Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis

Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis

David A. Forsyth
University of Illinois Urbana Champaign

Okan Arikan
University of Texas at Austin

Leslie Ikemoto
University of California, Berkeley

James O'Brien
University of California, Berkeley

Deva Ramanan
Toyota Technological Institute at Chicago

now
the essence of knowledge

Boston – Delft

Foundations and Trends® in Computer Graphics and Vision

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

A Cataloging-in-Publication record is available from the Library of Congress

The preferred citation for this publication is D.A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, D. Ramanan, Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis, Foundation and Trends® in Computer Graphics and Vision, vol 1, no 2/3, pp 77–254, 2005

Printed on acid-free paper

ISBN: 1-933019-72-7

Foundations and Trends® in Computer Graphics and Vision

Volume 1 Issue 2/3, 2005

Editorial Scope

Foundations and Trends® in Computer Graphics and Vision will publish survey and tutorial articles in the following topics:

- Rendering: Lighting models; Forward rendering; Inverse rendering; Image-based rendering; Non-photorealistic rendering; Graphics hardware; Visibility computation
- Shape: Surface reconstruction; Range imaging; Geometric modelling; Parameterization;
- Mesh simplification
- Animation: Motion capture and processing; Physics-based modelling; Character animation
- Sensors and sensing
- Image restoration and enhancement
- Segmentation and grouping
- Feature detection and selection
- Color processing
- Texture analysis and synthesis
- Illumination and reflectance modeling
- Shape Representation
- Tracking
- Calibration
- Structure from motion
- Motion estimation and registration
- Stereo matching and reconstruction
- 3D reconstruction and image-based modeling
- Learning and statistical methods
- Appearance-based matching
- Object and scene recognition
- Face detection and recognition
- Activity and gesture recognition
- Image and Video Retrieval
- Video analysis and event recognition
- Medical Image Analysis
- Robot Localization and Navigation

Information for Librarians

Foundations and Trends® in Computer Graphics and Vision, 2005, Volume 1, 4 issues. ISSN paper version 1572-2740. ISSN online version 1572-2759. Also available as a combined paper and online subscription.

Foundations and Trends® in
Computer Graphics and Vision
Vol. 1, No 2/3 (2005) 77–254
© 2006 D.A. Forsyth, O. Arikan, L. Ikemoto,
J. O'Brien, D. Ramanan
DOI: 10.1561/0600000005

Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis

David A. Forsyth[1], Okan Arikan[2], Leslie Ikemoto[3], James O'Brien[4] and Deva Ramanan[5]

[1] *University of Illinois Urbana Champaign*
[2] *University of Texas at Austin*
[3] *University of California, Berkeley*
[4] *University of California, Berkeley*
[5] *Toyota Technological Institute at Chicago*

Abstract

We review methods for kinematic tracking of the human body in video. The review is part of a projected book that is intended to cross-fertilize ideas about motion representation between the animation and computer vision communities. The review confines itself to the earlier stages of motion, focusing on tracking and motion synthesis; future material will cover activity representation and motion generation.

In general, we take the position that tracking does not necessarily involve (as is usually thought) complex multimodal inference problems. Instead, there are two key problems, both easy to state.

The first is lifting, where one must infer the configuration of the body in three dimensions from image data. Ambiguities in lifting can result in multimodal inference problem, and we review what little is known about the extent to which a lift is ambiguous. The second is data association, where one must determine which pixels in an image

come from the body. We see a tracking by detection approach as the most productive, and review various human detection methods.

Lifting, and a variety of other problems, can be simplified by observing temporal structure in motion, and we review the literature on data-driven human animation to expose what is known about this structure. Accurate generative models of human motion would be extremely useful in both animation and tracking, and we discuss the profound difficulties encountered in building such models. Discriminative methods – which should be able to tell whether an observed motion is human or not – do not work well yet, and we discuss why.

There is an extensive discussion of open issues. In particular, we discuss the nature and extent of lifting ambiguities, which appear to be significant at short timescales and insignificant at longer timescales. This discussion suggests that the best tracking strategy is to track a 2D representation, and then lift it. We point out some puzzling phenomena associated with the choice of human motion representation – joint angles vs. joint positions. Finally, we give a quick guide to resources.

Contents

1 Tracking: Fundamental Notions **1**

1.1 General observations 2
1.2 Tracking by detection 5
1.3 Tracking using flow 14
1.4 Tracking with probability 26

2 Tracking: Relations between 3D and 2D **33**

2.1 Kinematic inference with multiple views 34
2.2 Lifting to 3D 38
2.3 Multiple modes, randomized search and human tracking 49

3 Tracking: Data Association for Human Tracking **61**

3.1 Detecting humans 62
3.2 Tracking by matching revisited 74
3.3 Evaluation 85

4 Motion Synthesis **89**

4.1 Fundamental notions 90
4.2 Motion signal processing 101

4.3 Motion graphs 106
4.4 Motion primitives 116
4.5 Enriching a motion collection 121

5 Discussion 133

5.1 Representations 133
5.2 Generalization 145
5.3 Resources 149

Acknowledgements 151

References 153

1

Tracking: Fundamental Notions

In a tracking problem, one has some measurements that appear at each tick of a (notional) clock, and, from these measurements, one would like to determine the state of the world. There are two important sources of information. First, measurements constrain the possible state of the world. Second, there are dynamical constraints – the state of the world cannot change arbitrarily from time to time. Tracking problems are of great practical importance. There are very good reasons to want to, say, track aircraft using radar returns (good summary histories include [51, 53, 188]; comprehensive reviews of technique in this context include [32, 39, 127]).

Not all measurements are informative. For example, if one wishes to track an aircraft – where state might involve pose, velocity and acceleration variables, and measurements might be radar returns giving distance and angle to the aircraft from several radar aerials – some of the radar returns measured might not come from the aircraft. Instead, they might be the result of noise, of other aircraft, of strips of foil dropped to confuse radar apparatus (**chaff** or **window**; see [188]), or of other sources. The problem of determining which measurements are informative and which are not is known as **data association**.

Data association is the dominant difficulty in tracking objects in video. This is because so few of the very many pixels in each frame lie on objects of interest. It can be spectacularly difficult to tell which pixels in an image come from an object of interest and which do not. There are a very wide variety of methods for doing so, the details of which largely depend on the specifics of the application problem. Surprisingly, data association is not usually explicitly discussed in the computer vision tracking literature. However, whether a method is useful rests pretty directly on its success at data association – differences in other areas tend not to matter all that much in practice.

1.1 General observations

The literature on tracking people is immense. Furthermore, the problem has quite different properties depending on precisely what kind of representation one wishes to recover. The most important variable appears to be spatial scale. At a **coarse scale**, people are **blobs**. For example, we might view a plaza from the window of a building or a mall corridor from a camera suspended from the ceiling. Each person occupies a small block of pixels, perhaps 10–100 pixels in total. While we should be able to tell where a person is, there isn't much prospect of determining where the arms and legs are. At this scale, we can expect to recover representations of **occupancy** – where people spend time, for example [424] – or of **patterns of activity** – how people move from place to place, and at what time, for example [377].

At a **medium scale**, people can be thought of as blobs with attached **motion fields**. For example, a television program of a soccer match, where individuals are usually 50–100 pixels high. In this case, one can tell where a person is. Arms and legs are still difficult to localize, because they cover relatively few pixels, and there is motion blur. However, the motion fields around the body yield some information as to how the person is moving. One could expect to be able to tell where a runner is in the phase of the run from this information – are the legs extended away from the body, or crossing?

At a **fine scale**, the arms and legs cover enough pixels to be detected, and one wants to report the configuration of the body.

We usually refer to this case as **kinematic tracking**. At a fine spatial scale, one may be able to report such details as whether a person is picking up or handling an object. There are a variety of ways in which one could encode and report configuration, depending on the model adopted – is one to report the configuration of the arms? the legs? the fingers? – and on whether these reports should be represented in 2D or in 3D. We will discuss various representations in greater detail later.

Each scale appears to be useful, but there are no reliable rules of thumb for determining what scale is most useful for what application. For example, one could see ways to tell whether people are picking up objects at a coarse scale. Equally, one could determine patterns of activity from a fine scale. Finally, some quite complex determinations about activity can be made at a surprisingly coarse scale. Tracking tends to be much more difficult at the fine scale, because one must manage more degrees of freedom and because arms and legs can be small, and can move rather fast.

In this review, we focus almost entirely on the fine scale; even so, space will not allow detailed discussion of all that has been done. Our choice of scale is dictated by the intuition that good fine-scale tracking will be an essential component of any method that can give general reports on what people are doing in video. There are distinctive features of this problem that make fine scale tracking difficult:

- **State dimension:** One typically requires a high dimensional state vector to describe the configuration of the body in a frame. For example, assume we describe a person using a 2D representation. Each of ten body segments (torso, head, upper and lower arms and legs) will be represented by a rectangle of fixed size (that differs from segment to segment). This representation will use an absolute minimum of 12 state variables (position and orientation for one rectangle, and relative orientation for every other). A more practical version of the representation allows the rectangles to slide with respect to one another, and so needs 27 state variables. Considerably more variables are required for 3D models.

- **Nasty dynamics:** There is good evidence that such motions as walking have predictable, low-dimensional structure [335, 351]. However, the body can move extremely fast, with large accelerations. These large accelerations mean that one can stop moving predictably very quickly – for example, jumping in the air during a walk. For straightforward mechanical reasons, the body parts that move fastest tend to be small and on one end of a long lever which has big muscles at the other end (forearms, fingers and feet, for example). This means that the body segments that the dynamical model fails to predict are going to be hard to find because they are small. As a result, accurate tracking of forearms can be very difficult.
- **Complex appearance phenomena:** In most applications one is tracking clothed people. Clothing can change appearance dramatically as it moves, because the forces the body applies to the clothing change, and so the pattern of folds, caused by buckling, changes. There are two important results. First, the pattern of occlusions of texture changes, meaning that the apparent texture of the body segment can change. Second, each fold will have a typical shading pattern attached, and these patterns move in the image as the folds move on the surface. Again, the result is that the apparent texture of the body segment changes. These effects can be seen in Figure 1.4.
- **Data association:** There is usually no distinctive color or texture that identifies a person (which is why people are notoriously difficult to find in static images). One possible cue is that many body segments appear at a distinctive scale as extended regions with rather roughly parallel sides. This isn't too helpful, as there are many other sources of such regions (for example, the spines of books on a shelf). Textured backgrounds are a particularly rich source of false structures in edge maps. Much of what follows is about methods to handle data association problems for people tracking.

1.2 Tracking by detection

Assume we have some form of template that can detect objects reasonably reliably. A good example might be a face detector. Assume that faces don't move all that fast, and there aren't too many in any given frame. Furthermore, the relationship between our representation of the state of a face and the image is uncomplicated. This occurs, for example, when the faces we view are always frontal or close to frontal. In this case, we can represent the state of the face by what it looks like (which, in principle, doesn't change because the face is frontal) and where it is.

Under these circumstances, we can build a tracker quite simply. We maintain a pool of tracks. We detect all faces in each incoming frame. We match faces to tracks, perhaps using an appearance model built from previous instances and also – at least implicitly – a dynamical model. This is where our assumptions are important; we would like faces to be sufficiently well-spaced with respect to the kinds of velocities we expect that there is seldom any ambiguity in this matching procedure. This matching procedure should not require one-one matches, meaning that some tracks may not receive a face, and some faces may not be allocated a track. For every face that is not attached to a track, we create a new track. Any track that has not received a face for several frames is declared to have ended (Algorithm 1 breaks out this approach).

This basic recipe for tracking by detection is worth remembering. In many situations, nothing more complex is required, and the recipe is used without comment in a variety of papers. As a simple example, at coarse scales and from the right view, background subtraction and looking for dark blobs of the right size is sufficient to identify human heads. Yan and Forsyth use this observation in a simple track-by-detection scheme, where heads are linked across frames using a greedy algorithm [424]. The method is effective for obtaining estimates of where people go in public spaces.

The method will need some minor improvements and significant technical machinery as the relationship between state and image measurements grows more obscure. However, in this simple form, the

Assumptions: We have a detector which is reasonably reliable for all aspects that matter. Objects move relatively slowly with respect to the spacing of detector responses. As a result, a detector response caused either by another object or by a false positive tends to be far from the next true position of our object.

First frame:
Create a track for each detector response.

N'th frame:
Link tracks and detector responses. Typically, each track gets the closest detector response if it is not further away than some threshold. If the detector is capable of reporting some distinguishing feature (colour, texture, size, etc.), this can be used too.
Spawn a new track for each detector response not allocated to a track.
Reap any track that has not received a measurement for some number of frames.

Cleanup: We now have trajectories in space time. Link any where this is justified (perhaps by a more sophisticated dynamical or appearance model, derived from the candidates for linking).

Algorithm 1: *The simplest tracking by detection*

method gives some insight into general tracking problems. The trick of creating tracks promiscuously and then pruning any track that has not received a measurement for some time is a quite general and extremely effective trick. The process of linking measurements to tracks is the aspect of tracking that will cause us the most difficulty (the other aspect, inferring states from measurements, is straightforward though technically involved). This process is made easier if measurements have features that distinctively identify the track from which they come. This can occur because, for example, a face will not change gender from frame to frame, or because tracks are widely spaced with respect to the largest practical speed (so that allocating a measurement to the closest track is effective).

All this is particularly useful for face tracking, because **face detection** – determining which parts of an image contain human faces, without reference to the individual identity of the faces – is one of the substantial successes of computer vision. Neither space nor energy allow a comprehensive review of this topic here. However, the typical approach is: One searches either rectangular or circular image windows over translation, scale and sometimes rotation; corrects illumination within these windows by methods such as histogram equalization; then presents these windows to a classifier which determines whether a face is present or not. There is then some post-processing on the classifier output to ensure that only one detect occurs at each face. This general picture appears in relatively early papers [299, 331, 332, 382, 383]. Points of variation include: the details of illumination correction; appropriate search mechanisms for rotation (cf. [334] and [339]); appropriate classifiers (cf. [259, 282, 333, 339] and [383]); building an incremental classification procedure so that many windows are rejected early and so consume little computation (see [186, 187, 407, 408] and the huge derived literature). There are a variety of strategies for detecting faces using parts, an approach that is becoming increasingly common (compare [54, 173, 222, 253, 256] and [412]; faces are becoming a common category in so-called object category recognition, see, for example, [111]).

1.2.1 Background subtraction

The simplest detection procedure is to have a good model of the background. In this case, everything that doesn't look like the background is worth tracking. The simplest background subtraction algorithm is to take an image of the background and then subtract it from each frame, thresholding the magnitude of the difference (there is a brief introduction to this area in [118]). Changes in illumination will defeat this approach. A natural improvement is to build a moving average estimate of the background, to keep track of illumination changes (e.g. see [343, 417]; gradients can be incorporated [250]). In outdoor scenes, this approach is defeated by such phenomena as leaves moving in the wind. More sophisticated background models keep track of maximal and minimal values at each pixel [146], or build local statistical models at each pixel [59, 122, 142, 176, 177, 375, 376].

Under some circumstances, background subtraction is sufficient to track people and perform a degree of kinematic inference. Wren *et al.* describe a system, Pfinder, that uses background subtraction to identify body pixels, then identifies arm, torso and leg pixels by building "blobby" clusters [417]. Haritaoglu *et al.* describe a system called W4, which uses background subtraction to segment people from an outdoor view [146]. Foreground regions are then linked in time by applying a second order dynamic model (velocity and acceleration) to propagate median coordinates (a robust estimate of the centroid) forward in time. Sufficiently close matches trigger a search process that matches the relevant foreground component in the previous frame to that in the current frame. Because people can pass one another or form groups, foreground regions can merge, split or appear. Regions appearing, splitting or merging are dealt with by creating (resp. fusing) tracks. Good new tracks can be distinguished from bad new tracks by looking forward in the sequence: a good track continues over time. Allowing a tracker to create new tracks fairly freely, and then telling good from bad by looking at the future in this way is a traditional, and highly useful, trick in the radar tracking community (e.g. see the comprehensive book by Blackman and Popoli [39]). The background subtraction scheme is fairly elaborate, using a range of thresholds to obtain a good blob (Figure 1.1). The resulting blobs are sufficiently good that the contour can be parsed to yield a decomposition into body segments. The method then segments the contours using convexity criteria, and tags the segments using: distance to the head – which is at the top of the contour; distance to the feet – which are at the bottom of the contour; and distance to the median – which is reasonably stable. All this works because, for most configurations of the body, one will encounter body segments in the same order as one walks around the contour (Figure 1.2). Shadows are a perennial nuisance for background subtraction, but this can be dealt with using a stereoscopic reconstruction, as Haritaoglu *et al.* show ([147]; see also [178]).

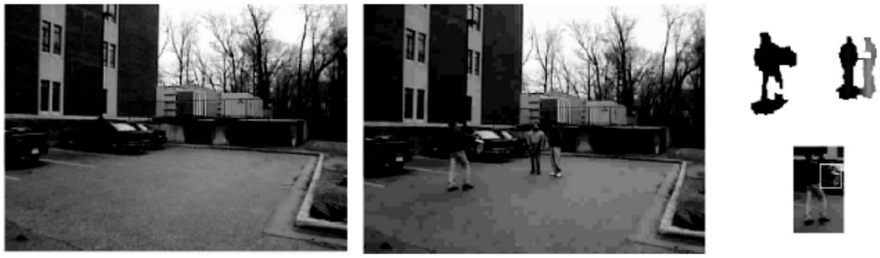

Fig. 1.1 *Background subtraction identifies groups of pixels that differ significantly from a background model. The method is most useful for some some cases of surveillance, where one is guaranteed a fixed viewpoint and a static background changing slowly in appearance. On the* **left**, *a background model; in the* **center**, *a frame; and on the* **right**, *the resulting image blobs. The figure is taken from Haritaoglu* et al. *[146]; in this paper, authors use an elaborate method involving a combination of thresholds to obtain good blobs. Figure 1.2 illustrates a method due to these authors that obtains a kinematic configuration estimate by parsing the blob.* Figure from "W4: Real-time surveillance of people and their activities", Haritaoglu *et al.*, IEEE Trans. Pattern Analysis and Machine Intelligence, 2000, © 2000 IEEE.

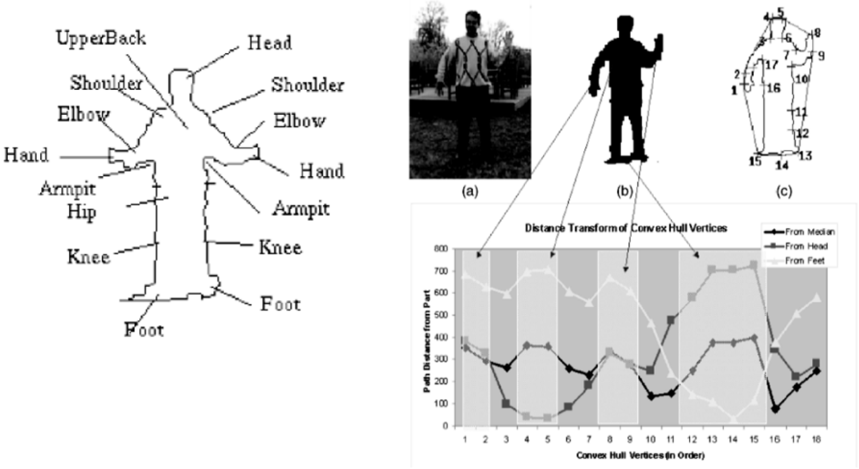

Fig. 1.2 *For a given view of the body, body segments appear in the outline in a predictable manner. An example for a frontal view appears on the* **left**. *Haritaoglu* et al *identify vertices on the outline of a blob using a form of convexity reasoning (***right (b)** *and* **right (c)***), and then infer labels for these vertices by measuring the distance to head (at the top), feet (at the bottom) and median (***below right***). These distances give possibly ambiguous labels for each vertex; by applying a set of topological rules obtained using examples of multiple views like that on the* **left**, *they obtain an unambiguous labelling.*Figure from "W4: Real-time surveillance of people and their activities", Haritaoglu *et al.*, IEEE Trans. Pattern Analysis and Machine Intelligence, 2000, © 2000 IEEE.

1.2.2 Deformable templates

Image appearance or **appearance** is a flexible term used to refer to aspects of an image that are being encoded and should be matched. Appearance models might encode such matters as: Edge position; edge orientation; the distribution of color at some scale (perhaps as a histogram, perhaps as histograms for each of some set of spatially localized buckets); or texture (usually in terms of statistics of filter outputs.

A **deformable template** or **snake** is a parametric model of image appearance usually used to localize structures. For example, one might have a template that models the outline of a squash [191, 192] or the outline of a person [33], place the template on the image in about the right place, and let a fitting procedure figure out the best position, orientation and parameters.

We can write this out formally as follows. Assume we have some form of template that specifies image appearance as a function of some parameters. We write this template – which gives (say) image brightness (or color, or texture, and so on) as a function of space $\mathbf{x}$ and some parameters θ – as $T(\mathbf{x}|\theta)$. We score a comparison between the image at frame n, which we write as $I(\mathbf{x}, t_n)$, and this template using the a scoring function ρ

$$\rho(T(\mathbf{x}|\theta), I(\mathbf{x}, t_n)).$$

A **point template** is built as a set of **active sites** within a model coordinate frame. These sites are to match **keypoints** identified in the image. We now build a model of acceptable sets of active sites obtained as shape, location, etc., changes. Such models can be built with, for example, the methods of **principal component analysis** (see, for example, [185]). We can now identify a match by obtaining image keypoints, building a correspondence between image keypoints and active sites on the template, and identifying parameters that minimize the fitting error.

An alternative is a **curve template**, an idea originating with the **snakes** of [191, 192]. We choose a parametric family of image curves – for example, a closed B-spline – and build a model of acceptable shapes, using methods like principal component analysis on the control points. There is an excellent account of methods in the book of Blake and Isard [41]. We can now identify a match by summing values of some image-based potential function over a set of sample points on the curve. A particularly important case occurs when we want the sample points to be close to image points where there is a strong feature response – say an edge point. It can be inconvenient to find every edge point in the image (a matter of speed) and this class of template allows us to search for edges only along short sections normal to the curve – an example of a **gate**.

Deformable templates have not been widely used as object detectors, because finding a satisfactory minimum – one that lies on the object of interest, most likely a global minimum – can be hard. The search is hard to initialize because one must identify the feature points that should lie within the gate of the template. However, in tracking problems this difficulty is mitigated if one has a dynamical model of some form. For example, the object might move slowly, meaning that the minimum for frame n will be a good start point for frame $n+1$. As another example, the object might move with a large, but near constant, velocity. This means that we can *predict* a good start point from frame $n+1$ given frame n. A significant part of the difficulty is caused by image features that don't lie on the object, meaning that another useful case occurs in the near absence of clutter – perhaps background subtraction, or the imaging conditions, ensures that there are few or no extra features to confuse the fitting process.

Baumberg and Hogg track people with a deformable template built using a B-spline as above, with principal components used to determine the template [33]. They use background subtraction to obtain an outline for the figure, then sample the outline. For this kind of template, correspondence is generally a nuisance, but in some practical applications, this information can be supplied from quite simple considerations. For example, Baumberg and Hogg work with background subtracted data of pedestrians at fairly coarse scales from fixed views [33]. In this case, sampling the outline at fixed fractions of length, and starting at the lowest point on the principal axis yields perfectly acceptable correspondence information.

1.2.2.1 Robustness

We have presented scoring a deformable template as a form of least squares fitting problem. There is a basic difficulty in such problems. Points that are dramatically in error, usually called **outliers** and traditionally blamed on typist error [153, 330], can be overweighted in determining the fit. Outliers in vision problems tend to be unavoidable, because nature is so generous with visual data that there is usually something seriously misleading in any signal. There are a variety of methods for managing difficulties created by outliers that are used in building deformable template trackers. An estimator is called **robust** if the estimate tends to be only weakly affected by outliers. For example, the average of a set of observations is not a robust estimate of the mean of their source (because if one observation is, say, mistyped, the average could be wildly incorrect). The median *is* a robust estimate, because it will not be much affected by the mistyped observation.

Gating – the scheme of finding edge points by searching out some distance along the normal from a curve – is one strategy to obtain robustness. In this case, one limits the distance searched. Ideally, there is only one edge point in the search window, but if there are more one takes the closest (strongest, *mutatis mutandis* depending on application details). If there is nothing, one accepts some fixed score, chosen to make the cost continuous. This means that the cost function, while strictly not differentiable, is not dominated by very distant edge points. These are not seen in the gate, and there is an upper bound on the error any one site can contribute.

An alternative is to use an **m-estimator**. One would like to score the template with a function of squared distance between site and measured point. This function should be close to the identity for small values (so that it behaves like the squared distance) and close to some constant for large values (so that large values don't contribute large biases). A natural form is

$$\rho(u) = \frac{u}{u+\sigma}$$

so that, for d^2 small with respect to σ, we have $\rho(d^2) \approx d^2$ and for d^2 large with respect to σ we have $\rho(d^2) \approx 1$. The advantage of this approach is that nearby edge points dominate the fit; the disadvantage is that even fitting problems that are originally convex are no longer convex when the strategy is applied. Numerical methods are consequently more complex, and one must use multiple start points. There is little hope of having a convex problem, because different start points correspond to different splits of the data set into "important" points and outliers; there is usually more than one such split. Again, large errors no longer dominate the estimation process, and the method is almost universally applied for flow templates.

1.2.2.2 The Hausdorff distance

The **Hausdorff distance** is a method to measure similarity between binary images (for example, edge maps; the method originates in Minkowski's work in convex analysis, where it takes a somewhat different form). Assume we have two sets of points P and Q; typically, each point is an edge point in an image. We define the Hausdorff distance between the two sets to be

$$H(P,Q) = max(h(P,Q), h(Q,P))$$

where

$$h(P,Q) = \max_{p \in P} \min_{q \in Q} \parallel p - q \parallel.$$

The distance is small if there is a point in Q close to each point in P and a point in P close to each point in P. There is a difficulty with robustness, as the Hausdorff distance is large if there are points with no good matches. In practice, one uses a variant of the Hausdorff distance (the **generalized Hausdorff distance**) where the distance used is the k-th ranked of the available distances rather than the largest. Define F_k^{th} to be the operator that orders the elements of its input largest to smallest, then takes the k'th largest. We now have

$$H_k(P,Q) = max(h_k(P,Q), h_k(Q,P))$$

where

$$h_k(P,Q) = F_k^{th}(\min_{q \in Q} \parallel p - q \parallel)$$

(for example, if there are $2n$ points in P, then $h_n(P,Q)$ will give the median of the minimum distances). The advantage of all this is that some large distances get ignored.

Now we can compare a template P with an image Q by determining some family of transformations $\mathcal{T}(\theta)$ and then choosing the set of parameters $\hat{\theta}$ that minimizes

$$H_k(\mathcal{T}(\theta) \circ P, Q).$$

This will involve some form of search over θ. The search is likely to be simplified if – as applies in the case of tracking – we have a fair estimate of $\hat{\theta}$ to hand.

Huttenlocher *et al.* track using the Hausdorff distance [165]. The template, which consists of a set of edge points, is itself allowed to deform. Images are represented by edge points. They identify the instance of the latest template in the next frame by searching over translations θ of the template to obtain the smallest value of $H_k(\mathcal{T}(\theta) \circ P, Q)$. They then translate the template to that location, and identify all edge points that are within some distance of the current template's edge points. The resulting points form the template for the next frame. This process allows the template to deform to take into account, say, the deformation of the body as a person moves. Performance in heavily textured video must depend on the extent to which the edge detection process suppresses edges and the setting of this distance parameter (a large distance and lots of texture is likely to lead to catastrophe).

1.3 Tracking using flow

The difficulty with tracking by detection is that one might not have a deformable template that fully specifies the appearance of an object. It is quite common to have a template that specifies the shape of the domain spanned by the object and the type of its transformation, but not what lies within. Typically, we don't know the pattern, but we do know how it moves. There are several important examples:

- **Human body segments** tend to look like a rectangle in any frame, and the motion of this rectangle is likely to be either Euclidean or affine, depending on imaging circumstances.

- **A face in a webcam** tends to fill a blob-like domain and undergo mainly Euclidean transformations. This is useful for those building user interfaces where the camera on the monitor views the user, and there are numerous papers dealing with this. The face is not necessarily frontal – computer users occasionally look away from their monitors – but tends to be large, blobby and centered.
- **Edge templates**, particularly those specifying outlines, are usually used because we don't know what the interior of the region looks like. Quite often, as we have seen, we know how the template can deform and move. However, we cannot score the interior of the domain because we don't know (say) the pattern of clothing being worn.

In each of these cases, we cannot use tracking by detection as above because we do not posess an appropriate template. As a matter of experience, objects don't change appearance much from frame to frame (alternatively, we should use the term appearance to apply to properties that don't change much from frame to frame). All this implies that parts of the previous image could serve as a template if we have a motion model and domain model. We could use a correspondence model to link pixels in the domain in frame n with those in the domain in frame $n+1$. A "good" linking should pair pixels that have similar appearances. Such considerations as camera properties, the motion of rigid objects, and computational expense suggest choosing the correspondence model from a small parametric family.

All this gives a formal framework. Write a pixel position in the n'th frame as $\mathbf{x}_n$, the domain in the n'th frame as $\mathcal{D}_n$, and the transformation from the n'th frame to the $n+1$'th frame as $\mathcal{T}_{n\to n+1}(\cdot;\theta_n)$. In this notation θ_n represent parameters for the transformation from the n'th frame to the $n+1$'th frame, and we have that $\mathbf{x}_{n+1} = \mathcal{T}_{n\to n+1}(\mathbf{x}_n;\theta_n)$.

We assume we know $\mathcal{D}_n$. We can obtain $\mathcal{D}_{n+1}$ from $\mathcal{D}_n$ as $\mathcal{T}_{n\to n+1}(\mathcal{D}_n;\theta_n)$. Now we can score the parameters θ_n representing the *change* in state between frames $n+1$ and n by comparing $\mathcal{D}_n$ with $\mathcal{D}_{n+1}$ (which is a function of θ_n). We compute some representation of image information $\mathbf{R}(\mathbf{x})$, and, within the domain $\mathcal{D}_{n+1}$ compare $\mathbf{R}(\mathbf{x}_{n+1})$ with $\mathbf{R}(\mathcal{T}_{n\to n+1}(\mathbf{x}_n;\theta_n))$, where the transformation is applied to the domain $\mathcal{D}_n$.

1.3.1 Optic flow

Generally, a frame-to-frame correspondence should be thought of as a **flow field** (or an **optic flow field**) – a vector field in the image giving local image motion at each pixel. A flow field is fairly clearly a correspondence, and a correspondence gives rise to a flow field (put the tail of the vector at the pixel position in frame n, and the head at the position in frame $n+1$). The notion of optic flow originates with Gibson (see, for example, [128]).

A useful construction in the optic flow literature assumes that image intensity is a continuous function of position and time, $I(\mathbf{x},t)$. We then assume that the intensity of image patches does not change with movement. While this assumption may run into troubles with illumination models, specularities, etc., it is not outrageous for small movements. Furthermore, it underlies our willingness to compare pixel values in frames. Accepting this assumption, we have

$$\frac{dI}{dt} = \nabla I \cdot \frac{d\mathbf{x}}{dt} + \frac{\partial I}{\partial t} = 0$$

(known as the **optic flow equation**, e.g. see [160]). Flow is represented by $d\mathbf{x}/dt$. This is important, because if we confine our attention to an appropriate domain, comparing $I(\mathcal{T}(\mathbf{x};\theta_n),t_{n+1})$ with $I(\mathbf{x},t_n)$ involves, in essence, estimating the total derivative. In particular,

$$I(\mathcal{T}(\mathbf{x};\theta_n),t_{n+1}) - I(\mathbf{x},t_n) \approx \frac{dI}{dt}.$$

Furthermore, the equivalence between correspondence and flow suggests a simpler form for the transformation of pixel values. We regard $\mathcal{T}(\mathbf{x};\theta_n)$ as taking $\mathbf{x}$ from the tail of a flow arrow to the head. At short timescales, this justifies the view that $\mathcal{T}(\mathbf{x};\theta_n) = \mathbf{x} + \delta\mathbf{x}(\theta_n)$.

1.3.2 Image stabilization

This form of tracking can be used to build boxes around moving objects, a practice known as **image stabilization**. One has a moving object on a fairly uniform background, and would like to build a domain such that the moving object is centered on the domain. This has the advantage that one can look at relative, rather than absolute, motion cues. For example, one might take a soccer player running around a field, and build a box around the player. If one then fixes the box and its contents in one place, the vast majority of motion cues within the box are cues to how the player's body configuration is changing. As another example, one might stabilize a box around an aerial view of a moving vehicle; now the box contains all visual information about the vehicle's identity.

Efros *et al.* use a straightforward version of this method, where domains are rectangles and flow is pure translation, to stabilize boxes around people viewed at a medium scale (for example, in a soccer video) [100]. In some circumstances, good results can be obtained by matching a rectangle in frame n with the rectangle in frame $n+1$ that has smallest sum-of-squared differences – which might be found by blank search, assisted perhaps by velocity constraints. This is going to work best if the background is relatively simple – say, the constant green of a soccer field – as then the background isn't a source of noise, so the figure need not be segmented (Figure 1.3). For more complex backgrounds, the approach may still work if one performs background subtraction before stabilization. At a medium scale it is very difficult to localize arms and legs, but they do leave traces in the flow field. The stabilization procedure means that the flow information can be computed with respect to a torso coordinate system, resulting in a representation that can be used to match at a kinematic level, without needing an explicit representation of arm and leg configurations (Figure 1.3).

1.3.3 Cardboard people

Flow based tracking has the advantage that one doesn't need an explicit model of the appearance of the template. Ju *et al.* build a model of legs in terms of a set of articulated rectangular patches ("cardboard people") [190]. Assume we have a domain D in the n'th image $I(\mathbf{x}, t_n)$

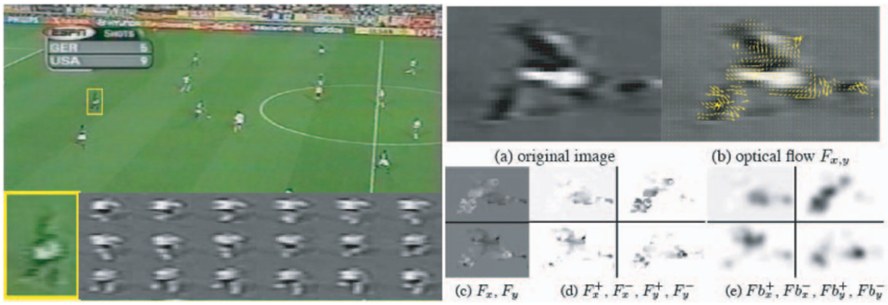
(a) original image
(b) optical flow $F_{x,y}$
(c) F_x, F_y
(d) $F_x^+, F_x^-, F_y^+, F_y^-$
(e) $Fb_x^+, Fb_x^-, Fb_y^+, Fb_y^-$

Fig. 1.3 *Flow based tracking can be useful for medium scale video. Efros* et al. *stabilize boxes around the torso of players in football video using a sum of squared differences (SSD) as a cost function and straightforward search to identify the best translation values. As the figure on the* **left** *shows, the resulting boxes are stable with respect to the torso. On the* **top right**, *larger versions of the boxes for some cases. Note that, because the video is at medium scale, it is difficult to resolve arms and legs, which are severely affected by motion blur. Nonetheless, one can make a useful estimate of what the body is doing by computing an estimate of optic flow (***bottom right**, F_x, F_y*), rectifying this estimate (***bottom right**, F_x^+, F_x^-, F_y^+, F_y^-*) and then smoothing the result (***bottom right**, Fb_x^+, *etc.). The result is a smoothed estimate of where particular velocity directions are distributed with respect to the torso, which can be used to match and label frames.* Figure from "Recognizing Action at a Distance", Efros *et al.*, IEEE Int. Conf. Computer Vision 2003, © 2003 IEEE.

and a flow field $\delta\mathbf{x}(\theta)$ parametrized by θ. Now this flow field takes D to some domain in the $n+1$'th image, and establishes a correspondence between pixels in the n'th and the $n+1$'th image. Ju *et al.* score

$$\sum_D \rho(I_{n+1}(\mathbf{x}+\delta\mathbf{x}(\theta)) - I_n(\mathbf{x}))$$

where ρ is some measure of image error, which is small when the two compare well and large when they are different. Notice that this is a very general approach to the tracking problem, with the difficulty that, unless one is careful about the flow model the problem of finding a minimum might be hard. To our knowledge, the image score is always applied to pixel values, and it seems interesting to wonder what would happen if one scored a difference in texture descriptors.

Typically, the score is not minimized directly, but is approximated with the optic flow equation and with a Taylor series. We have

$$\sum_D \rho(I(\mathbf{x}+\delta\mathbf{x}(\theta), t_{n+1}) - I_n(\mathbf{x}, t_n))$$

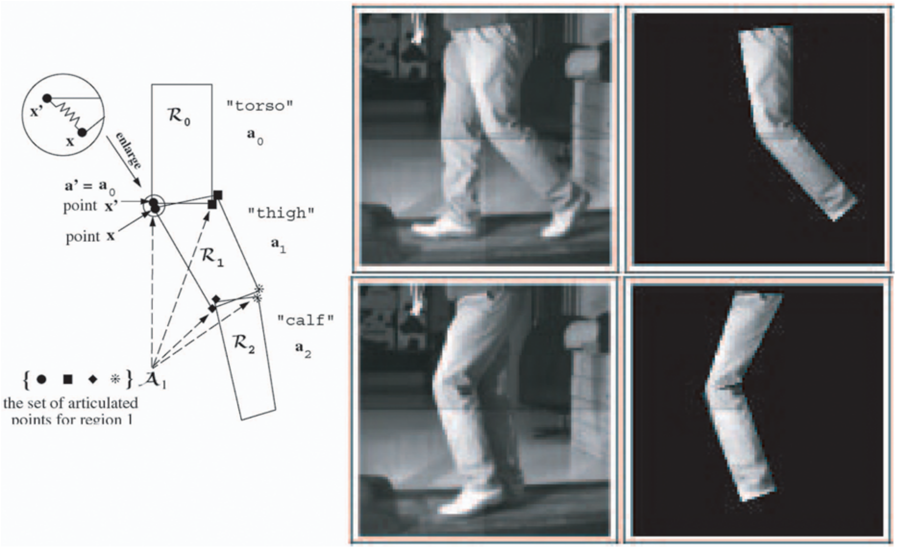

Fig. 1.4 *On the* **left**, *a 2D flow based model of a leg, called a "cardboard people" model by Ju* et al *[190]; there is a lower leg, an upper leg and a torso. Each domain is roughly rectangular, and the domains are coupled with an energy term to ensure they do not drift apart. The model is tracked by finding the set of deformation parameters that carve out a domain in the $n+1$'th frame that is most like the known domain in the n'th frame. On the* **right**, *two frames from a track, with the* **left column** *showing the original frame and the* **right column** *showing the track. Notice how the pattern of buckling folds on the trouser leg changes as the leg bends; this leads to quite significant changes in the texture and shading signal in the domain. These changes can be a significant nuisance.* Figure from "Cardboard People: A Parameterized Model of Articulated Image Motion", Ju *et al.*, IEEE Int. Conf. Face and Gesture, 1996, © 1996 IEEE.

is approximately equal to

$$\sum_D \rho(\frac{dI}{dt}) = \sum_D \rho(\frac{\partial I}{\partial x}\delta x(\theta_n) + \frac{\partial I}{\partial y}\delta y(\theta_n) + \frac{\partial I}{\partial t})$$

(this works because $\Delta t = 1$). Now assume that a patch has been marked out in a frame; then one can determine its configuration in the next by minimizing this error summed over the domain. The error itself is easily evaluated using smoothed derivative estimates. As we show below, we can further simplify error evaluation by building a flow model with convenient form. To track an articulated figure, Ju *et al.* attach a further term that encourages relevant vertices of each separate patch to stay close. Similarly, Black et al construct parametric families of flow fields and use them to track lips and legs, in the latter case yielding a satisfactory estimate of walk parameters [40]. In both cases, the flow model is view dependent. Yacoob and Davis build a view independent parametric flow field models to track views of walking humans [420]. As one would expect, this technique can be combined with others; for example, the W4S system of Haritaoglu *et al.* uses a "cardboard people" model to track torso configurations within the regions described above [147].

1.3.4 Building flow templates

We have seen how to construct tracks given parametric models of flow. But how is one to obtain good models? One strategy is to take a pool of examples of the types of flow one would like to track, and try to find a set of basis flows that explains most of the variation (for examples, see [190]). In this case, and writing θ_i for the i'th component of the parameter vector and $\mathbf{F}_i$ for the i'th flow basis vector field, one has

$$\delta\mathbf{x} = \sum_i \theta_i \mathbf{F}_i.$$

Now write ∇I for the image gradient and exploit the optic flow equation and a Taylor series as above. We get

$$\rho\left(\sum_i \theta_i((\nabla I)^T \mathbf{F}_i) + \frac{\partial I}{\partial t}\right).$$

As Ju *et al.* observe, this can be done with a singular value decomposition (and is equivalent to principal components analysis). A second strategy is to assume that flows involve what are essentially 2D effects – this is particularly appropriate for lateral views of human limbs – so that a set of basis flows that encodes translation, rotation and some affine effects is probably sufficient. One can obtain such flows by writing

$$\delta\mathbf{x} = \begin{pmatrix} u(\mathbf{x}) \\ v(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy \\ a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2 \end{pmatrix}.$$

This model is linear in the parameters (the a_i), which is convenient; it provides a reasonable encoding of flows resulting from 3D motions of a 2D rectangle (see Figure 1.5). One may also learn linearized flow models from example data [420].

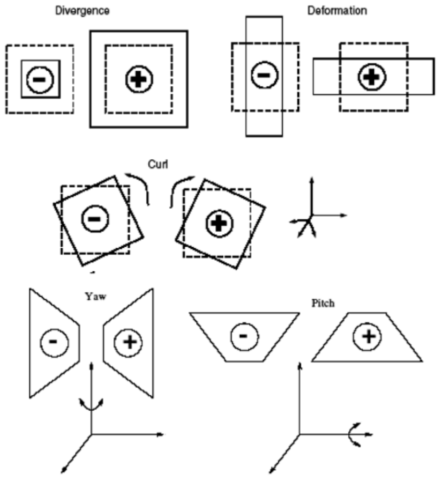

Fig. 1.5 *Typical flows generated by the model* $(u(\mathbf{x}), v(\mathbf{x})^T = (a_0 + a_1x + a_2y + a_6x^2 + a_7xy, a_3 + a_4x + a_5y + a_6xy + a_7y^2))$. *Different values of the* a_i *give different flows, and the model can generate flows typical of a 2D figure moving in 3D. We write* $\mathbf{a} = (a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7)$. **Divergence** *occurs when the image is scaled; for example,* $\mathbf{a} = (0,1,0,0,0,1,0,0)$. **Deformation** *occurs when one direction shrinks and another grows (for example, rotation about an axis parallel to the view plane in an orthographic camera); for example,* $\mathbf{a} = (0,1,0,0,0,-1,0,0)$. **Curl** *can result from in plane rotation; for example,* $\mathbf{a} = (0,0,-1,0,1,0,0,0)$. **Yaw** *models rotation about a vertical axis in a perspective camera; for example* $\mathbf{a} = (0,0,0,0,0,1,0)$. *Finally,* **pitch** *models rotation about a horizontal axis in a perspective camera; for example* $\mathbf{a} = (0,0,0,0,0,0,1)$. Figure from "Cardboard People: A Parameterized Model of Articulated Image Motion", Ju *et al.*, IEEE Int. Conf. Face and Gesture, 1996, © 1996 IEEE.

1.3.5 Flow models from kinematic models

An alternative method to build such templates is to work in 3D, and exploit the chain rule, as in the work of Bregler and Malik [49, 48]. We start with a segment in 3D, which is in some configuration and viewed with some camera. Each point on the segment produces some image value. We could think of the image values as a function – the **appearance map** – defined on the segment. This allows us to see viewing the segment as building a mapping from the points on the segment to the image domain. The image values are obtained by taking each point in the image, finding the corresponding point (if any) on the segment, and then evaluating the appearance map at this point.

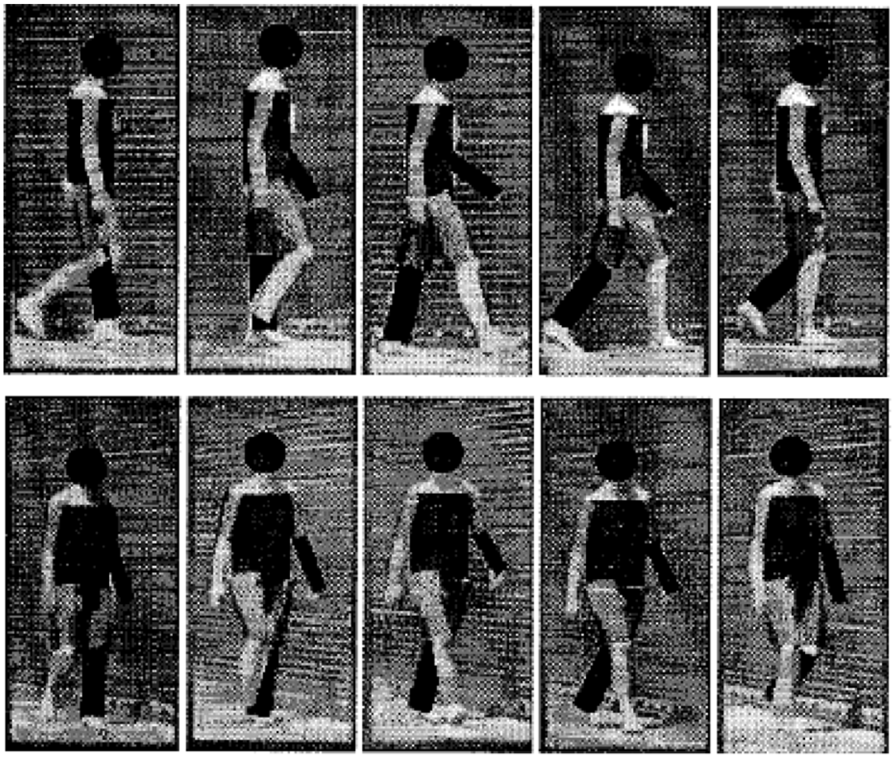

Fig. 1.6 *Bregler and Malik formulate parametric flow models by modelling a person as a kinematic chain and then differentiating the maps from segment to image [49]. They then track by searching for the parameter update that best aligns the current image pixels with those of the previous frame under this flow model. There is no dynamical model. This means that complex legacy footage, like these frames from the photographs of Eduard Muybridge [270, 269], can be tracked. Muybridge's frames are difficult to track because the frame-frame timing is not exact, and the figures can move in quite complex ways (see Figure 3.6).* Figure from "Tracking People with Twists and Exponential Maps", Bregler and Malik, Proc. Computer Vision and Pattern Recognition, 1998, © 1998 IEEE.

All this leads to an important formal model, again under the assumption that motions in 3D do not affect the appearance map in any significant way. We have a parametrized family of maps from points on the body to the image. A flow field in the image is a vector field induced by a change in the choice of parameters (caused by either a change in joint configuration or a camera movement). We will always assume that the change in parameters from frame to frame is small. At this point, we must introduce some notation. Write the map that takes points on the segment to points in the n'th image as $\mathcal{T}_{s\to I}(\cdot;\theta_n)$, where θ_n are parameters representing camera configuration, intrinsics, etc. The point $\mathbf{p}$ on the segment appears in image n at $\mathbf{x}_n = \mathcal{T}_{s\to I}(\mathbf{p};\theta_n)$ and in image $n+1$ at $\mathbf{x}_{n+1} = \mathcal{T}_{s\to I}(\mathbf{p};\theta_{n+1})$. The tail of the flow arrow is at $\mathbf{x}_n$ and the head is at $\mathbf{x}_{n+1}$. The change in parameters, $\Delta\theta = \theta_{n+1} - \theta_n$ is small. Then the flow is

$$\mathbf{x}_{n+1} - \mathbf{x}_n = \mathcal{T}_{s\to I}(\mathbf{p};\theta_{n+1}) - \mathcal{T}_{s\to I}(\mathbf{p};\theta_n) \approx \nabla_\theta \mathcal{T}_{s\to I} \cdot \Delta\theta$$

where the gradient, $\nabla_\theta \mathcal{T}_{s\to I}$, is evaluated at $(\mathbf{p},\theta_n)$.

1.3.5.1 Tracking a derivative flow model

The main point here is that the flow at $\mathbf{x}_n$ can be obtained by fixing the relevant point $\mathbf{p}$ on the object, then considering the map taking the *parameters* to the image plane – the derivative of $\mathcal{T}_{s\to I}(\mathbf{p};\cdot)$. This is important, because the flow $\nabla_\theta \mathcal{T}_{s\to I} \cdot \Delta\theta$ is a *linear* function of $\Delta\theta$. We now have the outline of a tracking algorithm:

- Start at frame $n = 0$ and some known configuration $\theta_0 = \hat{\theta}$.
- **Fit:** Fit the best value of $\Delta\theta$ to the flow between the frame n and frame $n+1$ using the flow model given by the derivative evaluated at θ_n.
- **Update:** Update the parameters by $\theta_{n+1} = \theta_n + \Delta\theta$ and set n to $n+1$.

This should be seen as a primitive integrator, using Euler's method and inheriting all the problems that come with it. This view confirms the reasonable suspicion that fast movements are unlikely to be tracked well unless that sampling rate is high.

1.3.5.2 The flow model from the chain rule

In the special case of segments lying on a **kinematic tree** – a series of links attached by joints of known parametric form, where there are no loops – the chain rule means that the derivative takes a special form. Each segment in a kinematic tree has its own coordinate system, and the joint is represented by a map from a link's world coordinate system to that of its parent. The parent of segment k is segment $k-1$. They are connected by a joint whose parameters at frame n are $\theta_{k,n}$. In general, in a kinematic tree, points on segments are affected by parameters at joints above them in the tree. Furthermore, we can obtain a transformation to the image by recursively concatenating transformations. Write the camera as $\mathcal{T}_{w\to i}$. Then the transformation taking a point of link k in frame n to the image can be written as

$$\mathcal{T}_{k\to i} = \mathcal{T}_{w\to i} \circ \mathcal{T}_{k-1\to w} \circ \mathcal{T}_{k\to k-1}.$$

Notice that the only transformation that depends on $\theta_{k,n}$ here is $\mathcal{T}_{k\to k-1}$.

There is an advantage to changing notation at this point. Write $\mathcal{T}_{k\to k-1}$ as $\mathcal{T}_k$. The root of the tree is at segment one, and we can write $\mathcal{T}_{1\to w}$ as $\mathcal{T}_1$ and $\mathcal{T}_{w\to i}$ as $\mathcal{T}_0$. We continue to divide up the parameters θ into components, $\theta_{k,n}$ being the components associated with segment k in the n'th frame ($\theta_{0,n}$ are viewing parameters in frame n). We can now see the map from point $\mathbf{p}$ on segment k to the image as

$$\mathcal{T}_{k\to i}(\mathbf{p};\theta) = \mathcal{T}_0(\mathcal{T}_1(\mathcal{T}_2(\ldots;\theta_2);\theta_1);\theta_0).$$

This is somewhat inconvenient to write out, and it is helpful to keep track of intermediate values. Introduce the notation

$$\mathbf{p}_l = \mathcal{T}_{k\to l}(\mathbf{p};\theta)$$

for the point $\mathbf{p}$ in the coordinate system of the l'th link.

Our transformations have two types of argument: the points in space, and the camera parameters. It is useful to distinguish between two types of derivative. Write the partial derivative of a transformation $\mathcal{T}$ *with respect to its spatial arguments* as $D\mathcal{T}$. In coordinates, $\mathcal{T}$ would take the form $(f_1(x_1,x_2,x_3,\theta), f_2(x_1,x_2,x_3,\theta), f_3(x_1,x_2,x_3,\theta))$, and this derivative would be the matrix whose i, j'th element is $\partial f_i/\partial x_j$. Similarly, write the partial derivative of a transformation $\mathcal{T}$ *with respect to parameters* θ as D_θ. If we regard θ as a vector of parameters whose j'th entry is θ_j, then in coordinates this derivative would be the matrix whose i, j'th element is $\partial f_i/\partial \theta_j$.

This orgy of notation leads to a simple form for the flow. Write the flow at point $\mathbf{x}$ – which is the image of point $\mathbf{p}$ on segment k – in frame n as $\mathbf{v}(\mathbf{x}, \theta_n)$. Then

$$\begin{aligned}\mathbf{v}(\mathbf{x}, \theta_n) = D_\theta \mathcal{T}_0(\mathbf{p}_0; \theta_0) \cdot \Delta\theta_0 + D_x \mathcal{T}_0 \circ D_\theta \mathcal{T}_1(\mathbf{p}_1; \theta_1) \Delta\theta_1 \\ \ldots + D_x \mathcal{T}_0 \circ D_x \mathcal{T}_1 \circ \ldots D_x \mathcal{T}_{k-1} \circ D_\theta \mathcal{T}_k(\mathbf{p}; \theta_k) \Delta\theta_k.\end{aligned}$$

Our indexing scheme hasn't taken into account the fact that we're dealing with a tree, but this doesn't matter; we need care only about links on the path from the relevant segment to the root. Furthermore, there is a relatively efficient algorithm for computing this derivative. We pass from the leaves to the root computing intermediate configurations $\mathbf{p}_l$ for each point $\mathbf{p}$ and the relevant parameter derivatives. We then pass from the root to the leaves concatenating spatial derivatives and summing.

1.3.5.3 Rigid-body transformations

All the above takes a convenient and simple form for rigid-body transformations (which are likely to be the main interest in human tracking). We use homogeneous coordinates to represent points in 3D, and so a rigid body transformation takes the form

$$\mathcal{T}(\mathbf{p}, \theta) = \begin{bmatrix} \mathcal{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \mathbf{p}$$

where $\mathcal{R}$ is an orthonormal matrix with determinant one (a rotation matrix). The parameters are the parameters of the rotation matrix and the coefficients of the vector $\mathbf{t}$. This means the spatial derivative is the same as the transformation, which is convenient.

The derivatives with respect to the parameters are also relatively easily dealt with. Recall the definition of the **matrix exponential** as an infinite sum,

$$\exp(\mathcal{M}) = \mathcal{I} + \mathcal{M} + \frac{1}{2}\mathcal{M}^2 + \mathcal{M}^3 + \ldots + \frac{1}{n}\mathcal{M}^n \ldots$$

where the sum exists. Now it is straightforward to demonstrate that if

$$\mathcal{M} = \begin{bmatrix} \mathcal{A} & \mathbf{t} \\ \mathbf{0} & 0 \end{bmatrix}$$

and if $\mathcal{A}$ is antisymmetric, then $\exp(\mathcal{M})$ is a rigid-body transformation. The elements of the antisymmetric matrix parametrize the rotation, and the rightmost column is the translation. This is useful, because

$$\frac{\partial\left(\exp \mathcal{M}(\theta)\right)}{\partial\theta} = \left(\frac{\partial \mathcal{M}(\theta)}{\partial\theta}\right) \exp \mathcal{M}(\theta)$$

which gives straightforward forms for the parameter derivatives.

1.4 Tracking with probability

It is convenient to see tracking as a probabilistic inference problem. In particular, we have a sequence of states $X_0, X_1, \ldots, X_N$ produced by some dynamical process. These states are unknown – they are sometimes called **hidden states** for this reason – but there are measurements $Y_0, Y_1, \ldots, Y_N$. Two problems follow naturally:

- **Tracking**, where we wish to determine some representation of $P(X_k | Y_0, \ldots, Y_k)$;
- **Filtering**, where we wish to determine some representation of $P(X_k | Y_0, \ldots, Y_N)$ (i.e. we get to use "future" measurements to infer the state).

These problems are massively simplified by two important assumptions.

- We assume measurements depend only the hidden state, that is, that $P(Y_k | X_0, \ldots, X_N, Y_0, \ldots, Y_N) = P(Y_k | X_k)$.
- We assume that the probability density for a new state is a function only of the previous state; that is, $P(X_k | X_0, \ldots, X_{k-1}) = P(X_k | X_{k-1})$, or, equivalently, that X_i form a **Markov chain**.

Now tracking involves three steps:

Prediction: where we construct some prediction of the future state given past measurements, or equivalently, construct a representation of $P(X_k | Y_0, \ldots, Y_{k-1})$. Straightforward manipulation of probability combined with the assumptions above yields that the **prior** or **predictive density** is

$$P(X_k | Y_0, \ldots, Y_{k-1}) = \int P(X_k | X_{k-1}) P(X_{k-1} | Y_0, \ldots, Y_{k-1}) dX_{k-1}.$$

Data association: where we use the predictive density – which is sometimes called the **prior** – and anything else likely to be helpful, to determine which of a pool of measurements contribute to the value of Y_k.

Correction: where we incorporate the new measurement into what is known, or, equivalently, construct a representation of $P(X_k|Y_0,\ldots,Y_k)$. Straightforward manipulation of probability combined with the assumptions above yields that the **posterior** is

$$P(X_k|Y_0,\ldots,Y_k) = \frac{P(Y_k|X_k)P(X_k|Y_0,\ldots,Y_{k-1})}{\int P(Y_k|X_k)P(X_k|Y_0,\ldots,Y_{k-1})dX_k}.$$

1.4.1 Linear dynamics and the Kalman filter

All this is much simplified in the case that the emission model is linear, the dynamic model is linear, and all noise is Gaussian. In this case, all densities are normal and the mean and covariance are sufficient to represent them. Both tracking and filtering boil down to maintenance of these parameters. There is a simple set of update rules (given in Algorithm 2; notation below), the **Kalman filter**.

Notation: We write $X \sim N(\mu;\Sigma)$ to mean that X is a normal random variable with mean μ and covariance Σ. Both dynamics and emission are linear, so we can write

$$X_k \sim N(\mathcal{A}_k X_{k-1};\Sigma_k^{(d)})$$

and

$$Y_k \sim N(\mathcal{B}_k X_k;\Sigma_k^{(m)}).$$

We will represent the mean of $P(X_i|y_0,\ldots,y_{i-1})$ as $\overline{X}_i^-$ and the mean of $P(X_i|y_0,\ldots,y_i)$ as $\overline{X}_i^+$ – the superscripts suggest that they represent our belief about X_i immediately before and immediately after the i'th measurement arrives. Similarly, we will represent the standard deviation of $P(X_i|y_0,\ldots,y_{i-1})$ as Σ_i^- and of $P(X_i|y_0,\ldots,y_i)$ as Σ_i^+. In each case, we will assume that we know $P(X_{i-1}|y_0,\ldots,y_{i-1})$, meaning that we know $\overline{X}_{i-1}^+$ and Σ_{i-1}^+.

Filtering is straightforward. We obtain a backward estimate by running the filter backward in time, and treat this as another

Dynamic Model:

$$\begin{aligned}\mathbf{x}_i &\sim N(\mathcal{D}_i\mathbf{x}_{i-1}, \Sigma_{d_i})\\ \mathbf{y}_i &\sim N(\mathcal{M}_i\mathbf{x}_i, \Sigma_{m_i})\end{aligned}$$

Start Assumptions: $\overline{\mathbf{x}}_0^-$ and Σ_0^- are known
Update Equations: Prediction

$$\begin{aligned}\overline{\mathbf{x}}_i^- &= \mathcal{D}_i\overline{\mathbf{x}}_{i-1}^+\\ \Sigma_i^- &= \Sigma_{d_i} + \mathcal{D}_i\sigma_{i-1}^+\mathcal{D}_i\end{aligned}$$

Update Equations: Correction

$$\begin{aligned}\mathcal{K}_i &= \Sigma_i^-\mathcal{M}_i^T\left[\mathcal{M}_i\Sigma_i^-\mathcal{M}_i^T + \Sigma_{m_i}\right]^{-1}\\ \overline{\mathbf{x}}_i^+ &= \overline{\mathbf{x}}_i^- + \mathcal{K}_i\left[\mathbf{y}_i - \mathcal{M}_i\overline{\mathbf{x}}_i^-\right]\\ \Sigma_i^+ &= \left[Id - \mathcal{K}_i\mathcal{M}_i\right]\Sigma_i^-\end{aligned}$$

Algorithm 2: *The Kalman filter updates estimates of the mean and covariance of the various distributions encountered while tracking a state variable of some fixed dimension using the given dynamic model.*

measurement. Extensive detail on the Kalman filter and derived methods appears in [32, 127].

1.4.2 Data association

Data association involves determining which pixels or image measurements should contribute to a track. That data association is a nuisance is a persistent theme of this work. Data association is genuinely difficult to handle satisfactorily – after all, determining which pixels contribute to which decision seems to be a core – and often very difficult – computer vision problem. The problem is usually particularly difficult when one wishes to track people, for several reasons. First, standard data association techniques aren't really all that much help, as for almost every aspect the image domain covered by a person changes shape very aggressively, and can do so very fast. Second, there seem to be a lot of background objects that look like some human body parts; for example, kinematic tracking of humans in office scenes is very often complicated by the fact that many book spines (or book shelves) can look like forearms.

In tracking by detection, almost all computation is directed at data association, which is achieved by minimizing ρ with respect to the template's parameters – the support of ρ identifies the relevant pixels. Similarly, in tracking using flow, data association is achieved by choosing the parameters of a flow model to get a good match between domains in frames n and $n+1$ – the definition of the domain cuts out the relevant pixels. When these methods run awry, it is because the underlying data association methods have failed. Either one cannot find the template, or one cannot get good parameters for the flow model.

There are a variety of simple data association strategies which exploit the presence of probability models. In particular, we have an estimate of $P(X_n|Y_0,\ldots,Y_{n-1})$ and we know $P(Y_n|X_n)$. From this we can obtain an estimate of $P(Y_n|Y_0,\ldots Y_{n-1})$, which gives us hints as to where the measurement might be.

One can use a **gate** – we look only at measurements that lie in a domain where $P(Y_n|Y_0,\ldots,Y_{n-1})$ is big enough. This is a method with roots in radar tracking of missiles and aeroplanes, where one must deal with only a small number (compared with the number of pixels in an image!) of returns, but the idea has been useful in visual tracking applications.

One can use **nearest neighbours**. In the classical version, we have a small set of possible measurements, and we choose the measurement with the largest value of $P(Y_n|Y_0,\ldots,Y_{n-1})$. This has all the dangers of wishful thinking – we are deciding that a measurement is valid because it is consistent with our track – but is often useful in practice. This strategy doesn't apply to most cases of tracking people in video because the search to find the maximising Y_n – which would likely be an image region – could be too difficult (but see Section 3). However, it could be applied when one is tracking markers attached to the body – in this case, we need to know which marker is which, and this information could be obtained by allocating a measurement to the marker whose predicted position is closest.

One can use **probabilistic data association**, where we use a weighted combination of measurements within a gate, weighted using (a) the predicted measurement and (b) the probability a measurement has dropped out. Again, this method has the dangers of wishful thinking, and again does not apply to most cases of tracking people; however, it could again be applied when one is tracking markers attached to the body.

1.4.3 Multiple modes

The Kalman filter is the workhorse of estimation, and can give useful results under many conditions. One doesn't need a guarantee of linearity to use a Kalman filter – if the logic of the application indicates that a linear model is reasonable, there is a good chance a Kalman filter will work. Rohr used a Kalman filter to track a walking person successfully, evaluating the measurement by matches to line segments on the outline [322, 323].

More recently, the method tends not to be used because of concerns about multiple modes. The representation adopted by a Kalman filter (the mean and covariance, sufficient statistics for a Gaussian distribution) tends to represent multimodal distributions poorly. There are several reasons one might encounter multiple modes.

First, nonlinear dynamics – or nonlinear measurement processes, or both – can create serious problems. The basic difficulty is that even quite innocuous looking setups can produce densities that are not normal, and are very difficult to represent and model. For example, let us look at only the hidden state. Assume that this is one dimensional. Now assume that state updates are *deterministic*, with $X_{i+1} = X_i + \epsilon \sin(X_i)$. If ϵ is sufficiently small, we have that for $0 < X_i < \pi$, $X_i < X_{i+1} < \pi$; for $-\pi < X_i < 0$, $-\pi < X_{i+1} < X_i$; and so on. Now assume that $P(X_0)$ is normal. For sufficiently large k, $P(X_k)$ will not be; there will be "clumps" of probability centered around the points $(2j + 1)\pi$ for j an integer. These clumps will be very difficult to represent, particularly if $P(X_0)$ has very large variance so that many clumps are important. Notice that what is creating a problem here is that quite small non-linearities in dynamics can cause probability to be concentrated in ways that are very difficult to represent. In particular, nonlinear dynamics are likely to produce densities with complicated sufficient statistics. There are cases where nonlinear dynamics does lead to densities that can be guaranteed to have finite-dimensional sufficient statistics (see [35, 83, 84]); to our knowledge, these have not been applied to human tracking.

Second, there are practical phenomena in human tracking that tend to suggest that non-normal distributions are a significant part of the problem. Assume we wish to track a 3D model of an arm in a single image. The elbow is bent; as it straightens, it will eventually run into an **end-stop** – the forearm can't rotate further without damage. At the end-stop, the posterior on state can't be a normal distribution, because a normal distribution would have some support on the wrong side of the end-stop, and this has a significant effect on the shape of the posterior (see Figure 2.5). Another case that is likely, but not guaranteed, to cause trouble is a **kinematic singularity**. For example, if the elbow is bent, we will be able to observe rotation about the humerus, but current observation models will make this unobservable if the elbow is straight (because the outline of the arm will not change; no current method can use the changes in appearance of the hand that will result). The dimension of the state space has collapsed. The posterior might be a normal distribution in this reduced dimension space, but that would require explicitly representing the collapse. The alternative, a covariance matrix of reduced rank, creates unattractive problems of representation. Deutscher *et al.* produce evidence that, in both cases, posteriors are not, in fact, normal distributions, and show that an extended Kalman filter can lose track in these cases [90].

Third, kinematic ambiguity in the relations between 3D and 2D are a major source of multiple modes. Assume we are tracking a human figure using a 3D representation of the body in a single view. If, for example, many 3D configurations correspond exactly to a single 2D configuration, then we expect the posterior to have multiple modes. Section 2 discusses this issue in extensive detail.

Fourth, the richest source of multiple modes is data association problems. An easy example illustrates how nasty this problem can be. Assume we have a problem with linear dynamics and a linear measurement model. However, at each tick of the clock we receive more than one measurement, exactly one of which comes from the process being studied. We will continue to write the states as $\mathbf{X}_i$, the measurements as $\mathbf{Y}_i$; but we now have δ_i, an indicator variable that tells which measurement comes from the process (and is unknown). $P(\mathbf{X}_N|\mathbf{Y}_{1..N},\delta_{1..N})$ is clearly Gaussian. We want $P(\mathbf{X}_N|\mathbf{Y}_{1..N}) = \sum_{histories} P(\mathbf{X}_N|\mathbf{Y}_{1..N},\delta_{1..N})P(\delta_{1..N}|\mathbf{Y}_{1..N})$, which is clearly a mixture of Gaussians. The number of components is exponential in the number of frames – there is one component per history – meaning that $P(\mathbf{X}_N|\mathbf{Y}_{1..N})$ could have a very large number of modes.

The following two sections discuss main potential sources of multimodal behaviour in great detail. Section 2 discusses the relations between 2D and 3D models of the body, which are generally agreed to be a source of multiple modes. Section 3 discusses data association methods. In this section, there is a brief discussion of the particle filter, a current favorite method for dealing with multi-modal densities. There are other methods: Beneš describes a class of nonlinear dynamical model for which the posterior can be represented with a sufficient statistic of constant finite dimension [35]. Daum extends the class of models for which this is the case ([83, 84]; see also [338] for an application and [106] for a comparison with the particle filter). Extensive accounts of particle filters appear in [93, 231, 319].

References

[1] Y. Abe, C. K. Liu, and Z. Popović, "Momentum-based parameterization of dynamic character motion," in *SCA '04: Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*, (New York, NY, USA), pp. 173–182, ACM Press, 2004.

[2] R. Abraham and J. E. Marsden, *Foundations of mechanics*. Addison-Wesley, 1978.

[3] A. Agarwal and B. Triggs, "Learning to track 3D human motion from silhouettes," in *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, (New York, NY, USA), p. 2, ACM Press, 2004.

[4] A. Agarwal and B. Triggs, "Tracking articulated motion using a mixture of autoregressive models," in *European Conference on Computer Vision*, pp. 54–65, 2004.

[5] A. Agarwal and B. Triggs, "Monocular human motion capture with a mixture of regressors," in *Workshop on Vision for Human Computer Interaction at CVPR'05*, 2005.

[6] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 44–58, 2006.

[7] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata, "Nonrigid motion analysis: articulated and elastic motion," *Computer Vision and Image Understanding*, vol. 70, no. 2, pp. 142–156, May 1998.

[8] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, March 1999.

[9] G. J. Agin and T. O. Binford, "Computer description of curved objects," in *Int. Joint Conf. Artificial Intelligence*, pp. 629–640, 1973.

[10] G. J. Agin and T. O. Binford, "Computer description of curved objects," *IEEE Trans. Computer*, vol. 25, no. 4, pp. 439–449, April 1976.

[11] R. M. Alexander, "Optimum timing of muscle activation for simple models of throwing," *J. Theor. Biol.*, vol. 150, pp. 349–372, 1991.

[12] F. C. Anderson and M. G. Pandy, "A dynamic optimization solution for vertical jumping in three dimensions," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 2, pp. 201–231, 1999.

[13] S. o. Anthropology Research Project, ed., *Anthropometric source book*. Webb Associates, 1978. NASA reference publication 1024, 3 Vols.

[14] W. A. Arentz and B. Olstad, "Classifying offensive sites based on image content," *Computer Vision and Image Understanding*, vol. 94, no. 1–3, pp. 295–310, April 2004.

[15] O. Arikan, D. A. Forsyth, and J. F. O'Brien, "Pushing people around," in *SCA '05: Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, (New York, NY, USA), pp. 59–66, ACM Press, 2005.

[16] O. Arikan and D. A. Forsyth, "Interactive motion generation from examples," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 483–490, ACM Press, 2002.

[17] O. Arikan, "Compression of motion capture databases," *ACM Transactions on Graphics: Proc. SIGGRAPH 2006*, to appear, 2006.

[18] O. Arikan, D. A. Forsyth, and J. O'Brien, "Motion synthesis from annotations," in *Proceedings of SIGGRAPH 95*, 2003.

[19] V. I. Arnold, *Mathematical methods of classical mechanics*. Springer-Verlag, 1989.

[20] V. Athitsos and S. Sclaroff, "An appearance-based framework for 3d hand shape classification and camera viewpoint estimation," in *Int. Conf. Automatic Face and Gesture Recognition*, pp. 40–45, 2002.

[21] V. Athitsos and S. Sclaroff, "Estimating 3D hand pose from a cluttered image," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 432–439, 2003.

[22] A. E. Atwater, "Biomechanics of overarm throwing movements and of throwing injuries," *Exerc. Sport. Sci. Rev.*, vol. 7, pp. 43–85, 1979.

[23] N. I. Badler, B. A. Barsky, and D. Zeltzer, eds., *Making them move: Mechanics, control, and animation of articulated figures*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1991.

[24] P. Baerlocher and R. Boulic, "An inverse kinematics architecture enforcing an arbitrary number of strict priority levels.," *The Visual Computer*, vol. 20, no. 6, pp. 402–417, 2004.

[25] H. H. Baker, "Building surfaces of evolution: the weaving wall," *Int. J. Computer Vision*, vol. 3, no. 1, pp. 51–72, May 1989.

[26] J. Barbič, A. Safonova, J.-Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard, "Segmenting motion capture data into distinct behaviors," in *GI '04: Proceedings of the 2004 conference on Graphics interface*, (School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada), pp. 185–194, Canadian Human-Computer Communications Society, 2004.

[27] C. D. Barclay, J. E. Cutting, and L. T. Kozlowski, "Temporal and spatial factors in gait perception that influence gender recognition," *Perception & Psychophysics*, vol. 23, no. 2, pp. 145–152, 1978.

[28] G. I. Barenblatt, *Scaling*. Cambridge University Press, 2003.

[29] C. Barron and I. A. Kakadiaris, "Estimating anthropometry and pose from a single image," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 669–676, 2000.

[30] C. Barron and I. A. Kakadiaris, "Estimating anthropometry and pose from a single uncalibrated image," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 269–284, March 2001.

[31] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," in *Int. Joint Conf. Artificial Intelligence*, pp. 659–663, 1977.

[32] Y. Bar-Shalom and X.-R. Li, *Estimation with applications to tracking and navigation*. New York, NY, USA: John Wiley & Sons, Inc., 2001.

[33] A. Baumberg and D. Hogg, "Learning flexible models from image sequences," in *European Conference on Computer Vision*, pp. 299–308, 1994.

[34] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, April 2002.

[35] V. E. Beneš, "Exact finite-dimensional filters with certain diffusion non linear drift," *Stochastics*, vol. 5, pp. 65–92, 1981.

[36] A. Berger, S. D. Pietra, and V. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, 1996.

[37] D. Bhat and J. K. Kearney, "On animating whip-type motions," *The Journal of Visualization and Computer Animation*, vol. 5, pp. 229–249, 1996.

[38] T. O. Binford, "Inferring surfaces from images," *Artificial Intelligence*, vol. 17, no. 1–3, pp. 205–244, August 1981.

[39] S. Blackman and R. Popoli, *Design and analysis of modern tracking systems*. Artech House, 1999.

[40] M. J. Black, Y. Yacoob, A. D. Jepson, and D. J. Fleet, "Learning parameterized models of image motion," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 561–567, 1997.

[41] A. Blake and M. Isard, *Active contours: The application of techniques from graphics, vision, control theory and statistics to visual tracking of shapes in motion*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1998.

[42] B. Bodenheimer, C. Rose, S. Rosenthal, and J. Pella, "The process of motion capture: Dealing with the data," in *Computer Animation and Simulation '97. Proceedings of the Eurographics Workshop*, 1997.

[43] A. Bosson, G. C. Cawley, Y. Chan, and R. Harvey, "Non-retrieval: Blocking pornographic images," in *Int. Conf. Image Video Retrieval*, pp. 50–59, 2002.

[44] J. E. Boyd and J. J. Little, "Phase in model-free perception of gait," in *IEEE Workshop on Human Motion*, pp. 3–10, 2000.

[45] M. Brand, "An entropic estimator for structure discovery," in *Proceedings of the 1998 conference on Advances in neural information processing systems II*, (Cambridge, MA, USA), pp. 723–729, MIT Press, 1999.

[46] M. Brand, “Structure learning in conditional probability models via an entropic prior and parameter extinction,” *Neural Comput.*, vol. 11, no. 5, pp. 1155–1182, 1999.

[47] M. Brand, “Shadow puppetry,” in *Int. Conf. on Computer Vision*, pp. 1237–1244, 1999.

[48] C. Bregler, J. Malik, and K. Pullen, “Twist based acquisition and tracking of animal and human kinematics,” *Int. J. Computer Vision*, vol. 56, no. 3, pp. 179–194, February 2004.

[49] C. Bregler and J. Malik, “Tracking people with twists and exponential maps,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 8–15, 1998.

[50] A. Broggi, M. Bertozzi, A. Fascioli, and M. Sechi, “Shape-based pedestrian detection,” in *Proc. IEEE Intelligent Vehicles Symposium*, pp. 215–220, 2000.

[51] L. Brown, *A radar history of world war II: Technical and military imperatives.* Institute of Physics Press, 2000.

[52] A. Bruderlin and L. Williams, “Motion signal processing,” in *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 97–104, ACM Press, 1995.

[53] R. Buderi, *The invention that changed the world.* Touchstone Press, 1998. reprint.

[54] M. C. Burl, T. K. Leung, and P. Perona, “Face localisation via shape statistics,” in *Int. Workshop on Automatic Face and Gesture Recognition*, 1995.

[55] Q. Cai and J. K. Aggarwal, “Automatic tracking of human motion in indoor scenes across multiple synchronized video streams,” in *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, (Washington, DC, USA), p. 356, IEEE Computer Society, 1998.

[56] B. Calais-Germain, *Anatomy of movement.* Eastland Press, 1993.

[57] M. Cardle, M. Vlachos, S. Brooks, E. Keogh, and D. Gunopulos, “Fast motion capture matching with replicated motion editing,” in *Proceedings of SIGGRAPH 2003 - Sketches and Applications*, 2003.

[58] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, “Free-viewpoint video of human actors,” *ACM Trans. Graph.*, vol. 22, no. 3, pp. 569–577, 2003.

[59] A. Cavallaro and T. Ebrahimi, “Video object extraction based on adaptive background and statistical change detection,” in *Proc. SPIE 4310*, pp. 465–475, 2000.

[60] J. Chai and J. K. Hodgins, “Performance animation from low-dimensional control signals,” *ACM Trans. Graph.*, vol. 24, no. 3, pp. 686–696, 2005.

[61] T. J. Cham and J. M. Rehg, “A multiple hypothesis approach to figure tracking,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 239–245, 1999.

[62] F. Cheng, W. Christmas, and J. V. Kittler, “Periodic human motion description for sports video databases,” in *Proceedings IAPR International Conference on Pattern Recognition*, pp. 870–873, 2004.

[63] K.-M. G. Cheung, S. Baker, and T. Kanade, “Shape-from-silhouette across time Part I: Theory and algorithms,” *Int. J. Comput. Vision*, vol. 62, no. 3, pp. 221–247, 2005.

[64] K.-M. G. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette across time Part II: Applications to human modeling and markerless motion tracking," *Int. J. Comput. Vision*, vol. 63, no. 3, pp. 225–245, 2005.

[65] K. M. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler, "A real time system for robust 3D voxel reconstruction of human motions," in *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, pp. 714 – 720, June 2000.

[66] J. chi Wu and Z. Popović, "Realistic modeling of bird flight animations," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 888–895, 2003.

[67] K. Choo and D. J. Fleet, "People tracking using hybrid Monte Carlo filtering," in *Int. Conf. on Computer Vision*, pp. 321–328, 2001.

[68] M. F. Cohen, "Interactive spacetime control for animation," in *SIGGRAPH '92: Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 293–302, ACM Press, 1992.

[69] D. Comaniciu and P. Meer, "Distribution free decomposition of multivariate data," *Pattern analysis and applications*, vol. 2, no. 1, pp. 22–30, 1999.

[70] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–97, 1995.

[71] L. S. Crawford and S. S. Sastry, "Biological motor control approaches for a planar diver," in *IEEE Conf. on Decision and Control*, pp. 3881–3886, 1995.

[72] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen, "Walking pedestrian recognition," *Intelligent Transportation Systems*, vol. 1, no. 3, pp. 155–163, September 2000.

[73] R. Cutler and L. S. Davis, "View-based detection and analysis of periodic motion," in *Proceedings IAPR International Conference on Pattern Recognition*, pp. 495–500, 1998.

[74] R. Cutler and L. S. Davis, "Real-time periodic motion detection, analysis, and applications," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 326–332, 1999.

[75] R. Cutler and L. S. Davis, "Robust periodic motion and motion symmetry detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 615–622, 2000.

[76] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and applications," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 781–796, August 2000.

[77] J. E. Cutting and L. T. Kozlowski, "Recognizing friends by their walk: Gait perception without familiarity cues," *Bulletin of the Psychonomic Society*, vol. 9, no. 5, pp. 353–356, 1977.

[78] J. E. Cutting, D. R. Proffitt, and L. T. Kozlowski, "A biomechanical invariant for gait perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 4, no. 3, pp. 357–372, 1978.

[79] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 886–893, 2005.

[80] T. J. Darrell, G. G. Gordon, M. Harville, and J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection," *Int. J. Computer Vision*, vol. 37, no. 2, pp. 175–185, June 2000.

[81] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Ann. Math. Statistics*, vol. 43, pp. 1470–1480, 1972.

[82] A. Dasgupta and Y. Nakamura, "Making feasible walking motion of humanoid robots from human motion capture data," in *1999 IEEE International Conference on Robotics & Automation*, pp. 1044–1049, 1999.

[83] F. E. Daum, "Beyond Kalman filters: practical design of nonlinear filters," in *Proc. SPIE*, pp. 252–262, 1995.

[84] F. E. Daum, "Exact finite dimensional nonlinear filters," *IEEE. Trans. Automatic Control*, vol. 31, pp. 616–622, 1995.

[85] Q. Delamarre and O. Faugeras, "3D articulated models and multi-view tracking with silhouettes," in *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*, (Washington, DC, USA), p. 716, IEEE Computer Society, 1999.

[86] Q. Delamarre and O. Faugeras, "3D articulated models and multiview tracking with physical forces," *Comput. Vis. Image Underst.*, vol. 81, no. 3, pp. 328–357, 2001.

[87] A. S. Deo and I. D. Walker, "Minimum effort inverse kinematics for redundant manipulators," *IEEE Transactions on Robotics and Automation*, vol. 13, no. 6, 1997.

[88] J. Deutscher, A. Blake, and I. D. Reid, "Articulated body motion capture by annealed particle filtering," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 126–133, 2000.

[89] J. Deutscher, A. J. Davison, and I. D. Reid, "Automatic partitioning of high dimensional search spaces associated with articulated body motion capture," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 669–676, 2001.

[90] J. Deutscher, B. North, B. Bascle, and A. Blake, "Tracking through singularities and discontinuities by random sampling," in *Int. Conf. on Computer Vision*, pp. 1144–1149, 1999.

[91] J. Deutscher and I. D. Reid, "Articulated body motion capture by stochastic search," *Int. J. Computer Vision*, vol. 61, no. 2, pp. 185–205, February 2005.

[92] M. Dimitrijevic, V. Lepetit, and P. Fua, "Human body pose recognition using spatio-temporal templates," in *ICCV workshop on Modeling People and Human Interaction*, 2005.

[93] A. Doucet, N. De Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[94] T. Drummond and R. Cipolla, "Real-time tracking of multiple articulated structures in multiple views," in *ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part II*, (London, UK), pp. 20–36, Springer-Verlag, 2000.

[95] T. Drummond and R. Cipolla, "Real-time tracking of highly articulated structures in the presence of noisy measurements.," in *ICCV*, pp. 315–320, 2001.

[96] T. Drummond and R. Cipolla, "Real-time tracking of complex structures with on-line camera calibration.," in *Proceedings of the British Machine Vision Conference 1999, BMVC 1999, Nottingham*, (T. P. Pridmore and D. Elliman, eds.), pp. 13–16, September 1999.

[97] T. W. Drummond and R. Cipolla, "Real-time visual tracking of complex structures," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 932–946, July 2002.

[98] A. D'Souza, S. Vijayakumar, and S. Schaal, "Learning inverse kinematics," in *Int. Conf. Intelligent Robots and Systems*, pp. 298–303, 2001.

[99] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Physics Letters B*, vol. 195, pp. 216–222, 1987.

[100] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, (Washington, DC, USA), pp. 726–733, IEEE Computer Society, 2003.

[101] A. E. Engin and S. T. Tumer, "Three-dimensional kinematic modelling of the human shoulder complex - Part I: Physical model and determination of joint sinus cones," *ASME Journal of Biomechanical Engineering*, vol. 111, pp. 107–112, 1989.

[102] P. Faloutsos, M. van de Panne, and D. Terzopoulos, "Composable controllers for physics-based character animation," in *Proceedings of ACM SIGGRAPH 2001*, pp. 251–260, August 2001. Computer Graphics Proceedings, Annual Conference Series.

[103] P. Faloutsos, M. van de Panne, and D. Terzopoulos, "The virtual stuntman: dynamic characters with a repertoire of autonomous motor skills," *Computers & Graphics*, vol. 25, no. 6, pp. 933–953, December 2001.

[104] A. C. Fang and N. S. Pollard, "Efficient synthesis of physically valid human motion," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 417–426, 2003.

[105] A. C. Fang and N. S. Pollard, "Efficient synthesis of physically valid human motion," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 417–426, July 2003.

[106] A. Farina, D. Benvenuti, and B. Ristic, "A comparative study of the Benes filtering problem," *Signal Processing*, vol. 82, pp. 133–147, 2002.

[107] A. Faul and M. Tipping, "Analysis of sparse Bayesian learning," in *Advances in Neural Information Processing Systems 14*, pp. 383–389, MIT Press, 2002.

[108] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient matching of pictorial structures," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 66–73, 2000.

[109] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Computer Vision*, vol. 61, no. 1, pp. 55–79, January 2005.

[110] X. Feng and P. Perona, "Human action recognition by sequence of movelet codewords," in *3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*, pp. 717–721, 2002.

[111] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 264–271, 2003.

[112] A. Fod, M. J. Matarić, and O. C. Jenkins, "Automated derivation of primitives for movement classification," *Auton. Robots*, vol. 12, no. 1, pp. 39–54, 2002.

[113] K. Forbes and E. Fiume, "An efficient search algorithm for motion data using weighted pca," in *Symposium on Computer Animation*, 2005.

[114] D. A. Forsyth, M. M. Fleck, and C. Bregler, "Finding naked people," in *European Conference on Computer Vision*, pp. 593–602, 1996.

[115] D. A. Forsyth and M. M. Fleck, "Body plans," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 678–683, 1997.

[116] D. A. Forsyth and M. M. Fleck, "Automatic detection of human nudes," *Int. J. Computer Vision*, vol. 32, no. 1, pp. 63–77, August 1999.

[117] D. A. Forsyth, J. Haddon, and S. Ioffe, "The joy of sampling," *Int. J. Computer Vision*, 2001.

[118] D. A. Forsyth and J. Ponce, *Computer vision: A modern approach.* Prentice-Hall, 2002.

[119] D. A. Forsyth, "Sampling, resampling and colour constancy," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 300–305, 1999.

[120] L. Fradet, M. Botcazou, C. Durocher, A. Cretual, F. Multon, J. Prioux, and P. Delamarche, "Do handball throws always exhibit a proximal-to-distal segmental sequence?," *Journal of Sports Sciences*, vol. 22, no. 5, pp. 439–447, 2004.

[121] J. Freyd, "Dynamic mental representations," *Psychological Review*, vol. 94, no. 4, pp. 427–438, 1987.

[122] D. S. Gao, J. Zhou, and L. P. Xin, "A novel algorithm of adaptive background estimation," in *IEEE Int. Conf. Image Processing*, pp. 395–398, 2001.

[123] D. M. Gavrila, J. Giebel, and S. Munder, "Vision-based pedestrian detection: the PROTECTOR system," in *Intelligent Vehicle Symposium*, pp. 13–18, 2004.

[124] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding: CVIU*, vol. 73, no. 1, pp. 82–98, 1999.

[125] D. M. Gavrila, "Sensor-based pedestrian protection," *Intelligent Transportation Systems*, pp. 77–81, 2001.

[126] D. Gavrila, "Pedestrian detection from a moving vehicle," in *European Conference on Computer Vision*, pp. 37–49, 2000.

[127] A. Gelb, *Applied optimal estimation.* MIT Press, 1974. written together with Staff of the Analytical Sciences Corporation.

[128] J. J. Gibson, *The perception of the visual world.* Houghton Mifflin, 1955.

[129] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds., *Markov chain Monte Carlo in practice.* Chapman and Hall, 1996.

[130] M. Girard and A. A. Maciejewski, "Computational modeling for the computer animation of legged figures," in *SIGGRAPH '85: Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 263–270, ACM Press, 1985.

[131] M. Girard, "Interactive design of 3-D computer-animated legged animal motion," in *SI3D '86: Proceedings of the 1986 workshop on Interactive 3D graphics*, (New York, NY, USA), pp. 131–150, ACM Press, 1987.

[132] M. Gleicher and N. Ferrier, "Evaluating video-based motion capture," in *CA '02: Proceedings of the Computer Animation*, (Washington, DC, USA), p. 75, IEEE Computer Society, 2002.

[133] M. Gleicher, H. J. Shin, L. Kovar, and A. Jepsen, "Snap-together motion: Assembling run-time animations," in *SI3D '03: Proceedings of the 2003 symposium on Interactive 3D graphics*, (New York, NY, USA), pp. 181–188, ACM Press, 2003.

[134] M. Gleicher, "Motion editing with spacetime constraints," in *Proceedings of the 1997 Symposium on Interactive 3D Graphics*, 1997.

[135] M. Gleicher, "Animation from observation: Motion capture and motion editing," *SIGGRAPH Comput. Graph.*, vol. 33, no. 4, pp. 51–54, 2000.

[136] M. Gleicher, "Comparing constraint-based motion editing methods," *Graphical Models*, 2001.

[137] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky, ""Dynamism of a dog on a leash" or behavior classification by eigen-decomposition of periodic motions," in *European Conference on Computer Vision*, p. 461 ff., 2002.

[138] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky, "Behavior classification by eigendecomposition of periodic motions," *Pattern Recognition*, vol. 38, no. 7, pp. 1033–1043, July 2005.

[139] H. Goldstein, *Classical mechanics.* Reading, MA: Addison Wesley, 1950.

[140] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *Proc. IEE-F*, vol. 140, pp. 107–113, 1993.

[141] K. Grauman, G. Shakhnarovich, and T. J. Darrell, "Virtual visual hulls: Example-based 3D shape inference from silhouettes," in *SMVP04*, pp. 26–37, 2004.

[142] W. E. L. Grimson, L. Lee, R. Romano, and C. Stauffer, "Using adaptive tracking to classify and monitor activities in a site," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 22–29, 1998.

[143] K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popović, "Style-based inverse kinematics," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 522–531, 2004.

[144] R. Grzeszczuk, D. Terzopoulos, and G. Hinton, "NeuroAnimator: Fast neural network emulation and control of physics-based models," in *Proceedings of SIGGRAPH 98*, pp. 9–20, July 1998.

[145] J. K. Hahn, "Realistic animation of rigid bodies," in *SIGGRAPH '88: Proceedings of the 15th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 299–308, ACM Press, 1988.

[146] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-time surveillance of people and their activities," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 809–830, 2000.

[147] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4S: A real-time system for detecting and tracking people in 2 1/2-D," in *European Conference on Computer Vision*, p. 877, 1998.

[148] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer Verlag, 2001.

[149] A. Hilton and J. Starck, "Multiple view reconstruction of people.," in *2nd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT 2004), 6–9 September 2004, Thessaloniki, Greece*, pp. 357–364, 2004.

[150] A. C. Hindmarsh and L. R. Petzold, "Algorithms and software for ordinary differential equations and differential-algebraic equations, Part I: Euler methods and error estimation," *Comput. Phys.*, vol. 9, no. 1, pp. 34–41, 1995.

[151] A. C. Hindmarsh and L. R. Petzold, "Algorithms and software for ordinary differential equations and differential-algebraic equations, Part II: Higher-order methods and software packages," *Comput. Phys.*, vol. 9, no. 2, pp. 148–155, 1995.

[152] G. E. Hinton, "Relaxation and its role in vision," Tech. Rep., University of Edinburgh, 1978. PhD Thesis.

[153] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds., *Understanding robust and exploratory data analysis*. John Wiley, 1983.

[154] J. K. Hodgins, J. F. O'Brien, and J. Tumblin, "Do geometric models affect judgments of human motion?," in *Graphics interface '97*, pp. 17–25, May 1997.

[155] J. K. Hodgins, J. F. O'Brien, and J. Tumblin, "Perception of human motion with different geometric models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 4, no. 4, pp. 307–316, October 1998.

[156] J. K. Hodgins and N. S. Pollard, "Adapting simulated behaviors for new characters," in *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 153–162, ACM Press/Addison-Wesley Publishing Co., 1997.

[157] J. K. Hodgins, W. L. Wooten, D. C. Brogan, and J. F. O'Brien, "Animating human athletics," in *Proceedings of SIGGRAPH 95*, pp. 71–78, August 1995.

[158] J. K. Hodgins, J. F. O'Brien, and R. E. Bodenheimer, "Computer animation," in *Wiley Encyclopedia of Electrical and Electronics Engineering*, (J. G. Webster, ed.), pp. 686–690, 1999.

[159] D. Hogg, "Model-based vision: A program to see a walking person," *Image and Vision Computing*, vol. 1, no. 1, pp. 5–20, 1983.

[160] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1–3, pp. 185–203, August 1981.

[161] B. K. P. Horn, "Closed form solutions of absolute orientation using orthonormal matrices," *J. Opt. Soc. America - A*, vol. 5, no. 7, pp. 1127–1135, 1987.

[162] B. K. P. Horn, "Closed form solutions of absolute orientation using unit quaternions," *J. Opt. Soc. America - A*, vol. 4, no. 4, pp. 629–642, April 1987.

[163] N. R. Howe, M. E. Leventon, and W. T. Freeman, "Bayesian Reconstruction of 3D Human Motion from Single-Camera Video," in *Advances in neural information processing systems 12*, (S. A. Solla, T. K. Leen, and K.-R. Müller, eds.), pp. 820–26, MIT Press, 2000.

[164] N. R. Howe, "Silhouette lookup for automatic pose tracking," in *IEEE Workshop on Articulated and Non-Rigid Motion*, p. 15, 2004.

[165] D. P. Huttenlocher, J. J. Noh, and W. J. Rucklidge, “Tracking non-rigid objects in complex scenes,” in *Int. Conf. on Computer Vision*, pp. 93–101, 1993.

[166] W. Hu, T. Tan, L. Wang, and S. Maybank, “A survey on visual surveillance of object motion and behaviors,” *IEEE transactions on systems, man, and cyberneticspart c: applications and reviews*, vol. 34, no. 3, 2004.

[167] L. Ikemoto, O. Arikan, and D. Forsyth, “Quick motion transitions with cached multi-way blends,” Tech. Rep. UCB/EECS-2006-14, EECS Department, University of California, Berkeley, February 13 2006.

[168] L. Ikemoto and D. A. Forsyth, “Enriching a motion collection by transplanting limbs,” in *Proc. Symposium on Computer Animation*, 2004.

[169] L. Ikemoto, O. Arikan, and D. A. Forsyth, “Knowing when to put your foot down,” in *Proc Symp. Interactive 3D graphics and Games*, 2006.

[170] S. Ioffe and D. A. Forsyth, “Learning to find pictures of people,” in *Proc. Neural Information Processing Systems*, 1998.

[171] S. Ioffe and D. A. Forsyth, “Human tracking with mixtures of trees,” in *Int. Conf. on Computer Vision*, pp. 690–695, 2001.

[172] S. Ioffe and D. A. Forsyth, “Probabilistic methods for finding people,” *Int. J. Computer Vision*, vol. 43, no. 1, pp. 45–68, June 2001.

[173] S. Ioffe and D. Forsyth, “Mixtures of trees for object recognition,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[174] M. Isard and A. Blake, “Icondensation: Unifying low-level and high-level tracking in a stochastic framework,” in *European Conference on Computer Vision*, p. 893, 1998.

[175] M. Isard and A. Blake, “C-conditional density propagation for visual tracking,” *IJCV*, vol. 29, no. 1, pp. 5–28, August 1998.

[176] Y. A. Ivanov, A. F. Bobick, and J. Liu, “Fast lighting independent background subtraction,” in *In Proc. of the IEEE Workshop on Visual Surveillance – VS’98*, pp. 49–55, 1998.

[177] Y. A. Ivanov, A. F. Bobick, and J. Liu, “Fast lighting independent background subtraction,” *Int. J. Computer Vision*, vol. 37, no. 2, pp. 199–207, June 2000.

[178] O. Javed and M. Shah, “Tracking and object classification for automated surveillance,” in *European Conference on Computer Vision*, p. 343 ff., 2002.

[179] O. C. Jenkins and M. J. Matarić, “Automated derivation of behavior vocabularies for autonomous humanoid motion,” in *AAMAS ’03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, (New York, NY, USA), pp. 225–232, ACM Press, 2003.

[180] O. C. Jenkins and M. J. Matarić, “A spatio-temporal extension to Isomap nonlinear dimension reduction,” in *ICML ’04: Proceedings of the twenty-first international conference on Machine learning*, (New York, NY, USA), p. 56, ACM Press, 2004.

[181] F. V. Jensen, *An introduction to bayesian networks.* London: UCL Press, 1996.

[182] C. Y. Jeong, J. S. Kim, and K. S. Hong, “Appearance-based nude image detection,” in *Proceedings IAPR International Conference on Pattern Recognition*, pp. 467–470, 2004.

[183] R. Jin, R. Yan, J. Zhang, and A. Hauptmann, "A faster iterative scaling algorithm for conditional exponential models," in *Proc. International Conference on Machine learning*, 2003.

[184] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[185] I. T. Joliffe, *Principal Component Analysis*. Springer-Verlag, 2002.

[186] M. J. Jones and P. Viola, "Face recognition using boosted local features," in *IEEE International Conference on Computer Vision (ICCV)*, 2003.

[187] M. J. Jones and P. Viola, "Fast multi-view face detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[188] R. V. Jones, *Most secret war*. Wordsworth Military Library, 1998. reprint.

[189] D. Jurafsky and J. H. Martin, *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition*. Prentice-Hall, 2000.

[190] S. X. Ju, M. J. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated image motion," in *Proc. Int. Conference on Face and Gesture*, pp. 561–567, 1996.

[191] M. Kass, A. P. Witkin, and D. Terzopoulos, "Snakes: active contour models," in *Int. Conf. on Computer Vision*, pp. 259–268, 1987.

[192] M. Kass, A. P. Witkin, and D. Terzopoulos, "Snakes: active contour models," *Int. J. Computer Vision*, vol. 1, no. 4, pp. 321–331, January 1988.

[193] R. Kehl, M. Bray, and L. V. Gool, "Full body tracking from multiple views using stochastic sampling," in *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, (Washington, DC, USA), pp. 129–136, IEEE Computer Society, 2005.

[194] E. J. Keogh, "Exact indexing of dynamic time warping.," in *VLDB*, pp. 406–417, 2002.

[195] E. J. Keogh, "Efficiently finding arbitrarily scaled patterns in massive time series databases.," in *7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 253–265, 2003.

[196] E. J. Keogh, T. Palpanas, V. B. Zordan, D. Gunopulos, and M. Cardle, "Indexing large human-motion databases," in *Proc. 30th VLDB Conf.*, pp. 780–791, 2004.

[197] V. Kettnaker and R. Zabih, "Counting people from multiple cameras," in *ICMCS '99: Proceedings of the IEEE International Conference on Multimedia Computing and Systems Volume II-Volume 2*, (Washington, DC, USA), p. 267, IEEE Computer Society, 1999.

[198] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 506–513, 2004.

[199] D. King, "Generating vertical velocity and angular momentum during skating jumps," in *23rd Annual Meeting of the American Society of Biomechanics*, 1999.

[200] A. G. Kirk, J. F. O'Brien, and D. A. Forsyth, "Skeletal parameter estimation from optical motion capture data," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005.

[201] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models," *Journal of Computational and Graphical Statistics*, vol. 5, pp. 1–25, 1996.

[202] K. Kondo, "Inverse kinematics of a human arm," Tech. Rep., Stanford University, Stanford, CA, USA, 1994.

[203] A. Kong, J. S. Liu, and W. H. Wong, "Sequential imputations and Bayesian missing data problems," *Journal of the American Statistical Association*, vol. 89, pp. 278–288, 1994.

[204] J. U. Korein and N. I. Badler, "Techniques for generating the goal-directed motion of articulated structures," *IEEE Computer Graphics and Applications*, pp. 71–81, 1982.

[205] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 473–482, ACM Press, 2002.

[206] L. Kovar and M. Gleicher, "Flexible automatic motion blending with registration curves," in *SCA '03: Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, (Aire-la-Ville, Switzerland, Switzerland), pp. 214–224, Eurographics Association, 2003.

[207] L. Kovar and M. Gleicher, "Automated extraction and parameterization of motions in large data sets," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 559–568, 2004.

[208] L. Kovar, J. Schreiner, and M. Gleicher, "Footskate cleanup for motion capture editing," in *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 97–104, ACM Press, 2002.

[209] L. T. Kozlowski and J. E. Cutting, "Recognizing the sex of a walker from a dynamic point-light display," *Perception & Psychophysics*, vol. 21, no. 6, pp. 575–580, 1977.

[210] L. T. Kozlowski and J. E. Cutting, "Recognizing the gender of walkers from point-lights mounted on ankles: Some second thoughts," *Perception & Psychophysics*, vol. 23, no. 5, p. 459, 1978.

[211] H. Ko and N. Badler, "Animating human locomotion with inverse dynamics," *IEEE Computer Graphics and Application*, vol. 16, no. 2, pp. 50–59, 1996.

[212] M. P. Kumar, P. H. S. Torr, and A. Zisserman, "Extending pictorial structures for object recognition," in *Proceedings of the British Machine Vision Conference*, 2004.

[213] T. Kwon and S. Y. Shin, "Motion modeling for on-line locomotion synthesis," in *SCA '05: Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, (New York, NY, USA), pp. 29–38, ACM Press, 2005.

[214] J. Lee and S. Y. Shin, "A hierarchical approach to interactive motion editing for human-like figures," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 39–48, ACM Press/Addison-Wesley Publishing Co., 1999.

[215] J. Lee, J. Chai, P. Reitsma, J. Hodgins, and N. Pollard, "Interactive control of avatars animated with human motion data," in *Proceedings of SIGGRAPH 95*, 2002.

[216] M. W. Lee and I. Cohen, "Human upper body pose estimation in static images," in *European Conference on Computer Vision*, pp. 126–138, 2004.

[217] M. W. Lee and I. Cohen, "Proposal maps driven MCMC for estimating human body pose in static images," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 334–341, 2004.

[218] M. W. Lee and R. Nevatia, "Dynamic human pose estimation using Markov Chain Monte Carlo approach," in *IEEE Workshop on Motion and Video Computing*, pp. 168–175, 2005.

[219] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *ECCV-04 Workshop on Stat. Learn. in Comp. Vis.*, pp. 17–32, 2004.

[220] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 878–885, 2005.

[221] A. Leonardis, A. Gupta, and R. Bajcsy, "Segmentation of range images as the search for geometric parametric models," *Int. J. Computer Vision*, vol. 14, no. 3, pp. 253–277, April 1995.

[222] T. K. Leung, M. C. Burl, and P. Perona, "Finding faces in cluttered scenes using random labelled graph matching," in *Int. Conf. on Computer Vision*, 1995.

[223] J. J. Little and J. E. Boyd, "Describing motion for recognition," in *International Symposium on Computer Vision*, pp. 235–240, 1995.

[224] J. J. Little and J. E. Boyd, "Recognizing people by their gait: The shape of motion," *Videre*, vol. 1, no. 2, 1998.

[225] J. J. Little and J. E. Boyd, "Shape of motion and the perception of human gaits," in *IEEE Workshop on Empirical Evaluation Methods in Computer Vision*, 1998.

[226] C. K. Liu, A. Hertzmann, and Z. Popović, "Learning physics-based motion style with nonlinear inverse optimization," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 1071–1081, 2005.

[227] C. K. Liu and Z. Popović, "Synthesis of complex dynamic character motion from simple animations," in *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 408–416, ACM Press, 2002.

[228] C. Liu, S. C. Zhu, and H. Y. Shum, "Learning inhomogeneous gibbs model of faces by minimax entropy," in *Int. Conf. on Computer Vision*, pp. 281–287, 2001.

[229] F. Liu and R. W. Picard, "Detecting and segmenting periodic motion," Media Lab Vision and Modelling TR-400, MIT, 1996.

[230] F. Liu and R. W. Picard, "Finding periodicity in space and time," in *Int. Conf. on Computer Vision*, pp. 376–383, 1998.

[231] J. S. Liu, *Monte Carlo strategies in scientific computing.* Springer, 2001.

[232] Z. Liu and M. F. Cohen, "Decomposition of linked figure motion: Diving," in *5th Eurographics Workshop on Animation and Simulation*, 1994.

[233] Z. Liu, S. J. Gortler, and M. F. Cohen, "Hierarchical spacetime control," in *SIGGRAPH '94: Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 35–42, ACM Press, 1994.

[234] Z. Liu, H. Chen, and H. Y. Shum, "An efficient approach to learning inhomogeneous Gibbs model," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 425–431, 2003.

[235] M. Liverman, *The Animator'S Motion Capture Guide: Organizing, Managing, Editing*. Charles River Media, 2004.

[236] B. Li and H. Holstein, "Recognition of human periodic motion: A frequency domain approach," in *Proceedings IAPR International Conference on Pattern Recognition*, pp. 311–314, 2002.

[237] Y. Li, T. Wang, and H.-Y. Shum, "Motion texture: A two-level statistical model for character motion synthesis," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 465–472, ACM Press, 2002.

[238] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[239] G. Loy, M. Eriksson, J. Sullivan, and S. Carlsson, "Monocular 3D reconstruction of human motion in long action sequences," in *European Conference on Computer Vision*, pp. 442–455, 2004.

[240] J. P. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," in *Int. Conf. on Computer Vision*, pp. 572–578, 1999.

[241] J. P. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," *Int. J. Computer Vision*, vol. 39, no. 1, pp. 57–71, August 2000.

[242] J. P. MacCormick and M. Isard, "Partitioned sampling, articulated objects, and interface-quality hand tracking," in *European Conference on Computer Vision*, pp. 3–19, 2000.

[243] A. A. Maciejewski, "Motion simulation: Dealing with the Ill-conditioned equations of motion for articulated figures," *IEEE Comput. Graph. Appl.*, vol. 10, no. 3, pp. 63–71, 1990.

[244] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. New York, NY, USA: Cambridge University Press, 2002.

[245] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[246] D. Marr and H. K. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes," *Proc. Roy. Soc. B*, vol. 200, pp. 269–294, 1978.

[247] M. J. Matarić, V. B. Zordan, and Z. Mason, "Movement control methods for complex, dynamically simulated agents: Adonis dances the Macarena," in *AGENTS '98: Proceedings of the second international conference on Autonomous agents*, (New York, NY, USA), pp. 317–324, ACM Press, 1998.

[248] M. J. Matarić, V. B. Zordan, and M. M. Williamson, "Making complex articulated agents dance," *Autonomous Agents and Multi-Agent Systems*, vol. 2, no. 1, pp. 23–43, 1999.

[249] M. McKenna and D. Zeltzer, "Dynamic simulation of autonomous legged locomotion," in *SIGGRAPH '90: Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 29–38, ACM Press, 1990.

[250] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Computer Vision and Image Understanding*, vol. 80, no. 1, pp. 42–56, October 2000.

[251] A. Menache, *Understanding Motion Capture for Computer Animation and Video Games*. Morgan-Kaufmann, 1999.

[252] A. Micilotta, E. Ong, and R. Bowden, "Detection and tracking of humans by probabilistic body part assembly," in *British Machine Vision Conference*, pp. 429–438, 2005.

[253] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *European Conference on Computer Vision*, pp. 69–82, 2004.

[254] K. Mikolajczyk, C. Schmid, and A. Zisserman, "Human detection based on a probabilistic assembly of robust part detectors," in *European Conference on Computer Vision*, pp. 69–82, 2004.

[255] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE T. Pattern Analysis and Machine Intelligence*, accepted, 2004.

[256] K. Mikolajczyk, "Face detector," Tech. Rep., INRIA Rhone-Alpes. Ph.D report.

[257] A. Mittal and L. S. Davis, "M2Tracker: A multi-view approach to segmenting and tracking people in a cluttered scene," *Int. J. Comput. Vision*, vol. 51, no. 3, pp. 189–203, 2003.

[258] T. B. Moeslund, "Summaries of 107 computer vision-based human motion capture papers," Tech. Rep. LLA 99-01, University of Aalborg, 1999.

[259] B. Moghaddam and A. P. Pentland, "Probabilistic visual learning for object detection," in *Int. Conf. on Computer Vision*, pp. 786–793, 1995.

[260] A. Mohan, C. P. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, pp. 349–361, April 2001.

[261] A. Mohr and M. Gleicher, "Building efficient, accurate character skins from examples," *ACM Trans. Graphics*, vol. 22, no. 3, pp. 562–568, 2003.

[262] G. Monheit and N. I. Badler, "A kinematic model of the human spine and torso," *IEEE Comput. Graph. Appl.*, vol. 11, no. 2, pp. 29–38, 1991.

[263] G. Mori and J. Malik, "Estimating human body configurations using shape context matching," in *European Conference on Computer Vision LNCS 2352*, pp. 666–680, 2002.

[264] G. Mori and J. Malik, "Recovering 3d human body configurations using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear, 2005.

[265] M. Müller, T. Röder, and M. Clausen, "Efficient content-based retrieval of motion capture data," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 677–685, 2005.

[266] F. Multon, L. France, M.-P. Cani, and G. Debunne, "Computer animation of human walking: a survey," *Journal of Visualization and Computer Animation (JVCA)*, vol. 10, pp. 39–54, Published under the name Marie-Paule Cani-Gascuel, 1999.

[267] J. L. Mundy and C.-F. Chang, "Fusion of intensity, texture, and color in video tracking based on mutual information," in *Applied Imagery Pattern Recognition Workshop*, pp. 10–15, 2004.

[268] K. Murphy, Y. Weiss, and M. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence*, pp. 467–475, 1999.

[269] E. Muybridge, *Animals in Motion*. Dover, 1957.

[270] E. Muybridge, *The Human Figure in Motion*. Dover, 1989.

[271] R. M. Neal, "Annealed importance sampling," *Statistics and Computing*, vol. 11, no. 2, pp. 125–139, 2001.

[272] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Computer science tech report CRG-TR-93-1, University of Toronto, 1993.

[273] R. M. Neal, "Sampling from multimodal distributions using tempered transitions," *Statistics and Computing*, vol. 6, pp. 353–366, 1996.

[274] R. M. Neal, "Annealed importance sampling," Tech. Rep., Technical Report No. 9805 (revised), Dept. of Statistics, University of Toronto, 1998.

[275] J. T. Ngo and J. Marks, "Physically realistic motion synthesis in animation," *Evol. Comput.*, vol. 1, no. 3, pp. 235–268, 1993.

[276] J. T. Ngo and J. Marks, "Spacetime constraints revisited," in *SIGGRAPH '93: Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 343–350, ACM Press, 1993.

[277] S. A. Niyogi and E. H. Adelson, "Analyzing gait with spatiotemporal surfaces," in *Proc. IEEE Workshop on Nonrigid and Articulated Motion*, pp. 64–69, 1994.

[278] S. A. Niyogi and E. H. Adelson, "Analyzing and recognizing walking figures in XYT," Media Lab Vision and Modelling TR-223, MIT, 1995.

[279] M. Oren, C. P. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 193–199, 1997.

[280] M. Oren, C. P. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "A trainable system for people detection," in *DARPA IU Workshop*, pp. 207–214, 1997.

[281] J. O'Rourke and N. I. Badler, "Model-based image analysis of human motion using constraint propagation," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 2, no. 6, pp. 522–536, November1980.

[282] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection.," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 130–6, 1997.

[283] C. J. Pai, H. R. Tyan, Y. M. Liang, H. Y. M. Liao, and S. W. Chen, “Pedestrian detection and tracking at crossroads,” in *IEEE Int. Conf. Image Processing*, pp. 101–104, 2003.

[284] C. J. Pai, H. R. Tyan, Y. M. Liang, H. Y. M. Liao, and S. W. Chen, “Pedestrian detection and tracking at crossroads,” *Pattern Recognition*, vol. 37, no. 5, pp. 1025–1034, May 2004.

[285] M. G. Pandy and F. C. Anderson, “Dynamic simulation of human movement using large-scale models of the body,” in *Proc. IEEE Intl. Conference on Robotics and Automation*, pp. 676–681, 2000.

[286] M. Pandy, F. C. Anderson, and D. G. Hull, “A parameter optimization approach for the optimal control of large-scale musculoskeletal systems,” *J. of Biomech. Eng.*, pp. 450–460, 1992.

[287] C. P. Papageorgiou, T. Evgeniou, and T. Poggio, “A trainable object detection system,” in *DARPA IU Workshop*, pp. 1019–1024, 1998.

[288] C. P. Papageorgiou, M. Oren, and T. Poggio, “A general framework for object detection,” in *Int. Conf. on Computer Vision*, pp. 555–562, 1998.

[289] C. P. Papageorgiou and T. Poggio, “A pattern classification approach to dynamical object detection,” in *Int. Conf. on Computer Vision*, pp. 1223–1228, 1999.

[290] C. P. Papageorgiou and T. Poggio, “Trainable pedestrian detection,” in *IEEE Int. Conf. Image Processing*, pp. 35–39, 1999.

[291] C. P. Papageorgiou, “A trainable system for object detection in images and video sequences constantine,” Tech. Rep., MIT, 2000. Ph. D.

[292] C. Papageorgiou and T. Poggio, “A trainable system for object detection,” *Int. J. Computer Vision*, vol. 38, no. 1, pp. 15–33, June 2000.

[293] V. Parenti-Castelli, A. Leardini, R. D. Gregorio, and J. J. O’Connor, “On the modeling of passive motion of the human knee joint by means of equivalent planar and spatial parallel mechanisms,” *Auton. Robots*, vol. 16, no. 2, pp. 219–232, 2004.

[294] S. I. Park, H. J. Shin, T. H. Kim, and S. Y. Shin, “On-line motion blending for real-time locomotion generation: Research Articles,” *Comput. Animat. Virtual Worlds*, vol. 15, no. 3–4, pp. 125–138, 2004.

[295] S. I. Park, H. J. Shin, and S. Y. Shin, “On-line locomotion generation based on motion blending,” in *SCA ’02: Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 105–111, New York, NY, USA: ACM Press, 2002.

[296] C. B. Phillips, J. Zhao, and N. I. Badler, “Interactive real-time articulated figure manipulation using multiple kinematic constraints,” in *SI3D ’90: Proceedings of the 1990 symposium on Interactive 3D graphics*, (New York, NY, USA), pp. 245–250, ACM Press, 1990.

[297] S. D. Pietra, V. D. Pietra, and J. Lafferty, “Inducing features of random fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380–393, 1997.

[298] R. Plänkers and P. Fua, “Tracking and modeling people in video sequences,” *Comput. Vis. Image Underst.*, vol. 81, no. 3, pp. 285–302, 2001.

[299] T. Poggio and K.-K. Sung, “Finding human faces with a gaussian mixture distribution-based face model,” in *Asian Conf. on Computer Vision*, pp. 435–440, 1995.

[300] R. Polana and R. C. Nelson, “Detecting activities,” in *DARPA IU Workshop*, pp. 569–574, 1993.

[301] R. Polana and R. C. Nelson, “Detecting activities,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2–7, 1993.

[302] R. Polana and R. C. Nelson, “Detecting activities,” *J. Visual Communication Image Representation*, vol. 5, pp. 172–180, 1994.

[303] R. Polana and R. C. Nelson, “Low level recognition of human motion,” in *IEEE Workshop on Articulated and Non-Rigid Motion*, 1994.

[304] R. Polana and R. C. Nelson, “Recognition of nonrigid motion,” in *ARPA94*, pp. 1219–1224, 1994.

[305] R. Polana and R. C. Nelson, “Detection and recognition of periodic, nonrigid motion,” *Int. J. Computer Vision*, vol. 23, no. 3, pp. 261–282, 1997.

[306] R. Polana and R. Nelson, “Recognizing activities,” in *Proceedings IAPR International Conference on Pattern Recognition*, pp. 815–818, 1994.

[307] N. S. Pollard and F. Behmaram-Mosavat, “Force-based motion editing for locomotion tasks,” in *In Proceedings of the IEEE International Conference on Robotics and Automation*, 2000.

[308] J. Popović, S. M. Seitz, and M. Erdmann, “Motion sketching for control of rigid-body simulations,” *ACM Trans. Graph.*, vol. 22, no. 4, pp. 1034–1054, 2003.

[309] Z. Popović and A. Witkin, “Physically based motion transformation,” in *SIGGRAPH ’99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 11–20, ACM Press/Addison-Wesley Publishing Co., 1999.

[310] K. Pullen and C. Bregler, “Motion capture assisted animation: Texturing and synthesis,” in *Proceedings of SIGGRAPH 95*, 2002.

[311] C. A. Putnam, “A segment interaction analysis of proximal-to-distal sequential segment motion patterns,” *Med. Sci. Sports. Exerc.*, vol. 23, pp. 130–144, 1991.

[312] D. Ramanan, D. A. Forsyth, and A. Zisserman, “Strike a pose: Tracking people by finding stylized poses,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 271–278, 2005.

[313] D. Ramanan and D. A. Forsyth, “Automatic annotation of everyday movements,” in *Proc. Neural Information Processing Systems*, 2003.

[314] D. Ramanan and D. A. Forsyth, “Finding and tracking people from the bottom up,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 467–474, 2003.

[315] D. Ramanan, *Tracking People and Recognizing their Activities.* PhD thesis, U.C. Berkeley, 2005.

[316] P. S. A. Reitsma and N. S. Pollard, “Evaluating motion graphs for character navigation,” in *Eurographics/ACM Symposium on Computer Animation*, pp. 89–98, 2004.

[317] L. Ren, A. Patrick, A. A. Efros, J. K. Hodgins, and J. M. Rehg, “A data-driven approach to quantifying natural human motion,” *ACM Trans. Graph.*, vol. 24, no. 3, pp. 1090–1097, 2005.

[318] L. Ren, G. Shakhnarovich, J. K. Hodgins, H. Pfister, and P. Viola, “Learning silhouette features for control of human motion,” *ACM Trans. Graph.*, vol. 24, no. 4, pp. 1303–1331, 2005.

[319] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications.* Artech House, 2004.

[320] J. Rittscher and A. Blake, “Classification of human body motion,” in *Int. Conf. on Computer Vision*, pp. 634–639, 1999.

[321] T. J. Roberts, S. J. McKenna, and I. W. Ricketts, “Human pose estimation using learnt probabilistic region similarities and partial configurations,” in *European Conference on Computer Vision*, pp. 291–303, 2004.

[322] K. Rohr, “Incremental recognition of pedestrians from image sequences,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 9–13, 1993.

[323] K. Rohr, “Towards model-based recognition of human movements in image sequences,” *CVGIP: Image Understanding*, vol. 59, no. 1, pp. 94–115, 1994.

[324] R. Ronfard, C. Schmid, and B. Triggs, “Learning to parse pictures of people,” in *European Conference on Computer Vision*, p. 700 ff., 2002.

[325] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff, “3D hand pose reconstruction using specialized mappings,” in *Int. Conf. on Computer Vision*, pp. 378–385, 2001.

[326] R. Rosenfeld, “A maximum entropy approach to adaptive statistical language modelling,” *Computer, Speech and Language*, vol. 10, pp. 187–228, 1996.

[327] C. Rose, M. F. Cohen, and B. Bodenheimer, “Verbs and adverbs: Multidimensional Motion Interpolation,” *IEEE Comput. Graph. Appl.*, vol. 18, no. 5, pp. 32–40, 1998.

[328] C. Rose, B. Guenter, B. Bodenheimer, and M. F. Cohen, “Efficient generation of motion transitions using spacetime constraints,” in *SIGGRAPH ’96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 147–154, ACM Press, 1996.

[329] S. Roth, L. Sigal, and M. J. Black, “Gibbs likelihoods for Bayesian tracking,” in *CVPR04*, pp. 886–893, 2004.

[330] P. J. Rousseeuw, *Robust Regression and Outlier Detection.* Wiley, 1987.

[331] H. A. Rowley, S. Baluja, and T. Kanade, “Human face detection in visual scenes,” in *Advances in Neural Information Processing 8*, (D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds.), pp. 875–881, 1996.

[332] H. A. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 203–8, 1996.

[333] H. A. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection,” *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[334] H. A. Rowley, S. Baluja, and T. Kanade, “Rotation invariant neural network-based face detection,” in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 38–44, 1998.

[335] A. Safonova, J. K. Hodgins, and N. S. Pollard, "Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 514–521, 2004.

[336] A. Safonova and J. K. Hodgins, "Analyzing the physical correctness of interpolated human motion," in *SCA '05: Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, (New York, NY, USA), pp. 171–180, ACM Press, 2005.

[337] S. Schaal, C. G. Atkeson, and S. Vijayakumar, "Scalable techniques from nonparametric statistics for real time robot learning," *Applied Intelligence*, vol. 17, no. 1, pp. 49–60, 2002.

[338] G. C. Schmidt, "Designing nonlinear filters based on Daum's theory," *Journal of Guidance, Control and Dynamics*, vol. 16, pp. 371–376, 1993.

[339] H. Schneiderman and T. Kanade, "A statistical method for 3d object detection applied to faces and cars," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 746–751, 2000.

[340] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA, USA: MIT Press, 2001.

[341] S. M. Seitz and C. R. Dyer, "Affine invariant detection of periodic motion," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 970–975, 1994.

[342] S. M. Seitz and C. R. Dyer, "View invariant analysis of cyclic motion," *Int. J. Computer Vision*, vol. 25, no. 3, pp. 231–251, December 1997.

[343] A. Senior, "Tracking people with probabilistic appearance models," in *IEEE Workshop on Performance Evaluation Tracking Surveillance*, pp. 48–55, 2002.

[344] A. Shahrokni, T. Drummond, and P. Fua, "Fast texture-based tracking and delineation using texture entropy," in *International Conference on Computer Vision*, 2005.

[345] A. Shahrokni, T. Drummond, V. Lepetit, and P. Fua, "Markov-based silhouette extraction for three–dimensional body tracking in presence of cluttered background," in *British Machine Vision Conference*, (Kingston, UK), 2004.

[346] A. Shahrokni, F. Fleuret, and P. Fua, "Classifier-based contour tracking for rigid and deformable objects," in *British Machine Vision Conference*, (Oxford, UK), 2005.

[347] G. Shakhnarovich, P. Viola, and T. J. Darrell, "Fast pose estimation with parameter-sensitive hashing," in *Int. Conf. on Computer Vision*, pp. 750–757, 2003.

[348] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis.* New York, NY, USA: Cambridge University Press, 2004.

[349] H. J. Shin, L. Kovar, and M. Gleicher, "Physical touch-up of human motions," in *PG '03: Proceedings of the 11th Pacific Conference on Computer Graphics and Applications*, (Washington, DC, USA), p. 194, IEEE Computer Society, 2003.

[350] H. J. Shin, J. Lee, S. Y. Shin, and M. Gleicher, "Computer puppetry: An importance-based approach," *ACM Trans. Graph.*, vol. 20, no. 2, pp. 67–94, 2001.

[351] H. Sidenbladh, M. J. Black, and D. J. Fleet, "Stochastic tracking of 3D human figures using 2D image motion," in *European Conference on Computer Vision*, 2000.

[352] H. Sidenbladh and M. J. Black, "Learning image statistics for bayesian tracking," in *Int. Conf. on Computer Vision*, pp. 709–716, 2001.

[353] H. Sidenbladh and M. J. Black, "Learning the statistics of people in images and video," *Int. J. Computer Vision*, vol. 54, no. 1, pp. 181–207, September 2003.

[354] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard, "Tracking loose-limbed people," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 421–428, 2004.

[355] M.-C. Silaghi, R. Plänkers, R. Boulic, P. Fua, and D. Thalmann, "Local and global skeleton fitting techniques for optical motion capture," in *Modelling and Motion Capture Techniques for Virtual Environments*, pp. 26–40, November 1998. Proceedings of CAPTECH '98.

[356] K. Sims, "Evolving virtual creatures," in *SIGGRAPH '94: Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 15–22, ACM Press, 1994.

[357] J. Sivic, M. Everingham, and A. Zisserman, "Person spotting: Video shot retrieval for face sets," in *International Conference on Image and Video Retrieval (CIVR 2005), Singapore*, 2005.

[358] C. Sminchisescu and A. Telea, "Human pose estimation from silhouettes: A consistent approach using distance level sets," in *WSCG02*, p. 413, 2002.

[359] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3D body tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 447–454, 2001.

[360] C. Sminchisescu and B. Triggs, "Building roadmaps of local minima of visual models," in *European Conference on Computer Vision*, p. 566 ff., 2002.

[361] C. Sminchisescu and B. Triggs, "Hyperdynamics importance sampling," in *European Conference on Computer Vision*, p. 769 ff., 2002.

[362] C. Sminchisescu and B. Triggs, "Estimating articulated human motion with covariance scaled sampling," *The International Journal of Robotics Research*, vol. 22, no. 6, pp. 371–391, 2003.

[363] C. Sminchisescu and B. Triggs, "Kinematic jump processes for monocular 3D human tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 69–76, 2003.

[364] C. Sminchisescu and B. Triggs, "Kinematic jump processes for monocular 3D human tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 69–76, 2003.

[365] C. Sminchisescu and B. Triggs, "Building roadmaps of minima and transitions in visual models," *Int. J. Computer Vision*, vol. 61, no. 1, pp. 81–101, January 2005.

[366] C. Sminchisescu, "Consistency and coupling in human model likelihoods," in *Proceedings International Conference on Automatic Face and Gesture Recognition*, pp. 22–27, 2002.

[367] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[368] Y. Song, X. Feng, and P. Perona, "Towards detection of human motion," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 810–817, 2000.

[369] Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 814–827, July 2003.

[370] N. Sprague and J. Luo, "Clothed people detection in still images," in *Proceedings IAPR International Conference on Pattern Recognition*, pp. 585–589, 2002.

[371] J. Starck, A. Hilton, and J. Illingworth, "Human shape estimation in a multi-camera studio.," in *BMVC*, 2001.

[372] J. Starck and A. Hilton, "Model-based multiple view reconstruction of people.," in *Int. Conf. on Computer Vision*, pp. 915–922, 2003.

[373] J. Starck and A. Hilton, "Spherical matching for temporal correspondence of non-rigid surfaces.," in *Int. Conf. on Computer Vision*, 2005.

[374] J. Starck and A. Hilton, "Virtual view synthesis of people from multiple view video sequences," *Graphical Models*, vol. 67, no. 6, pp. 600–620, 2005.

[375] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 246–252, 1999.

[376] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 246–252, 1999.

[377] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, August 2000.

[378] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler, "Speaking with hands: creating animated conversational characters from recordings of human performance," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 506–513, 2004.

[379] A. Sulejmanpašić and J. Popović, "Adaptation of performed ballistic motion," *ACM Trans. Graph.*, vol. 24, no. 1, pp. 165–179, 2005.

[380] J. Sullivan, A. Blake, and J. Rittscher, "Statistical foreground modelling for object localisation," in *European Conference on Computer Vision*, pp. 307–323, 2000.

[381] J. Sullivan and S. Carlsson, "Recognizing and tracking human action," in *European Conference on Computer Vision*, p. 629 ff., 2002.

[382] K.-K. Sung and T. Poggio, "Example based learning for view based face detection," AI Memo 1521, MIT, 1994.

[383] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 20, pp. 39–51, 1998.

[384] S. Tak and H. Ko, "Example guided inverse kinematics," in *International Conference on Computer Graphics and Imaging*, pp. 19–23, 2000.

[385] S. Tak and H.-S. Ko, "A physically-based motion retargeting filter," *ACM Trans. Graph.*, vol. 24, no. 1, pp. 98–117, 2005.

[386] S. Tak, O. Song, and H. Ko, "Motion balance filtering," *Computer Graphics Forum (Eurographics 2000)*, vol. 19, no. 3, pp. 437–446, 2000.

[387] C. J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 677–84, 2000.

[388] C. J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," *Computer Vision and Image Understanding*, vol. 80, no. 3, pp. 349–363, December 2000.

[389] A. Thangali and S. Sclaroff, "Periodic motion detection and estimation via space-time sampling," in *Motion05*, pp. 176–182, 2005.

[390] C. Theobalt, J. Carranza, M. A. Magnor, and H.-P. Seidel, "Enhancing silhouette-based human motion capture with 3d motion fields," in *PG '03: Proceedings of the 11th Pacific Conference on Computer Graphics and Applications*, (Washington, DC, USA), p. 185, IEEE Computer Society, 2003.

[391] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[392] M. E. Tipping, "The relevance vector machine," in *In Advances in Neural Information Processing Systems 12*, pp. 332–388, MIT Press, 2000.

[393] D. Tolani, A. Goswami, and N. I. Badler, "Real-time inverse kinematics techniques for anthropomorphic limbs," *Graphical Models*, vol. 62, pp. 353–388, 2000.

[394] D. Tolani and N. I. Badler, "Real-time inverse kinematics of the human arm," *Presence*, vol. 5, no. 4, pp. 393–401, 1996.

[395] N. Torkos and M. Van de Panne, "Footprint-based quadruped motion synthesis," in *Graphics Interface 98*, pp. 151–160, 1998.

[396] K. Toyama and A. Blake, "Probabilistic tracking in a metric space," in *Int. Conf. on Computer Vision*, pp. 50–57, 2001.

[397] K. Toyama and A. Blake, "Probabilistic tracking with exemplars in a metric space," *Int. J. Computer Vision*, vol. 48, no. 1, pp. 9–19, June 2002.

[398] S. T. Tumer and A. E. Engin, "Three-dimensional kinematic modelling of the human shoulder complex - Part II: Mathematical modelling and solution via optimization," *ASME Journal of Biomechanical Engineering*, vol. 111, pp. 113–121, 1989.

[399] Z. Tu and S. C. Zhu, "Image segmentation by data-driven Markov Chain Monte Carlo," in *Int. Conf. on Computer Vision*, pp. 131–138, 2001.

[400] Z. Tu and S. C. Zhu, "Image segmentation by data-driven Markov Chain Monte Carlo," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 657–673, May 2002.

[401] V. N. Vapnik, *The Nature of Statistical Learning Theory.* Springer Verlag, 1996.

[402] V. N. Vapnik, *Statistical Learning Theory.* John Wiley and Sons, 1998.

[403] D. D. Vecchio, R. M. Murray, and P. Perona, "Decomposition of human motion into dynamics-based primitives with application to drawing tasks," *Automatica*, vol. 39, no. 12, pp. 2085–2098, 2003.

[404] P. Viola and M. Jones, "Robust real-time face detection," in *Int. Conf. on Computer Vision*, p. 747, 2001.

[405] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Int. Conf. on Computer Vision*, pp. 734–741, 2003.

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

[407] P. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 511–518, 2001.

[408] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Computer Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[409] J. J. Wang and S. Singh, "Video analysis of human dynamics - a survey," *Real-Time Imaging*, vol. 9, no. 5, pp. 321–346, 2003.

[410] J. Wang and B. Bodenheimer, "An evaluation of a cost metric for selecting transitions between motion segments," in *SCA '03: Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, (Aire-la-Ville, Switzerland, Switzerland), pp. 232–238, Eurographics Association, 2003.

[411] J. Wang and B. Bodenheimer, "Computing the duration of motion transitions: An empirical approach," in *SCA '04: Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*, (New York, NY, USA), pp. 335–344, ACM Press, 2004.

[412] M. Weber, W. Einhauser, M. Welling, and P. Perona, "Viewpoint-invariant learning and detection of human heads," in *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 20–7, 2000.

[413] Y. Weiss, "Belief propagation and revision in networks with loops," Tech. Rep., Massachusetts Institute of Technology, Cambridge, MA, USA, 1997.

[414] D. J. Wiley and J. K. Hahn, "Interpolation synthesis of articulated figure motion," *IEEE Comput. Graph. Appl.*, vol. 17, no. 6, pp. 39–45, 1997.

[415] A. Witkin and M. Kass, "Spacetime constraints," in *SIGGRAPH '88: Proceedings of the 15th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 159–168, ACM Press, 1988.

[416] A. Witkin and Z. Popović, "Motion warping," in *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 105–108, ACM Press, 1995.

[417] C. R. Wren, A. Azarbayejani, T. J. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE T. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, July 1997.

[418] M.-Y. Wu, C.-Y. Chiu, S.-P. Chao, S.-N. Yang, and H.-C. Lin, "Content-based retrieval for Human Motion Data," in *16th IPPR Conference on Computer Vision, Graphics and Image Processing*, pp. 605–612, 2003.

[419] Y. Wu, T. Yu, and G. Hua, "A statistical field model for pedestrian detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1023–1030, 2005.

[420] Y. Yacoob and L. S. Davis, "Learned models for estimation of rigid and articulated human motion from stationary or moving camera," *Int. J. Computer Vision*, vol. 36, no. 1, pp. 5–30, January 2000.

[421] M. Yamamoto, A. Sato, S. Kawada, T. Kondo, and Y. Osaki, "Incremental tracking of human actions from multiple views," in *CVPR '98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (Washington, DC, USA), p. 2, IEEE Computer Society, 1998.

[422] K. Yamane and Y. Nakamura, "Natural motion animation through constraining and deconstraining at will," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 3, pp. 352–360, 2003.

[423] J. Yang, Z. Fu, T. N. Tan, and W. M. Hu, "A novel approach to detecting adult images," in *Proceedings IAPR International Conference on Pattern Recognition*, pp. 479–482, 2004.

[424] W. Yan and D. A. Forsyth, "Learning the behaviour of users in a public space through video tracking," in *CVPR*, 2004. In review.

[425] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," in *Exploring artificial intelligence in the new millennium*, pp. 239–269, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[426] J. Zhao and N. I. Badler, "Inverse kinematics positioning using nonlinear programming for highly articulated figures," *ACM Trans. Graph.*, vol. 13, no. 4, pp. 313–336, 1994.

[427] L. Zhao and N. Badler, "Gesticulation behaviors for virtual humans," in *PG '98: Proceedings of the 6th Pacific Conference on Computer Graphics and Applications*, (Washington, DC, USA), p. 161, IEEE Computer Society, 1998.

[428] L. Zhao and C. E. Thorpe, "Stereo- and neural network-based pedestrian detection," *Intelligent Transportation Systems*, vol. 1, no. 3, pp. 148–154, September 2000.

[429] S. C. Zhu, R. Zhang, and Z. Tu, "Integrating bottom-up/top-down for object recognition by data driven Markov Chain Monte Carlo," in *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 738–745, 2000.

[430] V. B. Zordan and J. K. Hodgins, "Motion capture-driven simulations that hit and react," in *ACM SIGGRAPH Symposium on Computer Animation*, pp. 89–96, July 2002.

[431] V. B. Zordan and J. K. Hodgins, "Tracking and modifying upper-body human motion data with dynamic simulation," in *Computer Animation and Simulation '99*, September 1999.

[432] M. Zyda, J. Hiles, A. Mayberry, C. Wardynski, M. Capps, B. Osborn, R. Shilling, M. Robaszewski, and M. Davis, "Entertainment R&D for defense," *IEEE Computer Graphics and Applications*, pp. 28–36, 2003.

[433] M. Zyda, A. Mayberry, C. Wardynski, R. Shilling, and M. Davis, "The MOVES institute's America's army operations game," in *Proceedings of the ACM SIGGRAPH 2003 Symposium on Interactive 3D Graphics*, pp. 217–218, 2003.