Estimating Human Pose from Occluded Images

Jia-Bin Huang and Ming-Hsuan Yang
Electrical Engineering and Computer Science
University of California at Merced
{jbhuang, mhyang}@ieee.org

Abstract. We address the problem of recovering 3D human pose from single 2D images, in which the pose estimation problem is formulated as a direct nonlinear regression from image observation to 3D joint positions. One key issue that has not been addressed in the literature is how to estimate 3D pose when humans in the scenes are partially or heavily occluded. When occlusions occur, features extracted from image observations (e.g., silhouettes-based shape features, histogram of oriented gradient, etc.) are seriously corrupted, and consequently the regressor (trained on un-occluded images) is unable to estimate pose states correctly. In this paper, we present a method that is capable of handling occlusions using sparse signal representations, in which each test sample is represented as a compact linear combination of training samples. The sparsest solution can then be efficiently obtained by solving a convex optimization problem with certain norms (such as l_1-norm). The corrupted test image can be recovered with a sparse linear combination of un-occluded training images which can then be used for estimating human pose correctly (as if no occlusions exist). We also show that the proposed approach implicitly performs relevant feature selection with un-occluded test images. Experimental results on synthetic and real data sets bear out our theory that with sparse representation 3D human pose can be robustly estimated when humans are partially or heavily occluded in the scenes.

1 Introduction

Estimating 3D articulated human pose from single view is of great interest to numerous vision applications, including human-computer interaction, visual surveillance, activity recognition from images, and video indexing as well as retrieval. Notwithstanding some demonstrated success in the literature, this problem remains very challenging for several reasons. First, recovering 3D human poses directly from 2D images is inherently ambiguous due to loss of depth information. This problem is alleviated with additional information such as temporal correlation obtained from tracking, dynamics of human motion and prior knowledge, or multiple interpretations conditioned on partial image observations. In addition, the shape and appearance of articulated human body vary significantly due to factors such as clothing, lighting conditions, viewpoints, and poses. The variation of background scenes also makes the pose estimation more difficult. Therefore, designing image representations that are invariant to these factors is critical for effective and robust pose estimation.

Human pose estimation algorithms can be categorized as generative (model-based) and discriminative (model-free). Generative methods employ a known model (e.g., tree structure) based on prior knowledge [1]. The pose estimation process includes two parts: 1) modeling: constructing the likelihood function and 2) estimation: predicting the most likely hidden poses based on image observations and the likelihood function. However, it is difficult to consider factors such as camera viewpoint, image representations and occlusion in the likelihood functions. Furthermore, it is computationally expensive to compute these functions and thus makes them unsuitable for inferring the hidden poses. In contrast, discriminative methods do not assume a particular human body model, and they can be further categorized as example-based [2] and learning-based [3–5]. Example-based approaches store a set of training samples along with their corresponding pose descriptors. For a given test image, a similarity search is performed to find similar candidates in training set and then obtain estimated poses by interpolating from their poses [2]. On the other hand, learning-based approaches learn the direct mapping from image observations to pose space using training samples [3–5]. While generative methods can infer poses with better precision than discriminative ones, discriminative approaches have the advantage in execution time.

Several image representations have been proposed in discriminative pose estimation algorithms such as shape context of silhouettes [6], signed-distance functions on silhouettes [7], binary principal component analysis of appearance [8], and mixture of probabilistic principal component analysis on multi-view silhouettes [2]. However, silhouettes are inherently ambiguous as different 3D poses can have very similar silhouettes. In addition, clean silhouette can be better extracted with robust background subtraction methods, which is not applicable in many real-world scenarios (e.g., videos with camera motion, dynamic background, sudden illumination change, etc.). To cope with this problem, appearance features like block SIFT descriptors [9], Haar-like features [10], Histogram of oriented gradients (HOG) [6, 11, 12] or bag-of-visual-words representations [13] have been proposed for pose estimation. These descriptors contain richer information than silhouette-based features, but they inevitably encode irrelevant background clutter into the feature vector. These unrelated feature dimensions may have accumulative negative effects on learning the image-to-pose mapping and thereby increase errors in pose estimation. Agarwal et al. [6] deal with this problem by using non-negative matrix factorization to suppress irrelevant background features, thereby obtaining most relevant HOG features. In [10], relevant features are selected from a predefined set of Haar-like features through multi-dimensional boosting regression. Okada and Soatto [12] observed that the components related to human pose in a feature vector are pose dependent. Thus, they first extract pose clusters using kernel support vector machine, and then train one local linear regressor for each cluster with features selected from the cluster.

Another important issue that has not been explicitly addressed in the literature is how to robustly estimate 3D pose when humans in the scenes are partially or heavily occluded. When parts of a human body are occluded, the extracted descriptors from image observation (e.g., shape features from silhouettes, block SIFT, HOG, or part-based features, etc.) are seriously corrupted. The learned regressor, induced from un-occluded images, is not able to estimate pose parameters correctly when a human is occluded in an image. While using tracking algorithm or making use of human motion prior may alleviate this problem, an effective approach is needed to explicitly handling occlusion.

In this paper, we show we are able to deal with such problems using sparse image representations in which each test sample can be represented as a compact linear combination of training samples, and the sparest solution can be obtained via solving a convex optimization problem with certain norms (such as l_1-norm). Within this formulation, the corrupted test image can be recovered with a linear combination of un-occluded training images which can then be used for estimating human pose correctly (as if no occlusions exist). The proposed algorithm exploits both the advantages of example-based and learning-based algorithms for pose estimation. In our algorithm, when we represent a given image as a linear combination of training samples and obtain a sparse solution, we are actually searching for a small number of candidates in the training data set that best synthesizes the test sample. It is similar to the idea of example-based approaches which perform efficient nearest neighbor search, but yet we use a more compact representation that has been proven to be effective in dealing with noise. We then learn a mapping between the compact representation and their corresponding pose space using regression functions. The major difference between sparse image representation and example-based approach (nearest neighbor search) is that we consider all possible supports and adaptively select the minimal number of training samples required for representing each test sample. Hence, with the recovered test sample, we can estimate 3D human pose when humans in the scenes are partially or heavily occluded. Moreover, by using sparse representations we can implicitly perform relevant feature selection. When representing each test sample as a compact linear combination of training samples, those mismatched components are treated as part of reconstruction error and discarded directly. Intuitively, we are replacing the background clutter in the test samples with backgrounds in the training images. In this way, we achieve pose-dependent feature selection without making any approximation (like clustering poses in [12] or bag-of-visual-words in [13]) and avoid the need to increase the complexity of the learning-based algorithms.

The contributions in this paper can be summarized in two main aspects. First, we propose an algorithm to handle occlusion in estimating 3D human pose by representing each test sample as a sparse linear combination of training samples. The prediction errors are significantly reduced by using the reconstructed test samples instead of the original ones when human in images are occluded. Second, we achieve pose-dependent feature selection by solving sparse solution with reconstruction error. Our approach improves over the learning-based algorithm without feature selection.

The remainder of this paper is organized as follows. Section 2 describes related works on human pose estimation. In Section 3, we introduce the proposed image representation scheme. We test our approach on both synthesized (INRIA) and real data set (HumanEva I) to demonstrate the ability to handle occlusion and feature selection in Section 4. We conclude this paper with comments on future work in Section 5.

2 Related Work

Due to its scope and potential applications, there has been a substantial amount of work on the general problem of human motion capture and understanding. As such, we find it useful to place the focus of our work within the taxonomy proposed by Moedlund and Granum [14] whereby the field of work is presented in the categories of person detection, tracking, pose estimation and recognition. Out approach fits best into the category of pose estimation where the goal is to accurately estimate the positions of the body parts. More specifically, our approach is to estimate 3D pose from a single image without the use of temporal information. We will focus on previous work with a similar goal and leave interested readers to consult one of the surveys for a more complete listing of work in this general area [14, 15].

Previous approaches to human pose estimation from a single image can be broadly categorized as model-based or model-free based. In model-based approaches a parametric model that captures the kinematics of the human body is explicitly defined. This model can be used in a predict-match-update paradigm in which maximal agreement between the model and the image measurements is sought. One method for this is to simultaneously detect body parts and assemble them in a bottom-up manner. Pictorial structures [16] presented a convenient discrete graphical form for this that can be adapted for people using an efficient dynamic programming minimization proposed by Felzenszwalb and Huttenlocher [17] and later used in various forms by a number of researchers [18–20]. Mori et al. followed a similar line of thought, but employed "superpixels" for the task of segmenting and detecting body parts [21]. Sigal et al. presented a bottom-up approach in a continuous parameter space using a modified particle filter for the minimization [1]. In contrast, Taylor developed a method to invert a kinematic model given an accurate labeling of joint coordinates that provides reconstruction up to a scale ambiguity [22]. This method was combined with shape-context matching in a fully automatic system by Mori and Malik [23].

Model-free based approaches, which include regression and example based methods, take a top-down approach to this problem and attempt to recover a mapping from image feature space to pose parameter space directly. An early approach of this type represented 3D pose space as a manifold that could be approximated by hidden Markov models [24]. Agarwal and Triggs advocated the relevance vector machine (RVM) [25] to learn this mapping where silhouette boundary points were used as features [26]. Rosales and Sclaroff used specialized maps in addition to an inverse rendering process to learn this mapping [27]. Along a different line, Shakhnarovich do not learn a regression function, but instead directly make use of training examples in a lookup table using an efficient hashing [28]. The feature space used for these types of methods, with few exceptions, is global in the sense that the features carry no information about the body region they describe. This provides a clean top-down approach that circumvents any need to implement part detectors. One exception to this is recent work by Agarwal and Triggs where the goal is pose estimation in cluttered environments that localized feature with respect to the window of interest [6].

Our approach uses a regression model to learn the mapping from image feature space to pose space, but differs from previous work in that sparse representations are learned from examples with demonstrated ability to handle occlusions.

3 Image Representation

We represent each input image observation as $\mathbf{x} \in \mathbb{R}^m$ and the output 3D human pose vector as $\mathbf{y} \in \mathbb{R}^k$. Given a training set of N labeled examples $\{(\mathbf{x}_i, \mathbf{y}_i)|i = 1, 2...N\}$, the goal of a typical learning-based approach in human pose estimation is to learn a smooth mapping function that generalizes well for unseen image observation $\mathbf{b}$ in the testing set. As mentioned in the Section 1, straightforward appearance features inevitably encode unwanted background information in $\mathbf{x}$, which may introduce significant errors in estimating pose from the test samples since the background clutters may be quite different. The performance of the learned mapping function will also be seriously degraded if humans are occluded in images because part of feature dimensions are corrupted. To address these two problems, we present a formulation to represent test samples such that the occluded or the irrelevant parts of the test samples can be recovered by solving convex optimization problems.

3.1 Test Image as a Linear Combination of Training Images

Given sufficient number of training samples, we model a test sample $\mathbf{b}$ by the linear combination of the N training samples:

$$\mathbf{b} = \omega_1 \mathbf{x}_1 + \omega_2 \mathbf{x}_2 + \cdots + \omega_N \mathbf{x}_N, \tag{1}$$

where $\omega_i, i \in \{1, 2, \ldots, N\}$ are the scalar coefficients denoting the weights of the i-th training sample contributing for synthesizing the test samples $\mathbf{b}$. By arranging the N training samples as columns of a matrix $A = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N] \in \mathbb{R}^{m \times N}$, the linear representation of $\mathbf{b}$ can be written compactly as

$$\mathbf{b} = A\omega, \tag{2}$$

where $\omega = [\omega_1, \omega_2, \ldots, \omega_N]^{\mathrm{T}}$ is the coefficient vector.

With this formulation, each test sample $\mathbf{b}$ can be represented using the corresponding coefficient vector ω by solving the linear system of equations $\mathbf{b} = A\omega$. If the number of the dimension of the image observation m is larger than the number of training samples N, then the unique solution for ω can usually be obtained by solving the overdetermined system. However, with data noise or if $N > m$, then the solution is not unique. Conventionally, the method of least squares can be used to find an approximate solution to this case by solving minimum l_2-norm solution:

$$\min \; ||\omega||_2 \quad \text{subject to} \quad A\omega = \mathbf{b}. \tag{3}$$

For the system $A\omega = \mathbf{b}$, the minimum l_2-norm solution can be obtained by $\hat{\omega}_2 = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\mathbf{b}$. However, the minimum l_2-norm (energy) solution $\hat{\omega}_2$ is usually dense (with many nonzero entries), thereby losing the discriminative ability to select the most relevant training samples to represent the test one. As the vectors of pose parameters for articulated human body pose reside in an high-dimensional space, the resulting pose variations are large and diverse. It is reasonable to assume that only very a small portion of training samples are needed to synthesize a test sample (i.e., only a few nonzero terms in the solution $\hat{\omega}$ for solving $A\omega = \mathbf{b}$). This is especially true when the training set contain a large number of examples that densely cover the pose space.

3.2 Finding Sparse Solutions via l_1-norm Minimization

To find the sparest solution to $A\omega = \mathbf{b}$, we can solve the optimization problem in (2) with l_0-norm

$$\min \; ||\omega||_0 \quad \text{subject to} \quad A\omega = \mathbf{b}, \tag{4}$$

where l_0-norm counts the nonzero entries in the vector ω. However, solving the l_0-norm minimization problem is both numerically unstable and NP-hard (no polynomial time solutions exist).

Recent theories from compressive sensing [29–32] suggest that if the solution of ω is sparse enough, then the sparsest solution can be exactly recovered via the l_1-norm optimization:

$$\min \ ||\omega||_1 \ \text{ subject to } \ A\omega = \mathbf{b}, \tag{5}$$

where the l_1-norm sums up the absolute weights of all entries in ω (i.e., $||\omega||_1 := \sum_i |\omega_i|$, where ω_i stands for the i-th entry in the vector). This is a convex optimization problem that can be solved by linear programming methods (e.g., generic path-following primal-dual algorithm) [33], also known as basis pursuit [34].

3.3 Coping with Background Clutter and Occlusion

Although sparse solution for the coefficient ω can be obtained by solving an l_1 optimization in (5), in the context of human pose estimation we may not find the sparest solution $\hat{\omega}_1$ that well explains the similarity between the test sample $\mathbf{b}$ and the training samples A. This can be explained with several factors. First, the background clutter may be quite different been training and testing samples, and thus there exist inevitable reconstruction errors when representing the test sample by training samples. For example, even the test sample contains pose exactly the same as one of the training samples, the background could be quite different, causing reconstruction error in representing the test sample. Second, when humans in the test images are occluded, the linear combination of training samples may not able to synthesize the occluded parts. Third, if we use dense holistic appearance features such as HOG or block SIFT, there may have misalignments within the detected image regions. To account for these errors, we introduce an error term $\mathbf{e}$ and then modify (2) as

$$\mathbf{b} = A\omega + \mathbf{e} = [A \ \ I] \begin{bmatrix} \omega \\ \mathbf{e} \end{bmatrix} = B\mathbf{v}, \tag{6}$$

where $B = [A \ \ I] \in \mathbb{R}^{m \times (N+m)}$ and $\mathbf{v} = [\omega \ \ \mathbf{e}]^{\mathrm{T}}$. If the vector $\mathbf{v}$ is sparse enough, the sparest representation can be obtained by solving the extended l_1-norm minimization problem:

$$\min \ ||\mathbf{v}||_1 \ \text{ subject to } \ B\mathbf{v} = \mathbf{b} \tag{7}$$

In this way, the first N entries of vector $\mathbf{v}$ obtained from solving (7) correspond to the coefficients of the training samples that can represent the test sample best using minimum nonzero entries. On the other hand, the latter m entries account for those factors (occlusion, misalignment, and background clutter) which can not be well explained by the training samples.

We validate the recovery ability of our approach using a synthetic data set [26] in which 1927 silhouette images are used for training and 418 images for testing. These images are first manually cropped and aligned to 128 × 64 pixels. For efficiency, we further downsample these images by a factor of 4 and add random blocks to simulate the occluded silhouettes. Fig. 1 shows that we can recover from the corrupted test feature (c) to (d). The reconstructed feature vector (d) can then be used for regressing the output 3D joint angle vector.

We also demonstrate that our algorithm, as a result of using sparse representation, is able to perform feature selection implicitly by discarding irrelevant background information in the feature vectors in Fig. 2. Fig. 2 shows the original test image, the

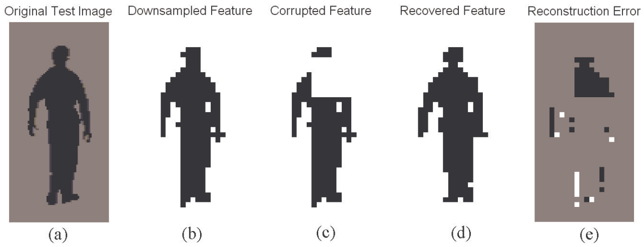


Fig. 1. Occlusion recovery on a synthetic dataset. (a)(b) The original input image and its feature. (c) Corrupted feature via adding random block. (d) Recovered feature via find the sparsest solution (7). (e) Reconstruction error.

corresponding HOG feature vector, and the recovered feature vector, and the reconstruction errors using our sparse representations (from (a) to (d)). Note that most of the reconstruction errors appear at the locations corresponded to background clutters, thereby validating our claim that the proposed sparse representation is able to filter out irrelevant noise.

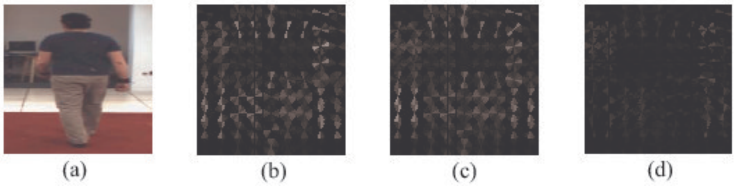


Fig. 2. Feature selection example. (a) Original test image. (b) The HOG feature descriptor computed from (a). (c) Recovered feature vector by our algorithm. (d) The reconstruction error.

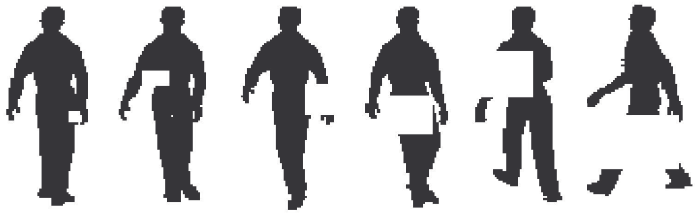
4 Experimental Results

We test the proposed algorithm on synthetic [26] and real [4] data sets for empirical validation. In all experiments, we use Gaussian process regressor [35] to learn the mapping between image features and the corresponding 3D pose parameters. We first demonstrate the proposed method is able to estimate human pose from images with occlusions. Even without occlusions, we show that the our algorithm still outperforms the baseline methods as a result of implicit feature selection within our formulation.

4.1 Robustness to Occlusion

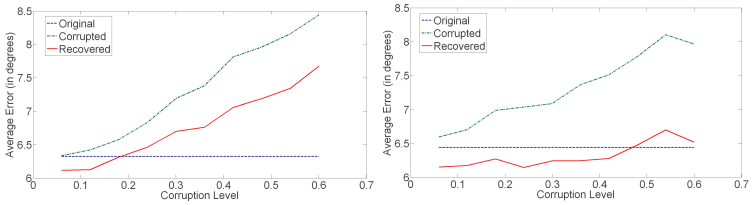
We use the synthetic data set in [26] to show that the proposed algorithm is able to recover the un-occluded silhouettes from occluded ones. We generate random blocks (with their width corresponds to the corruption level (CL)) to the all test sample to synthesize occluded image silhouettes (see Fig. 3 for some sample test images under various corruption level). We use two different feature representations in our experiment. The first one is the principle component analysis (PCA) where each test image is represent by its first 20 coefficients of principal components. The second image feature is based on the image appearance (i.e., pixel values) of the downsampled images.

Fig. 4 shows the average errors in angles (degree) for three experiment settings: 1) features extracted from original test image (baseline), 2) features computed from the corrupted images (see Fig. 3), and 3) recovered features using the proposed algorithm. First we note that in both PCA and appearance settings, the proposed algorithm improves the accuracy of pose estimation under occlusions. We also observe that our method with appearance features (e.g., downsampled images) performs better than that with holistic features (e.g., PCA). This can be explained by the fact holistic PCA is known to be sensitive to outliers. Thus, when a silhouette is occluded, the PCA coefficients computed from the occluded images are likely to be very different from the ones without occlusions. In contrast, only a small number of pixels of the occluded images have been changed or corrupted, thereby facilitating the process of recovering the un-occluded images. These results suggest that sparse and localized feature representations are suitable for pose estimation from occluded images.



(a) CL=0.1 (b) CL=0.2 (c) CL=0.3 (d) CL=0.4 (e) CL=0.5 (f) CL=0.6

Fig. 3. Sample test images under various corruption levels (CL) in the synthetic data set. The occlusions seriously corrupt the shape of the silhouette images.



(a)

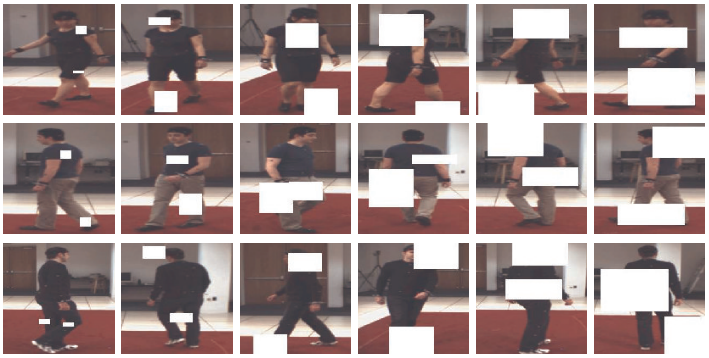
(b)

Fig. 4. Average error of pose estimation on synthetic data set using different features: (a) principle component analysis with 20 coefficients. (b) downsampled (20×20) images.

To further gauge the performance of the proposed method, we use the synchronized image and motion capture data from the HumanEva data sets [4] for experiments. The HumanEva I data set consists of 4 views of 4 subjects performing a set of 6 predefined actions (walking, jogging, gesturing, throwing/catching, boxing, combo) 3 times. For efficiency and performance analysis, we chose the common motion walking sequences of subjects S1, S2, S3 for experiments. Since we are dealing with pose estimation from one single view, we use the images (2950 frames) taken from the first camera (C1). The original HumanEva data set is partitioned into training, validation, and test subsets (where the test subset is held out by [4]). For each subject, we use a subset of the training set to train a Gaussian precess regressor [35] and test on a subset of the original validation where both the images and motion capture data are available.

As there are no occluded cases in the original HumanEva data set, we randomly generate two occluding blocks in the test images with various corruption level for synthesizing images with occlusions. The center locations of these blocks are randomly chosen within images and the block widths are correlated with the correction level. The aspect ratio of each block are sampled from a uniform distribution between 0 and 1.

In Fig. 5, we show sample images taken from the walking sequence of three subjects with various corruption levels. The corruption level ranges from 0.1 to 0.6. We can see that although human vision can still infer the underlying poses under occlusion, it is difficult for pose estimation algorithms to handle such test images due to heavy occlusions.

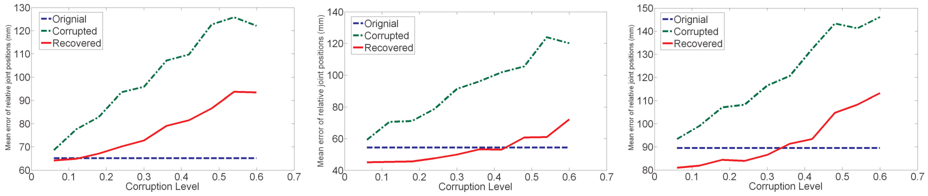


(a) CL=0.1 (b) CL=0.2 (c) CL=0.3 (d) CL=0.4 (e) CL=0.5 (f) CL=0.6

Fig. 5. Synthesized images with occlusions with HumanEva data set I (all walking sequence). Top row: Subject 1, Second row: Subject 2, and Third row: Subject 3. Each corrupted test image contains two randomly generated blocks with their widths equal to corruption level (CL) times original image width and with their centers located at the position from uniformly random sample from image. Each column shows the sample corruption at certain corruption level.

We use histograms of oriented gradients as our feature vectors to represent training and test images. In our experiments, we compute the orientation of gradients in $[0, \pi]$ (unsigned) and construct the histograms using 6 bins in each cell. We use 10×10 pixels per cell, 3×3 cells per block, and uniformly place 3×3 blocks overlapping with neighbor blocks by one cell. Thus, for each image window we obtain a 486-dimensional feature vector. We then learn the mapping function between the feature vectors and their corresponding pose parameters.

We carry out a series of experiments with three different settings: 1) HOG feature vectors from original testing images without synthetically generated occluding blocks, 2) corrupted HOG feature vectors computed from the occluded images (see Fig. 5), and 3) the recovered test feature vectors by solving the extended l_1-norm minimization problem (7). In the third setting, after solving (7), we discard the reconstruction error vector e and use $A\omega$ as our recovered feature vector. All feature vectors obtained in the above three settings are used to regress the pose vector using Gaussian process regressor. We present in Fig. 6 the mean errors of relative joint positions on the testing sub-set of HumanEva data set under various corruption levels (from 0.06 to 0.6). We show the increasing error curves on three settings in terms of joint position error in millimeters of our approach over the baseline (i.e., using HOG feature vectors computed from occluded images.) In all three subjects, we show that from occluded images our approach is able to recover the un-occluded images and then the pose parameters. It is also worth noting that our algorithm also often outperforms the baseline algorithm (trained and tested on un-occluded images). This can be explained by the fact that our algorithm also implicitly performs feature selection whereas the performance of the baseline algorithm is inevitably affected by noise contained in the training data.



(a) S1 (b) S2 (C) S3

Fig. 6. Results of pose estimation on HumanEva data set I in walking sequences. (a) Subject 1. (b) Subject 2. (c) Subject 3. Images in the first row show the 3D mean errors of relative joint position in millimeters (mm) under various corruption level (from 0.06 to 0.6). The blue lines indicate the results from the original test samples, thus the predicted errors are independent of corruption level. The green curves stand for the results from the corrupted test samples with different level of corruption and the red curve are the results from recovered test samples using sparse signal representation.

4.2 Robustness to Background clutter

In this section, we show that the proposed method is able to select relevant feature vectors. We use the same 486 dimensional HOG feature vectors to describe the image observation. We compare two settings: 1) HOG features computed from the original test image sequences, and 2) features extracted from our sparse representation. The mean 3D joint position errors (mm) for each frame are plotted in Fig. 7 for the whole test set. The blue and green error curves correspond to the results using the original HOG feature vectors and the ones extracted from our method, respectively. The improvements (i.e. reduction) of mean position errors (mm) of our method in three subjects are 4.89, 10.84, and 7.87 for S1, S2 and S3, respectively.

5 Conclusion

In this paper, we have presented a method capable of recovering 3D human pose when a person is partially or heavily occluded in the scene from monocular images. By representing the test images as a sparse linear combination of training images, the proposed method is able to recover the set of coefficients from the corrupted test image with minimum error via solving l_1-norm minimization problem, and therefore obtains robust pose estimation results. In addition, our algorithm improves the pose estimation accuracy even on images without occlusions by implicitly selecting relevant features and discarding unwanted noise from background clutter. Our future work includes more experiments with real image data where synchronized ground truth pose parameters and

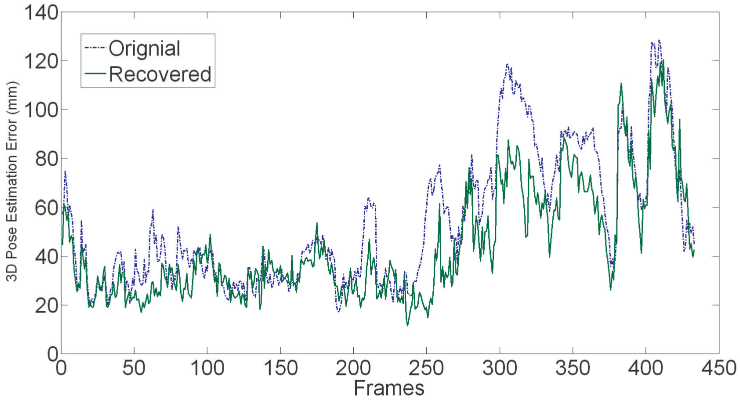


Fig. 7. Mean 3D error plots for the walking sequences (S2). The blue line indicates the errors by using the original test samples. The green line represents the error predicted from recovered feature vectors by the proposed algorithm. The results are comparable or better than the original test samples thanks to the ability of selecting relevant feature entries.

occluded images are available. We also plan to extend our sparse representation algorithm to temporal domain, making use of motion dynamics to further help disambiguate different poses with similar image observations.

References

1. Sigal, L., Isard, M., Sigelman, B., Black, M.: Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In: NIPS. (2004) 1539–1546
2. Grauman, K., Shakhnarovich, G., Darrell, T.: Inferring 3d structure with a statistical image-based shape model. In: ICCV. (2003) 641–647
3. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Discriminative density propagation for 3d human motion estimation. In: CVPR. (2005) 390–397
4. Sigal, L., Black, M.: Predicting 3d people from 2d pictures. In: Proceedings of the Fourth Conference on Articulated Motion and Deformable Objects. (2006) 185–194
5. Bo, L., Sminchisescu, C., Kanaujia, A., Metaxas, D.: Fast algorithms for large scale conditional 3d human pose prediction. In: CVPR. (2008)
6. Agarwal, A., Triggs, B.: A local basis representation for estimating human pose from cluttered images. In: ACCV. (2006) 50–59
7. Elgammal, A., Lee, C.: Inferring 3d body pose from silhouettes using activity manifold learning. In: CVPR. Volume 2. (2004) 681–688
8. Jaeggli, T., Koller-Meier, E., Gool, L.V.: Learning generative models for multi-activity body pose estimation. IJCV **83**(2) (2009) 121–134
9. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Bm^3e : Discriminative density propagation for visual tracking. PAMI **29**(11) (2007) 2030–2044
10. Bissacco, A., Yang, M.H., Soatto, S.: Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In: CVPR. (2007) 1–8
11. Poppe, R.: Evaluating example-based pose estimation: experiments on the HumanEva sets. In: IEEE Workshop on Evaluation of Articulated Human Motion and Pose Estimation. (2007)
12. Okada, R., Soatto, S.: Relevant Feature Selection for Human Pose Estimation and Localization in Cluttered Images. In: ECCV. Volume 2. (2008) 434–445

13. Ning, H., Xu, W., Gong, Y., Huang, T.: Discriminative learning of visual words for 3d human pose estimation. In: CVPR. (2008)
14. Moeslund, T., Granum, E.: A survey of computer vision-based human motion capture. Computer Vision and Image Understanding **81**(3) (2001) 231–268
15. Gavrila, D.: The visual analysis of human movement: A survey. Computer Vision and Image Understanding **73**(1) (1999) 82–98
16. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on Computers **22**(1) (1973) 67–92
17. Felzenszwalb, P., Huttenlocher, D.: Efficient matching of pictorial structures. In: CVPR. Volume 2. (2000) 2066–2073
18. Ronfard, R., Schmid, C., Triggs, B.: Learning to parse pictures of people. In: ECCV. Volume 4. (2002) 700–714
19. Ioffe, S., Forsyth, D.: Probabilistic methods for finding people. IJCV **43**(1) (2001) 45–68
20. Ramanan, D., Forsyth, D.: Finding and tracking people from the bottom up. In: CVPR. Volume 2. (2003) 467–474
21. Mori, G., Ren, X., Efros, A., Malik, J.: Recovering human body configurations: Combining segmentation and recognition. In: CVPR. Volume 2. (2004) 326–333
22. Taylor, C.J.: Reconstruction of articulated objects from point correspondence using a single uncalibrated image. In: CVPR. Volume 1. (2000) 667–684
23. Mori, G., Malik, J.: Estimating human body configurations using shape context matching. In: ECCV. Volume 3. (2002) 666–680
24. Brand, M.: Shadow puppetry. In: ICCV. (1999) 1237–1244
25. Tipping, M.: Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research **1** (2004) 211–244
26. Agarwal, A., Triggs, B.: Recovering 3d human pose from monocular images. PAMI **28**(1) (2006) 44–58
27. Rosales, R., Sclaroff, S.: Learning body pose via specialized maps. In: NIPS. (2001) 1263–1270
28. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: ICCV. (2003) 750–757
29. Candes, E., Romberg, J., Tao, T.: Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory **52**(2) (2006) 489–509
30. Candes, E., Tao, T.: Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Transactions on Information Theory **52**(12) (2006) 5406–5425
31. Donoho, D.: Compressed sensing. IEEE Transactions on Information Theory **52**(4) (2006) 1289–1306
32. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. PAMI **31**(2) (2009) 210–227
33. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge university press (2004)
34. Chen, S., Donoho, D., Saunders, M.: Automatic decomposition by basis pursuit. SIAM Journal of Scientific Computation **20**(1) (1998) 33–61
35. Rasmussen, C.E., Williams, C.K.I.: Gaussian processes for machine learning. MIT Press (2006)