# Reviewing Detections and Tracking Approaches[*]

Daniel Rowe
Computer Vision Centre

February 13, 2008

In spite of being a relatively new research area, a massive number of contributions related to HSE have been published in the last years[44, 43]. Undoubtedly, it represents an ambitious challenge, which is further raising important amounts of private and public funds due to the increasing number of attractive commercial applications.

The growing number of contributions in recent years has motivated the publication of multiple surveys [1, 17, 55, 43]. These review the state of the art, while proposing new domain taxonomies. Nevertheless, this field still lacks from a widely accepted taxonomy which arrange in a systematic way the different works. Thus, it would be interesting to show the relations between these, while including a hierarchical classification.

Here, the most relevant surveys are revisited, thereby putting into context the work here proposed. Further, a new taxonomy is also proposed. Subsequently, the focus is placed on detection and tracking methods. Thus, some of the most significant algorithms are discussed. The advantages of the different methods are explained and their drawbacks exposed.

## 1 A Review of Most Relevant Surveys and Taxonomies on HSE

The increasing number of papers —first related to people detection and tracking, then also to the analysis and understanding of human motion— in the last years has led to the publication of several surveys. Each of them has presented a taxonomy which arrange the most significant previous works according to different criteria.

Aggarwal and Cai presented a series of reviews in different workshops. Finally, this work resulted in what is probably the first relevant survey [1]. It reviews proposed approaches from 1980 to 1998, and 51 papers are referenced. Their taxonomy considers three main areas: (i) *body structure analysis*, (ii) *tracking moving humans*, and (iii) *recognition*, see Fig. 1.

[*] From *Towards Robust Multiple-Target Tracking in Unconstrained Human-Populated Environments.* Daniel Rowe. PhD Thesis, Chapter 2, Universitat Autònoma de Barcelona, Spain, 2008.

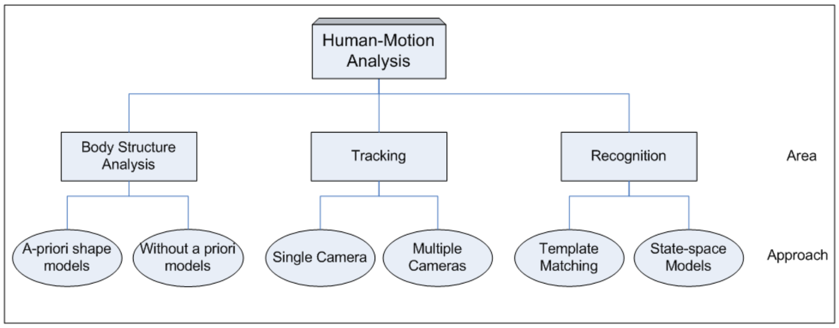


Figure 1: Taxonomy presented by Aggarwal and Cai in [1].

The first area concerns the structure of human-body parts. It is subdivided in two kind of approaches, depending on whether they rely on a-priori human shape models or not. Approaches from both categories can be grouped according to the representation used, namely, stick figures —the supporting bones— 2-D contours —the projection of the human figure— or volumetric models —modelling the flesh.

The second proposed area involves human tracking without considering its articulated configuration. Another subdivision is made based on whether a single camera or multiple perspectives are used. Papers from both approaches are also grouped depending on the representation, namely, points, 2-D blobs —that is, regions with similar properties— or 3-D volumes. The considered features are related to motion information (position, velocity), intensity values, etc.

The final area addresses human-activity recognition. Papers are grouped depending on whether they use *template-matching techniques* or *state-space models*. The former uses representations based on points, lines and blobs, while the latter uses point and meshes.

Another survey covering the time period from 1973 to 1997 —which references 81 papers— was presented by Gavrila [17]. Here, the classification is based on two criteria: the type of model, and the space dimensionality. Thus, this survey distinguishes three categories: (i) *2-D approaches without an explicit shape model*, (ii) *2-D approaches with explicit shape models*, and (iii) *3-D approaches*, see Fig. 2.

The first kind of approach relies on statistical descriptions based on low-level features and heuristics such as image moments, orientation histograms, and skin colour. The second one assumes a known point of view and a defined motion model. Representations are based on sticks and 2-D blobs. The third kind of approaches are mainly based on stick figures which model the skeleton, and 2-D surfaces or volumes which model the flesh. Features such as joint angles are considered. The three categories aim to provide results for all the required functionalities at the moment, that is, detection, tracking and recognition.

In addition, Gavrila provided an application classification altogether with

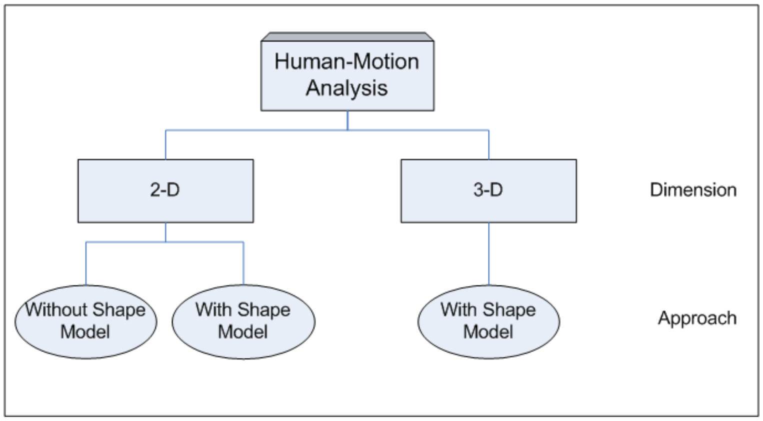


Figure 2: Taxonomy presented by Gavrila in [17].

the system required capabilities. Six fields are considered: virtual reality, smart surveillance, advanced user interfaces, motion analysis, and model-based coding. Among the capabilities, presence detection, identification, tracking, action recognition, and gesture or expression recognition can be found.

Moeslund and Granum [44] gave the most comprehensive survey, covering the years between 1980 and 2000 and citing 154 papers. Further, some previous surveys are discussed and compared. The covered period is later extended in [43], where contributions from 2000 to 2006 are included, and 337 papers are referenced.

In their work, a novel taxonomy based on functionalities is proposed: (i) *initialisation*, (ii) *tracking*, (iii) *pose estimation* and (iv) *recognition*, see Fig. 3. However, facial expression and hand gestures are not covered.

The first considered task concerns the camera, scene and target model initialisation, that is to say, calibration, manual or automatic parameter tuning, target initial pose, etc.

Then, tracking is addressed. The process is divided in three main tasks, i.e., target *segmentation*, *representation* and *tracking*. The former is divided in *temporal* and *spatial* approaches. According to the authors, on the one hand, temporal approaches can be subdivided into subtraction —which includes frame differencing and background subtraction— and optical flow techniques. On the other hand, spatial approaches may rely on thresholding, or on statistical methods.

Secondly, the representation of segmented entities is reviewed. Two categories are given, namely, *object-based* —points, boxes, silhouettes or active contours, and blobs— and *image-based* —spatial, spatio-temporal, edges, and features such as length, area, etc. Finally, the tracking task is discussed considering *model-based approaches* opposed to *probabilistic learnt models*; and single camera against multiple-camera approaches.

The third main functionality concerns the pose estimation. It is here con-

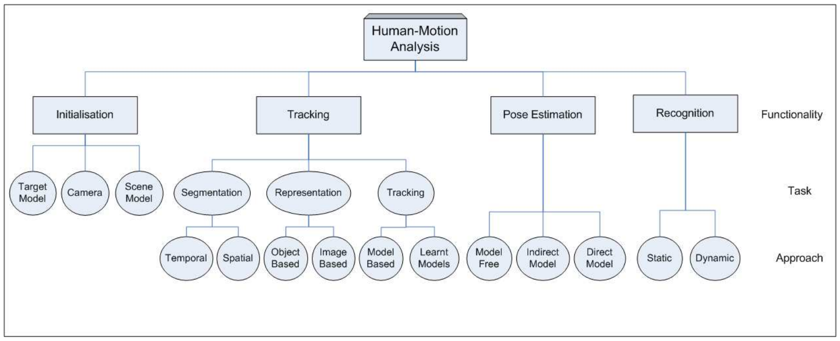


Figure 3: Taxonomy presented by Moeslund and Granum in [44].

sidered as either a tracking post-processing, or as an active part of it. Three categories are then given: *model-free*, *indirect model* and *direct model*. The former builds a representation without the use of an a-priori model. It can be based on a point, box or stick representation. The second category considers approaches which use a model as a guide to interpret the given data. The latter includes those approaches which use a direct model, that is, a detailed a-priori human model.

This last category is discussed in a comprehensive way. A large number of papers are classified according to their *abstraction level* —edges, silhouettes, sticks and joints, blobs, depth, texture, movement— the *dimension* —2-D, $2\frac{1}{2}$–D, 3-D— or the *model type* —cylinders, stick figures, patches, cones, ellipsoids, scaled prisms, CAD model, boxes, etc.

The way in which the results are evaluated is also taken into account: quantitative such as ground truth or manually segmented data, and qualitative such as visual inspection or animation.

Subsequently, the recognition task is addressed. Two distinction are made: *static* and *dynamic* recognition. Among the former, techniques such as template matching, normalised silhouettes or postures can be found in the literature. The latter includes low-level methods, such as spatio-temporal templates or motion templates, and high level ones such as Hidden Markov Models (HMM) or Neural Networks (NN).

Finally, a classification of applications is also proposed by considering three main areas: *surveillance*, *control* and *analysis*. A taxonomy relative to the assumptions made in the field is as well given, which consists of movement, environment and subject assumptions.

In 2003, Wang et al. presented an extensive and one of the most interesting surveys [55]. The time period from 1992 to 2001 is covered by citing 164 papers. Applications are classified under three categories, namely, *visual surveillance*, *advanced user interfaces,* and *motion-based diagnosis and identification*. Previous surveys are also revisited. This review presented a taxonomy based on functionalities organised in a hierarchical manner. The proposed framework

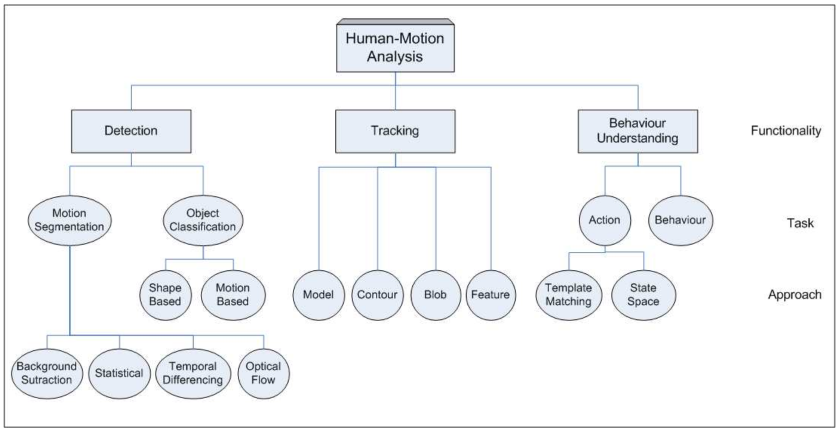


Figure 4: Taxonomy presented by Wang et al. in [55].

consist of three levels corresponding to low-level vision, intermediate-level vision and high-level vision. Each level is focused on one of the following task: *detection*, *tracking* and *behaviour understanding*, see Fig. 4.

The detection level aims to segment and group moving pixels corresponding to people. It is divided in two sub-processes: (i) *motion segmentation* and (ii) *object classification*. The former includes several approaches which are organised under four categories, namely, *background subtraction*, *statistical methods*, *temporal differencing* and *optical flow*. The latter is subdivided into two categories, which are *shape-based classification* and *motion-based classification*.

The goal of the tracking level is to establish coherent relations of image features between frames. Present-day approaches are classified according to whether they are *model-based*, *contour-based*, *region-based* or *feature-based*. With respect to the former, human-body models can be represented by stick figures, 2-D contours or volumetric models. The second and third kind of approaches aim to track detected contours and blobs, respectively. Finally, the last one aims to track sub-features as points or lines.

The highest level involves action recognition and description, and the analysis and understanding of human behaviours. The usual techniques are dynamic time warping, hidden Markov models or neural networks. The recognition is carried out under two groups of approaches, namely, *template matching* and *state-space methods*. Semantic descriptions are also receiving increasing attention from the community, as is stated by the authors.

Finally, Pentland [46] presented a paper which, without aiming to classify explicitly the up-to-time approaches, touches a diversity of human-motion analysis methods and applications. This domain was called in the paper “Looking at People”, and this term have been subsequently widely used[1]. A review of related

[1] As an example, the search of the terms “looking at people” plus “tracking” through the Internet yields more than 24000 hits.

mathematical techniques, and a domain taxonomy based in channels, scales and intentionality is provided. The state-of art of face recognition, surveillance, 3-D methods and perceptual user interfaces is revisited.

In order to put the presented work into context, it is worth to locate it within the taxonomies above revisited. Thus, it lies within the *tracking* area, and the *single-camera* approach category of the taxonomy proposed by Aggarwal and Cai [1]; within the *2D* area, and *without-shape-model* approach category of the one proposed by Gavrila in [17]; in the taxonomy proposed by Moeslund and Granum in [44], it lies within the *tracking* functionality, covering all *segmentation*, *representation*, and *tracking* tasks, and following *temporal* segmentation approaches, *object-based* representation, and *probabilistic learnt models*; finally, it the taxonomy presented by Wang et al. in [55], our work is covers both *detection* and *tracking* functionalities, and it addresses *motion-segmentation* and *tracking* tasks by following *statistical* approaches for the former, and *blob* ones for the latter.

## 2 State of the Art of Target Detection and Tracking

In this section, a review of the most relevant papers published in recent times relative to segmentation, detection and tracking approaches is presented. The different proposals are here outlined, and their advantages and drawbacks discussed. However, despite the huge efforts made, and the fact that achieving robust and accurate tracking is the first basic task to HSE, the problem is still open.

From the author point of view, target segmentation and tracking tasks are so linked that they should be considered together. Thus, a proper segmentation is, at least, essential for tracking initialisation and error recovery. And without applying a tracking scheme, it is not possible to keep a temporal consistency on detected targets. Further, it is really unusual to find a relevant paper specific to just segmentation or tracking. Papers are here inscribed in one of the following categories or another according to their main contribution, albeit they usually cover several tasks

This review implicitly presents a taxonomy according to the information flow. Thus, tracking is usually carried out using either bottom-up or top-down approaches. The formers rely on foreground segmentation, and a subsequent target association, which is usually followed by a state filtering; on the contrary, the latters are based on a prior complex motion, shape and/or appearance modelling, and a posterior state prediction. Thus, bottom-up approaches generate hypothesis according to the results of image processing, whereas top-down ones specify a-priori generated hypotheses according to current image data.

In this taxonomy, each of the bottom-up tasks is subsequently divided according to the different techniques used —which in some cases coincide with the ones stated by the aforementioned surveys.

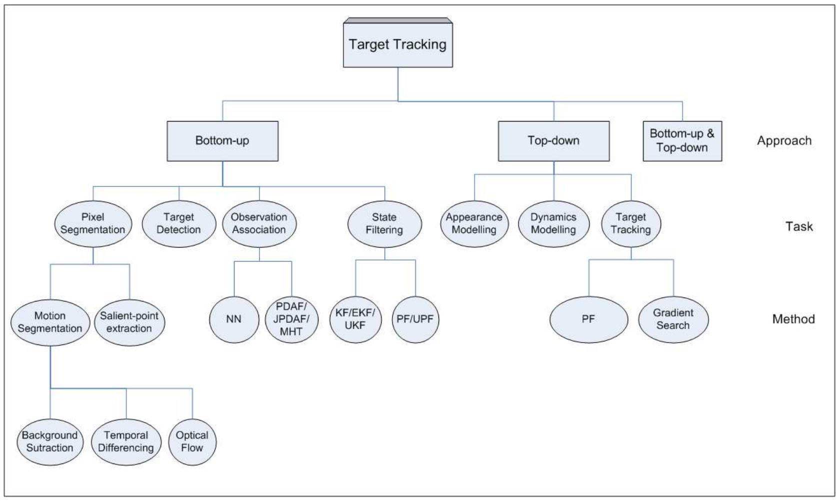


Figure 5: Proposed tracking taxonomy. Tracking approaches are classified in bottom-up and top-down methods. Bottom-up ones usually perform target segmentation, observation association, and state filtering tasks. Top-down approaches require an off-line appearance and dynamic modelling, and then perform target tracking according to the chosen methods.

Top-down approaches are split taking into account the tracking technique used, although it is subsequently detailed the feature in which the particular proposal rely. A sketch of this taxonomy is shown in Fig. 5.

Finally, some research groups have developed structured architectures which aim not to be restricted to a particular task, but to perform a global scene analysis [32, 49]. These contributions usually combine several techniques.

## 2.1 Bottom-up Tracking

Bottom-up tracking approaches are usually based on motion segmentation in order to extract foreground entities from the background [56, 42, 50]. This can be performed by means of *background subtraction*, *frame differencing*, a combination of both, or *optical flow*.

Alternatively, detection can be achieved by means of detection of salient features [21, 38, 6]. In this case, regions with high curvature in space-scale images —blobs— regions with large gradients —corners—- and other significant image characteristics are extracted. However, by using this kind of approaches, any salient background point is selected as a potential target.

### 2.1.1 Pixel Segmentation

This task involves separating image regions that do not belong to the background, and extracting them. Although this issue is closely related to movement, foreground objects could remain static for an unknown number of frames while the background may be in motion[2].

Motion segmentation algorithms face multiple difficulties. These can be classified into two categories, since some of them are intrinsic to the problem domain, whereas others may be seen as drawbacks of the approach used, see Table 1. Thus, the main difficulties are the following:

- **Bootstrapping**. It refers to the problems that arise when the method requires and initialisation period, and a scene free of moving objects cannot be assured.
- **Foreground aperture**. In this case, homogeneous object in motion cause that the inner part is not segmented.
- **Ghosts**. The relocation of a background object implies changes in both the old and the new location. However, only the latter should be identified as foreground region.
- **Stopped object**. Some motion segmentation methods requires significant changes between frames to segment any pixel. Thus, if a target stop motion, the segmentation fails.
- **Illumination changes**. These completely alter the pixels characteristics, thereby resulting in a drastic increase of pixel segmentation. They may be global —thereby yielding a general highlight or shadow— or local —which are mainly caused by target shadows. Further, they can also be sudden —such as those due to changes in weather conditions, or by turning on/off a light— or gradual.
- **Camouflage**. In this case, some of the pixel features between the background and the foreground are too similar to disambiguate them.
- **Clutter in motion**. Any approach that relies on motion to perform segmentation is liable to consider as foreground any moving background pixel.
- **Camera motion.** In this case, the whole scene seems to be in motion.

In the following, papers are classified according to the approach used, and how the different difficulties are addressed is explained.

[2] Think about a person stopped momentarily at a traffic light. He or she must still be considered as foreground and, therefore detected and tracked. On the other hand, waving branches and leaves or flowing water must not be segmented, although they are in motion.

| Drawbacks of common approaches | Intrinsic difficulties |
|---|---|
| Bootstrapping | Illumination changes |
| Foreground aperture | Camouflage |
| Ghost | Clutter in motion |
| Stopped Objects | Camera motion |

Table 1: Motion-segmentation difficulties.

**Background Subtraction** Background subtraction is one of the most commonly used approaches for motion segmentation [47, 35]. Pixels in motion are segmented by comparing the current image and a reference one, namely, the background model. In the early days, simple methods consisted in differencing each image and a reference one, and subsequently compare the result with an a-priori set threshold [22]:

$$|\mathbf{B}_t - \mathbf{I}_t| \quad > \quad \tau, \tag{1}$$

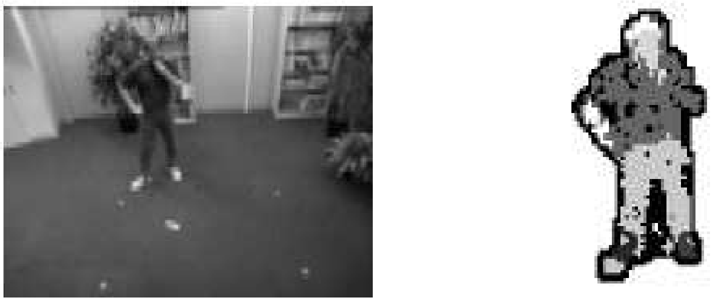
where $\mathbf{B}_t$ is the reference background at time $t$, $\mathbf{I}_t$ the current frame, and $\tau$ a pre-set threshold. The model could be subsequently updated following a Infinite-Impulse Response filter (IIR) :

$$\mathbf{B}_{t+1} \quad = \quad (1 - \alpha)\,\mathbf{B}_t + \alpha \mathbf{I}_t, \tag{2}$$

being $\alpha$ the adaptation rate that weights the current model versus the new observation. However, this method was extremely sensitive to changes in the background conditions such as lightning or due to background in motion, as well as to the camera noise. More recent approaches model either each pixel or group of pixels statistically. This allows building adaptive background models while providing robustness to the above-stated background conditions. Usually, model statistics are continuously updated in order to provide an adaptive approach.

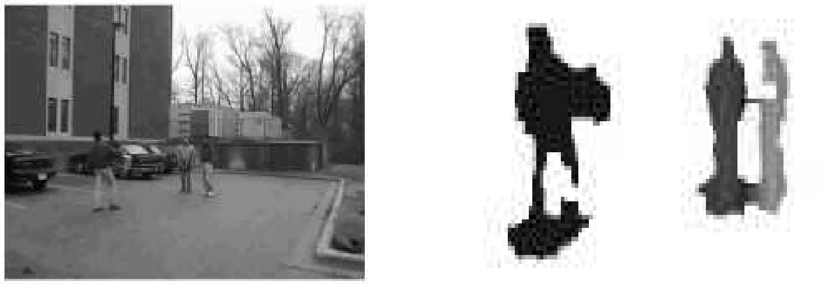
Among the background-subtraction approaches, Wren et al. developed the *Pfinder* algorithm [56]. Each scene pixel is modelled using a Gaussian colour distribution. Thus, outliers are assumed to be foreground pixels, and are therefore segmented. Visible pixels are updated using a single adaptive filter. Segmented pixels are grouped into blobs and each blob is modelled using spatial and colour components. Blobs are associated with body parts using a log likelihood measure and tracked by means of Kalman Filters (KF). However, it just attempts to detect and track one person, in upright posture, in indoor scenes. A sample frame is shown in Fig. 6.

Haritaoglu et al. presented the *W4* method [20, 19]. Unlike Pfinder, it aims to detect and track people, isolated or in groups, in outdoor scenes, and considering several poses. Each pixel is modelled with a range of intensity



(a) Sample frame (b) Obtained segmentation

Figure 6: Sample frame using the approach published in [56] by Wren et al.



(a) Sample frame (b) Obtained segmentation

Figure 7: Sample frame using the approach published in [19] by Haritaoglu et al.

values given by minimum and maximum intensity values, and the maximum intensity difference between frames during a training period. Pixels whose values are placed outside the interval which is given by the minimum value minus a multiple of the maximum difference and the maximum value plus a multiple of the maximum difference are considered as foreground pixels. A sample frame is shown in Fig. 7.

The model is periodically updated considering both pixel-based and object-based methods: the former updates the values of the pixels classified as background, and the latter replaces the model parameters for those pixels classified as static foreground. Neighbour pixels are grouped and blobs are classified using heuristics. Poses are identified by means of projection histograms. KFs and textural temporal templates are used to track detected targets. However, this approach is rather sensitive to shadows and lighting changes, since the only cue is the pixel intensity.

Horprasert et al. [23, 24] implemented an statistical colour background algo-

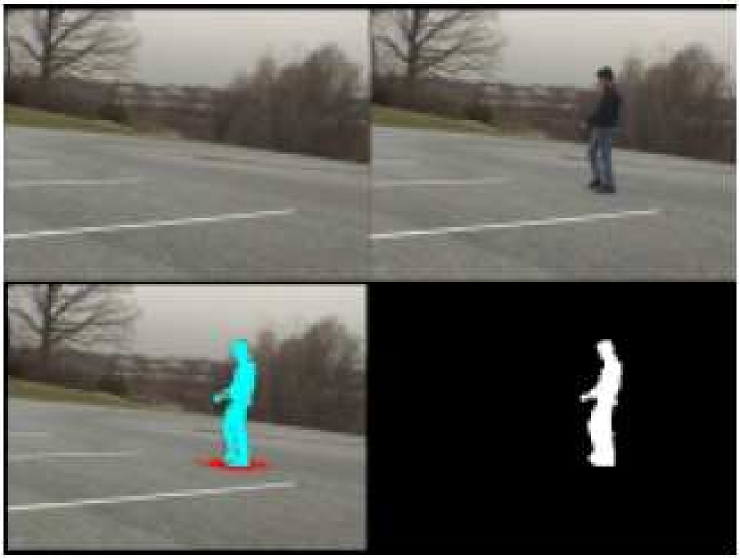


Figure 8: Sample frame using the approach published in [24] by Horprasert et al.

rithm, which models each pixel based on both brightness and colour distortion. It still needs a static background scene, but it's able to handle strong shadows and highlights. The proposed algorithm is able to classify the image pixels into four categories, namely, original, shadowed and highlighted background, and moving foreground. A sample result is shown in Fig. 8.

McKenna et al. [42] combined colour and gradient information in their adaptive background subtraction approach. Each pixel chrominance —given by the normalised red and green channels— is modelled using two Gaussians, one on each channel. The Gaussian parameters are updated using an adaptive filter. If one of the current chrominance values is farther from the mean more than three times the standard deviation, the pixel is marked as foreground. Using chrominance instead of RGB values, shadow detection is avoided, but it cannot cope with foregrounds of the same chrominance as the background. Thus, they also modelled the background pixels using the spatial RGB gradients, and pixels are also flagged as foreground if the gradient of any of the channels is out of the scope of the corresponding Gaussian. As a result, albeit foreground pixels with the same chrominance as the background can now be segmented, hard-edge shadows are also segmented. Tracking is done by means of data association.

Three levels of representation are used, namely *regions* —stable connected components— *people* —groups of regions that satisfy conditions relative to overlapping and area— and *groups* —people that share regions. People appearance is modelled using colour histograms. Visibility indexes —obtained from the probabilities that the pixels correspond to unoccluded people— are used to disambiguate occlusions. However, problems arise when several people and the background have a similar appearance. It is also assumed that the target appearance do not significantly change while the targets are grouped.

Still, shadow removal has not be properly addressed yet within a target detection framework, where shadows are considered to yield just changes in intensity, but not in chrominance. Last advances in the field —such as those contributions of Finlayson et al. [16]— need to be incorporated.

Nevertheless, none of these models can cope with background in motion. Stauffer and Grimson presented in [51] an approach focused on this issue. A colour background model is built using a Mixture of Gaussians (MoG) to represent each pixel. Thus, each Gaussian models the pixel colour distribution for one of the possible backgrounds learnt in a training period. Pixels which do not match any of the distributions are considered as foreground. The distribution weights are periodically updated according to the one that has matched the current pixel value. The least probable distribution is replaced in case none of them match the value, thereby, including long-term still foregrounds. The adaptive scheme apparently also copes with lighting and scenes changes, as well as motion from clutter. Tracking is performed by implementing a set of KFs.

Javed et al. [30] presented a method that aimed to solve most of the common segmentation difficulties: bootstrapping, ghosts, quick illumination changes, background in motion, and camouflage. It uses both colour and gradient cues. A hierarchical system is build based on three levels: *pixel*, *region* and *frame*.

At the pixel level, statistical models of pixel colour and gradients based on mixture of Gaussians are independently used to classify each pixel as potential background or foreground. At the region level, foreground pixels obtained from the colour model are grouped into regions, and the gradient model is then used to eliminate regions corresponding to highlights or ghosts. Pixel-based models are updated based on decisions made at the this level. Finally, the frame level ignores the colour-based segmentation if more than 50 percent of the image pixels are considered foreground. In this case, a global illumination change in considered, and segmentation is performed according to gradient information. Nevertheless, the ghosts are not eliminated if the background contains a high number of edges.

**Frame Differencing and Hybrid Algorithms** A typical temporal differencing approach segments motion by subtracting the current image from the previous one pixel by pixel. Then, pixels are segmented if the result is over a pre-defined threshold:

$$|\mathbf{I}_t - \mathbf{I}_{t-1}| \quad > \quad \tau. \tag{3}$$

It can also be done by considering several consecutive frames. For example, Collins et al. [9, 11] implemented an hybrid algorithm for target detection that combines an adaptive background subtraction and a three-frame differencing approach. Background subtraction techniques can provide good segmentation results, but they are extremely sensitive to scene changes due to dynamic background, lighting or extraneous events. In addition, *ghost* are usually detected when long-term stationary objects start moving —albeit statistical models eventually adapt to this situation. On the other hand, temporal differencing is very adaptive to dynamic environments and do not generate false alarms caused by ghosts, but it cannot segment all relevant pixels, and it may be rather sensitive to camera noise.

In that work, pixel intensity is taken as the representing feature. Thus, pixels whose intensity varies significantly from both the last frame and the next-to-last one are marked as moving. These pixels are clustered and a background subtraction method is applied to the inner region. Both background model and threshold are updated over time for non-moving pixels.

The approach is adapted to pan-tilt camera platforms by collecting a set of background references for known camera settings and registering the images according to selective pixel integration. They also introduced a *layered detection* algorithm: pixels are classified as stationary, transient or background according to two measures, namely, a motion trigger and a stability measure. These point out if the pixel belongs to a moving object, a stopped object or the "motion" is due to lightning changes. Foreground pixels are clustered into regions and classified as moving or stationary ones. Stationary regions constitute layers which are used to determine occlusions and motion resuming. Tracking is done by predicting next positions according to the estimated dynamic model, and convolving the object templates with candidate regions. Several scenarios are described according to the results of the two previous stages and hypotheses are launched accordingly. Finally, clutter in motion is rejected if the cumulative object displacement indicates changes in direction.

Thus, this system use a network of cooperative active cameras to detect and track people and vehicles in cluttered environments. Targets are classified into semantic categories and their activities are monitored. Once the geo-locations are extracted, symbolic data are inserted into a synthetic scene visualisation.

The algorithm proposed in [50] is also a good example of hybrid algorithms which combines frame differencing and background subtraction techniques to achieve motion segmentation. Segmentation is performed in two sequential steps. First, a fuzzy classification is carried out by according to current pixel motion on each RGB channel. Then, results are enhanced taken into account the previous segmentation result, and a background model. Finally, HSI colour space is used to eliminate shadows.

In addition to frame differencing and background subtraction, optical flow techniques have also been used to perform motion segmentation. These describe coherent feature motion between frames. These techniques independently segment moving objects, even in presence of camera motion. However, this approach is rather sensitive to noise and background in-motion, and it requires huge computational resources.

| Addressed difficulty | References |
|---|---|
| Sudden illumination changes | [56, 24, 42, 51, 50, 30] |
| Gradual illumination changes | [19, 24, 42, 51, 11] |
| Camouflage | [42, 30] |
| Clutter in motion | [51, 11, 30] |
| Camera motion | [7, 11] |
| Bootstrapping | [51, 19, 30] |
| Stopped Objects | [51, 19, 11, 30] |
| Ghosts | [51, 11, 19, 50, 30] |

Table 2: Motion-segmentation methods.

**Optical Flow** These methods look for coherent motion of points or features between frames. Bregler [7] presented a human-dynamics recognising method where motion is segmented according to optical flow results. An affine motion model is used for this purpose. Blobs are extracted by means of the *Expectation-Maximisation* (EM) algorithm, where the likelihood of each pixel of belonging to a particular blob depends on the coherent affine motion, HSV colour values, and spatial proximity. In order to incorporate past estimates, a bank of KFs provides priors for the EM initialisation, resulting in a MoG propagation.

Summarising, multiple techniques have been developed to tackle motion segmentation. They usually address a limited of the numerous difficulties expected. The way of solution may come from a smart combination of techniques. The different algorithms here described are summed up in Table 2, while pointing out the difficulties addressed.

### 2.1.2 Target Detection and Observation Association

Segmented pixels are grouped into blobs, which could be considered as an entity of interest. This is usually done according to a connected component analysis, and a subsequent spatial filtering process. Then, some features can be extracted to represent a target observation, thereby classifying the target, and concluding its detection.

However, as it has been above stated, in some cases this process is enhanced by taking into account the probability of a given pixel of belonging to the target according to some statistical model.

In general, once detection has been performed, several approaches arise to keep track of the targets. New observations can be just associated to previous ones. This process can be done taking into account different cues like spatial proximity or appearance similarity. The latter may consist of a template matching between newly detected targets and the models of the previous ones. In both cases several problems must be expected due to detection failures. These mainly occur because of segmentation errors —such as those due to background clutter which mimics the target appearance, and illumination changes— and target occlusions or merging.

Depending on whether several targets and measurements are expected, the association is accomplished using nearest-neighbour techniques, or by means of Data Association Filters —such as the *Probabilistic Data Association Filter* (PDAF), the *Joint Probabilistic Data Association Filter* (JPDAF), or the *Multiple Hypotheses Tracking* (MHT) [4].

### 2.1.3 State Filtering

Usually, a prediction stage is also incorporated after associating the observation, thereby providing better chances of tracking success. Filters such as the KF [34], or subsequent extensions and improvements such as the *Extended Kalman Filter* (EKF) [2] or *Unscented Kalman Filter* (UKF) [31, 54] are commonly used.

The KF is a linear recursive estimator which predicts the next state according to a dynamic model, and updates this result in agreement with the obtained measurement. Although it has been widely used, it presents important drawbacks:

1. it requires strong assumptions about the linearity and Gaussianity of the transition model and the likelihood function;
2. it cannot cope with multiple targets and measurements;
3. and, it relies on a previous segmentation in order to provide the measurement.

These requisites are often not feasible in MTT scenarios, specially during target grouping and occlusions, or in cluttered backgrounds. Therefore, several approaches have been implemented in order to avoid these restrictions. The EKF linearises both transition and likelihood models using Taylor series expansions. The system Jacobian is computed for the predicted states, and the results are used in the updating stage. However, the EKF keeps several drawbacks:

1. posterior densities are still modelled as Gaussians;
2. the series approximation can lead to poor representations of the posterior distribution —this is specially the case on highly non-linear systems, because only the mean is propagated through the non-linearity;
3. and, although the models do not need to be linear, they still must be differentiable.

The UKF aims to propagate high-order moments through non-linear functions. A set of deterministic sample points —called *sigma points*— are selected around the mean and subsequently propagated. It can be analytically proved that it yields better approximations of the mean and covariance than the EKF. Further, there is no need to compute expensive, computationally speaking, Jacobians. However, it cannot be applied to general non-Gaussian distributions.

More general dynamics and measurement functions can be dealt with by means of *Particle Filters* (PF) [15, 3]—which are also known as *Sequential Importance Re-sampling* (SIR)— and further evolutions, such as the *Unscented Particle Filter* (UPF) [53]. These address the filtering problem when no assumption about linearity or Gaussianity is made on almost all involved probability density functions. Since the seminal paper by Gordon et al. [18], PFs have been widely used to perform stochastic estimation. The algorithm is based on Bayesian filters. Therefore, they compute a posterior probability density function (pdf) which undergoes a diffusion-reinforcement process making use of Monte Carlo simulation techniques. The reinforcement stage is accomplished by means of factored sampling. Thus, the PF approach provides a complete representation of the posterior pdf. Therefore, any statistical estimate can be computed despite non-linearities and non-Gaussianity of the involved distributions. Multiple hypotheses can simultaneously be considered, and they can be propagated even when no evidence is obtained from the current image. However, the search region is reduced, which may increase the processing speed, but the robustness could as well be cut down.

Although the asymptotic correctness of the algorithm is proved, it has several drawbacks [36]:

1. there is no information about the number of samples required for a requested precision, specially for undefined times lengths;

2. it suffers from several intrinsic problems such as *sample degeneration* or *sampling impoverishment*, depending on the whether re-sampling is used or not;

3. and finally, PFs were initially designed to keep multiple hypotheses but only for a single target; further extensions which combine information about all targets in every sample usually cause the curse of dimensionality.

In every PF approach, samples are drawn from a proposal distribution. Usually, the transition model is used as such proposal. However, problems may arise if the samples are placed in the tail of the temporal prior or if the likelihood is very peaked. De Freitas et al. [13] used the results provided by EKF as a proposal distribution. More recently, given that the UKF outperforms the EKF, this filter has been used to generate the prior samples [53].

## 2.2 Top-down Tracking

Despite these efforts, there are many situations where segmentation-from-motion, and the subsequent observation-tracker correspondence, is not possible, like in target grouping or target occlusion. Top-down approaches incorporate a-priori knowledge about the targets and the context in order to tackle these situations. Thus, these methods rely on accurate target modelling. Hence, complex templates, which should cope with an important degree of deformation, are predefined. Further, high-level motion patterns are a-priori learnt, and used to reduce the state-space search region in agreement to some state prediction.

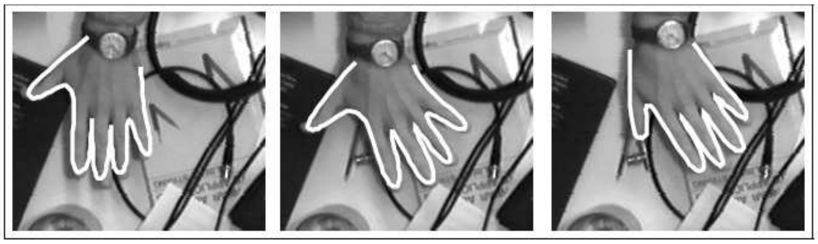


Figure 9: Sample frame using the approach published in [25] by Isard and Blake.

Further, targets can be localised following an appearance segmentation, instead of a motion segmentation. This relies on feature extraction, and a subsequent exhaustive search of some feature patterns learnt during a classifier training process.

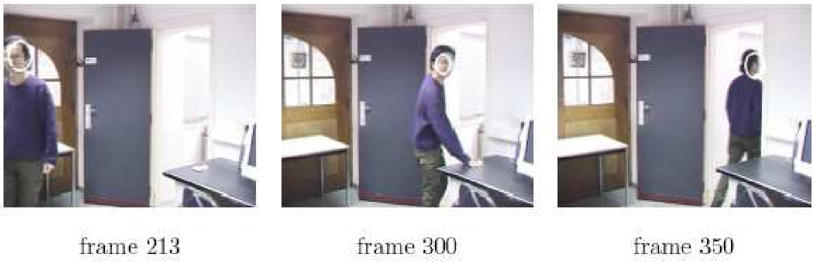
Nevertheless, model-based high-level tracking is not feasible in case this information is not available there is not enough a-priori knowledge about either the scene or the targets. Also, an accurate initialisation is often not possible. The need of adaptation when target appearances considerably evolve over time usually leads to the phenomenon known as *model drift*. In those cases, motion-based tracking usually outperforms model-based appearance or shape tracking.

Notwithstanding, numerous proposals have been presented to perform model-based tracking, while trying to overcome these drawbacks.

### 2.2.1 Particle Filtering

The aforementioned PF techniques —together with complex dynamic and appearance models— have constituted a common approach [25, 39, 41, 37, 52, 14]. These techniques were introduced in the Computer-Vision field in Condensation [25, 27] by Isard and Blake, albeit they were already known in some other areas, such as Automatic Control or Artificial Intelligence. This algorithm is based on a PF framework combined with edge-based image features. Subsequently, contour tracking have been widely researched within this framework [26, 40], although this may not be the best approach in crowded scenarios because of the potential multiple occlusions. A sample performance is shown Fig. 9.

Nummiaro et al. [45] applied PFs using colour distributions as image features. These are approximated using histograms, which are supposed to be less sensitive to partial occlusions and rotations in depth than other appearance models such as templates. They used the HSV colour space since they claimed that it can provide robustness to changes in lightning conditions. Histograms are calculated inside an elliptic region, once the pixels have been weighted according to a kernel. A similarity function is implemented using the Bhattacharyya Coefficient (BC) [5]. Samples are represented using the centroid position in image coordinates, its speed, the length of the ellipsis axes, and a scale change.



frame 213 frame 300 frame 350

Figure 10: Sample frame using the approach published in [45] by Nummiaro et at.

The tracker is initialised placing samples —assuming a known target model— at strategic positions. Models are only updated when the likelihood of the estimated state is over a pre-defined threshold. However, no MTT is considered —which implies that no event such as target grouping or occlusion can be analysed— and it lacks from an independent observation process, since samples are evaluated according to the histograms of the predicted image region. A sample frame is shown in Fig. 10.

Perez et al. [48] proposed also a PF based on a colour-histogram likelihood. They introduced interesting extensions in multiple-part modelling, incorporation of background information, and MTT. Nevertheless, it may require an extremely large number of samples, since one sample contains information about the state of all targets, dramatically increasing the state dimensionality. Further, no appearance model updating is performed, what leads to target loss in dynamic scenes.

Deutscher and Reid [14] presented an attractive approach called *Annealing Particle Filter* to recover full body motion. It aims to reduce the required number of samples. A series of weighting functions is designed from the original one by raising to a series of decreasing exponents, thereby defining a series of layers. One annealing run is performed at each time slice. The run started using the broader weighting function. At each layer, $N$ particles are weighted, re-sampled with replacement, and used to yield a particle set for the next layer by applying Gaussian diffusion. As a result, all particles are spread around the global maximum. This final set is used to initialise the broader layer at the next time slice. Thus, the number or required samples is considerably reduced. However, pruning hypotheses with lower likelihood may lead to a single hypothesis, and therefore it could be inappropriate in cluttered environments.

The weighting function is built taken into account two image features: edges and silhouettes. Edges are obtained using a gradient-based mask over the entire image. Silhouettes are produced using a background-subtraction algorithm. Pixel weight maps are built taken into account both the proximity to an edge, and its enclosing into an extracted silhouette. In addition, two enhancements are introduced. Firstly, a soft-partition sampling is implementing by adding an amount of randomness to each parameter proportional to the variance of that

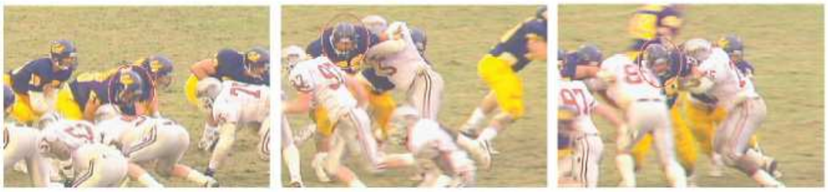


Figure 11: Sample frame using the approach published in [12] by Comaniciu et al.

parameter. In this way, samples are not wasted and the effort is concentrated on those parameters whose uncertainty is bigger. Secondly, a cross-over operator is used by combining selected particles, and thereby, tracking in parallel different sections of the search space. As they focus on motion analysis, multiple targets and unconstrained environments are not explored.

BraMBLe [29] is an appealing approach to multiple-blob tracking which models both background and foreground using MoG. However, no model updating is performed, there is a common foreground model for all targets, and it suffers from the curse of dimensionality —as all PF-based methods which tackle MTT combining information about all targets in every sample.

Occlusion events present particular difficulties which should be explicitly addressed. Wu et al. [57] address these issues using a PF by implementing a Dynamic Bayesian Network (DBN) with an extra hidden process for occlusion handling.

### 2.2.2 Gradient-descent Search

Target localisation following a gradient-descent search —*Mean-shift* tracking— has also been commonly used [8, 12, 10]. The search is performed in the basin of attraction of a spatially-smooth similarity function given by a weighted image region. Thus, in this case the search is deterministic. This is usually done according to a measure of histogram similarity between both model and candidate distributions related to the BC.

However, these methods do not work in unconstrained situations. The main drawbacks of the algorithm consist of the assumptions that the target candidate do not drastically change its appearance between time steps, and that its new location is in the basin of attraction of the similarity function, which is defined by the kernel size. Further, it is assumed that the similarity function presents a unique local maximum within the basin of attraction. In addition, only one hypothesis is considered, thereby limiting its effectiveness in case of occlusions or heavy cluttered backgrounds.

For instance, Comaniciu et al. [12] represented a target by an elliptic regions defined at given location, and a target model. This is obtained from the features of the normalised-to-unit-circle pixels locations, once applied an isotropic kernel. Colour is selected as image feature, and the target model pdf is approximated

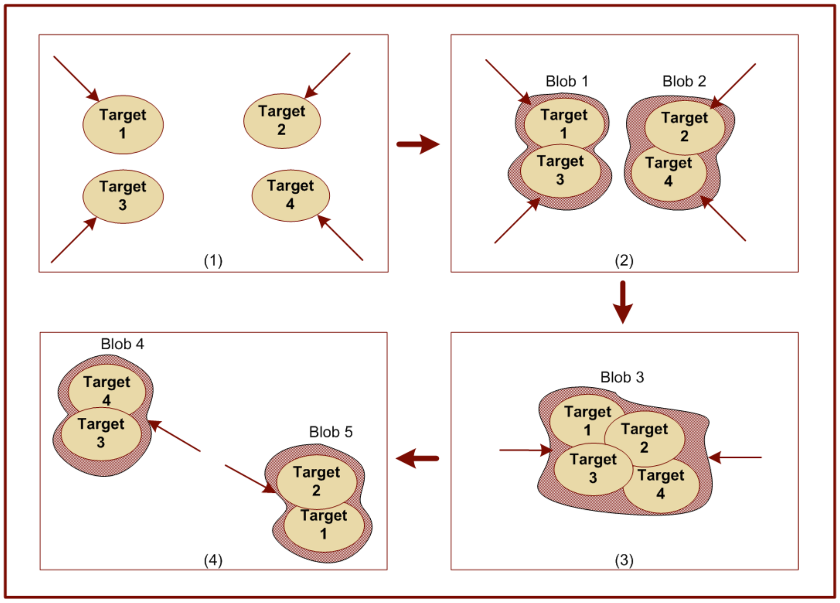


Figure 12: Target interaction. Keeping the identity of multiple targets which cannot be independently segmented is a challenging task. Notice the different group membership of targets in blob 1 and 4.

by means of histograms. However, it tracks just one target, initialised by hand, and the appearance model is never updated. A sample performance is shown in Fig 11.

Collins et al. [10] presented an appealing tracker, based also on the mean-shift algorithm, with on-line feature selection of discriminative features. It aims to maximise the distinction between the target appearance and its surroundings. Still, it tracks just one target, and may suffer from model drift, although models are anchored to the first frame, which is manually segmented. It still tracks rigid targets (or rigid regions of them), appearance changes are limited, and since MTT is not considered, interaction events are not studied. These facts cannot be seen as minor issues in real applications such as video-surveillance.

## 2.3 Bottom-up and Top-down Tracking

Algorithms which combine both bottom-up and top-down approaches have also been proposed [28, 49]. Most appealing approaches rely on the combination of several techniques. Senior et al. presented a two-level tracking system with template-based appearance models [49]. These are used in conjunction with probability masks to infer depth ordering and detect occlusions. Nonetheless, appearance ambiguities among grouped targets have not been addressed.

In [28], the probabilistic top-down tracking framework developed for Con-

DENSATION [27] is extended by means of *importance sampling* in order to generate samples according to a bottom-up process.

Yang et al. [58] proposed a system which specifically tackles grouping situations, albeit no filtering is carried out, and grouped targets are not independently tracked. Thus, during grouping events, just a coarse localisation can be obtained by considering that the targets are inside the group region. Therefore, grouped targets are not accurately tracked, and no complex situation can satisfactorily be faced —for instance, those in which a group of more-than-two members merge and split, see Fig. 12.

Kahn et al. [32, 33] developed a system called Perseus. It is a visual purposive architecture which aims to recognise gestures. The way in which the structure is modularised was surprisingly novel, allowing the system to use knowledge about context and task at every stage and providing it with redundancy and independence of assumptions. It also provides an interface to higher-level systems. It consisted of six components: a *planner* is located at the higher level. It called *visual routines* which aim to detect and track selected objects. *Object representations (OR)* —background objects, light, people, objects, etc.— can be instantiated, which involves registering it at the *long term visual memory*. The object methods, such as *segment*, keep a *global segmentation map* using the image features maps located at the lower level. The considered features are intensity, edges, disparity, colour and motion. All higher levels made use of these maps to carry out their functionalities. Features parameters can be tuned according to the task and context. All object representations are also associated to *markers* which track the segmented objects.

Alternatively, several approaches take advantage of 3D information by making use of a known camera model and assuming that agents move on a known ground plane. These and other assumptions relative to a known Sun position or constrained standing postures allow the system presented in [59] to initialise trackers on people who do not enter the scene isolated.

## 3 Discussion

Summarising, an evolution in the perception of the analysis of the human motion task can certainly be noticed. Taxonomies have being refined from mere classifications according to the aim of the task, or even to criteria such as the model dimension or the sensor used, to hierarchical structures which cope which all the required functionalities. These are spread through different levels which are task-oriented.

However, this area is sill in a transition step between Image Processing and Pattern Recognition, and a more advanced view in which Cognitive Sciences provide a global understanding of the scene. The latter supplies also interactive capabilities, such as a natural language communication between a user and the system, or synthetic scene visualisations.

With respect to segmentation, it can be concluded that although remarkable advances have been achieved by presenting a wide set of different approaches, the segmentation task is still an open problem. These techniques must be enhanced to cope successfully with the numerous difficulties expected, specially in outdoor scenes. Among these difficulties, we can include lighting changes, different weather conditions, background in motion, or camouflage. Further, it is still not clear how to deal with background objects which unexpectedly move at a given moment, with the *ghost* they leave, or with foreground objects which stop momentarily. The solution may come from the combination and development of some of the existing approaches, thereby providing the system with redundancy. Taking advantage of context knowledge and making use of high-level information may also be a way of solution.

With respect to tracking, numerous approaches have been proposed to perform this task. Data-association techniques on their own are not reliable enough, since they completely depend on a proper segmentation. Prediction-updating approaches should be flexible and general enough to cope with complex environments. The combination of several of the aforementioned techniques may lead to a way of solution. Thus, for instance, EKF/UKF approaches may enhance system predictions; mean shift techniques could adjust final estimates; and several segmentation methods may be combined with prediction-updating techniques in order to provide the system with error recovery capabilities.

In our opinion, it is clear that some sort of structured architecture with cooperative levels is needed in order to cope with a such a complex problem as the analysis of human motion.

# References

[1] J.K. Aggarwal and Q. Cai. Human Motion Analysis: A Review. *CVIU*, 73(3):428–440, 1999. (Cited on pages 1, 2, and 6)

[2] B. Anderson and J. Moore. *Optimal Filtering*. Prentice Hall, 1979. (Cited on page 15)

[3] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A Tutorial on PFs for On-line Non-linear/Non-Gaussian Bayesian Tracking. *Signal Processing*, 50(2):174–188, 2002. (Cited on page 16)

[4] Y. Bar-Shalom and T. Fortran. *Tracking and Data Association*. A. Press, 1988. (Cited on page 15)

[5] A. Bhattacharyya. On a Measure of Divergence Between Two Statistical Populations Defined by Probability Distributions. *Bull. Calcutta Math. Soc*, 35:99–109, 1943. (Cited on page 17)

[6] C. Bibby and I. Reid. Visual Tracking at Sea. In *International Conference in Robotics and Automation*, pages 1841– 1846. IEEE, 2005. (Cited on page 7)

[7] C. Bregler. Learning and Recognising Human Dynamics in Video Sequences. In *CVPR, Puerto Rico*, pages 568–574. IEEE, 1997. (Cited on page 14)

[8] R. Collins. Mean-shift Blob Tracking through Scale Space. In *CVPR, Madison, WI, USA*, volume 2, pages 234–240. IEEE, 2003. (Cited on page 19)

[9] R. Collins, A. Lipton, and T. Kanade. A System for Video Surveillance and Monitoring. In *8th International Topical Meeting on Robotics and Remote Systems, Pittsburgh, USA*, pages 1–15. American Nuclear Society, 1999. (Cited on page 12)

[10] R. Collins, Y. Liu, and M. Leordeanu. Online Selection of Discriminative Tracking Features. *PAMI*, 27(10):1631–1643, 2005. (Cited on pages 19 and 20)

[11] R.T. Collins, A.J. Lipton, and T. Kanade. A System for Video Surveillance and Monitoring: VSAM Final Report. Technical Report TR00-12, CMU, 2000. (Cited on pages 12 and 14)

[12] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based Object Tracking. *PAMI*, 25(5):564–577, 2003. (Cited on page 19)

[13] N. de Freitas, A. Gee M. Niranjan, and A. Doucet. Sequential Monte Carlo Methods for Optimisation of Neural Network Models. Technical Report TR 328, Cambridge University, 1998. (Cited on page 16)

[14] J. Deutscher and I. Reid. Articulated Body Motion Capture by Stochastic Search. *IJCV*, 61(2):185–205, 2005. (Cited on pages 17 and 18)

[15] A. Doucet. On Sequential Simulation-Based Methods for Bayesian Filtering. Technical Report TR310, Cambridge University, 1998. (Cited on page 16)

[16] G.D. Finlayson, S.D. Hordley, C. Lu, and M.S. Drew. On the removal of shadows from images. *Pattern Analysis and Machine Intelligence*, 28(1):59–68, 2006. (Cited on page 12)

[17] D. M. Gavrila. The Visual Analysis of Human Movement: A Survey. *CVIU*, 73(1):82–98, 1999. (Cited on pages 1, 2, 3, and 6)

[18] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-gaussian Bayesian state estimation. In *IEE Proceedings-F*, volume 140, pages 107–113, 1993. (Cited on page 16)

[19] I. Haritaoglu, D. Harwood, and L. Davis. W4: real-time surveillance of people and their activities. *PAMI*, 22(8):809–830, 2000. (Cited on pages 9, 10, and 14)

[20] I. Haritaoglu, D. Harwood, and L.S. Davis. W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People. In *Third International Conference on Automatic Gesture and Face Recognition, Nara, Japan*, pages 222–227. IEEE, 1998. (Cited on page 9)

[21] C. G. Harris and M. Stephens. A Combined Corner and Edge Detector. In *4th Alvey Vision Conf. Manchester, UK*, pages 147–151, 1988. (Cited on page 7)

[22] J. Heikkila and O. Silven. A real-time system for monitoring of cyclists and pedestrians. In *2nd Workshop on Visual Surveillance, Washington DC, USA*, pages 74–81. IEEE, 1999. (Cited on page 9)

[23] T. Horprasert, I. Haritaoglu, D. Harwood, L. Davis, C. Wren, and A. Pentland. Real-Time 3D Motion Capture. In *2nd Workshop Perceptual Interfaces*, 1998. (Cited on page 10)

[24] T. Horprasert, D. Harwood, and L. Davis. A Robust Background Subtraction and Shadow Detection. In *4th ACCV, Taipei, Taiwan*, volume 1, pages 983–988, 2000. (Cited on pages 10, 11, and 14)

[25] M. Isard and A. Blake. Contour Tracking by Stochastic Propagation of Conditional Density. In *4th ECCV, Cambridge UK*, volume 1, pages 343–356. Springer-Verlang, 1996. (Cited on page 17)

[26] M. Isard and A. Blake. A Mixed State Condensation Tracker with Automatic Model Switching. In *6th ICCV, Bombay, India*, pages 107–112, 1998. (Cited on page 17)

[27] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *IJCV*, 29(1):5–28, 1998. (Cited on pages 17 and 21)

[28] M. Isard and A. Blake. Icondensation Unifying Low-level and High-level Tracking in a Stochastic Framework. In *5th ECCV, Freiburg, Germany*, volume 1, pages 893–908, 1998. (Cited on page 20)

[29] M. Isard and J. MacCormick. BraMBLe: A Bayesian Multiple-Blob Tracker. In *8th ICCV, Vancouver, Canada*, volume 2, pages 34–41. IEEE, 2001. (Cited on page 19)

[30] O. Javed, K. Shafique, and M. Shah. A hierarchical approach to robust background subtraction using color and gradient information. In *Workshop on Motion and Video Computing*, pages 22–27. IEEE, 2002. (Cited on pages 12 and 14)

[31] S. Julier and J. Uhlmann. A New Extension of the Kalman Filter to Nonlinear Systems. In *11th AeroSense, Orlando, Florida*, volume 3068, pages 182–193, 1997. (Cited on page 15)

[32] R. Kahn, M. Swain, P. Prokopowicz, and R. Firby. Gesture Recognition Using the Perseus Architecture. In *CVPR, San Francisco, USA*, pages 734–741. IEEE, 1996. (Cited on pages 7 and 21)

[33] R. E. Kahn, M. J. Swain, P.N. Prokopowicz, and R.J. Firby. Real-time Gesture Recognition with the Perseus System. Technical Report TR96-04, University of Chicago, 1996. (Cited on page 21)

[34] R. Kalman. A New Approach to Linear Filtering and Prediction Problems. *ASME–Journal of Basic Engineering*, 82(D):35–45, 1960. (Cited on page 15)

[35] M. Karaman, L. Goldmann, D. Yu, and T. Sikora. Comparison of static background segmentation methods. In *Visual Communications and Image Processing*, volume 5960, pages 2140–2151. SPIE, 2005. (Cited on page 9)

[36] O. King and D. Forsyth. How Does CONDENSATION Behave with a Finite Number of Samples? In *6th ECCV, Ireland*, volume 1, pages 695–709, 2000. (Cited on page 16)

[37] E.B. Koller-Meier and F. Ade. Tracking Multiple Objects Using the Condensation Algorithm. *Robotics and Autonomous Systems*, 34:93–105, 2001. (Cited on page 17)

[38] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. (Cited on page 7)

[39] J. MacCormick and A. Blake. A Probabilistic Exclusion Principle for Tracking Multiple Objects. In *7th ICCV, Kerkyra, Greece*, volume 1, pages 572–578. IEEE, 1999. (Cited on page 17)

[40] J. MacCormick and A. Blake. A Probabilistic Exclusion Principle for Tracking Multiple Objects. *IJCV*, 39(1):57–71, 2000. (Cited on page 17)

[41] J. MacCormick and M. Isard. Partioned Sampling, Articulated Objects, and Interface-quality Hand Tracking. In *6th ECCV, Dublin, Ireland*, volume 2, pages 3–19. Springer-Verlang, 2000. (Cited on page 17)

[42] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *CVIU*, 80(1):42–56, 2000. (Cited on pages 7, 11, and 14)

[43] T. Moeslund, A. Hilton, and V. Krüger. A Survey of Advances in Vision-Based Human Motion Capture and Analysis. *CVIU*, 104:90–126, 2006. (Cited on pages 1 and 3)

[44] T. B. Moeslund and Erik Granum. A Survey of Computer Vision-Based Human Motion Capture. *CVIU*, 81(3):231–268, 2001. (Cited on pages 1, 3, 4, and 6)

[45] K. Nummiaro, E. Koller-Meier, and L. Van Gool. An Adaptive Color-Based Particle Filter. *Image and Vision Computing*, 21(1):99–110, 2003. (Cited on pages 17 and 18)

[46] A. Pentland. Looking at people: Sensing for Ubiquitous and Wearable Computing. *PAMI*, 22(1):107–119, 2000. (Cited on page 5)

[47] M. Piccardi. Background subtraction techniques: a review. In *International Conference on Systems, Man and Cybernetics*, volume 4, pages 3099–3104. IEEE, 2004. (Cited on page 9)

[48] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based Probabilistic Tracking. In *7th ECCV, Copenhaguen, Denmark*, pages 661–675. Springer-Verlang, 2002. (Cited on page 18)

[49] A. Senior, A. Hampapur, Y.L. Tian, L. Brown, S. Pankanti, and R. Bolle. Appearance Models for Occlusion Handling. *Image and Vision Computing*, 24:1233–1243, 2006. (Cited on pages 7 and 20)

[50] J. Shen. Motion detection in color image sequence and shadow elimination. In *Visual Communications and Image Processing, California, USA*, volume 5308, pages 731–740. SPIE, 2004. (Cited on pages 7, 13, and 14)

[51] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR, Fort Collins, CO, USA*, volume 2, pages 246–252. IEEE, 1999. (Cited on pages 12 and 14)

[52] J. Sullivan, A. Blake, M. Isard, and J. MacCormick. Bayesian Object Localisation in Images. *International Journal of Computer Vision*, 44(2):111–135, 2001. (Cited on page 17)

[53] R. van der Merwe, N. de Freitas, A. Doucet, and E. Wan. The Unscented Particle Filter. Technical Report TR380, Cambridge University, 2000. (Cited on page 16)

[54] E. Wan and R. van der Merwe. The Unscented Kalman Filter for Nonlinear Estimation. In *Adaptive Systems for Signal Processing, Communication and Control, Lake Louise, Canada*, pages 153–158. IEEE, 2000. (Cited on page 15)

[55] L. Wang, W. Hu, and T. Tan. Recent Developments in Human Motion Analysis. *Pattern Recognition*, 36(3):585–601, 2003. (Cited on pages 1, 4, 5, and 6)

[56] C. R. Wren, A. Azarbayejani, T. Darrell, and A.Pentland. Pfinder: Real-Time Tracking of the Human Body. *PAMI*, 19(7):780–785, 1997. (Cited on pages 7, 9, 10, and 14)

[57] Y. Wu, T. Yu, and G. Hua. Tracking Appearances with Occlusions. In *CVPR, Wisconsin, USA*, volume 1, pages 789–795. IEEE, 2003. (Cited on page 19)

[58] T. Yang, S. Li, Q. Pan, and J. Li. Real-time Multiple Object Tracking with Occlusion Handling in Dynamic Scenes. In *CVPR, San Diego, USA*, volume 1, pages 970–975. IEEE, 2005. (Cited on page 21)

[59] T. Zhao and R. Nevatia. Tracking Multiple Humans in Complex Situations. *PAMI*, 26(9):1208–1221, 2004. (Cited on page 21)