Learning shape models for monocular human pose estimation from the Microsoft Xbox Kinect

James Charles and Mark Everingham
School of Computing
University of Leeds
{j.charles|m.everingham}@leeds.ac.uk

Abstract

We propose a method for learning shape models enabling accurate articulated human pose estimation from a single image. Where previous work has typically employed simple geometric models of human limbs e.g. *cylinders which lead to rectangular projections, we propose to learn a generative model of limb shape which can capture the wide variation in shape due to varying anatomy and pose. The model is learnt from silhouette, depth and 3D pose data provided by a Microsoft Xbox Kinect, such that no manual annotation is required. We employ the learnt model in a pictorial structure model framework and demonstrate improved pose estimation from single silhouettes compared to using conventional rectangular limb models.*

1. Introduction

We tackle the problem of 2D articulated human pose estimation from monocular images – recovering the 2D joint positions of a human skeleton from a single silhouette. This task has received much attention in the literature [1, 7, 9, 14] because of the many applications including high-level understanding of images and video and markerless motion capture. Despite the robust 3D human pose estimates achievable using consumer depth sensors such as the Microsoft Xbox Kinect [21], the task of pose estimation from monocular visible-light images in unconstrained scenarios remains important and challenging, for example to enable interpretation of consumer photographs, archive video or surveillance footage from monocular, low framerate cameras. The problem is particularly challenging because of large variation in human body shape due *e.g.* to anatomy, clothing, lighting or camera parameters and the many degrees of freedom in pose space which the human body exhibits. Using an 'impoverished' imaging modality, *i.e.* a single visible-light image or silhouette, compounds these challenges compared to the use of a depth sensor.

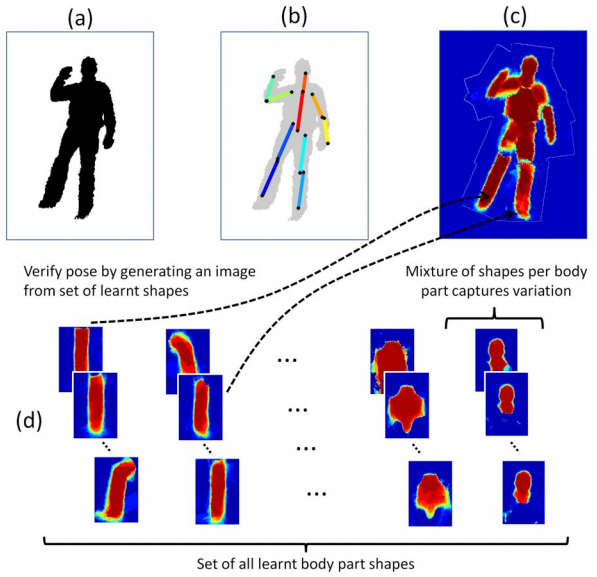


Figure 1. Outline of proposed method. From a binary silhouette (a) we infer 2D human pose (b) using a generative model of shape (c). Mixture models of probabilistic shape templates for each limb (d) are learnt from Kinect output by inferring segmentation of silhouettes into limbs (Fig. 2).

Most previous work has assumed simple fixed models of 3D human limb shape consisting of cylinders or conic sections [15, 6, 19, 8], which lead to rectangular or quadrilateral projected shapes. Such models clearly capture poorly the true shape of human limbs given wide variations in anatomy or clothing, and indeed we demonstrate that this has a negative effect on the accuracy of pose estimation using an approach based on the popular pictorial structure model (PSM) [8]. Fig. 1 outlines our proposal: using the PSM we infer 2D human pose (b) from a single silhouette (a). The pose estimate is driven by an accurate generative model of human shape (c) where the shape of each limb is modelled by a mixture of probabilistic shape templates (d) capturing variation in limb shape. As outlined in

Fig. 2, the limb shape models are learnt automatically from Kinect output by inferring segmentation of the human silhouette into limbs. We demonstrate that using learnt shape models *c.f*. simple parametric models (*e.g*. rectangles) improves the fidelity of the generative image model (Fig. 1(c)), and yields corresponding improvements in quantitative pose estimation accuracy.

Related work. Most previous work on modelling human body shape has used simple geometric primitives *e.g*. 2D rectangles [15], ellipses [22] or conic sections [6]. While such models can be fitted to approximately match an individual's anatomy *e.g*. limb length, they only crudely represent the shape of human limbs, especially when clothed.

Roberts *et al*. [20] proposed a 2D probabilistic region template to model limb shape, where each pixel of the template represents the probability that a pixel belongs to the limb versus background. The templates were learnt from 20 manually segmented images. We use a similar representation but learn a mixture of templates automatically without manual segmentation, and correctly treat the case of self-occlusion. Buehler *et al*. [5] use a set of manually-segmented templates to represent head and torso shape, combining these with a conventional rectangle model of upper/lower arms. In contrast to the related work by Felzenszwalb *et al*. [8] they do not use limb models with a "center-surround" response (requiring pixels around the limb to be background), since these work poorly in the case of self-occlusion *e.g*. arms in front of the torso, and note that this introduces much ambiguity. In our approach we adopt the same "sample and verify" approach [8, 5], but use mixture models learnt for all body parts without manual segmentation. Some recent works have used 3D scans of humans to learn variations in body shape [12, 2, 13], but are limited by lack of sufficient clothed human data and only capable of representing the unclothed human body. The 3D PCA-based model "SCAPE" [2], built from a set of 3D body scans, has been applied to pose estimation [3], estimating body shape under clothing and building part-based 2D shape models [10]. Recent work by Guan *et al*. [11] addressed the limitations of the model in explaining clothed humans, and augmented the model with a PCA model of clothing, representing deviations from the underlying body contour, and learnt from meshes of virtual clothing and clothed/unclothed images. Our proposed method is complementary in that we learn shape models applicable directly in image space *c.f*. models parameterized in shape space.

In contrast to methods using explicit models of limb shape, much recent work has adopted a discriminative approach, effectively using "sliding-window" limb/non-limb classifiers [18, 1, 14, 4]. While undoubtedly such approaches, which yield strong appearance terms, are an important component for pose estimation, use of an explicit shape model allows natural exploitation of image and motion edges, and gives a more detailed image interpretation. For instance, the output reveals a segmentation of the image into human/background [16, 11] and potentially pixel-wise segmentation of the body into different limbs [5, 17]. Whereas discriminative approaches typically yield only a point estimate of the limb center and orientation.

Outline. Sec. 2 reviews the PSM and defines our proposed shape model. Sec. 3 describes model learning. Sec. 4 reports experimental results, and we conclude in Sec. 5.

2. Pictorial structure model

We adopt a PSM [8] of the human body, motivated by computational efficiency and its success in previous work [19, 18, 5, 1, 14]. A graphical model is defined by ten nodes corresponding to the limbs, and edges E representing a prior distribution over the relative location and orientation of the limbs. Pose is parameterized by $Z = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{10}\}$, where $\boldsymbol{z}_l = (x_l, y_l, \theta_l)$ represents the 2D location and orientation of a limb. The posterior probability of a pose given a binary silhouette image B is defined as

$$P(Z|B,\Theta) \propto p(B|Z,\Theta)p(Z|\Theta) \tag{1}$$

where Θ represents the model parameters to be learnt. Given the tree structure of the PSM and the assumption of independence between the shape of each limb, which enable efficient sampling and inference [8], the posterior is factorized into unary shape terms and binary prior terms:

$$p(B|Z,\Theta)p(Z|\Theta) = \prod_i p(\boldsymbol{b}_i|\boldsymbol{z}_i,\Theta_i) \prod_{(z_i,z_j)\in E} p(\boldsymbol{z}_i|\boldsymbol{z}_j,\Theta_{ij}) \tag{2}$$

where $\boldsymbol{b}_i$ defines the region of B covered by part i. The prior terms $p(\boldsymbol{z}_i|\boldsymbol{z}_j,\Theta_{ij})$ are modelled as Gaussian enabling efficient inference using distance transforms [8]. Sec. 2.1 discusses shape terms $p(\boldsymbol{b}_i|\boldsymbol{z}_i,\Theta_i)$ which capture the agreement between a pose hypothesis and the underlying image.

Sample and verify approach. While the posterior probability in Eqn. 1 can be maximized efficiently using dynamic programming methods [8], this can yield poor pose estimates because of limitations in the PSM: (i) by assuming independent shape terms the PSM has no notion of self-occlusion; (ii) the PSM only considers the image region covered by limbs (Eqn. 2), not the entire image. Together this can lead to pose estimates where (a) multiple limbs are placed at the same location; (b) limbs are missed entirely. These limitations can be overcome effectively by using a two step "sample and verify" approach [8, 5] which we adopt here: (i) sample poses from the PSM posterior distribution (Eqn. 1); (ii) verify these samples by scoring the

pose using an extended model which treats self-occlusion correctly and considers the entire image. This approach retains the efficiency of the PSM for sampling, while allowing complete freedom in the choice of verification model.

2.1. Shape modelling

As shown in Fig. 1(d) we learn models of limb shape in the form of mixture models over templates where pixels in a template represent the probability of a pixel being foreground *vs.* background. This is similar to previous work using rectangle templates [5] which only model foreground pixels, or center-surround templates [8] which assume all pixels around a rectangular limb should be background, except that we *learn* the appropriate probabilistic shapes.

In the "sample and verify" approach described above, the limb shape models play two roles: (i) they are used to define the shape terms in Eqn. 2 for sampling likely poses, assuming shape independence between each limb; (ii) they are used ro "render" an entire hypothesized silhouette (Fig. 1(c)) with which to verify a pose sample, taking into account self-occlusion. Because of these two roles it is appropriate to learn two connected shape models per limb, assuming independence for sampling, or no independence for verification. We describe these models here.

Model for sampling. A mixture of K probabilistic masks for limb i is defined as $\boldsymbol{\tau}_i = \{\boldsymbol{\tau}_i^1, \ldots, \boldsymbol{\tau}_i^K\}$, where each element τ_{ij}^k of $\boldsymbol{\tau}_i^k$ represents the probability of a pixel j being foreground. Masks are defined in a canonical coordinate system relative to the principal axis of the limb (see Fig. 2). To evaluate the likelihood of a limb being at a given image location and orientation (Eqn. 2), the corresponding image silhouette region $\boldsymbol{b}_i$ is evaluated under the generative model represented by the mixture of masks by maximizing over the mixture component k:

$$p(\boldsymbol{b}_i|\boldsymbol{z}_i, \Theta_i) = \max_k P(k|i) \prod_j \left[\tau_{ij}^k\right]^{b_j} \left[1 - \tau_{ij}^k\right]^{1-b_j} \quad (3)$$

where $P(k|i)$ is the prior probability for mixture component k and b_j is the jth pixel in region $\boldsymbol{b}_i$. To simplify notation throughout the paper we assume that the masks for each limb i have undergone a similarity transformation T_i to place them in the canonical coordinate system (see Fig. 2). Note that the pixels b_j in the input image region simply act as "switch" variables which select whether a pixel should be assigned the corresponding foreground τ_{ij}^k or background $(1 - \tau_{ij}^k)$ probability defined by the mask.

Model for verification. Given a pose sampled from the PSM, it is verified by a generative model of the *entire* human silhouette which overcomes the PSM problems of self-occlusion and missed evidence. A probabilistic mask for the entire human is generated by rendering masks $\boldsymbol{\mu}_i^k$ for each limb such that each pixel is assigned the maximum probability of being foreground under any limb at that position. The likelihood of the observed silhouette B is defined thus:

$$p(B|Z, \Theta) = \max_{i,k} \prod_j \left[\mu_{ij}^k\right]^{B_j} \left[1 - \mu_{ij}^k\right]^{(1-B_j)} \quad (4)$$

where masks $\boldsymbol{\mu}_i^k$ have the same form as those used for sampling, but because here there is no assumption of independence between limbs they should be learnt taking into account self-occlusion (Sec. 3). Note that this likelihood is defined in a *pixel-centric* as opposed to limb-centric manner, such that self-occlusion is treated correctly (pixels can only given evidence for a single limb) and all pixels contribute evidence, not only those lying in predicted limb regions.

Note that the maximization in Eqn. 4 in terms of the mixture component used for each limb is strictly intractable. In practice this is achieved by sampling mixture components

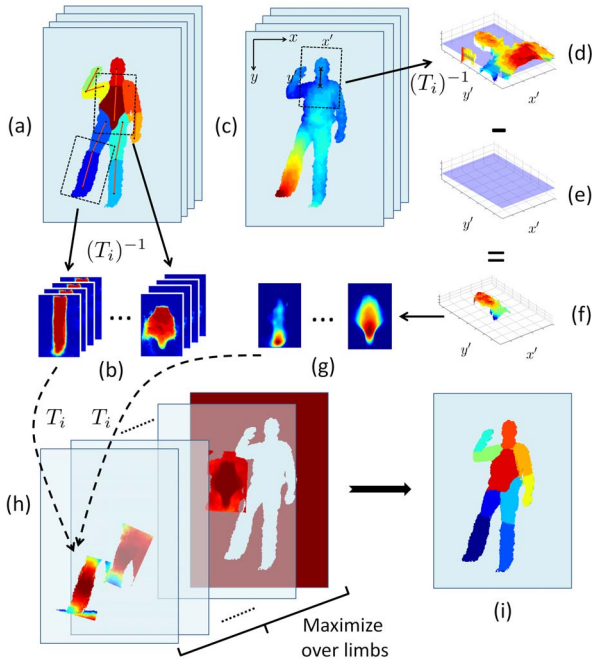


Figure 2. Learning limb shape models. An initial segmentation of the silhouette is estimated (a) using known joint positions. Limb segments are transformed into canonical form and used to construct a mixture of probability masks per limb (b). Limb depth values from the depth image (c) are transformed into canonical form (d) and an offset plane (e) is subtracted to form a depth image per limb (f). Averaging over limb depth images forms a depth profile per limb (g). The shape models (b) and depth profiles (g) are transformed and constrained to foreground/background regions using the input silhouette (h). Re-segmentation is achieved by maximizing over limb labels for each pixel (i) while accounting for self-occlusion. The process is repeated until convergence.

from the PSM which are likely to yield high corresponding verification likelihoods.

3. Model learning

We now describe the procedure for learning the limb shape masks. We first cover the case of learning models for the verification stage, since this requires correct handling of self-occlusion, then turn to the simpler case of the models for sampling. Fig. 2 outlines the learning procedure. Training data is provided by the Kinect, in the form of silhouettes, 3D joint positions and depth images. Note that the Kinect does not provide a segmentation of the image into *limbs* – this must be inferred during the learning process. Although our learnt models are applied to 2D imagery we exploit the depth image provided by the Kinect at the learning stage.

A generative model of the depth image D (Fig. 2(c)) is used as a driving force for learning limb shape models. For each limb i a depth profile $\boldsymbol{d}_i$ (Fig. 2(g)) is learnt alongside the shape models $\boldsymbol{\mu}_i^k$ (Fig. 2(b)) and used to assist with inferring a correct segmentation of the limb, and resolving self-occlusion ambiguities. The silhouette image is used to constrain limb segmentations such that pixels inside/outside the silhouette are assigned to be limb/background respectively. The observed depth values of pixels are assigned likelihoods based on the segmentation label (limb) and corresponding depth profile and shape model. Segmentation of the image into limb regions (Fig. 2(i)) is defined by a set of binary latent variables $S = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{10}\}$, where pixels j with $s_{ij} = 1$ belong to limb i and pixel assignments to limbs are mutually exclusive. The depth image likelihood can then be written as:

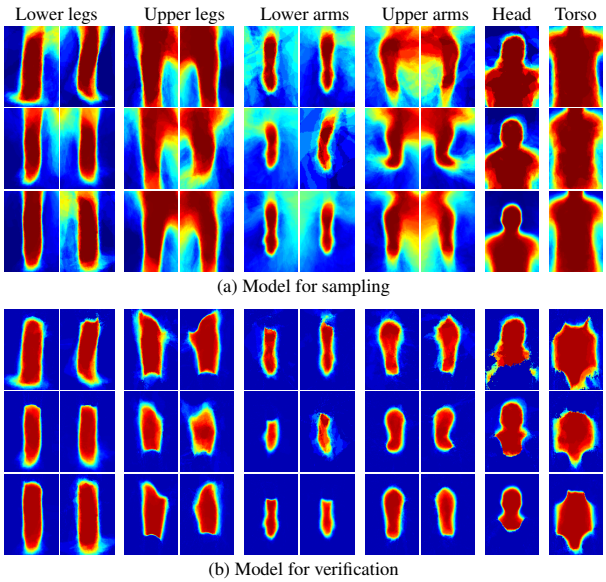
$$P(D|S,\Phi) = \prod_j \sum_i P(D_j|d_{ij}) \sum_{k=1}^{K} P(s_{ij}|\mu_{ij}^k) P(k|i) \tag{5}$$

where Φ collects the parameters to be learnt for each limb $\Phi_i = \{\boldsymbol{d}_i, \boldsymbol{\mu}_i\}$ corresponding to the depth profile (Fig. 2(g)) and shape masks (Fig. 2(b)) respectively. Only one limb's depth profile should explain a pixel, and this is enforced by using s_{ij} to select which depth profile in Eqn. 5 to use within the sum over limbs i. Parameters are learnt by maximizing the product of Eqn. 5 over all training images using an alternation strategy described below.

Shape term. The term $P(s_{ij}|\mu_{ij}^k)$ models the likelihood of labelling a given pixel as one of ten limbs dependent on the pixel location within a shape model. The segmentation likelihood is defined as

$$P(s_{ij}|\mu_{ij}^k) = \left[\mu_{ij}^k\right]^{s_{ij}} \left[1 - \mu_{ij}^k\right]^{(1-s_{ij})} \tag{6}$$

where s_{ij} is used to select either foreground probabilities μ_{ij}^k or background probabilities $(1 - \mu_{ij}^k)$.



(a) Model for sampling

(b) Model for verification

Figure 3. Learnt shape models. Each row represents a single mixture component. Note that the sampling model (a), which assumes independence between limbs, has substantial differences to the verification model (b), capturing the likelihood of neighboring limbs being present.

Depth term. The term $P(D_j|d_{ij})$ models the likelihood of the observed depth value given the expected depth value provided by the latent depth profiles. We assume a Gaussian model for noise in the depth image and correspondingly define the depth likelihood as

$$P(D_j|d_{ij}) \propto \exp -\frac{1}{2\sigma^2}\left[d_{ij} + o_{ij} - D_j\right]^2 \tag{7}$$

where o_{ij} is the offset for limb i which maps depth values relative to joint location to depth values relative to the origin of the depth camera. The offset values form a plane fitted through the joint coordinates (Fig. 2(d)–(f)).

Learning model for verification. Using an alternation strategy for learning, we update segmentations while keeping model parameters fixed then update model parameters while keeping segmentations fixed. This is repeated until convergence (Fig. 2). Initially we assign training images to mixture components by clustering the 3D poses using K-means. Segmentations are estimated by assigning silhouette pixels to the nearest skeleton edge by Voronoi tessellation. Limb masks are initialized by averaging across limb segmentations for each mixture assignment (Fig. 2(a)–(b)) and depth profiles are computed by averaging across all profiles per limb (Fig. 2(g)). All model parameters are held fixed and Eqn. 5 is maximized by updating limb segmentations while constraining pixels outside the silhouette to be background and pixels inside to be foreground (Fig. 2(h)). Fixing segmentations, we then update the models by averaging

Table 1. Pose estimation accuracy per part (in percentages) scored at 50% slack value [9]. Results are shown for method using rectangular limbs and the proposed method using learnt shapes with 1–4 mixture components.

Shape model	Upper arms	Lower arms	Upper legs	Lower legs	Head	Torso	Total ± sd
Rectangles	69.5	66.8	85.6	79.2	66.4	92.5	76.1 ± 0.2
Learnt 1 Mix	83.9	69.9	90.5	85.4	96.0	98.7	85.4 ± 0.2
Learnt 2 Mix	**84.3**	66.2	90.7	84.7	96.5	98.7	84.7 ± 0.2
Learnt 3 Mix	82.7	66.8	90.7	84.3	96.5	**99.6**	84.5 ± 0.2
Learnt 4 Mix	84.1	**72.1**	**91.4**	**85.8**	**96.5**	98.2	**86.2 ± 0.2**

segmentations and depth values, treating self-occlusion correctly by using only those limb pixels not occluded by another limb *i.e. iff* $s_{ij} = 0$ and $s_{lj} = 1$ for some limb l with lower depth order than i then we know limb i is occluded by limb j. Depth order can be determined by sorting limbs according to mean depth per limb segmentation. This process is repeated until no change in segmentation. The final mixture component assignment C is used to learn the model for sampling, as described below.

Learning model for sampling. Mixture component assignments are tied between the shape model used for sampling and the shape model used for verification so that we can sample mixture components from the PSM and use them for maximizing Eqn. 4. The masks $\boldsymbol{\tau}_i^k$ (Eqn. 3) are computed using the silhouette regions $\boldsymbol{b}_i$ covered by part i. Using the set of training silhouettes we assign each region a mixture component according to C learnt earlier. Because the model for sampling assumes limbs are independent, the masks can be computed simply by averaging across training silhouette regions for each limb i and mixture component k.

4. Experimental results

A dataset of 226 diverse poses was collected using the Kinect and the PrimeSense OpenNI drivers. We used "static" poses to avoid inaccurate joint positions produced by lag in the PrimeSense skeleton tracker. Our method was tested using 5-fold cross-validation. Training uses the silhouette, depth image and joint positions output by Kinect. Testing uses only the color image output by Kinect – the silhouette is estimated by background subtraction, to replicate conditions using a single visible-light camera.

Learnt models. Fig. 3 shows an example of learnt limb shape models, for a mixture model with three components. The different mixture components approximate variations in scale and configuration of the limbs. As seen, the limb shapes learnt for verification naturally represent jigsaw-like body part pieces which can be neatly joined together at joint locations to form a full body silhouette. The models learnt for sampling, which assume independence between limbs, differ in that they capture the likelihood of neighboring limbs being present *e.g.* the "upper arm" model captures

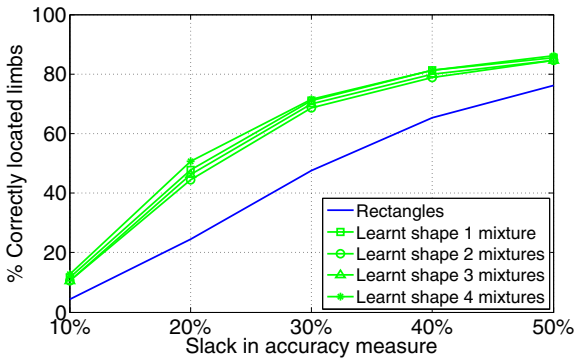


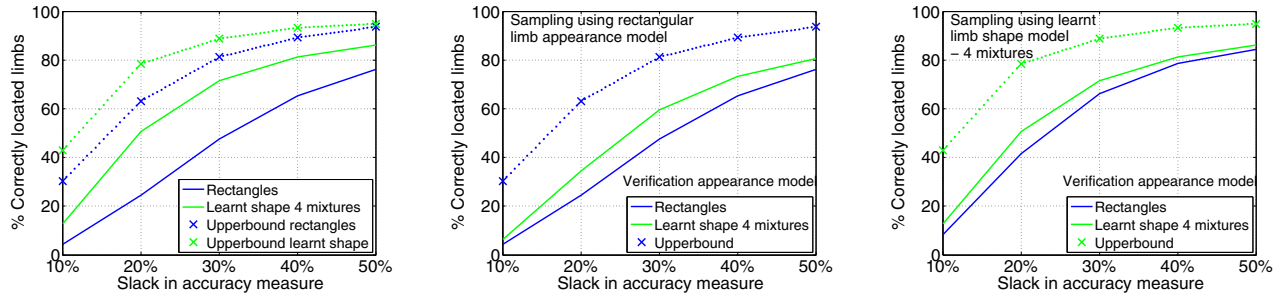
Figure 4. Results for pose estimation precision and use of learnt *vs.* rectangular models. Learnt shapes produce better pose estimates and increased precision.

that there is torso and lower arm in the neighboring region. Note that this differs significantly to center-surround models used in some previous work [8] which inaccurately assume that limbs are surrounded by background.

Quantitative comparison. We adopt a method proposed by Ferrari *et al.* [9] for quantitative evaluation of pose estimation accuracy: for a given pose a limb is considered correctly located if its endpoints are within $x\%$ of the part length from the corresponding ground truth points. The "slack" value x (50% in [9]) controls the accuracy of pose estimate required to be considered correct.

We compare our method using learnt shape models with 1–4 mixture components per limb to the use of rectangular center-surround limb models [8]. For both models the same "sample and verify" approach is taken, drawing 5,000 pose samples from the PSM. For the rectangular limb models, samples are verified by rendering the corresponding silhouettes and picking the sample which minimizes chamfer distance to the input silhouette [8]; for the proposed method the sample is chosen which maximizes Eqn. 4.

Tbl. 1 shows mean accuracy per body part and mean/standard deviation of overall accuracy for each model, using a slack value of 50% [9]. Using learnt shape models gives a substantial improvement over rectangular models, with four mixture components proving best, improving the overall accuracy from 76.1% (rectangles) to 86.2% (learnt shape), a relative improvement of 13%. This improvement is statistically significant (sd of 0.2% for both models). Increasing the number of mixture components beyond four did not improve results further but we would expect this situation to change with increasing dataset size. The accuracy improves for all limbs, but the increase is very modest for the lower arms; this is likely due to high variation in lower arm pose which has low probability under the PSM prior, and their small size which can contribute little to the verification likelihood, discussed later. Failure modes



(a) Influence of sampling model (b) Verification of rectangle samples (c) Verification of learnt samples

Figure 5. Evaluation of using learnt *vs*. rectangular limb models for sampling and verification. (a) learnt shapes produce better samples as indicated by a higher upperbound; (b)/(c) learnt shapes improve verification, approaching the upper-bound from the sampling stage. See text for discussion.

of the approach are largely cases of extreme self-occlusion and poses from behind where the left-right ambiguity is not resolvable from silhouette input.

Fig. 6 shows example results obtained with the rectangle models and learnt shape model with four mixture components, with the rendered shape (red) and estimated pose (green) overlaid on the input silhouette (blue outline). As shown, the learnt shape models give a much more faithful reproduction of the image silhouette, which yields more accurate pose estimates. This is particularly noticeable for the cases where the legs are bent and the limbs foreshortened.

Precision of pose estimates. The precision of the different models was analyzed by varying the slack threshold in the range 10–50%, requiring more precise limb locations for a pose to be considered correct. Fig. 4 plots percentage of correct poses as a function of the slack threshold. The results show that using learnt shape models particularly improves the pose estimates at higher precisions (lower slack thresholds), where the rectangle model gives only an approximate pose estimate. Results do not vary greatly with the number of mixture components, although this is likely to change with a larger dataset.

Influence of shape models on sampling *vs*. verification. Obtaining accurate pose estimates with the "sample and verify" approach relies on two factors: (i) an accurate sample must be generated by the PSM; (ii) the verification method must score such a sample above less accurate samples. We evaluated the influence of using learnt *vs*. rectangular limb models on these two aspects. First we evaluate the influence on the sampling stage by computing an upper-bound on the pose accuracy achieved by selecting the sample which best matches the ground truth (effectively simulating a perfect verification stage). The results in Fig. 5(a) show that using learnt shape models improves the upper-bound over the rectangle model, especially for high precision (low slack).

Fig. 5(b)/(c) show results when the model used for sampling is fixed and the verification model changed. As shown, when using either rectangular or learnt shapes for sampling, use of learnt shapes improves the verification stage such that the selected samples are more accurate, regardless of the sampling model, and a tighter fit to the upper-bound from the sampling stage is obtained. The gap between the upper-bound and verification using shape is largely due to misplaced lower arms, which sometimes contribute little to the verification likelihood due to their small size, something which might be overcome in future work by incorporating discriminative limb models.

5. Conclusions

We have proposed an approach for learning 2D limb shape models in the form of a mixture over probabilistic masks, and demonstrated the application to 2D articulated pose estimation from a single image silhouette. By using the Kinect to provide training data we learn shape models without any ground truth segmentation of limbs, exploiting the depth image to assist with automatic segmentation. We showed that using such learnt shape models in a PSM-based framework substantially improves the accuracy of pose estimation due to improvements in the fidelity of the models to the observed silhouettes.

In future work we aim to greatly extend our dataset to cover substantial variation in anatomy and clothing, and to investigate coupling between the mixture components between limbs to capture correlations in shape. We will also combine our models with discriminative limb detectors to further improve the pose estimation accuracy, and incorporate integrated image segmentation to remove the requirement for background segmentation. We believe that the combination of discriminative detectors with generative models such as that proposed here shows great promise in yielding detailed pixel-level interpretations of human pose in general images.

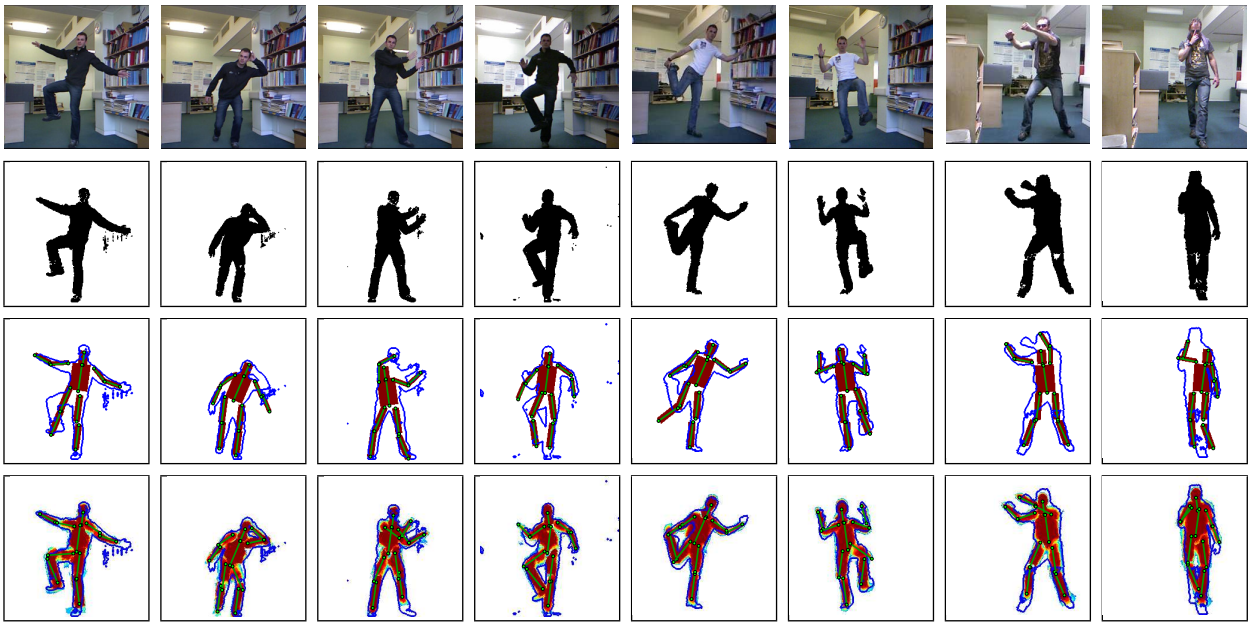


Figure 6. Example results. Rows from top to bottom show original color image, background removed silhouette image, results using rectangle models and results using learnt shape models. The estimated skeleton (green), rendered shape (red) and input silhouette (blue outline) are shown. Use of learnt shape models improves fidelity to the input silhouette, and corresponding pose accuracy.

Acknowledgements. James Charles is supported by EPSRC project EP/H035885/1. Mark Everingham is supported by an RCUK Academic Fellowship.

References

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

[2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. *ACM Transactions on Graphics*, 24(3), 2005.

[3] A. Balan, L. Sigal, M. Black, J. Davis, and H. Haussecker. Detailed human shape and pose from images. In *CVPR*, 2007.

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

[5] P. Buehler, M. Everingham, D. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *BMVC*, 2008.

[6] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *CVPR*, 2000.

[7] M. Eichner, V. Ferrari, and S. Zurich. Better appearance models for pictorial structures. In *BMVC*, 2009.

[8] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1), 2005.

[9] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[10] O. Freifeld, A. Weiss, S. Zuffi, and M. Black. Contour people: A parameterized model of 2D articulated human shape. In *CVPR*, 2010.

[11] P. Guan, O. Freifeld, and M. Black. A 2D human body model dressed in eigen clothing. In *ECCV*, 2010.

[12] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormahlen, and H. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *CVPR*, 2010.

[13] N. Hasler, T. Thormahlen, B. Rosenhahn, and H. Seidel. Learning skeletons for shape and pose. In *SIGGRAPH*, 2010.

[14] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[15] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In *FG*, 1996.

[16] Z. Lin, L. Davis, D. Doermann, and D. DeMenthon. An interactive approach to pose-assisted and appearance-based segmentation of humans. In *ICCV*, 2007.

[17] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *CVPR*, 2004.

[18] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2007.

[19] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. In *CVPR*, 2003.

[20] T. Roberts, S. McKenna, and I. Ricketts. Human pose estimation using partial configurations and probabilistic regions. *IJCV*, 73(3), 2007.

[21] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[22] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *PAMI*, 19(7), 1997.