3D Human Motion Analysis in Monocular Video
Techniques and Challenges

Cristian Sminchisescu

TTI-C, University of Chicago Press,
1427 East 60th Street, Chicago, IL 60637
crismin@nagoya.uchicago.edu
http://nagoya.uchicago.edu/~crismin

Abstract. Extracting meaningful 3D human motion information from video sequences is of interest for applications like intelligent human-computer interfaces, biometrics, video browsing and indexing, virtual reality or video surveillance. Analyzing videos of humans in unconstrained environments is an open and currently active research problem, facing outstanding scientific and computational challenges. The proportions of the human body vary largely across individuals, due to gender, age, weight or race. Aside from this variability, any single human body has many degrees of freedom due to articulation and the individual limbs are deformable due to moving muscle and clothing. Finally, real-world events involve multiple interacting humans occluded by each other or by other objects and the scene conditions may also vary due to camera motion or lighting changes. All these factors make appropriate models of human structure, motion and action difficult to construct and difficult to estimate from images. In this chapter we give an overview of the problem of reconstructing *3D human motion* using sequences of images acquired with a *single video camera*. We explain the difficulties involved, discuss ways to address them using generative and discriminative models and speculate on open problems and future research directions.

Key words: computer vision, statistical models, video analysis, human motion tracking, 3D reconstruction, Bayesian models, numerical optimization.[1]

1 The problem

The problem we address is the reconstruction of full-body 3D human motion in monocular video sequences. This can be formulated either as an *incremental* or as a *batch* problem. In incremental methods, images are available one at a time and one updates estimates of the human pose after each new image observation. This is known as filtering. Batch approaches estimate the pose at each timestep, using a sequence of images, prior and posterior to it. This is known as smoothing.

[1] Chapter in in *Human Motion Understanding, Modeling, Capture and Animation*, R. Kleete, D. Metaxas and B. Rosenhahn Eds., Springer-Verlag, 2007.

It is legitimate to ask why one should restrict attention to only one camera, as opposed to several, in order to attack an already difficult 3D inference problem? The answers are both practical and philosophical. On the practical side, often only a single image sequence is available, when processing and reconstructing movie footage, or when cheap devices are used as interface tools devoted to gesture or activity recognition. A more stringent practical argument is that, even when multiple cameras are available, general 3d reconstruction is complicated by occlusion from other people or scene objects. A robust human motion perception system has to necessarily deal with incomplete, ambiguous and noisy measurements. Fundamentally, these difficulties persist irrespective of how many cameras are used. From a philosophical viewpoint, reconstructing 3D structure using only one eye or a photograph is something that we, as humans, can do. We don't yet know how much is direct computation on 'objective' image information, and how much is prior knowledge in such skills, or how are these combined. But it is probably their conjunction that makes biological vision systems flexible and robust, despite being based on one eye or many. By attacking the 'general' problem instead of focusing on problem simplifications, we hope to make progress towards identifying components of such robust and efficient visual processing mechanisms.

Two general classes of strategies can be used for 3D inference: (*i*) *Generative (top-down) methods* optimize volumetric and appearance-based 3d human models for good alignment with image features. The objective is encoded as an observation likelihood or cost function with optima (ideally) centered at correct pose hypotheses; (*ii*) *Conditional (bottom-up) methods (also referred as discriminative or recognition-based)* predict human poses directly from images, typically using training sets of (pose, image) pairs. Difficulties exist in each case. Some of them, like data association are generic. Others are specific to the class of techniques used: optimizing generative models is expensive and many solutions may exist, some of which spurious, because human appearance is difficult to model accurately and because the problem is non-linear; discriminative methods need to model complex multivalued image-to-3d (inverse) relations.

Organization: The chapter is organized as follows. In §1 and §2 we review the problem of 3d human motion reconstruction and its difficulties. In §3 we introduce generative and conditional models. Learning and inference algorithms are detailed in §4 and §5. In §6 we introduce techniques for combining top-down and bottom-up processing and learning generative and recognition models jointly. We review open problems and conclude in §8.

2 Difficulties

Extracting monocular 3D human motion poses several difficulties that we review. Some are inherent to the use of a single camera, others are generic computer vision difficulties that arise in any complex image understanding problem.

Depth 3D-2D Projection Ambiguities: Projecting the 3D world into images suppresses depth information. This difficulty is fundamental in computer vision.

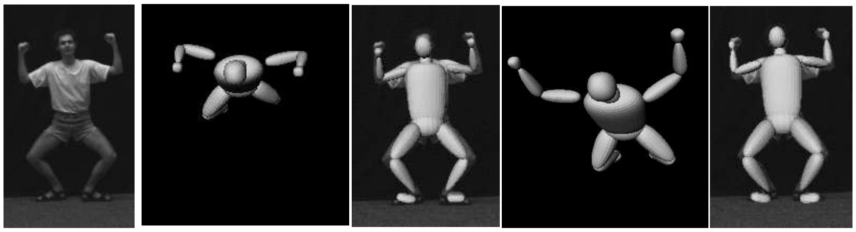

Fig. 1. Reflective Ambiguities (a,b,c,d, e). Original image (a). Two very different configurations of a 3D model (b and d) have image projections that align well with the contour of the imaged human subject (c and e).

Inferring the world from *only one camera*, firmly places our research in the class of science dealing with inverse and ill-posed problems [5]. The non-uniqueness of solution when estimating human pose in monocular images is apparent in the 'forward-backward ambiguities' produced when positioning the human limbs, symmetrically, forwards or backwards, with respect to the camera 'rays of sight' (see fig. 1). Reflecting the limb angles in the frontoparallel plane leaves the image unchanged to first order. For generative models, ambiguities can lead to observation likelihood functions with multiple peaks of somewhat comparable magnitude. The distinction between a global and a local optimum becomes narrow – in this case, we are interested in all optima that are sufficiently good. For discriminative models, the ambiguities lead to multivalued image-pose relations that defeat function approximations based on neural networks or regression. The ambiguity is temporally persistent both under general smooth dynamical models [48] and under dynamics learned from typical human motions [47].

High-Dimensional Representation: Reconstructing 3D human motion raises the question as of what information is to be recovered and how to represent it. A-priori, a model where the 3D human is discretized as densely as possible, with a set of 3D point coordinates, with independent structure and motion is as natural as any other, and could be the most realistic one. Nevertheless, in practice, this would be difficult to constrain since it has excess degrees of freedom for which the bare monocular images cannot account. Representing the human as a blob with centroid coordinates is the opposite extreme, that can be efficient and simpler to estimate at the price of not being particularly informative for 3D reasoning[2]. Consequently, a middle-ground has to be found. At present, this selection is based mostly on intuition and on facts from human structural anatomy. For 3D human tracking the preferred choice remains a kinematic representation with a skeletal structure covered with 'flesh' of more or less complex type (cones, cylinders, globally deformable surfaces). For motion estimation, the model can have, depending on the level of detail, in the order of 30-60 joint angle variables – enough to reproduce a reasonable class of human motions with accuracy. How-

[2] Apart from tractability constraints, the choice of a representation is also application dependent. For many applications, a hierarchy of models with different levels of complexity, depending on context, may be the most appropriate.

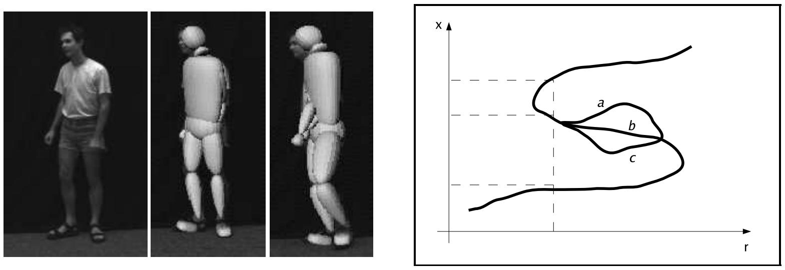

Fig. 2. *(Left)* Physical constraint violations when joint angle limits or body part non-penetration constraints are not enforced. *(Right)* Illustrative example of ambiguities during dynamic inference, for a model with 1d state x and observation r. The S-like distribution implies that multiple state hypotheses (shown in dashed) may exists for certain observations. The ambiguity persists for observations sequences commonly falling under each individual 'S-branch' (up, middle, bottom), see also fig. 6. The close loops created by the splitting-merging of trajectories a, b and c abstract real imaging situations, as in fig. 1, see also [48]. Due to their loopy nature, these ambiguities cannot be resolved even when considering long observation time-scales.

ever, estimation in high-dimensional spaces is computationally expensive, and exhaustive or random search is practically infeasible. Existing algorithms rely on approximations or problem-dependent heuristics: temporal coherency, dynamical models, and symmetries (*e.g.* hypotheses generated using forward-backward flips of limbs, from a given configuration). From a statistical perspective, more rigorous is to follow a learned data-driven approach *i.e.* a minimal representation with intrinsic dimension based on its capacity to synthesize the variability of human shapes and poses present in the tracking domain. Sections §4.3 and §2 discuss techniques for learning low-dimensional models and for estimating their intrinsic dimensionality.

Appearance Modeling, Clothing: Not operating with a anatomically accurate human body models is in most applications offset by outer clothing that deforms. This exhibits strong variability in shape and appearance, both being difficult to model.

Physical Constraints: Physically inspired models based on kinematic and volumetric parameterizations can be used to reason about the physical constraints of real human bodies. For consistency, the body parts have to not penetrate eachother and the joint angles should only have limited intervals of variation (see fig. 2). For estimation, the presence of constraints is both good and bad news. The good news is that the admissible state space volume is smaller than initially designed, because certain regions are not reachable, and many physically unrealistic solutions may be pruned. The bad news is that handling the constraints automatically is non-trivial, especially for continuous optimization methods used in generative models.

Self-Occlusion: Given the highly flexible structure of an articulated human body, self-occlusion between different body parts occurs frequently in monocular views and has to be accounted for. Self-occlusion is an observation ambiguity (see section below). Several aspects are important. First is occlusion detection or prediction, so as to avoid the mis-attribution of image measurements to occluded model regions that have not generated any contribution to image appearance. The second aspect is the management of uncertainty in the position of the body parts that are not visible. Improperly handled this can produce singularities. It is appropriate to use prior-knowledge acquired during learning in order to constrain the uncertainty of unobserved body parts, based on the state of visible ones. Missing data is filled-in using learned correlations typically observed in natural human motions.

For generative models, occlusion raises the additional problem of constructing of an observation likelihood that realistically reflects the probability of different configurations under partial occlusion and viewpoint change. Independence assumptions are often used to fuse likelihoods from different measurements, but this conflicts with occlusion, which is a relatively coherent phenomenon. For realistic likelihoods, the probabilities of both occlusion and measurement have to be incorporated, but this makes the computations intractable.

General Unconstrained Motions: Humans move in diverse, but also highly structured ways. Certain motions have a repetitive structure like running or walking, others represent 'cognitive routines' of various levels of complexity, *e.g.* gestures during a discussion, or crossing the street by checking for cars to the left and to the right, or entering one's office in the morning, sitting down and checking e-mail. It is reasonable to think that if such routines could be identified in the image, they would provide strong constraints for tracking and reconstruction with image measurements serving merely to adjust and fine tune the estimate. However, human activities are not simply preprogrammed – they are parameterized by many cognitive and external un-expected variables (goals, locations of objects or obstacles) that are difficult to recover from images and several activities or motions are often combined.

Kinematic Singularities: These arise when the kinematic Jacobian looses rank and the associated numerical instability can lead to tracking failure. An example is the non-linear rotation representation used for kinematic chains, for which no singularity-free minimal representation exists[3].

Observation Ambiguities: Ambiguities arise when a subset of the model state cannot be directly inferred from image observations. They include but are by no means limited to kinematic ambiguities. Observability depends on the design of the observation model and image features used. (Prior knowledge becomes important and the solutions discussed for self-occlusion are applicable.) For instance when an imaged limb is straight and an edge-based observation likelihood is used with a symmetric body part model, rotations around the limb's own axis cannot be observed – the occluding contour changes little when the limb rotates around its own axis. Only when the elbow moves the uncertain axial parameter values can be constrained. This may *not* be ambiguous under an intensity-based model, where the texture flow can make the rotation observable.

[3] Non-singular over-parameterizations exist, but they are not unique.

Data Association Ambiguities: Identifying which image features belong to the person and which to the background is a general vision difficulty known as data association. For our problem this is amplified by distracting clutter elements that resemble human body parts, *e.g.* various types of edges, ridges or pillars, trees, bookshelves, encountered in man-made and natural environments.

Lighting and Motion Blur: Lighting changes form another source of variability whenever image features based on edge or intensity are used. Artificial edges are created by cast shadows and inter-frame lighting variations could lead to complicated, difficult to model changes in image texture. For systems with a long shutter time, or during rapid motion, image objects appear blurred or blended with the background at motion boundaries. This has impact on the quality of both static feature extraction methods, and of frame to frame algorithms, such as the ones that compute the optical flow.

3 Approaches: Generative and Conditional Models

Approaches to tracking and modeling can be broadly classified as **generative** and **discriminative**. They are similar in that both require a state representation $\mathbf{x}$, here a 3D human model with kinematics (joint angles) or shape (surfaces or joint positions), and both use a set of image features as observations $\mathbf{r}$ for state inference. Often, a training set, $\mathcal{T} = \{(\mathbf{r}_i, \mathbf{x}_i) \mid i = 1 \dots N\}$ sampled from the *joint distribution* is available. (For unsupervised problems, samples from *only* the state or *only* the observation distribution may be available to use.) The computational goal for both approaches is common: the conditional distribution, or a point estimate, for the model state, given observations.[4] Clearly, an important design choice is the state representation and the observation descriptor. The state should have representation and dimensionality well-calibrated to the variability of the task, whereas the observation descriptor is subject to selectivity-invariance trade-offs: it needs to capture not only discriminative, subtle image detail, but also the strong, stable dependencies necessary for learning and generalization. Currently, these are by and large, obtained by combining a-priori design and off-line unsupervised learning. But once decided upon, the representation (model state + observation descriptor) is no longer free, but known and fixed for subsequent learning and inference stages. This holds notwithstanding of the method type, be it generative or discriminative.

Generative algorithms typically model the joint distribution using a constructive form of the observer – the observation likelihood, with maxima ideally centered at correct pose hypotheses. Inference involves complex state space search in order to locate the likelihood peaks, using either non-linear optimization or sampling. Bayes' rule is then used to compute the state conditional from

[4] This classification and statement of purpose is quite general. Methods may deviate from it in a way or another and shortcuts may be taken. But this shouldn't undermine the usefulness of a framework for formal reasoning where to state the assumptions made and the models used, as well as the circumstances when these are expected to perform optimally – see fig. 3.

the observation model and the state prior. Learning can be both supervised and unsupervised. This includes state priors [8, 21, 13, 44], low-dimensional models [47, 64] or learning the parameters of the observation model, *e.g.* texture, ridge or edge distributions, using problem-dependent, natural image statistics [42, 38]. Temporal inference is framed in a clear probabilistic and computational framework based on mixture filters or particle filters [23, 13, 12, 56, 57, 59, 44].

It has been argued that generative models can flexibly reconstruct complex unknown motions and can naturally handle problem constraints. It has been counter-argued that both flexibility and modeling difficulties lead to expensive, uncertain inference [13, 43, 57, 48], and a constructive form of the observer is both difficult to build and somewhat indirect with respect to the task, which requires conditional state estimation and not conditional observation modeling. These arguments motivate the complementary study of **discriminative algorithms** [37, 34, 41, 63, 2, 18] which model and predict the state conditional directly in order to simplify inference. Prediction however involves missing (state) data, unlike learning which is supervised. But learning is also difficult because modeling perceptual data requires adequate representations of highly multimodal distributions. The presence of multiple solutions in the image-to-pose mapping implies that, strictly, this is multivalued and cannot be functionally or globally approximated. However, several authors made initial progress using single hypothesis schemes [41, 34, 63, 2, 18]. *E.g.* nearest-neighbor [34, 41, 63] and regression [2, 18] have been used with good results. Others used mixture models [37, 2] to cluster the joint distribution of (observation, state) pairs and fitted function approximators (neural network or regressor) to each partition. In §5, we will review our BM^3E, a formal probabilistic model based on mixture of experts and conditional temporal chains [49, 51, 52].

Notation: We discuss generative and conditional models based on the graphical dependency in fig. 3. These have continuous temporal states $\mathbf{x}_t$, observations $\mathbf{r}_t$, observation model $p(\mathbf{r}_t|\mathbf{x}_t)$, and dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, $t = 1 \dots T$ (for generative models). For conditional models, we model the conditional state distribution $p(\mathbf{x}_t|\mathbf{r}_t)$ and a previous state/current observation-based density $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$. $\mathbf{X}_t = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t)$ is the model joint state estimated based on a time series of observations $\mathbf{R}_t = (\mathbf{r}_1, \dots, \mathbf{r}_t)$.

4 Generative Methods

Consider a non-linear generative model $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})$ with $d = \dim(\mathbf{x})$, and parameters $\boldsymbol{\theta}$. Without loss of generality, assume a robust observation model:

$$p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x}) = (1 - w) \cdot \mathcal{N}(\mathbf{r}; \mathcal{G}(\mathbf{x}), \Sigma_{\boldsymbol{\theta}}) + o_{\boldsymbol{\theta}} \cdot w \tag{1}$$

This corresponds to a mixture of a Gaussian having mean $\mathcal{G}(\mathbf{x})$ and covariance $\Sigma_{\boldsymbol{\theta}}$, and a uniform background of outliers $o_{\boldsymbol{\theta}}$ with proportions given by w. The outlier process is truncated at large values, so the mixture is normalizable.

In our case, the state space $\mathbf{x}$ represents human joint angles, the parameters $\boldsymbol{\theta}$ may include the Gaussian observation noise covariance, the weighting of outliers, the human body proportions, *etc.* $\mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x})$ is a non-linear transformation that predicts human contours, internal edges and possibly appearance (it includes non-linear kinematics, occlusion analysis and perspective projection), according to consistent kinematic constraints. Alternatively, we also use an equivalent energy-based model – the maxima in probability or the minima in energy have similar meaning and are used interchangeably:

$$p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r}) = p_{\boldsymbol{\theta}}(\mathbf{r}|\mathbf{x})p(\mathbf{x}) = \frac{1}{Z_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})}\exp(-E_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})) \tag{2}$$

$$E_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r}) = -\log[(1-w)\mathcal{N}(\mathbf{r};\mathcal{G}(\mathbf{x}),\mathit{\Sigma}_{\boldsymbol{\theta}}) + o_{\boldsymbol{\theta}}w] + E_{\boldsymbol{\theta}}(\mathbf{x}) - \log Z_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r}) \tag{3}$$

with prior $E_{\boldsymbol{\theta}}(\mathbf{x})$ and normalization constant $Z_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r}) = \int_{(\mathbf{x},\mathbf{r})}\exp(-E_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r}))$. Notice that $Z_{\boldsymbol{\theta}}(\mathbf{x}) = \int_{\mathbf{r}}\exp(-E_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r}))$ can be easily computed by sampling from the mixture of Gaussian and uniform outlier distribution, but computing $Z_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})$ and $Z_{\boldsymbol{\theta}}(\mathbf{r}) = \int_{\mathbf{x}}\exp(-E_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})$ is intractable because the averages are taken w.r.t. the unknown state distribution.[5]

4.1 Density Propagation using Generative Models

For filtering, we compute the optimal state distribution $p(\mathbf{x}_t|\mathbf{R}_t)$, conditioned by observations $\mathbf{R}_t$ up to time t. The recursion can be derived as [20, 22, 24, 25, 46] (fig. 3b):

$$p(\mathbf{x}_t|\mathbf{R}_t) \;=\; \frac{1}{p(\mathbf{r}_t|\mathbf{R}_{t-1})}p(\mathbf{r}_t|\mathbf{x}_t)\int p(\mathbf{x}_t|\mathbf{x}_{t-1})\,p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})\mathbf{d}\mathbf{x}_{t-1} \tag{4}$$

The joint distribution factorizes as:

$$p(\mathbf{X}_T,\mathbf{R}_T) = p(\mathbf{x}_1)\prod_{t=2}^{T}p(\mathbf{x}_t|\mathbf{x}_{t-1})\prod_{t=1}^{T}p(\mathbf{r}_t|\mathbf{x}_t) \tag{5}$$

4.2 Optimization and Temporal Inference Algorithms

Several general-purpose sampling and optimization algorithms have been proposed in order to efficiently search the high-dimensional human pose space. In a temporal framework the methods keep a running estimate of the posterior distribution over state variable (either sample-based or mixture-based) and update it based on new observations. This works time-recursively, the starting point(s) for the current search being obtained from the results at the previous time step, perhaps according to some noisy dynamical model. To the (often limited) extent that the dynamics and the image matching cost are statistically realistic,

[5] The choice of predicted and measured image features, hence the exact specification of the observation model, albeit very important, will not be further discussed.

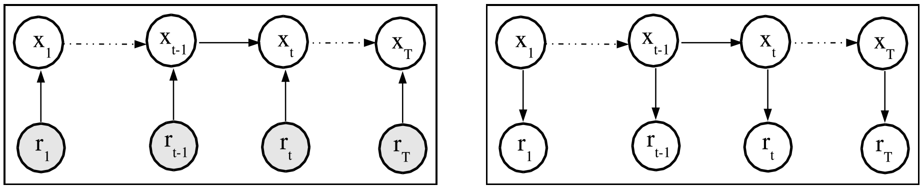

Fig. 3. A conditional/discriminative temporal chain model *(a, left)* reverses the direction of the arrows that link the state and the observation, compared with a generative one *(b, right)*. The state conditionals $p(\mathbf{x}_t|\mathbf{r}_t)$ or $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ can be learned using training pairs and directly predicted during inference. Instead, a generative approach *(b)* will model and estimate $p(\mathbf{r}_t|\mathbf{x}_t)$ and do a more complex probabilistic inversion to compute $p(\mathbf{x}_t|\mathbf{r}_t)$ via Bayes' rule. Shaded nodes reflect variables that are not modeled but conditioned upon.

Bayes-law propagation of a probability density for the true state is possible. For linearized unimodal dynamics and observation models under least squares / Gaussian noise, this leads to Extended Kalman Filtering. For likelihood-weighted random sampling under general multimodal dynamics and observation models, bootstrap filters [20] or Condensation [23] result. In either case various model parameters must be tuned and it sometimes happens that physically implausible settings are needed for acceptable performance. In particular, to control mistracking caused by correspondence errors, selection of slightly incorrect inverse kinematics solutions, and similar model identification errors, visual trackers often require exaggerated levels of dynamical noise. The problem is that even quite minor errors can pull the state estimate a substantial distance from its true value, especially if they persist over several time steps. Recovering from such an error requires a state space jump greater than any that a realistic random dynamics is likely to provide, whereas using an exaggeratedly noisy dynamics provides an easily controllable degree of local randomization that often allows the mistracked estimate to jump back onto the right track. Boosting the dynamical noise does have the side effect of reducing the information propagated from past observations, and hence increasing the local uncertainty associated with each mode. But this is a small penalty to pay for reliable tracking lock, and in any case the loss of accuracy is often minor in visual tracking, where weak dynamical models (*i.e.* short integration times: most of the state information comes from current observations and dynamical details are unimportant) are common. The critical component in most nowday trackers remains the method that searches the observation likelihood at a given timestep based on initializations from the previous one.

General Search Algorithms: Importance sampling [43] and annealing [35, 13] have been used to construct layered particle filters which sample with increased sensitivity to the underlying observation likelihood in order to better focus samples in probable regions. Methods based on Hybrid Monte-Carlo [17, 12, 55] use the gradient of the sampling distribution in order to generate proposals that are accepted more frequently during a Markov Chain Monte Carlo simulation. Hyperdynamic Sampling [55] modifies the sampling distribution based on its local gradient and curvature in order to avoid undesirable trapping in local optima. This creates bumps in the regions of negative curvature in the core of the maxima. Samples are specifically repelled towards saddle-points, so to make inter-maxima transitions occur more frequently. Hyperdynamic Sampling is complementary and can be used in conjunction with both Hybrid-Monte Carlo and/or annealing. Non-parametric belief propagation [59, 44] progressively computes partial sample-based state estimates at each level of a temporal (or spatial, *e.g.* body like structured) graphical model. It uses belief propagation and fits compact mixture approximations to the sample-estimated conditional posteriors at each level along the way.

Eigenvector Tracking and Hypersurface Sweeping [54] are saddle-point search algorithms. They can start at any given local minimum and climb uphill to locate a first-order saddle point – a stable point with only one negative curvature, hence a local maximum in one state space dimension and a local minimum in all the other dimensions. From the saddle it is easy to slide downhill to a nearby optimum using gradient descent and recursively resume the search. For high-dimensional problems many saddle points with different patterns of curvature exist, but the first-order ones are potentially the most useful. They are more likely to lead to low-cost nearby local minima because, from any given one, only one dimension is climbed uphill.

Problem Specific Algorithms: Covariance Scaled Sampling (CSS) [56] is a probabilistic method which represents the posterior distribution of hypotheses in state space as a mixture of long-tailed Gaussian-like distributions whose weights, centers and scale matrices ('covariances') are obtained as follows. Random samples are generated, and each is optimized (by nonlinear local optimization, respecting any joint constraints, *etc.*) to maximize the local posterior likelihood encoded by an image- and prior-knowledge based cost function. The optimized likelihood value and position give the weight and center of a new component, and the inverse Hessian of the log-likelihood gives a scale matrix that is well adapted to the contours of the cost function, even for very ill-conditioned problems like monocular human tracking. However, when sampling, particles are deliberately scattered more widely than a Gaussian of this scale matrix (covariance) would predict, in order to probe more deeply for alternative minima.

Kinematic Jump Sampling (KJS) [57] is a domain-specific sampler, where each configuration of the skeletal kinematic tree has an associated *interpretation tree* — the tree of all fully- or partially-assigned 3D skeletal configurations that can be obtained from the given one by forwards/backwards flips. The tree contains only, and generically all, configurations that are image-consistent in the sense that their joint centers have the same image projections as the given one. (Some of these may still be inconsistent with other constraints: joint limits, body self-intersection or occlusion). The interpretation tree is constructed by traversing the kinematic tree from the root to the leaves. For each link, we construct the 3D sphere centered on the currently hypothesized position of the link's root, with radius equal to link length. This sphere is pierced by the camera ray of sight through the observed image position of the link's endpoint to give (in general) two possible 3D positions of the endpoint that are consistent with the image observation and the hypothesized parent position (see fig. 1). Joint angles are then recovered for each position using simple closed-form inverse kinematics. KJS can be used in conjunction with CSS in order to handle data association ambiguities. Both CSS and KJS can be used in conjunction with non-linear mixture smoothers [48] in order to optimally estimate multiple human joint angle *trajectory hypotheses* based on video sequences.

4.3 Learning

We review unsupervised and supervised methods for learning generative human models. These are applicable to obtain both model representations (state and observation) and parameters.

Learning Representations Unsupervised methods have recently been used to learn state representations that are lower-dimensional, hence better adapted for encoding the class of human motions in a particular domain, *e.g.* walking, running, conversations or jumps [47, 64, 31]. We discuss methods trained on sequences of high-dimensional joint angles obtained from human motion capture, but other representations, *e.g.* joint positions can be used. The goal is to reduce standard computations like visual tracking in the human joint angle state space – referred here as *ambient space*, to better constrained low-dimensional spaces referred as *perceptual (or latent)*. Learning couples otherwise independent variables, so changes in *any* of the perceptual coordinates change *all* the ambient high-dimensional variables (fig. 4). The advantage of perceptual representations is that image measurements collected at *any* of the human body parts constrain *all* the body parts. This is useful for inference during partial visibility or self-occlusion. A disadvantage of perceptual representations is the loss of physical interpretation – joint angle limit constraints are simple to express and easy to enforce as per-variable, localized inequalities in ambient space, but hard to separate in a perceptual space, where they involve (potentially complex) relations among all variables. The following aspects are important when designing latent variable models:

(*i*) *Global perceptual coordinate system:* To make optimization efficient in a global coordinate system is necessary. This can be obtained with any of several dimensionality reduction methods including Laplacian Eigenmaps, ISOMAP, LLE, etc [4, 61, 39, 14]. The methods represent the training set as a graph with local connections based on Euclidean distances between high-dimensional points. Local embeddings aim to preserve the local geometry of the dataset whereas ISOMAP conserves the global geometry (the geodesics on the manifold approximated as shortest paths in the graph). Learning the perceptual representation involves embedding the graph with minimal distortion. Alternatively the perceptual space can be represented with a mixture of low-dimensional local models with separate coordinate systems. In this case, one either has to manage the transition between coordinate systems by stitching their boundaries, or to align, post-hoc, the local models in a global coordinate system. The procedure is more complex and the coordinates not used to estimate the alignment, or out of sample coordinates, may still not be unique. This makes global optimization based on gradient methods non-trivial.

(*ii*) *Preservation of intrinsic curvature:* The ambient space may be intrinsically curved due to the physical constraints of the human body or occlusion [15]. To preserve the structure of the ambient space when embedding, one needs to use methods that preserve the local geometry. *e.g.* Laplacian eigenmaps, LLE or Hessian embeddings [4, 39, 14]. ISOMAP would not be adequate, because geodesics running around a curved, inadmissible ambient region, will be mapped, at curvature loss, to straight lines in perceptual space.

(*iii*) *Intrinsic Dimensionality:* It is important to select the optimal number of dimensions of a perceptual model. Too few will lead to biased, restricted models that cannot capture the variability of the problem. Too many dimensions will lead to high variance estimates during inference. A useful sample-based method to estimate the intrinsic dimensionality is based on the Hausdorff dimension, and measures the rate of growth in the number of neighbors of a point as the size of its neighborhood increases. At the well calibrated dimensionality, the increase should be exponential in the intrinsic dimension. This is illustrated in fig. 4, which shows analysis of walking data obtained using human motion capture. Fig. 4(a) shows Hausdorff estimates for the intrinsic dimensionality: $d = \lim_{r \to 0} \frac{\log N(r)}{\log(1/r)}$, where r is the radius of a sphere centered at each point, and $N(r)$ are the number of points in that neighborhood (the plot is averaged over many nearby points). The slope of the curve in the linear domain $0.01 - 1$ corresponds roughly to a 1d hypothesis. Fig. 4(b) plots the embedding distortion, computed as the normalized Euclidean SSE over each neighborhood in the training set. Here, 5-6 dimensions appear sufficient for a model with low-distortion.

(*iv*) *Continuous generative model:* Continuous optimization in a low dimensional, perceptual space based on image observations requires not only a global coordinate system but also a global continuous mapping between the perceptual and observation spaces. Assuming the high-dimensional ambient model is continuous, the one obtained by reducing its dimensionality should also be. For example, a smooth mapping between the perceptual and the ambient space can be estimated using function approximation (*e.g.* kernel regression, neural networks) based on high-dimensional points in both spaces (training pairs are available once the embedding is computed). A perceptual continuous generative model enables the use of continuous methods for high-dimensional optimization [12, 58, 56, 57]. Working in perceptual spaces indeed targets dimensionality reduction but for many complex processes, even reduced representations would still

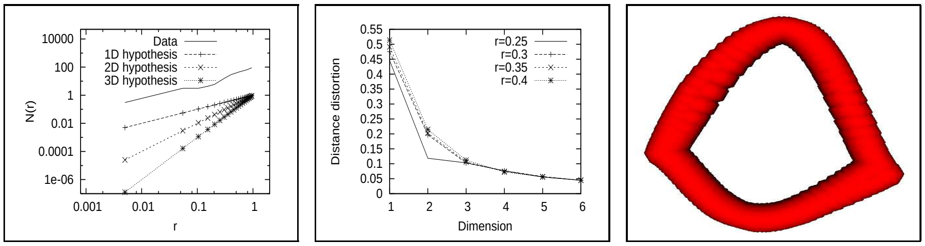

Fig. 4. Analysis of walking data. **(a)** Estimates of intrinsic dimensionality based on the Hausdorff dimension. **(b)** Geometric distortion vs. neighborhood size for a Laplacian embedding method. **(c)** Embedding of a walking data set of 2500 samples in 2d. Also shown, the Gaussian mixture prior (3 stdev), modeling the data density in perceptual space.

have large dimensionality (*e.g.* 10d–15d) – efficient optimizers are still necessary.

(v) *Consistent estimates* impose not only a prior on probable regions in perceptual space, as measured by the typical training data distribution, but also the separation of holes produced by insufficient sampling from genuine intrinsic curvature, *e.g.* due to physical constraints. The inherent sparsity of high-dimensional training sets makes the disambiguation difficult, but analytic expressions can be derived using a prior transfer approach. Ambient constrains can be related to perceptual ones, under a change of variables. If physical constraints are given as priors in ambient space $p_a(\mathbf{x}_a)$ and there exist a continuous perceptual-to-ambient mapping $\mathbf{x}_a = \mathbf{F}(\mathbf{x}), \forall \mathbf{x}$, with Jacobian $\mathbf{J_F}$, an equivalent prior in latent space is:

$$p(\mathbf{x}) \propto p_a(\mathbf{F}(\mathbf{x}))\sqrt{|\mathbf{J_F}\mathbf{J_F}^\top|} \tag{6}$$

Low-dimensional generative models based on principles *(i)-(v)* (or a subset of them) have been convincingly demonstrated for 3D human pose estimation [47, 64, 31].

Learning Parameters Generative models are based on normalized probabilities parameterized by $\boldsymbol{\theta}$, that may encode the proportions of the human body, noise variances, feature weighting in the observation model, or the parameters of the dynamical model. For inference, the normalization is not important. For learning, the normalizer is essential in order to ensure that inferred model state distributions peak in the correct regions when presented with typical image data. Here, we only review learning methods for a static generative model $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})$, learning in video will instead use the joint distribution at multiple timesteps $p_{\boldsymbol{\theta}}(\mathbf{X}_T, \mathbf{R}_T)$. It is convenient to work with probabilistic quantities given as Boltzmann distributions, with uniform state priors, *c.f.* (2). Assuming a supervised training set of state-observation pairs, $\{\mathbf{x}^i, \mathbf{r}^i\}_{i=1\ldots N}$, one can use Maximum

Likelihood to optimize the model parameters using a free energy cost function:

$$\mathcal{F} = -\frac{1}{N}\sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{x}^n, \mathbf{r}^n) = \langle E_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})\rangle_{data} + \log Z_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r}) \tag{7}$$

To minimize the free energy we need to compute its gradients:

$$\frac{d\mathcal{F}}{\mathbf{d}\boldsymbol{\theta}} = \left\langle \frac{dE_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})}{\mathbf{d}\boldsymbol{\theta}} \right\rangle_{data} - \left\langle \frac{dE_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})}{\mathbf{d}\boldsymbol{\theta}} \right\rangle_{model} \tag{8}$$

where the second term is equal to the negative derivative of the log-partition function w.r.t. $\boldsymbol{\theta}$. Note that the only difference between the two terms in (8) is the distribution used to average the energy derivative. In the first term we use the empirical distribution, *i.e.* we simply average over the available data-set. In the second term however we average over the model distribution as defined by the current setting of the parameters. Computing the second average analytically is typically too complicated, and approximations are needed.[6] An unbiased estimate can be obtained by replacing the integral by a sample average, where the sample is to be drawn from the model $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})$. Any of the approximate optimization or inference methods described in §4.2 can be used. The goal of learning is to update the model parameters in order to make the training data likely. Normalizing using the partition function $Z_{\boldsymbol{\theta}}$ ensures discrimination: making the true solution likely automatically makes the incorrect competing solutions unlikely. ML learning iteratively reshapes the model state probability distribution to (at least!) infer the correct result on the training set. Results obtained using this learning method to estimate the parameters of a generative model (noise variances, weighting of the image features and the variance of a Gaussian dynamical model) are shown in fig. 5. This corresponds to the video sequence in [48], which films a person walking towards the camera and doing a bow.

5 Conditional and Discriminative Models

In this section we describe BM^3E, a Conditional $\underline{B}$ayesian $\underline{M}$ixture of $\underline{E}$xperts $\underline{M}$arkov $\underline{M}$odel for probabilistic estimates in discriminative visual tracking. The framework applies to temporal, uncertain inference for *continuous state-space models*, and represents the bottom-up counterpart of pervasive top-down generative models estimated with Kalman filtering or particle filtering (§4).[7] But instead of inverting a generative observation model at run-time, we learn to cooperatively predict complex state distributions directly from descriptors encoding image observations. These are integrated in a conditional graphical model in

[6] The problem is simpler if the prior energy $E_{\boldsymbol{\theta}}(\mathbf{x})$ is fixed and not learned and only the 'easier' partition function $Z_{\boldsymbol{\theta}}(\mathbf{x})$ needs to be computed. The problem remains hard ($Z_{\boldsymbol{\theta}}(\mathbf{r})$) for a hybrid conditional model expressed using generative energies.

[7] Unlike most generative models, systems based on BM^3E can automatically initialize and recover from failure – an important feature for reliable 3D human pose tracking.

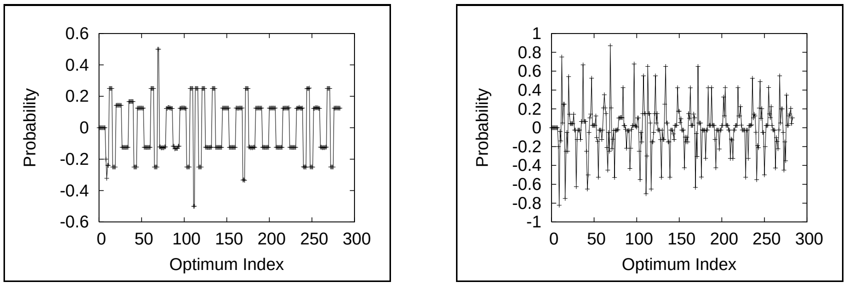

Fig. 5. We show the trajectory probability through each optimum of the observation model at each timestep in a video sequence before (left) and after ML learning (right). The video films a person walking towards a camera and doing a bow [48]. The time is unfolded on the x axis and we switch sign in-between successive timesteps for visualization (the values are all normally positive). Before learning, the temporal trajectory distribution collapses to fewer components in regions where the uncertainty of the model-image matching cost diminishes, but is multimodal and has high entropy. The distribution has lower entropy after learning, showing the usefulness of this procedure. The ambiguity diminishes significantly, but does not disappear. The entropy of the state posterior after learning reflects some of the limits of modeling and gives intuition about run-time speed and accuracy.

order to enforce temporal smoothness constraints and allow a principled management of uncertainty. The algorithms combine sparsity, mixture modeling, and non-linear dimensionality reduction for efficient computation in high-dimensional continuous state spaces. We introduce two key technical aspects: (1) The density propagation rules for *discriminative inference* in continuous, temporal chain models; (2) Flexible algorithms for *learning* feedforward, multimodal state distributions based on compact, conditional Bayesian mixture of experts models.

5.1 The BM^3E Model

Discriminative Density Propagation We work with a conditional model having chain structure, as in fig. 3a. The filtered density can be derived using the conditional independence assumptions in the graphical model in fig. 3a [33, 51, 52]:

$$p(\mathbf{x}_t|\mathbf{R}_t) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{r}_t)p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})\mathbf{d}\mathbf{x}_{t-1} \tag{9}$$

The conditional joint distribution for T timesteps is:

$$p(\mathbf{X}_T|\mathbf{R}_T) = p(\mathbf{x}_1|\mathbf{r}_1)\prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{r}_t) \tag{10}$$

In fact, (9) and (10) can be derived even more generally, based on a predictive conditional that depends on a larger window of observations up to time t [49], but the advantage of these models has to be contrasted to: (i) Increased amount of data required for training due to higher dimensionality. (*ii*) Increased difficulty to generalize due to sensitivity to timescale and / or alignment with a long sequence of past observations.

In practice, one can model $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ as a conditional Bayesian mixture of M experts (*c.f.* §2). The prior $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ is also represented as a Gaussian mixture with M components. To compute the filtered posterior, one needs to integrate M^2 pairwise products of Gaussians analytically, and use mixture of Gaussian simplification and pruning methods to prevent the posterior from growing exponentially [46, 48].

A discriminative corrective conditional $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ can be more sensitive to incorrect previous state estimates than 'memoryless' distributions like $p(\mathbf{x}_t|\mathbf{r}_t)$. However we assume, as in any probabilistic approach, that the training and testing data are representative samples from the true underlying distributions in the domain. In practice, for improved robustness it is straightforward to include an importance sampler based on $p(\mathbf{x}_t|\mathbf{r}_t)$ to eq. (9) – as necessary for initialization or for recovery from transient failure. Equivalently, a model based on a mixture of memoryless and dynamic distributions can be used.

Conditional Bayesian Mixture of Experts Model This section describes the methodology for learning multimodal conditional distributions for discriminative tracking ($p(\mathbf{x}_t|\mathbf{r}_t)$ and $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$ in §5.1). Many perception problems like 3d reconstruction require the computation of inverse, intrinsically multivalued mappings. The configurations corresponding to different static or dynamic estimation ambiguities are peaks in the (multimodal) conditional state distribution (fig. 6). To represent them, we use several 'experts' that are simple function approximators. The experts transform their inputs[8] to output predictions, combined in a probabilistic mixture model based on Gaussians centered at their mean value. The model is consistent across experts and inputs, *i.e.* the mixing proportions of the experts reflect the distribution of the outputs in the training set and they sum to 1 for every input. Some inputs are predicted competitively by multiple experts and have multimodal state conditionals. Other 'unambiguous' inputs are predicted by a single expert, with the others effectively switched-off, having negligible probability (see fig. 6). This is the rationale behind a *conditional* Bayesian mixture of experts, and provides a powerful mechanism for contextually modeling complex multimodal distributions. Formally this is described by:

$$Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) = p(\mathbf{x}|\mathbf{r}, \mathbf{W}, \boldsymbol{\Omega}, \boldsymbol{\lambda}) = \sum_{i=1}^{M} g(\mathbf{r}|\boldsymbol{\lambda}_i) p(\mathbf{x}|\mathbf{r}, \mathbf{W}_i, \boldsymbol{\Omega}_i^{-1}) \tag{11}$$

where:

$$g(\mathbf{r}|\boldsymbol{\lambda}_i) = \frac{f(\mathbf{r}|\boldsymbol{\lambda}_i)}{\sum_{k=1}^{M} f(\mathbf{r}|\boldsymbol{\lambda}_k)} \tag{12}$$

[8] The 'inputs' can be either observations $\mathbf{r}_t$, when modeling $p(\mathbf{x}_t|\mathbf{r}_t)$ or observation-state pairs $(\mathbf{x}_{t-1}, \mathbf{r}_t)$ for $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$. The 'output' is the state throughout. Notice that temporal information is used to learn $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$.

$$p(\mathbf{x}|\mathbf{r}, \mathbf{W}_i, \boldsymbol{\Omega}_i) = \mathcal{N}(\mathbf{x}|\mathbf{W}_i \boldsymbol{\Phi}(\mathbf{r}), \boldsymbol{\Omega}_i^{-1}) \tag{13}$$

Here $\mathbf{r}$ are input or predictor variables, $\mathbf{x}$ are outputs or responses, g are *input dependent* positive gates, computed in terms of functions $f(\mathbf{r}|\boldsymbol{\lambda}_i)$, parameterized by $\boldsymbol{\lambda}_i$. f needs to produce gates g within $[0, 1]$, the exponential and the softmax functions being natural choices: $f_i(\mathbf{r}|\boldsymbol{\lambda}_i) = \exp(\boldsymbol{\lambda}_i^\top \mathbf{r})$. Notice how g are normalized to sum to 1 for consistency, by construction, for any given input $\mathbf{r}$. We choose p as Gaussians (13) with covariances $\boldsymbol{\Omega}_i^{-1}$, centered at different expert predictions, here kernel ($\boldsymbol{\Phi}$) regressors with weights $\mathbf{W}_i$. Both the experts and the gates are learned using sparse Bayesian methods, which provide an automatic relevance determination mechanism [32, 62] to avoid overfitting and encourage compact models with fewer non-zero weights for efficient prediction. The parameters of the model, including experts and gates are collectively stored in $\boldsymbol{\nu} = \{(\mathbf{W}_i, \boldsymbol{\alpha}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i, \boldsymbol{\beta}_i) \mid i = 1 \ldots M\}$.

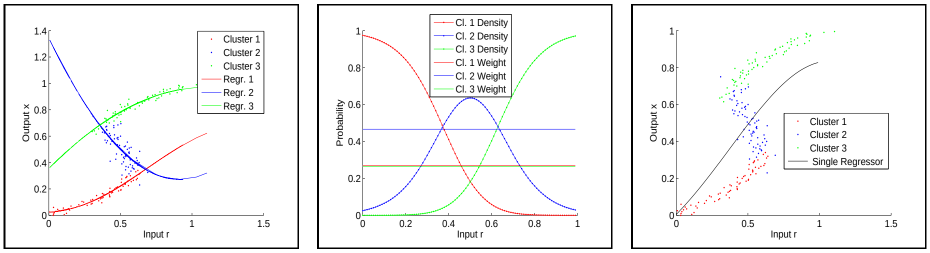

Fig. 6. An illustrative dataset [6] consists of about 250 values of x generated uniformly in $(0, 1)$ and evaluated as $r = x + 0.3\sin(2\pi x) + \epsilon$, with ϵ drawn from a zero mean Gaussian with standard deviation 0.05. Notice that $p(x|r)$ is multimodal. *(a) Left* shows the data colored by the posterior membership probability h (assignment of points to experts) of three expert kernel regressors. *(b) Middle* shows the gates g (12), as a function of the input, but also the three uniform probabilities (of the joint distribution) that are computed by a clusterwise regressor [37]. *(c) Right* shows how a single kernel regressor cannot represent a multivalued dependency (it may either average the different values or commit to an arbitrary one, depending on the kernel parameters).

Learning the conditional mixture of experts involves two layers of optimization. As in many prediction problems, one optimizes the parameters $\boldsymbol{\nu}$ to maximize the log-likelihood of a data set, $\mathcal{T} = \{(\mathbf{r}_i, \mathbf{x}_i) \mid i = 1 \ldots N\}$, *i.e.* the accuracy of predicting $\mathbf{x}$ given $\mathbf{r}$, averaged over the data distribution. For learning, we use a double-loop EM algorithm. This proceeds as follows. In the *E-step* we estimate the posterior over assignments of training points to experts (there is one hidden variable h for each expert-training pair). This gives the probability that the expert i has generated the data n, and requires knowledge of both inputs and outputs. In the *M-step*, two optimization problems are solved: one for each expert and one for its gate. The first learns the expert parameters $(\mathbf{W}_i, \boldsymbol{\Omega}_i)$, based on training data $\mathcal{T}$, weighted according to the current h estimates (the covariances

$\boldsymbol{\Omega}_i$ are estimated from expert prediction errors [66]). The second optimization teaches the gates g how to predict h.[9] The solutions are based on ML-II, with greedy (expert weight) subset selection. This strategy aggressively sparsifies the experts by eliminating inputs with small weights after each iteration [62, 68]. The approximation can can be interpreted as a limiting series of variational approximations (Gaussians with decreasing variances), via dual forms in weight space [68]. **Inference** (state prediction) is straightforward using (11). The result is a conditional mixture distribution with components and mixing probabilities that are input-dependent. In fig. 6 we explain the model using an illustrative toy example, and show the relation with clusterwise and (single-valued) regression.

Learning Conditional Bayesian Mixtures over Kernel Induced State Spaces For many human visual tracking tasks, low-dimensional models are appropriate, because the components of the human state and of the image observation vector exhibit strong correlations, hence low intrinsic dimensionality. In order to efficiently model conditional mappings between high-dimensional spaces with strongly correlated dimensions, we rely on kernel non-linear dimensionality reduction and conditional mixture prediction, as introduced in §2. One can use nonlinear methods like kernel PCA [40, 67] and account for the structure of the

$$\begin{tikzcd}
\mathbf{z} \in \mathcal{P}(\mathcal{F}_r) \arrow[r, "p(\mathbf{y}|\mathbf{z})"] & \mathbf{y} \in \mathcal{P}(\mathcal{F}_x) \arrow[dr] & \\
\boldsymbol{\Phi}_r(\mathbf{r}) \subset \mathcal{F}_r \arrow[u, "kPCA"] & \boldsymbol{\Phi}_x(\mathbf{x}) \subset \mathcal{F}_x \arrow[u, "kPCA"] & \mathbf{x} \approx \mathrm{PreImage}(\mathbf{y}) \arrow[d] \\
\mathbf{r} \in \mathcal{R} \subset \mathbb{R}^r \arrow[u, "\boldsymbol{\Phi}_r"] & \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^x \arrow[u, "\boldsymbol{\Phi}_x"] & p(\mathbf{x}|\mathbf{r}) \approx p(\mathbf{x}|\mathbf{y})
\end{tikzcd}$$

Fig. 7. A learned *conditional Bayesian mixture of low-dimensional kernel-induced experts* predictor to compute $p(\mathbf{x}|\mathbf{r}) \equiv p(\mathbf{x}_t|\mathbf{r}_t), \forall t$. (One can similarly learn $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{r}_t)$, with input $(\mathbf{x}, \mathbf{r})$ instead of $\mathbf{r}$ – here we illustrate only $p(\mathbf{x}|\mathbf{r})$ for clarity.) The input $\mathbf{r}$ and the output $\mathbf{x}$ are decorrelated using Kernel PCA to obtain $\mathbf{z}$ and $\mathbf{y}$ respectively. The kernels used for the input and output are $\boldsymbol{\Phi}_r$ and $\boldsymbol{\Phi}_x$, with induced feature spaces $\mathcal{F}_r$ and $\mathcal{F}_x$, respectively. Their principal subspaces obtained by kernel PCA are denoted by $\mathcal{P}(\mathcal{F}_r)$ and $\mathcal{P}(\mathcal{F}_x)$, respectively. A conditional Bayesian mixture of experts $p(\mathbf{y}|\mathbf{z})$ is learned using the low-dimensional representation $(\mathbf{z}, \mathbf{y})$. Using learned local conditionals of the form $p(\mathbf{y}_t|\mathbf{z}_t)$ or $p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{z}_t)$, temporal inference can be efficiently performed in a *low-dimensional kernel induced state space* (see (9) where $\mathbf{y} \leftarrow \mathbf{x}$ and $\mathbf{z} \leftarrow \mathbf{r}$). For visualization and error measurement, the filtered density $p(\mathbf{y}_t|\mathbf{Z}_t)$ can be mapped back to $p(\mathbf{x}_t|\mathbf{R}_t)$ using a pre-image calculation.

[9] Prediction based on the input *only* is essential for output prediction (state inference), where membership probabilities h cannot be computed because the output is missing.

problem, where both the inputs and the outputs are likely to be low-dimensional and their mapping multivalued (fig. 7). Since temporal inference is performed in the low-dimensional kernel induced state space, backtracking to high-dimensions is only necessary for visualization or error reporting.

6 Learning Joint Generative-Recognition Models

In the previous sections we have reviewed both generative (top-down) and conditional (bottom-up, recognition) models. Despite being a natural way to model the appearance of complex articulated structures, the success of generative models (§4)) has been partly shadowed because it is computational demanding to infer the distribution on their hidden states (human joint angles) and because their parameters are unknown and variable across many real scenes. In turn, conditional models are simple to understand and fast, but often need a generative model for training and could be blind-sighted by the lack of feedback for self-assessing accuracy. In summary, what appears to be necessary is a mechanism to consistently integrate top-down and bottom-up processing: the flexibility of 3d generative modeling (represent a large set of possible poses of human body parts, their correct occlusion and foreshortening relationships and their consistency with the image evidence) with the speed and simplicity of feed-forward processing. In this section we sketch one possible way to meet these requirements based on a bidirectional model with both recognition and generative sub-components – see [53] for details. *Learning* the parameters alternates self-training stages in order to maximize the probability of the observed evidence (images of humans). During one step, the recognition model is trained to invert the generative model using samples drawn from it. In the next step, the generative model is trained to have a state distribution close to the one predicted by the recognition model. At local equilibrium, which is guaranteed, the two models have consistent, registered parameterizations. During *on-line inference*, the estimates can be driven mostly by the fast recognition model, but may include generative (consistency) feedback.

The goal of both learning and inference is to maximize the probability of the evidence (observation) under the data generation model:

$$\log p_{\boldsymbol{\theta}}(\mathbf{r}) = \log \int_{\mathbf{x}} p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r}) = \log \int_{\mathbf{x}} Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) \frac{p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})}{Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})} \tag{14}$$

$$\geq \int_{\mathbf{x}} Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})}{Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})} = KL(Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})||p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})) \tag{15}$$

which is based on Jensen's inequality [25], and KL is the Kullback-Leibler divergence between two distributions. For learning, (14) will sum over the observations in the training set, omitted here for clarity. We have introduced a variational distribution $Q_{\boldsymbol{\nu}}$ and have selected it to be exactly the recognition model. This is the same as maximizing a lower bound on the log-marginal (observation) probability of the generative model, with equality when $Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{r})$.

$$\log p_{\boldsymbol{\theta}}(\mathbf{r}) - KL(Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})||p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{r})) = KL(Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})||p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{r})) \tag{16}$$

Algorithm for Bidirectional Model Learning

E-step: $\boldsymbol{\nu}^{k+1} = \arg\max_{\boldsymbol{\nu}} \mathcal{L}(\boldsymbol{\nu}, \boldsymbol{\theta}^k)$
Train the *recognition* model using samples from the current *generative* model.

M-step: $\boldsymbol{\theta}^{k+1} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\nu}^{k+1}, \boldsymbol{\theta})$
Train the *generative* model to have state posterior close to the one predicted by the current *recognition* model.

Fig. 8. Variational Expectation-Maximization (VEM) algorithm for jointly learning a generative and a recognition model.

According to (14) and (16), optimizing a variational bound on the observed data is equivalent to minimizing the KL divergence between the state distribution inferred by the generative model $p(\mathbf{x}|\mathbf{r})$ and the one predicted by the recognition model $Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})$. This is equivalent to minimizing the KL divergence between the recognition distribution and the joint distribution $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})$ – the cost function we work with:

$$KL(Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r})||p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r})) = -\int_{\mathbf{x}} Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) \log Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) \tag{17}$$

$$+ \int_{\mathbf{x}} Q_{\boldsymbol{\nu}}(\mathbf{x}|\mathbf{r}) \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{r}) = \mathcal{L}(\boldsymbol{\nu}, \boldsymbol{\theta}) \tag{18}$$

The cost $\mathcal{L}(\boldsymbol{\nu}, \boldsymbol{\theta})$ balances two conflicting goals: assign values to states that have high probability under the generative model (the second term), but at the same time be as uncommitted as possible (the first term measuring the entropy of the recognition distribution). The gradient-based learning algorithm is summarized in fig. 8 and is guaranteed to converge to a locally optimal solution for the parameters. The procedure is, in principle, self-supervised (one has to only provide the image of a human *without* the corresponding 3d human joint angle values), but one can initialize by training the recognition and the generative models separately using techniques described in §4 and §5.

Online inference (3d reconstruction and tracking) is straightforward using the E-step in fig. 8. But for efficiency one can work only with the recognition model *c.f.* (11) and only do generative inference (full E-step) when the recognition distribution has high entropy. The model then effectively switches between a discriminative density propagation rule [51, 52] and a generative propagation rule [24, 13, 42, 47]. This offers a natural 'exploitation-exploration' or prediction-search tradeoff. An integrated 3d temporal predictor based on the model operates similarly to existing 2d object detectors. It searches the image at different locations and uses the recognition model to hypothesize 3d configurations. Feedback from the generative model helps to downgrade incorrect competing 3d hypotheses and to decide on the detection status (human or not) at the analyzed image sub-window. In fig. 9 we show results of this model for the *automatic* reconstruction of 3d human motion in environments with background clutter. The framework provides a uniform treatment of human detection, 3d initialization and 3d recovery from transient failure.

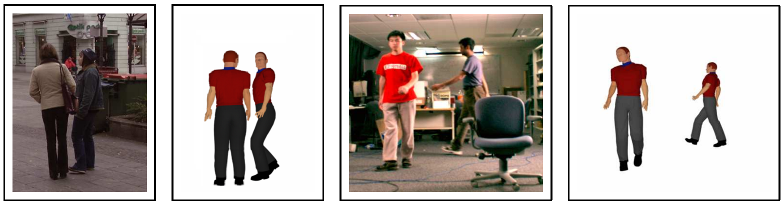

Fig. 9. Automatic human detection and 3d reconstruction using a learned generative-recognition model that combines bottom-up and top-down processing [53]. This shows some of difficulties of *automatically* detecting people and reconstructing their 3d poses in the real world. The background is cluttered, the limb constrast is often low, and there is occlusion from other objects (*e.g.* the chair) or people.

7 Training Sets and Representation

It is difficult to obtain ground truth for 3D human motion and even harder to train using many viewpoints or lighting conditions. In order to gather data one can use packages like Maya (Alias Wavefront) with realistically rendered computer graphics human surface models, animated using human motion capture [37, 41, 18, 47, 51, 52, 2, 63]. 3D human data capture databases have emerged more recently for both motion capture [1, 38] and for human body laser-scans [3]. Alternatively, datasets based on photo-realistic multicamera human reconstruction algorithms can be used [10]. The human representation (**x**) is usually based on an articulated skeleton with spherical joints, and may have 30-60 d.o.f.

8 Challenges and Open Problems

One of the main challenges for the human motion sensing community today is to automatically understand people *in-vivo*. We need to find where the people are, infer their poses, recognize what they do and perhaps what objects do they use or interact with. However, many of the existing human tracking systems tend to be complex to build and computationally expensive. The human structural and appearance models used are often built off-line and learned only to a limited extent. The algorithms cannot seamlessly deal with high structural variability, multiple interacting people and severe occlusion or lighting changes, and the resulting full body reconstructions are often qualitative yet not photorealistic. An entirely convincing transition between the laboratory and the real world remains to be realized.

In the long run, in order to build reliable human models and algorithms for complex, large scale tasks, it is probable that learning will play a major role.

Central themes are likely to be the choice of representation and its generalization properties, the role of bottom-up and top-down processing, and the importance of efficient search methods. Exploiting the problem structure and the scene context can be critical in order to limit inferential ambiguities. Several directions may be fruitful to investigate in order to advance existing algorithms:

- The role of representation. Methods to automatically extract complex, possibly hierarchical models (of structure, shape, appearance and dynamics) with the optimal level of complexity for various tasks, from typical, supervised and unsupervised datasets. Models that can gracefully handle partial views and multiple levels of detail.
- Cost functions adapted for learning human models with good generalization properties. Algorithms that can learn reliably from small training sets.
- Relative advantages of bottom-up (discriminative, conditional) and top-down (generative) models and ways to combine them for initialization and for recovery from tracking failure.
- Inference methods for multiple people and for scenes with complex data association. Algorithms and models able to reliably handle occlusion, clutter and lighting changes. The relative advantages of 2d and 3d models and ways to jointly use them.
- The role of context in resolving ambiguities during state inference. Methods for combining recognition and reconstruction.

Acknowledgments This work has been supported in part by the National Science Foundation under award IIS-0535140. We thank our collaborators: Allan Jepson, Dimitris Metaxas, Bill Triggs and Atul Kanaujia, for their support with different parts of the research described in this chapter.

References

1. CMU Human Motion Capture DataBase. Available online at http://mocap.cs.cmu.edu/search.html, 2003.
2. A. Agarwal and B. Triggs. Monocular human motion capture with a mixture of regressors. In *Workshop on Vision for Human Computer Interaction*, 2005.
3. B. Allen, B. Curless, and Z. Popovic. The space of human body shapes: reconstruction and parameterization from range scans. In *SIGGRAPH*, 2003.
4. M. Belkin and P. Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *Advances in Neural Information Processing Systems*, 2002.
5. M. Bertero, T. Poggio, and V. Torre. Ill-posed Problems in Early Vision. *Proc. of IEEE*, 1988.
6. C. Bishop and M. Svensen. Bayesian mixtures of experts. In *Uncertainty in Artificial Intelligence*, 2003.
7. A. Blake and M. Isard. *Active Contours.* Springer, 2000.
8. M. Brand. Shadow Puppetry. In *IEEE International Conference on Computer Vision*, pages 1237–44, 1999.

9. C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1998.
10. J. Carranza, C. Theobalt, M. Magnor, and H. Seidel. Free-viewpoint video of human actors. In *SIGGRAPH*, 2003.
11. T. Cham and J. Rehg. A Multiple Hypothesis Approach to Figure Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 239–245, 1999.
12. K. Choo and D. Fleet. People Tracking Using Hybrid Monte Carlo Filtering. In *IEEE International Conference on Computer Vision*, 2001.
13. J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2000.
14. D. Donoho and C. Grimes. Hessian Eigenmaps: Locally Linear Embedding Techniques for High-dimensional Data. *Proc. Nat. Acad. Arts and Sciences*, 2003.
15. D. Donoho and C. Grimes. When Does ISOMAP Recover the Natural Parameterization of Families of Articulated Images? Technical report, Dept. of Statistics, Stanford University, 2003.
16. T. Drummond and R. Cipolla. Real-time Tracking of Highly Articulated Structures in the Presence of Noisy Measurements. In *IEEE International Conference on Computer Vision*, 2001.
17. S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.
18. A. Elgammal and C. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.
19. D. Gavrila. The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.
20. N. Gordon, D. Salmond, and A. Smith. Novel Approach to Non-linear/Non-Gaussian State Estimation. *IEE Proc. F*, 1993.
21. N. Howe, M. Leventon, and W. Freeman. Bayesian Reconstruction of 3D Human Motion from Single-Camera Video. *Advances in Neural Information Processing Systems*, 1999.
22. M. Isard and A. Blake. A Smoothing Filter for CONDENSATION. In *European Conference on Computer Vision*, 1998.
23. M. Isard and A. Blake. CONDENSATION – Conditional Density Propagation for Visual Tracking. *International Journal of Computer Vision*, 1998.
24. M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *European Conference on Computer Vision*, 1998.
25. M. Jordan. *Learning in graphical models*. MIT Press, 1998.
26. I. Kakadiaris and D. Metaxas. Model-Based Estimation of 3D Human Motion with Occlusion Prediction Based on Active Multi-Viewpoint Selection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 81–87, 1996.
27. R. Kehl, M. Bray, and L. V. Gool. Full body tracking from multiple views using stochastic sampling. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.
28. X. Lan and D. Huttenlocher. Beyond trees: common factor models for 2d human pose recovery. In *IEEE International Conference on Computer Vision*, 2005.
29. H. J. Lee and Z. Chen. Determination of 3D Human Body Postures from a Single View. *Computer Vision, Graphics and Image Processing*, 30:148–168, 1985.

30. M. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.
31. R. Li, M. Yang, S. Sclaroff, and T. Tian. Monocular Tracking of 3D Human Motion with a Coordianted Mixture of Factor Analyzers. In *European Conference on Computer Vision*, 2006.
32. D. Mackay. Bayesian interpolation. *Neural Computation*, 4(5):720–736, 1992.
33. A. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In *International Conference on Machine Learning*, 2000.
34. G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *European Conference on Computer Vision*, 2002.
35. R. Neal. Annealed Importance Sampling. *Statistics and Computing*, 11:125–139, 2001.
36. D. Ramanan and C. Sminchisescu. Training Deformable Models for Localization. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.
37. R. Rosales and S. Sclaroff. Learning Body Pose Via Specialized Maps. In *Advances in Neural Information Processing Systems*, 2002.
38. S. Roth, L. Sigal, and M. Black. Gibbs Likelihoods for Bayesian Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.
39. S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 2000.
40. B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.
41. G. Shakhnarovich, P. Viola, and T. Darrell. Fast Pose Estimation with Parameter Sensitive Hashing. In *IEEE International Conference on Computer Vision*, 2003.
42. H. Sidenbladh and M. Black. Learning Image Statistics for Bayesian Tracking. In *IEEE International Conference on Computer Vision*, 2001.
43. H. Sidenbladh, M. Black, and D. Fleet. Stochastic Tracking of 3D Human Figures Using 2D Image Motion. In *European Conference on Computer Vision*, 2000.
44. L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking Loose-limbed People. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2004.
45. C. Sminchisescu. Consistency and Coupling in Human Model Likelihoods. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 27–32, Washington D.C., 2002.
46. C. Sminchisescu and A. Jepson. Density propagation for continuous temporal chains. Generative and discriminative models. Technical Report CSRG-401, University of Toronto, October 2004.
47. 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.
48. C. Sminchisescu and A. Jepson. Variational Mixture Smoothing for Non-Linear Dynamical Systems. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 608–615, Washington D.C., 2004.
49. C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Learning to reconstruct 3D human motion from Bayesian mixtures of experts. A probabilistic discriminative approach. Technical Report CSRG-502, University of Toronto, October 2004.

50. C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *IEEE International Conference on Computer Vision*, volume 2, pages 1808–1815, 2005.
51. C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 390–397, 2005.
52. C. Sminchisescu, A. Kanaujia, and D. Metaxas. BM^3E: Discriminative Density Propagation for Visual Tracking. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.
53. C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning Joint Top-down and Bottom-up Processes for 3D Visual Inference. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.
54. C. Sminchisescu and B. Triggs. Building Roadmaps of Local Minima of Visual Models. In *European Conference on Computer Vision*, volume 1, pages 566–582, Copenhagen, 2002.
55. C. Sminchisescu and B. Triggs. Hyperdynamics Importance Sampling. In *European Conference on Computer Vision*, volume 1, pages 769–783, Copenhagen, 2002.
56. C. Sminchisescu and B. Triggs. Estimating Articulated Human Motion with Covariance Scaled Sampling. *International Journal of Robotics Research*, 22(6):371–393, 2003.
57. C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular 3D Human Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 69–76, Madison, 2003.
58. C. Sminchisescu and M. Welling. Generalized Darting Monte-Carlo. In *9th International Conference on Artificial Intelligence and Statistics*, 2007.
59. E. Sudderth, A. Ihler, W. Freeman, and A.Wilsky. Non-parametric belief propagation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2003.
60. C. J. Taylor. Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 677–684, 2000.
61. J. Tenenbaum, V. Silva, and J. Langford. A Global Geometric Framewok for Nonlinear Dimensionality Reduction. *Science*, 2000.
62. M. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 2001.
63. C. Tomasi, S. Petrov, and A. Sastry. 3d tracking = classification + interpolation. In *IEEE International Conference on Computer Vision*, 2003.
64. R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking in small training sets. In *IEEE International Conference on Computer Vision*, 2005.
65. S. Wachter and H. Nagel. Tracking Persons in Monocular Image Sequences. *Computer Vision and Image Understanding*, 74(3):174–192, 1999.
66. S. Waterhouse, D.Mackay, and T.Robinson. Bayesian methods for mixtures of experts. In *Advances in Neural Information Processing Systems*, 1996.
67. J. Weston, O. Chapelle, A. Elisseeff, B. Scholkopf, and V. Vapnik. Kernel dependency estimation. In *Advances in Neural Information Processing Systems*, 2002.
68. D. Wipf, J. Palmer, and B. Rao. Perspectives on Sparse Bayesian Learning. In *Advances in Neural Information Processing Systems*, 2003.