Human pose estimation

Leonid Sigal, Disney Research, Pittsburgh

Synonyms

- Articulated pose estimation
- Body configuration recovery

Related Concepts

- Human pose tracking
- People tracking
- Articulated pose tracking
- Body parsing
- People parsing

Definition

Human pose estimation is the process of estimating the configuration of the body (pose) from a single, typically monocular, image.

Background

Human pose estimation is one of the key problems in computer vision that has been studied for well over 15 years. The reason for its importance is the abundance of applications that can benefit from such a technology. For example, human pose estimation allows for higher level reasoning in the context of human-computer interaction and activity recognition; it is also one of the basic building blocks for marker-less motion capture (MoCap) technology. MoCap technology is useful for applications ranging from character animation to clinical analysis of gait pathologies.

Despite many years of research, however, pose estimation remains a very difficult and still largely unsolved problem. Among the most significant challenges are: (1) variability of human visual appearance in images, (2) variability in lighting conditions, (3) variability in human physique, (4) partial occlusions due to self articulation and layering of objects in the scene, (5) complexity of human skeletal structure, (6) high dimensionality of the pose, and (7) the loss of 3d information that results from observing the pose from 2d planar image projections. To date, there is no approach that can produce satisfactory results in general, unconstrained settings while dealing with all of the aforementioned challenges.

Theory and Application

Human pose estimation is typically formulated probabilistically to account for ambiguities that may exist in the inference (though there are notable exceptions, *e.g.* [11]). In such cases, one is interested in estimating the posterior

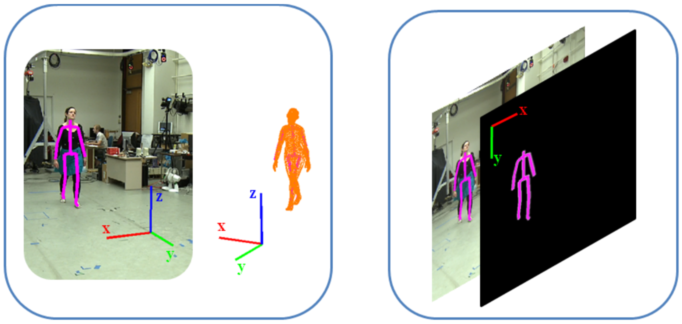

Fig. 1. Skeleton Representation: Illustration of the 3d and 2d kinematic tree skeleton representation on the left and right, respectively.

distribution, $p(\mathbf{x}|\mathbf{z})$, where $\mathbf{x}$ is the pose of the body and and $\mathbf{z}$ is a feature set derived from the image. The key modeling choices that affect the inference are:

- The representation of the pose – $\mathbf{x}$
- The nature and encoding of image features – $\mathbf{z}$
- The inference framework required to estimate the posterior – $p(\mathbf{x}|\mathbf{z})$

Next, the primary lines of research in pose estimation with respect to these modeling choices are reviewed. It is worth noting that these three modeling choices are not always independent. For example, some inference frameworks are specifically designed to utilize a given representation of the pose.

Representation: The configuration of the human body can be represented in a variety of ways. The most direct and common representation is obtained by parameterizing the body as a *kinematic tree* (see Figure 1), $\mathbf{x} = \{\tau, \theta_\tau, \theta_1, \theta_2, ..., \theta_N\}$, where the pose is encoded using position of the root segment (to keep the kinematic tree as short as possible, the pelvis is typically used as the root segment), τ, orientation of the root segment in the world, θ_τ, and a set of relative joint angles, $\{\theta_i\}_{i=1}^N$, that represent the orientations of body parts with respect to their parents along the tree (*e.g.*, the orientation of the thigh with respect to the pelvis, shin with respect to the thigh, *etc.*).

Kinematic tree representation can be obtained for 2d, 2.5d, and 3d body models. In 3d, $\tau \in \mathbb{R}^3$ and $\theta_\tau \in SO(3)$; $\theta_i \in SO(3)$ for spherical joints (*e.g.*, neck), $\theta_i \in \mathbb{R}^2$ for saddle joints (*e.g.*, wrist), and $\theta_i \in \mathbb{R}^1$ for hinge joints (*e.g.*, knee) and represents the pose of the body in the world. Note, the actual representation of the rotations in 3d is beyond the scope of this entry. In 2d, $\tau \in \mathbb{R}^2$ and $\theta_\tau \in \mathbb{R}^1$; $\theta_i \in \mathbb{R}^1$ corresponds to pose of the *cardboard* person in the image plane. 2.5d representations are the least common and are extensions of the 2d representation such that the pose, $\mathbf{x}$, is augmented with (typically discrete) variables encoding the relative depth (layering) of body parts with respect to one

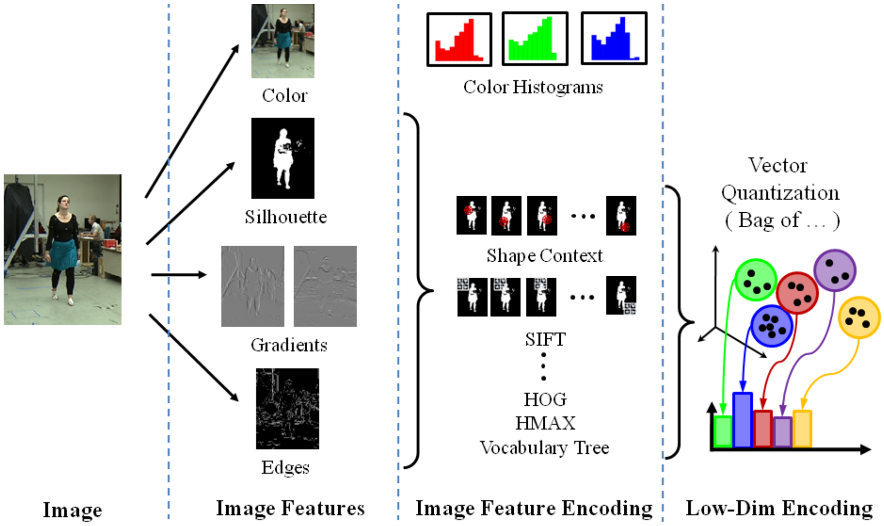

Fig. 2. Image Features: Illustration of common image features and encoding methods used in the literature.

another in the 2d *cardboard* model. In all cases, be it in 2d or 3d, this representation results in a high-dimensional pose vector, $\mathbf{x}$, in $\mathbb{R}^{30} - \mathbb{R}^{70}$, depending on the fidelity and exact parameterization of the skeleton and joints. Alternatively, one can parameterize the pose of the body by 2d or 3d locations of the major joints [6]. For example, $\mathbf{x} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N\}$, where $\mathbf{p}_i$ is the joint location in the world, $\mathbf{p}_i \in \mathbb{R}^3$, or in an image, $\mathbf{p}_i \in \mathbb{R}^2$. This latter representation is less common, however, because it is not invariant to the morphology (body segment lengths) of a given individual.

A typical alternative to the kinematic tree models is to model the body as a set of parts, $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$, each with its own position and orientation in space, $\mathbf{x}_i = \{\tau_i, \theta_i\}$, that are connected by a set of statistical or physical constraints that enforce skeletal (and sometimes image) consistency. Because part-based parameterization is redundant, it results in an even higher-dimensional representation. However, it does so in a way that makes it efficient to infer the pose, as will be discussed in a later section. Methods that utilize such a parameterization are often called *part-based*. As in kinematic tree models, the parts can be defined in 2d [2,4,7,8,11,16] or in 3d [21], with 2d parameterizations being significantly more common. In 2d, each part's representation is often augmented with an additional variable, s_i, that accounts for uniform scaling of the body part in the image, *i.e.*, $\mathbf{x}_i = \{\tau_i, \theta_i, s_i\}$ with $\tau_i \in \mathbb{R}^2$, $\theta_i \in \mathbb{R}^1$ and $s_i \in \mathbb{R}^1$.

Image features: Performance of any pose estimation approach depends substantially on the observations, or image features, that are chosen to represent salient parts of the image with respect to the human pose. A related and equally important issue is one of how these features are encoded. In addition to using different encodings, some approaches propose to reduce the dimensionality of the resulting feature vectors through vector quantization or *bag-of-words* representations. These coarser representations simplify feature matching, but at the expense of losing spatial structure in the image. Common features and encoding methods are illustrated in Figure 2.

Over the years, many features have been proposed by various authors. The most common features include: image silhouettes [1], for effectively separating the person from background in static scenes; color [16], for modeling un-occluded skin or clothing; edges [16], for modeling external and internal contours of the body; and gradients [5], for modeling the texture over the body parts. Less common features include shading and focus [14]. To reduce dimensionality and increase robustness to noise, these raw features are often encapsulated in image descriptors, such as shape context [1,2,6], SIFT [6] and histogram of oriented gradients [5]. Alternatively, hierarchical multi-level image encodings can be used, such as HMAX [12], spatial pyramids [12], and vocabulary trees [12]. The effectiveness of different feature types on pose estimation has been studied in the context of several inference architectures; see [2] and [12] for discussions and quantitative analyses.

Inference (regression models): Characterizing the posterior distribution, $p(\mathbf{x}|\mathbf{z})$, can be done in a number of ways. Perhaps the most intuitive way is to define a parametric [1,6,12] or non-parametric [15,18,22,25] form for the conditional distribution $p(\mathbf{x}|\mathbf{z})$ and learn the parameters of that distribution from a set of training exemplars. This class of models is more widely known as *discriminative methods*, and they have been shown to be very effective for pose estimation. Such methods directly learn $p(\mathbf{x}|\mathbf{z})$ from a labelled dataset of poses and corresponding images, $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^N$, which can be produced artificially using computer graphics software packages (*e.g.*, Poser) [1,12,18]. The inference takes a form of probabilistic regression. Once a regression function is learned, a scanning window approach is typically used at test time to detect a portion of the image (bounding box) where the person resides; $p(\mathbf{x}|\mathbf{z})$ is then used to characterize the configuration of the person in that target window.

The simplest method in this category is the one of linear regression [1], where the body configuration, $\mathbf{x}$, is assumed to be a linear combination of the image features, $\mathbf{z}$, with additive Gaussian noise,

$$\mathbf{x} = A[\mathbf{z} - \mu_z] + \mu_x + \nu; \qquad \nu \sim \mathcal{N}(0, \Sigma);$$

$\mu_x = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$ and $\mu_z = \frac{1}{N}\sum_{i=1}^N \mathbf{z}_i$ are means computed over the training samples to center the data. Alternatively, this can be written as:

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(A[\mathbf{z} - \mu_z] + \mu_x, \Sigma). \tag{1}$$

The regression coefficients, A, can be learned easily from paired training samples, $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^N$, using the least squares formulation (see [1] for details).

Parametric vs. Non-parametric: Parametric discriminative methods [1,6,12] are appealing because the model representation is fixed with respect to the size of the training dataset $\mathcal{D}$. However, simple parametric models, such as Linear Regression [1] or Relevance Vector Machine [1], are unable to deal with complex non-linear relationships between image features and poses. Non-parametric methods, such as Nearest Neighbor Regression [18] or Kernel Regression [18], are able to model arbitrary complex relationships between input features and output poses. The disadvantage of these non-parametric methods is that the model and inference complexity are both functions of the training set size. For example, in Kernel Regression,

$$p(\mathbf{x}|\mathbf{z}) = \sum_{i=1}^{N} \mathcal{K}_x(\mathbf{x}, \mathbf{x}_i) \frac{\mathcal{K}_z(\mathbf{z}, \mathbf{z}_i)}{\sum_{k=1}^{N} \mathcal{K}_z(\mathbf{z}, \mathbf{z}_k)}, \tag{2}$$

where $\mathcal{K}_x(\cdot, \cdot)$ and $\mathcal{K}_z(\cdot, \cdot)$ are kernel functions measuring the similarity of the arguments (*e.g.*, Gaussian kernels), the inference complexity is $O(N)$ (where N is the size of the training dataset). More sophisticated non-parametric methods, such as Gaussian Process Latent Variable Models (GPLVMs), can have even higher complexity; GPLVMs have $O(N^3)$ learning and $O(N^2)$ inference complexity. In practice, non-parametric methods tend to perform better but are slower.

Dealing with ambiguities: If one assumes that $p(\mathbf{x}|\mathbf{z})$ is uni-modal [1], conditional expectation can be used to characterize the plausible configuration of the person in an image given the learned model. For example, for linear regression in Eq. (1), $E[\mathbf{x}|\mathbf{z}] = A[\mathbf{z} - \mu_z] + \mu_x$; for kernel regression in Eq. (2),

$$E[\mathbf{x}|\mathbf{z}] = \sum_{i=1}^{N} \mathbf{x}_i \frac{\mathcal{K}_z(\mathbf{z}, \mathbf{z}_i)}{\sum_{k=1}^{N} \mathcal{K}_z(\mathbf{z}, \mathbf{z}_k)}. \tag{3}$$

In practice, however, most features under standard imaging conditions are ambiguous, resulting in multi-modal distributions. Ambiguities naturally arise in image projections, where multiple poses can result in similar, if not identical, image features (*e.g.*, front- and back-facing poses yield nearly identical silhouette features). To account for these ambiguities, parametric mixture models were introduced in the form of Mixture of Experts [6,12]. Non-parametric alternatives, such as Local Gaussian Process Latent Variable Models (LGPLVM) [25], cluster the data into convex local sets and make uni-modal predictions within each cluster, or search for prominent modes in $p(\mathbf{x}|\mathbf{z})$ [15,22].

Learning: Obtaining the large datasets that are required for learning discriminative models that can generalize across motions and imaging conditions is challenging. Synthetic datasets often do not exhibit the imaging characteristics present in real images, and real fully-labeled datasets are scarce. Furthermore, even if large datasets could be obtained, learning from vast amounts of data is not a trivial task [6]. To address this issue, two solutions were introduced: (1)

learning from small datasets by discovering an intermediate low dimensional latent space for regularization [15,22] and (2) learning in semi-supervised settings, where a relatively small dataset of paired samples is accompanied by a large amount of unlabeled data [12,15,22].

Limitations: Despite popularity and lots of successes, discriminative methods do have limitations. First, they are only capable of recovering a relative 3d configuration of the body and not its position in 3d space. The reason for this is practical, as reasoning about position in 3d space would require prohibitively large training datasets that span the entire 3d volume of the space visible from the camera. Second, their performance tends to degrade as the distributions of test and training data start to diverge; in other words, generalization remains one of the key issues. Lastly, learning discriminative models efficiently from large datasets that cover wide range of realistic activities and postures remains a challenging task.

Inference (generative): Alternatively, one can take a generative approach and express the desired posterior, $p(\mathbf{x}|\mathbf{z})$, as a product of a likelihood and a prior:

$$p(\mathbf{x}|\mathbf{z}) \propto \underbrace{p(\mathbf{z}|\mathbf{x})}_{likelihood} \underbrace{p(\mathbf{x})}_{prior}. \tag{4}$$

Characterizing this high-dimensional posterior distribution is typically hard; hence, most approaches rely on *a posteriori* (MAP) solutions that look for the most probable configurations that are both typical (have high *prior* probability) and can explain the image data well (have high *likelihood*):

$$\mathbf{x}_{MAP} = \arg\max p(\mathbf{x}|\mathbf{z}). \tag{5}$$

Searching for such configurations, however, in the high-dimensional (40+) articulation space is very challenging and most approaches frequently get stuck in local optima. Global hierarchical search methods, such as Annealed Particle Filter [10], have shown some promising results for simple skeletal configurations, where body is mostly upright, and when observations from multiple cameras are available. For the more general articulations and monocular observations, that are often the focus of pose estimation algorithms, this class of methods has not been very successful to date.

Inference (part-based models): To battle the inference complexity of generative models, part-based models have been introduced. These methods originate in the object recognition community with formulation of Fischler and Elschlager (1973) and assume that a body can be represented as an assembly of parts that are connected by constraints imposed by the joints within the skeletal structure (and, sometimes, by the image constraints imposed by projections onto an image plane that account for occlusions). This formulation reduces the inference complexity because likely body part locations can be searched for independently, only considering the nearby body parts that constrain them, which significantly prunes the total search space.

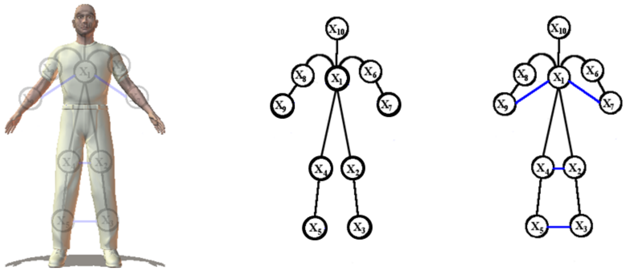

Fig. 3. Pictorial Structures Model: Illustrated is the depiction of the 10-part tree-structured pictorial structures model (middle) and a non-tree-structured (loopy) pictorial structures model (right). In the non-tree-structured model additional constraints encoding occlusions are illustrated in blue.

Among the earliest successes along this line of research is the work of Lee and Cohen [13]. Their approach focused on obtaining proposal maps for the locations of individual joints within an image. These proposal maps were obtained based on a number of features that were computed densely over the image. For example, face detection was used to obtain hypotheses for the location of the head; head-shoulder contour matching, obtained using a deformable contour model and gradient descent, was used as evidence for shoulder joint locations; elliptical skin regions, obtained using skin-color segmentation, were used to determine the locations of the lower arms and lower legs. In addition, second-derivative (ridge) observations were used as evidence for other limbs of the body. Given proposals for the different joints, weighted by the confidence of corresponding detectors, a data-driven Markov Chain Monte Carlo (MCMC) approach was used to recover 3d configurations of the skeleton. This inference relied on direct inverse kinematics (IK) obtained from 2d proposal maps. To further improve the results, a kinematic jump proposal process was also introduced. The kinematic jump proposal process involves flipping a body part or a set of parts (*i.e.*, the head, a hand, or an entire arm) in the depth direction around its pivotal joint.

Other part-based approaches try to assemble regions of an image into body parts and successively construct those parts into a body. Prime examples of such methods are introduced by Mori *et al*. [14] and Ren *et al*. [17]. In [14] super-pixels were first assembled into body parts based on the evaluation of low-level image cues, including contour, shape, shading, and focus. The part-proposals were then pruned and assembled together using length, body part adjacency, and clothing symmetry. A similar approach was taken in [17], but line segments were used instead of assembling super-pixels. Parallel lines were assembled into candidate parts using a set of predefined rules, and the candidate parts were in turn assembled into the body with a set of joint, scale, appearance, and orientation consistency constraints. Unlike [14], the search for the most probable body configurations was formulated as a solution to an Integer Quadratic Programming (IQP) problem.

The most traditional and successful approach, however, is to represent the body using a Markov Random Field (MRF) with body parts corresponding to the nodes and constraints between parts encoded by potential functions that account for physical and statistical dependencies (see Figure 3). Formally, the posterior, $p(\mathbf{x}|\mathbf{z})$, can be expressed as:

$$\begin{aligned} p(\mathbf{x}|\mathbf{z}) &\propto p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) \\ &= p(\mathbf{z}|\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\})p(\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}) \\ &\approx \underbrace{\prod_{i=1}^{M} p(\mathbf{z}|\mathbf{x}_i)}_{likelihood} \underbrace{p(\mathbf{x}_1) \prod_{(i,j)\in E} p(\mathbf{x}_i, \mathbf{x}_j)}_{prior} . \end{aligned} \tag{6}$$

In this case, pose estimation takes the form of inference in a general MRF network. The inference can be solved efficiently using message-passing algorithms, such as Belief Propagation (BP). BP consists of two distinct phases: (1) a set of message-passing iterations are executed to propagate consistent part estimates within a graph, and (2) marginal posterior distributions are estimated for every body part [2,8,16]. A typical formulation looks at the configuration of the body in the 2d image plane and assumes discretization of the pose for each individual part, *e.g.*, $\mathbf{x}_i = \{\tau_i, \theta_i, s_i\}$ where $\tau_i \in \mathbb{R}^2$ is the location and $\theta_i \in \mathbb{R}^1$ and $s_i \in \mathbb{R}^1$ are orientation and scale of the part i (represented as a rectangular patch) in the image plane. As a result, the inference is over a set of discrete part configurations $\mathbf{l}_i \in \mathbb{Z}$ (for part i), where $\mathbb{Z}$ is the enumeration of poses for a part in an image ($\mathbf{l}_i$ is a discrete version of $\mathbf{x}_i$). With an additional assumption of pair-wise potentials that account for kinematic constraints, the model forms a tree-structured graph known as the *Tree-structured Pictorial Structures* (PS) model. An approximate inference with continuous variables is also possible [20,21].

Inference in the tree-structured PS model first proceeds by sending recursively defined messages of the form:

$$m_{i\to j}(\mathbf{l}_j) = \sum_{\mathbf{l}_i} p(\mathbf{l}_i, \mathbf{l}_j)p(\mathbf{z}|\mathbf{l}_i) \prod_{k\in A(i)\setminus j} m_{k\to i}(\mathbf{l}_i), \tag{7}$$

where $m_{i\to j}$ is the message from part i to part j, with $p(\mathbf{l}_i, \mathbf{l}_j)$ measuring the compatibility of poses for the two parts and $p(\mathbf{z}|\mathbf{l}_i)$ the likelihood, and $A(i) \setminus j$ is the set of parts in the graph adjacent to i except for j. Compatibility, $p(\mathbf{l}_i, \mathbf{l}_j)$, is often measured by the physical consistency of two parts at the joint, or by their statistical (*e.g.*, angular) co-occurrence with respect to one another. In a tree-structured PS graph, these messages are sent from the outermost extremities inward and then back outward.

Once all of the message updates are complete, the marginal posteriors for all of the parts can be estimated as:

$$p(\mathbf{l}_i|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{l}_i) \prod_{j\in A(i)} m_{j\to i}(\mathbf{l}_i). \tag{8}$$

Similarly, the most likely configuration can be obtained as a MAP estimate:

$$\mathbf{l}_{i,MAP} = \arg\max_{\mathbf{l}_i} p(\mathbf{l}_i|\mathbf{z}). \tag{9}$$

One of the key benefits of the Pictorial Structures (PS) paradigm is its simplicity and efficiency. In PS exact inference is possible in the time linear to the number of discrete configurations a given part can assume. Because of this property, recent implementations [2] can handle the pixel-dense configurations of parts that result in millions of potential discrete states for each body part. The linear complexity comes from the observation that a generally complex non-Gaussian prior over neighboring parts, $p(\mathbf{x}_i, \mathbf{x}_j)$, can be expressed as a Gaussian prior over the transformed locations corresponding to joints, mainly $p(\mathbf{x}_i, \mathbf{x}_j) = \mathcal{N}(T_{ij}(\mathbf{x}_i); T_{ji}(\mathbf{x}_j), \Sigma_{ij})$. This is done by defining a transformation, $T_{ij}(\mathbf{x}_i)$, that maps a common joint between parts i and j, defined in the part i's coordinate frame, to its location in the image space. Similarly, $T_{ji}(\mathbf{x}_j)$ defines the transformation from the same common joint defined in the j's coordinate frame to the location in the image plane. This transformation allows the inference to use an efficient solution that involves convolution (see [8] for more details).

Performance: Recently, it has been shown that the effectiveness of a PS model is closely tied to the quality of the part likelihoods [2]. Discriminatively trained models [16] and more complex appearance models [2] tend to outperform models defined by hand [8]. Methods that learn likelihood cascades, corresponding to better and better features tuned to a particular image, have also been explored for both superior speed and performance [16]. Most recent discriminative formulation of PS model allows joint learning of part appearances and model structure [26] using structural Support Vector Machine (SVM).

Speed: Cascades of part detectors serve not only to improve performance, but also to speed up the inference (*e.g.*, [23]). Fast likelihoods can be used to prune away large parts of the search space before applying more complex and computationally expensive likelihood models. Other approaches to speed up performance include data-driven methods (*e.g.*, data-driven Belief Propagation). These methods look for the parts in an image first and then assemble a small set of the part candidates into the body (akin to the methods of [14,17]). The problem with such approaches is that any occluded parts are missed altogether because they cannot be detected by the initial part detectors. Inference can also be sped up by using progressive search refinement methods [9]. For example, some methods use upper body detectors to restrict the search to promising parts of the image instead of searching the whole image.

Non-tree structured extensions: Although tree-structured PS models are computationally efficient and exact, they generally are not sufficient to model all the necessary constraints imposed by the body. More complex relationships among the parts that fall outside of the realm of these models include non-penetration constraints and occlusion constraints [20]. Incorporating such relationships into the model adds loops corresponding to long-range dependencies between body parts. These loops complicate inference because: (1) no optimal solutions can be found efficiently (message passing algorithms, like BP, are not guaranteed to converge in loopy graphs) and (2) even approximate inference is typically computationally expensive. Despite these challenges, it has been argued that adding such constraints is necessary to improve performance [4]. To alleviate some of the inference complexities with these non-tree-structured models, a number of competing methods have been introduced. Early attempts used sampling techniques from the tree-structured posterior as proposals for evaluation of a more complex non-tree-structured model [7,8]. To obtain optimality guarantees, branch-and-bound search was recently proposed by Tian *et al.* [24], with the tree-structured solutions as a lower bound on the more complex loopy model energy.

Open problems

Despite much progress in the field, pose estimation remains a challenging and still largely unsolved task. Progress has been made in estimating the configurations of mostly unoccluded and isolated subjects. Open problems include dealing with multiple, potentially interacting people (*e.g.*, [7]), and tolerance to unexpected occlusions. Future research is also likely to expand on the types of postures and imaging conditions that the current algorithms can handle.

To date, most successful pose estimation approaches have been bottom-up. This observation applies to both discriminative approaches and part-based approaches. However, it seems short-sighted to assume that the general pose estimation problem can be solved purely in a bottom-up fashion. Top-down information may be useful for enforcing global pose consistency, and a combination of top-down and bottom-up inference is likely to lead to success faster. The recent success of combining bottom-up part-based models with 3d top-down priors [3] is encouraging and should be built upon to produce models that can deal with more complex postures and motions. Earlier attempts at building hierarchical models [27] may also be worth revisiting with the newfound insights.

Finally, there is significant evidence suggesting that successfully estimating pose independently at every frame is a very ill-posed problem. Spatio-temporal models that aggregate information over time [3] are emerging as a way to regularize performance obtained in individual frames and smooth out the noise in the estimates. Leveraging all sources of generic prior knowledge, such as spatial layout of the body and temporal consistency of poses, and rich image observation models is critical in advancing the state-of-the-art.

Recommended Readings

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

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[3] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[4] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnorr. A study of parts-based object class detection using complete graphs. *Internation Journal of Computer Vision*, 2010.

[5] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, 2010.

[6] L. Bo, C. Sminchisescu, A. Kanaujia, and D. Metaxas. Fast algorithms for large scale conditional 3d prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[7] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *European Conference on Computer Vision*, 2010.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[9] V. Ferrari, M. J. Marn-Jimnez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[10] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *International Journal of Computer Vision*, 87(1–2):75–92, 2010.

[11] H. Jiang. Human pose estimation using consistent max-covering. In *IEEE International Conference on Computer Vision*, 2009.

[12] A. Kanaujia, C. Sminchisescu, and D. Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[13] M. W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[14] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[15] R. Navaratnam, A. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *IEEE International Conference on Computer Vision*, 2007.

[16] D. Ramanan. Learning to parse images of articulated bodies. In *Neural Information and Processing Systems*, 2006.

[17] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pair-wise constraints between parts. In *International Conference on Computer Vision*, 2005.

[18] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *International Conference on Computer Vision*, 2003.

[19] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Neural Information and Processing Systems*, 2007.

[20] L. Sigal and M. J. Black. Measure locally, reason globally:occlusion-sensitive articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[21] L. Sigal, M. Isard, B. H. Sigelman, and M. J. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *Advances in Neural Information Processing Systems*, 2003.

[22] L. Sigal, R. Memisevic, and D. J. Fleet. Shared kernel information embedding for discriminative inference. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[23] V. K. Singh, R. Nevatia, and C. Huang. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In *European Conference on Computer Vision*, pages 314–327, 2010.

[24] T.-P. Tian and S. Sclaroff. Fast globally optimal 2d human detection with loopy graph models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[25] R. Urtasun and T. Darrell. Sparse probabilistic regression for activity-independent human pose inference. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[26] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixture-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[27] J. Zhang, J. Luo, R. Collins, and Y. Liu. Body localization in still images using hierarchical models and hybrid search. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.