Human Pose Regression Through Multiview Visual Fusion

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

***Abstract*—We consider the problem of estimating 3-D human body pose from visual signals within a discriminative framework. It is challenging because there is a wide gap between complex 3-D human motion and planar visual observation, which makes this a severely 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 scale-invariant feature transform (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 (ML) 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 meaningful insights into the three crucial aspects 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, and 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 is 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 approach is an accurate and robust solution to the human body pose estimation problem.**

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

Manuscript received September 23, 2008; revised March 5, 2009 and June 25, 2009. Date of publication March 18, 2010; date of current version July 16, 2010. This work was funded by the National Basic Research Program of China (973 Program), under Grant No. 2006CB303103, the Key Program of National Natural Science Foundation of China, under Grant No. 60833009, and the National High Technology Research and Development Program of China (863 Program), under Grant No. 2009AA01Z330. This paper was recommended by Associate Editor R. Green.

X. Zhao and Y. Liu are with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhaoxu@sjtu.edu.cn; whomliu@sjtu.edu.cn).

Y. Fu is with the Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY 14260 USA (e-mail: yunfu@buffalo.edu).

H. Ning is with AKiiRA Media Systems, Inc., Palo Alto, CA 94301 USA (e-mail: hning2@ifp.uiuc.edu).

T. S. Huang is with Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: huang@ifp.uiuc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2010.2045916

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 and computer. Existing advanced technologies have been derived from a wide spectrum of real-world applications [2], such as behavior understanding, content-based image retrieval, visual surveillance, rehabilitation engineering, and humanoid robotics. Some robust solutions to this problem have been provided. However, recovering human pose, especially, 3-D pose, from planar visual information is still extremely challenging due to the complicated nature of human motion and limited available information in 2-D images [1].

In general, the state-of-the-art technologies for human pose estimation can be summarized as two categories: 1) generative, and 2) discriminative [3]. Generative methods [4]–[8] 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 the high-dimensional state space. Discriminative methods [3], [9]–[12] model the state posterior distribution conditioned on observations directly. The models are usually constructed 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 the discriminative framework for 3-D human pose estimation, as we did in [1] before.

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: 1) feature extraction, 2) learning algorithm, and 3) camera utilization.

Many existing discriminative methods [3], [9], [11], [13]–[15] 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

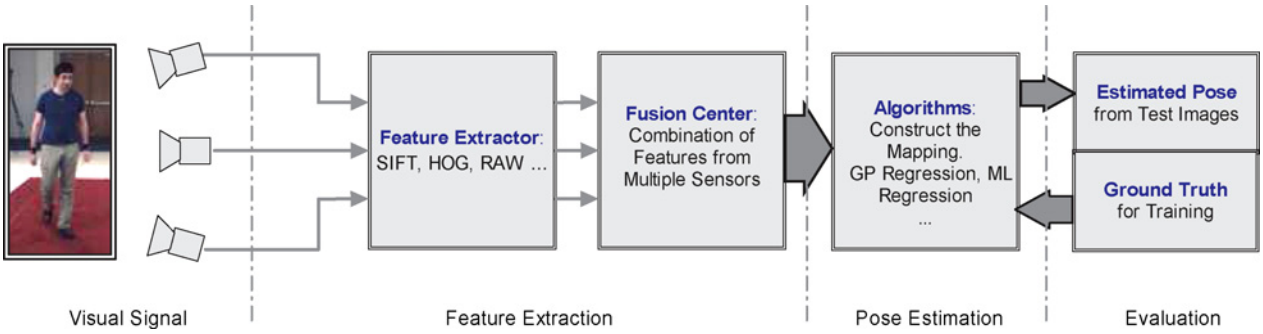


Fig. 1. Framework of human pose estimation by fusing visual information from multiple views. The whole framework consists of four main modules: visual signal, feature extraction, pose estimation, and evaluation.

appearance and lighting. The silhouettes can 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 multimodal 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 by experiments in a few recent works [16]–[19]. However, this ambiguity cannot be entirely resolved when only utilizing monocular image information, even if the image descriptor is perfect and has no loss of any local details. For such a highly nonlinear and severely ill-conditioned problem, introducing multiview visual information is a radical way to further enhance the performance of pose estimation. Most of existing works [20]–[26] using multiview information are based on a generative framework. The approaches of utilizing multiview information mainly include photogrammetric techniques [22], [23], [26] and integrating the multiview information into the computation of likelihood functions [20], [21], [25]. Derived from our previous work [1], we propose to develop a discriminative framework for body pose estimation by fusing multiview camera inputs.

We describe the image features using the salient interest points, represented by the SIFT descriptor [27]. The distribution of the feature space, containing these sparse and local image descriptors, is modeled by the bag-of-words representation, which essentially embodies the discriminative property. This representation captures the spatial co-occurrence and context information of the local structure, and also encodes the relative spatial positions. In addition, as computed on overlapped patches instead of pixels, it can tolerate a range of illumination and position variations. After extracting the image features, the fusing strategy is straightforward. We concatenate the feature vectors from 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.

As long as the discriminative features from multiple views are extracted, modeling the mapping between pose label space and feature space becomes more important. There are variety of approaches that can provide reasonable solutions, such as neural networks [13], fast nearest neighbor retrieval [6], [28], regression methods [9], [29], and Bayesian mixture of experts [3], [16]. We present a technique based on the nonparametric Gaussian processes (GP) regression, since it is flexible, fully probabilistic, and effective to handle the small-sample-size problem in the particular scenario of human pose estimation. As a contrast to GP regression, ML regression is also tested in this paper.

Extensive evaluations on the HumanEva database [30], [31] are provided across all the three aspects of the pose estimation system. In multiview scenarios, it is significant to find how the quantity and quality of image information cast impacts on the pose estimation performance. We present comparative research on different combinations of multiple views. The comparison between GP and ML regression algorithms actually demonstrates the difference in efficiency between the nonlinear nonparametric and the linear parametric algorithm, for the problem of pose estimation. Both feature extraction and choice of algorithm are crucial, but comprehensive experiments give some interesting insights into the situation when effective algorithms dominate the system performance for different choices 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 sent 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 3-D 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 corner-interest-point-based SIFT (CP-SIFT) feature, 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 [1] from single-view input to multiple-view input, which brings satisfying improvement on the performance.
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 system performance, which provide useful guidance for the future system design.

The remainder of the 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 ML regression, and their application to 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 this 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 can 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 [1] to specify the following feature extraction procedure.

1) *Human Detection:* The background substraction 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 [32] is used to detect interest points. Fig. 2(a) 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 [27] is applied at each interest point, which is denoted as a vector $\mathbf{p}$.

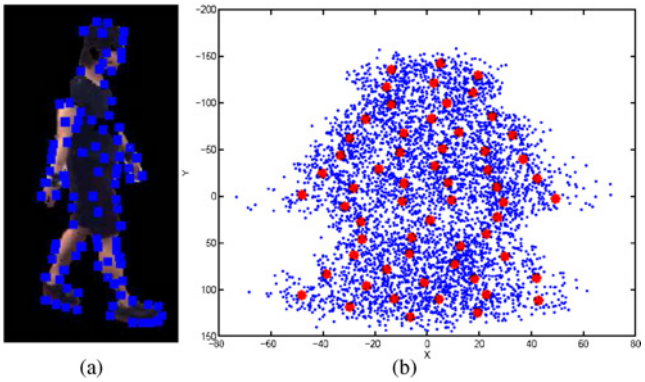


Fig. 2. Feature extraction (originally shown in [1]). (a) Interest points on an image frame. (b) Relative coordinates of interest points from all images (marked as "$\cdot$") and the visual words (marked as "$*$").

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 the contributions in this paper. The feature is scale invariant and partially illumination invariant due to the introduction of Harris corner 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 multimodal 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 [33] especially by our previous "X–Y patch" work [34], 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 hierarchal structure with fixed relative connections between different body parts.

After we calculate all the local descriptors, the unsupervised bag-of-words model [35] 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 $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$. In this paper, the number of visual words is empirically set as 60 in the experiments, which are orders of magnitude lower than other works [16], [35]. Once the code book is available, each descriptor in the descriptor set $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_m\}$ 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 accumulated score of all descriptors on the K visual

words. Each image finally is represented as a K-dimensional feature vector. In Fig. 2(b), the relative coordinates of interest points and the visual words are displayed as a 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 this paper, these raw features are used as a baseline feature for the comparisons with CP-SIFT.

As for multiview visual data input, we use the simplest fusion method by concatenating the feature vectors of all the synchronized views. According to our previous work [36], a more sophisticated fusion method can be adopted. To demonstrate the advantage of multiview 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 3-D human poses.

A. Nonparametric Regression: Gaussian Process Regression

GP [37] 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 $\mathcal{S} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$ 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

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

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

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

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

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

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

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

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

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

$\mathbf{C}_N$ has elements

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

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

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

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

where $\mathbf{Y} = [y_1, \ldots, y_N]^T$ 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

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

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

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

B. Parametric Regression: ML Regression

To evaluate the efficiency of different regression algorithms on the pose estimation task, we take the ML regression model [38] as a comparative baseline in the algorithm level. The ML regression model can be formulated as

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

where $\mathbf{Y}$ is the joint angle vector over all training samples. $\widetilde{\mathbf{X}}$ 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 $\widetilde{\mathbf{X}}$ are all 1s 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 $\hat{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{Y}$ [38]. But the normal equations are often badly conditioned relative to the original system. So the orthogonal decomposition of $\widetilde{\mathbf{X}}$ is used to find the solution.

Once the regression model is trained, we have

$$\hat{\mathbf{Y}} = \widetilde{\mathbf{X}}\hat{\boldsymbol{\beta}} \tag{11}$$

with

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

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

TABLE I

AVERAGE RMS ERROR (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

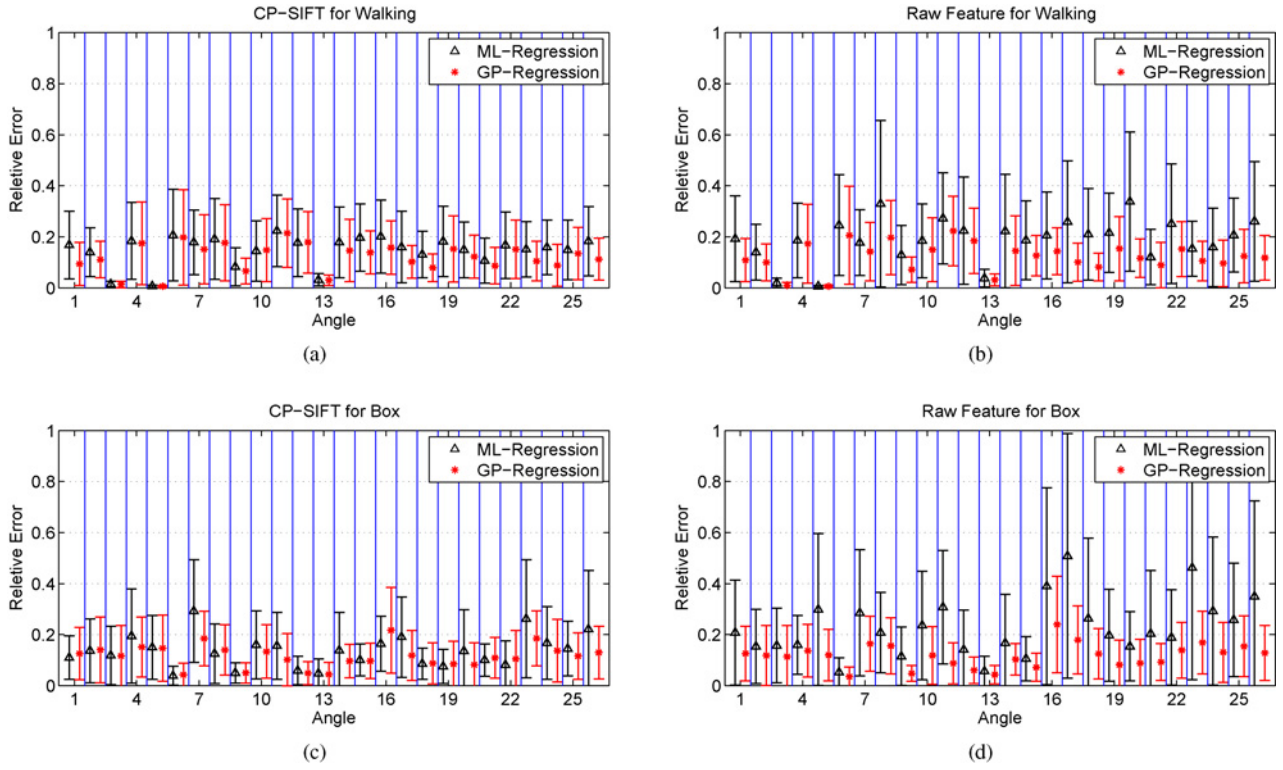


Fig. 3. Performance comparison between GP and ML regression on (a) CP-SIFT for *Walking*, (b) raw feature for *Walking* (c) CP-SIFT feature for *Box*, and (d) raw feature for *Box*. Here, (a) and (b) are for the *Walking* action, (c) and (d) are for the *Box* action. Both mean and standard deviation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

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 multiviews. All the experiments are conducted on the publicly available HumanEva dataset [30], [31] for the evaluation of human pose estimation, collected at Brown University, Providence, RI.

A. Database

The HumanEva data were captured simultaneously using multiple high-speed video capture systems and a calibrated marker-based motion capture system, whose video and motion streams were well synchronized. In the recording, multiple subjects were asked to perform a set of predefined actions repetitively. This database originally provides 3-D locations of the body parts in the world coordinate system for the motion capture data. In total, there are 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 3-D locations to the 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 are normalized to be a zero-mean unit variance process. An original partition on the database generates the training, validation, and test subsets. As there is no motion data provided for the test set, we use sequences in the training set for training and those in the validation set for testing. As a result, 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 multi-view visual information we used are from cameras C1, C2, and C3.

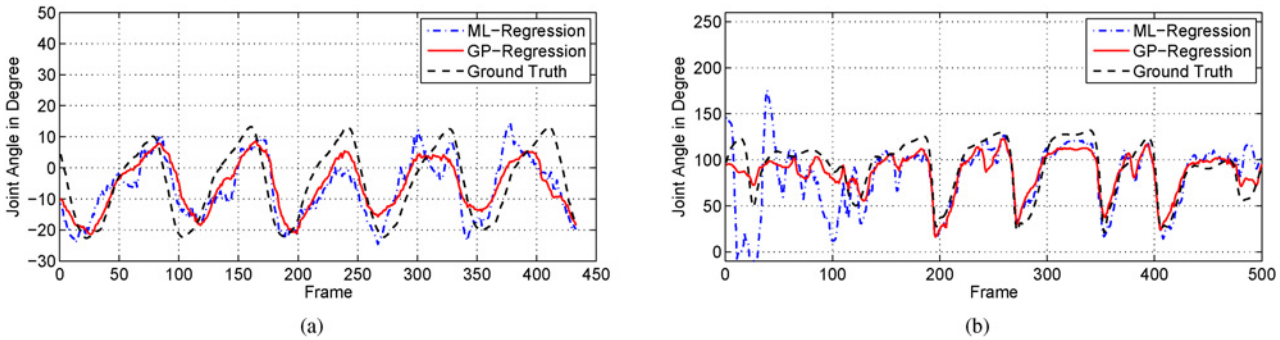


Fig. 4. Curve comparisons of joint angles: ground truth and estimations with GP and ML regression. (a) Right hip (x-axis) of subject S2 in *Walking*. (b) Right elbow of subject S3 in boxing.

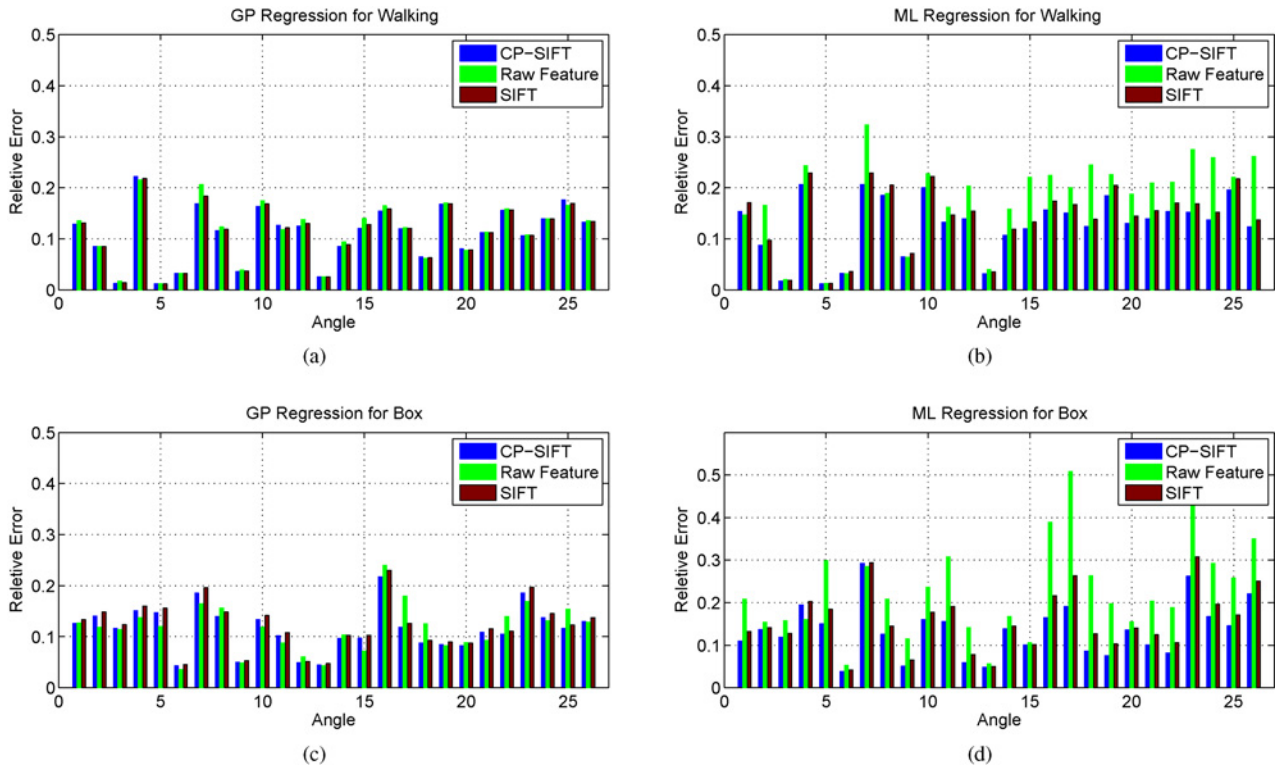


Fig. 5. Performance comparison between CP-SIFT feature, SIFT feature and raw feature for (a) and (c) GP regression and (b) and (d) ML regression. Here, (a) and (b) are for the *Walking* action, (c) and (d) are for the *Box* action. RMS error of each individual angle, normalized by the range of that angle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

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 further comparisons between the two classes of algorithms (nonparametric nonlinear and parametric linear regression algorithm) on the pose estimation problem, we also use the ridge regression algorithm in the experiments. For multiview visual data, we choose to discuss the combination of cameras C1 and C2 since similar results and conclusions can be obtained from other camera combinations in our investigation. For the raw feature, we reduce the dimension to 100 with principal component analysis (PCA) after fusion.

In Table I, we report the mean (over all joint angles) root mean square (RMS) absolute difference errors [9] between the ground-truth and estimated joint angles, in degrees

$$D(\mathbf{y}, \mathbf{y}') = \frac{1}{d}\sum_{i=1}^{d} |(y_i - y_i')\text{mod} \pm 180^{\text{deg}}|. \tag{12}$$

From the table, we can see that the performance of GP regression largely outperforms ML regression and ridge regression for all the three features. The ML regression and ridge regression get close performance and the performance differ-

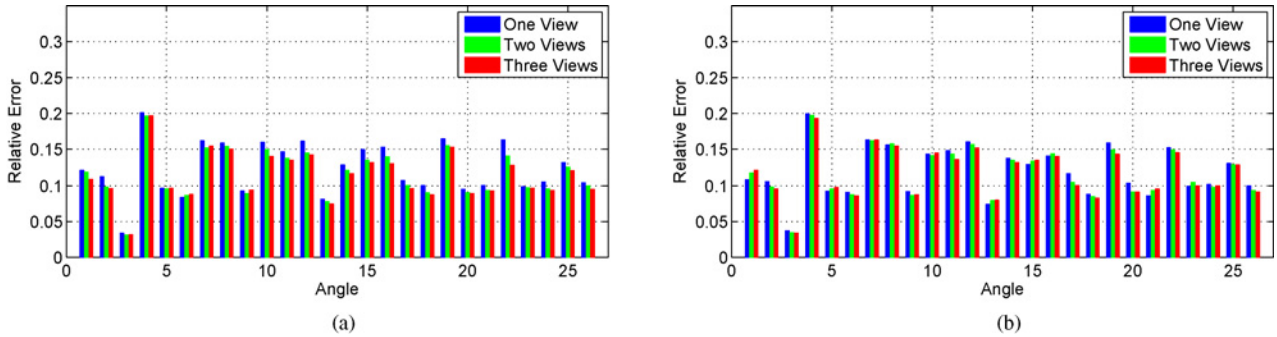


Fig. 6. Performance comparison between different view combinations. Average relative errors on all the *Walking* and *Jog* sequences for one-view, two-view, and three-view combinations. The algorithm and features used here are (a) GP and raw feature, and (b) GP and CP-SIFT feature.

ence is statistically insignificant. These results demonstrate the efficacy of nonparametric nonlinear regression algorithm on pose estimation.

We can also observe that the performance of the CP-SIFT feature is significantly better than that of the raw feature for all 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 do not 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. 5(a) and (c) for GP regression, the superiority of CP-SIFT feature over raw feature and SIFT feature is small. However, as we can see in Fig. 5(b) and (d), this superiority is salient for ML regression. It is an interesting observation, which is also consistent with that indicated by the data shown in Table I. It demonstrates that the performance difference among algorithms is much larger than that among features. In other words, the choice of regression algorithm plays a more important role than the choice of feature for 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 vectors from each single camera together as the complete representation of the visual signal. To avoid over fitting, we 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 in consideration. We represent these combinations as C1, C2, C3, C1–C2, C1–C3, C2–C3, and C1–C2–C3.

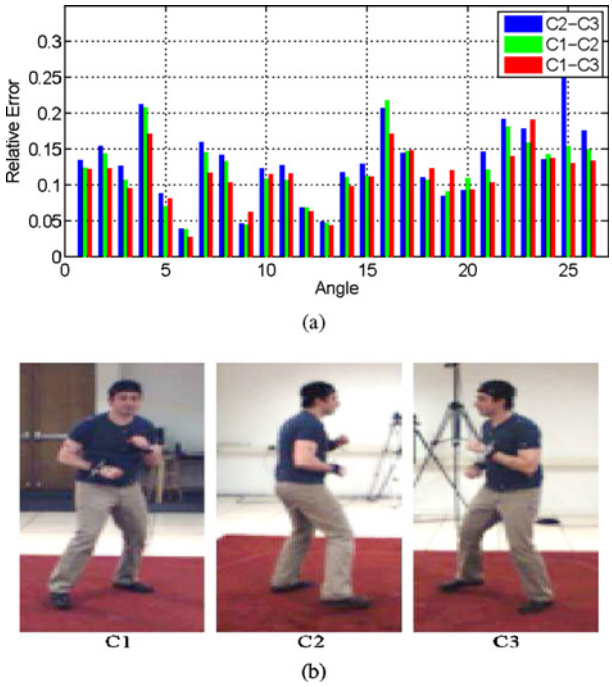


Fig. 7. Multiview comparison. (a) Average relative errors on all the Box sequences for different two-view combinations C1–C2, C1–C3, C2–C3. (b) Sample images from camera C1, C2, and C3 for the Box action.

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

TABLE II

AVERAGE RMS ERROR (IN DEGREE) OVER ALL JOINT ANGLES, ALL SUBJECTS FOR *Walking*, *Box*, *Jog*, AND *Gestures* ACTIONS

		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

The evaluation is for the all-round comparisons of algorithm, feature, and view combination.

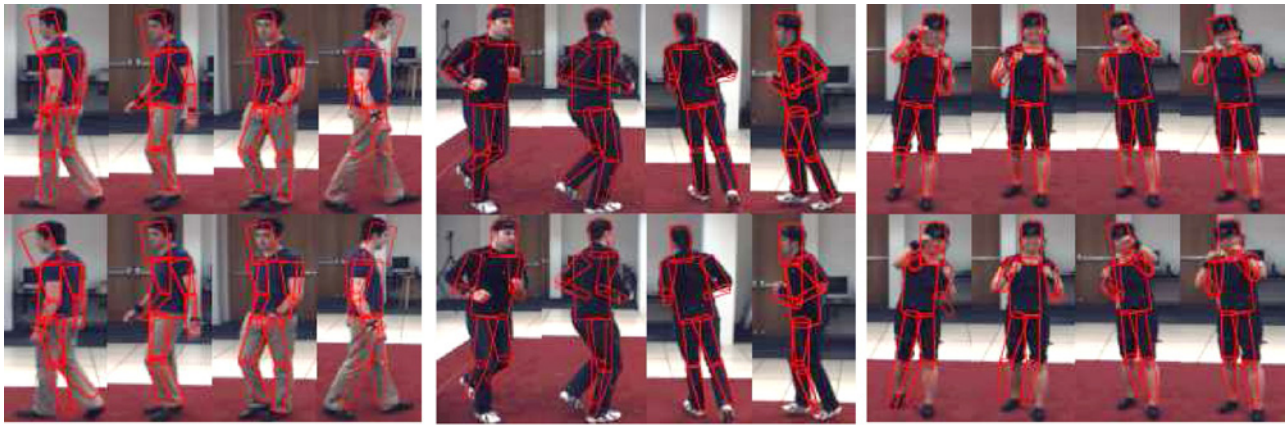


Fig. 8. Some of the sample estimation results for action (a) *Walking* (b) *Jog*, and (c) *Box* (originally shown in [1]). The first row shows the provided ground truth projected onto the camera C1, and the second row shows the estimated pose projected onto the same camera. Each column corresponds to a frame.

action, we make the comparisons on two classes of actions: 1) people moving around, and 2) 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. 7(b). 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 when the information from multiple cameras is involved in the process of pose estimation, the ill-condition of this problem is mitigated. 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 multiview combination is related to the quality of image information. Fig. 7(a) 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. 7(b) 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.

D. Discussion

In the presented evaluations, we evaluated how and to what extent the three critical factors, feature extraction, regression algorithm, and multiview utilization impact 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 impacts 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 and Tables I and II, we can see that if the regression algorithm is not powerful enough, the impact of the choice of feature will be remarkable. But once the regression algorithm performs better, e.g., in this paper 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 is 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. 7, 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 aspects. In the feature extraction module, we introduced the 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 comparison 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 [39]. Another interesting observation is about the information fusion of multiple views. In the process of pose estimation, before fusing multiview information, one has 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 fusion strategies from our recent work [36] on the multiple feature combination. The temporal-spatial local Gaussian process experts model [40] will also be developed to handle multimodality.

Acknowledgment

The authors would like to thank Brown University, Providence, RI, for providing the HumanEva database [30], [31].

References

[1] X. Zhao, H. Ning, Y. Liu, and T. Huang, "Discriminative estimation of 3-D human pose using Gaussian processes," in *Proc. Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[2] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vision Image Understand.*, vol. 104, nos. 2–3, pp. 90–126, 2006.

[3] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Discriminative density propagation for 3-D human motion estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2005, pp. 390–397.

[4] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, vol. 2. 2000, pp. 126–133.

[5] H. Ning, T. Tan, L. Wang, and W. Hu, "People tracking based on motion model and motion constraints with automatic initialization," *Pattern Recognit.*, vol. 37, no. 7, pp. 1423–1440, 2004.

[6] G. Mori and J. Malik, "Recovering 3-D human body configurations using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1052–1062, Jul. 2006.

[7] M. Lee and I. Cohen, "A model-based approach for estimating human 3-D poses in static images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 905–916, Jun. 2006.

[8] X. Zhao and Y. Liu, "Generative tracking of 3-D human motion by hierarchical annealed genetic algorithm," *Pattern Recognit.*, vol. 41, no. 8, pp. 2470–2483, 2008.

[9] A. Agarwal and B. Triggs, "Recovering 3-D human pose from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 44–58, Jan. 2006.

[10] M. Brand, "Shadow puppetry," in *Proc. IEEE Int. Conf. Comput. Vision*, vol. 2. 1999, pp. 1237–1244.

[11] A. Elgammal and C.-S. Lee, "Inferring 3-D body pose from silhouettes using activity manifold learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2004, pp. 681–688.

[12] C. Lee and A. Elgammal, "Modeling view and posture manifolds for tracking," in *Proc. IEEE Int. Conf. Comput. Vision*, 2007, pp. 1–8.

[13] R. Rosales and S. Sclaroff, "Learning body pose via specialized maps," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 2. 2002, pp. 1263–1270.

[14] C. H. Ek, P. H. Torr, and N. D. Lawrence, "Gaussian process latent variable models for human pose estimation," in *Lecture Notes in Computer Sciences*, vol. 4892. Berlin, Germany: Springer, 2008, pp. 132–143.

[15] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel, "Optimization and filtering for human motion capture: A multilayer framework," *Int. J. Comput. Vision*, vol. 87, nos. 1–2, pp. 75–92, 2010.

[16] H. Ning, X. Wei, Y. Gong, and T. Huang, "Discriminative learning of visual words for 3-D human pose estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2008, pp. 1–8.

[17] A. Bissacco, M.-H. Yang, and S. Soatto, "Fast human pose estimation using appearance and motion via multidimensional boosting regression," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2007, pp. 1–8.

[18] J. Gall, B. Rosenhahn, and H.-P. Seidel, "Drift-free tracking of rigid and articulated objects," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2008, pp. 1–8.

[19] R. Urtasun and T. Darrell, "Local probabilistic regression for activity-independent human pose inference," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2008, pp. 1–8.

[20] Q. Delamarre and O. Faugeras, "3-D articulated models and multiview tracking with silhouettes," in *Proc. IEEE Int. Conf. Comput. Vision*, vol. 99. 1999, pp. 716–721.

[21] D. Gavrila and L. Davis, "3-D model-based tracking of humans in action: A multiview approach," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 1996, pp. 73–80.

[22] F. Remondino, "3-D reconstruction of static human body shape from image sequence," *Comput. Vision Image Understand.*, vol. 93, no. 1, pp. 65–85, 2004.

[23] I. Mikić, M. Trivedi, E. Hunter, and P. Cosman, "Human body model acquisition and tracking using voxel data," *Int. J. Comput. Vision*, vol. 53, no. 3, pp. 199–223, 2003.

[24] R. Plankers and P. Fua, "Articulated soft objects for video-based body modeling," in *Proc. IEEE Int. Conf. Comput. Vision*, vol. 1. 2001, pp. 394–401.

[25] S. Dockstader and A. Tekalp, "Multiple camera tracking of interacting and occluded human motion," *Proc. IEEE*, vol. 89, no. 10, pp. 1441–1455, Oct. 2001.

[26] I. Kakadiaris and D. Metaxas, "3-D human body model acquisition from multiple views," *Int. J. Comput. Vision*, vol. 30, no. 3, pp. 191–218, 1998.

[27] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[28] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter-sensitive hashing," in *Proc. IEEE Int. Conf. Comput. Vision*, 2003, pp. 750–757.

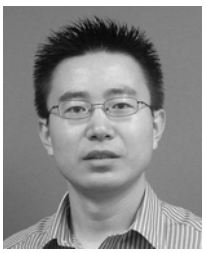
[29] A. Agarwal and B. Triggs, "Tracking articulated motion using a mixture of autoregressive models," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2004, pp. 54–65.

[30] L. Sigal and M. Black, "HumanEva: Synchronized video and motion

capture dataset for evaluation of articulated human motion," Dept. Comput. Sci., Brown Univ., Providence, RI, Tech. Rep. CS-06-08, 2006.
[31] L. Sigal, A. Balan, and M. Black, "HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion," *Int. J. Comput. Vision*, vol. 87, nos. 1–2, pp. 4–27, 2010.
[32] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vision Conf.*, 1988, pp. 147–151.
[33] A. Kelman, M. Sofka, and C. Stewart, "Keypoint descriptors for matching across multiple image modalities and nonlinear intensity variations," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2007, pp. 1–7.
[34] M. Liu, S. Yan, Y. Fu, and T. Huang, "Flexible X–Y patches for face recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 2113–2116.
[35] L. Fei-Fei and P. Perona, "A Bayesian heirarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2005, pp. 524–531.
[36] Y. Fu, L. Cao, G. Guo, and T. Huang, "Multiple feature fusion by subspace learning," in *Proc. Assoc. Comput. Machinery Int. Conf. Image Video Retrieval (CIVR)*, 2008, pp. 127–134.
[37] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.
[38] S. Weisberg, *Applied Linear Regression*. Hoboken, NJ: Wiley-Interscience, 2004.
[39] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[40] X. Zhao, Y. Fu, and Y. Liu, "Temporal–Spatial Local Gaussian Process Experts for Human Pose Estimation," in *Proc. 9th Asian Conf. Comput. Vision*, 2009.

Xu Zhao received the B.S. degree in electrical engineering from China Agricultural University, Beijing, China, in 1997, the M.S. degree in electrical engineering from China Ship Research and Development Academy, Beijing, China, in 2004, and is currently pursuing the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China.

From 2007 to 2008, he was a Visiting Student with the Beckman Institute for Advanced Science and Technology at University of Illinois, Urbana. His current research interests include visual analysis of human motion, machine learning, and image/video processing.



Yun Fu (S'07–M'08) received the B.Eng. degree in information engineering in 2001, the M.Eng. degree in pattern recognition and intelligence systems in 2004, both from Xi'an Jiaotong University, Xi'an, China, the M.S. degree in statistics in 2007, and the Ph.D. degree in electrical and computer engineering in 2008, both from the University of Illinois at Urbana-Champaign, Champaign.

He was a Research Intern with Mitsubishi Electric Research Laboratories, Cambridge, MA, in 2005, and with Multimedia Research Lab of Motorola Labs, Schaumburg, IL, in 2006. He jointed BBN Technologies, Cambridge, MA, as a Scientist in 2008. He was a part-time Lecturer with the Department of Computer Science, Tufts University, Medford, MA, in 2009. Since 2010, he has been an Assistant Professor with the Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY. His current research interests include human-centered computing, image and video analysis, pattern recognition, smart environments.

Dr. Fu is the recipient of the 2002 Rockwell Automation Master of Science Award, Edison Cups of the 2002 GE Fund Edison Cup Technology Innovation Competition, the 2003 Hewlett-Packard Silver Medal and Science Scholarship, the 2007 Chinese Government Award for Outstanding Self-Financed Students Abroad, the 2007 DoCoMo USA Labs Innovative Paper Award (IEEE International Conference on Image Processing 2007 Best Paper Award), the 2007–2008 Beckman Graduate Fellowship, and the 2008 M. E. Van Valkenburg Graduate Research Award. He is a Life Member of the Institute of Mathematical Statistics, and a Beckman Graduate Fellow.

Huazhong Ning received the B.S. degree in computer science from University of Science and Technology of China, Hefei, China, in 2000, the M.S. degree in pattern recognition and intelligence systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China in 2003, and the Ph.D. degree in electrical engineering at University of Illinois, Urbana, in 2008.

From 2003 to 2004, he was a 3G Software Engineer with Alcatel Shanghai Bell Co., Shanghai, China. From 2008 to 2009, he was an Applied Researcher with Microsoft AdCenter Labs, Redmond, WA. Since 2009, he has been a Research Scientist with AKiiRA Media Systems, Inc., Palo Alto, CA. His current research interests include video/image processing, machine learning, clustering, audio analysis, data mining, and so on.

Yuncai Liu (M'94) received the Ph.D. degree from the Department of Electrical and Computer Science Engineering, University of Illinois, Urbana, in 1990.

From 1990 to 1991, he was an Associate Researcher with Beckman Institute of Science and Technology, Urbana, IL. Since 1991, he had been a System Consultant and then a Chief Consultant of Research with Sumitomo Electric Industries Ltd., Tokyo, Japan. In 2000, he joined the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, and is currently a Distinguished Professor. His current research interests include image processing and computer vision, especially in motion estimation, feature detection and matching, and image registration. He also made many progresses in the research of intelligent transportation systems.

Thomas S. Huang (S'61–M'63–SM'76–F'79–LF'01) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, and the M.S. and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He was a Faculty Member with the Department of Electrical Engineering, MIT, from 1963 to 1973, and a Faculty Member with the School of Electrical Engineering and the Director of its Laboratory for Information and Signal Processing, Purdue University, West Lafayette, IN, from 1973 to 1980. In 1980, he joined the University of Illinois at Urbana-Champaign, Urbana, where he is now William L. Everitt Distinguished Professor of Electrical and Computer Engineering, a Research Professor with the Coordinated Science Laboratory, the Head of the Image Formation and Processing Group with the Beckman Institute for Advanced Science and Technology, and the Co-Chair of the Institute's major research theme: human-computer intelligent interaction. His professional interests lie in the broad area of information technology, especially the transmission and processing of multidimensional signals. He has published 20 books and over 500 papers in network theory, digital filtering, image processing, and computer vision.

Dr. Huang is a Member of the National Academy of Engineering, a Foreign Member of the Chinese Academies of Engineering and Science, and a Fellow of the International Association of Pattern Recognition and the Optical Society of America. He has received a Guggenheim Fellowship, an A. von Humboldt Foundation Senior U.S. Scientist Award, and a Fellowship from the Japan Association for the Promotion of Science. He received the IEEE Signal Processing Society's Technical Achievement Award in 1987, and the Society Award in 1991. He was awarded the IEEE Third Millennium Medal in 2000. Also in 2000, he received the Honda Lifetime Achievement Award for "contributions to motion analysis." In 2001, he received the IEEE Jack S. Kilby Medal. In 2002, he received the King-Sun Fu Prize, International Association of Pattern Recognition, and the Pan Wen-Yuan Outstanding Research Award. He is a Founding Editor of the *International Journal of Computer Vision, Graphics, and Image Processing* and the Editor of the *Springer Series in Information Sciences* (Berlin, Germany: Springer).