarity measure used for comparing two shape sequences is an extension of the dynamic time warping (DTW) [104] algorithm. To have a more efficient description of shapes along the time, the Kendall's statistical shape [105] is further used as a sparse shape descriptor to model the dynamics of shape of human objects in a video sequence [14].

Furthermore, both unsupervised and supervised methods can be used for human activity recognition. For the unsupervised method, the features proposed by Schindler and Gool [55], including shape-based and flow-based features, are extracted in a biologically-inspired fashion. On the other hand, for the supervised method, the features proposed by Danafar and Gheissari [56] are histograms of optical flow, which capture both local and global information of actions. The disadvantage of shape-based features is that the internal body-part movement of human object is difficult to be detected in the silhouette region. Moreover, the shaped-based features need an accurate silhouette segmentation, which is difficult to achieve even by the state-of-the-art background subtraction methods, especially in dynamic environment. However, flow-based features can complement these disadvantages with no need of background subtraction.

3.3.6. Appearance-Based Features

Compared with shaped-based features, appearance-based features can provide more discriminative information, such as color, and can be more robust in dealing with occlusion. Several appearance descriptors have been proposed in the literatures [52,53]. Sedai *et al.* [52] propose a novel noise-resilient appearance descriptor, called histogram of local appearance context (HLAC) for 3D human pose estimation. The local appearance context (LAC) descriptors are first computed on the locations of human objects in an image based on HOG, and PCA is further applied for dimensionality reduction, followed by the histogramming of LAC to obtain HLAC. The experimental results showed that HLAC features can have superior performance for human activity recognition due to its incorporation of local appearance and the minimum clutter effect since no background subtraction is needed.

Moreover, Ramanan *et al.* [53] also use appearance-based features for human tracking. A discriminative model of appearance (color) is built either by a bottom-up approach [106] that looks for candidate body parts in each frame, or by a top-down approach [107] that looks for the entire human body. In [53], an instance-specific model is built first to capture a person's appearance (color). The person and body parts are then tracked by detecting the model in each frame. However, the main limitation of the appearance model is that it is sensitive to clothing and illumination changes.

3.4. Body Modeling

Besides the global and local features, some methods are proposed to model human body for human body tracking, pose estimation and human activity recognition. Depending on whether the use or not an explicit *a priori* body model [90,108], the body modeling can be roughly categorized in three classes [90,108], *i.e.*, model-free, indirect model and direct model. Normally, before going to classification/recognition stage, the coordinates of the body modeling can be further converted or dimensionality reduced into some features with more efficient and discriminative forms, such as polar coordinate representation [71], Boolean features [72] and geometric relational features (GRFs) [73,74].

3.4.1. Model-Free

There is no *a priori* model, only simple blobs to represent the human poses. In [39,60], algorithms are developed to detect the foreground, detect the blobs and track the blobs, which are used to represent head/hands. Nakazawa *et al.* [61] use an ellipse to represent the human body and perform ellipse tracking by four steps, i.e., extraction of the human region from the image, generation of simulated image, matching and updating of human position. Furthermore, Iwasawa *et al.* [62] use stick-figures, resembling the human skeleton, to represent human structure information so as to enable the human body tracking. Generally, this model-free class relies on extensive training using ground truth data obtained by commercial motion capture systems.

3.4.2. Indirect Model

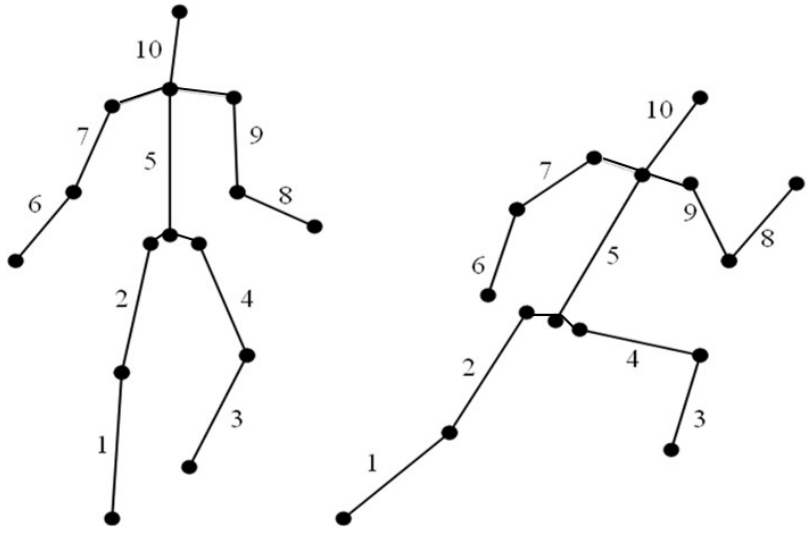
In an indirect model scheme*, a priori* model is only used indirectly to guide the interpretation of measured data. Leung and Yang [63] use U-shaped edges to describe the outline of a moving human body from a video sequence. A moving edge detection technique is proposed to generate a more

complete outline of the moving object, based on the difference picture and the coincidence edges which are edges of both the difference picture and the original intensity picture. Moreover, Huo *et al.* [15] use a head-shoulder-upper-body model to detect and track humans by particle filtering [109]. The 2D and 3D coordinates of the torso and both hands are then transformed into normalized feature space for pose estimation. In this indirect model class, the estimated poses are generally not very detailed, and it is not easy to handle occlusion or to impose kinematics constraints for human body configuration.

3.4.3. Direct Model

In the direct model scheme*, a priori* human model (i.e., the explicit 3D geometric representation of human space and kinematic structure) is directly used as the model, which represents the observed subject and is continuously updated by the observations [52,64–70]. Sedai *et al.* [52] use an HLAC image descriptor (as discussed in Section 3.3.6) for 3D human pose estimation from monocular video sequence (see a 3D human model shown in Figure 9 [52]). The orientation of each body part is represented by three Euler angle components. Leong *et al.* [64] propose a body feature extraction method to extract 21 feature points and 35 feature lines on the scanned human torso by a full body scanner. This approach requires rather accurate torso estimation and is very sensitive to noise and occlusion. Rogez *et al.* [68] use a series of view-based shape-skeleton models for video surveillance systems. Various viewpoints of shape-skeleton models are established by projecting the input image frames onto the training plane and eventually the selected view-based models are employed for feature extraction in the warped image. The major concern of this method is the extensive training dataset needed due to the use of various viewpoints.

Figure 9. A 3D human body with orientations of the 10 body parts, namely torso, head, upper/lower left/right arm/leg [52].

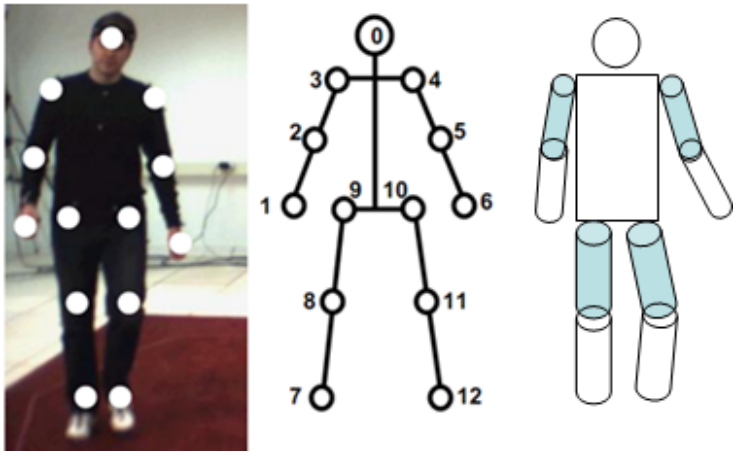


Moreover, Lee and Cohan [65] use a human kinematics model for 3D human pose estimation. The human model contains three components including kinematic, shape and clothing. A data-driven approach based on Markov Chain Monte Carlo (DD-MCMC) [110,111] is used to recover the 3D human poses. Lee and Nevatia [66,67] use a multi-level structure model for 3D human pose estimation

from monocular video sequence. The three-level method addresses several issues including automatic initialization, data association, self- and inter-occlusion. At the first level, the positions and sizes of human objects are coarsely estimated. In the second level, body parts such as faces, shoulders and limbs are detected, and the results are combined by a grid-based belief propagation algorithm to infer 2D joint positions. At the last level, the derived belief maps are formatted as proposal functions to infer 3D human poses by using DD-MCMC. The performance is quite promising but the computation cost is extremely high.

Furthermore, Ke *et al.* [69,70] propose a method to do body-part tracking by integrating skeleton, color and temporary information and estimate 3D human poses with a predefined 3D human model as shown in Figure 10 [70]. The proposed system can do real-time front-view 3D human pose estimation [69], and view-invariant 3D human pose estimation [70], which requires close to real-time computational cost. The main drawback of the proposed systems is that it has limited capability to recover the human poses while tracking errors happen.

Figure 10. Human body models [70] (**a**) human image, (**b**) 13-point model, (**c**) 3D model.



(**a**) (**b**) (**c**)

4. Activity Detection and Classification Algorithms

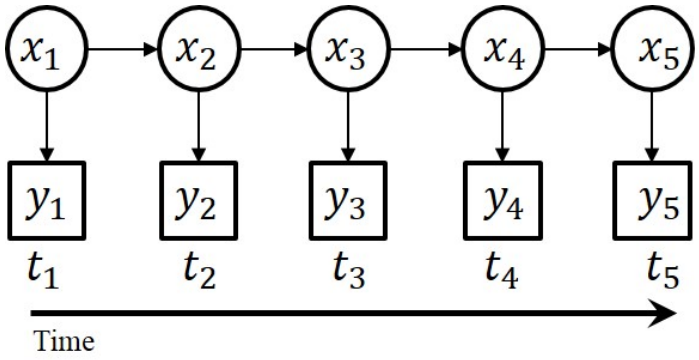
After selecting proper features from image or video, activity detection and classification algorithms are the next stage under consideration for human activity recognition. To achieve good recognition performance, it is essential to choose a proper classification algorithm using the selected feature representation. One of the well-known classification algorithms is dynamic time warping (DTW) [104], which is a similarity measure for two sequences, possibly with different length and different rate of occurrence. However, the main drawback of DTW is that it needs extensive templates. For example, only for speech recognition of the English character alphabets, it might need thousands of templates based on different accents. Therefore, DTW has issues of extensive templates for dataset, resulting in high computation cost. To overcome the issues of DTW, many model-based methods are thus proposed. Basically, the model-based algorithms can be divided into generative models and discriminative models. Generative models, that explicitly simulate the generation process of the data sequences as achieved by the hidden Markov model (HMM) [75] and dynamic Bayesian network (DBN) [78], learn the joint probability distribution $P(X,Y)$over observation X and label sequence Y, or

equivalently the likelihood $P(X|Y)$. On the other hand, discriminative models learn the conditional (posterior) probability distribution $P(Y|X)$over an unobserved fixed-length class label Y on a given observed fixed-length feature vector X, such as support vector machines (SVMs) [79,80], relevance vector machines (RVMs) [81,82] and artificial neural networks (ANNs) [84]. Some other popular classification algorithms are also introduced here including Kalman filter [86], binary tree, multidimensional indexing [21] and K nearest neighbor (K-NN) [22].

4.1. Dynamic Time Warping (DTW)

DTW [104] is one of dynamic programming algorithms to measure similarity (distance) between two sequences, i.e., one kind of template matching algorithms. Veeraraghavan *et al.* [14] suggest a modification of DTW algorithm to include the non-Euclidean space, in which the shape deformations take place, to match shape sequences for human movement. Moreover, Sempena *et al.* [16] use DTW to recognize various human activities such as waving, punching and clapping. An exemplar-based sequential single-layered approach (as categorized in [95], where the approaches for human activity recognition are categorized into single-layered approaches and hierarchical approaches) for DTW is proposed in [16], and is used to attack speed variation. The advantage of DTW is that it is fast and easy, but it might need extensive templates for various situations, resulting in high computation cost to match with these extensive templates.

Figure 11. A hidden Markov Model (HMM) inference graph [6].



4.2. Generative Models

One of the most popular generative models is the hidden Markov model (HMM) [75]. As stated in Rabiner and Juang [75], the classical tutorial to HMMs, an HMM is defined as a doubly stochastic process with an underlying hidden stochastic process and an observed stochastic process which can produce the sequence of observed symbols. The underlying hidden stochastic process is a first-order Markov process; that is, each hidden state depends only on the previous hidden state. Moreover, in the observed stochastic process, each observed measurement (symbol) depends only on the current hidden state. An HMM inference graph is shown in Figure 11 [6]. The circular nodes denote the hidden state variables, and the square nodes represent the observed variables. On the other hand, a dynamic Bayesian network (DBN) [78] relaxing the assumptions made by HMMs by allowing the state space in

a factored form, rather than a single-layer stochastic process. A DBN provides a more general system model but with higher complexity and computation cost.

4.2.1. Hidden Markov Model (HMM)

An HMM is specified by three terms [75], $\Phi = (\pi,\ A,\ B)$. The first term (π) is the initial probability of hidden states. The second term (A) is the transition matrix, which specifies a transition probability from one hidden state to another hidden state. The third term (B) is the observation matrix, which specifies the probability of the observed symbol given a hidden state. As addressed in [76], three types of problems of HMMs are addressed.

- *The evaluation problem*: Given a model $Y = \Phi$ and an observation sequence X, what is the probability (likelihood), $P(X \mid \Phi)$, of X given the specified model Φ? This problem can be efficiently solved by the forward/backward algorithm.
- *The decoding problem*: Given a model Φ and an observation sequence X, what is the most likely underlying hidden state sequence that produces the observation sequence? This problem can be efficiently solved by the Viterbi algorithm.
- *The learning problem*: Given a model Φ and an observation sequence X, how can we adjust the model parameters $\Phi = (\pi,\ A,\ B)$ to maximize the conditional probability (likelihood) $\prod_X P(X \mid \Phi)$? This problem can be efficiently solved by the Baum-Welch re-estimation algorithm.

HMMs have been popularly used to model time-sequential data such as speech and video. Yamato *et al.* [4] train HMMs to recognize actions of different tennis strokes by the Baum-Welch re-estimation algorithm. Besides the conventional HMMs, many variations of HMMs are proposed to deal with different scenarios of the problems. Brand *et al.* [6] propose a coupled hidden Markov model (CHMM) to model the coupling probabilities of two-handed interacting actions, such as single whip, cobra and brush knee. The state transition graph of traditional left-to-right HMM and CHMM are shown in Figure 12 [6]. The main difference between HMMs and CHMMs is that an HMM has only one hidden layer of Markov process, while a CHMM has two hidden layers of Markov processes which couple with each other.

Besides, to effectively model the temporal causality of human activity, the HMMs with the left-to-right state-transition structure are commonly used. However, many human actions exhibit the quasi-period cycles of body movements, which cannot be easily modeled by the standard left-to-right HMMs. To overcome this drawback, Thuc *et al.* [77] proposed a cyclic HMM to effectively adapt to most quasi-periodic human activity recognition tasks. The cyclic HMM is a left-to-right HMM model with a return transition from the ending state to the beginning state as shown in Figure 13 [77].

Figure 12. The state transition graph for left-to-right HMM (the upper graph) and coupled hidden Markov model (CHMM) (the lower graph) [6].

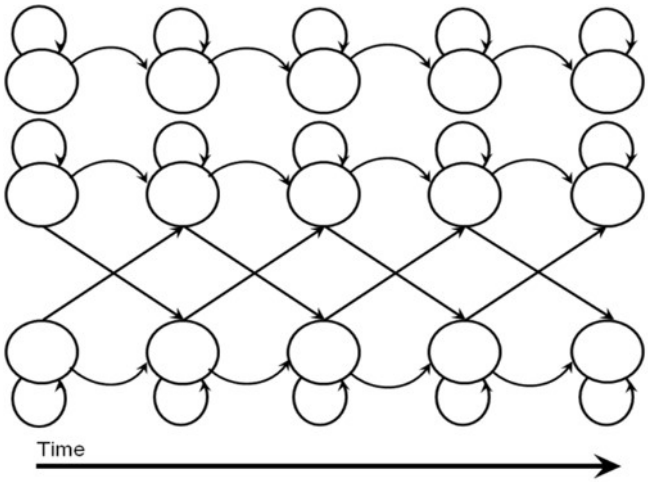
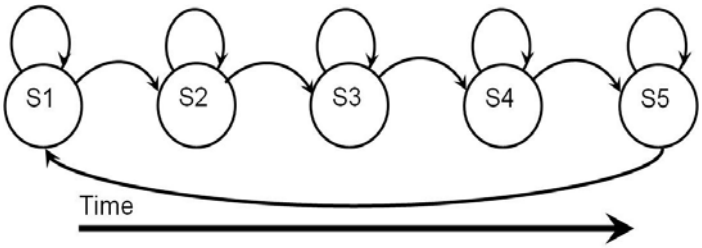


Figure 13. A cyclic HMM with the left-to-right state transition structure [77].



Moreover, Duong *et al.* [1] introduce the Switching Hidden Semi-Markov Model (S-HSMM) to recognize human daily life activities and detect abnormality. An HMM has an implicit duration distribution which is geometrically distributed, while an HSMM is a version of HMM with explicit duration modeling with any specific distribution. Further, an S-HSMM is a two-layered extension of the hidden semi-Markov model (HSMM) with the bottom layer represents atomic activities and their duration, and the top layer as a sequence of high-level activities, where each high-level activity is made of a sequence of atomic activities with specified durations. As some examples of atomatic activities, people spend different amounts of time at the cupboard, stove, fridge, or moving between kitchen and living room. With the help of high-layer HSMM, some of these atomic actions can be combined to form high-level composite activities including making/eating breakfast, making coffee and washing dishes.

Furthermore, Natarajan and Nevatia [17] introduce a Hierarchical Variable Transition Hidden Markov Model (HVT-HMM) to simultaneously track and recognize articulated full-body human motion. As proposed in [17], HVT-HMM is a three-layered extension of the Variable Transition Hidden Markov Model (VTHMM), where a single Markov chain is used to denote the composite actions at the top layer, a VTHMM with dynamically changing transition probability is applied to the action units at the middle layer, and a conventional HMM is used to represent the body pose transitions at the bottom layer. The graph structure of a HVT-HMM is shown in Figure 14 [17].

Figure 14. The graphical structure of a HVT-HMM [17].

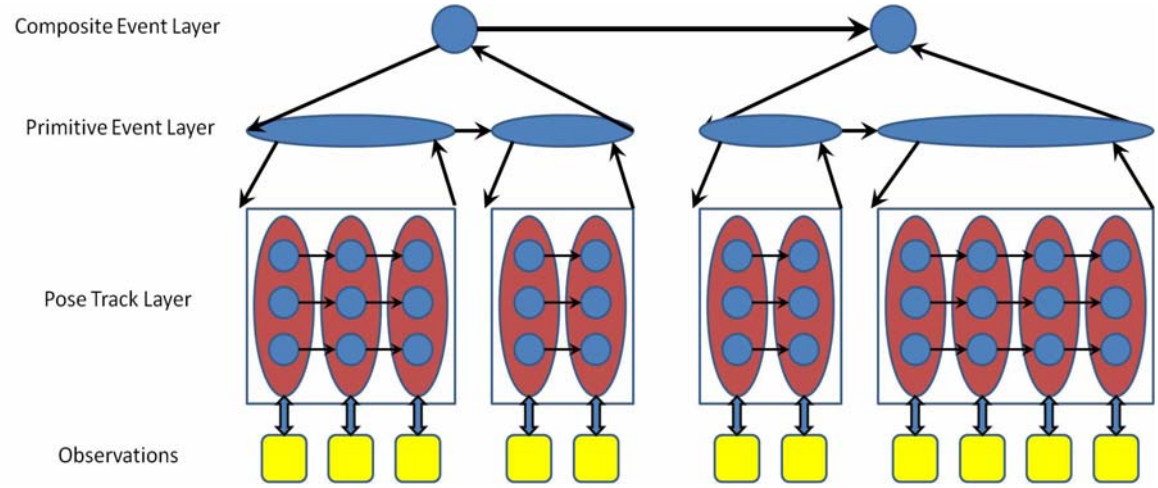
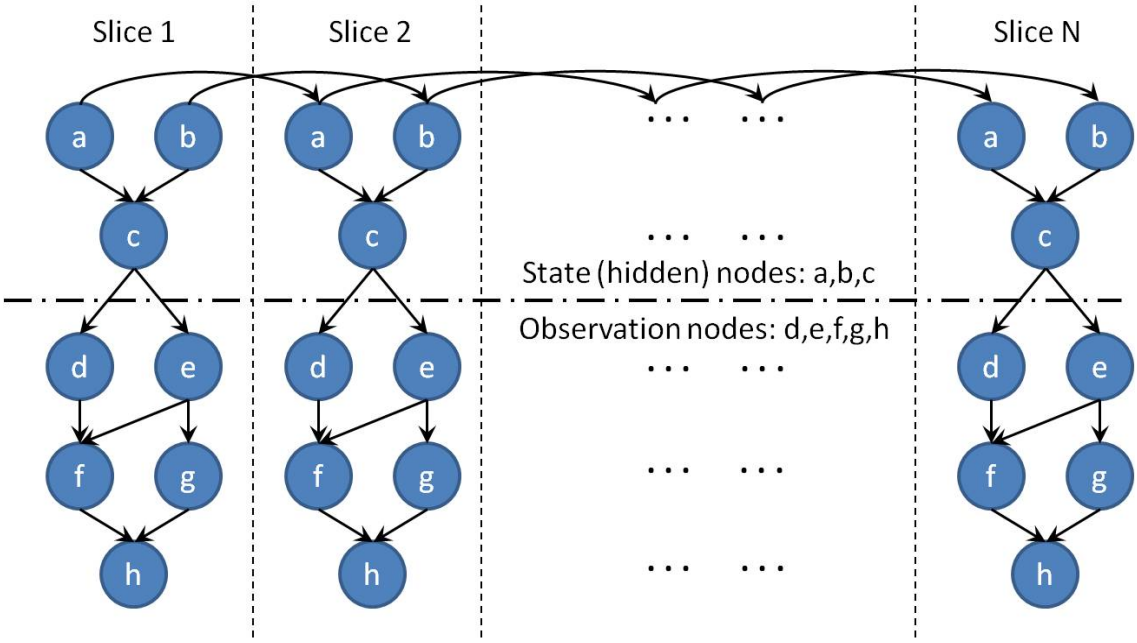


Figure 15. An example of a dynamic Bayesian network (DBN) is unrolled in time axis with state (hidden) notes and observation nodes [7].



4.2.2. Dynamic Bayesian Network (DBN)

A DBN [78] is a Bayesian network with the same structure unrolled in the time axis as shown in Figure 15 [7]. An HMM is proven to be a special type of DBN with a fixed structure of inference graph. Luo *et al.* [7] use DBNs to characterize the spatio-temporal nature of semantic video objects in several sport events, including downhill skiing, golf swing, baseball pitching, bowling, and ski jump. Moreover, a variation of DBN, named Coupled Hierarchical Duration-State DBN (CHDS-DBN), is proposed by Du *et al.* [9] to model human interactions as a multiple stochastic process. The CHDS-DBN represents two scales of human motions in video sequences, more specifically, the global activity state scale and the local activity state scale. The CHDS-DBN has some advantages. First, multiple scales of motions can be represented in a CHDS-DBN to consider large and small scales at

the same time. Secondly, a CHDS-DBN can capture the important motion details of human interaction. Thirdly, based on motion decomposition, a CHDS-DBN can have a low-dimensional feature space and a small size of state space. However, one main drawback of DBN is the inevitable need of very long training time.

4.3. Discriminative Models

Discriminative models can learn the conditional (posterior) probability distribution $P(Y \mid X)$, of a specific class label *Y* given the observed variable *X*. Compared with generative models, discriminative models cannot be used to generate samples with the joint probability $P(X,Y)$, since the original varying-length data sequence needs to be transformed to a fixed length feature data for discriminative model based classifiers. But for some classification tasks where the categorization of the data is the only purpose instead of knowing the additional hidden state transition of the data sequences, joint probability may not be required. The examples of discriminative models are described below, including support vector machines (SVMs) [18,19,79,80], relevance vector machines (RVMs) [81,82] and artificial neural networks (ANNs) [84].

4.3.1. Support Vector Machine (SVM)

The SVM [18,79,80] is one of the most popular margin-based supervised classifier in the pattern recognition. It is used to separate a data set into two classes. The goal of designing an SVM is to find the optimal dichotomic hyperplane which can maximize the margin (the largest separation) of two classes. The data points on the margin of the hyperplane are called support vectors. Schuldt *et al.* [18] apply SVMs to recognize human activities by extracting local space-time features in a video. Moreover, Laptev *et al.* [19] use a nonlinear SVM with a multi-dimensional Gaussian kernel for recognition of various natural human activities, including AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp and StandUp by building spatial-temporal bag-of-features (space-time grids). The main drawback of an SVM is its higher computation burden for the constrained optimization programming used in the learning phase.

4.3.2. Relevance Vector Machine (RVM)

Tipping [81,82] proposes an RVM, which is an SVM-like probabilistic model with fewer kernel functions needed. Compared with an SVM, the RVM can relax the limitation of the SVM such as point prediction by providing a full predictive distribution and still retain the sparseness property of the SVM [82]. Agarwal and Triggs [54] further apply RVMs to shape context descriptor features for 3D pose estimation. The main advantage of RVMs is that they allow sparse sets of highly relevant features or training examples to be selected for the training of the separation functions. But the main disadvantage is still the high computation cost of learning an RVM.

4.3.3. Artificial Neural Network (ANN)

An ANN [84] is a mathematical model to describe the problems in a network of directed graphs, whose nodes are represented as artificial neurons and the weighted directed edges in the graphs are

connections between artificial neurons. As mentioned in [84], an ANN enables to learn complex nonlinear input-output relationships with a universal approximation (a black box) model. With a systematically sequential training procedure, an ANN can train the model to adapt to the input data. ANNs use the concept of biological neuron networks to supply nonlinear capability via more discriminative feature transformation through hidden layers, resulting in more effective classification through multilayer perceptrons (MLPs), which apply the back propagation (BP) supervised learning algorithms to compute suitable connection weights and biases of the network. Fiaz and Ijaz [83] propose a method to design an ANN-based intelligent human activity monitoring system to detect and track suspicious activity in a surveillance environment. A three-layer perceptron is used for the classification of the human activities from the information of the distance vectors and the motion vectors for each frame in a video sequence. Moreover, Foroughi *et al.* [29] apply four-layer MLPs to learn eigen-motions, which are the movement patterns extracted by an eigenspace technique, for motion classification and falling detection. Although ANNs have the ability to implicitly describe complex nonlinear relationships between dependent and independent variables, they have higher computation burden during the learning, and are prone to over-fitting of the data.

4.4. Others

Among classification algorithms for human activities, in addition to DTW, generative models such as HMMs and DBNs, and discriminative models such as SVMs, RVMs and ANNs, there are also many other potential techniques can be adopted, such as Kalman filters [10,86], binary trees [20,87], multidimensional indexing [21], and k nearest neighbors [22].

4.4.1. Kalman Filter (KF)

A KF [86] is a set of mathematical equations to minimize the mean square errors (MSEs) and estimate the state of the dynamic process. Two phases in a Kalman filter are performed recursively, i.e., the estimation (prediction) of the process state and the update of the process state by the measurement [86]. More specifically, firstly, time update (predictor) phase equations are to predict the current state of the dynamic process and error covariance, and to estimate *a priori* for the next time index. Secondly, the measurement update (corrector) phase equations are to update the state of the process by the measurement of the estimate of *a priori*. These two phases (prediction and correction) are run recursively to predict and correct the current estimate on all of the past measurements. Bodor *et al.* [10] use a KF for tracking pedestrians through the development of a position and velocity path characteristics for each pedestrian. The main limitation of Kalman filter is that it needs good foreground segmentation; hence, it has little ability to handle the occlusion.

4.4.2. Binary Tree

A binary tree is a tree structure with a maximum of two children for each internal node. The binary tree structure can be applied in classification problems, called classification trees. Stauffer and Grimson [87] build a hierarchical classification binary tree by the accumulated joint co-occurrence statistics of the representations in a video sequence. The classification trees take the whole set of

prototypes and the co-occurrence matrix to determine two distributions, or called probability mass functions (pmfs), through the prototypes of codebook that best represent the co-occurrence matrix. Moreover, Ribeiro and Santos-Victor [20] investigate the use of hierarchical binary tree classifiers on human activity recognition. For each node of the binary tree classifier, a Bayesian classifier is used and the likelihood functions are modeled and systematically learned as Gaussian mixtures. Binary tree classification is simple and fast, but the separation rules in each node are difficult to be general for other cases, making it difficult for complex scenarios.

4.4.3. Multidimensional Indexing

Ben-Arie *et al.* [21] develop a multi-dimensional indexing method for view-based recognition of human activity from video sequences. More specifically, the human poses and velocity of hands, legs and torso are used to represent an activity and stored in a set of multidimensional hash tables. The activity is then recognized by indexing and sequencing a few pose vectors in the multidimensional harsh table. To be invariant to the speed of the activity, a sequenced-based voting approach is designed for the activity recognition. The computation of the indexing approach is comparatively low. The experimental results in [21] show that the multidimensional indexing method is invariant to viewpoint variations to the extent of ± 30 degrees in azimuth.

4.4.4. K-Nearest Neighbor (K-NN)

The K-nearest neighbor (K-NN) [22] algorithm is a classification method based on the K, a predefined constant, closest training data in the feature space. A point/vector is classified to one label, which is the most frequent label among K nearest training points/vectors. Because K-NN classification decision is based on K neighborhood points/vectors, therefore K-NN can be easily used in multi-modal classification tasks. As described in [22], there are some advantages for K-NN. First, K-NN is a simple model with few parameters. Secondly, the computation time for testing phase is independent of the number of classes. Thirdly, K-NN is robust in the search space even for nonlinearly separable data. Kumari and Mitra [22] use DFT of the small image blocks as feature selection, and apply K-NN as the classifier for human activity recognition. The main drawback for K-NN is that the classification performance is sensitive to the selection of K. Different K values can be evaluated and validated during the training phase to decide the best K before performing classification.

5. Human Activity Recognition Systems

The low-level core technology can be effectively integrated into human activity recognition systems, which includes single person activity recognition, multiple people interaction and crowd behavior, and abnormal activity recognition.

5.1. Single Person Activity Recognition

5.1.1. Trajectory

A trajectory is the path that a person moves as a function of time. The trajectory of a tracked person in a scene is often used to analyze the activity or behavior of the tracked person. Lu and Little [5] use a PCA-HOG descriptor to track and recognize sports videos, such as hockey and soccer. When the player runs with a ball, the trajectory is used to analyze the player's run-left, run-right, runs in, or run-out. Moreover, the loitering behavior can be easily inferred by analyzing trajectories. Bodor *et al.* [10] use Kalman filters to analyze pedestrian location and velocity, which are used to classify either a walking pedestrian, a running pedestrian, a loitering pedestrian or a falling-down pedestrian. Bird *et al.* [112] use an appearance-based method to detect loitering pedestrians by adopting a linear discriminative method based on the clothing color. The time stamps for each pedestrian are used to judge the present duration for a pedestrian, and a strip-shifting algorithm is designed to detect the loitering cases. Furthermore, the stalking behavior is another popular topic by tracking the trajectories. Niu *et al.* [113] propose a framework to detect the stalking behavior for surveillance environment based on simple motion detection algorithms such as frame differencing and feature correlation. SVM with the Gaussian kernel function is then used to recognize three behaviors including following, following-and-gaining and stalking behavior. The following behavior is characterized by an almost constant relative position and a nearly zero relative velocity; the following-and-gaining behavior is characterized by a linearly-shrinking change in the relative position and a nearly constant, but non-zero relative velocity; the stalking behavior is similar to the following-and-gaining behavior, but with a much larger variance in both relative position and velocity.

5.1.2. Falling Detection

Another popular topic of single person activity recognition is falling detection. Falling detection is significantly critical for security and safety environments, especially for the elderly who live alone. Töreyin *et al.* [114] model human motions with HMMs, and fuse audio channel data with the results of HMMs to detect a falling event. The audio information is essential to distinguish a falling person from a person simply sitting down or sitting on the floor. Moreover, Shieh and Huang [115] propose a human-shape-based algorithm to extract the one-pixel-wide edge features, and then a string matching algorithm [116,117] is applied for falling detection. But the building of the templates for falling posture might be costly, due to the various postures of falling. Furthermore, Sengto and Leauhatong [118] propose a falling detection algorithm by using a back-propagation neural network (BPNN) and a tri-axial accelerometer. Four daily activities including walking, jumping, getting into bed and rising from the bed, and four falling actions including front fall, back fall, left fall and right fall, are recognized by the BPNN. However, the limitation of the accelerometer is needing to wear it, as it is cumbersome and easy to forget to put on. Foroughi *et al.* [26–29] conduct some methods for falling detection based on human shape information and multi-class SVMs, integrated with time motion images and BPNNs. Without the additional tri-axial accelerometer, falling detection can be made more natural and smooth to incorporate into daily life.

5.1.3. Human Pose Estimation

Human pose estimation is also a popular topic in the computer vision community. Based on the results of the human poses, human activity can be efficiently recognized. Huo *et al.* [15] propose a system for human pose recognition. First, a 2D model is used for torso detection and tracking, and a skin color model is used for hands tracking. 3D reconstruction is done by multi-view from synchronized multiple cameras. The 2D and 3D coordinates are then converted into a normalized feature space, and classified by nearest mean classifiers (NMC) for recognizing key poses. However, it might be an enormous task to accumulate all predefined key poses, because human poses vary tremendously. Lee and Nevatia [66,67] propose a three-stage method for 3D human pose estimation including foreground blobs tracking, 2D joints and body-parts tracking, and 3D pose estimation by using data-driven Markov chain Monte Carlo (DD-MCMC). The sneaking-in and meeting-room scenes are analyzed with 3D human poses. Nevertheless, the computation cost is very high. Ke *et al.* [70] propose another method for 3D pose estimation with much lower time complexity. The shape, color and time continuity information are jointly considered to track 2D body parts, and the 3D pose estimation is achieved by the downhill simplex algorithm based on an analysis-by-synthesis methodology. Thuc *et al.* [73] further convert the estimated 3D poses into a feature space and the human actions can be recognized by using HMMs.

5.2. Multiple People Interaction and Crowd Behavior

Multiple people interaction and the crowd behavior have drawn much attention recently due to the needs of environment security. Jacques Junior *et al.* [119] reviewed crowd analysis literature and tackled three important issues including people counting, people tracking, and crowd behavior understanding.

With regard to people counting, Subburaman *et al.* [120] investigate the count estimate of people in a crowded scene by detecting the head region, based on the state-of-art cascade of boosted integral features. The PETS 2012 and Turin metro station dataset are used to evaluate the performance of the head counting system. Merad *et al.* [121] propose a fast people counting method by using skeleton graph. The skeleton silhouette is decomposed into the head, torso and limbs, and the 3D head pose is estimated by finding the rigid transformation, minimizing the sum of square errors with an Orthogonal Iteration (OI) algorithm [122]. Head counting is achieved by detecting the heads of people in a scene in order to count the number of pedestrians passing an indoor area.

Moreover, in the aspect of people tracking, McKenna *et al.* [123] propose a method to track groups of people by an adaptive background subtraction method combining color and gradient information to remove shadows and unreliable color cues. The case of a group splitting up into several groups, and the case of several groups merging into one group are dealt with by color histogram information. Some interactions with objects can also be detected. If a person removes or deposits an object, a new region will be split from the person. However, only groups of people are tracked, instead of individual persons. Chu *et al.* [124] propose a method for human tracking by adaptive Kalman filtering and multiple kernels tracking with projected gradients. The multiple kernels tracking is applied to deal with the occlusion cases. A person is modeled in two or four kernels, e.g., the upper body-part kernel and

the lower body-part kernel for two-kernel cases. The CAVIAR, PETS 2010, i-LIDS, and Crowd Analysis datasets are used to evaluate the performance of the people tracker.

Furthermore, with regard to crowd behavior understanding, Saxena *et al.* [125] propose to model crowd events for specific end-user scenarios by the use of an extended Scenario Recognition Engine (SRE) [126] based on a Kanade-Lucas-Thomasi (KLT) tracker for multiple-frame feature point detection and tracking. The speed, direction and location of a crowd can be estimated in a scene and used to build event models. In the fighting event, the normal behaviors and abnormal behaviors are detected, including crowd moving forward (normal), opposite movement in a crowd (abnormal), crowd stopped (normal), and strong lateral crossing to right/left (robbery by crowd, abnormal). Szczodrak *et al.* [127] use optical flow combined with artificial neural network (ANN) to evaluate the types of crowd behavior. The crowd exit path observation is used to detect the people flow interruption caused by slowdown or stopping, based on the measurement of speed of the crowd flow. The normal behavior (walking), diverging, and abnormal behavior are classified by ANN. Besides, Cho and Kang *et al.* [128] propose a method to detect abnormal crowd behavior using integrated multiple behavior models and optical flow. The method can reflect not only social properties [129] but also personal properties such as speed and direction; therefore, it can effectively detect abnormal behavior in a crowd scene.

5.3. Abnormal Activity Recognition

It is very difficult to explicitly define what abnormal or unusual activities are, since their definition depends on the contexts and surrounding environments. Moreover, by definition, they are not frequently observed; otherwise they are normal or usual. Thus, rather than building an individual abnormal activity model, a deviation approach has been commonly adopted. It builds a normal model as in background subtraction using given examples or previously seen data, and considers new observation as abnormal or unusual if they deviate too much from the trained model.

Boiman and Irani [130] assume that spatio-temporal patches of regular events can be accurately constructed by the combination of large continuous patches from previously seen data or their spatial neighborhoods. If not, they are classified as regions where irregular events occur. Kim and Grauman [131] employ histograms of optical flow as primary features and build a normal model using the bag of visual words approach and Markov Random Filed (MRF). They also provide an efficient updating method so that their system can adapt to environmental changes over time. Mahadevan *et al*. [132] construct probabilistic dynamic model using Mixtures of Dynamic Textures (MDT) [133]. They successfully incorporate appearance and motion information of crowded scenes in a unified way, whereas most other works utilize motion information only. Adam *et al* [134] focus on real time systems and provide an approximated optical flow technique, which is simple and efficient to estimate motion direction and speed. However, they cannot detect abnormal behaviors which consist of sequences of normal actions, such as loitering. Kratz and Nishino [135] propose the distributions of spatio-temporal gradients as motion features. They successfully extract motion patterns from very crowded scenes, where traditional optical flows often fail. Then, they use HMMs to capture spatial and temporal relationships between motion patterns.

6. Applications

The goal of the last stage, i.e., applications, is to analyze classified activities so that their semantic meaning can be understood in specific domains. Activities can be simple actions such as walking, waving; complex single-person actions, such as ballet dancing, doing aerobics; the interactions between persons, such as hand shaking, hugging; or the interactions between humans and objects, such as preparing a meal, kicking a car door. Activity understanding requires expert knowledge to characterize the uniqueness accurately and to build the scenario suitable to each specific domain of applications. This makes the activity recognition techniques more valuable and widely used in diversified applications of our daily lives. In this section, we focus on three dominant applications, including surveillance environments, entertainment environments and healthcare systems.

6.1. Surveillance Environments

The application of human activity recognition in surveillance systems mainly focus on automatically tracking individuals and crowds, so as to support security personnel to observe and understand activities, resulting in recognition of the criminal and detecting suspicious activities.

Most security surveillance systems are equipped with several cameras and require laborious human monitoring on screens for video content understanding. By applying automatic human activity recognition techniques to video-based surveillance systems, we can effectively reduce the workload of security staff as well as systematically creating an alert immediately when security events are detected in order to prevent potentially dangerous situations. Some typical scenarios are as follows.

People detecting and tracking is one of the first objectives of a security surveillance system. Nakazawa *et al.* [61] propose a wide area human tracking method using network-connected vision systems that each consists of a camera and an image processing module. Each vision system carries out two tasks, the tracking task and the acquisition task. If there are humans in a visible region of a vision system, the tracking task system tracks the human position and broadcasts the results to the other systems. In case the vision system has no person in view, then the acquisition task system must find the person in their image to be tracked. Bodor *et al.* [10] use Kalman filters to track the position and velocity path for each pedestrian in high pedestrian traffic areas, then the tracking results are further exploited to detect suspicious behaviors, such as entering a "secured area," running or moving erratically, loitering or moving against traffic, or dropping a bag or other items. Similarly, Fiaz and Ijaz [83] also perform suspicious activity detection and tracking using ANNs. Besides this, some researchers perform detection of various kinds of violent behaviors such as fighting, punching, stalking, *etc*. [16,20,113,136,137].

Loitering is a suspicious behavior which attracts many researchers, due to the fact that loitering often leads to abnormal situations, such as suspected drug-dealing activity, bank robbery, and pickpocketing, *etc.* [138]. Bird *et al.* [112] present a vision-based method to detect individuals loitering at inner-city bus stops. Using a stationary camera to monitor a bus stop, the system takes snapshots of individuals and then uses them to classify the individuals using an appearance-based method. The features used to correlate individuals are only based on clothing colors. To determine if a given individual is loitering at the same bus stop, instead of using human tracking technique from videos,

time stamps collected with the snapshots in their corresponding database class can be used to judge how long an individual has been present. Zin *et al.* [138] detect loitering by using a 2D Markov random walk model, including both motion and appearance features. The model is also made less sensitive to obstructions in the region of interest from irrelevant objects, by incorporating an occlusion effect on the observation probabilities of the Markov random walk model.

In video-based surveillance systems, gait-based person authentication is an attractive application, because it can be performed in distance and surreptitiously, whereas the other biometric methods would require physical touch such as fingerprint or close distance such as face recognition. The gait recognition method by Ran *et al.* [139] is based on decomposing a video sequence into *x–t* slices, which are periodic patterns of double helical signature (DHS). From video sequences, DHS is extracted by applying an iterative local curve embedding algorithm to segment and label the body parts in cluttered scenes so as to detect variations in the gait due to carrying load. In [140], Hu *et al.* propose a novel incremental framework based on the local binary pattern (LBP) feature, which is used to describe the texture information of optical flow. This flow-based method is more robust to noise and greatly improves the usability of gait traits in video surveillance applications when compared to other dominant silhouette-based approaches.

Beside the main goal of video-based surveillance systems in detecting criminals and suspicious security events, they also aim to ensure safety of swimmers in pools. For example, Poseidon [141] is a commercial system developed for drowning detection. The system, equipped with a network of cameras mounted either above or below the surface of the water, can help lifeguards to systematically monitor swimmers' trajectories and can alert them in seconds to a swimmer in trouble. This drowning detection system can increase the chance of saving a life and reduce the likelihood that a person will suffer long-term damage as a result of a drowning incident [141].

In addition, video-based human activity recognition systems can also be applied in marketing analysis, such as detecting customers' interest and interactions with various products while shopping, based on existing methods to detect motion and track the detected objects [142].

6.2. Entertainment Environments

Human activity recognition can also be used to recognize entertainment activities, such as sport [3–5,7], dance [2,51] and gaming [10,15,39], in order to enrich lifestyles. Various video-based entertainment activity recognition systems have been proposed.

For the purpose of recognizing sportive activities, Yamato *et al.* [4] might well be pioneers in applying HMMs to recognize the time-sequential images of tennis scenes, including six tennis strokes such as forehand stroke, backhand stroke, forehand volley, backhand volley, smash and serving. In [7], Luo *et al.* developed an object-based method for video analysis and interpretation of sports video sequences. The sport behaviors are effectively recognized by using DBNs, which can generate a hierarchical description for video events, including bowling, downhill skiing, golf swing, pitching, and ski jumps recorded from real scenarios, with cluttered background and moving cameras. Ke *et al.* [3] exploit the use of volumetric features for the recognition of actions such as serve, run right and return serve actions in the tennis sequences when combined with flow-based correlation techniques. By using an extended behavior-based similarity measure, Shechtman and Irani [51] were successful in detecting

dives into a pool during a swimming relay match. Despite the numerous simultaneous activities and despite the severe noise, this method is able to separate most dives from other activities.

The 3D shapes induced by the 2D silhouettes in the space-time volume are used by Blank *et al.* [2] for action detection in a ballet movie. The method utilizes properties of the solution to the Poisson equation to extract space-time features, such as local space-time salience, action dynamics, shape structure and orientation. The effectiveness and robustness of the method is demonstrated via an example of finding all the locations in the movie where an action called "cabriole" (beating feet together at an angle in the air) is performed by either a male or a female dancer. Besides the ability of detecting diving in a pool, the method developed by Shechtman and Irani [51] is also successful in detecting the single turn of a dancer in a very fast moving ballet video sequence.

One of the most popular leisure activities is playing video games. A number of methods are developed for this purpose [10,15,39]. Richard *et al.* [39] developed Pfinder as a real-time system for tracking people and interpreting their actions by using a multi-class statistical model of color and shape to obtain the head, hands and feet positions in different viewing conditions. Pfinder has been successfully used in several different human interface applications, e.g., to navigate a 3D virtual game environment or to place the player at a particular place in a virtual room, which is populated by virtual occupants from real-time 3D computer graphics based on live video. In [15], Huo *et al.* present a method for human motion capture and pose recognition. The human torso and the hands are segmented from the whole body and tracked over time. A 2D model is used for the torso detection and tracking, while a skin color model is utilized for the hands tracking. Moreover, 3D location of these body parts are calculated and further used for pose recognition. The implementation of the proposed approach is simple, easy to realize, and suitable for real gaming applications. Ke *et al.* [69] also perform a similar 3D human pose estimation task, but they only use one camera rather than multiple cameras as in [15]. Moreover, their system can estimate 3D poses not only the upper body part as in [15] but also the lower body part including knees and feet. A simple video game is implemented, i.e., a 3D avatar, generated by real-time, based on the derived 3D poses of the proposed system, to hit balls so as to get scores or to avoid attacks from flying balls.

6.3. Healthcare Systems

The applications for activity recognition in healthcare systems analyze and understand patients' activities, so as to facilitate health workers to diagnose, treat and care for patients, resulting in improving the reliability of diagnosis, decreasing the working load for the medical personnel, shortening the hospital stay for patients, and improving patients' quality of life, *etc*. [1,23–38].

6.3.1. Daily Life Activity Monitoring

Daily life activity monitoring mainly focuses on learning and recognizing the daily life activities of seniors at home. The proposed systems are to provide seniors an opportunity to live safely, independently and comfortably. In order to accomplish this, most proposed systems continuously capture the movements of individual senior or multiple seniors at home, automatically recognizing their activities, and detecting gradual changes in baseline activities such as mobility functional

disabilities, mental problems, as well as the urgent warning signs of abnormal activities such as falling down or having a stroke. Some of these scenarios can be summarized as follows.

Respiration behavior can be critical in diagnosing a patient's illness or recognizing distress during sleep. Many diseases, such as obstructive sleep-apnea syndrome, cardiovascular disease, and stroke, can induce abnormal respiration. Automated respiration monitoring is performed by Kuo *et al.* [23], where near-IR images are captured to measure the sleeper's respiration based on the periodic rising and falling motions of their chest or abdomen.

Gao *et al.* [24] measure feeding difficulties of nursing home residents with severe dementia, by automatically measuring the number of hand movements to the mouth using motion feature vectors and an HMM to identify the start and end of individual dining events.

Huynh *et al.* [25] present a video monitoring method for detecting and tracking face, mouth, hands and medication bottles in the context of medication intake. This aims to monitor medicine intake behavior of elderly at home to avoid the inappropriate use of medicine.

Falling is a major health risk of elderly as it is known to be the leading cause of injury and deaths among seniors. Foroughi *et al.* [26–29] conduct some methods to detect the fall, e.g., based on human shape variation. Extracted features, including combination of best-fit approximated ellipse around the human body, projection histograms of the segmented silhouette and temporal changes of head pose, are fed to a multi-class SVM [26] or an *MLP* ANN [27] for reliable classification of motions and determination of a fall event. Other features, also widely used for fall detection, are based on the combination of integrated time motion images (ITMI) and Eigen space technique [28,29].

In order to recognize the activities at a higher semantic level, the activity duration, the position of human, the interaction between people and person-object are the essential elements to be analyzed.

For the activity duration, Luhr *et al.* [30] use the explicit state duration HMM (ESD-HMM), in which a duration variable is introduced in a standard HMM. Duong *et al.* [1] introduce the switching hidden semi-Markov model (S-HSMM), which implicitly exploits the benefit of both the inherent hierarchical organization of the activities and their typical duration. In [31], similar to [1], Duong *et al.* further explicitly add the time duration information to a standard HMM, called hidden semi-Markov model (HSMM), to model the state duration by using the generic exponential family.

For the human position, in [31], the door, the stove, the fridge, the sink, the cupboard, and the table areas are used to identify the activities in a kitchen room. For example, meal preparation and consumption consists of twelve steps: *take-food-from-fridge* → *bring-food-to-stove* → *wash-vegetable* → *come-back-to-stove-for-cooking* → *take-plates/cup-from-cupboard* → *return-to-stove-for-food* → *bring-food-to-table* → *take-drink-from-fridge* → *have-meal-at-table* → *clean-stove* → *wash-dishes-at-sink* → *leave-the-kitchen*. In [1], the kitchen is quantized into 28 square cells of 1 m^2 each and the position of the human is captured by four cameras mounted at the ceiling corners, and the tracking system returns the list of cells visited by the person as the moving trajectory path.

For the interaction, Liu *et al.* [32] propose an interaction-embedded hidden Markov model (IE-HMM) framework, for detecting and classifying individual human activities and group interactions in a nursing home environment.

6.3.2. Rehabilitation Applications

Traditional rehabilitation systems often require patients to undergo several clinical visits for the physical therapy exercises and the scheduled evaluation until his/her full recovery of mobility function for daily activities. Such clinical visits can be avoided by using innovative rehabilitation systems, which are home-centered and self-health care with the help of video-based activity recognition techniques. Moreover, by continuously monitoring the daily activities and gaits, the early symptoms of some diseases can be timely detected so that the diagnosis and interventions are more appropriate. Some of these scenarios can be summarized as follows.

Stroke is a major cause of disability and health care expenditure around the world. Ghali *et al.* [33] design a system to provide real-time feedback to stroke patients performing daily activities necessary for independent living. More specifically, they envisage a situation in which a stroke patient stands in a standard kitchen and makes a cup of coffee in the usual way. The position and movement of the patient's hands and the objects he/she manipulates are captured by overhead cameras and monitored using histogram-based recognition methods. The key events (e.g., picking up a cup) are recognized and interpreted in the context of a model of the coffee-making task.

In order to objectively evaluate the improvement of motor functions of the elders at home, as well as to reduce burden on fitness instructors, Ryuichi *et al.* [34] propose a "multimedia fitness exercise progress notes" system, where the video capturing exercise movements of the elders are sent to an analysis center. Snapshots of the captured videos are used to semi-automatically measure many kinds of exercise parameters, such as lap time, distances and angles.

Goffredo *et al.* [35] propose the Gauss-Laguerre transform-based (GLT-based) motion estimation method in order to analyze the sit-to-stand (STS) motion from monocular videos. STS movement mainly involves hip and knee flexion-extension, and ankle plantar flexion-dorsiflexion is analyzed by utilizing a 2D human body model that considers the projections of body segments on the sagittal plane.

Besides sit-to-stand, walking gait is another human activity of great interest to many researchers due to the fact that the loss of ability to walk correctly can be caused by a serious health problem, such as pain, injury, paralysis, muscle damage, or even mental problems. Liao *et al.* [36] present a video-based system for analyzing four posture features of human walking, including body line, neck line, center of gravity (COG) and gait width based on the extracted silhouettes from front view and side view. Similarly, Leu *et al.* [37] also use two cameras and the feedback control structure at the segmentation level to extract the gait features, such as torso angle, left and right thigh angles, and left and right shank angles. With the aim of reducing the inconvenience of using many methods that require the captured images of human walking from either front or side view, Li *et al.* [38] propose a system with much less restrictions on walking direction. The system successfully extracts gait features, such as COG and pace length, from images obtained from two cameras with orthogonal views.

7. Conclusions and Future Direction

Although progress in recent video-based human activity recognition has been encouraging, there are still some apparent performance issues that make it challenging for real-world deployment. More specifically:

- The viewpoint issue remains the main challenge for human activity recognition. In real world activity recognition systems, the video sequences are usually observed from arbitrary camera viewpoints; therefore, the performance of systems needs to be invariant from different camera viewpoints. However, most recent algorithms are based on constrained viewpoints, such as the person needs to be in front-view (i.e., face a camera) or side-view. Some effective ways to solve this problem have been proposed, such as using multiple cameras to capture different view sequences then combining them as training data or a self-adaptive calibration and viewpoint determination algorithm can be used in advance. Sophisticated viewpoint invariant algorithms for monocular videos should be the ultimate objective to overcome these issues.
- Since most moving human segmentation algorithms are still based on background subtraction, which requires a reliable background model, a background model is needed that can be adaptively updated and can handle some moving background or dynamic cluttered background, as well as inconsistent lighting conditions. Learning how to effectively deal with the dynamic cluttered background as well as how to systematically understand the context (when, what, where, *etc.*), should enable better and more reliable segmentation of human objects. Another important challenge requiring research is how to handle occlusion, in terms of body–body part, human–human, human–objects, *etc.*
- Natural human appearance can change due to many factors such as walking surface conditions (e.g., hard/soft, level/stairs, *etc.*), clothing (e.g., long dress, short skirt, coat, hat, *etc.*), footgear (e.g., stockings, sandals, slippers, *etc.*), object carrying (e.g., handbag, backpack, briefcase, *etc.*) [143]. The change of human action appearance leads researchers to a new research direction, i.e., how to describe the activities that are less sensitive to appearance but still capture the most useful and unique characteristics of each action.
- Unlike speech recognition systems, where the features are more or less unified to be the mel-frequency cepstral coefficients (MFCCs) for HMM classifiers, there are still no clear winners on the features for human activity recognitions, nor the corresponding classifier designs. It can be expected that 3D viewpoint invariant modeling of human poses would be a good starting point for a unified effort.

Finally, human activity recognition tasks constitute the foundation of human behavior understanding, which requires additional contextual information such as W5+ (who, where, what, when, why, and how) [144]. The same activity may have different behavior interpretations depending on the context in which it is performed. More specifically, the “where” (place) context can provide the location information to be used to detect abnormal behaviors. For example, lying down on the bed or a sofa is interpreted as taking a rest or sleeping, but in inappropriate places such as the floor of the bathroom or kitchen, it can be interpreted as a fall or a sign of a stroke. Moreover, the “when” (time) context also plays another important contextual role for behavior understanding. For example, a person usually watching TV after midnight can be regarded as an insomniac. Another example is that a person will be detected as picking something up if he/she squats and stands up quickly. However, if he/she squats for a longer period, there might be a motion difficulty due to osteoarthritis or senility. Furthermore, the number of repetitions of an action can also be informative. For example, eating too many times or too little a day can be an early symptom of depression. The interaction between people

or between person and objects is also a good indicator to identify the meaning of the activity. For example, if a person is punching a punch-bag, he might be doing exercise. But if he is punching the wall, it can indicate anger or a mental disorder.

In conclusion, this review provides an extensive survey of existing research efforts on video-based human activity recognition systems, covering all critical modules of these systems such as object segmentation, feature extraction and representation, and activity detection and classification. Moreover, three application domains of video-based human activity recognition are reviewed, including surveillance, entertainment and healthcare. In spite of the great progress made on the subject, many challenges are raised herein together with the related technical issues that need to be resolved for real-world practical deployment. Furthermore, generating descriptive sentences from images [145] or videos is a further challenge, wherein objects, actions, activities, environment (scene) and context information are considered and integrated to generate descriptive sentences conveying key

Acknowledgments

This work was supported by the Industrial Strategic Technology Development Program, 10039149, funded by the Ministry of Knowledge Economy (MKE, Korea) and also supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0093828).

Conflict of Interest

The authors declare no conflict of interest.

References

1. Duong, T.V.; Bui, H.H.; Phung, D.Q.; Venkatesh, S. Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 838–845.
2. Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; Basri, R. Actions as Space-time Shapes. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV), Beijing, China, 17–21 October 2005; Volume 2, pp. 1395–1402.
3. Ke, Y.; Sukthankar, R.; Hebert, M. Spatio-temporal Shape and Flow Correlation for Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
4. Yamato, J.; Ohya, J.; Ishii, K. Recognizing Human Action in Time-sequential Images using Hidden Markov Model. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Champaign, IL, USA, 15–18 June 1992; pp. 379–385.
5. Lu, W.; Little, J.J. Simultaneous tracking and action recognition using the PCA-HOG descriptor. In Proceedings of the 3rd Canadian Conference on Computer and Robot Vision, Quebec, PQ, Canada, 7–9 June 2006; p. 6.

6. Brand, M.; Oliver, N.; Pentland, A. Coupled hidden Markov Models for Complex Action Recognition. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Juan, PR, USA, 17–19 June 1997; pp. 994–999.
7. Luo, Y.; Wu, T.; Hwang, J. Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks. *Comput. Vis. Image Underst.* **2003**, *92*, 196–216.
8. Lu, X.; Liu, Q.; Oe, S. Recognizing Non-rigid Human Actions using Joints Tracking in Space-Time. In Proceedings of the IEEE International Conference on Information Technology: Coding and Computing (ITCC), Las Vegas, NV, USA, 5–7 April 2004; Volume 1; pp. 620–624.
9. Du, Y.; Chen, F.; Xu, W. Human interaction representation and recognition through motion decomposition. *IEEE Signal Process. Lett.* **2007**, *14*, 952–955.
10. Bodor, R.; Jackson, B.; Papanikolopoulos, N. Vision-based Human Tracking and Activity Recognition. In Proceedings of the 11th Mediterranean Conference on Control and Automation, Rhodes, Greece, 18–20 June 2003; Volume 1, pp. 18–20.
11. Dollár, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior Recognition via Sparse Spatio-Temporal Features. In Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15–16 October 2005; pp. 65–72.
12. Scovanner, P.; Ali, S.; Shah, M. A 3-dimensional SIFT Descriptor and Its Application to Action Recognition. In Proceedings of the 15th International Conference on Multimedia, ACM, Augsburg, Germany, 23–28 September 2007; pp. 357–360.
13. Lin, C.; Hsu, F.; Lin, W. Recognizing human actions using NWFE-based histogram vectors. *EURASIP J. Adv. Signal Process.* **2010**, *2010*, 9.
14. Veeraraghavan, A.; Roy-Chowdhury, A.K.; Chellappa, R. Matching shape sequences in video with applications in human movement analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1896–1909.
15. Huo, F.; Hendriks, E.; Paclik, P.; Oomes, A.H.J. Markerless Human Motion Capture and Pose Recognition. In Proceedings of the 10th IEEE Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), London, UK, 6–8 May 2009; pp. 13–16.
16. Sempena, S.; Maulidevi, N.U.; Aryan, P.R. Human Action Recognition Using Dynamic Time Warping. In IEEE International Conference on Electrical Engineering and Informatics (ICEEI), Bandung, Indonesia, 17–19 July 2011; pp. 1–5.
17. Natarajan, P.; Nevatia, R. Online, Real-time Tracking and Recognition of Human Actions. In Proceedings of IEEE Workshop on Motion and Video Computing (WMVC), Copper Mountain, CO, USA, 8–9 January 2008; pp. 1–8.
18. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing Human Actions: A Local SVM Approach. In Proceedings of the 17th IEEE International Conference on Pattern Recognition (ICPR), Cambridge, UK, 23–26 August 2004; Volume 3, pp. 32–36.
19. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning Realistic Human Actions from Movies. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.

20. Ribeiro, P.C.; Santos-Victor, J. Human Activity Recognition from Video: Modeling, Feature Selection and Classification Architecture. In Proceedings of the International Workshop on Human Activity Recognition and Modelling (HAREM), Oxford, UK, 9 September 2005; Volume 1, pp. 61–70.
21. Ben-Arie, J.; Wang, Z.; Pandit, P.; Rajaram, S. Human activity recognition using multidimensional indexing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1091–1104.
22. Kumari, S.; Mitra, S.K. Human Action Recognition Using DFT. In Proceedings of the third IEEE National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Hubli, India, 15–17 December 2011; pp. 239–242.
23. Kuo, Y.; Lee, J.; Chung, P. A visual context-awareness-based sleeping-respiration measurement system. *IEEE Trans. Inf. Technol. Biomed.* **2010**, *14*, 255–265.
24. Gao, J.; Hauptmann, A.G.; Bharucha, A.; Wactlar, H.D. Dining Activity Analysis Using a Hidden Markov Model. In Proceedings of the 17th IEEE International Conference on Pattern Recognition (ICPR), Cambridge, UK, 23–26 August 2004; Volume 2, pp. 915–918.
25. Huynh, H.H.; Meunier, J.; Sequeira, J.; Daniel, M. Real time detection, tracking and recognition of medication intake. *World Acad. Sci. Eng. Technol.* **2009**, *60*, 280–287.
26. Foroughi, H.; Rezvanian, A.; Paziraee, A. Robust Fall Detection Using Human Shape and Multi-Class Support Vector Machine. In Proceedings of the IEEE Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP), Bhubaneswar, India, 16–19 December 2008; pp. 413–420.
27. Foroughi, H.; Aski, B.S.; Pourreza, H. Intelligent Video Surveillance for Monitoring Fall Detection of Elderly in Home Environments. In Proceedings of the IEEE 11th International Conference on Computer and Information Technology (ICCIT), Khulna, Bangladesh, 24–27 December 2008; pp. 219–224.
28. Foroughi, H.; Yazdi, H.S.; Pourreza, H.; Javidi, M. An Eigenspace-based Approach for Human Fall Detection Using Integrated Time Motion Image and Multi-class Support Vector Machine. In Proceedings of IEEE 4th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 28–30 August 2008; pp. 83–90.
29. Foroughi, H.; Naseri, A; Saberi, A.; Yazdi, H.S. An Eigenspace-based Approach for Human Fall Detection Using Integrated Time Motion Image and Neural Network. In Procedings of IEEE 9th International Conference on Signal Processing (ICSP), Beijing, China, 26–29 October 2008; pp. 1499–1503.
30. Lühr, S.; Venkatesh, S.; West, G.; Bui, H.H. Explicit state duration HMM for abnormality detection in sequences of human activity. *PRICAI 2004: Trends Artif. Intell.* **2004**, *3157*, 983–984.
31. Duong, T.V.; Phung, D.Q.; Bui, H.H.; Venkatesh, S. Human Behavior Recognition with Generic Exponential Family Duration Modeling in the Hidden Semi-Markov Model. In Proceedings of IEEE 18th International Conference on Pattern Recognition (ICPR), Hong Kong, China, 20–24 August 2006; Volume 3, pp. 202–207.
32. Liu, C.; Chung, P.; Chung, Y.; Thonnat, M. Understanding of human behaviors from videos in nursing care monitoring systems. *J. High Speed Netw.* **2007**, *16*, 91–103.

33. Ghali, A.; Cunningham, A.S.; Pridmore, T.P. Object and Event Recognition for Stroke Rehabilitation. In Proceedings of Visual Communications and Image Processing, Lugano, Switzerland, 8–11 July 2003; pp. 980–989.
34. Ayase, R.; Higashi, T.; Takayama, S.; Sagawa, S.; Ashida, N. A Method for Supporting At-home Fitness Exercise Guidance and At-home Nursing Care for the Elders, Video-based Simple Measurement System. In Proceedings of IEEE 10th International Conference on e-health Networking, Applications and Services (HealthCom), Singapore, 7–9 July 2008; pp. 182–186.
35. Goffredo, M.; Schmid, M.; Conforto, S.; Carli, M.; Neri, A.; D'Alessio, T. Markerless human motion analysis in Gauss–Laguerre transform domain: An application to sit-to-stand in young and elderly people. *IEEE Trans. Inf. Technol. Biomed.* **2009**, *13*, 207–216.
36. Liao, T.; Miaou, S.; Li, Y. A vision-based walking posture analysis system without markers. In IEEE 2nd International Conference on Signal Processing Systems (ICSPS), Dalian, China, 5–7 July 2010; Volume 3, pp. 254–258.
37. Leu, A.; Ristic-Durrant, D.; Graser, A. A Robust Markerless Vision-based Human Gait Analysis System. In Proceedings of 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, 19–21 May 2011; pp. 415–420.
38. Li, Y.; Miaou, S.; Hung, C.K.; Sese, J.T. A Gait Analysis System Using two Cameras with Orthogonal View. In Proceedings of IEEE International Conference on Multimedia Technology (ICMT), Hangzhou, China, 26–28 July 2011; pp. 2841–2844.
39. Wren, C.R.; Azarbayejani, A.; Darrell, T.; Pentland, A.P. Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 780–785.
40. Cucchiara, R.; Grana, C.; Piccardi, M.; Prati, A. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1337–1342.
41. Seki, M.; Fujiwara, H.; Sumi, K. A Robust Background Subtraction Method for Changing Background. In Proceedings of Fifth IEEE Workshop on Applications of Computer Vision, Palm Springs, CA, USA, 4–6 December 2000; pp. 207–213.
42. Permuter, H.; Francos, J.; Jermyn, I. A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recogn.* **2006**, *39*, 695–706.
43. Yoon, S.; Won, C.S.; Pyun, K.; Gray, R.M. Image Classification Using GMM with Context Information and with a Solution of Singular Covariance Problem. In IEEE Proceedings of Data Compression Conference (DCC), Snowbird, UT, USA, 25–27 March 2003; p. 457.
44. Horprasert, T.; Harwood, D.; Davis, L.S. A statistical approach for real-time robust background subtraction and shadow detection. *IEEE ICCV* **1999**, *99*, 1–19.
45. Brendel, W.; Todorovic, S. Video Object Segmentation by Tracking Regions. In proceedings of IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 833–840.
46. Yu, T.; Zhang, C.; Cohen, M.; Rui, Y.; Wu, Y. Monocular Video Foreground/Background Segmentation by Tracking Spatial-color Gaussian Mixture Models. In Proceedings of IEEE Workshop on Motion and Video Computing (WMVC), Austin, TX, USA, 23–24 February 2007; p. 5.
47. Murray, D.; Basu, A. Motion tracking with an active camera. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 449–459.

48. Kim, K.K.; Cho, S.H.; Kim, H.J.; Lee, J.Y. Detecting and Tracking Moving Object Using an Active Camera. In Proceedings of IEEE 7th International Conference on Advanced Communication Technology (ICACT), Phoenix Park, Dublin, Ireland, 21–23 February 2005; Volume 2, pp. 817–820.
49. Daniilidis, K.; Krauss, C.; Hansen, M.; Sommer, G. Real-time tracking of moving objects with an active camera. *Real-Time Imaging* **1998**, *4*, 3–20.
50. Huang, C.; Chen, Y.; Fu, L. Real-time Object Detection and Tracking on a Moving Camera Platform. In Proceedings of IEEE ICCAS-SICE, Fukuoka, Japan, 18–21 August 2009; pp. 717–722.
51. Shechtman, E.; Irani, M. Space-time Behavior Based Correlation. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 405–412.
52. Sedai, S.; Bennamoun, M.; Huynh, D. Context-based Appearance Descriptor for 3D Human Pose Estimation from Monocular Images. In Proceedings of IEEE Digital Image Computing: Techniques and Applications (DICTA), Melbourne, VIC, Australia, 1–3 December 2009; pp. 484–491.
53. Ramanan, D.; Forsyth, D.A.; Zisserman, A. Tracking people by learning their appearance. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 65–81.
54. Agarwal, A.; Triggs, B. Recovering 3D human pose from monocular images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 44–58.
55. Schindler, K.; Gool, L.V. Action Snippets: How Many Frames Does Human Action Recognition Require? In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 24–26 June 2008; pp. 1–8.
56. Danafar, S.; Gheissari, N. Action recognition for surveillance applications using optic flow and SVM. *Comput. Vis.–ACCV 2007*, **2007**, *4844*, 457–466.
57. Lowe, D.G. Object Recognition from Local Scale-invariant Features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–25 September 1999; Volume 2, pp. 1150–1157.
58. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
59. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
60. Dargazany, A.; Nicolescu, M. Human Body Parts Tracking Using Torso Tracking: Applications to Activity Recognition. In Proceedings of IEEE Ninth International Conference on Information Technology: New Generations (ITNG), Las Vegas, NV, USA, 16–18 April 2012; pp. 646–651.
61. Nakazawa, A.; Kato, H.; Inokuchi, S. Human Tracking Using Distributed Vision Systems. In Proceedings of IEEE Fourteenth International Conference on Pattern Recognition, Brisbane, Qld., Australia, 20 August 1998; Volume 1, pp. 593–596.
62. Iwasawa, S.; Ebihara, K.; Ohya, J.; Morishima, S. Real-time Estimation of Human Body Posture from Monocular Thermal Images. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 17–19 June 1997; pp. 15–20.

63. Leung, M.K.; Yang, Y. First sight: A human body outline labeling system. *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 359–377.
64. Leong, I.; Fang, J.; Tsai, M. Automatic body feature extraction from a marker-less scanned human body. *Comput.-Aided Des.* **2007**, *39*, 568–582.
65. Lee, M.W.; Cohen, I. A model-based approach for estimating human 3D poses in static images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 905–916.
66. Lee, M.W.; Nevatia, R. Body Part Detection for Human Pose Estimation and Tracking. In Proceedings of IEEE Workshop on Motion and Video Computing (WMVC), Austin, TX, USA, 23–24 February 2007 ; pp. 23–23.
67. Lee, M.W.; Nevatia, R. Human pose tracking in monocular sequence using multilevel structured models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 27–38.
68. Rogez, G.; Guerrero, J.J.; Orrite, C. View-invariant Human Feature Extraction for Video-Surveillance Applications. In Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), London, UK, 5–7 September 2007; pp. 324–329.
69. Ke, S.; Zhu, L.; Hwang, J.; Pai, H.; Lan, K.; Liao, C. Real-time 3D Human Pose Estimation from Monocular View with Applications to Event Detection and Video Gaming. In Proceedings of Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Boston, MA, USA, 29 August–1 September 2010; pp. 489–496.
70. Ke, S.; Hwang, J.; Lan, K.; Wang, S. View-invariant 3D Human Body Pose Reconstruction Using a Monocular Video Camera. In Proceedings of Fifth ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), Ghent, Belgium, 23–26 August 2011; pp. 1–6.
71. Campbell, L.W.; Becker, D.A.; Azarbayejani, A.; Bobick, A.F.; Pentland, A. Invariant Features for 3-D Gesture Recognition. In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, Killington, VT, USA, 14–16 October 1996; pp. 157–162.
72. Müller, M.; Röder, T.; Clausen, M. Efficient content-based retrieval of motion capture data. *ACM Trans. Graph. (TOG)* **2005**, *24*, 677–685.
73. Hoang, L.U.T.; Ke, S.; Hwang, J.; Yoo, J.; Choi, K. Human Action Recognition based on 3D Body Modeling from Monocular Videos. In Proceedings of Frontiers of Computer Vision Workshop, Tokyo, Japan, 2–4 February 2012; pp. 6–13.
74. Hoang, L.U.T.; Tuan, P.V.; Hwang, J. An Effective 3D Geometric Relational Feature Descriptor for Human Action Recognition. In Proceedings of IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), Ho Chi Minh City, Vietnam, 27 February–1 March 2012; pp. 1–6.
75. Rabiner, L.; Juang, B. An introduction to hidden Markov models. *IEEE ASSP Mag.* **1986**, *3*, 4–16.
76. Huang, X.; Acero, A.; Hon, H. *Spoken Language Processing*; Prentice Hall PTR: Upper Saddle River, NJ, USA, 2001; Volume 15.
77. Hoang, L.U.T.; Ke, S.; Hwang, J; Tuan, P.V.; Chau, T.N. Quasi-periodic Action Recognition from Monocular Videos via 3D Human Models and Cyclic HMMs. In Proceedings of IEEE International Conference on Advanced Technologies for Communications (ATC), Hanoi, Vietnam, 10–12 October 2012; pp. 110–113.

78. Murphy, K.P. Dynamic Bayesian networks: Representation, inference and learning. PhD diss., University of California, Berkeley, CA, USA, 2002.
79. Vapnik, V . *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1999.
80. Vapnik, V.; Golowich, S.E.; Smola, A. Support vector method for function approximation, regression estimation, and signal processing. *Adv. Neural Inf. Process. Syst.* **1997**, *9*, 281–287.
81. Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
82. Tipping, M.E. The relevance vector machine. *Adv. Neural Inf. Process. Syst.* **2000**, *12*, 652–658.
83. Fiaz, M.K.; Ijaz, B. Vision based Human Activity Tracking using Artificial Neural Networks. In Proceedings of IEEE International Conference on Intelligent and Advanced Systems (ICIAS), Kuala Lumpur, Malaysia, 15–17 June 2010; pp. 1–5.
84. Jain, A.K.; Duin, R.P.W.; Mao, J. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 4–37.
85. Jordan, A. On discriminative *vs.* generative classifiers: A comparison of logistic regression and naive Bayes. *Adv. Neural Inf. Process. Syst. (NIPS)*, **2002**, *14*, 841.
86. Welch, G.; Bishop, G. An Introduction to the Kalman Filter. In *Technical Report TR 95–041*; Department of Computer Science, University of North Carolina at Chapel Hill: Chapel Hill, NC, USA, 1995.
87. Stauffer, C.; Grimson, W.E.L. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 747–757.
88. Aggarwal, J. K.; Park, S. Human Motion: Modeling and Recognition of Actions and Interactions. In Proceedings of IEEE 2nd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), Thessaloniki, Greece, 6–9 September 2004; pp. 640–647.
89. Valera, M.; Velastin, S.A. Intelligent distributed surveillance systems: A review. *IEE Proc. Vis. Image Signal Process.* **2005**, *152*, 192–204.
90. Moeslund, T.B.; Hilton, A; Krüger, V. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* **2006**, *104*, 90–126.
91. Krüger, V.; Kragic, D.; Ude, A.; Geib, C. The meaning of action: A review on action recognition and mapping. *Adv. Robot.* **2007**, *21*, 1473–1501.
92. Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1473–1488.
93. Enzweiler, M.; Gavrila, D.M. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2179–2195.
94. Candamo, J.; Shreve, M.; Goldgof, D.B.; Sapper, D.B.; Kasturi, R. Understanding transit scenes: A survey on human behavior-recognition algorithms. *IEEE Trans. Intell. Transp. Syst.* **2010**, *11*, 206–224.
95. Aggarwal, J.K.; Ryoo, M.S. Human activity analysis: A review. *ACM Comput. Surv. (CSUR)* **2011**, *43*, 16.
96. Jiang, Y.; Bhattacharya, S.; Chang, S.; Shah, M. High-level event recognition in unconstrained videos. In *International Journal of Multimedia Information Retrieval*, **2013**, *2*, 73–101
97. Enzweiler, M.; Gavrila, D.M. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2179–2195.

98. Piccardi, M. Background Subtraction Techniques: A Review. In Proceedings of IEEE International Conference on Systems, Man and Cybernetics, The Hague, The Netherlands, 10–13 October 2004; Volume 4, pp. 3099–3104.
99. Zhang, Z.; Li, M.; Li, S.Z.; Zhang, H. Multi-view Face Detection with Floatboost. In Proceedings of Sixth IEEE Workshop on Applications of Computer Vision (WACV), Orlando, FL, USA, 3–4 December 2002; pp. 184–188.
100. Lucas, B.D.; Kanade, T. An Iterative Image Registration Technique with An Application to Stereo Vision. In Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, B.C., Canada, 24–28 August 1981.
101. Shi, J.; Tomasi, C. Good Features to Track. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 21–23 June 1994; pp. 593–600.
102. Dedeoğlu, Y.; Töreyin, B.U.; Güdükbay, U.; Çetin, A.E. Silhouette-based method for object classification and human action recognition in video. In Proceedings of the 9th European Conference on Computer Vision (ECCV) in Human-Computer Interaction, Graz, Austria, 7–13 May 2006; pp. 64–77.
103. Cherla, S.; Kulkarni, J.; Kale, A.; Ramasubramanian, V. Towards Fast, View-invariant Human Action Recognition. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR), Anchorage, AK, USA, 24–26 June 2008; pp. 1–8.
104. Rabiner, L.; Juang, B. *Fundamentals of Speech Recognition*; Prentice Hall: Englewood Cliffs, NJ, USA, 1993.
105. Dryden, I.L.; Mardia, K.V. *Statistical Analysis of Shape*. Wiley: Chichester, UK, 1998.
106. Ramanan, D.; Forsyth, D.A. Finding and Tracking People from the Bottom Up. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 16–22 June 2003; Volume 2, pp. II-467–II-474.
107. Ramanan, D.; Forsyth, D.A.; Zisserman, A. Strike a Pose: Tracking People by Finding Stylized Poses. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; vol. 1, pp. 271–278.
108. Moeslund, T.B.; Granum, E. A survey of computer vision-based human motion capture. *Comput. Vis. Image Underst.* **2001**, *81*, 231–268.
109. Isard, M.; Blake, A. Condensation—conditional density propagation for visual tracking. *Int. J. Comput. Vis.* **1998**, *29*, 5–28.
110. Gilks, W.R.; Richardson, S.; Spiegelhalter, D.J. *Markov Chain Monte Carlo in Practice*; Chapman & Hall/CRC: London, UK, 1996; Volume 2.
111. Zhu, S.; Zhang, R.; Tu, Z. Integrating Bottom-up/Top-down for Object Recognition by Data Driven Markov Chain Monte Carlo. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, SC, USA, 13–15 June 2000; Volume 1, pp. 738–745.
112. Bird, N.D.; Masoud, O.; Papanikolopoulos, N.P.; Isaacs, A. Detection of loitering individuals in public transportation areas. *IEEE Trans. Intell. Transp. Syst.* **2005**, *6*, 167–177.

113. Niu, W.; Long, J.; Han, D.; Wang, Y. Human Activity Detection and Recognition for Video Surveillance. In Proceedings of IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 27–30 June 2004; Volume 1, pp. 719–722.
114. Töreyin, B.U.; Dedeoğlu, Y.; Çetin, A.E. HMM based falling person detection using both audio and video. In Proceedings of the 2005 International Conference on Computer Vision (ICCV) in Human-Computer Interaction, Beijing, China, 17–20 October; pp. 211–220.
115. Shieh, W.; Huang, J. Speedup the Multi-Camera Video-Surveillance System for Elder Falling Detection. In Proceedings of IEEE International Conference on Embedded Software and Systems (ICESS), HangZhou, Zhejiang, China, 25–27 May 2009; pp. 350–355.
116. Ristad, E.S.; Yianilos, P.N. Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 522–532.
117. Hall, P.A.; Dowling, G.R. Approximate string matching. *ACM Comput. Surv. (CSUR)* **1980**, *12*, 381–402.
118. Sengto, A.; Leauhatong, T. Human Falling Detection Algorithm Using Back Propagation Neural Network. In Proceedings of IEEE Biomedical Engineering International Conference (BMEiCON), Ubon Ratchathani, Thailand, 5–7 December 2012; pp. 1–5.
119. Jacques, J.C.S., Jr.; Musse, S.R.; Jung, C.R. Crowd analysis using computer vision techniques. *IEEE Signal Process. Mag.* **2010**, *27*, 66–77.
120. Subburaman, V.B.; Descamps, A.; Carincotte, C. Counting People in the Crowd Using a Generic Head Detector. In Proceedings of IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), Beijing, China, 18–21 September 2012; pp. 470–475.
121. Merad, D.; Aziz, K.E.; Thome, N. Fast People Counting Using Head Detection from Skeleton Graph. In Proceeidngs of Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Boston, MA, USA, 29 August–1 September 2010; pp. 233–240.
122. Lu, C.P.; Hager, G.D.; Mjolsness, E. Fast and Globally Convergent Pose Estimation from Video Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 610–622.
123. McKenna, S.J.; Jabri, S.; Duric, Z.; Rosenfeld, A.; Wechsler, H. Tracking groups of people. *Comput. Vis. Image Underst.* **2000**, *80*, 42–56.
124. Chu, C.; Hwang, J.; Wang, S.; Chen, Y. Human Tracking by Adaptive Kalman Filtering and Multiple Kernels Tracking with Projected Gradients. In Proceedings of IEEE Fifth ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), Ghent, Belgium, 23–26 August 2011; pp. 1–6.
125. Saxena, S.; Brémond, F.; Thonnat, M.; Ma, R. Crowd behavior recognition for video surveillance. In Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS), Juan-les-Pins, France, 20–24 October 2008; pp. 970–981.
126. Vu, V.; Bremond, F.; Thonnat, M. Automatic video interpretation: A novel algorithm for temporal scenario recognition. *Int. Jt. Conf. Artif. Intell.* **2003**, *18*, 1295–1302.
127. Szczodrak, M.; Kotus, J.; Kopaczewski, K.; Lopatka, K.; Czyzewski, A.; Krawczyk, H. Behavior Analysis and Dynamic Crowd Management in Video Surveillance System. In Proceedings of IEEE 22nd International Workshop on Database and Expert Systems Applications (DEXA), Toulouse, France, 29 August–2 September 2011; pp. 371–375.

128. Cho, S.; Kang, H. Integrated Multiple Behavior Models for Abnormal Crowd Behavior Detection. In Proceedings of IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), Santa Fe, NM, USA, 22–24 April 2012; pp. 113–116.
129. Mehran, R.; Oyama, A.; Shah, M. Abnormal Crowd Behavior Detection Using Social Force Model. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 935–942.
130. Boiman, O.; Irani, M. Detecting Irregularities in Images and in Video. In Proceedings of Tenth IEEE International Conference on Computer Vision (ICCV), Beijing, China, 17–20 October 2005; Volume 1, pp. 462–469.
131. Kim, J.; Grauman, K. Observe Locally, Infer Globally: A Space-time MRF for Detecting Abnormal Activities with Incremental Updates. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 2921–2928.
132. Mahadevan, V.; Li, W.; Bhalodia, V.; Vasconcelos, N. Anomaly Detection in Crowded Scenes. In Proccedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1975–1981.
133. Chan, A.B.; Vasconcelos, N. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 909–926.
134. Adam, A.; Rivlin, E.; Shimshoni, I.; Reinitz, D. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 555–560.
135. Kratz, L.; Nishino, K. Anomaly Detection in Extremely Crowded Scenes Using Spatio-temporal Motion Pattern Models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 1446–1453.
136. Moënne-Loccoz, N.; Brémond, F.; Thonnat, M. Recurrent Bayesian network for the recognition of human behaviors from video. In Proceedings of the 3rd International Conference on Computer Vision Systems (ICVS), Graz, Austria, 1–3 April 2003; pp. 68–77.
137. Lin, W.; Sun, M.; Poovandran, R.; Zhang, Z. Human Activity Recognition for Video Surveillance. In Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), Seattle, WA, USA, 18–21 May 2008; pp. 2737–2740.
138. Zin, T.T.; Tin, P.; Toriu, T.; Hama, H. A Markov Random Walk Model for Loitering People Detection. In Proceedings of IEEE Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), Darmstadt, Germany, 15–17 October 2010; pp. 680–683.
139. Ran, Y.; Zheng, Q.; Chellappa, R.; Strat, T.M. Applications of a simple characterization of human gait in surveillance. *IEEE Trans. Syst. Man, Cybern. Part B: Cybern.* **2010**, *40*, 1009–1020.
140. Hu, M.; Wang, Y.; Zhang, Z.; Zhang, D.; Little, J.J. Incremental learning for video-based gait recognition with LBP flow. *IEEE Trans. Syst. Man, Cybern. Part B: Cybern.* **2012**, *43*, 77–89.
141. Poseidon. The lifeguard's third eye. 2006. Available online: http://www.poseidon-tech.com/us/system.html (accessed on 22 November 2012).

142. Sicre, R.; Nicolas, H. Shopping Scenarios Semantic Analysis in Videos. In Proceedings of the 8th IEEE International Workshop on Content-Based Multimedia Indexing (CBMI), Grenoble, France, 23–25 June 2010; pp. 1–6.
143. Gafurov, D. A survey of biometric gait recognition: Approaches, security and challenges. In Proceedings of Norwegian Symposium on Informatics 2007 (NIK 2007), Oslo, Norway, 19–21 November 2007.
144. Pantic, M.; Pentland, A.; Nijholt, A.; Huang, T.S. Human computing and machine understanding of human behavior: A survey. *Artif. Intell. Hum. Comput.* **2007**, *4451*, 47–71.
145. Farhadi, A.; Hejrati, M.; Sadeghi, M.A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; Forsyth, D. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision (ECCV), Heraklion, Crete, Greece, 5–11 September 2010; pp. 15–29.