Discriminative Learning of Visual Words for 3D Human Pose Estimation

Huazhong Ning[1] Wei Xu[2] Yihong Gong[2] Thomas Huang[1]
[1]Dept. of ECE, U. of Illinois at Urbana-Champaign, USA. {hning2,huang}@ifp.uiuc.edu.
[2]NEC Laboratories America, Inc., USA. {xw,ygong}@sv.nec-labs.com.

Abstract

This paper addresses the problem of recovering 3D human pose from a single monocular image, using a ***discriminative*** *bag-of-words approach. In previous work, the visual words are learned by unsupervised clustering algorithms. They capture the most common patterns and are good features for coarse-grain recognition tasks like object classification. But for those tasks which deal with subtle differences such as pose estimation, such representation may lack the needed discriminative power. In this paper, we propose to jointly learn the visual words and the pose regressors in a supervised manner. More specifically, we learn an individual distance metric for each visual word to optimize the pose estimation performance. The learned metrics rescale the visual words to suppress unimportant dimensions such as those corresponding to background. Another contribution is that we design an* Appearance and Position Context (APC) *local descriptor that achieves both selectivity and invariance while requiring no background subtraction. We test our approach on both a quasi-synthetic dataset and a real dataset (HumanEva) to verify its effectiveness. Our approach also achieves fast computational speed thanks to the integral histograms used in APC descriptor extraction and fast inference of pose regressors.*

1. Introduction

Robust recovery of 3D human pose in monocular images or videos is an actively growing field. Effective solutions would lead to breakthroughs in a wide range of applications spanning visual surveillance, video indexing and retrieval, human-computer interfaces, and so on. Unfortunately, this problem is extremely challenging due to both the internal complexity of the articulated human body and the external variations of the scene.

There are two general classes of approaches for human pose estimation: *generative methods* and *discriminative methods*. The generative methods recover the hidden states (human pose) within an analysis-by-synthesis loop. They are natural and flexible to represent the hidden states and appearance of the human body, but their applicability is partly prohibited by the high computational cost to infer the distribution on the hidden states and by the difficulties of constructing the observation models [21]. These disadvantages have motivated the advent of *discriminative methods* that learn direct image-to-pose mappings by training on a dataset with labeled human poses. Compared to generative models, the discriminative models, once trained, have the advantage of much faster test speed, although in some cases they cannot obtain estimates as precise as generative methods do. Our purpose of pose estimation is to recognize human actions in monocular videos without requiring precise estimates for each frame. Therefore, the fast discriminative methods exactly fit our purpose.

Among the image representations used by the discriminative methods, the *bag-of-words* model has produced superior results in the literature [7, 20]. However, among the majority works to date, the bag of visual words are usually obtained by unsupervised clustering methods such as K-means. Visual words obtained this way actually capture the most common patterns in the entire training set, and are good features for coarse-grain recognition tasks such as object detection and classification. However, such representations may lack the needed power to discriminate subtle differences in recognition tasks such as pose estimation.

In this paper, we propose to use a supervised method to learn visual words for the specific problem (human pose estimation). We start with the visual words that are initially obtained by an unsupervised clustering algorithm, and then learn a separate metric for each visual word from the labeled image-to-pose pairs through a supervised learning process. We use the Bayesian mixtures of experts (BME) to represent the multi-modal distribution of the 3D human pose space conditioned on the feature space. The metric learning and the BME model are jointly optimized by an iterative gradient ascent algorithm.

In essence, the visual words obtained by an unsupervised clustering method represent the general frequent patterns existing in all training images, and the visual words obtained by our supervised learning method capture the patterns that are particularly informative for pose estima-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

tion. More specifically, the learned distance metric implicitly transforms the visual word to a new space so that (1) it can better represent the local structures (*e.g.*, bent elbow) useful for pose estimation; (2) it can suppress the unimportant dimensions of the visual words, especially the dimensions corresponding to background. When the background varies, these dimensions might introduce nontrivial errors if they are treated uniformly.

A successful bag-of-words approach heavily relies on the design of local image descriptors that possess such preferable features as high discriminative power and invariance to scale, rotation, illumination, and background to some extent. In this paper, we design an sparse and local image descriptor that 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 scale and position variations because it is computed on small cells, instead of pixels. We call it Appearance and Position Context (APC) descriptor. It is superior to the shape context descriptor in that it requires no background subtraction and silhouette extraction. It also outperforms the SIFT descriptor [12] in our experiments (see Section 5.1).

The contributions of our work is summarized as follows. (1) We jointly learn the visual words and the pose estimators in a supervised manner. The learned metrics rescale the visual words to suppress those unimportant dimensions that correspond to background. (2) Our APC descriptor achieves both the discriminative power and the invariance while requiring no background subtraction and silhouette extraction. We have constructed a quasi-synthetic human database that is much larger and more complex than the previous ones [22, 1, 9, 11], trained and tested our approach on this dataset to verify its effectiveness. We have also tested our approach using the real dataset HumanEva [18] and have achieved the state of the art performance. Our approach achieves fast computational speed thanks to the integral histograms used in APC descriptor extraction and fast inference of pose estimators.

2. Related Work

The research we present relates to topics including image descriptors, discriminative methods for human pose estimation, generative methods, and a combination of both generative and discriminative methods.

The image descriptor is a compact representation of an image that is expected to preserve both selectivity and invariance. Most of the commonly used image descriptors for discriminative human pose estimation are either silhouette-based descriptors, such as bag of shape context descriptors [1], Gaussian mixture models of silhouette [9], and signed-distance functions on silhouette [5], or dense holistic features, such as block SIFT [22], HOG [14], hierarchical features [11], and Hu moments [15]. These descriptors are successful, but the silhouette-based descriptors rely on accurate silhouette extraction, and the dense holistic features require alignment of human region in detection window. We use bag-of-words representation as [1] did to resist misalignment, and design an APC descriptor that can represent the subtle differences in pose estimation while requiring no background subtraction and silhouette extraction.

The discriminative methods learn direct image-to-pose mappings by training on labeled data. The learned mappings differ in the organization of training set and in the runtime hypothesis selection [22], varying from linear/nonlinear regression [1], Bayesian mixture of experts (BME) [22, 9, 19], manifold embedding [5, 11], nearest-neighbor retrieval from typical examples [6, 16], mixture of probabilistic PCA [8], to mixture of multi-layer perceptrons for each pose cluster [15]. We choose the BME model because it has been verified to be able to accurately represent the multi-modal image-to-pose distributions and also can be jointly optimized with the distance metric learning.

The discriminative methods usually have fast computational speed, while the estimates by the generative methods are often more precise. Therefore, researchers have attempted to combine them and expected to explore the advantages of both of them [17, 21]. However, both the generative and the combinative methods usually require high computational cost in inference. This paper only focuses on the fast pose recovery by discriminative algorithms.

3. Image Representation

3.1. Appearance & Position Context Descriptor

With the human vision system, it is highly probable that we recognize human poses in 2D images by identifying the shapes and positions of the informative local structures (*e.g.*, bent elbow, stretched arm, and lifted leg). This observation motivates us to design an Appearance and Position Context (APC) descriptor specifically for human pose estimation.

The APC descriptor is extracted in the following steps. (1) For each image, the human window is detected and rescaled to a fixed size. (2) Centered at each point that has large gradient in the human window, the local region is partitioned into log-polar sectors (top row at Fig. 1(a)) [13], making the descriptor more sensitive to positions of nearby sample points than to those far away. (3) Suppose from inner to outer, the sectors are numbered $1, 2, ..., B$, and θ_i, m_i is the orientation and magnitude of the dominant gradient in sector i. Then the local descriptor is represented as $(x, y, \theta_1, r_1, ..., \theta_B, r_B)$ where x, y is the relative position in the human window and, $r_i = m_i/m_1$ is the normalized magnitude that basically removes the contrast of the image.

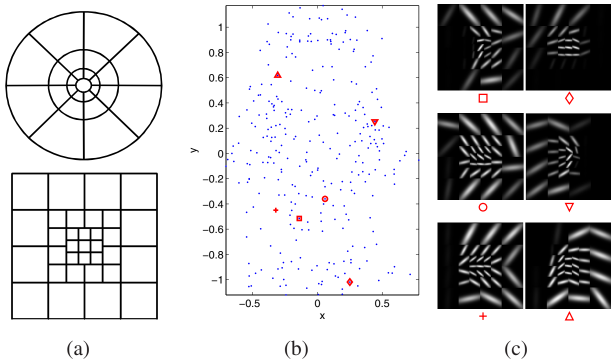


Figure 1. (a) Partition of local descriptor region. Top: log-polar partition; Bottom: rectangular partition. (b) The x, y values (relative positions) of K visual words. It has a human shape. (c) Six sample visual words. We draw the dominant orientations whose magnitudes are rescaled by their learned metrics. Their corresponding x, y are marked in (b). From left to right and top to down, their marks in (b) and (c) are □, ◇, ○, ▽, +, and △, respectively, and their represented local structures are belt knee, ankle, bent knee, bent elbow, bent knee, and shoulder, respectively. Enlarge for better visualization.

In implementation, the log-polar sectors are approximated by rectangular cells so that fast computation is allowed by integral histograms. The bottom row at Fig. 1(a) is an exemplar partition. The size of the local region is chosen to exactly cover the average length of human limbs. To calculate the dominant gradient, an orientation histogram is computed for each cell where the votes are weighted by the gradient magnitude and interpolated bilinearly between neighboring histogram entries. The dominant gradient corresponds to the maximum histogram entry.

Our descriptor is inspired by the shape context (SC) descriptor proposed by [13] in the aspect of capturing co-occurrence information in local regions. The SC descriptor has been successfully applied to human pose estimation by Agarwal and Triggs [1]. But such a silhouette-based representation is prone to left-right ambiguities and cannot be applied to the cases where background subtraction is unavailable. While our descriptor encodes richer information to disambiguate hard poses, and requires no background subtraction. Our descriptor also outperforms the sparse SIFT descriptor [12] in testing accuracy in our experiments (see Section 5.1), mainly due to that (1) our descriptor encodes the relative position x, y that helps to locate the local structures. Experiments show that it makes a significant contribution to the accurate estimation; (2) our partition is larger enough to capture the context information; and (3) we utilize the dominant gradient, instead of the entire histogram, in each cell, which suppresses noise and enables invariance.

3.2. Bag-of-Words

Our bag-of-words model is initially obtained by an unsupervised method as most of the previous work did [7, 20, 1]. First, the APC descriptors extracted from all training images are clustered by K-means, and the K cluster centers, called *visual words*, form a set $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_K\}$ that is the so called *codebook*. Fig. 1(b) shows the x, y values (relative positions) of all visual words that forms a human shape, *i.e.*, the visual words basically cover the key points of the human images. Fig. 1(c) gives six sample visual words that are typical local structures. For each visual word, we draw the dominant orientations whose magnitudes are rescaled by their learned metrics (see Section 4.2). Their x, y coordinates are marked in Fig. 1(b).

After the codebook is available and given a testing image I and its APC descriptor set $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_m\}$, each descriptor votes softly with respect to the visual words. The bag-of-words representation, denoted as $\mathbf{x}$, is the accumulating scores of all descriptors. The i-th element x_i of $\mathbf{x}$ is:

$$x_i = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{d} \in \mathcal{D}} e^{-\rho^2(\mathbf{c}_i, \mathbf{d}, A_i)}, i = 1, 2, \cdots, K \qquad (1)$$

where $\rho(\mathbf{c}, \mathbf{d}, A) = \sqrt{(\mathbf{c} - \mathbf{d})^T A (\mathbf{c} - \mathbf{d})}$, and A is positive semi-definite, *i.e.*, $A \succeq 0$, parameterizing a family of Mahalanobis distance.

In most of the previous work, A_i's are empirically chosen. In this paper, A_i's are obtained from the labeled image-to-pose data through a supervised learning process. This distinguishes our approach from most of the previous work.

4. Joint Learning of Metrics and BME

As mentioned in Section 1, the visual words obtained by an **unsupervised** method may lack discriminative power for those problems that deal with subtle differences such as pose estimation. Thus, we propose to obtain the visual words through a **supervised** learning process so as to make them particularly informative to the specific problem of pose estimation. This is done by learning a sperate distance metric for each visual word from the labeled image-to-pose pairs. More specifically, we start with the visual words initially obtained by an unsupervised algorithm, and then jointly learn the distance metrics and the BME model through a supervised learning process.

4.1. Bayesian Mixtures of Experts

The image-to-pose relation is highly non-linear. Fortunately, close observation of human images shows that human appearance changes very fast as the human global orientation changes, while the appearance changes relatively slowly in a fixed orientation. Therefore, we may assume that the image-to-pose distribution in a fixed orientation can be well modelled by a single or a combination of linear regressor(s). This leads us to use the Bayesian mixtures of experts (BME) [2, 10] to model the multi-modal image-to-pose distributions. Suppose $\mathbf{x}$ is the bag-of-words representation of the image and $\mathbf{y}$ is the human pose, the model with M experts is:

$$p(\mathbf{y}|\mathbf{x}, \Theta) = \sum_{i=1}^{M} g(\mathbf{x}, \nu_i) p(\mathbf{y}|\mathbf{x}, T_i, \Lambda_i) \qquad (2)$$

where

$$\begin{aligned} g(\mathbf{x}, \nu_i) &= \frac{e^{\nu_i^T \mathbf{x}}}{\sum_j e^{\nu_j^T \mathbf{x}}} && (3) \\ p(\mathbf{y}|\mathbf{x}, T_i, \Lambda_i) &\sim \mathcal{N}(T_i \mathbf{x}, \Lambda_i) && (4) \end{aligned}$$

Here $\Theta = \{\nu_i, T_i, \Lambda_i | i = 1, 2, \cdots, M\}$ consists of the parameters of the BME model. $p(\mathbf{y}|\mathbf{x}, T_i, \Lambda_i)$ is an Gaussian distribution with mean $T_i\mathbf{x}$ and covariance matrix Λ_i, and it is an *expert* that transforms the input into output prediction. Then the predictions from different experts are combined in a probabilistic mixture model. Note that the mixing proportions of the experts, $g(\mathbf{x}, \nu_i)$, are *input dependent* and normalized to 1 by the softmax construction. They reflect the distributions of the outputs in the training set. They work like gates that can competitively switch-on multiple experts for some input domains, allowing multimodal conditionals. They can also pick a single expert for unambiguous inputs by switching-off other experts.

The parameter Θ can be estimated by Maximum Likelihood $L = \sum_k \ln p(\mathbf{y}^{(t)}|\mathbf{x}^{(t)}, \Theta)$ where $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ are labeled image-to-pose pairs. This can be achieved through EM algorithm. Interested readers are referred to [2, 10].

4.2. Learning Distance Metrics for Visual Words

Among the majority work to date, the visual words are learned by unsupervised clustering methods. They represent the most frequent patterns existing in the entire training images, so they contain much information unrelated to the specific problem of human pose estimation. This information may introduce nontrivial errors since pose estimation requires to deal with subtle differences. This motivates a supervised learning process to suppress the unrelated information so as to make the visual words particularly informative to the specific problem. On the other hand, the basic mechanism of bag-of-words involves a step of voting the local descriptors to the visual words according to the distances between the descriptors and the visual words. Eqn. 1 gives a softmax voting. And the distance metrics ($\{A_i\}_{i=1}^K$ of the Mahalanobis distance in Eqn. 1) are equivalent to a rescaling of the visual words that replace each visual word $\mathbf{c}$ with $A^{1/2}\mathbf{c}$ and applying the standard Euclidian distance to the rescaled visual words [23]. Therefore, we can rescale the visual word to suppress the unrelated information by learning a separate metric for each visual word from the labeled image-to-pose pairs. Fig. 1(c) gives six sample visual words that are typical informative local structures. For each visual word, we draw the dominant orientations whose magnitudes are rescaled by their learned metrics.

In this paper, the metric learning is jointly optimized with the learning of BME model by an iterative gradient ascent algorithm. Let $\mathcal{A} = \{A_i\}_{i=1}^K$ consisting of metrics for all visual words. Suppose the parameter set Θ is currently available for the BME model $p(\mathbf{y}|\mathbf{x}, \Theta)$, and the visual words are initially obtained by K-means. Then a simple way of defining a criterion for the desired metrics $\mathcal{A}$ is to demand that the BME model gives maximum log-likelihood on the training data. This gives the optimization problem:

$$\max_{\mathcal{A}} \; H(\mathcal{A}) = \ln p(\mathbf{y}|\mathbf{x}, \Theta) - \xi \sum_{i=1}^{K} \|I - A_i\|^2 \qquad (5)$$

$$\text{s.t. } A_i \succeq 0, i = 1, \cdots, K. \qquad (6)$$

Here $H(\mathcal{A})$ is the objective function, and $-\xi \sum_{i=1}^K \|I - A_i\|^2$ is a penalty that constrains A_i to approach diagonal as much as possible so as to reduce the complexity of the metric A_i. The penalty term also prevents $\mathcal{A}$ from drifting too much. We use a gradient ascent step to optimize $H(\mathcal{A})$,

$$\Delta_{A_i} H(\mathcal{A}) = \frac{1}{p}\frac{\partial p}{\partial x_i}\frac{\partial x_i}{\partial A_i} + 2\xi(I - A_i), i = 1, \cdots, K \quad (7)$$

where x_i, the i-th element of $\mathbf{x}$, is defined in Eqn. 1. We take derivatives on x_i, instead of $\mathbf{x}$, because x_j is independent of A_i when $j \neq i$. The BME model $p(\mathbf{y}|\mathbf{x}, \Theta)$ is differentiable with respect to x_i because both the experts and gates g are differentiable. $\partial x_i / \partial A_i$ is computed by differentiating Eqn. 1:

$$\frac{\partial x_i}{\partial A_i} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{d} \in \mathcal{D}} e^{-\rho^2(\mathbf{c}_i, \mathbf{d}, A_i)} (\mathbf{c}_i - \mathbf{d})(\mathbf{c}_i - \mathbf{d})^T \qquad (8)$$

where $\mathcal{D}$ is the set of APC descriptors, and $\mathbf{c}_i$ is the i-th visual word.

Algorithm 1. Joint Learning of Metrics and BME

```
Initialization: $A_i \leftarrow I, i = 1, \cdots, K$
repeat
  Estimate $\Theta$ for the BME model using EM
  repeat
    for each input-output pair $(\mathbf{x}, \mathbf{y})$ do
      $A_i := A_i + \alpha \Delta_{A_i} H(\mathcal{A}), i = 1, \cdots, K$
      $A_i := \arg\min_{A'} \{\|A' - A_i\|_F | A' \in \mathcal{P}\}$
    end for
  until convergence
until convergence
```

We take a gradient step $A_i := A_i + \alpha \Delta_{A_i} H(\mathcal{A})$ to update $\{A_i\}_{i=1}^K$, and then project A_i onto the set $\mathcal{P} = \{A | A \succeq 0\}$ to ensure that the constraint $A_i \succeq 0$ holds,

$$A_i := \arg\min_{A'} \{\|A' - A_i\|_F | A' \in \mathcal{P}\} \tag{9}$$

The projection step onto $\mathcal{P}$ is done by first finding the decomposition $A_i = VSV^T$ where $S = diag(\lambda_1, \cdots, \lambda_n)$ is A_i's eigenvalues and the columns of V contains A_i's eigenvectors, and then taking $A' = VS'V^T$ where $S' = diag(max\{\lambda_1, 0\}, \cdots, max\{\lambda_n, 0\})$ [23]. After obtaining the metrics $\{A_i\}_{i=1}^K$, we re-estimate the parameters for the BME model using the new metrics, and this procedure is repeated until convergence. This gives the Algorithm 1 that jointly learns the metrics and the BME model.

Until now, we consider only *on-line* learning–taking one input-output pair $(\mathbf{x}, \mathbf{y})$ for each iteration. It can be extended to *batch* learning by putting all training samples in the objective function, *i.e.*, $H(\mathcal{A}) = \sum_t \ln p(\mathbf{y}^{(t)} | \mathbf{x}^{(t)}, \Theta) - \xi \sum_{i=1}^K \|I - A_i\|^2$. But we choose on-line learning in this paper because it is much faster. Notice that our framework of jointly learning metrics can also be extended to other tasks where bag-of-words representation is used, like object recognition, as long as the cost function is differentiable with respect to $\{A_i\}_{i=1}^K$ (*e.g.*, BME and *least-square-error*).

4.3. Inference

After the BME model and distance metrics are ready, inference (state prediction) is straightforward using Eqn. 2. Giving a testing image, we extract the APC descriptors and compute the bag-of-words representation $\mathbf{x}$ by Eqn. 1. Eqn. 2 takes $\mathbf{x}$ as input, and the output is a conditional mixture distributions with components and mixing proportions that are input-dependent.

5. Experiments

We test our approach on both a quasi-synthetic dataset and a real dataset. Our quasi-synthetic dataset is generated by Poser 5, covering large appearance variations. The real dataset is the publicly available HumanEva dataset for the evaluation of human pose estimation, collected at Brown University [18]. Our approach achieves the state of the art performance on this dataset

5.1. On Quasi-synthetic Dataset

A robust and reliable human pose regressor requires training on a labeled database that is large enough to cover the variations of pose, background, illuminations, clothes, body shapes, hair style, and so on. However, collecting realistic pose labeled human databases (pairs of human image and its 3D pose) with large variations is extremely difficult because no existing system can capture accurate 3D ground truth for humans in the real world without wearing any instruments. The current available realistic databases are usually captured by commercial motion acquisition systems in engineered environments where the subjects are wearing costumes and markers, the backgrounds are indoor scenes, and the number of subjects is limited by the economic cost [11]. Therefore, we constructed a quasi-synthetic human database with large variations, by animating computer graphic human avatars using real motion data and placing the synthetic images on real backgrounds.

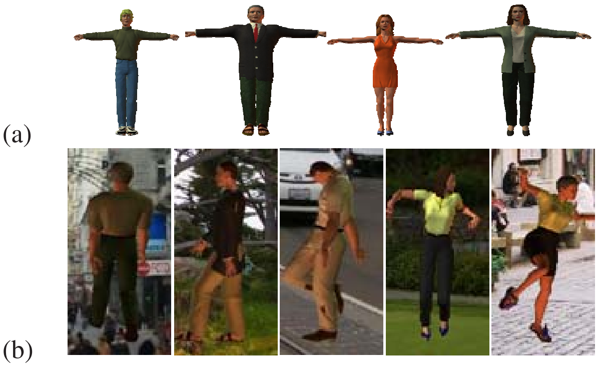


Figure 2. Quasi-synthetic dataset. (a) Some sample avatars with varying clothes, body shapes, and hair style. (b) Some sample synthetic human images. Only the human region is cropped out. See sample videos in the supplemental materials.

We constructed 376 computer graphic avatars with varying clothes, body shapes, and hair style (Fig. 2 (a) gives some sample avatars), and collected a background image pool covering natural, indoor, and street scenes. The 3D human pose has 52 degrees of freedom (DOF), 1 for global orientation and 51 for 17 joints (each upper limb has 4 joints, lower limb has 3, and chest, neck, and head has one, respectively). For each human action, we randomly choose view angles, avatars, lighting conditions, and backgrounds, and use the commercial software Poser to synthesize a human motion video (Fig. 2 (b) gives some sample images and a sample video is included in the supplemental materials). Our dataset contains various human actions, consisting of about 131,468 labeled samples, much larger and more complex than the previous quasi-synthetic datasets, like 8,262 samples in [22], 2,500 in [1], 1,200 in [9], and 9,741 in [11].

The experiment is set up as follows. We choose 60% sequences of the dataset for training and 40% are left for testing. The human detector[1] proposed in [4] is run on each image in the dataset to detect the bounding box of the human in the image. Then APC descriptors are extracted inside the bounding boxes. Both human detector and APC descriptor require no background subtraction. The human regions in

[1] The human detector is trained on a dataset containing both INRIAPerson data and synthetic data.

Table 1. Average RMS error in degrees over all angles for four settings: (1) full approach, (2) no x, y (relative positions) in APC descriptors but with metric learning, (3) no metric learning but with x, y information, and (4) using SIFT instead of APC descriptors.

	full	no x, y	no metric	SIFT
error	6.04^o	7.08^o	7.67^o	6.97^o

the bounding boxes have misalignments in some challenging images (this is common for currently available human detectors). The bag-of-words representation can handle this problem because it is invariant to translation. But the misalignment may pose difficulties on other holistic features like HOG [14]. We train a codebook of 200 visual words, and use 8 experts for the BME model.

We report mean (over all 52 angles or an individual angle) RMS absolute difference errors between the true and estimated joint angle (vectors), in degrees as in [1]:

$$D(\mathbf{y}, \mathbf{y}') = \frac{1}{m} \sum_{i=1}^{m} |(y_i - y'_i) \text{mod} \pm 180^o|. \tag{10}$$

We compare the performances on four settings: (1) full approach, (2) no x, y (relative positions) in APC descriptors but with metric learning, (3) no metric learning but with x, y information, (4) using SIFT instead of APC descriptors. Table 1 gives the average RMS errors over all angles for the four settings. Fig. 3 shows the RMS error of each individual angle normalized by the range of variation of that angle. In Fig. 3, we select only the global orientation and one angle of each joint with the biggest variation for better displaying. Table 1 and Fig. 3 show that the full approach achieves the best performance, having about 17% relative improvement. This demonstrates that the learned metrics, the APC descriptor, and the encoded position information (x, y) make a significant contribution to the pose recovery.

From Table 1, the average RMS error over all angles of our full approach is 6.04^o, but the error for individual joint angle varies depending on the range and discernibility of each joint angle. The RMS errors obtained for some key body angles are listed as follows, with that the ranges of variation of these angles in the test set are given in parentheses: global orientation: 19.65^o (360^o), right shoulder angle: 5.77^o (34.27^o), and left hip: 9.03^o (45.26^o). Our performance are numerically comparable to that in [22] (see lower part of Table 1 in [22]) and in [1] (see Fig. 8 in [1])[2]. But both [22] and [1] are based on near perfect background substraction (on their quasi-synthetic datasets), while our APC descriptors are extracted from images with cluttered background. And also our quasi-synthetic dataset is much larger and more complex than those in [22, 1].

5.2. On HumanEva Dataset

We also test the effectiveness of our approach on a real human motion dataset–HumanEva–made publicly available by the Brown Group [18]. The 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 by software. It contains multiple subjects performing a set of predefined actions with repetitions (See Fig. 6 for some sample frames). To facilitate comparison with other state of the art methods [24, 3], our first experiment uses only the walking sequences having a total of 2950 frames (first trial of subject S1, S2, and S3), as [3] did. All of the images are taken from a single camera (C1) because our approach recovers human pose from a single view. The HumanEva dataset was originally partitioned into *training*, *validation*, and *testing* sub-sets. We use walking sequences in the original training sub-set for training and those in the original validation sub-set for testing. The original testing sub-set is not used because motion data were not provided for it.

The original motion data provided by HumanEva were (x, y, z) locations of the body parts in the world coordinate system. There is a total of 10 parts: torso, head, upper and lower arms, and upper and lower legs. In this paper, we discard the internal parameters of the human body model (like limb length) as [3] did, and convert the (x, y, z) locations to global orientation of torso and relative orientation of adjacent body parts. Each orientation is represented by 3 Euler angles. As we did on the Quasi-synthetic dataset, the human region of each image is automatically cropped out by the human detector. Given a set of APC descriptors with the associated joint angles, we train a codebook of 200 visual words, learn a separate metric for each visual word, and learn a BME model with 8 experts to represent the image-to-pose distribution.

To facilitate the comparison with [24, 3], we normalize the joint angle trajectories so that $\mathbf{y}$ is a zero-mean unit

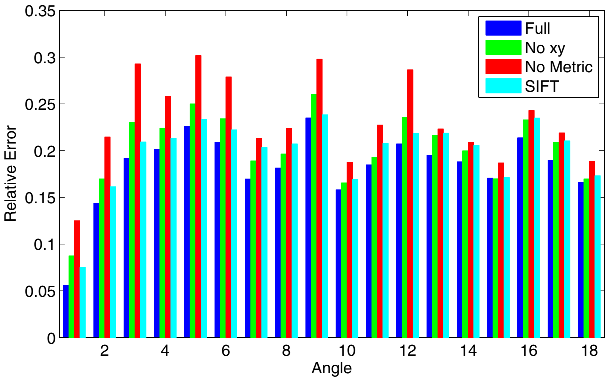


Figure 3. RMS error of each individual angle normalized by the range of variation of that angle. We select only the global orientation and one angle of each joint with the biggest variation.

[2] Actually the performances are not exactly comparable since the tested datasets are different.

Table 2. Comparison of pose estimation errors on the walking sequences. The table gives the mean and standard deviation of the relative L_2 error norm.

Algorithm	Mean	Standard Deviation	Time(s)
Zhou [24] (Walking)	0.303	0.075	40.55
Bissacco [3](Walking)	0.274	0.116	3.28
Ours (Walking)	**0.241**	**0.158**	**0.21**

Table 3. Average RMS error over all joints and over only global orientation, for sequences of walking, boxing, jogging, and combination of the three.

Sequence	Walking	Box	Jog	Combine
Ave RMS	6.68^o	5.50^o	4.12^o	6.17^o
Global	5.75^o	7.20^o	5.93^o	6.67^o

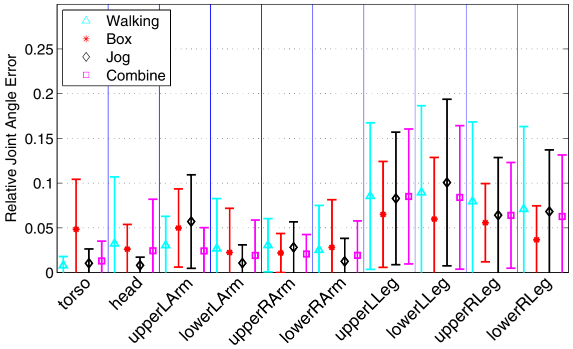


Figure 4. Mean and standard deviation of RMS error over the individual joint, normalized by the range of variation of that joint.

variance process. In this way, each angle in $\mathbf{y}$ contributes equally to the error function. We use the relative L_2 error norm [3]: $\|\hat{\mathbf{y}} - \mathbf{y}\|/\|\mathbf{y}\|$ where $\mathbf{y}$ is the ground truth and $\hat{\mathbf{y}}$ is the estimation. Table 2 shows the mean and standard deviation of the relative L_2 pose error norms on the walking sequences. Our approach outperforms the other state of the art algorithms [24, 3] in estimation accuracy[3]. And the computational speed of our approach[4] is 15 times faster than [24, 3] thanks to the integral histograms used in APC descriptor extraction and fast inference of human pose by the discriminative model (BME).

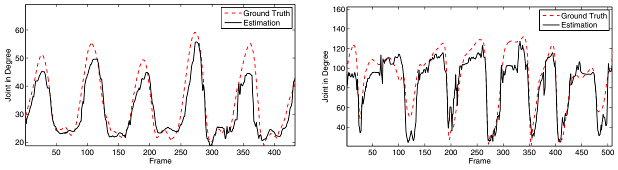
Besides the walking action, we also train and test our model on individual boxing and jogging actions, and on the combination of all three actions. Table 3 reports the average RMS error over all joints (here we use RMS error, instead of L_2 error norm, because RMS is more intuitive), and Fig. 4 reports RMS error over individual joint but the error is normalized by the range of variation of that joint. Both [24, 3] and our approach report a small error on the global orientation. The RMS error on the global orientation is 5.75^o for walking (see Table 3).

Fig. 5 plots the estimation (by the regressor trained on the combination of three actions) and ground truth of two joint angles in walking and boxing action respectively. The curves of estimation are close to the ground truth although they are less smooth. The smoothness is expected to be achieved and even accuracy improved if temporal information is added to the regressors. We leave this for future work. Fig. 6 shows some sample frames together with the estimated pose represented as the outline of a cylinder-based human model superimposed onto the original images (again the regressor is trained on the combined data). We visualize the estimated pose on cameras: C1, C2, and C3, and the ground truth on camera C1 only. Note that estimations are obtained only from images captured by camera C1. Please view the videos in the supplemental materials for better visualization.

6. Discussions and Conclusion

We stress that our approach currently does not employ any temporal information in human pose recovery. This is due to three reasons: (1) temporal information is unavailable for still images; (2) employing temporal information in pose estimation requires much extra computation cost that is a nontrivial challenge to our final goal of human action recognition; and (3) temporal smoothness can be easily achieved after the pose sequence is estimated. However, estimation accuracy is expected to be significantly improved if temporal information is employed as in [22, 3, 1]. We leave this for future work.

This paper attempts to address the problem of 3D human pose estimation from monocular images, using a **discriminative** bag-of-words approach. Unlike previous work that learns general visual words using unsupervised clustering algorithms, we use a supervised approach to learn a separate distance metric for each visual word, and the learned metrics rescale the visual words to better represent the frequent patterns existing in images that are particularly useful for the specific problem of pose estimation. The metric learning and BME model are jointly optimized by an iterative gradient ascent algorithm. We also designed a local descriptor (APC) that achieves both selectivity and in-



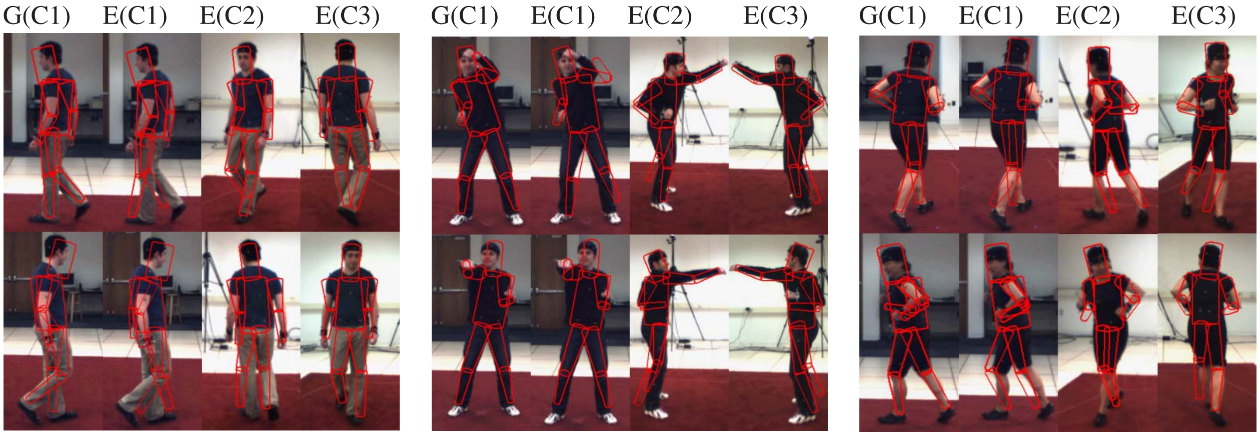
(a)

(b)

Figure 5. Joint angles: ground truth and estimation. (a) Left elbow of subject S1 in walking; (b) Right elbow of subject S3 in boxing.

[3] There are papers (*e.g.* [14]) reporting estimation errors of (x, y, z) locations of the body parts. They are not comparable with [3] and ours

[4] Our system is implemented by c code running on a PC desktop with 3GHZ Intel CPU, 2G RAM.



Walking (S2) Boxing (S3) Jogging (S1)

Figure 6. Sample estimation results. Each column shows the provided ground truth projected to camera C1 and estimation projected to cameras: C1, C2 and C3. Each row corresponds to a frame in that action sequence. G: ground truth; E: estimation. Please view the videos in the supplemental materials for better visualization.

variance for the purpose of pose estimation and requires no background subtraction. We tested our approach on both a quasi-synthetic dataset and a real dataset (HumanEva) and achieved a performance better than, or at least comparable to the other state of the art approaches.

Acknowledgment

This work was supported in part by U.S. Government VACE program, and in part by a Grant from NEC Laboratories America, Inc., USA.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *PAMI*, 2006. 2, 3, 5, 6, 7

[2] C. Bishop and M. Svensen. Bayesian mixtures of experts. *Uncertainty in Artificial Intelligence*, 2003. 4

[3] A. Bissacco, M.-H. Yang, and S. Soatto. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. *CVPR*, 2007. 6, 7

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 5

[5] A. Elgammal and C. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. *CVPR*, 2004. 2

[6] A. Fathi and G. Mori. Human pose estimation using motion exemplars. *ICCV*, 2007. 2

[7] L. Fei-Fei and P. Perona. A bayesian heirarchical model for learning natural scene categories. *CVPR*, 2005. 1, 3

[8] K. Grauman, G. Shakhnarovich, and T. Darell. Inferring 3d structure with a statistical image-based shape model. *ICCV*, 2003. 2

[9] F. Guo and G. Qian. Learning and inference of 3d human poses from gaussian mixture modeled silhouettes. *ICPR*, 2006. 2, 5

[10] M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6, 1994. 4

[11] A. Kanaujia, C. Sminchisescu, and D. Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction. *CVPR*, 2007. 2, 5

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2, 3

[13] G. Mori and J. Malik. Estimating human body configurations using shape context matching. *ECCV*, 2002. 2, 3

[14] R. Poppe. Evaluating example-based pose estimation: Experiments on the humaneva sets. *CVPR 2nd Workshop on EHuM2*, 2007. 2, 6, 7

[15] R. Rosales and S. Sclaroff. Learning body pose via specialized maps. *NIPS*, 2002. 2

[16] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. *ICCV*, 2003. 2

[17] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. *NIPS*, 2007. 2

[18] L. Sigal and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Technical Report CS-06-08, Brown University*, 2006. 2, 5, 6

[19] L. Sigal and M. J. Black. Predicting 3d people from 2d pictures. *AMDO*, 2006. 2

[20] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. *ICCV*, 2005. 1, 3

[21] C. Sminchisescu, A. Kanaujia, and D.Metaxas. Learning joint top-down and bottom-up processes for 3d visual inference. *CVPR*, 2006. 1, 2

[22] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Bm^{3}e: Discriminative density propagation for visual tracking. *PAMI*, 2007. 2, 5, 6, 7

[23] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. *NIPS*, 2005. 4, 5

[24] S. K. Zhou, B. Georgescu, X. S. Zhou, and D. Comaniciu. Image based regression using boosting method. In *ICCV*, 2005. 6, 7