A local basis representation for estimating human pose from cluttered images

Ankur Agarwal, Bill Triggs

To appear in Proceedings of the 7th Asian Conference on Computer Vision, 2006.

A Local Basis Representation for Estimating Human Pose from Cluttered Images

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. Recovering the pose of a person from single images is a challenging problem. This paper discusses a bottom-up approach that uses local image features to estimate human upper body pose from single images in cluttered backgrounds. The method takes the image window with a dense grid of local gradient orientation histograms, followed by non negative matrix factorization to learn a set of bases that correspond to local features on the human body, enabling selective encoding of human-like features in the presence of background clutter. Pose is then recovered by direct regression. This approach allows us to key on gradient patterns such as shoulder contours and bent elbows that are characteristic of humans and carry important pose information, unlike current regressive methods that either use weak limb detectors or require prior segmentation to work. The system is trained on a database of images with labelled poses. We show that it estimates pose with similar performance levels to current example-based methods, but unlike them it works in the presence of natural backgrounds, without any prior segmentation.

1. Introduction

The ability to identify objects or their parts in the presence of cluttered backgrounds is critical to the success of many computer vision algorithms, but finding descriptors that can distinguish objects of interest from the background is often very difficult. We address this problem in the context of understanding human body pose from general images. Images of people are seen everywhere. A system that was capable of reliably estimating the configuration of a person's limbs from images would have applications spanning from human computer interaction to activity recognition from images to annotating video content. In this paper, we focus on recognizing upper body gestures. Human arm gestures often convey a lot of information — *e.g.* during communication — and automated inference and interpretation of these could allow for critical understanding of a person's behaviour.

Current methods for human pose inference usually rely on background subtraction to isolate the subject. This limits their applicability to fixed environments. Model-based approaches use a manual/heuristic initialization of pose as a starting point to optimize over image likelihoods, or to track through subsequent frames in a video sequence. The application of such methods to 3D pose recovery requires camera parameter estimates

and realistic human body models. We prefer to take a bottom-up approach to the problem, considering pose inference from general images in terms of two interdependent sub-problems: *(i)* identifying/localizing the human parts of interest in the image, and *(ii)* estimating 3D pose from them. We combine methods that are currently used mainly for object and pedestrian detection with recent advances in example-based pose estimation from human silhouettes or segmented images, implicitly using the knowledge contained in human body configurations to learn to localize body parts in the presence of cluttered backgrounds and to infer 3D pose.

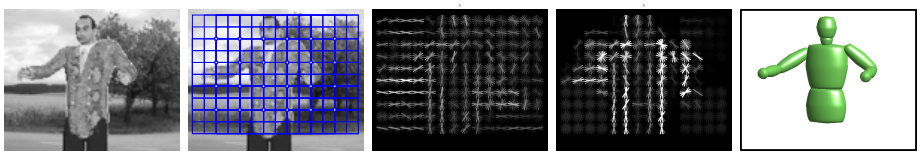
Our approach to modeling human body parts is based on using SIFT-like histograms [5] computed on a uniform grid of overlapping patches on an image to encode the image content as an array of 128-d feature vectors. This scheme encodes local image content in terms of gradient patterns invariant to illumination changes, while still retaining spatial position information. It allows us to key on gradient patterns such as head/shoulder contours or bent elbows that are characteristic of humans and that contain important pose information, in contrast to limb based representations that either key on skin colour and face detection (*e.g.* [11]), or learn individual limb detectors and then apply kinematic tree based constraints [16, 20].

As the human body is highly articulated, it is a complicated object to detect, particularly at the resolution of individual body parts. Although explicit kinematic tree based structures can be an effective tool in this regard, we avoid such assumptions, instead learning characteristic spatial configurations directly from images. Our patch based representation allows us to work on the scale of small body parts, and besides providing spatial information for each of these parts, enables us to mix and match part combinations for modeling generic appearance.

Previous work: There are currently only a few bottom up approaches to the estimation of human pose from images and video. Many of these methods use combinations of weak limb detectors to detect the presence of a person [16, 9], but are not capable of deducing 3D poses accurately enough to infer actions and gestures. Similarly, in [15], loose 2D configurations of body parts are used to coarsely track people in video by filtering potential limb-like objects based on motion and color statistics.

Most methods for precise pose estimation adopt top-down approaches in the sense that they try to minimize projection errors of kinematic models, either using numerical optimization [21] or by generating large number of pose hypotheses [11]. With suitable initialization or sufficiently fine sampling such methods can produce accurate results, but the computational cost is high. Efficient matching methods such as [6] fall back to the assumption of having pre-segmented images. [20] discusses an interesting approach that combines weak responses from bottom-up limb detectors based on a statistical model of image likelihoods with a full articulated body model using belief propagation. However, this approach uses background subtraction and it also relies on multiple calibrated cameras.

A recent work that addresses upper body pose from single images in clutter is [11]. This is based on the use of heuristic image cues including a clothes model and skin color detection; and relies on generating and testing large numbers of pose hypotheses using a 3D body model. Here we adopt an example based approach inspired by [19] and [1]. Both of these approaches infer pose from edge feature representations of the



(a) (b) (c) (d) (e)

Fig. 1. An overview of our method of pose estimation from cluttered images. (a) original image, (b) a grid of fixed points where the descriptors are computed (each descriptor block covers an array of 4x4 cells, giving a 50% overlap with it's neighbouring blocks), (c) SIFT descriptors computed at these points, the intensity of each line representing the weight of the corresponding orientation bin in that cell, (d) Suppressing background using a sparse set of learned (NMF) bases encoding human-like parts, (e) final pose obtained by regression

input image using a model learned from a number of labeled training examples (image-pose pairs). However, both require clean backgrounds for their representations. Here we develop a more general approach that works with cluttered backgrounds. Our image representation is based on local appearance descriptors extracted from a uniformly spaced grid of image patches. This notion, in the form of superpixels, or image sites, has previously been used in several different contexts, *e.g.* [4, 13, 17]. We also take inspiration from the image coding and object localization methods described in [22, 14].

2. Regression based approach

Example based methods often have problems when working in high dimensional spaces as it is difficult to create or incorporate enough examples to densely cover the space. This is particularly true for human pose estimation which must recover many articular degrees of freedom from a complex image signal. The sparsity of examples is usually tackled by smoothly interpolating between nearby examples. Learning a single smooth inference model in the form of a regressor was suggested in [1]. This has the advantage of directly recovering pose parameters from image observations, which obviates the need to attach explicit meanings or attributions to image features (*e.g.* labels designating the body parts seen). However it requires a robust and discriminative representation of the input image. Following [1], we take a regression based approach, extending it to deal with the presence of cluttered image background. Encoding pose by the 3D locations of 8 key upper body joint centres, we regress a 24-d output pose vector $\mathbf{y}$ on a set of image features $\mathbf{x}$:

$$\mathbf{y} = \mathbf{A}\,\boldsymbol{\phi}(\mathbf{x}) + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\phi}(\mathbf{x})$ is a vector of basis functions, $\mathbf{A}$ is a matrix of weight vectors, and $\boldsymbol{\epsilon}$ is a residual error vector. The matrix $\mathbf{A}$ is estimated by minimizing least squares error while applying a regularization term to control overfitting.

The method turns out to be relatively insensitive to the choice of regression methods. Here we work with a classical single-valued regressor as frontal upper body gestures have relatively few multimodality problems in comparison to the full body case, but the multimodal multi-valued regression method of [2] could also be used if necessary. Our

main focus is on exploring suitable image representations and mechanisms for dealing with background clutter.

3. Image Features

Image information can be encoded in many different ways. Given the variability of clothing and the fact that we want to be able to use black and white images, we do not use colour information. Silhouette shape and body contours have proven effective in cases where segmentations are available, but with current segmentation algorithms they do not extend reliably to images with cluttered backgrounds [12]. Furthermore, more local, part-based representations are likely to be able to adapt better to the highly non-rigid structure of the human body. To allow the method to key on important body contours, we based our representation on local image gradients. For effective encoding, we use histograms of gradient orientations in small spatial cells. The relative coarseness of the spatial coding provides some robustness to small position variations, while still capturing the essential spatial position and limb orientation information. Note that owing to loose clothing, the *positions* of limb contours do not in any case have a very precise relation to the pose, whereas *orientation* of body edges is a much more reliable cue. Hence a SIFT-like representation is appropriate. We compute these histograms in the same way as SIFT descriptors [5], quantizing gradient orientations into discrete values in small spatial cells and normalizing these distributions over local blocks of cells to achieve insensitivity to illumination changes. To retain the information about image location that is indispensable for pose estimation, the descriptors are computed at fixed grid locations in the image window. Figure 1(c) shows the features extracted from a sample image. We denote the descriptor vectors at each of these L locations as $\mathbf{v}^l, l \in \{1 \dots L\}$, and represent the complete image as a large vector $\mathbf{x}$, a concatenation of the individual descriptors: $\mathbf{x} \equiv (\mathbf{v}^{1^\top}, \mathbf{v}^{2^\top}, \dots \mathbf{v}^{L^\top})^\top$.

An alternate approach that failed to provide convincing results in our experiments is a *bag of features* style of representation. In the absence of reliable salient points on the human body, we computed SIFT descriptors at all edge points in the image and added spatial information by appending image coordinates to the descriptor vector. For effective pose estimation, though, it seems that coding location precisely is extremely important and extracting descriptors on a fixed grid of locations is preferable.

3.1. Similarity based encoding

Representations based on collections of local parts are commonly used in object recognition [18, 3, 7]. A common scheme is to identify a representative set of parts as a vocabulary for representing new images. In an analogous manner, the human body can be represented as a collection of limbs and other key body parts in particular configurations. To test this, we independently clustered patches at each image location to identify representative configurations of the body parts that are seen in these locations. Each image patch was then represented by *softly* vector quantizing the SIFT descriptor by voting into each of its corresponding k-means centers, *i.e.* as a sparse vector of similarity weights computed from each cluster center. Results from this and other representations are summarized in figure 4 and discussed in the experimental section.

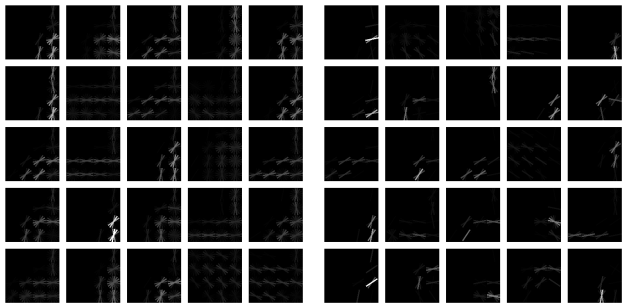


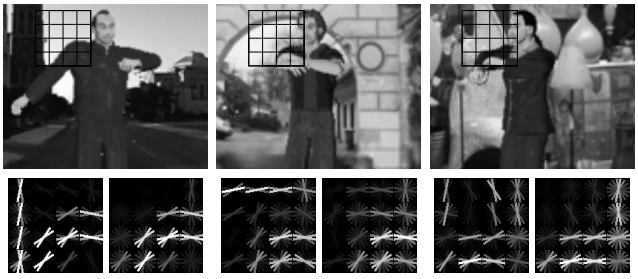
Fig. 2. Exemplars, or basis vectors, extracted from SIFT descriptors over 4000 image patches located close to the right shoulder. The corresponding block is shown in figure 3. *(left)* Representative examples selected by k-means. *(right)* Much sparser basis vectors obtained by non-negative matrix factorization. These capture important contours encoding a shoulder, unlike the denser examples given by k-means.

3.2. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) is a recent method that can exploit latent structure in data to find part based representations [10,8]. NMF factorizes a non-negative data matrix $\mathbf{V}$ as $\mathbf{V} \sim \mathbf{WH}$, where $\mathbf{W}$ and $\mathbf{H}$ are both constrained to be non-negative. If the columns of $\mathbf{V}$ consist of feature vectors, $\mathbf{W}$ can be interpreted as a set of basis vectors, and $\mathbf{H}$ as corresponding coefficients needed to reconstruct the original data. Each entry of $\mathbf{V}$ is thus represented as $\mathbf{v}_i = \sum_j \mathbf{w}_j h_{ji}$. Unlike other linear decompositions such as PCA or ICA [23], this purely additive representation (there is no subtraction) tends to pull out local fragments that occur consistently in the data, giving a sparse set of basis vectors. The results of applying NMF to the 128-d descriptor space at a given patch location are shown in figure 2.

Besides capturing the local edges representative of human contours, the NMF bases allow us to compactly code each 128-d SIFT descriptor directly by its corresponding vector $\mathbf{h}$ of basis coefficients. This serves as a nonlinear image coding that retains good locality for each patch: $\phi(\mathbf{x}) \equiv (\mathbf{h}^{1^\top}, \mathbf{h}^{2^\top}, \ldots \mathbf{h}^{L^\top})^\top$ in (1). Having once estimated the basis $\mathbf{W}$ (for each image location) from a training set, we keep it fixed when we compute the coefficients for test images. In our case, we find that the performance tends to saturate at about 30-40 basis elements per grid patch.

Selectively removing clutter: An interesting advantage of using NMF to represent images is its ability to selectively encode only the foreground of regions of interest, hence effectively rejecting background. We find that by learning the bases $\mathbf{W}$ from a set of clean images (containing no background clutter), and using these only additively (with NMF) to reconstruct images with clutter, only the edge features corresponding to the foreground are reconstructed, while suppressing features in unexpected parts of the image. This happens because the bases are constructed from clean human images and hence forced to contain mass only in regions containing human-like features. Some examples illustrating this phenomenon are shown in figure 3.



(a) (b) (c)

Fig. 3. To selectively encode foreground features and suppress unwanted background, we use NMF bases learned on clean images (with no clutter) to reconstruct the cluttered image patches. For each image, the original SIFT feature and its representation using the bases extracted using NMF are shown for the patch marked. Features corresponding to background edges such as those of the building on the left in (a) and the arch in (b) are clearly suppressed, while background clutter in (c) is downweighted

4. Experimental Performance

We trained and evaluated the methods on two different databases of human pose examples. The first is a set of randomly generated human poses using a human model rendering package, POSER from Curious Labs. This is a subset of the data used in [19], kindly supplied to us by its authors. The second dataset contains motion capture data from human recordings of several sets of arm movements. It was obtained from http://mocap.cs.cmu.edu. Unfortunately neither set has significant background clutter, nor are we aware of any existing dataset that combines images of human poses with background clutter and motion capture data for training and ground truth. However, as all of this data was created under controlled conditions, we were able to artificially add random backgrounds to the images while retaining their 3D pose ground truth information for comparative testing with and without background clutter. So we have *clean* and *cluttered* versions of both image sets, albeit with somewhat artificial poses (for set 1) and backgrounds.

For descriptor computation, we quantized gradient orientations into 8 orientation bins (in $[0, \pi]$) in 4×4 spatial cells, as described in [5], using blocks 32 pixels across. Our images are centered and resized to 118×95 pixels. The descriptor histograms are computed on a 4×6 grid of 24 uniformly spaced overlapping blocks on each image, giving rise to 3072-d image descriptor vectors $\mathbf{x}$.

Figure 4 shows the performance of different feature encodings over all combinations of training and testing on clean and cluttered images. The regularization parameter of the regressor was optimized using cross validation. These figures are reported for 4000 training and 1000 test points from the POSER dataset. The errors reported indicate, in centimeters, the RMS deviations for the 3D locations of shoulder, elbow, wrist, neck and pelvis joints. The best performance, as expected, is obtained by training and testing on clean, background-free images, irrespective of the descriptor encoding used. Training on clean images does not suffice for generalization to clutter. Using cluttered images for training provides reasonably good generalization to unseen backgrounds, but the result-

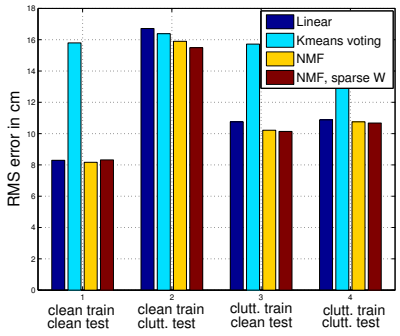


Fig. 4. A comparison of the performance of different feature encodings in regressing 3D pose, over different combinations of training and testing on clean and cluttered data. See text.

ing errors are larger by 2-3 cms on both clean and cluttered test sets than the best case. Surprisingly, a linear regressor on the vector $\mathbf{x}$ performs very well despite the clutter — an examination of the elements of the weight matrix $\mathbf{A}$ reveals this is due to automatic downweighting of descriptor elements that usually contain only background. On average, the k-means based representation performs the worst of all and the NMF-based representation gives the best performance. To study the space of encodings 'between' an extreme exemplar based k-means representation and the set of basis vectors obtained by NMF, we tested NMF with constraints on the sparsity level of the basis vectors and coefficients [8]. Varying the sparsity of the basis vectors $\mathbf{W}$ has very little effect on the performance, while varying the sparsity of the coefficients $\mathbf{H}$ gives results spanning the range of performances from k-means to unconstrained NMF. As the sparsity prior on $\mathbf{H}$ is increased to a maximum, NMF is forced to use only a few basis vectors for each training example, in the extreme case giving a solution very similar to k-means.

To see the effect of depth ambiguities on these results, we computed errors separately in the x and y coordinates corresponding to the image plane and z, corresponding to depth. We find that errors in depth estimation are a little higher than those in lateral displacement. *E.g.*, of the 10.88 cm of error obtained in the experiment on cluttered images, 9.65 cm comes from x and y, while 12.97 cm from errors in z. In the absence of clutter, we obtain errors of $\sim$8 cm. This is similar to the performance reported in [19] on this dataset (when transformed into the angle based error measure used in that paper), showing that regression based methods can match the performance of nearest-neighbourhood based ones, while avoiding having to store and search through excessive amounts of training data. Examples of pose estimation on the cluttered test set are shown in figure 5.

For our second set of experiments, we use $\sim$1600 images from 9 video sequences of motion capture data. Performance on a test set of 300 images from a 10th sequence gives an error of 7.4 cm in the presence of clutter. We attribute this slightly improved performance to the similarity of the gestures performed in the test set to those in the training sequences, although we emphasize that in the test set they were performed by a different subject. Figure 6 shows sample reconstructions over test examples from the second database and from some natural images found with Google. We find that train-

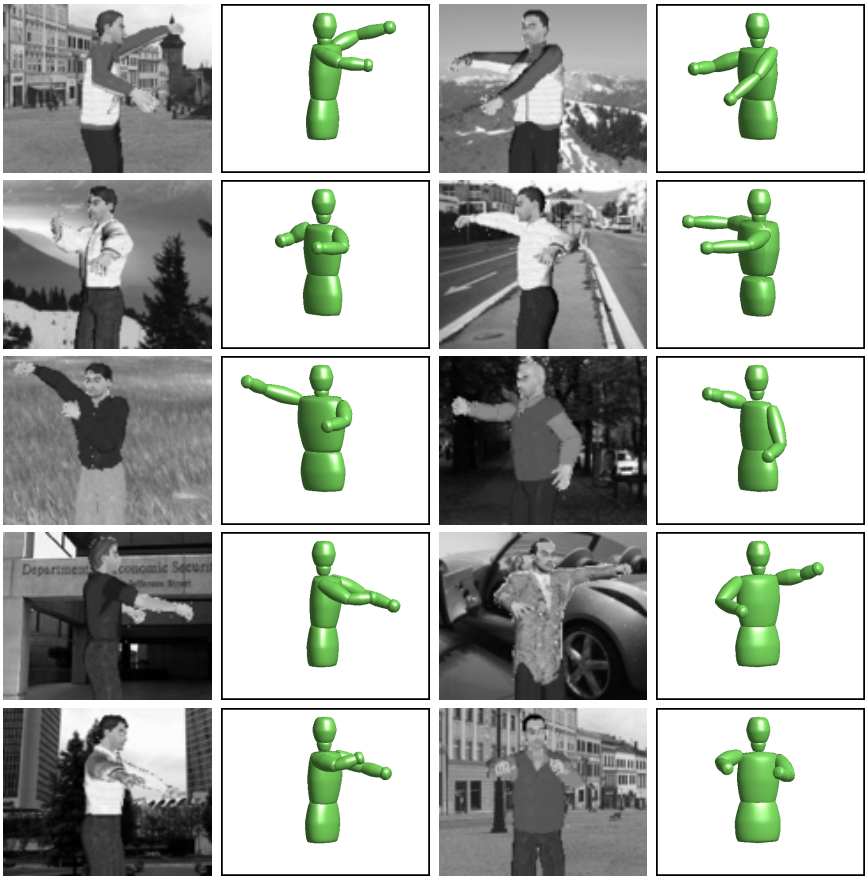


Fig. 5. Sample pose estimates from a test set of 1000 images in cluttered backgrounds. No knowledge of segmentation is used in the process.

ing on the second dataset also gives qualitatively better performance on a set of randomly selected real images. This suggests that it is important to include more 'natural', human-like poses in the training set, which are not covered by randomly sampling over the space of possible poses. We are currently collecting more training data to improve performance on typical human gestures.

5. Conclusion

We have presented a method that is capable of estimating 3D human upper body pose from a single image. To the best of our knowledge, this is the first totally bottom-up approach to this problem that works in the presence of background clutter. An image representation based on a set of local descriptors computed at known locations in the image allows us to model the appearance of different parts independently, before combining the information for pose regression. The regression based approach eliminates the need to store large numbers of training examples. We have also demonstrated a

Fig. 6. Pose reconstructions on real unseen images. The first 3 images are taken from a test sequence in our motion capture dataset which includes similar gestures made by another person, while the last 3 are example images obtained using Google. The results on the real images are not very precise if overlayed on the images, but they do capture the general appearance of the subject's gestures fairly well. They would probably improve considerably given more training data for common gestures.

novel application of non-negative matrix factorization that allows us to discriminate features of interest from background. This is likely to prove useful in other applications including segmentation and recognition.

Future work: We currently work with centered images of people. The framework could be applied as it is on the output of a person detector to estimate pose or infer activity of multiple people in a scene. In fact, we are hoping to construct a unified person detector and pose estimator that uses a knowledge of human body configurations for complete detection. As regards immediate extensions, the method will be trained on a larger database of common gestures and extended to incorporate motion information for tracking full body motion in cluttered backgrounds.

Acknowledgements

This work was supported by a MENRT Doctoral Research Fellowship from the French Education Ministry and in part by the IST Programme of the European Community, under the PASCAL Network of Excellence.

References

[1] A. Agarwal and B. Triggs. 3D Human Pose from Silhouettes by Relevance Vector Regression. In *Int. Conf. Computer Vision & Pattern Recognition*, 2004.

[2] A. Agarwal and B. Triggs. Monocular Human Motion Capture with a Mixture of Regressors. In *IEEE Workshop on Vision for Human-Computer Interaction*, 2005.

[3] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490, 2004.

[4] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Int. Conf. Computer Vision & Pattern Recognition*, 2005.

[5] D.Lowe. Distinctive Image Features from Scale-invariant Keypoints. *International Journal of Computer Vision*, 60, 2:91–110, 2004.

[6] P. Felzenszwalb and D. Huttenlocher. Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61 (1), 2005.

[7] R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *Int. Conf. Computer Vision & Pattern Recognition*, 2003.

[8] P. Hoyer. Non-negative Matrix Factorization with Sparseness Constraints. *J. Machine Learning Research*, 5:1457–1469, 2004.

[9] K.Mikolajczyk, C. Schmid, and A. Zisserman. Human Detection based on a Probabilistic Assembly of Robust Part Detectors. In *European Conference on Computer Vision*, volume I, pages 69–81, 2004.

[10] D. D. Lee and H. S. Seung. Learning the Parts of Objects by Non–negative Matrix Factorization. *Nature*, 401:788–791, 1999.

[11] M. Lee and I. Cohen. Human Upper Body Pose Estimation in Static Images. In *European Conference on Computer Vision*, 2004.

[12] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7–27, 2001.

[13] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering Human Body Configurations: Combining Segmentation and Recognition. In *Int. Conf. Computer Vision & Pattern Recognition*, 2004.

[14] B. Olshausen and D. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–339, 1996.

[15] D. Ramanan and D. Forsyth. Finding and Tracking People from the Bottom Up. In *Int. Conf. Computer Vision & Pattern Recognition*, 2003.

[16] R. Ronfard, C. Schmid, and B. Triggs. Learning to Parse Pictures of People. In *European Conference on Computer Vision*, pages IV 700–714, Copenhagen, 2002.

[17] S. Kumar and M. Hebert. Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification. In *Int. Conf. Computer Vision*, 2003.

[18] E. Sali and S. Ullman. Combining Class-specific Fragments for Object Classification. In *British Machine Vision Conference*, 1999.

[19] G. Shakhnarovich, P. Viola, and T. Darrell. Fast Pose Estimation with Parameter Sensitive Hashing. In *Int. Conf. Computer Vision*, 2003.

[20] L. Sigal, M. Isard, B. Sigelman, and M. Black. Assembling Loose-limbed Models using Non-parametric Belief Propagation. In *NIPS*, 2003.

[21] C. Sminchisescu and Bill Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research*, 22(6):371–391, June 2003. Special issue on Visual Analysis of Human Movement.

[22] J. Sullivan, A. Blake, M. Isaard, and J. MacCormick. Object Localization by Bayesian Correlation. In *Int. Conf. Computer Vision*, 1999.

[23] J. van Haateran and vander Schaaf A. Independent component filters of natural images compared with simlpe cells in preimary visual cortex. *Proc. R. Soc. Lond.*, B 265:359–366, 1998.