Contents

1 Introduction

These notes are a heavily precised version of [104]; we did not get figure permissions sorted out in time to sell copies of those notes here. Most of the citations of that review are preserved.

2 Tracking: Fundamental Notions

Good summary histories include [52, 54, 159]; comprehensive reviews of technique in this context include [22, 35, 115]). Coarse scale trackers include [370, 328].

2.1 Tracking by detection

Faces are natural candidates for tracking by detection. Background subtraction is described in [363, 287, 208, 129, 60, 109, 124, 152, 153, 326, 327]. 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 [35]).

Shadows are a perennial nuisance for background subtraction, but this can be dealt with using a stereoscopic reconstruction, as Haritaoglu *et al.* show ([128]; see also [154]).

2.2 Tracking using Flow

Flow based tracking has the advantage that one doesn't need an explicit model of the appearance of the template. Cardboard people is due to [160]. Families of flow due to walking can be found in [34]. Yacoob and Davis build a view independent parametric flow field models to track views of walking humans [366]. 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 [128].

2.3 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, 50].

2.4 Tracking with Probability

There is an account in [105]. Standard references are [272, 115, 35, 22].

The particle filter is a current favorite method for dealing with multi-modal densities; see [86, 150, 193, 272]. 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 [31]. Daum extends the class of models for which this is the case ([80, 81]; see also [283] for an application and [95] for a comparison with the particle filter).

3 Tracking: Relations between 3D and 2D

3.1 Kinematic Inference with Multiple Views

Important examples include [64, 162, 59, 252, 83, 84, 49, 50, 88, 87, 89, 289, 290, 288, 320, 131, 323, 321, 322, 367, 59, 339]. The most comprehensive and recent discussion of 3D reconstruction from multiple views appears in two papers [65, 66]. Curiously, although Mori and Malik have shown that one can obtain landmark positions automatically [220], there appears to be no multiple view reconstruction work that identifies landmarks in several views and builds a geometric reconstruction this way. Reducing configuration ambiguity is one reason

to use multiple cameras; another is to keep track of individuals who move out of view of a particular camera. Currently, this is done at a coarse scale, where people are blobs (e.g. [55, 166, 213]).

3.2 Lifting to 3D

The way that people are imaged means that there are very few cases where a scaled orthographic camera model is not appropriate. One such case to keep in mind is a person pointing towards the camera; if the hand is quite close, compared with the length of the arm, one may see distinct perspective effects over the hand and arm and in extreme cases the hand can occlude much of the body.

Regard each body segment as a cylinder, for the moment of known length. If we know the camera scale, and can mark each end of the body segment — we might do this by hand, as Taylor [335, 336] does and Barrón and Kakadiaris [25, 26] do, or by a strategy of matching image patches to marked up images as Mori and Malik do [220, 221] — then we know the cosine of the angle between the image plane and the axis of the segment, which means we have the segment in 3D up to a twofold ambiguity and translation in depth. We can reconstruct each separate segment and obtain an ambiguity of translation in depth (which is important and often forgotten) and a two-fold ambiguity at each segment.

For the moment, assume we know all segment lengths and the camera scale. We can now reconstruct the body by obtaining a reconstruction for each segment, and joining them up. Each segment has a single missing degree of freedom (depth), but the segments must join up, meaning that we have a discrete set of ambiguities. Depending on circumstances, one might work with from nine to eleven body segments (the head is often omitted; the torso can reasonably be modelled with several segments), yielding from 512 to 2048 possible reconstructions. These ambiguities persist for perspective images. Mori and Malik deal with discrete ambiguities by matching [220, 221].

Sminchicescu and Telea compare silhouette to projection to produce a reconstruction from a single view ([310]; see also [305]). Randomized search is a reasonable strategy for attacking the minimization. Sminchisescu and Triggs describe various methods to bias the likelihood function searched by a sampler so that the state will move freely between local minima [313, 312, 316]. Sminchisescu and Triggs exploit an explicit representation of kinematic ambiguities to help this search, by making proposals for large changes of state that have a strong likelihood of being good [315]. Lee and Cohen use a markov chain Monte Carlo method to search the likelihood, using both a set of image detectors and a model of kinematic ambiguities to propose moves; this gives a set of possible reconstructions for the upper body [179] and the whole body [180].

One can lift by nearest neighbour regression. Athitsos and Sclaroff determine 20 kinematic configuration parameters from an image of a hand by matching the image to a set of examples [17, 18]. A match from a short 2D track to a short 3D track might not be ambiguous [140, 141, 267, 266].

Rosales and Sclaroff use of a collection of local experts ("specialized mappings") to regress hand configuration against image appearance [277]. Shakhnarovich *et al*. use parameter sensitive hashing [291]; a version of this approach can produce full 3D shape estimates [123]. Liu *et al*. demonstrate a full body reconstruction from silhouettes in five views using a similar regression model; the reconstruction is not evaluated directly, but is used to control motion synthesis [271].

Agarwal and Triggs observe that the pose in the previous frames, if correctly computed, should give a good guide to the current pose — one is unlikely to jump from sheet to sheet in a single frame [1, 4].

3.3 Multiple Modes, Randomized Search and Human Tracking

The core method is the particle filter. We have refrained from an exposition, as the idea is described in detail in several recent publications (e.g. [86, 272, 193, 150]).

Particle filters should be seen as a form of randomized search. One starts a set of points that tend to be concentrated around large values of the posterior. These are pushed through the dynamical model, to predict possible configurations in the data. The result is a sampled representation of the prior. The predictions are compared to the data, and those that compare well are given higher weights, yielding a sampled representation of the posterior. This simple view provides some insight into why particle filters in their most basic form are not particularly well adapted to kinematic tracking.

There is a problem with dimension. The state vector for most kinematic tracking problems must be high dimensional. One expects to encounter at least 20 degrees of freedom (one at each knee, two at each hip, three at each shoulder, one at each elbow and six for the root) and quite possibly many more. This means that mismatches between the prior and the likelihood can generate serious problems. Such mismatches are likely for three reasons.

First, the body can move quickly and unexpectedly, meaning that probability must be quite widely spread in the prior to account for large accelerations. It is hard to be clear on how much uncertainty there is in the state of the body at some time given the past, and there are fair arguments either way. However, fast movements do occur, and current methods are forced to have fairly diffuse dynamical models to cope with them.

Second, the likelihood has multiple peaks, which can be very narrow. Narrow peaks occur because some body segments — forearms are a particularly nasty example — have relatively small cross-section in the image, and so only a small range of body states will place these segments in about the right image configuration. Multiple peaks occur because there tend to be numerous objects that look somewhat like body segments (long, narrow, parallel sides, constant colour). We are now using the predictions of the prior to find the largest narrow peak in a high-dimensional likelihood — for this to have any hope of success, the predictions need to be good or to occur in very large numbers. But we know the predictions will be poor, because we know people can generate fast, unexpected movements.

Third, detectors used to produce a likelihood model may be inaccurate. This can result in small errors in inferred state, which in turn produce potentially large changes in state from frame to frame. As Sminchicescu and Triggs point out ([314], p. 372), this suggests using a relatively diffuse dynamical model as an insurance policy.

The key idea in particle filters is the randomized search. One might abandon, or at least de-emphasize, probabilistic semantics, and focus on building an effective search of the likelihood. The key difficulties are that the peaks in the likelihood are narrow (and so easy to miss) and that the configuration space is high-dimensional (so that useful search probes may be difficult to find). The narrow peaks in the likelihood could be dealt with by annealing, and good search probes may be found by considering the ambiguity of 3D reconstructions.

There are a series of approaches to deal with problems created by the dimension of the state space. First, we could refine the search using importance sampling methods. Second, we could use sequential inference methods to obtain more efficient samples of the prior. Third, we could build lower-dimensional dynamical models. Finally, we could build more complex searches of the likelihood.

Isard and Blake use importance sampling to track hands and forearms [151], using a skin detector to build an importance function. Rittscher and Blake use importance sampling methods to track contours of motions drawn from two classes (pure jump and half star jump); the tracker maintains a representation of posterior on the motion class, which can be used to distinguish between motion classes successfully [273]. Forsyth uses edge detector responses as a source of proposal mechanisms to find simple boundaries [100], and Zhu *et al*. — who call the approach data driven MCMC — use image observations to propose segmentations [345, 344, 378]. We are not aware of the method being used for kinematic tracking.

Sigal *et al*. use loopy propagation ([225, 356, 373]), representing messages passed between nodes using a set of particles [299], to track a 2D template with links in both space and time.

MacCormick and Isard track hands using partitioned sampling [201]. MacCormick and Blake use this method to track multiple objects [199, 200], where one needs a method to avoid both tracks lying on the same object.

Sidenbladh *et al*. build a 3D model of a human as a kinematic chain, with state encoded as the configuration and velocity of each element of this chain with respect to its parent, and the root with respect to the camera [298].

Choo and Fleet implement a more extensive search of the posterior using a Markov chain Monte Carlo (MCMC) method [67].

One difficulty with a sampled model of the posterior is that we don't know if there are larger values of the posterior close to each sample. We could regard each sample as a plausible start point for a search of the posterior. We are now no longer building a set of particles that explicitly represents the posterior in the sense above, but are using multiple states to represent the prospect that the posterior is multi-modal. Each state lies on a mode in the posterior, and we attempt to ensure that all modes have a state. The origins of this approach lie with Cham and Rehg [62], who use it to track a 2D kinematic model of the body.

More complex models of the posterior appear in [308, 307, 306].
Sminchisescu and Triggs elaborate this search by analysis of the Hessian of the log-posterior [311, 314].

4 Tracking: Data Association for Human Tracking

Early human trackers, which used quite straightforward matching methods, (for example, Hogg's 1983 paper [137]; Rohr's 1994 tracker [275]) could produce kinematic tracks for people moving without sudden accelerations on reasonably simple, high-contrast backgrounds if started manually. The advantages of a known, simple background have been thoroughly explored. The more recent trackers we have described use more complex inference machinery, but without any great change in competence.

Improvements in competence seem to have come with increased attention paid to tracking by detection schemes. These are well established in, say, face tracking. For example, one can build a fairly satisfactory face tracker by simply running a face detector on frames, and linking over time; smart linking schemes built around affine invariant feature patches can result in very satisfactory tracks [304]. Tracking by detection is now capable of building good human kinematic tracks, without relying on background subtraction.

4.1 Detecting Humans

Approximately half-a-million pedestrians are killed by cars each year (1997 figures, in [112]). Papageorgiou and Poggio represent 128x64 image windows with a modified wavelet expansion, and present the expansion to a **support vector machine** (SVM), which determines whether a pedestrian is present [244]. SVM's are classifiers, trained with positive and negative examples. For a brief informative discussion of SVM's see [347] or [69]. More extensive information appears in [284, 292, 346], and discussion in the context of a variety of other classifiers is in [130]. The training data consists of windows with and without people in them; each positive example is scaled such that the person spans approximately 80 pixels from shoulder to foot. A variety of image representations are tested, with the modified wavelet expansion applied to colour images performing significantly better than wavelet expansions applied to grey-level images, low resolution pixel values for grey-level images, principal components analysis representations of grey-level images, and the like. The strength of these wavelet features appears to be that they emphasize points that are, rather roughly, outline points. This yields a method for exploiting the restricted range of contours without explicitly encoding contour templates. The wavelet expansion can be reduced in dimension to obtain a faster, though somewhat less accurate, matcher. There are several variants of this approach in the literature [233, 234, 239, 240, 241, 243].

Zhao and Thorpe use stereopsis to segment the image into blocks, then present each block to a neural network [376]. Gavrila describes an approach that matches image contours against a hierarchy of contour templates using a chamfer distance [111]. Gavrila *et al.* describe an improved version of this method, using stereo cues and temporal integration [113]. Broggi *et al.* describe a method that uses vertical edges, the characteristic appearance of the head and shoulders, and background subtraction to identify pedestrians [51].

Wu *et al.* build random field models of image windows with and without a pedestrian, and then detect using a likelihood ratio [365]. Dalal and Triggs give a comprehensive study of features and their effects on performance for the pedestrian detection problem [76]. The method that performs best involves a histogram of oriented gradient responses (a **HOG** descriptor). The paper compares HOG descriptors with the original method of Papageorgiou and Poggio [244]; with an extended version of the Haar wavelets of Mohan *et al.* [215]; with the PCA-Sift of Ke and Sukthankar ([161]; see also [211]); and with the shape contexts of Belongie *et al.* [28]. There is considerable detailed information on tuning of features.

Pedestrians also tend to move in quite restricted ways — they are typically either standing or walking. Niyogi and Adelson point out that, if one forms an **XYT image** — a stack of frames, registered as to camera motion, originally due to Baker [21] — these motions produce quite distinctive structures, which can be used to identify motions [230] or recover some gait parameters [229]. Polana and Nelson consider spatial patterns of motion energy, which also have a characteristic structure [259]. There is a substantial literature on the characteristic appearance of human motion fields; a good start is [45, 185, 186, 187, 253, 254, 255, 257, 258, 260]. Particular

efforts have been directed to periodic motion; one might consult [63, 72, 73, 74, 75, 121, 122, 183, 191, 192, 285, 286, 337].

This characteristic structure can be used to detect pedestrians in a variety of ways. Papageorgiou and Poggio compute spatial wavelet features for the frame of interest and the four previous frames, stack these into a feature vector, and present this feature vector to an SVM, as above [242]. The result is a fairly significant improvement in detection rate for a given false positive rate.

Viola *et al.* use explicit motion features — obtained by computing spatial averages of differences between a frame and a previous frame, possibly shifted spatially — and obtain dramatic improvements in detection rates over static features ([349, 350]; see also the explicit use of spatial features in [71, 236, 237], which prunes detect hypotheses by looking for walking cues).

Dimitrijevic *et al.* build a spatio-temporal template as a list of spatial templates in time-order [85].

One might build templates with complex internal kinematics. The core idea is very old (for example, one might consult [6, 7, 33, 132, 203, 235]) but the details are hard to get right and important novel formulations are a regular feature of the current research literature. The advantage of these 2D kinematic templates is that they are relatively easy to learn.

The first difficulty is that simply identifying the body parts can be hard. This is simplified if people are not wearing clothing, because skin has a quite distinctive appearance in images. Forsyth *et al.* then search for naked people by finding extended skin regions, and testing them to tell whether they are consistent with body kinematics [102, 103]. The method is effective on their dataset (and can be extended to find horses [101]), but is not competitive with more recent methods for finding "adult" images (which typically use whole-image features [12, 43, 157, 371]). Ioffe and Forsyth formalize this process of testing, and apply it to relatively simple images of clothed people [147, 149]. Their procedure builds a classifier that accepts or rejects whole assemblies of body components; this is then projected onto factors to obtain derived classifiers that can reject partial assemblies that could never result in acceptable complete assemblies. Sprague and Luo use this approach to find clothed people in more complex images, by reasoning about image segments [319].

Mohan *et al.* use a discriminative approach not only to identify good assemblies of parts (as above), but also to find body parts [215].

Felzenswalb and Huttenlocher show how one may use **distance transforms** to speed this process up substantially [96, 97]. Kumar *et al.* extend this model to incorporate boundaries into the likelihood and use loopy belief propagation to apply it to arbitrary graphs (rather than trees); the method is applied to pictures of cows and horses [174].

Ronfard *et al.* use a discriminative model to identify body parts, and then a form of generative model to construct and evaluate assemblies [276]. Mikolajczyk *et al.* use discriminative part detectors, applied to orientation images and built using methods similar to those of Viola and Jones, to identify faces, head-and-shoulders, and legs [212]. Micilotta *et al.* use discriminative methods to detect hands, face and legs; a randomized search through assemblies is used to identify one with a high likelihood, which is tested against a threshold [210]. Similarly, Roberts *et al.* use a randomized search to assemble parts; parts are scored with a generative model, which is used to obtain a proposal distribution for joints [274].

Representing a body by segments may not, in fact, be natural; our goal is effective encoding for recognition, rather than disarticulation. One might represent people by image patches chosen to be good at representing people. Leibe *et al.* have built the best known pedestrian detection system using this approach [182].

4.2 Tracking by Matching Revisited

Most probabilistic tracking algorithms must compute the likelihood of some image patch conditioned on the presence of a model at some point. The easy model to adopt is to produce a template for the patch from the model parameters, subtract that template from the image, and assume that the result consists of independent noise — that is, that the value at each pixel is independent. Whether it is wise to use this model or not depends on how the template is produced — for example, a template that does not encode illumination effects is going to result in a residual whose pixel values are not independent from one another (see Sullivan *et al.* for this example [330]), and so the likelihood model is going to significantly misestimate the image likelihood.

The problem occurs in a variety of forms. For example, if one represents an image patch with a series of filter outputs (after, say, [296, 297]), each element is unlikely to be independent and errors are unlikely to be independent. Sullivan *et al*. describe the problem, and demonstrate a set of actions (including building an illumination model and estimating correlation between filter outputs) that tend to ameliorate it, in the context of face finding [330]. Roth *et al*. build likelihood models for vectors of filter outputs using a **Gibbs model** (known in other circles as a **maximum entropy model** or a **conditional exponential model**) [281]. Their method is trained using an algorithm due to Liu *et al*. ([194]; see also [188], and one might compare variants of iterative scaling [32, 78, 158, 249, 280]). There is some evidence that the likelihood produced using this model is more tightly tuned to — in their example — the presence and location of a leg. The model is used by Sigal *et al*. [299].

While it is clear that there is an issue here, it is a bit uncertain how significant it is. I am not aware of clear evidence that better tracking or localization results from being careful about this point, and am inclined to believe that the rough-and-ready nature of current likelihood models is not a major problem.

Toyama and Blake encode image likelihoods using a mixture built out of templates, which they call **exemplars** [343, 342].

Spatial templates can be used to identify key points on the body. Sullivan and Carlsson encode a motion sequence (of a tennis player) using a small set of templates, chosen to represent many frames well [331]. Loy *et al*. show that such transferred keypoints can be used to produce a three dimensional reconstruction of the configuration of the body [198].

The advantage of a tree-structured kinematic model, that one can use dynamic programming for detection, extends to a mixture of such trees. However, adding temporal dependencies produces a structure that does not allow for simple exact inference, because the state of a limb in frame t has two parents: the state in time $t-1$, and the state of its parent in frame t. Ioffe and Forsyth attack this problem with a form of coordinate ascent on $P(\mathbf{X}_0, \ldots, \mathbf{X}_k | \mathbf{Y}_0, \ldots, \mathbf{Y}_k)$ [148].

This difficulty is quite often ignored, apparently without major consequences. Mori and Malik use no dynamical model, detecting joints repeatedly in each frame; the result is a fair track of a fast-moving skater [220]. Lee and Nevatia use a Markov model of configuration (but not of appearance), where each body configuration depends only on the previous configuration [181].

Agarwal and Triggs build a set of dynamical models, each of which explains a cluster of motion data well; a mixture of these models is then used to propose the 2D configuration in the $i+1$'th frame from the state in the i'th frame [2]. The predictions are refined by an optimization method, as in [314].

Some advantages of a tracking by detection framework and the difficulties that result from relying on a dynamical model are: First, recovery from occlusion, people leaving frame or dropped frames is straightforward; because we know what each individual looks like, we can detect the individual when they reappear and link the tracks (this point is widely acknowledged; see, for example, [77, 224]). Second, track errors don't propagate; when a segment is misidentified in a frame, this doesn't fatally contaminate the appearance model. Difficulties occur if different individuals look the same (although one may be able to deal with this by instancing) or if we fail to build a model.

Ramanan *et al*. demonstrate an alternative method of building a model using a detector [269].

Song *et al*. use a variant of tree-structured models to identify human motion. They identify local image flows at interest points in an image, using the Lucas-Tomasi-Kanade procedure for identifying and tracking localizable points [317, 318].

Sminchisescu *et al*. see tracking by matching as a discriminative problem [309].

4.3 Evaluation

There is no current consensus on how to evaluate a tracker, and numerical evaluations are relatively rare (there are several numerical evaluations of lifting to 3D; see, for example, [1, 180, 179]). In our opinion, it is insufficient to simply apply it to several video sequences and show some resulting frames (a practice fairly widespread until recently). Counting the number of frames until the tracker fails is unhelpful: First, the tracker may not fail. Second, the causes of failure are more interesting than the implicit estimate of their frequency, which may be poor. Third, this sort of test should be conducted on a very large scale to be informative, and that is seldom practical. Trackers are — or should be — a means to a larger end, and evaluation should most likely focus on this

point. In this respect, trackers are probably like edge-detectors, in that detailed evaluation is both very difficult and not wholly relevant. What matters is whether one can use the resulting representation for other purposes without too much incovenience.

A fair proxy for this criterion is to regard the tracker as a detector, and test its accuracy at detection and localization. In particular, if one has a pool of frames each containing a known number of instances of a person, one can (a) compare the correct count with the tracker's count and (b) check that the inferred figure is in the right place. The first test can be conducted on a large scale without making unreasonable demands on human attention, but the second test is difficult to do on a large scale. Ramanan and Forsyth use these criteria; their criterion for whether a particular body segment is in the right place is to check the predicted segment intersects the image segment (which is a generous test) [268, 299].

Lee and Nevatia evaluate reprojection error for the tracked person [181]. There might be some difficulty in using this approach on a large scale. Sigal *et al* construct a 3D reconstruction, and so can report the distance in millimetres between the true and expected positions (predicted from the posterior) of markers [299]. Agarwal and Triggs give the RMS error in joint angles compared to motion capture on a 500 frame sequence [3].

There is little consensus on what RMS errors actually mean in terms of the quality of reported motion. There is some information in [13], which evaluates compression of motion capture; this boils down to the fact that very small RMS errors in joint position indicate that the motion is acceptable, but quite large errors are hard to evaluate. There is no information on what errors in joint angle mean.

5 Motion Synthesis and Animation

Variations in rendering style alter a viewer's perception of motions [133, 134]. As the characters' appearance improves so too does viewer expectations concerning the characters' motion. More realistic characters with a more interesting range of behaviors present substantial challenges.

5.1 Motion capture

Reviews of available techniques in motion capture appear in, for example [41, 119, 197, 209, 214, 300]. Some very fast motions can be captured only with specialized stroboscopic equipment [338].

5.2 Footskate

Kovar *et al*. assume that constraints that identify whether heel or toe of which foot is planted in which frame (but not where it is planted) are available [173] and then clean up with inverse kinematics (see [117, 116, 127, 202, 19, 340, 374, 20, 295, 219, 245].

Ikemoto *et al*. demonstrate that one can clean up footskate introduced by editing and so on automatically [143].

5.3 Resolving Kinematic Ambiguities with Examples

The danger here is that one may obtain poses that do not look human. Motion editing deals with this by being interactive, so that an animator who doesn't like the results can fiddle with the constraints until something better appears (see also [248]). An alternative is to allow relatively few degrees of freedom — for example, allow the animator to adjust only one limb at a time — or to require similarity to some reference pose [332, 368, 375]. This isn't always practical. An alternative, as Grochow *et al*. demonstrate, is to build a probabilistic model of poses and then obtain the best pose [125].

While motion editing does not offer direct insight into representing motion, the artifacts produced by this work have been useful, and it has produced several helpful insights. The first is that it is quite dangerous to require large changes in a motion signal; typically, the resulting motion path does not look human (e.g. [119]). The second is that enforcing some criteria — for example, conservation of momentum and angular momentum [294]; requiring the zero-moment point lies within the support polygon [79, 169, 294] — can improve motion editing results quite significantly. However, note that one can generate bad motions without violating any of these constraints, because motion is the result of extremely complex considerations. The third is that requiring motion lie close to examples can help produce quite good results.

5.4 Motion Signal Processing

A variety of signal processing operations on motion are successful, an observation originating with Bruderlin and Williams [53]. These include temporal scaling, alignment [170], blending [170, 144, 53, 278, 358, 171, 175, 246, 247], constant offsets [145] and filtering [53]. One may blend motions with simple physical models [16]. It remains hard to know when two motions will blend successfully.

There is considerable recent interest in finding motions that are similar to some query [171, 99, 223, 364, 57, 164, 163].

5.5 Motion Graphs

Core motion graph papers include [172, 177, 15]. Annotation based synthesis is from [14]. One can evaluate motion graphs [270]

5.6 Motion Primitives

There are numerous attempts to encode motion in terms of primitives [98, 278, 204, 205, 348]. Motion clusters and segments well [23, 14]. One can apply dimension reduction techniques to produce such an encoding [156]. Chai and Hodgins demonstrate a form of **video puppetry** — where an animated figure is controlled by observations of an actor — using relatively few markers; this approach most likely works because motions tend to be confined to a low dimensional subspace [61]. Safonova *et al.* are able to produce plausible figure animations using optimization techniques confined to a low-dimensional space (see [282]).

Li *et al.* segment and model motions simultaneously using a linear dynamical system model of each separate primitive and a Markov model to string the primitives together by specifying the likelihood of encountering a primitive given the previous primitive [184].

5.7 Enriching a Motion Collection

Gleicher shows that one can usefully edit motions — typically, so that they meet constraints that are a small revision of constraints met by the original motion — by adding a displacement [118]. Lee and Shin obtain a more manageable optimization problem by representing the motion as a hierarchical B-spline [178]. Witkin and Popović modify motions using parametric warps, so that they pass through keyframes specified by an animator [362]. Shin *et al.* use similar methods to touchup motion to meet physical constraints (for example, motion not in contact is ballistic and preserves angular momentum), while sacrificing physical rigor in the formulation for speed [294]; see also [333, 334]). Motion editing in this way is useful, and there are several other systems; a review appears in [120].

Ikemoto and Forsyth build new motions from old by cutting arms or upper bodies off one motion and attaching them to another [145]. Pullen and Bregler [264] built a motion synthesis system that allows animators to sketch part of the motion of the body, and then uses a non-parametric regression method to fill in the details. Controllers that track motion data provide a useful mechanism for smoothing recorded errors while also adjusting for disturbances not present in the recorded motion [93, 261, 380, 381]. Other approaches make use of hand designed or optimized controllers that operate independently from recorded motion [91, 92, 126, 136, 263]. Building controllers that generate human-like motion remains an open research problem.

5.8 Motion from Physical Considerations

Witkin and Kass introduced the use of **variational methods**, widely known as **spacetime constraints** [361]. The actual minimization process might be extremely difficult. There is some reason to believe that a coarse-to-fine representation is useful [196]. One may simplify optimization difficulties by choosing simplified characters (e.g. [94, 261, 263, 341]; freefall diving is a particular interest [70, 195]) or by exploiting interaction with an animator (e.g. [68]). Ngo and Marks produce motions for quite complex characters using spacetime optimization by building motions out of **stimulus-response** pairs — parametric packets of motion that are triggered by some parametric test ([227, 228]; see also [207] for other motions built out of packets). The precise set of packets, and the parameters of those packets, are chosen using search by a genetic algorithm (see also the work of Sims [301]). There is no claim that these motions necessarily appear human.

Liu *et al.* show a method to obtain simulation parameters from examples [189]. Rose *et al.* generate **motion transitions** — short sequences of motion that join specified frames "naturally"— using an optimization procedure that minimizes the total squared torque moving the upper body [279]. Anderson and Pandy describe a simulation of one step of a walk for a highly detailed dynamic model that produces (using months of supercomputer time) a pattern of muscle activations that minimize an effort criterion and also look like human muscle activation patterns ([11]; see also [238]).

5.8.1 Simplified Characters

Popović and Witkin use characters with simplified kinematics, and model muscle forces explicitly (the muscle is modelled as a **proportional-derivative controller** attempting to drive a degree of freedom to a setpoint) [263].

5.8.2 Modified Physics

Liu and Popović produce character animations from rough initial sketches using an optimization method by breaking the motion into phases, simplifying the physical constraints, and, where necessary, exploiting the animator's input [190].

There is a real advantage to not constraining forces and torques and not allowing them to participate in the objective function: one does not need to compute them. This means that computing various Jacobians that arise in the optimization procedure can be made linear (rather than quadratic) in the number of degrees of freedom, as Fang and Pollard show [94].

5.8.3 Reduced Dimensions

Safonova *et al*. describe a method for synthesizing motions from variational considerations using a dimension reduced representation of configuration [282].

5.8.4 Modifying Existing Motions

Hodgins and Pollard describe scaling rules that allow a motion that applies to one character to be transferred to another character, using methods of dimensional analysis ([135]; for dimensional analysis, see [24]). Sulejmanpašić and Popović modify existing motions to obtain revised motions that meet animator demands using a full dynamical model ([329]; see also [262], which describes a search method to obtain parameters of a rigid body simulation that is similar to a sketch).

6 Describing Activities

These are early notes for a draft chapter on activities, for a projected volume following the lines of [104]. As a result, they tend to emphasize older material and are not really assembled according to a plan. However, the citations are useful, which is why I circulate. I've added a short list of recent citations at the end. DAF

Understanding what people are doing is one of the great unsolved problems of computer vision. A fair solution opens tremendous application possibilities, including: improved surveillance systems; a better understanding of what people do in public; better architectural design; and better human computer interfaces. While there has been extensive study of this topic, it still isn't terribly well understood. One can obtain statistics of some behaviours from coarse scale tracks (e.g. for car parks, see [328]; for architectural domains, see [370]). But understanding activities that depend on detailed information about the body is still hard. We contend that the major difficulties have been (a) that good kinematic tracking is hard; (b) that models typically have too many parameters to be learned directly from data; and (c) for much everyday behaviour, there isn't a clear taxonomy into which to classify observations.

There is a long tradition of research on interpreting activities in the vision community (see, for example, the extensive survey in [142]). There are three major threads. First, one can use temporal logics to represent crucial order relations between states that constrain activities. Second, one can use spatio-temporal templates to identify instances of activities. Third, one can use (typically, hidden Markov) models of dynamics.

6.1 What should an Activity Representation do?

There appear to be a series of quite different cases in activity recognition. First, we distinguish between short, medium and long timescales. Second, we distinguish between motions that can be sustained (walking, running, waving) and motions that have a localizable character (catch, throw, punch, kick). Since we want our complex, composite motions to share a vocabulary of base units, we use the kinematic configuration of the body, limb velocities, and perhaps accelerations as distinctive features at **short** timescales — which might be of the order of a small number of frames. We define **acts** to be frame labels that can be decided on such very short timescale features — such labels, (for example, walk-right-leg-stance-left-leg-swing) tend not to have directly useful semantics.

At **medium** timescales, we have **activities** — motions like walking, running, jumping, standing, waving, whose temporal extent can be short (but may be long); such motions are typically composites of multiple acts. Furthermore, activities can be sustained for long periods. We use the term **actions** for motions that have a localizable character and require medium timescales to identify. Both actions and activities may be difficult to identify with only a few frames but are relatively easy to identify from hundreds to thousands of frames. Both actions and activities allow a degree of composition — for example, one could walk and scratch at the same time.

One's interpretation of a view of moving humans is strongly affected by objects nearby. For example, a person standing in an isolated field may be behaving strangely; the same person in the same configuration next to a bus stop is waiting for a bus. We believe that the most natural level at which to start inserting considerations of context into activity recognition is that of activities — where one can pool object detector responses over a long enough sequence of frames to expect quite good behaviour — and define the next layer of the representation to be **motions in context**. Context applies to both activities and actions. These occur at medium timescales, but the nature of the motion in context is determined by both the actions in the sequence and the response of object detectors. We use the term **behaviour** to cover motions at **long** timescales — typically, behaviours such as fighting, exercising or visiting an ATM might be composed of a selection of different motions in context, linked up by activities, and organized in a variety of possible ways and meeting a variety of constraints on temporal ordering. It has been recognized for some time that there are other helpful distinctions (e.g. Bobick [37] distinguishes between movements, activity and actions, corresponding to longer timescales and increasing complexity of representation; some variants are described in two useful review papers [5, 114]).

6.1.1 Necessary Properties of an Activity Representation

The big goal is a theory and mechanism for recognizing a wide range of behaviours. There are some important constraints on solutions to this problem. First, we expect that typical behaviours are a composite of many activities, and this composite is not unique — the same behaviour may be represented by multiple sequences of actions, as long as these sequences observe an internal structure. For example, one may scratch or groom at any time while visiting an ATM, but one must type a PIN before retrieving money, and insert a card before typing a PIN (notice that one can't retrieve a card before inserting it, but at some machines one might retrieve the card before typing a PIN; at others, the card is retrieved after typing the PIN and before recovering money). Each activity may itself be one of several different composites of multiple actions, in the same way. Each action might also have compositional properties — for example, one may walk with three-quarters of one's body while scratching with the fourth limb. The modelling strategy must respect both this hierarchical structure and the compositional nature of motion.

Second, we expect that there is not labelled data for each possible case; we cannot simply learn models without any human interaction. This applies to models of actions, activities and behaviours. This difficulty is created by the compositional nature of human motion; the sheer richness of available motions defeats pure data-driven strategies. An important criterion for choosing a modelling strategy is that it be easy for humans to author and to assess rich models quickly. Such models should be amenable to parameter learning from data, but it should not be necessary to see an example of every possible instance of a behaviour to build a model.

Third, we expect that the supervised data that is available may be marked up somewhat inaccurately. Typically, a behaviour will be marked up with activity names (an activity with actions, respectively), but the boundaries of the markup are unlikely to be accurate. We expect the learning algorithm to be robust to some segmentation noise.

Finally, we models should have the property that basic activities with the model — model building, composition, and inference — is relatively straightforward. In this, we follow the experience of the statistical natural language community, that trading expressiveness in models for simplicity of authoring and inference is often advantageous.

6.1.2 What Data is Available?

An important part of design here is to keep into account what kinds of data are easy to obtain and what difficult, so as to plan model authoring around what is practical. Experience suggests that it is possible to get from minutes to hours of reasonable quality motion capture data; relatively few minutes of video labelled as to actions (these labels are very difficult to produce because they require frame accuracy); minutes to hours of video labelled in reasonable detail with respect to activities and behaviours, accepting poor temporal resolution in the labels; and of the order of months of public observation video. It is relatively straightforward to look at large volumes of labelled motion capture data and correct labels, not least because one can observe many frames simultaneously (e.g. see [14]).

One important source of difficulty is that it is hard to tell which aspects of behaviour should be modelled accurately in order to perform useful tasks. Resolving this requires (a) study of ideas in sufficient generality that they transfer between tasks and (b) some example tasks. But the selection of example tasks is not innocuous. In particular, a distinctive feature of everyday activity is the number of behaviours that appear familiar, but for which the observer may not know a word or even a compact description. In contrast, in some domains (e.g. ballet [56]; gymnastics [90];tai chi [46, 48]; tennis [331]; walking [58]) there are quite specific vocabularies that refer to very precisely delineated behaviours. This is an advantage for building demonstration systems, because one can evaluate them, but may avoid the real difficulty, which is that for most activities we want to classify the activity without knowing a precise or canonical set of classes.

6.2 Miscellaneous Methods

6.2.1 Activity Representation Methods based around Temporal Logics

Pinhanez and Bobick [250, 251] describe a method for detecting what we have called behaviours using a representation derived from Allen's interval algebra [10], a method for representing temporal relations between a set of intervals. One authors a description of the behaviour in terms of primitives, which are indivisible and occupy temporal intervals. The description incorporates a set of legal relations between the primitive intervals; a description is consistent if at least one set of intervals, together with an allocation of those intervals to primitives, satisfies it. One determines whether an event is past, now or future by solving a consistent labelling problem, allowing temporal propagation. There is no dynamical model — sets of intervals produced by processes with quite different dynamics could be a consistent labelling; this can be an advantage at the behaviour level, but probably is a source of difficulties at the action/activity level. These papers do not show the method applied to noisy detectors; there are results using simulated detectors on real data.

Siskind [303, 302] describes methods to infer activities related to objects — such as throw, pick up, carry, and so on — from an event logic formulated around a set of physical primitives —- such as translation, support relations, contact relations, and the like — from a representation of video. A combination of spatial and temporal criteria are required to infer both relations and events, using a form of logical inference. The methods are focussed on activity representation, and do not use real video data; there is no mechanism to account for missing or noisy interpretations of video.

6.2.2 Activity Representation Methods based on Templates

The notion that a motion produces a characteristic spatio-temporal pattern dates at least to Polana and Nelson [253, 254, 256, 255, 260]. Spatio-temporal patterns are used to recognize actions in work by Bobick and Davis [38] and Davis and Bobick [82]. Ben-Arie et al [30, 29] recognize actions by first finding and tracking body parts using a form of template matcher and voting on lifted tracks; the tracks are lifted to 3D and a spatio-temporal representation of each body segment votes separately for an action. The action with the most votes is chosen. The method is successful, and has the advantage that it is robust to composition — if all but the left arm is walking, the action will still be recognized. However, the vocabulary consists of eight items (jump, kneel, pick, put, run, sit, stand, walk) and the vocabulary cannot be composed. An alternative is to match gestural information directly, incorporating a timewarp to improve the match. Bobick and Wilson [39, 40] use a state-based method that encodes gestures as a string of vector-quantized observation segments; this preserves order, but drops dynamical information. The advantage is relatively fast training.

6.3 Activity Representation using Hidden Markov Models and Finite State Representations

6.4 The Speech Analogy

Hidden Markov models (HMM's) pervade studies of motion, gesture and activity, and a complete review of their applications here may now be impossible. HMM's are models of sequences, and at their heart is a clock. One has a set of hidden states; at each tick of the clock, a Markov process chooses a new state, dependent on the previous state and nothing else; and an emission process produces an observation from the new state. There are clean solutions for the standard problems of learning (determining an appropriate state transition model and emission model for a given state model) and inference (determine which hidden states occurred given a set of observed states). HMM's have been used for understanding human behaviour but typically with quite small state models.

Very large state models are common in speech recognition, where HMM's have been hugely influential. This area is a useful source of inspiration by analogy. Viewed from a great height, a typical speech system has a series of components: a language model showing how words are built up into sentences; a pronunciation dictionary, giving sequences of context independent phones that correspond to words; a context dependency model, showing how local influences produce context dependent phones (cphones hereafter) from context independent phones; an acoustic observation model showing how acoustic observations result from context dependent phones (this is an

extremely compact description of a highly sophisticated area; more extensive descriptions appear in [155, 265]). The resulting object is a vast HMM — in our example, states can be thought of as being tagged with word-cphone-phone-sample — to explain each sample.

This HMM has some important, attractive features. Learning and authoring can be broken into tractable subproblems — the language model might be learned with one kind of dataset, the pronunciation dictionary with another — and as a result, we obtain an HMM on a massive scale, but with little difficulty in authoring it. While the state space is so big that dynamic programming must be sacrificed for a beam search, the state transition model is not impossible to learn, because most state transitions don't occur. Furthermore, the model is forced to share parameters in important ways — a phoneme in one word has the same model as that phoneme in a different word.

6.4.1 Finite State Transducers

Finite state models have had considerable success in the speech and language community. We introduce some terminology here, from the reviews by Mohri and others [218, 217, 216]. A *finite state automaton* is a directed graph, whose nodes are known as states. There is at least one final state and one initial state; each edge is labelled with an element of an alphabet. The automaton accepts any string corresponding to a path from an initial state to a final state. In a *finite state transducer*, transitions are labelled with both an element of an input alphabet and an element of an output alphabet; any string accepted by the transducer results in a string of output symbols, and so the transducer can be seen as representing a relation between families of strings. Transducers (representing relations between strings) can be composed, and there are efficient algorithms for computing the composition of two transducers.

In a *string-to-weight transducer*, the output alphabet consists of weights (typically, in a semiring or better; non-negative reals with addition and min is common, because it corresponds to the case of Viterbi and negative log-probabilities); there are initial and final weights. If a string-to-weight transducer accepts some string, its output for that string is defined as the minimum sum of weights over the paths accepting the string. Particularly attractive are subsequential string-to-weight transducers, where there is only one path accepting any given string. Not all transducers can be transformed to this form; there are algorithms for this process, known as *determinization* when it is possible. Furthermore, there are *minimization* algorithms, that can produce the unique (up to automorphism) smallest transducer that implements the same set of mappings as a given transducer.

6.4.2 Why should we Care?

Each of the components of a speech architecture (language model; a pronunciation dictionary; context dependency model; acoustic observation model) is a string-to-weight transducer. In principle, one could compose the lot to produce a single, enormous string-to-weight transducer, determinize it, minimize the result, and search that (this is equivalent to recognizing that, in the final analysis, the composition of each component produces an HMM with an enormous state space). In practice, the object involved is far too large. Instead, one uses a *beam search* to produce a reduced string-to-weight transducer (*the word lattice*) that contains a reduced pool of higher probability paths. Determinizing and minimizing this transducer is practical and useful; the result is very much faster searches.

There are two reasons that this material is of interest to us. First, the trick of reducing a speech signal to a (determinized and minimized) word lattice produces a highly compact representation of a large number of different transcriptions (each corresponding to a path through the string-to-weight transducer) that is easy to search and manage. We argue below that we can produce act, action and activity models which will allow reduction of video to an action/activity lattice with the same attractive properties. Second, a finite state automaton (whose states represent actions and activities) is a reasonable representation for a behaviour. If one determinizes and minimizes this, standard algorithms allow one to identify weights associated with instances of such a transducer in a word lattice extremely fast. This means we could be able to engage in fast searches for behaviours.

LeCun et al identify other useful building blocks associated with finite state models [176]. Their *graph transformers* take (weighted directed) graphs as inputs and produce graphs as outputs; an example of a transformer would be composition with a fixed transducer. Particularly useful is the idea of a Viterbi transformer, a process

that (using our terminology) takes a string-to-weight transducer and applies a beam search to produce a reduced string-to-weight transducer which is effectively a word lattice. They demonstrate that gradient based learning can usefully be applied to architectures of such objects.

6.5 Activity Recognition Methods based around HMM's

HMM's have been very widely adopted in activity recognition, but the models used have tended to be small (for example, one sees three and five state models in [46, 48]). Yamato *et al.* describe recognizing tennis strokes with HMM's [369]. Wilson and Bobick describe the use of HMM's for recognizing gestures such as pushes [359]. Yang et al use HMM's to recognize handwriting gestures [372]. Feng and Perona [98] call actions "movelets", and build a vocabulary by vector quantizing a representation of image shape, as a collection of rectangle, varying over time. These codewords are then strung together by an HMM, representing activities; there is one HMM per activity. We can then identify a new video by computing the image representation for each frame, obtaining the movelets, and choosing the particular model that generated the keyword sequence by a form of maximum likelihood. The method is not view invariant, depending on an image centered representation.

There has been a great deal of interest in models obtained by modifying the HMM structure. The intention is to improve the expressive power of the model without complicating the processes of learning or inference. Brand et al use coupled HMM's (CHMM's), which involve some number of simultaneous HMM's operating to the same clock, where the choice of a particular model's hidden state is affected by all other model's states [46, 48]. Such an object is clearly itself an HMM, but authors demonstrate a training method that reduces the number of parameters to learn by coupling but with very much enlarged state space; however, instead of estimating the parameters of that object, one projects the parameter estimates to transition parameters for each separate model. This means that one learns parameters for each separate model that tend to couple the two models. They show these models can distinguish between a set of T'ai Chi moves.

Oliver et al [232, 231] represent behaviours using layered hidden Markov models (LHMM's). These models involve a bank of HMM's at the lowest level, each generating some portion of the observation. The observations at higher levels are the maximum likelihood hidden state sequences for the lower levels. One then obtains for each HMM the maximum likelihood hidden state sequence. At the next level, the observations are these states, and this continues recursively. The resulting object is an HMM, but of complex structure; the LHMM form offers authoring advantages. This representation outperforms a straightforward HMM in recognizing such activities as phone conversation from both vision and acoustic data.Similarly, Mori et al build a hierarchical representation out of HMM's to recognize everyday gesture [222].

Wilson and Bobick [360] use a form of HMM where an unknown, global parameter applies to all emission models (which they call a parametric hidden Markov model or PHMM) to model gestures with a parametric form (such as might accompany "it was *this* big"). Data is from stereo or a Polhemus. There are recognition results for classes of gesture such as pointing. Kettnaker and Brand [165](also, Brand and Kettnaker, [47]) fit an HMM while penalizing model entropy; this tends to reduce the number of non-zero parameters, so that one can fit models with quite large state spaces satisfactorily (such models are sometimes known as Entropic HMM's or EHMM's). Galata *et al.* use variable length Markov models (VLMM's: a model that generates a state stochastically based on a variable but bounded length history) to encode behaviour and obtain a reduction in perplexity by doing so [107, 108].

Building variant HMM's is a way to simplify learning the state transition process from data (if the state space is large, the number of parameters is a problem). But there is an alternative — one could author the state transition process in such a way that it has relatively few free parameters, despite a very large state space, and then learn those parameters.

Finite state methods have been used directly. Hongeng *et al.* demonstrate recognition of multiperson activities from video of people at coarse scales (few kinematic details are available); activities include conversing and blocking [139]. Zhao and Nevatia use a finite-state model of walking, running and standing, built from motion capture [377]. Hong *et al.* use finite state machines to model gesture [138]. We are not aware of material that attempts to build large hierarchical finite state machines, patterned after speech recognition programs, and using opportunistic learning, as we propose to do.

6.6 Sign Language Recognition

The best-known system for sign matching is due to Starner and Pentland [324, 325]. Features are image moments of the hand region; signers either wear coloured gloves, or hands are identified using a skin filter. A Hidden Markov Model (HMM) is used to model individual signs; signs are strung together with a rigid language model (`pronoun verbnoun adjective pronoun`). Authors report a recognition rate of 90% with a vocabulary of 40 signs. Grobel and Assan recognize isolated signs under similar conditions for a 262-word vocabulary using HMM's [167]. This work was extended to recognize continuous German sign language with a vocabulary of 97 signs by Bauer and Hienz [27]. Vogler and Metaxas have built a system that uses estimates of arm position, recovered either from a physical sensor mounted on the body or from a system of three cameras that measures arm position fairly accurately [351, 352, 355]. For a vocabulary of 53 words, and an independent word language model, they report a word recognition accuracy of the order of 90%. A more recent system attempted to recognize phonemes with HMM's; Vogler and Metaxas were able to recognize signs from a 22 word vocabulary with similar recognition rates for phoneme and word models (without handshapes in [353], with handshapes in [354]).

Kadous transduced isolated Australian sign language signs with a powerglove, reporting a recognition rate of 80% using decision trees [226]. Matsuo *et al* transduced Japanese sign language with stereo cameras, using decision tree methods to recognize a vocabulary of 38 signs [206]. Kim *et al*. transduce Korean sign language using datagloves, reporting 94% accuracy in recognition for 131 Korean signs [168]. Al-Jarrah and Halawani report high recognition accuracy for 30 Arabic manual alphabet signs recognized from monocular views of a signer using a fuzzy inference system [9]. Gao *et al*. describe recognizing isolated signs drawn from a vocabulary of 5177 using datagloves and an HMM model [110, 357]. Their system is not speaker-independent: they describe relatively high accuracy for the original signer, and a significant reduction in performance for other signers. Similarly, Zieren and Kraiss report high, but not speaker independent, accuracy for monocular recognition of German sign language drawn from a vocabulary of 152 signs [379]. Akyol and Canzler describe an information terminal which can recognize 16 signs with a high, user-independent, recognition rate; their system uses HMM's to infer signs from monocular views of users wearing coloured gloves [8]. Bowden *et al*. use independent component analysis to obtain state estimates from a set of discriminative visual features; each sign is encoded as a Markov chain, learned from a single example [44]. They report high accuracy recognition from a lexicon of 49 signs using a very small training set.

6.7 More recent material

Low resolution activity recognition appears in [90]. Motion cues for computer games are in [106]. The EyeToy is one of computer vision's greatest commercial successes, and much underappreciated by the vision community; I had a long chat with its core visionary at CVPR 05. He had lots of time, because few people were talking to him, largely because few people knew what the EyeToy was or appreciated its significance. Work on location and activity in football appears in [146]. Correlation matching of activities is in [293]. Encoding complex actions in space time appears in [36]. Spotting irregular actions using this form of encoding appears in [42].

References

[1] 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*, page 2, New York, NY, USA, 2004. ACM Press.

[2] A. Agarwal and B. Triggs. Tracking articulated motion using a mixture of autoregressive models. In *European Conference on Computer Vision*, pages Vol III: 54–65, 2004.

[3] 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.

[4] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE T. Pattern Analysis and Machine Intelligence*, 28(1):44–58, 2006.

[5] J. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, March 1999.

[6] G. Agin and T. Binford. Computer description of curved objects. In *Int. Joint Conf. Artificial Intelligence*, pages 629–640, 1973.

[7] G. Agin and T. Binford. Computer description of curved objects. *IEEE Trans. Computer*, 25(4):439–449, April 1976.

[8] S. Akyol and U. Canzler. An information terminal using vision based sign language recognition. In *ITEA Workshop on Virtual Home Environments, VHE Middleware Consortium*, pages 61–68, 2002.

[9] O. Al-Jarrah and A. Halawani. Recognition of gestures in arabic sign language using neuro-fuzzy systems. *Artificial Intelligence*, 133(1-2):117–138, December 2001.

[10] J. F. Allen. Towards a general theory of action and time. *Artif. Intell.*, 23(2):123–154, 1984.

[11] F. Anderson and M. Pandy. A dynamic optimization solution for vertical jumping in three dimensions. *Computer Methods in Biomechanics and Biomedical Engineering*, 2:201–231, 1999.

[12] W. Arentz and B. Olstad. Classifying offensive sites based on image content. *Computer Vision and Image Understanding*, 94(1-3):295–310, April 2004.

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

[14] O. Arikan, D. Forsyth, and J. O'Brien. Motion synthesis from annotations. In *Proceedings of SIGGRAPH 95*, 2003.

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

[16] 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*, pages 59–66, New York, NY, USA, 2005. ACM Press.

[17] 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*, pages 40–45, 2002.

[18] V. Athitsos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 432–439, 2003.

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

[20] P. Baerlocher and R. Boulic. An inverse kinematics architecture enforcing an arbitrary number of strict priority levels. *The Visual Computer*, 20(6):402–417, 2004.

[21] H. Baker. Building surfaces of evolution: The weaving wall. *Int. J. Computer Vision*, 3(1. May 1989):51–72, May 1989.

[22] Y. Bar-Shalom and X.-R. Li. *Estimation with Applications to Tracking and Navigation*. John Wiley & Sons, Inc., New York, NY, USA, 2001.

[23] 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*, pages 185–194, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2004. Canadian Human-Computer Communications Society.

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

[25] C. Barron and I. Kakadiaris. Estimating anthropometry and pose from a single image. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 669–676, 2000.

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

[27] B. Bauer and H. Hienz. Relevant features for video-based continuous sign language recognition. In *IEEE Workshop on Automatic Face and Gesture Recognition*, pages 440–445, 2000.

[28] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE T. Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002.

[29] J. Ben-Arie, P. Pandit, and S. Rajaram. View-based human activity recognition by indexing and sequencing. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II:78–83, 2001.

[30] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram. Human activity recognition using multidimensional indexing. *IEEE T. Pattern Analysis and Machine Intelligence*, 24(8):1091–1104, August 2002.

[31] V. Beneš. Exact finite-dimensional filters with certain diffusion non linear drift. *Stochastics*, 5:65–92, 1981.

[32] A. Berger, S. D. Pietra, and V. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1), 1996.

[33] T. Binford. Inferring surfaces from images. *Artificial Intelligence*, 17(1-3):205–244, August 1981.

[34] M. Black, Y. Yacoob, A. Jepson, and D. Fleet. Learning parameterized models of image motion. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 561–567, 1997.

[35] S. Blackman and R. Popoli. *Design and Analysis of Modern Tracking Systems*. Artech House, 1999.

[36] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *IEEE International Conference on Computer Vision (ICCV)*, 2005.

[37] A. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. *Proc. Roy. Soc. B*, 352:1257–1265, 1997.

[38] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE T. Pattern Analysis and Machine Intelligence*, 23(3):257–267, March 2001.

[39] A. Bobick and A. Wilson. A state based technique for the summarization and recognition of gesture. In *Int. Conf. on Computer Vision*, pages 382–388, 1995.

[40] A. Bobick and A. Wilson. A state based approach to the representation and recognition of gesture. *IEEE T. Pattern Analysis and Machine Intelligence*, 19(12):1325–1337, December 1997.

[41] 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.

[42] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *IEEE International Conference on Computer Vision (ICCV)*, 2005.

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

[44] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. A linguistic feature vector for the visual interpretation of sign language. In *European Conference on Computer Vision*, pages Vol I: 390–401, 2004.

[45] J. Boyd and J. Little. Phase in model-free perception of gait. In *IEEE Workshop on Human Motion*, pages 3–10, 2000.

[46] M. Brand. Coupled hidden markov models for complex action recognition. Media lab vision and modelling tr-407, MIT, 1997.

[47] M. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *IEEE T. Pattern Analysis and Machine Intelligence*, 22(8):844–851, August 2000.

[48] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 994–999, 1997.

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

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

[51] A. Broggi, M. Bertozzi, A. Fascioli, and M. Sechi. Shape-based pedestrian detection. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 215–220, 2000.

[52] L. Brown. *A Radar History of World War II: Technical and Military Imperatives*. Institute of Physics Press, 2000.

[53] A. Bruderlin and L. Williams. Motion signal processing. In *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 97–104, New York, NY, USA, 1995. ACM Press.

[54] R. Buderi. *The Invention that Changed the World*. Touchstone Press, 1998. reprint.

[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*, page 356, Washington, DC, USA, 1998. IEEE Computer Society.

[56] L. W. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. In *ICCV*, pages 624–630, 1995.

[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] S. Carlsson. Recognizing walking people. In *European Conference on Computer Vision*, pages I: 472–486, 2000.

[59] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3):569–577, 2003.

[60] A. Cavallaro and T. Ebrahimi. Video object extraction based on adaptive background and statistical change detection. In *Proc. SPIE 4310*, pages 465–475, 2000.

[61] J. Chai and J. K. Hodgins. Performance animation from low-dimensional control signals. *ACM Trans. Graph.*, 24(3):686–696, 2005.

[62] T. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 239–245, 1999.

[63] F. Cheng, W. Christmas, and J. Kittler. Periodic human motion description for sports video databases. In *Proceedings IAPR International Conference on Pattern Recognition*, pages III: 870–873, 2004.

[64] 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)*, volume 2, pages 714 – 720, June 2000.

[65] K.-M. G. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette across time part i: Theory and algorithms. *Int. J. Comput. Vision*, 62(3):221–247, 2005.

[66] 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*, 63(3):225–245, 2005.

[67] K. Choo and D. Fleet. People tracking using hybrid monte carlo filtering. In *Int. Conf. on Computer Vision*, pages II: 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*, pages 293–302, New York, NY, USA, 1992. ACM Press.

[69] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–97, 1995.

[70] L. Crawford and S. Sastry. Biological motor control approaches for a planar diver. In *IEEE Conf. on Decision and Control*, pages 3881–3886, 1995.

[71] C. Curio, J. Edelbrunner, T. Kalinke, C. Tzomakas, and W. von Seelen. Walking pedestrian recognition. *Intelligent Transportation Systems*, 1(3):155–163, September 2000.

[72] R. Cutler and L. Davis. View-based detection and analysis of periodic motion. In *Proceedings IAPR International Conference on Pattern Recognition*, pages Vol I: 495–500, 1998.

[73] R. Cutler and L. Davis. Real-time periodic motion detection, analysis, and applications. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 326–332, 1999.

[74] R. Cutler and L. Davis. Robust periodic motion and motion symmetry detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 615–622, 2000.

[75] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE T. Pattern Analysis and Machine Intelligence*, 22(8):781–796, August 2000.

[76] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 886–893, 2005.

[77] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. *Int. J. Computer Vision*, 37(2):175–185, June 2000.

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

[79] 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*, pages 1044–1049, 1999.

[80] F. Daum. Beyond kalman filters: practical design of nonlinear filters. In *Proc. SPIE*, volume 2561, pages 252–262, 1995.

[81] F. Daum. Exact finite dimensional nonlinear filters. *IEEE. Trans. Automatic Control*, 31:616–622, 1995.

[82] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 928–934, 1997.

[83] 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*, page 716, Washington, DC, USA, 1999. IEEE Computer Society.

[84] Q. Delamarre and O. Faugeras. 3d articulated models and multiview tracking with physical forces. *Comput. Vis. Image Underst.*, 81(3):328–357, 2001.

[85] 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.

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

[87] T. Drummond and R. Cipolla. Real-time tracking of complex structures with on-line camera calibration. In T. P. Pridmore and D. Elliman, editors, *Proceedings of the British Machine Vision Conference 1999, BMVC 1999, Nottingham, 13-16 September 1999*, 1999.

[88] 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*, pages 20–36, London, UK, 2000. Springer-Verlag.

[89] T. Drummond and R. Cipolla. Real-time visual tracking of complex structures. *IEEE T. Pattern Analysis and Machine Intelligence*, 24(7):932–946, July 2002.

[90] 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*, pages 726–733, Washington, DC, USA, 2003. IEEE Computer Society.

[91] P. Faloutsos, M. van de Panne, and D. Terzopoulos. Composable controllers for physics-based character animation. In *Proceedings of ACM SIGGRAPH 2001*, Computer Graphics Proceedings, Annual Conference Series, pages 251–260, Aug. 2001.

[92] P. Faloutsos, M. van de Panne, and D. Terzopoulos. The virtual stuntman: dynamic characters with a repertoire of autonomous motor skills. *Computers & Graphics*, 25(6):933–953, Dec. 2001.

[93] A. C. Fang and N. S. Pollard. Efficient synthesis of physically valid human motion. *ACM Transactions on Graphics*, 22(3):417–426, July 2003.

[94] A. C. Fang and N. S. Pollard. Efficient synthesis of physically valid human motion. *ACM Trans. Graph.*, 22(3):417–426, 2003.

[95] A. Farina, D. Benvenuti, and B. Ristic. A comparative study of the benes filtering problem. *Signal Processing*, 82:133–147, 2002.

[96] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 66–73, 2000.

[97] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int. J. Computer Vision*, 61(1):55–79, January 2005.

[98] 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*, pages 717–721, 2002.

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

[100] D. Forsyth. Sampling, resampling and colour constancy. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 300–305, 1999.

[101] D. Forsyth and M. Fleck. Body plans. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 678–683, 1997.

[102] D. Forsyth and M. Fleck. Automatic detection of human nudes. *Int. J. Computer Vision*, 32(1):63–77, August 1999.

[103] D. Forsyth, M. Fleck, and C. Bregler. Finding naked people. In *European Conference on Computer Vision*, pages 593–602, 1996.

[104] D. Forsyth, O.Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies in human motion 1: Tracking and animation. *Foundations and Trends in Computer Vision*, 2006. In press.

[105] D. Forsyth and J. Ponce. *Computer Vision: a modern approach*. Prentice-Hall, 2002.

[106] W. Freeman, D. Anderson, and P. B. et al. Computer vision for interactive computer graphics. *Computer Graphics and Applications*, pages 42–53, 1998.

[107] A. Galata, N. Johnson, and D. Hogg. Learning behaviour models of human activities. In *British Machine Vision Conference*, page Modelling Human Behaviour, 1999.

[108] A. Galata, N. Johnson, and D. Hogg. Learning structured behavior models using variable length markov models. In *IEEE Workshop on Modelling People*, 1999.

[109] D. Gao, J. Zhou, and L. Xin. A novel algorithm of adaptive background estimation. In *IEEE Int. Conf. Image Processing*, pages II: 395–398, 2001.

[110] W. Gao, J. Ma, X. Chen, et al. Handtalker: a multimodal dialog system using sign language and 3d virtual human. In *Proc. Third Int. Conf. Multimodal Interface*, pages 564–571, 2000.

[111] D. Gavrila. Pedestrian detection from a moving vehicle. In *European Conference on Computer Vision*, pages II: 37–49, 2000.

[112] D. Gavrila. Sensor-based pedestrian protection. *Intelligent Transportation Systems*, pages 77–81, 2001.

[113] D. Gavrila, J. Giebel, and S. Munder. Vision-based pedestrian detection: the protector system. In *Intelligent Vehicle Symposium*, pages 13–18, 2004.

[114] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.

[115] A. Gelb. *Applied Optimal Estimation*. MIT Press, 1974. written together with Staff of the Analytical Sciences Corporation.

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

[117] 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*, pages 263–270, New York, NY, USA, 1985. ACM Press.

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

[119] M. Gleicher. Animation from observation: Motion capture and motion editing. *SIGGRAPH Comput. Graph.*, 33(4):51–54, 2000.

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

[121] 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*, page I: 461 ff., 2002.

[122] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky. Behavior classification by eigendecomposition of periodic motions. *Pattern Recognition*, 38(7):1033–1043, July 2005.

[123] K. Grauman, G. Shakhnarovich, and T. Darrell. Virtual visual hulls: Example-based 3d shape inference from silhouettes. In *SMVP04*, pages 26–37, 2004.

[124] W. 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*, pages 22–29, 1998.

[125] K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popović. Style-based inverse kinematics. *ACM Trans. Graph.*, 23(3):522–531, 2004.

[126] R. Grzeszczuk, D. Terzopoulos, and G. Hinton. Neuroanimator: Fast neural network emulation and control of physics-based models. In *Proceedings of SIGGRAPH 98*, Computer Graphics Proceedings, Annual Conference Series, pages 9–20, July 1998.

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

[128] I. Haritaoglu, D. Harwood, and L. Davis. W4s: A real-time system for detecting and tracking people in 2 1/2-d. In *European Conference on Computer Vision*, page I: 877, 1998.

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

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

[131] 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*, pages 357–364, 2004.

[132] G. Hinton. Relaxation and its role in vision. Technical report, University of Edinburgh, 1978. PhD Thesis.

[133] J. K. Hodgins, J. F. O'Brien, and J. Tumblin. Do geometric models affect judgments of human motion? In *Graphics Interface '97*, pages 17–25, May 1997.

[134] 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*, 4(4):307–316, Oct. 1998.

[135] 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*, pages 153–162, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

[136] J. K. Hodgins, W. L. Wooten, D. C. Brogan, and J. F. O'Brien. Animating human athletics. In *Proceedings of SIGGRAPH 95*, Computer Graphics Proceedings, Annual Conference Series, pages 71–78, Aug. 1995.

[137] D. Hogg. Model-based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[138] P. Hong, M. Turk, and T. Huang. Gesture modeling and recognition using finite state machines. In *Int. Conf. Automatic Face and Gesture Recognition*, pages 410–415, 2000.

[139] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding*, 96(2):129–162, November 2004.

[140] N. Howe. Silhouette lookup for automatic pose tracking. In *IEEE Workshop on Articulated and Non-Rigid Motion*, page 15, 2004.

[141] N. R. Howe, M. E. Leventon, and W. T. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. In S. Solla, T. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 820–26. MIT Press, 2000.

[142] 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*, 34(3), 2004.

[143] L. Ikemoto, O. Arikan, and D. Forsyth. Knowing when to put your foot down. In *Proc Symp. Interactive 3D graphics and Games*, 2006.

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

[145] L. Ikemoto and D. Forsyth. Enriching a motion collection by transplanting limbs. In *Proc. Symposium on Computer Animation*, 2004.

[146] S. S. Intille and A. F. Bobick. Closed-world tracking. In *ICCV*, pages 672–678, 1995.

[147] S. Ioffe and D. Forsyth. Learning to find pictures of people. In *Proc. Neural Information Processing Systems*, 1998.

[148] S. Ioffe and D. Forsyth. Human tracking with mixtures of trees. In *Int. Conf. on Computer Vision*, pages I: 690–695, 2001.

[149] S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *Int. J. Computer Vision*, 43(1):45–68, June 2001.

[150] M. Isard and A. Blake. C-conditional density propagation for visual tracking. *IJCV*, 29(1):5–28, August 1998.

[151] M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *European Conference on Computer Vision*, page I: 893, 1998.

[152] Y. Ivanov, A. Bobick, and J. Liu. Fast lighting independent background subtraction. In *In Proc. of the IEEE Workshop on Visual Surveillance – VS'98*, pages 49–55, 1998.

[153] Y. Ivanov, A. Bobick, and J. Liu. Fast lighting independent background subtraction. *Int. J. Computer Vision*, 37(2):199–207, June 2000.

[154] O. Javed and M. Shah. Tracking and object classification for automated surveillance. In *European Conference on Computer Vision*, page IV: 343 ff., 2002.

[155] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1998.

[156] 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*, pages 225–232, New York, NY, USA, 2003. ACM Press.

[157] C. Jeong, J. Kim, and K. Hong. Appearance-based nude image detection. In *Proceedings IAPR International Conference on Pattern Recognition*, pages IV: 467–470, 2004.

[158] 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.

[159] R. V. Jones. *Most Secret War*. Wordsworth Military Library, 1998. reprint.

[160] 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*, pages 561–567, 1996.

[161] Y. Ke and R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 506–513, 2004.

[162] 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*, pages 129–136, Washington, DC, USA, 2005. IEEE Computer Society.

[163] E. J. Keogh. Exact indexing of dynamic time warping. In *VLDB*, pages 406–417, 2002.

[164] 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*, pages 253–265, 2003.

[165] V. Kettnaker and M. Brand. Minimum-entropy models of scene activity. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I:281–286, 1999.

[166] 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*, page 267, Washington, DC, USA, 1999. IEEE Computer Society.

[167] K.Grobel and M.Assan. Isolated sign language recognition using hidden markov models. In *Proc. Int. Conf. System Man and Cybernetics*, pages 162–167, 1997.

[168] J. Kim, W. Jang, and Z. Bien. A dynamic gesture recognition system for the korean sign language (ksl). *Systems, Man and Cybernetics-B*, 26(2):354–359, April 1996.

[169] H. Ko and N. Badler. Animating human locomotion with inverse dynamics. *IEEE Computer Graphics and Application*, 16(2):50–59, 1996.

[170] 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*, pages 214–224, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

[171] L. Kovar and M. Gleicher. Automated extraction and parameterization of motions in large data sets. *ACM Trans. Graph.*, 23(3):559–568, 2004.

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

[173] 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*, pages 97–104. ACM Press, 2002.

[174] 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.

[175] 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*, pages 29–38, New York, NY, USA, 2005. ACM Press.

[176] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[177] 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.

[178] 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*, pages 39–48. ACM Press/Addison-Wesley Publishing Co., 1999.

[179] M. Lee and I. Cohen. Human upper body pose estimation in static images. In *European Conference on Computer Vision*, pages Vol II: 126–138, 2004.

[180] M. 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*, pages II: 334–341, 2004.

[181] M. Lee and R. Nevatia. Dynamic human pose estimation using markov chain monte carlo approach. In *IEEE Workshop on Motion and Video Computing*, pages 168–175, 2005.

[182] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 878–885, 2005.

[183] B. Li and H. Holstein. Recognition of human periodic motion: A frequency domain approach. In *Proceedings IAPR International Conference on Pattern Recognition*, pages I: 311–314, 2002.

[184] 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*, pages 465–472. ACM Press, 2002.

[185] J. Little and J. Boyd. Describing motion for recognition. In *International Symposium on Computer Vision*, pages 235–240, 1995.

[186] J. Little and J. Boyd. Recognizing people by their gait: The shape of motion. *Videre*, 1(2), 1998.

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

[188] C. Liu, S. Zhu, and H. Shum. Learning inhomogeneous gibbs model of faces by minimax entropy. In *Int. Conf. on Computer Vision*, pages I: 281–287, 2001.

[189] C. K. Liu, A. Hertzmann, and Z. Popović. Learning physics-based motion style with nonlinear inverse optimization. *ACM Trans. Graph.*, 24(3):1071–1081, 2005.

[190] 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*, pages 408–416, New York, NY, USA, 2002. ACM Press.

[191] F. Liu and R. Picard. Detecting and segmenting periodic motion. Media lab vision and modelling tr-400, MIT, 1996.

[192] F. Liu and R. Picard. Finding periodicity in space and time. In *Int. Conf. on Computer Vision*, pages 376–383, 1998.

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

[194] Z. Liu, H. Chen, and H. Shum. An efficient approach to learning inhomogeneous gibbs model. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 425–431, 2003.

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

[196] 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*, pages 35–42, New York, NY, USA, 1994. ACM Press.

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

[198] 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*, pages Vol IV: 442–455, 2004.

[199] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Int. Conf. on Computer Vision*, pages 572–578, 1999.

[200] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. *Int. J. Computer Vision*, 39(1):57–71, August 2000.

[201] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *European Conference on Computer Vision*, pages II: 3–19, 2000.

[202] A. A. Maciejewski. Motion simulation: Dealing with the ill-conditioned equations of motion for articulated figures. *IEEE Comput. Graph. Appl.*, 10(3):63–71, 1990.

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

[204] 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*, pages 317–324, New York, NY, USA, 1998. ACM Press.

[205] M. J. Matarić, V. B. Zordan, and M. M. Williamson. Making complex articulated agents dance. *Autonomous Agents and Multi-Agent Systems*, 2(1):23–43, 1999.

[206] H. Matsuo, S. Igi, S. Lu, Y. Nagashima, Y. Takata, and T. Teshima. The recognition algorithm with non-contact for japanese sign language using morphological analysis. In *Proc. Int. Gesture Workshop*, pages 273–284, 1997.

[207] 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*, pages 29–38, New York, NY, USA, 1990. ACM Press.

[208] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, October 2000.

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

[210] A. Micilotta, E. Ong, and R. Bowden. Detection and tracking of humans by probabilistic body part assembly. In *British Machine Vision Conference*, volume 1, pages 429–438, 2005.

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

[212] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, pages Vol I: 69–82, 2004.

[213] A. Mittal and L. S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. *Int. J. Comput. Vision*, 51(3):189–203, 2003.

[214] T. Moeslund. Summaries of 107 computer vision-based human motion capture papers. Technical Report LLA 99-01, University of Aalborg, 1999.

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

[216] M. Mohri. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2), 1997.

[217] M. Mohri, F. C. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16(1):69–88, 2002.

[218] M. Mohri and M. Riley. Weighted finite-state transducers in speech recognition (tutorial). In *Proceedings of the International Conference on Spoken Language Processing 2002 (ICSLP '02)*, 2002.

[219] G. Monheit and N. I. Badler. A kinematic model of the human spine and torso. *IEEE Comput. Graph. Appl.*, 11(2):29–38, 1991.

[220] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *European Conference on Computer Vision LNCS 2352*, volume 3, pages 666–680, 2002.

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

[222] T. Mori, Y. Segawa, M. Shimosaka, and T. Sato. Hierarchical recognition of daily human actions based on continuous hidden markov models. In *Int. Conf. Automatic Face and Gesture Recognition*, pages 779–784, 2004.

[223] M. Müller, T. Röder, and M. Clausen. Efficient content-based retrieval of motion capture data. *ACM Trans. Graph.*, 24(3):677–685, 2005.

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

[225] 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*, pages 467–475, 1999.

[226] M.W.Kadous. Machine recognition of auslan signs using powergloves: towards large lexicon integration of sign language. In *Proc. Workshop on the Integration of Gesture in Language and Speech*, pages 165–174, 1996.

[227] J. T. Ngo and J. Marks. Physically realistic motion synthesis in animation. *Evol. Comput.*, 1(3):235–268, 1993.

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

[229] S. Niyogi and E. Adelson. Analyzing gait with spatiotemporal surfaces. In *Proc. IEEE Workshop on Nonrigid and Articulated Motion*, pages 64– 69, 1994.

[230] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in xyt. Media lab vision and modelling tr-223, MIT, 1995.

[231] N. Oliver, A. Garg, and E. Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96(2):163–180, November 2004.

[232] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pages 3–8, 2002.

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

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

[235] J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE T. Pattern Analysis and Machine Intelligence*, 2(6):522–536, November 1980.

[236] C. Pai, H. Tyan, Y. Liang, H. Liao, and S. Chen. Pedestrian detection and tracking at crossroads. In *IEEE Int. Conf. Image Processing*, pages II: 101–104, 2003.

[237] C. Pai, H. Tyan, Y. Liang, H. Liao, and S. Chen. Pedestrian detection and tracking at crossroads. *Pattern Recognition*, 37(5):1025–1034, May 2004.

[238] M. Pandy and F. Anderson. Dynamic simulation of human movement using large-scale models of the body. In *Proc. IEEE Intl. Conference on Robotics and Automation*, pages 676–681, 2000.

[239] C. Papageorgiou. A trainable system for object detection in images and video sequences constantine. Technical report, MIT, 2000. Ph. D.

[240] C. Papageorgiou, T. Evgeniou, and T. Poggio. A trainable object detection system. In *DARPA IU Workshop*, pages 1019–1024, 1998.

[241] C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Int. Conf. on Computer Vision*, pages 555–562, 1998.

[242] C. Papageorgiou and T. Poggio. A pattern classification approach to dynamical object detection. In *Int. Conf. on Computer Vision*, pages 1223–1228, 1999.

[243] C. Papageorgiou and T. Poggio. Trainable pedestrian detection. In *IEEE Int. Conf. Image Processing*, pages IV:35–39, 1999.

[244] C. Papageorgiou and T. Poggio. A trainable system for object detection. *Int. J. Computer Vision*, 38(1):15–33, June 2000.

[245] 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*, 16(2):219–232, 2004.

[246] 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*, 15(3-4):125–138, 2004.

[247] 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*, pages 105–111, New York, NY, USA, 2002. ACM Press.

[248] 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*, pages 245–250, New York, NY, USA, 1990. ACM Press.

[249] S. D. Pietra, V. D. Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

[250] C. Pinhanez and A. Bobick. Pnf propagation and the detection of actions described by temporal intervals. In *DARPA IU Workshop*, pages 227–234, 1997.

[251] C. Pinhanez and A. Bobick. Human action detection using pnf propagation of temporal constraints. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 898–904, 1998.

[252] R. Plänkers and P. Fua. Tracking and modeling people in video sequences. *Comput. Vis. Image Underst.*, 81(3):285–302, 2001.

[253] R. Polana and R. Nelson. Detecting activities. In *DARPA IU Workshop*, pages 569–574, 1993.

[254] R. Polana and R. Nelson. Detecting activities. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2–7, 1993.

[255] R. Polana and R. Nelson. Detecting activities. *J. Visual Communication Image Representation*, 5:172–180, 1994.

[256] R. Polana and R. Nelson. Detecting activities. In *Proceedings IAPR International Conference on Pattern Recognition*, pages A:815–818, 1994.

[257] R. Polana and R. Nelson. Low level recognition of human motion. In *IEEE Workshop on Articulated and Non-Rigid Motion*, 1994.

[258] R. Polana and R. Nelson. Recognition of nonrigid motion. In *ARPA94*, pages II:1219–1224, 1994.

[259] R. Polana and R. Nelson. Recognizing activities. In *Proceedings IAPR International Conference on Pattern Recognition*, pages A:815–818, 1994.

[260] R. Polana and R. Nelson. Detection and recognition of periodic, nonrigid motion. *Int. J. Computer Vision*, 23(3):261–282, 1997.

[261] 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.

[262] J. Popović, S. M. Seitz, and M. Erdmann. Motion sketching for control of rigid-body simulations. *ACM Trans. Graph.*, 22(4):1034–1054, 2003.

[263] Z. Popović and A. Witkin. Physically based motion transformation. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 11–20, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

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

[265] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

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

[267] D. Ramanan and D. Forsyth. Automatic annotation of everyday movements. In *Proc. Neural Information Processing Systems*, 2003.

[268] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 467–474, 2003.

[269] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 271–278, 2005.

[270] P. Reitsma and N. Pollard. Evaluating motion graphs for character navigation. In *Eurographics/ACM Symposium on Computer Animation*, pages 89–98, 2004.

[271] L. Ren, G. Shakhnarovich, J. K. Hodgins, H. Pfister, and P. Viola. Learning silhouette features for control of human motion. *ACM Trans. Graph.*, 24(4):1303–1331, 2005.

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

[273] J. Rittscher and A. Blake. Classification of human body motion. In *Int. Conf. on Computer Vision*, pages 634–639, 1999.

[274] T. Roberts, S. McKenna, and I. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. In *European Conference on Computer Vision*, pages Vol IV: 291–303, 2004.

[275] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, 59(1):94–115, 1994.

[276] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. In *European Conference on Computer Vision*, page IV: 700 ff., 2002.

[277] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff. 3d hand pose reconstruction using specialized mappings. In *Int. Conf. on Computer Vision*, pages I: 378–385, 2001.

[278] C. Rose, M. F. Cohen, and B. Bodenheimer. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Comput. Graph. Appl.*, 18(5):32–40, 1998.

[279] 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*, pages 147–154, New York, NY, USA, 1996. ACM Press.

[280] R. Rosenfeld. A maximum entropy approach to adaptive statistical language modelling. *Computer, Speech and Language*, 10:187–228, 1996.

[281] S. Roth, L. Sigal, and M. Black. Gibbs likelihoods for bayesian tracking. In *CVPR04*, pages I: 886–893, 2004.

[282] A. Safonova, J. K. Hodgins, and N. S. Pollard. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Trans. Graph.*, 23(3):514–521, 2004.

[283] G. Schmidt. Designing nonlinear filters based on Daum's theory. *Journal of Guidance, Control and Dynamics*, 16:371–376, 1993.

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

[285] S. Seitz and C. Dyer. Affine invariant detection of periodic motion. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 970–975, 1994.

[286] S. Seitz and C. Dyer. View invariant analysis of cyclic motion. *Int. J. Computer Vision*, 25(3):231–251, December 1997.

[287] A. Senior. Tracking people with probabilistic appearance models. In *IEEE Workshop on Performance Evaluation Tracking Surveillance*, pages 48–55, 2002.

[288] A. Shahrokni, T. Drummond, and P. Fua. Fast Texture-Based Tracking and Delineation Using Texture Entropy. In *International Conference on Computer Vision*, 2005.

[289] 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.

[290] A. Shahrokni, F. Fleuret, and P. Fua. Classifier-based Contour Tracking for Rigid and Deformable Objects. In *British Machine Vision Conference*, Oxford, UK, 2005.

[291] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Int. Conf. on Computer Vision*, pages 750–757, 2003.

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

[293] E. Shechtman and M. Irani. Space-time behavior based correlation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[294] 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*, page 194, Washington, DC, USA, 2003. IEEE Computer Society.

[295] H. J. Shin, J. Lee, S. Y. Shin, and M. Gleicher. Computer puppetry: An importance-based approach. *ACM Trans. Graph.*, 20(2):67–94, 2001.

[296] H. Sidenbladh and M. Black. Learning image statistics for bayesian tracking. In *Int. Conf. on Computer Vision*, pages II: 709–716, 2001.

[297] H. Sidenbladh and M. Black. Learning the statistics of people in images and video. *Int. J. Computer Vision*, 54(1):181–207, September 2003.

[298] 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.

[299] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 421–428, 2004.

[300] 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*, pages 26–40, Nov. 1998. Proceedings of CAPTECH '98.

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

[302] J. M. Siskind. Grounding language in perception. *Artificial Intelligence Review*, 8:371–391, 1995.

[303] J. M. Siskind. Reconstructing force-dynamic models from video sequences. *Artificial Intelligence*, 151:91–154, 2003.

[304] 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.

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

[306] C. Sminchisescu and A. Jepson. Generative Modeling for Continuous Non-Linearly Embedded Visual Inference. In *International Conference on Machine Learning*, pages 759–766, Banff, 2004.

[307] C. Sminchisescu and A. Jepson. Variational Mixture Smoothing for Non-Linear Dynamical Systems. In *IEEE Computer Vision and Pattern Recognition*, volume 2, pages 608–615, Washington D.C., 2004.

[308] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional Visual Tracking in Kernel Space. In *Neural Information Processing Systems*, 2005.

[309] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *IEEE Computer Vision and Pattern Recognition*, volume 1, pages 390–397, 2005.

[310] C. Sminchisescu and A. Telea. Human pose estimation from silhouettes: A consistent approach using distance level sets. In *WSCG02*, page 413, 2002.

[311] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3d body tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I:447–454, 2001.

[312] C. Sminchisescu and B. Triggs. Building roadmaps of local minima of visual models. In *European Conference on Computer Vision*, page I: 566 ff., 2002.

[313] C. Sminchisescu and B. Triggs. Hyperdynamics importance sampling. In *European Conference on Computer Vision*, page I: 769 ff., 2002.

[314] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *The International Journal of Robotics Research*, 22(6):371–391, 2003.

[315] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 69–76, 2003.

[316] C. Sminchisescu and B. Triggs. Building roadmaps of minima and transitions in visual models. *Int. J. Computer Vision*, 61(1):81–101, January 2005.

[317] Y. Song, X. Feng, and P. Perona. Towards detection of human motion. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 810–817, 2000.

[318] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE T. Pattern Analysis and Machine Intelligence*, 25(7):814–827, July 2003.

[319] N. Sprague and J. Luo. Clothed people detection in still images. In *Proceedings IAPR International Conference on Pattern Recognition*, pages III: 585–589, 2002.

[320] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. In *Int. Conf. on Computer Vision*, pages 915–922, 2003.

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

[322] J. Starck and A. Hilton. Virtual view synthesis of people from multiple view video sequences. *Graphical Models*, 67(6):600–620, 2005.

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

[324] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. Technical report, MIT, 1996.

[325] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE T. Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[326] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 246–252, 1999.

[327] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 246–252, 1999.

[328] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE T. Pattern Analysis and Machine Intelligence*, 22(8):747–757, August 2000.

[329] A. Sulejmanpašić and J. Popović. Adaptation of performed ballistic motion. *ACM Trans. Graph.*, 24(1):165–179, 2005.

[330] J. Sullivan, A. Blake, and J. Rittscher. Statistical foreground modelling for object localisation. In *European Conference on Computer Vision*, pages II: 307–323, 2000.

[331] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *European Conference on Computer Vision*, page I: 629 ff., 2002.

[332] S. Tak and H. Ko. Example guided inverse kinematics. In *International Conference on Computer Graphics and Imaging*, pages 19–23, 2000.

[333] S. Tak and H.-S. Ko. A physically-based motion retargeting filter. *ACM Trans. Graph.*, 24(1):98–117, 2005.

[334] S. Tak, O. Song, and H. Ko. Motion balance filtering. *Computer Graphics Forum (Eurographics 2000)*, 19(3):437–446, 2000.

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

[336] C. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80(3):349–363, December 2000.

[337] A. Thangali and S. Sclaroff. Periodic motion detection and estimation via space-time sampling. In *Motion05*, pages II: 176–182, 2005.

[338] C. Theobalt, I. Albrecht, J. Haber, M. Magnor, and H.-P. Seidel. Pitching a baseball: tracking high-speed motion with multi-exposure images. *ACM Trans. Graph.*, 23(3):540–547, 2004.

[339] 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*, page 185, Washington, DC, USA, 2003. IEEE Computer Society.

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

[341] N. Torkos and M. V. de Panne. Footprint-based quadruped motion synthesis. In *Graphics Interface 98*, pages 151–160, 1998.

[342] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *Int. Conf. on Computer Vision*, pages II: 50–57, 2001.

[343] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric space. *Int. J. Computer Vision*, 48(1):9–19, June 2002.

[344] Z. Tu and S. Zhu. Image segmentation by data-driven markov chain monte carlo. In *Int. Conf. on Computer Vision*, pages II: 131–138, 2001.

[345] Z. Tu and S. Zhu. Image segmentation by data-driven markov chain monte carlo. *IEEE T. Pattern Analysis and Machine Intelligence*, 24(5):657–673, May 2002.

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

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

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

[349] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Int. Conf. on Computer Vision*, pages 734–741, 2003.

[350] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *Int. J. Computer Vision*, 63(2):153–161, July 2005.

[351] C. Vogler and D. Metaxas. ASL recognition based on a coupling between HMMs and 3D motion analysis. In *Int. Conf. on Computer Vision*, pages 363–369, 1998.

[352] C. Vogler and D. Metaxas. Parallel hidden markov models for American sign language recognition. In *Int. Conf. on Computer Vision*, pages 116–122, 1999.

[353] C. Vogler and D. Metaxas. Toward scalability in asl recognition: breaking down signs into phonemes. In *Gesture workshop 99*, 1999.

[354] C. Vogler and D. Metaxas. Handshapes and movements: Multiple-channel american sign language recognition. In *Proc. Gesture Workshop*, pages 247–258, 2003.

[355] C. Vogler, H. Sun, and D. Metaxas. A framework for motion recognition with applications to American sign language and gait recognition. In *IEEE Workshop on Human Motion*, 2000.

[356] Y. Weiss. Belief propagation and revision in networks with loops. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA, 1997.

[357] W.Gao, J. Ma, J. Wu, and C. Wang. Sign language recognition based on hmm/ann/dp. *Int. J. Pattern Recognition and Artificial Intelligence*, 14(5):587–602, 2000.

[358] D. J. Wiley and J. K. Hahn. Interpolation synthesis of articulated figure motion. *IEEE Comput. Graph. Appl.*, 17(6):39–45, 1997.

[359] A. Wilson and A. Bobick. Learning visual behavior for gesture analysis. In *IEEE Symposium on Computer Vision*, pages 229–234, 1995.

[360] A. Wilson and A. Bobick. Parametric hidden markov models for gesture recognition. *IEEE T. Pattern Analysis and Machine Intelligence*, 21(9):884–900, September 1999.

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

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

[363] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE T. Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.

[364] 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*, pages 605–612, 2003.

[365] Y. Wu, T. Yu, and G. Hua. A statistical field model for pedestrian detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 1023–1030, 2005.

[366] Y. Yacoob and L. Davis. Learned models for estimation of rigid and articulated human motion from stationary or moving camera. *Int. J. Computer Vision*, 36(1):5–30, January 2000.

[367] 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*, page 2, Washington, DC, USA, 1998. IEEE Computer Society.

[368] K. Yamane and Y. Nakamura. Natural motion animation through constraining and deconstraining at will. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):352–360, 2003.

[369] J. Yamato, J. Ohya, and K. Ishii. Recognising human action in time sequential images using hidden markov model. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 379–385, 1992.

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

[371] J. Yang, Z. Fu, T. Tan, and W. Hu. A novel approach to detecting adult images. In *Proceedings IAPR International Conference on Pattern Recognition*, pages IV: 479–482, 2004.

[372] J. Yang, Y. Xu, and C. S. Chen. Human action learning via hidden markov model. *IEEE Transactions on Systems Man and Cybernetics*, 27:34–44, 1997.

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

[374] J. Zhao and N. I. Badler. Inverse kinematics positioning using nonlinear programming for highly articulated figures. *ACM Trans. Graph.*, 13(4):313–336, 1994.

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

[376] L. Zhao and C. Thorpe. Stereo- and neural network-based pedestrian detection. *Intelligent Transportation Systems*, 1(3):148–154, September 2000.

[377] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE T. Pattern Analysis and Machine Intelligence*, 26(9):1208–1221, September 2004.

[378] S. 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*, pages I: 738–745, 2000.

[379] J. Zieren and K.-F. Kraiss. Non-intrusive sign language recognition for human computer interaction. In *Proc. IFAC/IFIP/IFORS/IEA symposium on analysis, design and evaluation of human machine systems*, 2004.

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

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