Estimating 3D Human Pose from Single Images using Iterative Refinement of the Prior

Ben Daubney and Xianghua Xie

Department of Computer Science, Swansea University, UK
{B.Daubney, X.Xie}@swansea.ac.uk

Abstract

This paper proposes a generative method to extract 3D human pose using just a single image. Unlike many existing approaches we assume that accurate foreground background segmentation is not possible and do not use binary silhouettes. A stochastic method is used to search the pose space and the posterior distribution is maximized using Expectation Maximization (EM). It is assumed that some knowledge is known a priori about the position, scale and orientation of the person present and we specifically develop an approach to exploit this. The result is that we can learn a more constrained prior without having to sacrifice its generality to a specific action type. A single prior is learnt using all actions in the HumanEva dataset [9] and we provide quantitative results for images selected across all action categories and subjects, captured from differing viewpoints.

1. Introduction

Recently, there has been much progress made in the development of discriminative methods that are capable of not only accurately detecting people in cluttered scenes at varying scales [3], but also of being able to estimate orientation [7]. Whilst discriminative methods have also been shown capable of estimating 3D pose from features such as binary silhouettes [1] it is not clear how well these approaches generalize to unseen actions or scenes. It is in this context that generative methods have been shown to be advantageous, since these are not so heavily dependent on the training data used. It seems that a marriage of discriminative and generative methods could provide a good solution to 3D pose estimation from single images. Discriminative methods could be used for location, scale and orientation estimation of the root node (i.e. pelvis location) and generative methods for estimating pose of individual parts. In this work we suppose that such a discriminative detector does exist and explore what benefit this would bring, in particular we present a generative method explicitly designed to exploit this information. Specifically, we are able to move away from "loose limbed" models [5, 10] where limbs are not constrained to be connected at specific joints. Instead of defining our graphical model over parts (or limbs) as is most commonplace [5, 10, 2] we define a model over joints, which are typically of interest.

Whilst prior knowledge of location, scale and orientation considerably constrains the task of 3D pose estimation the remaining problem is by no means trivial. This has been highlighted by recent attempts to extract 2D pose in cluttered images using only local appearance and edge cues where despite relative accurate detection of the torso (81%) detection of the lower limbs is far more difficult (55%) [2].

One of the principal problems with estimating pose is that the human body has many degrees of freedom and as such the search space in which pose is located is extremely large. The approach taken by many existing algorithms is to decompose the object into its principal parts and assume Markovian properties between connected parts. This problem can then be solved using methods such as Dynamic Programming [5] and Belief Propagation [8] for 2D pose estimation and particle methods such as Non-Parametric Belief Propagation [10] and Variational MAP [6] for 3D pose estimation. In the presented method a part based approach is also taken and particle sets are also used to approximate probability distributions. As the position and orientation of a single node in the graph is manually initialized and therefore known with certainty, we propose a novel method to search and grow the pose space. To maximize the posterior we use an iterative method based on Expectation Maximization (EM). At each iteration of the algorithm we maximize the current posterior by re-estimating the prior distribution. As a result the prior converges towards a maxima, which empirically appears to be global.

In this work a single prior is learnt to represent all

actions contained in the HumanEva dataset [9] and the presented algorithm is tested on images taken from all action categories filmed from different viewpoints; we do not as is commonly the case just constrain our model to a single action filmed from a single viewpoint and do not rely on binary silhouette extraction.

2. Model Representation

Typically the conditional probability distribution between two connected parts $p(x_i|x_j, \theta_{ij})$, where x_i is the child of x_j and θ_{ij} is a model parameter, is approximated as a distribution over their relative location and orientation i.e. $p(x_{ij}|\theta_{ij})$ where $x_{ij} = x_j - x_i$ [10, 5]. However, instead we learn a conditional distribution over the orientation of a part measured in the global frame of reference defined by the root node. This will result in a much more representative prior that will help to prevent problems such as self intersection between parts, which would clearly not be present in the training set but can still occur using the approximation described above. To achieve this given a set of training data for two connected joints $\mathbf{X}_i = \{\mathbf{x}_i^1, .., \mathbf{x}_i^l\}$ and $\mathbf{X}_j = \{\mathbf{x}_j^1, .., \mathbf{x}_j^l\}$, we concatenate the data such that the training set becomes $\mathbf{X}_{ij} = \{\mathbf{x}_{ij}^1, .., \mathbf{x}_{ij}^l\}$, where $\mathbf{x}_{ij} = (\mathbf{x}_i, \mathbf{x}_j)$. A Gaussian Mixture Model (GMM) can then be learnt using this data and a separate model is learnt for each pair of connected joints. Each GMM represents the distribution $p(\mathbf{x}_i, \mathbf{x}_j|\theta_{ij})$ and below it is described how to use this distribution to draw a sample $\mathbf{x}_i$ conditioned on a sample $\mathbf{x}_j$.

We first need to determine the number of components to use in each GMM. As the position of the root node is fixed it would be expected that the complexity of the distribution at joints located at a further depth from this root node would increase and more components are required. To represent this we employ the following scheme: Joints connected directly to the root node are given three components and at every subsequent increase in depth a further two components are added. Under our model the maximum number of components is assigned to the wrists with 9 components.

Samples are drawn moving outwards from the root node and for efficiency all GMM's learnt have only diagonal covariance matrices. Given a sample for the jth node $\mathbf{x}_j$, we can create a sample conditioned on this by first calculating the marginal likelihood of observing this value $p(\mathbf{x}_j|m_{ij}^k)$ for the kth component in the GMM. The connection parameters $m_{ij}^k = \{\mu_{ij}^k, \Sigma_{ij}^k, \lambda_{ij}^k\}$ define the mean, covariance and weighting of the component, respectively. Given that all covariance matrices are diagonal, i.e. $\Sigma_{ij}^k = diag(\Lambda_{ii}^k, \Lambda_{jj}^k)$, the marginal likelihood can be calculated as $p(\mathbf{x}_j|m_{ij}^k) = \lambda_{ij}^k \mathcal{N}(\mathbf{x}_j; \mu_j^k, \Lambda_{jj}^k)$. Once this has been calculated for all components the resultant distribution is normalized to give the conditional distribution $p(m_{ij}^k|\mathbf{x}_j)$. A GMM component can then be sampled from this distribution $k^* \sim p(m_{ij}^k|\mathbf{x}_j)$, from which a sample for $\mathbf{x}_i$ can be drawn from the selected component $\mathbf{x}_i \sim \mathcal{N}(\mu_i^{k^*}, \Lambda_{ii}^{k^*})$.

To effectively search the pose space we exponentially grow the number of particles as we get further from the root. For each particle $\mathbf{x}_j$ we draw N child samples $[\mathbf{x}_i^l]_{l=1}^N$ so that we search a larger space with more samples for less constrained limbs. As very few particles are needed to describe the prior for nodes near to the root this exponential growth is not problematic, we set $N = 8$, which will result in 4096 samples being generated for each of the wrists.

3. Pose Estimation

Typically an articulated object can be written as a graph where the set of n hidden nodes $v_i \in \mathcal{V}$ represent the set of parts used to represent the object and $\{v_i, v_j\} \in \mathcal{E}$ represent the edges that connect the nodes of the graph together. Given a set of proposal values for each node $X = \{x_i, .., x_n\}$ and a set of observations for each node $Z = \{z_i, .., z_n\}$ the posterior can then be calculated as

$$p(X|Z, \theta) = \prod_{\{i,j\} \in \mathcal{E}} p(x_i|x_j, \theta_{ij}) \prod_{i \in \mathcal{V}} p(z_i|x_i) \quad (1)$$

where $p(x_i|x_j, \theta_{ij})$ represents the prior and $p(z_i|x_i)$ represents the observational likelihood.

The problem in defining a model over joints as apposed to parts is that there does not exist one-to-one correspondences between joints and observations; we can not directly observe a joint only the parts to which it is connected. To accommodate this we define a set of m observable parts $p_i \in P$, where $m \neq n$. We further define $v_j \in p_i$ as being the set of joints defining the ith part and conversely $p_j \in v_i$ as being the set of parts of which the ith joint is a member. The set of observations made for the parts are defined by $Z = \{z_i, .., z_m\}$. The observational likelihood for the ith part can now be written as $p(z_i|\{x_{j \in p_i}\})$ and it becomes clear that this distribution is dependent on a number of joint positions. This is an intuitive result, for example the appearance of the forearm must be dependent on the location of both the wrist and elbow. To estimate $p(z_i|x_j)$ from $p(z_i|\{x_j, x_{k \in p_i|j}\})$ the nodes $x_{k \in p_i|j}$ can be treated as nuisance parameters and marginalized over. In practice this is cumbersome to calculate and instead the following approximations are used: If the $x_{k \in p_i|j}$ are child nodes to x_j we use the expectation of the set of particles drawn from x_j. If they are parent nodes we use the sample of $x_{k \in p_i|j}$ from which x_j was drawn. For the torso we use the expectation of the shoulder and hips since

these joints are not directly connected and do not share child/parent relationships. This method then allows an approximation of the term $p(z_i|x_j)$ to be calculated.

We further need to account for that a joint may be a member of several parts, for example the elbow defines both the upper arm and forearm. To accommodate this the likelihood terms $p(z_i|x_j)$ are combined for all parts to which that joint is a member $p_i \in v_j$ assuming the appearance likelihood for each part is conditionally independent. This can be calculated as

$$p(z_{i\in v_j}|x_j) = \prod_{i\in v_j} p(z_i|x_j). \qquad (2)$$

This again represents an intuitive result that to estimate the position of a joint you must observe all parts to which it is connected.

Maximizing the posterior is achieved using EM where a new prior is estimated at each iteration given the posterior calculated using the old prior, a new set of particles is then generated from the prior and the posterior re-estimated. Given a set of particles for the jth joint $[\mathbf{x}_j^k]_{k=1}^N$ each is assigned a weight proportional to $p(z_{i\in v_j}|x_j^k)$, which are then used to update the prior. At each iteration simulated annealing is used to ensure the distribution converges so that $w_j^k = p(z_{i\in v_j}|x_j^k)^\beta$, where β is calculated at each iteration so approximately half of the particles would be discarded if resampling were performed [4].

4. Limb Likelihoods

A part is represented by a rectangular patch and defined by the joints that it is composed from. We use two image cues, edges and color. Edge cues are exploited using a set of M overlapping HOG features [3] placed along the edges of the part. Each feature is represented as a single normalized histogram of the local image gradients at that location and they are combined such that $p(z_i|\{x_{j\in p_i}\})_{edge} = \frac{1}{M}\sum_{m=1}^M H(\theta_\perp)$, where $H(\theta_\perp)$ returns the value in the histogram bin that is perpendicular to the edge of the part, $\theta_\perp$.

Color is exploited by placing a bounding box at the location of the root node and then learning a foreground model using the pixel values within the box and a model for the background using pixels outside the box. The models are learnt using a GMM. This creates a very crude and noisy foreground probability map (see Figure 2 (a) (i-ii)). Given the location of a part its foreground likelihood is calculated as the average value of the probability map in the region encompassed by that part. This average can be efficiently approximated by creating an Integral Image (II) for the foreground probability map. The II representation allows the integral over a rectangular region to be calculated requiring just 4 memory accesses provided the

edges of the rectangle are aligned with that of the image (i.e. is axis aligned). Since the rectangular patches used to represent the model's parts may be orientated at any angle, the following approximation can be used: Firstly, the integral is calculated over the minimum axis aligned bounding box that encompasses the orientated part, denoted F_{bb}. Between the mid-point along the two longest edges of the part and the two nearest corners of the bounding box, two smaller boxes are defined over which the average probability μ_{bg} for the region they encompass can be calculated; this estimates the average probability for the background region of the bounding box. Given the area of the bounding box A_{bb} and the area of the orientated part A_{part} the average probability for the part can be approximated as $p(z_i|\{x_{j\in p_i}\})_{col} = \frac{F_{bb}-\mu_{bg}A_{bb}}{A_{part}} + \mu_{bg}$. This requires just 12 memory accesses of the integral probability map and is illustrated in Figure 1. The individual likelihoods for each cue are then combined as $p(z_i|\{x_{j\in p_i}\}) = p(z_i|\{x_{j\in p_i}\})_{edge}p(z_i|\{x_{j\in p_i}\})_{col}$.
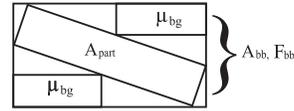

Figure 1. Approximating the average probability for an arbitrarily orientated patch, see text for details.

5. Experiments

The prior was learnt using 3D motion capture data from all subjects and actions contained in the HumanEva dataset. 200 images were then randomly selected across all subjects, actions and camera views (using color cameras only) as a testing set. In each image the root node, which corresponded to the pelvis, was initialized using the ground truth data and the scale set as the difference between the feet and the head. The algorithm was then iterated ten times, under the current Matlab implementation each iteration requires just 7 seconds of processing. The focal length is assumed to be unknown and an orthographic projection is used. The errors are presented as the average difference between the extracted pose and ground truth, both 3D and 2D errors are presented in Table 1. Results are shown using different cues and as can be seen both the 2D and 3D errors improve when color cues are used in conjunction with edge cues. It is difficult to compare our 3D reconstruction errors directly with those of others since we use different assumptions, for example our method uses only a single image. In comparison to discriminative methods trained using only a single action type [7] (38.0 mm) our method performs poorly. However, learning just a single action effectively constrains the
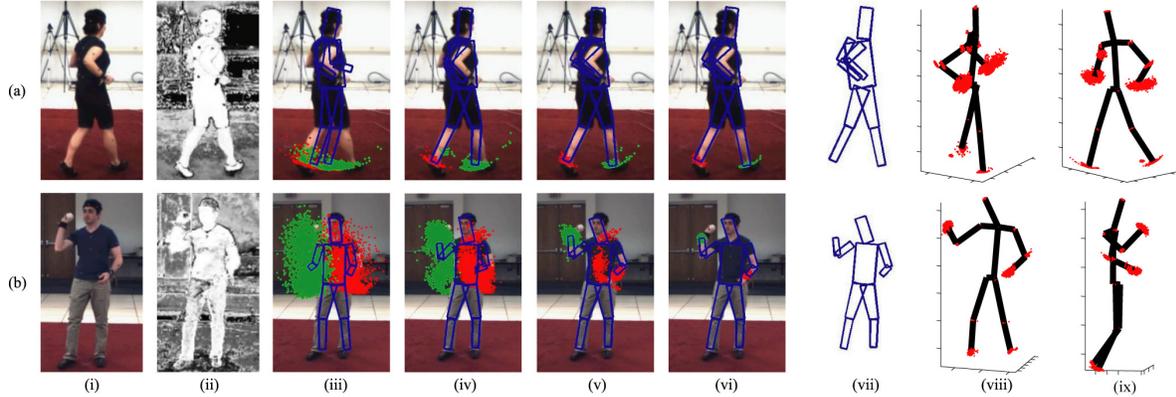
Figure 2. (i) Original image. (ii) Foreground probability map. (iii)-(vi) (a) Samples for left ankle (red) and right ankle (green); (b) Samples for left wrist (red) and right wrist (green) drawn from prior for iteration 1, 4, 6 and 10 respectively. The expected pose for each iteration is also shown. (vii) 2D expected pose as shown in (vi). (viii) & (ix) Final 3D reconstruction viewed from different orientations. Samples drawn from the final prior are also shown.

Table 1. Pose estimation errors for 2D and 3D pose estimation. 2D errors have units of $pixels$ and 3D errors mm. E - using edge cues, C - using color cues.

	2D E	2D E + C	3D E	3D E + C
Neck	15.0	13.0	128.6	114.7
Hips	2.1	2.1	12.1	12.2
Head	21.8	17.9	212.8	195.7
Shoulders	15.5	13.2	127.5	117.8
Elbows	23.3	20.4	188.3	173.2
Wrist	47.3	44.3	331.4	311.5
Knees	10.7	10.1	117.5	115.3
Feet	18.4	16.4	202.4	197.8
mean	19.4	17.4	164.3	154.7

pose space to a single dimension (i.e. gait phase) with strong correlations between parts. In comparison to a monocular tracking method without an action specific prior [9] (654 mm) we perform considerably better. However, it should be noted that whilst in [9] the full pose is manually initialized in the first frame, in the presented method the root position is effectively initialized in all frames. The results we present as perhaps is expected, fall somewhere in the middle of the two and are provided as a baseline for future work.

Figure 2 (iii)-(vi) shows the expectant pose for each iteration and as shown they converge towards the correct solution. In Figure 2 (b)(v), it clearly shows that the GMM is more than capable of representing multi-modal hypotheses as seen in the sample distribution for the left arm.

6. Conclusions

We presented a method that is capable of estimating pose from a single color image requiring minimal initialization. This has been achieved by creating strong conditional models between joints so that the prior is more representative of the initial training data.

A method of sampling from the prior has been presented and an EM approach was used to maximize the posterior. We also have shown that HOG features are suitable to represent edge cues in generative approaches and that color cues are beneficial even given uncertainty in the class of the training data used to learn initial foreground/background models. In future work we will employ a better camera model and attempt to automatically estimate orientation, scale and the subject's location.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *PAMI*, 28(1):44–58, 2006.

[2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[4] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 2005.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[6] G. Hua and Y. Wu. Variational maximum a posteriori by annealed mean field analysis. *PAMI*, 27(11):1747–1761, 2005.

[7] R. Okada and S. Soatto. Relevant feature selection for human pose estimation and localization in cluttered images. In *ECCV*, 2008.

[8] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006.

[9] L. Sigal, A. Balan, and M. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 2009.

[10] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *CVPR*, 2004.