3D Human Pose from Silhouettes by Relevance Vector Regression

Ankur Agarwal, Bill Triggs

To appear in IEEE International Conference on Computer Vision and Pattern Recognition, 2004.

3D Human Pose from Silhouettes by Relevance Vector Regression

Ankur Agarwal and Bill Triggs
GRAVIR-INRIA-CNRS, 655 avenue de l'Europe, Montbonnot 38330, France
{Ankur.Agarwal,Bill.Triggs}@inrialpes.fr, http://lear.inrialpes.fr

Abstract

We describe a learning based method for recovering 3D human body pose from single images and monocular image sequences. Our approach requires neither an explicit body model nor prior labelling of body parts in the image. Instead, it recovers pose by direct nonlinear regression against shape descriptor vectors extracted automatically from image silhouettes. For robustness against local silhouette segmentation errors, silhouette shape is encoded by histogram-of-shape-contexts descriptors. For the main regression, we evaluate both regularized least squares and Relevance Vector Machine (RVM) regressors over both linear and kernel bases. The RVM's provide much sparser regressors without compromising performance, and kernel bases give a small but worthwhile improvement in performance. For realism and good generalization with respect to viewpoints, we train the regressors on images resynthesized from real human motion capture data, and test it both quantitatively on similar independent test data, and qualitatively on a real image sequence. Mean angular errors of 6–7 degrees are obtained — a factor of 3 better than the current state of the art for the much simpler upper body problem.

1. Introduction

We consider the problem of estimating and tracking the 3D configurations of complex articulated objects from monocular images, *e.g.* for applications requiring 3D human body pose and hand gesture analysis. There are two main schools of thought on this. *Model-based approaches* presuppose an explicitly known parametric body model, and estimate the pose either by directly inverting the kinematics (which requires known image positions for each body part) [15], or by numerically optimizing some form of model-image correspondence metric over the pose variables, using a forward rendering model to predict the images (which is expensive and requires a good initialization, and the problem always has many local minima [13]). An important subcase is *model-based tracking*, which focuses on tracking the pose estimate from one time step to the next starting from a known initialization, based on an approximate dynamical model [5, 12]. In contrast, *learning based approaches* try to avoid the need for explicit initialization and accurate 3D modelling and rendering, and to capitalize on the fact that the set of *typical* human poses is far smaller than the set of kinematically possible ones, by estimating (learning) a model that directly recovers pose estimates from observable image quantities. In particular, *example based methods* explicitly store a set of training examples whose 3D poses are known, and estimate pose by searching for training image(s) similar to the given input image, and interpolating from their poses [2, 14, 9, 11].

In this paper we take a learning based approach, but instead of explicitly storing and searching for similar training examples, we use sparse Bayesian nonlinear regression to distill a large training database into a single compact model that has good generalization to unseen examples. Given the high dimensionality and intrinsic ambiguity of the monocular pose estimation problem, the selection of appropriate image features and good control of overfitting is critical for success. We are not aware of previous work on pose estimation that directly addresses these issues. Our strategy is based on the sparsification and generalization properties of our nonlinear regression algorithm, which is a form of the *Relevance Vector Machine (RVM)* [16]. RVM's have been used, *e.g.*, to build kernel regressors for 2D displacement updates in correlation-based patch tracking [18]. Human pose recovery is significantly harder — more ill-conditioned and nonlinear and much higher dimensional — but by selecting a sufficiently rich set of image descriptors, it turns out that we can still obtain enough information for successful regression. However a good descriptor set is needed: *e.g.*, the 10-D moment descriptors used in [1] are not discriminative enough for good results on full body pose.

Formally, we regress 55-D output vectors $\mathbf{y}$ representing 3D full body poses (including 3 joint angles for each of the 18 major body joints) against 100-D input vectors $\mathbf{x}$ encoding the local shapes (distribution of shape contexts) of a human image silhouette. Given a set of labelled training examples, $\{(\mathbf{x}_i, \mathbf{y}_i) \,|\, i = 1 \ldots n\}$, the RVM learns a smooth reconstruction function $\mathbf{y} = \mathbf{r}(\mathbf{x})$ valid over the region spanned by the training points. The function is a weighted linear combination $\mathbf{r}(\mathbf{x}) \equiv \sum_k \mathbf{a}_k \, \phi_k(\mathbf{x})$ of a prespecified set of scalar basis functions $\{\phi_k(\mathbf{x}) \,|\, k = 1 \ldots p\}$. The solution is regularized in the sense that the weight vectors $\mathbf{a}_k$ are well-damped, and sparse in the sense that many of them are zero. Sparsity implies that the method has selected only the *most relevant* basis functions — the ones that really need to have nonzero coefficients to complete the regression successfully. For a linear basis ($\phi_k(\mathbf{x}) = k^{th}$ component of $\mathbf{x}$), relevant input *features* (components) are selected, and for a kernel basis ($\phi_k(\mathbf{x}) \equiv K(\mathbf{x}, \mathbf{x}_k)$ for some kernel $K(\mathbf{x}, \mathbf{y})$ and centres $\mathbf{x}_k$), relevant *examples* $\mathbf{x}_k$ are selected.

Previous work: There is a good deal of prior work on human pose analysis, but relatively little on directly learning 3D pose from image measurements. Brand [4] models a dynamical manifold of human body configurations with a Hidden Markov Model and learns using entropy minimization, Sclaroff [1] learns a perceptron mapping between the appearance and parameter spaces, and Shakhnarovich *et al* [11] use an interpolated-k-nearest-neighbor learning method. Human pose is hard to ground truth, so most papers in this area [4, 1, 9] use only heuristic visual inspection to judge their results. However Shakhnarovich *et al* [11] used a human model rendering package (POSER from Curious Labs) to synthesize ground-truthed training and test images of 13 d.o.f. upper body poses with a limited (±40°) set of random torso movements and view points, obtaining RMS estimation errors of about 20° per d.o.f. In comparison, our regression algorithm estimates full body pose and orientation (54 d.o.f.) — a problem whose high dimensionality would really stretch the capacity of an example based method such as [11] — with mean errors of only 6–7°. Like [11, 6], we used POSER to synthesize a large set of training and test images from different viewpoints, but rather than using random synthetic poses, we used poses taken from real human motion capture sequences. Our results thus relate to real data. The motion capture data was taken from the public website www.ict.usc.edu/graphics/animWeb/humanoid.

Another approach is to use the image locations of the centre of each body joint as an intermediate representation, first estimating these centres, then recovering the 3D pose from them. Howe *et al* [7] develop a Bayesian learning framework to recover 3D pose from known centres, based on a training set of pose-centre pairs obtained from resynthesized motion capture data. Mori & Malik [9] estimate the centres using shape context image matching against a set of training images with pre-labelled centres, then reconstruct 3D pose using the algorithm of [15]. These works show that passing via centres can be an effective strategy, but we have preferred to estimate pose directly from the underlying local image descriptors as we feel that this is likely to prove both more accurate and more robust.

Organization: §2 describes our image descriptors, §3 our regression methods. §4 discusses RVM's feature selection properties. §5 gives our experimental results. §6 concludes.

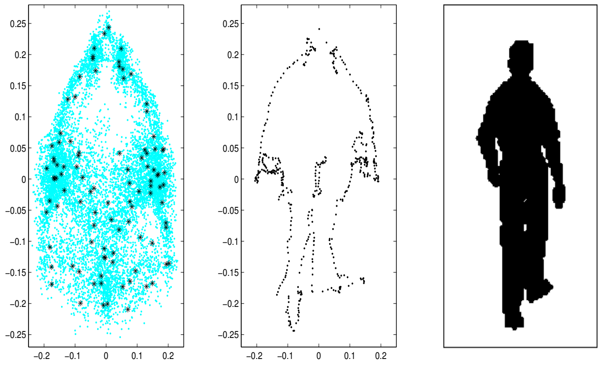

Figure 1. (Left) The first two principal components of the distribution of all shape context vectors from a training data sequence, with the k-means centres superimposed. The average-over-human-silhouettes like form arises because (besides finer distinctions) the context vectors encode approximate spatial position on the silhouette: a context at the bottom left of the silhouette receives votes only in its upper right bins, *etc*. (Centre) The same projection for the edge-points of a single silhouette (shown on the right).

2. Image Descriptors

Silhouettes: Of the many different image descriptors that could be used for human pose estimation, and in line with [4, 1], we have chosen to base our system on image silhouettes, because: *(i)* they can be extracted moderately reliably from images, at least when robust background- or motion-based segmentation is available and problems with shadows are avoided; *(ii)* they are insensitive to irrelevant surface attributes like clothing colour and texture; *(iii)* they clearly encode a great deal of useful information about 3D pose. Two factors limit the performance attainable from silhouettes:

(i) Artifacts such as shadow attachment and poor background segmentation tend to distort their local form. This often causes problems when global descriptors such as shape moments are used (as in [1, 4]), as each local error pollutes every component of the descriptor: to be robust, shape descriptors need to have good locality. Also, any representation (Fourier coefficients, *etc*) based on treating the silhouette as a continuous parametrized curve is unacceptable: silhouettes frequently change topology (*e.g.* when a hand's silhouette touches the torso's one), so any curve-based parametrization necessarily has discontinuities w.r.t shape.

(ii) Silhouettes make several discrete and continuous degrees of freedom invisible or poorly visible. It is difficult to tell frontal views from back ones, whether a person seen from the side is stepping with the left leg or the right one, and what are the exact poses of arms or hands that fall within (are "occluded" by) the torso's silhouette. We expect that including interior edge information within the silhouette [11] would provide a useful degree of disambiguation, but we have not yet tested this.

Shape Context Distributions: Histogramming edge information is a good way to encode local shape robustly [8, 3]. Here, we use shape contexts (histograms of local edge pixels into log-polar bins [3]) to encode silhouette shape quasi-locally over a range of scales. Shape contexts w.r.t. silhouette edges are evaluated at regularly spaced points along the silhouette edge[1]. The silhouette shape is thus encoded as a distribution (in fact, as a noisy multibranched curve, but we treat it as a distribution) in the 60-D shape context space (12 angular × 5 radial bins). Matching silhouettes is reduced to matching shape context distributions. To implement this, we reduce the distributions of all points on each silhouette to 100-D histograms by vector quantizing the shape context space. Silhouette comparison is thus finally reduced to a comparison of 100-D histograms. The 100 centre codebook is learned by running k-means on the combined context vectors of all training silhouettes (see fig. 1), but other centre selection methods give similar results. Histograms are built by allowing context vectors to vote softly into the few centres nearest to them, with Gaussian weights. This softening reduces the effects of spatial quantization, allowing us to compare histograms using simple Euclidean distance[2] rather than, say, Earth Movers Distance [10]. This histogram-of-shape-contexts scheme gives us some degree of robustness to occlusions and local silhouette segmentation failures.

3. Pose Regression

This section describes the regression methods that we have evaluated for recovering 3D human body pose from the above image descriptors. Poses are represented by real vectors $\mathbf{y} \in \mathbb{R}^m$. In our case (for a full body model) these are 55-dimensional vectors including 3 joint angles for each of the 18 major body joints[3]. This is not a minimal representation of the true human pose degrees of freedom, but it corresponds to our motion capture based training data, and our regression methods handle such redundant output representations without problems. We regress poses against image descriptor vectors $\mathbf{x} \in \mathbb{R}^d$, which in our case represent probability densities of silhouette points in shape context space, vector quantized to 100-D histograms.

We assume that the relationship between $\mathbf{x}$ and $\mathbf{y}$ — which a priori, given the ambiguities of pose recovery, might be multi-valued and hence non-functional — can be approximated functionally as a linear combination over a prespecified set of basis functions:

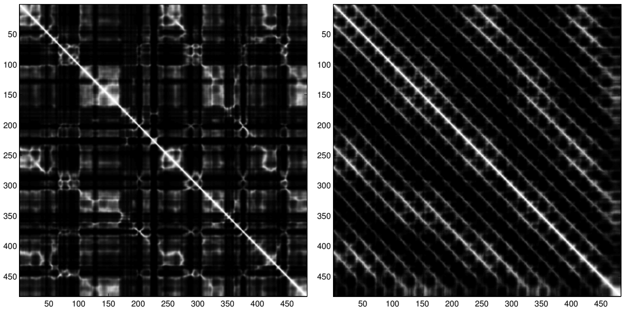

Figure 2. Pairwise similarity matrices for (left) image silhouette descriptors and (right) true 3D poses, for a 483-frame sequence of a person walking in a decreasing spiral. The light off-diagonal bands that are visible in both matrices denote regions of comparative similarity linking corresponding poses on different cycles of the spiral. This indicates that our silhouette descriptors do indeed capture a significant amount of pose information. (The light SW-NE ripples in the 3D pose matrix just indicate that the standing-like poses at the middle of each stride have mid-range joint values, and hence are closer on average to other poses than the 'stepping' ones at the end of strides).

$$\mathbf{y} \;=\; \mathbf{A}\,\mathbf{f}(\mathbf{x}) + \boldsymbol{\epsilon} \;\equiv\; \sum_{k=1}^{p} \mathbf{a}_k\,\phi_k(\mathbf{x}) + \boldsymbol{\epsilon} \qquad (1)$$

Here, $\{\phi_k(\mathbf{x})\,|\,k = 1\ldots p\}$ are the basis functions, $\mathbf{a}_k$ are $\mathbb{R}^m$-valued weight vectors, and $\boldsymbol{\epsilon}$ is a residual error vector. For compactness, we gather the weight vectors into an $m\times p$ weight matrix $\mathbf{A} \equiv (\mathbf{a}_1\;\mathbf{a}_2\;\cdots\;\mathbf{a}_p)$ and the basis functions into a $\mathbb{R}^p$-valued function[4] $\mathbf{f}(\mathbf{x}) = (\phi_1(\mathbf{x})\;\;\phi_2(\mathbf{x})\;\cdots\;\phi_p(\mathbf{x}))^\top$.

To train the model (estimate $\mathbf{A}$), we are given a set of training pairs $\{(\mathbf{y}_i, \mathbf{x}_i)\,|\,i = 1\ldots n\}$ (in our case, 3D poses and the corresponding image silhouette descriptors). All of the regressors that we test here use the Euclidean norm to measure $\mathbf{y}$-space prediction errors, so the estimation problem always takes the form:

$$\mathbf{A} \;:=\; \arg\min_{\mathbf{A}} \left\{ \sum_{i=1}^{n} \|\mathbf{A}\,\mathbf{f}(\mathbf{x}_i) - \mathbf{y}_i\|^2 \;+\; R(\mathbf{A}) \right\} \qquad (2)$$

where $R(-)$ is a regularizer on $\mathbf{A}$. Gathering the training points into an $m\times n$ output matrix $\mathbf{Y} \equiv (\mathbf{y}_1\;\mathbf{y}_2\;\cdots\;\mathbf{y}_n)$ and a $p\times n$ feature matrix $\mathbf{F} \equiv (\mathbf{f}(\mathbf{x}_1)\;\;\mathbf{f}(\mathbf{x}_2)\;\cdots\;\mathbf{f}(\mathbf{x}_n))$, the estimation problem takes the form:

$$\mathbf{A} \;:=\; \arg\min_{\mathbf{A}} \left\{ \|\mathbf{A}\,\mathbf{F} - \mathbf{Y}\|^2 + R(\mathbf{A}) \right\} \qquad (3)$$

Note that the dependence on $\{\phi_k(-)\}$ and $\{\mathbf{x}_i\}$ is encoded entirely in the numerical matrix $\mathbf{F}$.

[1] The scene vertical is always preserved in our application, so it turns out to be more discriminant to preserve the vertical, *i.e.* not to normalize contexts with respect to their dominant local orientations.

[2] We have also tested the normalized cellwise distance $\|\sqrt{\mathbf{p}_1} - \sqrt{\mathbf{p}_2}\|^2$, with very similar results.

[3] The subject's overall azimuth (compass heading angle) θ can wrap around through 360°. We maintain continuity by regressing $(a, b) = (\cos\theta, \sin\theta)$ rather than θ, using $\mathrm{atan2}(b, a)$ to recover θ from the not-necessarily-normalized vector returned by regression. $55 = 3\times 18 + 1$.

[4] To allow for a constant offset $\mathbf{A}\mathbf{f} + \mathbf{b}$, we can include $\phi(\mathbf{x}) \equiv 1$ in $\mathbf{f}$.

3.1 Damped Least Squares Regression

Pose estimation is a high dimensional and intrinsically ill-conditioned problem, so simple least squares estimation — setting $R(\mathbf{A}) \equiv \mathbf{0}$ and solving for $\mathbf{A}$ in least squares — typically produces severe overfitting and hence poor generalization. To reduce this, we need to add a smoothness constraint on the learned mapping, for example by including a damping or regularization term $R(\mathbf{A})$ that penalizes large values in the coefficient matrix $\mathbf{A}$. First consider the simplest choice, $R(\mathbf{A}) \equiv \lambda\,\|\mathbf{A}\|^2$, where λ is a regularization parameter. This gives the *damped least squares* regressor, which minimizes

$$\|\mathbf{A}\,\tilde{\mathbf{F}} - \tilde{\mathbf{Y}}\|^2 \;:=\; \|\mathbf{A}\,\mathbf{F} - \mathbf{Y}\|^2 + \lambda\,\|\mathbf{A}\|^2 \qquad (4)$$

where $\tilde{\mathbf{F}} \equiv (\mathbf{F} \;\; \lambda\,\mathbf{I})$ and $\tilde{\mathbf{Y}} \equiv (\mathbf{Y} \;\; \mathbf{0})$. The solution can be obtained by solving the linear system $\mathbf{A}\,\tilde{\mathbf{F}} = \tilde{\mathbf{Y}}$ (*i.e.* $\tilde{\mathbf{F}}^\top\,\mathbf{A}^\top = \tilde{\mathbf{Y}}^\top$) for $\mathbf{A}$ in least squares, using QR decomposition or the normal equations. λ must be set large enough to control ill-conditioning and overfitting, but not so large as to cause overdamping (forcing $\mathbf{A}$ towards $\mathbf{0}$ so that the regressor systematically underestimates the solution).

3.2 Relevance Vector Regression

Relevance Vector Machines (RVM's) [16, 17] are a sparse Bayesian approach to classification and regression. They introduce Gaussian priors on each parameter or group of parameters, each prior being controlled by its own individual scale hyperparameter. Integrating out the hyperpriors (which can be done analytically) gives singular, highly nonconvex total priors of the form $p(a) \sim \|a\|^{-\nu}$ for each parameter or parameter group a, where ν is a hyperprior parameter. Taking log likelihoods gives an equivalent regularization penalty of the form $R(a) = \nu \log \|a\|$. Note the effect of this penalty. If $\|a\|$ is large, the 'regularizing force' $dR/da \sim \nu/\|a\|$ is small so the prior has little effect on a. But the smaller $\|a\|$ becomes, the greater the regularizing force becomes. At a certain point, the data term no longer suffices to hold the parameter at a nonzero value against this force, and the parameter rapidly converges to zero. Hence: *(i)* The fitted model is sparse — the RVM automatically selects a subset of 'relevant' basis functions that suffices to describe the problem. *(ii)* The regularizing effect is invariant to rescalings of $\mathbf{f}()$ or $\mathbf{Y}$. (*E.g.* scaling $\mathbf{f} \to \alpha\mathbf{f}$ forces a rescaling $\mathbf{A} \to \mathbf{A}/\alpha$ with no change in residual error, so the regularization forces $1/\|a\| \propto \alpha$ track the data-term gradient $\mathbf{A}\,\mathbf{F}\,\mathbf{F}^\top \propto \alpha$ correctly). *(iii)* ν serves both as a sparsity parameter and as a scale-free regularization parameter. *(iv)* The complete RVM model is highly nonconvex with many local minima. Optimizing it is problematic because relevant parameters can easily become accidentally 'trapped' in the singularity at zero. In practice, these caveats do not prevent RVM's from giving useful results: setting ν to optimize the estimation error on a validation set, one typically finds that RVM's give sparse regressors with performance very similar to the much denser ones from analogous methods with milder priors.

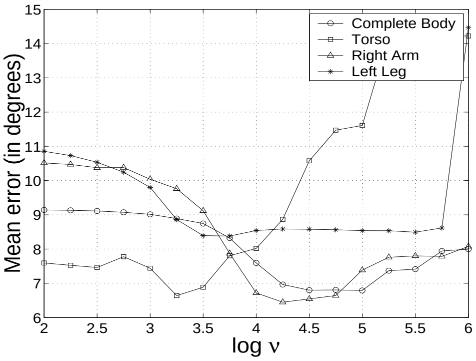

Figure 3. Mean test-set fitting error for different combinations of body parts, versus the linear RVM spareseness parameter ν. The minima indicate the optimal sparsity / regularization settings for each body part. Limb regressors are sparser than body or torso ones: whole body, 23 features; torso, 31; right arm, 10; left leg, 7.

To train our RVM's, we do not use Tipping's algorithm [16], but a continuation method based on successively approximating the $\nu \log \|a\|$ regularizers with quadratic "bridges" $\nu\,(\|a\|/a_{\text{scale}})^2$ chosen to match the prior gradient at a_{scale}, a running scale estimate for a. The bridging allows parameters to pass through zero if they need to, with less risk of premature trapping. Details are omitted for lack of space.

We have tested both *componentwise* priors $R(\mathbf{A}) = \nu \sum_{jk} \log |\mathbf{A}_{jk}|$, which effectively allow a different set of relevant basis functions to be selected for each dimension of $\mathbf{y}$, and *columnwise* ones $R(\mathbf{A}) = \nu \sum_k \log \|\mathbf{a}_k\|$ where $\mathbf{a}_k$ is the k^{th} column of $\mathbf{A}$, which select a common set of relevant basis functions for all components of $\mathbf{y}$. Both priors give similar results. The experiments shown use columnwise priors, as one of the main advantages of sparsity is in reducing the number of basis functions (support features or examples) that need to be evaluated.

3.3 Choice of Basis

We tested two kinds of regression bases $\mathbf{f}(\mathbf{x})$. *(i) Linear bases* $\mathbf{f}(\mathbf{x}) \equiv \mathbf{x}$ simply return the input vector, so the regressor is linear in $\mathbf{x}$ and the RVM selects relevant *features* (components of $\mathbf{x}$). *(ii) Kernel bases* $\mathbf{f}(\mathbf{x}) = (K(\mathbf{x}, \mathbf{x}_1) \;\cdots\; K(\mathbf{x}, \mathbf{x}_n))^\top$ are based on a kernel function $K(\mathbf{x}, \mathbf{y})$ instantiated at training examples $\mathbf{x}_i$, so the RVM effectively selects relevant *examples*. Our experiments with various kernels and combinations of kernels and linear functions show that kernelization gives a slight improvement in performance — about 0.8° per body angle, out of a total mean error of around 7°. The form and parameters of the kernel have remarkably little influence on the results. The exper-

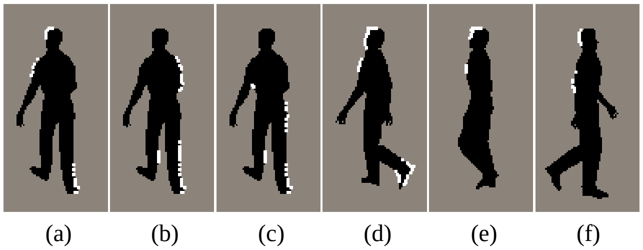

Figure 4. Silhouette points whose context bins are retained by the RVM for regression on (a) torso & neck angles (b) left arm angles, and (c) right leg angles; shown on a sample silhouette. (d,e,f): Silhouette points encoding torso parameter values over different view points and poses. On average, about 10 features covering about 10% of the silhouette suffice to estimate the pose of each body part.

iments shown use a Gaussian kernel $K(\mathbf{x}, \mathbf{x}_i) = e^{-\beta\|\mathbf{x}-\mathbf{x}_i\|^2}$ with β estimated from the scatter matrix of the training data, but other β values give very similar results.

Damped Least Squares and Relevance Vector Regression give very similar test-set errors, but the RVM regressors are much sparser. For example, in our best whole-body method to date, the final RVM selects just 156 of the 2636 training points as basis kernels, to give a mean test-set error of $6.0°$.

4. Implicit Feature Selection

Kernel based RVM regression gives reliable pose estimates while retaining only about 6% of the training examples, but working in kernel space hides information associated with individual input features (components of $\mathbf{x}$-vectors). Conversely, linear-basis RVM regression ($\mathbf{f}(\mathbf{x}) = \mathbf{x}$) provides less flexible modelling of the relationship between $\mathbf{y}$ and $\mathbf{x}$, but reveals which of the original input features encode useful pose information, as the RVM directly selects relevant components of $\mathbf{x}$.

One might expect that, *e.g.* the pose of the arms was mainly encoded by (shape-context histogram bins receiving contributions from) features on the arms, and so forth, so that the arms could be regressed from fewer features than the whole body, and could be regressed robustly even if the legs were occluded. To test this, we divided the body joints into five subsets — torso & neck, the two arms, and the two legs — and trained separate linear RVM regressors for each subset. Fig. 3 shows that similar validation-set errors are attained for each part, but the optimal regularization level is significantly smaller (there is less sparsity) for the torso than for the other parts. Fig. 4 shows the silhouette points whose contexts contribute to the features (histogram bins) that were selected as relevant, for several parts and poses. The two main observations are that the regressors are indeed sparse — only about 10 of the 100 histogram bins were classed as relevant on average, and the points contributing to these tend to be well localized in important-looking regions of the silhouette — but that there is very little spatial association between the points at which observations are made and the parts of the body being estimated. This nonlocality is somewhat surprising. It is perhaps only due to the extent to which the motions of different body segments are synchronized during natural walking motion, but if it turns out to be true for larger training sets containing less orchestrated motions, it may suggest that the localized calculations of model-based pose recovery actually miss a good deal of the information most relevant for pose.

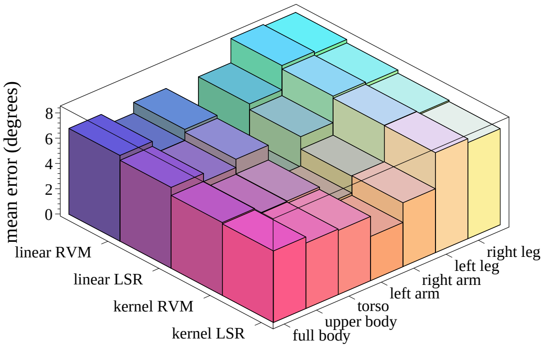

Figure 5. A summary of our various regressors' performance on different combinations of body parts for the spiral walking test sequence.

5. Overall Performance

We conducted experiments using a database of motion capture data for a 54 d.o.f. body model (3 angles for each of 18 joints, including body orientation w.r.t the camera). We report mean (over all angles) RMS (over time) absolute difference errors between the true and estimated joint angle vectors, in degrees ($m = 54$):

$$D(\mathbf{y}, \mathbf{y}') = \frac{1}{m} \sum_{i=1}^{m} |(y_i - y_i') \bmod \pm 180°| \qquad (5)$$

The training silhouettes were created by using POSER to render the motion captured poses, and reduced to 100-D histograms by vector quantizing their shape context distributions using centres selected by k-means.

Fig. 5 summarizes the final test-set performance of the various regression methods studied — kernelized and linear basis versions of regularized least squares and RVM regression, for the full body model and various subsets of it — at optimal regularizer settings. RVM regression gives very slightly higher errors than damped least squares, but much more sparsity. Kernelization brings only a small advantage over purely linear regression against our (highly nonlinear) descriptor set. The largest estimation errors occur for the leg angles, and the left arm has consistently lower error than the right one (perhaps because the subject kept it well separated from his torso).

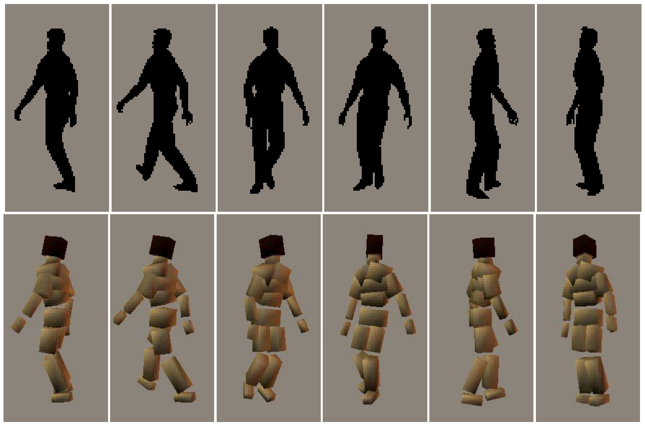

Figure 6. Some sample pose reconstructions for a spiral walking sequence not included in the training data. The reconstructions were computed with a Gaussian kernel RVM, using only 156 of the 2636 training examples. The mean angular error per d.o.f. over the whole sequence is $6.0°$.

Fig. 6 shows some sample pose estimation results, on silhouettes from a spiral-walking motion capture sequence that was not included in the training set. The mean estimation error over all joints for the Gaussian RVM in this test is $6.0°$, but the error for individual joints varies depending on the range and discernibility of each joint angle. The RMS errors obtained along with the ranges of variation (in the same test set) for some angles are as follows: body heading angle: 17° (360°), left shoulder angle: 7.5° (50.8°), and right hip angle: 4.2° (47.4°). Fig. 7 (top) plots the estimated and actual values of the overall body heading angle θ during the test sequence, showing that much of the error is due to occasional glitches. These are associated with ambiguous cases where the silhouette might easily arise from any of several possible poses. As one sign of this, recall that to allow for 360° wrap around of the heading angle θ, we actually regress $(a, b) = (\cos\theta, \sin\theta)$ rather than θ. In ambiguous cases, the regressor tends to compromise between several possible solutions, and hence returns an (a, b) vector whose norm is significantly less than one. These events are strongly correlated with large estimation errors in θ, as illustrated in fig. 7.

Fig. 8 shows reconstruction results on some real images. The reconstruction quality demonstrates the method's robustness to imperfect visual features, as a quite naive background subtraction method was used to extract somewhat imperfect body silhouettes from these images. The last example demonstrates the problem of silhouette ambiguity: the left knee is estimated to be bent instead of the right one, as the silhouette looks the same in the two cases. To reduce such errors, we plan to incorporate stronger features such as internal body edges within the silhouette, and also enforce temporal smoothness.

Note that although our results are significantly better than others presented in the literature, our pose reconstructions do still contain a significant amount of temporal jitter. This is to be expected given that each image is processed independently. It can be reduced by temporal filtering (simple smoothing or Kalman filtering), and also by adding a temporal dimension to the regressor, an avenue that we are currently exploring.

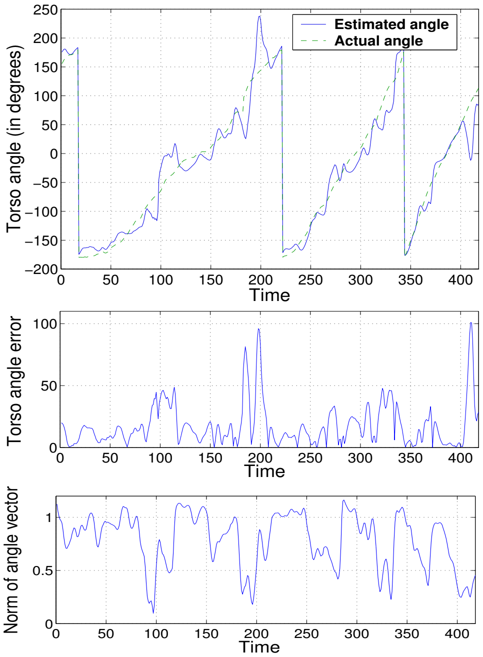

Figure 7. (Top): The estimated body heading (azimuth θ) over 418 frames of the spiral test sequence, compared with its actual value from motion capture. (Middle, Bottom): Episodes of high estimation error are strongly correlated with periods when the norm of the $(\cos\theta, \sin\theta)$ vector that was regressed to estimate θ becomes small. These occur when similar silhouettes arise from very different poses, so that the regressor is forced into outputting a compromise solution.

6. Discussion & Conclusions

We have presented a method that recovers 3D human body pose from monocular silhouettes by direct nonlinear regression of joint angles against histogram-of-shape-context silhouette shape descriptors. No 3D body model or labelling of image positions of body parts is required, making the method easily adaptable to different people or appearances. (In principle, the approach could be used to regress any set of control variables associated with any kind of image observations). The regression is done in either linear or kernel space, using either damped least squares or Relevance Vector Machines. The main advantage of RVM's is that they allow sparse sets of highly relevant features or training examples to be selected for the regression. Our kernelized RVM regressors retain only about 6% of their training examples, thus giving a large effective reduction in storage space compared

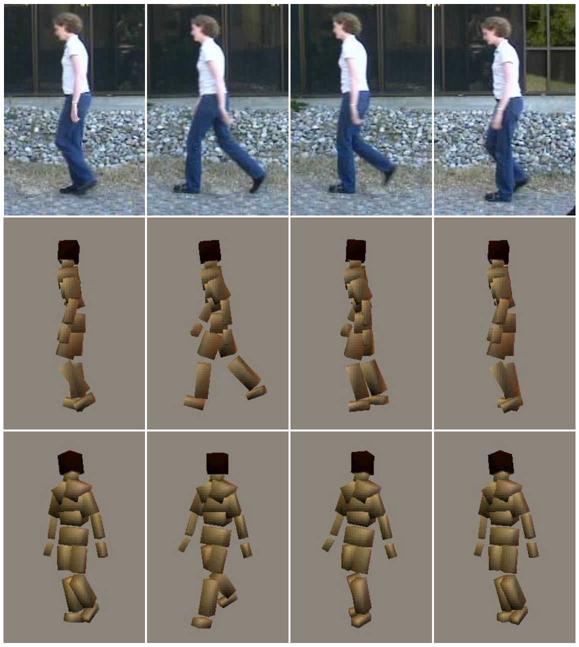

Figure 8. 3D poses reconstructed from some real test images (part of a sequence from www.nada.kth.se/∼hedvig/data.html). The middle and lower rows respectively show the estimates from the original viewpoint and from a new one. The last two columns illustrate limitations of our current system. In the third column, a noisy silhouette causes slight mis-estimation of the lower right leg, while the final column demonstrates a case of left-right ambiguity in the silhouette.

to nearest neighbour methods, which must retain the whole training database. Our methods show promising results, being about three times more accurate than the current state of the art [11].

Future work: Our linear RVM's directly select relevant features in the image descriptor space. This property may be useful for identifying better feature sets, not only for pose recovery and tracking, but also for human detection tasks.

At present, we estimate pose separately in each image. As a result, our tracking results show significant temporal jitter and some instances of incorrect poses. Temporal smoothing helps here, but we are currently investigating the effects of regressing pose $\mathbf{y}_t$ against a sequence of the last few silhouettes $(\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots)$, and of explicitly adding dynamical models. Our framework handles both of these extensions gracefully. We also intend to include richer features, such as internal edges in addition to silhouette boundaries.

Acknowledgments

This work was supported by the European Union projects VIBES and LAVA, and the research network PASCAL.

References

[1] V. Athitsos and S. Sclaroff. Inferring Body Pose without Tracking Body Parts. In *International Conference on Computer Vision and Pattern Recognition*, 2000.

[2] V. Athitsos and S. Sclaroff. Estimating 3D Hand Pose From a Cluttered Image. In *International Conference on Computer Vision and Pattern Recognition*, 2003.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(24), 2002.

[4] M. Brand. Shadow Puppetry. In *International Conference on Computer Vision*, pages 1237–1244, 1999.

[5] C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. In *International Conference on Computer Vision and Pattern Recognition*, pages 8–15, 1998.

[6] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3D Structure with a Statistical Image-Based Shape Model. In *International Conference on Computer Vision*, 2003.

[7] N. Howe, M. Leventon, and W. Freeman. Bayesian Reconstruction of 3D Human Motion from Single-Camera Video. In *Neural Information Processing Systems*, 1999.

[8] D. Lowe. Object recognition from local scale-invariant features. In *Int. Conf. Computer Vision*, pages 1150–1157, 1999.

[9] G. Mori and J. Malik. Estimating Human Body Configurations Using Shape Context Matching. In *European Conference on Computer Vision*, volume 3, pages 666–680, 2002.

[10] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. In *Int. Conf. Computer Vision*, Bombay, 1998.

[11] G. Shakhnarovich, P. Viola, and T. Darrell. Fast Pose Estimation with Parameter Sensitive Hashing. In *International Conference on Computer Vision*, 2003.

[12] H. Sidenbladh, M. Black, and L. Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In *European Conference on Computer Vision*, volume 1, 2002.

[13] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *IEEE Int. Conf. Computer Vision and Pattern Recognition*, June 2003.

[14] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Filtering Using a Tree-Based Estimator. In *International Conference on Computer Vision*, 2003.

[15] C. Taylor. Reconstruction of Articulated Objects from Point Correspondances in a Single Uncalibrated Image. In *Computer Vision and Image Understanding*, 2000.

[16] M. Tipping. The Relevance Vector Machine. In *Neural Information Processing Systems*, 2000.

[17] M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[18] O. Williams, A. Blake, and R. Cipolla. A Sparse Probabilistic Learning Algorithm for Real-Time Tracking. In *International Conference on Computer Vision*, 2003.