

Human Pose Estimation with Regression by Fusing Multi-View Visual Information

Xu Zhao, Yun Fu, *Member, IEEE*, Huazhong Ning, *Member, IEEE*, Yuncai Liu *Member, IEEE*,
and Thomas S. Huang, *Life Fellow, IEEE*

Abstract—We consider the problem of estimating 3D human body pose from visual signals within a discriminative framework. It is challenging because there is a wide gap between complex 3D human motion and planar visual observation, which makes this a severally ill-conditioned problem. In this paper, we focus on three critical factors to tackle human body pose estimation, namely, feature extraction, learning algorithm and camera utilization. On the feature level, we describe images using the salient interest points represented by SIFT-like descriptors, in which the position, appearance, and local structural information are encoded simultaneously. On the learning algorithm level, we propose to use Gaussian processes and multiple linear regression to model the mapping between poses and features. Fusing image information from multiple cameras in different views is of great interest to us on the camera level. We make a comprehensive evaluation on the HumanEva database and get two new insights into the three crucial issues for human pose estimation: (1) Although the choice of feature is very important to the problem, once the learning algorithm becomes efficient, the choice of feature is no longer critical; (2) The impact of information combination from multiple cameras on pose estimation is closely related to not only the quantity of image information, but also its quality. In most cases it's true that the more information is involved, the better results can be achieved. But when the information quantity is the same, the differences in quality will lead to totally different performance. Furthermore, dense evaluations demonstrate that our approaches are an accurate and robust solution to the human body pose estimation problem.

Index Terms—Human pose estimation, Gaussian processes regression, multiple views, image feature.

I. INTRODUCTION

HUMAN body pose estimation from visual signals has long been an active research topic in the computer vision society, especially for the past two decades. As one of the most common pieces of content in visual media, human motion carries a lot of meaningful information for social communication between humans and interactions between human

This work was funded in part by the China National 973 Program of 2006CB303103 and in part by the China NSFC Key Program of 60833009. This research was also funded in part by the U.S. Government VACE program and the NSF Grant CCF 04-26627. The views and conclusions are those of the authors, not of the US Government or its Agencies.

Xu Zhao and Yuncai Liu are with the School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (email: zhaoxu@sjtu.edu.cn; whomliu@sjtu.edu.cn).

Yun Fu is with the Department of Computer Science and Engineering, University at Buffalo (SUNY), 201 Bell Hall Buffalo, NY 14260, USA (email: raymondyunfu@gmail.com).

Huazhong Ning and Thomas S. Huang are with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA (email: hning2@uiuc.edu; huang@ifp.uiuc.edu).

and computer. A wide spectrum of potential applications [1] such as behavior understanding, content-based image retrieval, visual surveillance, rehabilitation engineering, and humanoid robotics motivate endeavors to find robust solutions to this problem. However, recovering human pose, especially, 3D pose, from planar visual information is extremely challenging due to the complicated nature of human motion and limited available information in 2D images.

In general, the state-of-the-art technologies for human pose estimation can be summarized as two categories: *generative* and *discriminative* [2]. Generative methods [3]–[7] follow the prediction-match-update philosophy embedded into the framework of bottom-up Bayes' rule and model the state posterior density using the observation likelihood or a cost function. This class of methods can handle unknown and complex motions but suffer from the expensive computation cost for the unavoidable search in high-dimensional state space. Discriminative methods [2], [8]–[11] model the state posterior distribution conditioned on observations directly. The models are constructed usually by finding the direct mapping from the image feature space to the pose label space based on training samples. Once the training process is completed, pose estimation will be computationally effective. In this paper, we choose to focus on estimating 3D human pose within the discriminative framework.

A. Discriminative Human Pose Estimation

The critical pose estimation problem typically utilizes redundant sensory inputs, e.g. images, to capture valid pose information. A general discriminative pose estimation system is mainly constrained by three aspects: *feature extraction*, *algorithm*, and *camera utilization*.

Many existing discriminative methods [2], [8], [10], [12], [13] extract image features from human body silhouettes. These kinds of features have the advantage of containing strong shape cues for pose estimation while being invariant to appearance and lighting. The silhouettes could be conveniently extracted by simple background subtraction. However, these methods are mainly applicable to the discrete pose case with large pose intervals between labels, because the information loss of interior appearance may introduce one-to-many ambiguities to the mapping from silhouettes to poses. Such multi-modal ambiguity of state posterior distribution is one of the main error sources of pose estimation. Intuitively, this problem can be alleviated more or less by effectively utilizing the interior appearance information. This belief was proven

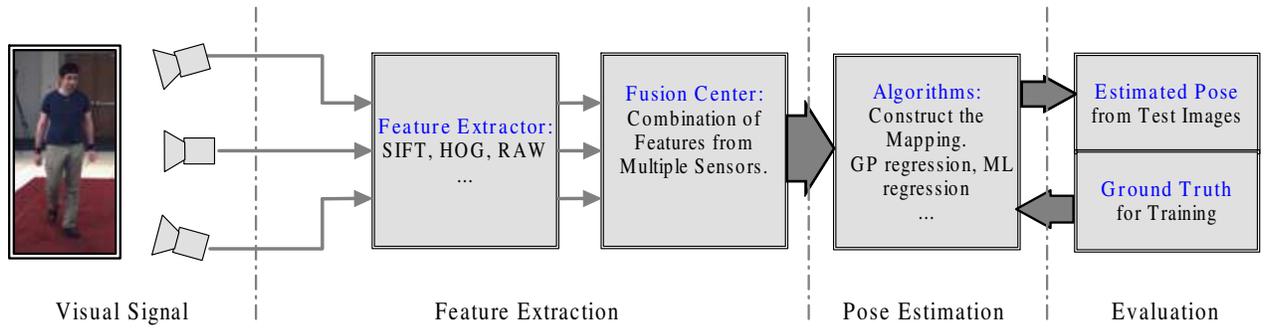


Fig. 1: Framework of human pose estimation by fusing visual information from multiple views. The whole framework consists of three main modules: feature extraction, pose regression and evaluation.

by experiments in a few recent works [14]–[17]. However, sometimes this ambiguity cannot be entirely resolved, even if the image descriptor is perfect and has no loss of any local details when only monocular image information is used. For such a highly non-linear and severely ill-conditioned problem, introducing multi-view visual information is a radical way to further enhance the performance of pose estimation. Most of the existing works [18]–[23] using multi-view information use a generative framework. The approaches of utilizing multi-view information mainly include photogrammetric techniques [20], [21], [23] and integrating the multi-view information into the computation of likelihood functions [18], [19], [22]. Derived from our previous work [24], we propose to develop a discriminative framework for body pose estimation by fusing multi-view camera inputs.

We describe the image features using the salient interest points, represented by the SIFT descriptor [25]. The bag-of-words representation is used to model the distribution of feature space. This sparse and local image descriptor attempts to not only capture the spatial co-occurrence and context information of the local structure, but also encode their relative spatial positions. These properties make the descriptor discriminative for the task of pose estimation. The descriptor also tolerates a range of illumination and position variations because it is computed on overlapped patches, instead of pixels. After extracting the image features, the fusing strategy is straightforward. We concatenate the feature vectors from the multiple views together as the complete representation of the visual signal. We also extract raw features, namely, the original image pixel intensity as the comparative benchmark feature.

After extracting discriminative features from multiple views, modeling the mapping between pose label space and feature space becomes more important. The approaches to solving this problem varies from neural networks [26], fast nearest-neighbor retrieval [5], [27], regression methods [8], [28], to Bayesian mixture of experts [2], [14]. Our method is based on the non-parametric Gaussian processes (GP) regression. GP is flexible, fully probabilistic, and effective to deal with the small-sample-size problem. Such properties are desirable to pose estimation. As a contrast of GP regression, multiple linear (ML) regression is also tested in our work.

Extensive evaluations on the HumanEva database [29] are

provided across all the three aspects of the pose estimation system. In multi-view scenarios, it's significant to find how the quantity and quality of image information cast impacts on the pose estimation performance. We present comparative research on different combination of multiple views. The comparison between GP and ML regression algorithms actually demonstrates the difference in efficiency between non-linear non-parametric and linear parametric algorithm, for the problem of pose estimation. Both feature extraction and choice of algorithm is crucial, but comprehensive experiments give some interesting insights into the situation when effective algorithms dominate the system performance for a different choice of features.

B. System Framework and Contributions

As illustrated in Fig. 1, our whole framework is composed of four main parts. The second part is focused on feature extraction. After the images captured by multiple cameras are imported in the first part, the system extracts the features for each camera respectively. According to the demands of evaluation, we can form the features by combining a different number of cameras in the fusion step. In the training process, the combined image features are input into the algorithm module. The parameters of the algorithm for estimating human pose are learned by using the ground truth. Once the training process is completed, given a test image, the system estimates the pose by directly applying the learned parameters. These steps are completed mainly in the first three parts. The last part serves for the performance evaluation.

The contributions of the paper are fourfold:

- 1) We develop a discriminative 3D pose estimation framework in a systematic way, in which three critical factors, feature extraction, regression algorithm, and camera utilization, are jointly considered.
- 2) We design a novel corner-interest-point-based SIFT (CP-SIFT) features, in which the body position, appearance, and local structural information are encoded simultaneously. The Gaussian process regression is exploited to build the mapping from visual feature space to pose label space. This feature descriptor and regression algorithm are demonstrated to be sufficiently effective and robust in the realistic evaluations.

- 3) We extend our previous pose estimation work [24] from single view input to multiple views input, which brings satisfying performance improvement.
- 4) We conduct comprehensive experimental studies on the HumanEva database using our proposed framework. We obtain some interesting insights into the impacts of feature, regression algorithm and information fusion of multiple views on the performance of the system. It can provide useful guidance for the system design.

The rest of this paper is organized as follows. In section II, we briefly introduce the visual features used in our pose estimation system. The regression algorithms, Gaussian process and multiple linear regression, and their application in the framework are described in section III. In section IV, extensive experiments, evaluation results, method comparisons, and case-dependent analysis are presented. Finally, we conclude the paper with some interesting and useful insights and future directions in section V.

II. FEATURE EXTRACTION

Image-based body pose regression heavily relies on an efficient feature extraction algorithm. The silhouettes or contours of human body contain strong shape cues for pose estimation and are invariant to appearance and lighting variations. However, the appearance information within the human body cannot be simply neglected because adjacent poses could get ambiguous from each other by pure shape, contour, or silhouette. The feature representation should be designed to contain interior body appearance information, which is sensitive to the subtle change of human pose. The local structures and interior relative positions of body parts also play important roles in determining the pose labels for most ambiguous cases. Under these considerations, we design a specific descriptor to specify the following feature extraction procedure.

- 1) *Human detection.* The background subtraction is used to determine the bounding window for human detection in each input image. This bounding window is then rescaled to a fixed size.
- 2) *Interest point detection.* Within the bounding window, the Harris corner detector [30] is used to detect interest points. Fig. 2a shows the example of interest points labeled on an image frame. The background subtraction here can slightly improve the performance of interest point detection.
- 3) *SIFT feature extraction.* The SIFT descriptor [25] is applied at each interest point, which is denoted as a vector \mathbf{p} .
- 4) *CP-SIFT feature representation.* Find the relative coordinate (u, v) of each corner interest point. The final descriptor of each interest point is represented as $\mathbf{d} = (u, v, \mathbf{p})^T$.

We call the feature as CP-SIFT feature because it is a SIFT like feature based on corner interest points with *position* information. The combination of SIFT descriptor and Harris corner points with position information is one of our contributions in this work. The feature is scale invariant and partially illumination invariant due to the introduction of Harris corner

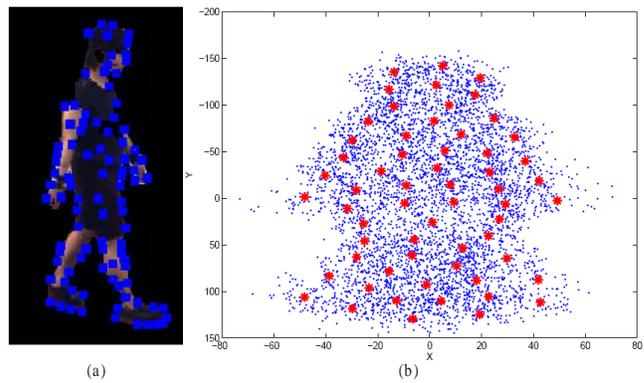


Fig. 2: Feature extraction. (a) Interest points on an image frame. (b) The relative coordinates of interest points from all images (marked as “.”) and the visual words (marked as “*”).

detector and SIFT descriptor. Specifically, the local region around each interest point is first partitioned into nine cells, and a nine-orientation histogram is calculated on each cell. In total the descriptor vector has 83 dimensions, including the dimension of relative coordinates of the interest point. Technically, eliminating the left-right ambiguity of human body motion is crucial to the accuracy of pose estimation. The proposed descriptor encodes the appearance, edge, and position information into a vector. In doing so, the multi-modal ambiguities of posterior pose distribution can be alleviated to a large extent. Actually, the CP-SIFT feature in which the position information is encoded, is a variation of SIFT. This idea is inspired by the previous work [31] especially by our previous “X-Y patch” work [32], which demonstrates the importance of local feature coordinates when pose variation is distinct. It is meaningful for body pose estimation because in nature the human body is a hierarchical structure with fixed relative connections between different body parts.

After we calculate all the local descriptors, the unsupervised bag-of-words model [33] is used to represent the distribution of visual feature space. The descriptors extracted from all training images are clustered by K -means. The K cluster centers, called visual words, form a code book $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K\}$. In this paper, the number of visual words is empirically set as 60 in the experiments. Once the code book is available, each descriptor in the descriptor set $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_m\}$ of a given testing image votes softly with respect to the visual words by calculating the distances. The bag-of-words representation, denoted as \mathbf{x} , is the accumulating scores of all descriptors on the K visual words. Each image finally is represented as a K -dimensional feature vector. In Fig. 2b, the relative coordinates of interest points and the visual words are displayed as an example, where the relative coordinates are calculated by subtracting the coordinate mean of all the interest points.

To test how and to what extent the choice of image feature can impact pose estimation, we also extract the raw features (appearance) in which the original image pixel intensities are kept. In our work, these raw features are used as a baseline feature for the comparisons with CP-SIFT.

As for multiple view visual data input, we use the simplest fusion method by concatenating the feature vectors of all the synchronized views. According to our previous work [34], a more sophisticated fusion method can be adopted. To demonstrate the advantage of multi-view feature fusion, we will compare the performances of single view and multiple views in the experiments.

III. POSE REGRESSION

In this section, we introduce the regression methods for estimating human pose from image features proposed in the foregoing section. We denote the pose label vector as $\mathbf{y} \in \mathbb{R}^d$ and the image feature vector as $\mathbf{x} \in \mathbb{R}^K$. Both GP regression and ML regression will be evaluated to estimate 3D human pose.

A. Non-Parametric Regression: Gaussian Process Regression

Gaussian process (GP) [35] is the generalization of Gaussian distributions defined over infinite index sets. It can be used to specify a distribution over functions. Given a training sample set $\mathcal{S} = \{(\mathbf{x}_i, y_i), i = 1, \dots, N\}$ where (\mathbf{x}_i, y_i) is an image-to-pose pair and y_i 's are the components of \mathbf{y}_i which are normalized to be zero-mean unit variance process. Suppose the relationship between \mathbf{x}_i and y_i is modeled by regression

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad (1)$$

where $\epsilon_i \sim (0, \xi^{-1})$ denotes noise and hyper-parameter ξ represents the precision of the noise. Define a GP prior over functions f_i , we have

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K}), \quad (2)$$

where $\mathbf{f} = [f_1, \dots, f_N]^T$ is the function values, $f_i = f(\mathbf{x}_i)$, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$, and \mathbf{K} is a covariance matrix whose entries are given by a covariance kernel function, $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. Here, we choose the kernel function as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \theta_0 \exp \left\{ -\frac{\theta_1}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right\} + \theta_2 + \theta_3 \mathbf{x}_i^T \mathbf{x}_j. \quad (3)$$

For an unseen observation \mathbf{x}_{N+1} , the joint distribution is therefore written as

$$p(\mathbf{Y}_{N+1}) = \mathcal{N}(\mathbf{Y}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1}), \quad (4)$$

where $\mathbf{Y}_{N+1} = [y_1, \dots, y_N, y_{N+1}]^T$ and the covariance matrix \mathbf{C}_{N+1} is given by

$$\mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^T & c \end{pmatrix}. \quad (5)$$

\mathbf{C}_N has elements

$$C(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \xi^{-1} \delta_{ij}, \quad (6)$$

where δ_{ij} is the Kronecker delta function, the vector \mathbf{k} has elements $k(\mathbf{x}_n, \mathbf{x}_{N+1})$ for $n = 1, \dots, N$, and the scalar $c = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \xi^{-1}$.

During training, the hyper-parameters $\Theta = \{\theta_0, \dots, \theta_3, \xi\}$ of GP are learned by minimizing

$$-\ln p(\mathbf{X}, \Theta|\mathbf{Y}) = \frac{1}{2} \ln |\mathbf{C}_N| + \frac{1}{2} \mathbf{Y}^T \mathbf{C}_N^{-1} \mathbf{Y} + r, \quad (7)$$

where $\mathbf{Y} = [y_1, \dots, y_N]^T$ and $r = N \ln(2\pi)/2$ is a constant. Once the GP model is learned, the conditional distribution $p(y_{N+1}|\mathbf{Y}, \mathbf{X})$ is a Gaussian distribution with mean and covariance given by

$$\mu(\mathbf{x}_{N+1}) = \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{Y}, \quad (8)$$

$$\sigma^2(\mathbf{x}_{N+1}) = c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k}. \quad (9)$$

New test samples can be easily and efficiently inferred by (8) and (9).

B. Parametric Regression: Multiple Linear Regression

To evaluate the efficiency of different regression algorithms on the pose estimation task, we take the multiple linear regression model [36] as a comparative baseline in the algorithm level. The ML regression model can be formulated as

$$\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{e}, \quad \text{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}, \quad (10)$$

where \mathbf{Y} is the joint angle vector over all training samples. $\tilde{\mathbf{X}}$ is the design matrix whose columns are the model terms evaluated at the components of image feature vector. In this model, the elements of the first column in $\tilde{\mathbf{X}}$ are all 1's for the intercept and the other columns include linear terms and pure-quadratic terms. The vector $\boldsymbol{\beta}$ encodes the regression coefficients that need to be estimated during the training process. The error vector \mathbf{e} consists of zero mean and independent random variables with common variance σ^2 . To fit the model to the data, $\boldsymbol{\beta}$ can be estimated by ordinary least squares (OLS) $\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}' \mathbf{Y}$ [36]. But the normal equations are often badly conditioned relative to the original system. So the orthogonal decomposition of $\tilde{\mathbf{X}}$ is used to find the solution.

Once the regression model is trained, we have

$$\hat{\mathbf{Y}} = \tilde{\mathbf{X}} \hat{\boldsymbol{\beta}} \quad (11)$$

with

$$\hat{\mathbf{Y}} = [\hat{y}_1 \dots \hat{y}_N]^T, \quad \hat{\boldsymbol{\beta}} = [\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\beta}_2]^T, \\ \tilde{\mathbf{X}} = [\mathbf{1}_{N \times 1} \quad [\mathbf{x}_1 \dots \mathbf{x}_N]^T \quad [\mathbf{x}_1^2 \dots \mathbf{x}_N^2]^T],$$

where \hat{y}_i represents the estimated joint angle for the image feature \mathbf{x}_i , $\hat{\beta}_0$ is the learned intercept term, $\hat{\beta}_1, \hat{\beta}_2 \in \mathbb{R}^m$ are the learned parameter vectors, and \mathbf{x}_i^2 is the array-wise square of \mathbf{x}_i .

IV. EXPERIMENTS

In the experiments, we aim to find the intrinsic relationships between the accuracy of pose estimation and image features, regression algorithm and the utilization of information from multi-views. All the experiments are conducted on the publicly available HumanEva dataset [29] for the evaluation of human pose estimation, collected at Brown University.

TABLE I: Average RMS error (in degree) over all joint angles, all subjects for action walking, box, jog, and gestures.

	CP-SIFT Feature				SIFT Feature				Raw Feature			
	Walking	Box	Jog	Gestures	Walking	Box	Jog	Gestures	Walking	Box	Jog	Gestures
ML-Regression	7.0365	7.9200	4.0853	7.7505	7.3596	8.1303	4.2951	7.9723	7.9774	9.6856	4.7837	8.8647
Ridge Regression	7.1249	7.8601	3.8912	6.9338	7.5327	9.3216	4.3094	6.8351	8.2553	8.3954	4.9216	7.6933
GP-Regression	6.0934	4.7904	3.7766	4.5056	6.4211	5.2236	4.1923	4.4981	6.9798	5.2662	4.2656	4.7938

A. Database

The HumanEva dataset was captured simultaneously using a calibrated marker-based motion capture system and multiple high-speed video capture systems. The video and motion capture streams were synchronized. It contains multiple subjects performing a set of predefined actions with repetitions.

The original motion capture data provided by HumanEva were (x, y, z) locations of the body parts in the world coordinate system. There is a total of ten parts: torso, head, upper and lower arms (left and right), and upper and lower legs (left and right). In this paper, we convert the (x, y, z) locations to global orientation of torso and relative orientation of adjacent body parts. Each orientation is represented by three Euler angles. We have in total 26 whole body degrees of freedom by discarding the coordinates that have a constant value in the performed motions. The set of joint angle trajectories is normalized to be a zero-mean unit variance process.

The HumanEva dataset was originally partitioned into training, validation, and test sub-sets. We use sequences in the training sub-set for training and those in the validation sub-set for testing. The original test sub-set is not used because there is no motion data provided for it. A total of 2950 frames (first trial of subjects S1, S2, and S3) for walking motion, 2345 frames for jog motion, 2486 frames for box motion and 2850 frames for Gestures motion are used. The image information we used is from camera C1, C2 and C3.

B. Evaluation: Feature and Regression Algorithm

Theoretically, the accuracy of pose estimation is closely related to the choice of image feature and regression algorithm. To evaluate the impacts of both factors on pose estimation, we test the pose regression on CP-SIFT feature, SIFT feature and raw feature. We choose GP and ML regression as the regression algorithms. Furthermore, to verify the efficiency of GP regression and make the further comparisons between two class algorithms (non-parametric non-linear and parametric linear regression algorithm) on pose estimation problem, we also use the ridge regression algorithm in our experiments. For multi-view visual data, we choose to use the combination of camera C1 and C2 since similar results and conclusions can be obtained from other camera combinations. For the raw feature, we reduce the dimensionality to 100 with PCA after fusion.

In Table I, we report the mean (over all joint angles) RMS absolute difference errors [8] between the ground-truth and estimated joint angles, in degrees.

$$D(\mathbf{y}, \mathbf{y}') = \frac{1}{d} \sum_{i=1}^d |(y_i - y'_i) \bmod \pm 180^\circ|. \quad (12)$$

From the table, we can see that the performance of GP regression largely outperforms ML regression and ridge regression for the three features. The ML regression and ridge regression get close performance and the performance difference is statistically insignificant. This results demonstrate the efficacy of non-parametric non-linear regression algorithm on pose estimation.

We also can find that the performance of the CP-SIFT feature is significantly better than that of the raw feature for the three regression algorithms. The performance of SIFT feature is lower than CP-SIFT but better than raw feature.

In Fig. 3, we compare the performances of GP and ML regression on the CP-SIFT and raw features. Because the performance of ridge regression is very close to that of ML regression, we don't show it in the figures to save space. Actions shown in the figure are walking and box, which are the representative actions for moving around and standing at a fixed place respectively. The mean and standard derivation of RMS error over all the 26 joint angles, normalized by the range of variation, are reported respectively. It can be seen that GP regression achieves superior performance for both features by mean and standard derivation. And the superiority of GP over ML is much apparent for raw features than the CP-SIFT feature. In Fig. 4, the estimations and ground truth of two joint angles in walking and box actions are plotted respectively. The curves of estimation with GP regression are closer to the ground truth and smoother than that of ML regression although there exist jitters in some segments.

We also compare the relative errors of individual angles in Fig. 5 on the feature level. Similar to Fig. 3, walking and box actions are selected to show in the figure respectively. As shown in Fig. 5a, 5c, for GP regression, the superiorities of CP-SIFT feature over raw feature and SIFT feature are small. However, as we can see in Fig. 5b, 5d, this superiority is salient for ML regression. It is an interesting observation for us, which is consistent with that indicated by the data in Table I. It demonstrates that the performance difference in algorithms is much bigger than that in features. In other words, the choice of regression algorithm plays a more important role than the choice of feature in this problem. We will discuss this further in section IV-D.

C. Evaluation: Multiple Views

To evaluate the relationship between the quantity and quality of image information and pose estimation, we conduct the experiments combining information from multiple views. The combination strategy is simple. We concatenate the feature vector from each single camera together as the complete representation of the visual signal. To avoid overfitting, we

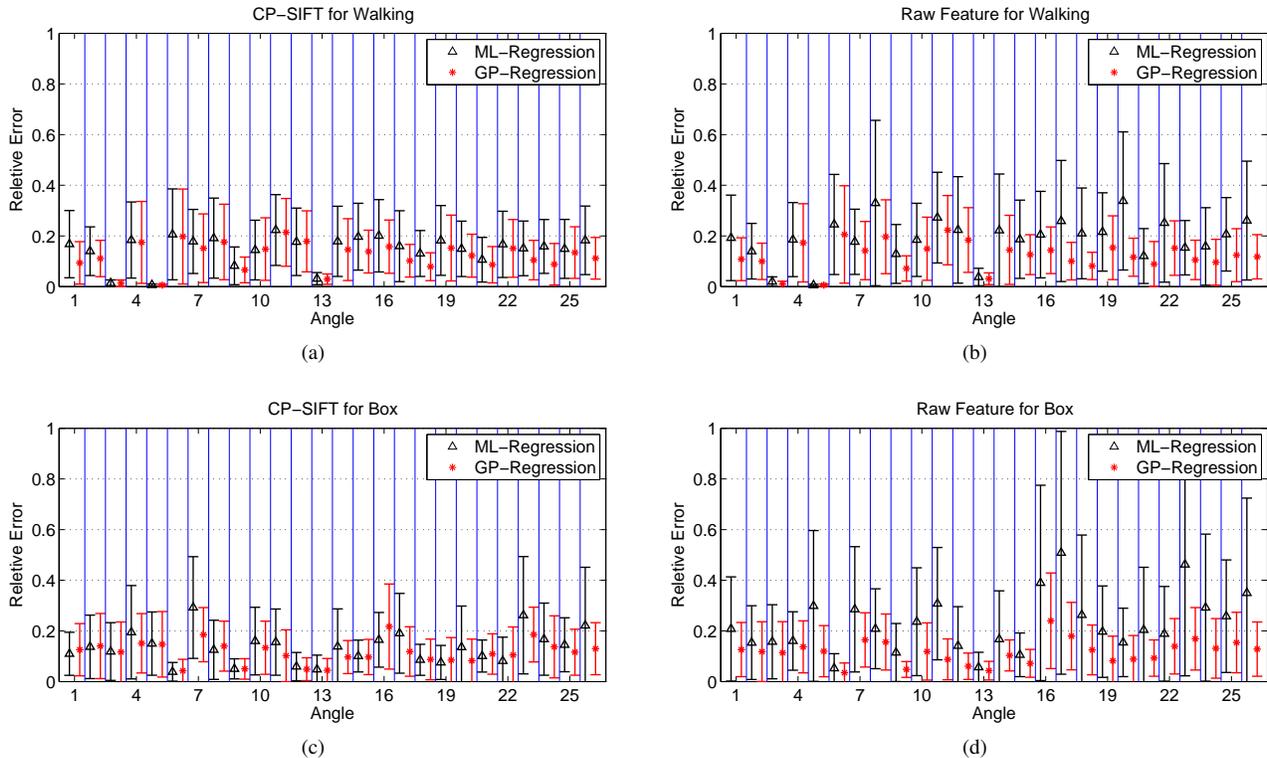


Fig. 3: Performance comparison between GP and ML regression on (a), (c) CP-SIFT feature and (b), (d) raw feature. Here, (a) and (b) are for the walking action, (c) and (d) are for the box action. Both mean and standard deviation of RMS error over all the individual joints, normalized by the range of that joint variation, are reported.

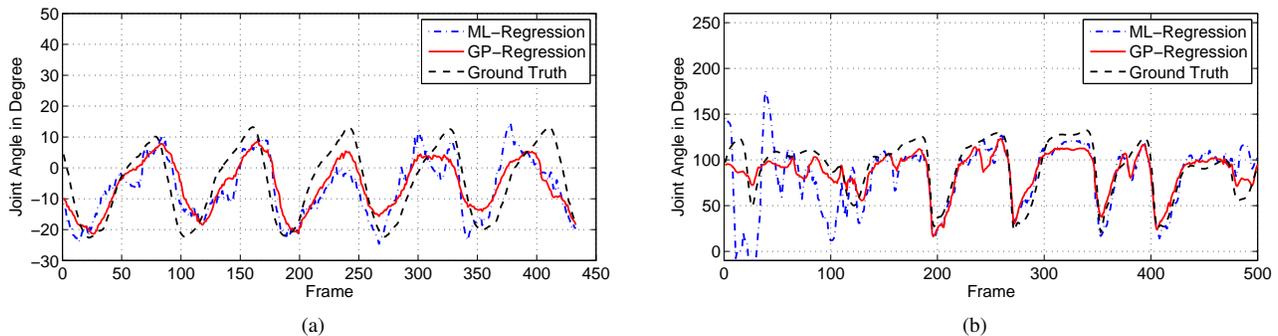


Fig. 4: Curve comparisons of joint angles: ground truth and estimations with GP and ML regression. (a) Right hip (x-axis) of subject S2 in walking. (b) Right elbow of subject S3 in boxing.

reduce the dimension of the concatenated feature vector to 100 with PCA. In general, at this point, over 95% of the data variance can be kept. In the experiments, there are in total seven camera combinations considered. We represent these combinations as C1, C2, C3, C1-C2, C1-C3, C2-C3 and C1-C2-C3.

The relative errors of all the joint angles with different camera combinations are shown in Fig. 6, which are averaged on all the three subjects (S1, S2, S3). For the one view scenario, the errors are averaged on C1, C2, and C3. For the two views scenario, the errors are averaged on the combinations of C1-C2, C1-C3, and C2-C3. The GP regression algorithm is used

to combine with raw feature and CP-SIFT feature respectively. Because the performance comparison between different view combinations is closely related to the style of action, we make the comparisons on two classes of actions: people moving around and people standing in a fixed place. For the moving around actions, such as walking and jog, the contributions of different views are roughly similar over the whole sequence. But for the actions with small global motion with respect to the cameras, such as box, the viewpoint of different cameras is quite different, just as shown in Fig. 7b. From Fig. 6, we can see that when more views are combined, more accurate results of pose estimation can be achieved. It is reasonable because

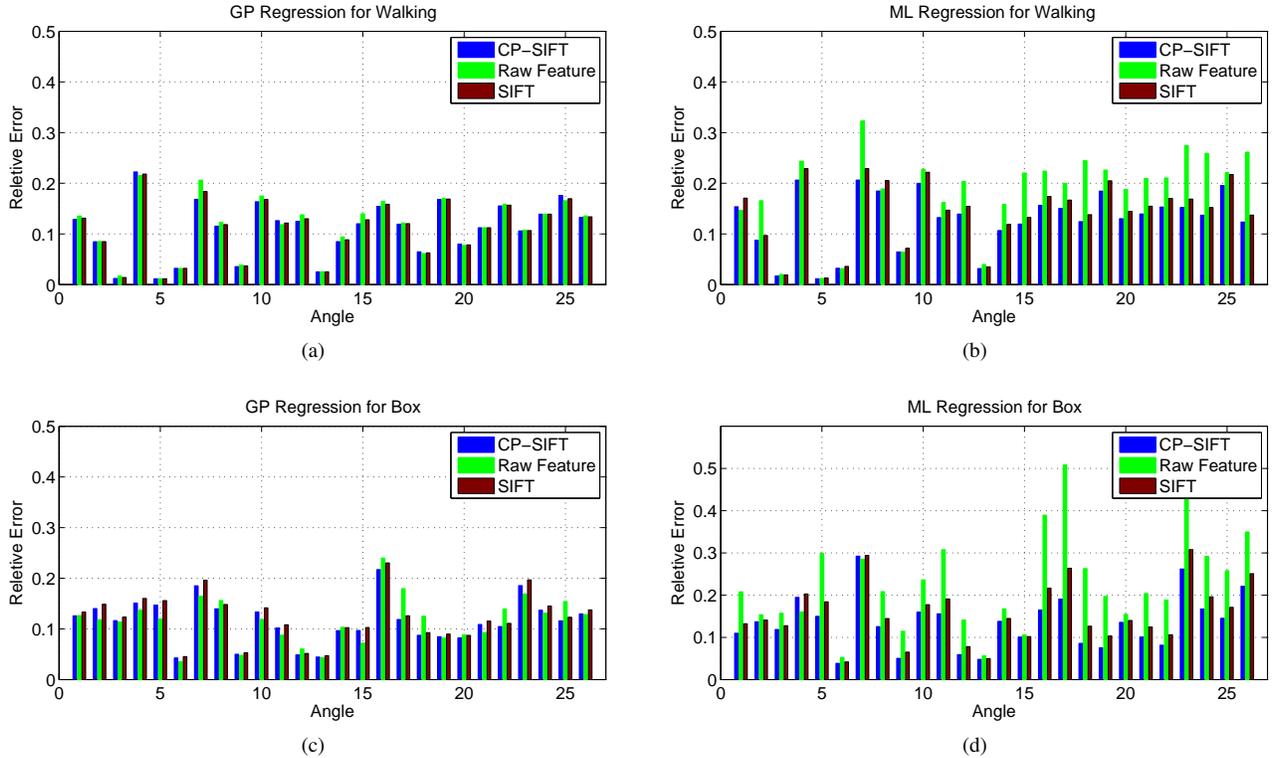


Fig. 5: Performance comparison between CP-SIFT feature, SIFT feature and raw feature for (a), (c) GP regression and (b), (d) ML regression. Here, (a) and (b) are for the walking action, (c) and (d) are for the box action. RMS error of each individual angle, normalized by the range of that angle variation, is reported.

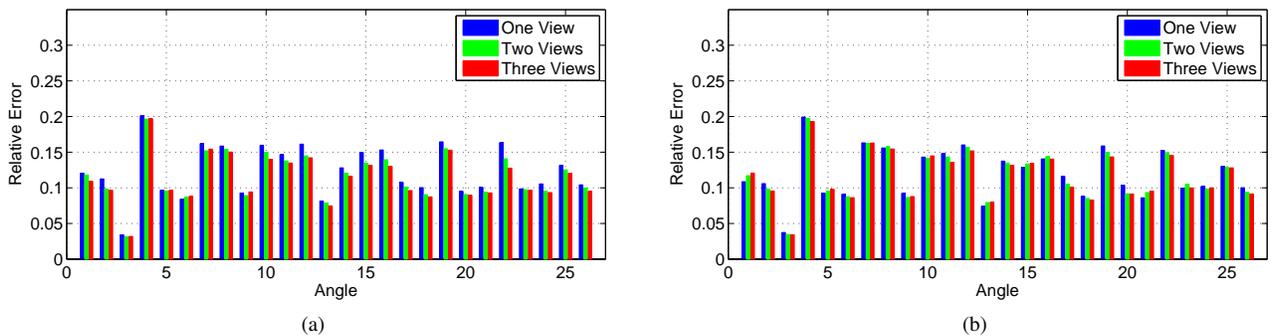


Fig. 6: Performance comparison between different view combinations. Average relative errors on all the walking and jog sequences for one view, two views and three views combination. The algorithm and features used here are (a) GP and RAW feature, (b) GP and CP-SIFT feature.

when the information from multiple cameras is involved into the process of pose estimation, the ill-condition of this problem is alleviated. The improvement is due to the contribution from the quantity of image information. Fig. 8 shows some sample image frames of camera C1, on which the ground truth and estimated pose represented as the outline of a cylinder based human model are superimposed.

Another interesting observation about the multi-view combination is related to the quality of image information. Fig. 7a shows the relative errors of three combinations of two views for all the box sequences. It can be seen that the performance

of C1-C3 combination outperforms that of the C1-C2 and C2-C3 combinations for most of the joint angles. Fig. 7b shows the sample images from three single views. We can see that the camera C1 captures the frontal view of box action and the other two, C2 and C3, capture the side view of the action. Therefore, the combination of C1-C2 and C1-C3 can get better results than the combination of C2-C3. And, the action is more observable to camera C3 than camera C2 because camera C3 can capture some part of the frontal view. This is the reason why C1-C3 combination has some superiority over the combination of C1-C2.

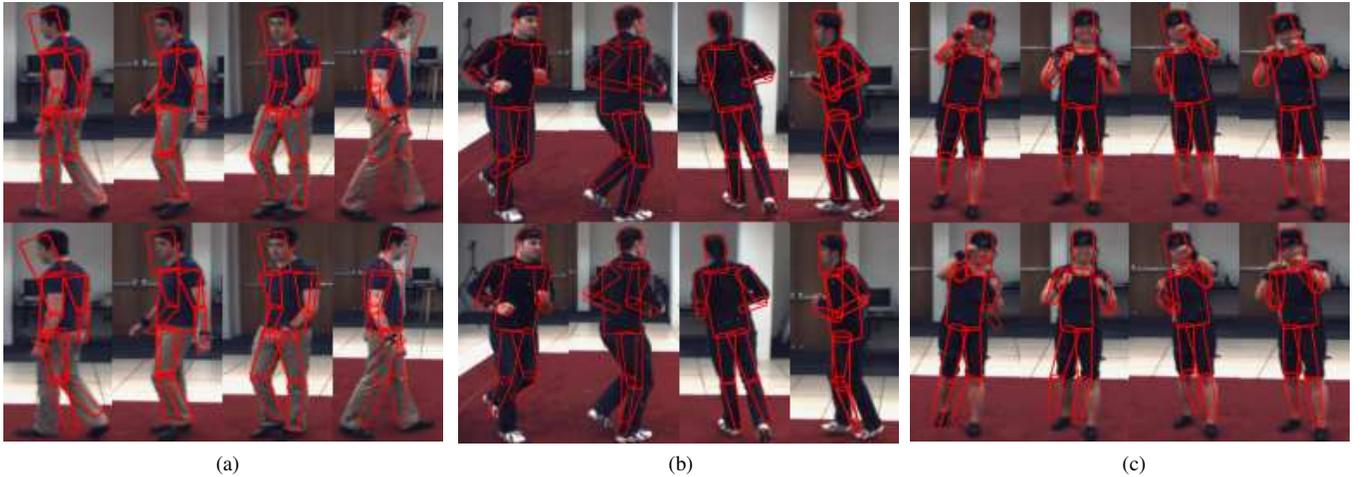


Fig. 8: Some of the sample estimation results for action (a) Walking (b) Jog and (c) Box. The first row shows the provided ground truth projected onto the camera C1, and the second row shows the estimated pose projected onto the same camera. Each column corresponds to a frame.

TABLE II: Average RMS error (in degree) over all joint angles, all subjects for walking, box, jog, and gestures actions. The evaluation is for the all-round comparisons of algorithm, feature and view combination.

		GP-Regression				ML-Regression			
		Walking	Box	Jog	Gestures	Walking	Box	Jog	Gestures
One View	CP-SIFT Feature	6.2781	5.3892	3.7319	4.7831	7.0436	9.7649	4.1527	8.6803
	Raw Feature	6.0982	5.4391	4.0843	4.9828	8.6213	11.4148	5.1201	12.5518
Two Views	CP-SIFT Feature	6.0584	4.8527	3.8153	4.6328	7.0384	6.3836	4.1061	7.6316
	Raw Feature	7.4333	5.6369	3.8751	4.8845	7.7915	9.9176	4.5428	10.6549
Three Views	CP-SIFT Feature	6.0564	4.8101	3.7297	4.4676	6.9867	6.2322	4.0575	8.3205
	Raw Feature	6.0199	5.3332	3.8287	4.8977	8.0366	9.6856	5.1291	8.8647

D. Discussion

In the presented evaluations, we evaluated how and to what extent the three critical factors, feature extraction, regression algorithm, and multi-view utilization impacts on the problem of pose estimation within the discriminative framework. More details about the all-around comparisons of the three factors are presented in Table II.

We found in the evaluation of feature versus regression algorithm that, as the representation of visual signals, the choice of feature has important effects on the accuracy of pose estimation (see Fig. 5 and Table I). However, compared to the regression algorithm, feature is not the most important factor. This conclusion is validated in our experiments for the problem of pose estimation. Actually, from Fig. 5, Table I and Table II, we can see that if the regression algorithm is not powerful enough, the effect of the choice of feature will be remarkable. But once the regression algorithm performs better, e.g. in our work the GP regression is used, the performance difference between features is reduced dramatically. So, this phenomena indicates that an effective algorithm can extract more useful information from any features and dominate the whole system performance.

Considering the view combination, it is intuitively believed that more views will provide more information and more accurate pose estimation performance. However, in the

evaluation, we found that sometimes the situation is more complicated. The final results depend not only on the quantity of information but also its quality. From Table II, we can see in most cases it's true that when more visual information is involved, the performance is much better. However, there exist some cases violate this belief. For the outliers, some of the combined information may introduce unexpected noise to the feature extraction module. On the other hand, the importance of information quality is well demonstrated in Fig. 7a and Fig. 7b, where the information quantity is the same, but the difference in quality leads to totally different performances.

V. CONCLUSION

We have presented methods to solve the human body pose estimation problem in a discriminative framework. Our interests are in finding out not only the state-of-the-art solution to this problem, but also the impact of the three critical factors, namely, regression algorithm, feature extraction and camera utilization on the problem. We made comprehensive evaluations on the HumanEva database and got some interesting insights into the relationship of these crucial issues. In the feature extraction module, we introduced our new CP-SIFT feature in which the position, appearance, and local structural information are all captured and encoded. The efficiency of the CP-SIFT feature has been demonstrated by the compar-

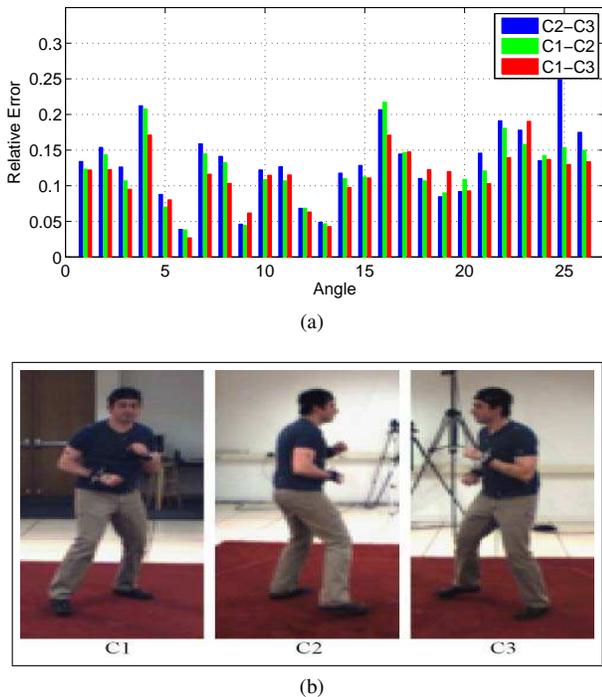


Fig. 7: Multi-view comparison. (a) Average relative errors on all the box sequences for different two views combination C1-C2, C1-C3, C2-C3. (b) Sample images from camera C1, C2 and C3 for the box action.

ison to raw features in the evaluations. For the regression algorithm, GP regression, as we chose, showed remarkable superiority over ML regression. By the evaluation, we found that although the choice of feature is very important, but when an efficient regression algorithm is chosen, it is no longer critical. We noticed specially that this observation is fairly consistent with the finding in recent works on sparse representation [37]. Another interesting observation is about the information fusion of multiple views. In the process of pose estimation, before fusing multi-view information, one had to consider the important roles of both quality and quantity of image information at the same time. For the future work, we plan to explore more sophisticated fusing strategies from our recent work [34] on the multi-view feature combination. The theoretical explanation of our evaluation results is also of our major interest.

ACKNOWLEDGMENT

The authors would like to thank Brown University for providing the HumanEva database [29].

REFERENCES

- [1] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.
- [2] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Discriminative density propagation for 3d human motion estimation," *CVPR*, pp. 217–323, 2005.
- [3] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," *CVPR*, 2000.
- [4] H. Ning, T. Tan, L. Wang, and W. Hu, "People tracking based on motion model and motion constraints with automatic initialization," *Pattern Recognition*, 2004.
- [5] G. Mori and J. Malik, "Recovering 3D Human Body Configurations Using Shape Contexts," *PAMI*, 2006.
- [6] M. Lee and I. Cohen, "A model-based approach for estimating human 3D poses in static images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 6, pp. 905–916, 2006.
- [7] X. Zhao and Y. Liu, "Generative tracking of 3D human motion by hierarchical annealed genetic algorithm," *Pattern Recognition*, vol. 41, no. 8, pp. 2470–2483, 2008.
- [8] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *PAMI*, 2006.
- [9] M. Brand, "Shadow puppetry," *ICCV*, vol. 2, p. 1237, 1999.
- [10] A. Elgammal and C. Lee, "Inferring 3D body pose from silhouettes using activity manifold learning," *CVPR*, 2004.
- [11] C. Lee and A. Elgammal, "Modeling view and posture manifolds for tracking," *ICCV*, 2007.
- [12] C. H. Ek, P. H. Torr, and N. D. Lawrence, "Gaussian process latent variable models for human pose estimation," *Lecture Notes in Computer Sciences (LNCS)*, 2007.
- [13] J. Gall, B. Rosenhahn, T. Brox, and H. Seidel, "Optimization and Filtering for Human Motion Capture: A Multi-Layer Framework," *International Journal of Computer Vision*, 2008.
- [14] H. Ning, X. Wei, Y. Gong, and T. Huang, "Discriminative Learning of Visual Words for 3D Human Pose Estimation," *CVPR*, 2008.
- [15] A. Bissacco, M.-H. Yang, and S. Soatto, "Fast human pose estimation using appearance and motion via multi-dimensional boosting regression," *CVPR*, 2007.
- [16] J. Gall, B. Rosenhahn, and H. Seidel, "Drift-free Tracking of Rigid and Articulated Objects," *CVPR*, 2008.
- [17] R. Urtasun and T. Darrell, "Local probabilistic regression for activity-independent human pose inference," *CVPR*, 2008.
- [18] Q. Delamarre and O. Faugeras, "3D articulated models and multi-view tracking with silhouettes," *Proc. of ICCV*, vol. 99, pp. 716–721, 1999.
- [19] D. Gavrilu and L. Davis, "3-D model-based tracking of humans in action: a multi-view approach," *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pp. 73–80, 1996.
- [20] F. Remondino, "3-D reconstruction of static human body shape from image sequence," *Computer Vision and Image Understanding*, vol. 93, no. 1, pp. 65–85, 2004.
- [21] I. Mikić, M. Trivedi, E. Hunter, and P. Cosman, "Human Body Model Acquisition and Tracking Using Voxel Data," *International Journal of Computer Vision*, vol. 53, no. 3, pp. 199–223, 2003.
- [22] S. Docksstader and A. Tekalp, "Multiple camera tracking of interacting and occluded human motion," *Proceedings of the IEEE*, vol. 89, no. 10, pp. 1441–1455, 2001.
- [23] I. Kakadiaris and D. Metaxas, "Three-Dimensional Human Body Model Acquisition from Multiple Views," *International Journal of Computer Vision*, vol. 30, no. 3, pp. 191–218, 1998.
- [24] X. Zhao, H. Ning, Y. Liu, and T. Huang, "Discriminative Estimation of 3D Human Pose Using Gaussian Processes," *ICPR*, 2008.
- [25] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *IJCV*, 2004.
- [26] R. Rosales and S. Sclaroff, "Learning Body Pose via Specialized Maps," *NIPS*, 2001.
- [27] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter-sensitive hashing," *ICCV*, 2003.
- [28] A. Agarwal and B. Triggs, "Tracking articulated motion using a mixture of autoregressive models," *ECCV*, 2004.
- [29] L. Sigal and M. Black, "Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion," *Technical Report CS-06-08, Brown University*, 2006.
- [30] C. Harris and M. Stephens, "A combined corner and edge detector," *Alvey Vision Conference*, vol. 15, p. 50, 1988.
- [31] A. Kelman, M. Sofka, and C. Stewart, "Keypoint descriptors for matching across multiple image modalities and non-linear intensity variations," *CVPR*, 2007.
- [32] M. Liu, S. Yan, Y. Fu, and T. S. Huang, "Flexible X-Y Patches for Face Recognition," *ICASSP*, pp. 2113–2116, 2008.
- [33] L. Fei-Fei and P. Perona, "A bayesian heirarchical model for learning natural scene categories," *CVPR*, 2005.
- [34] Y. Fu, L. Cao, G. Guo, and T. S. Huang, "Multiple Feature Fusion by Subspace Learning," *ACM CIVR*, pp. 127–134, 2008.
- [35] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

- [36] S. Weisberg, *Applied Linear Regression*. Hoboken, NJ: Wiley Interscience, 2004.
- [37] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2008.

Xu Zhao

Yun Fu

Huazhong Ning

Yuncai Liu

Thomas S. Huang