Multimed Tools Appl (2014) 73:327–344
DOI 10.1007/s11042-013-1617-3

Motionlet LLC coding for discriminative human pose estimation

Li Sun · Mingli Song · Dapeng Tao · Jiajun Bu · Chun Chen

Published online: 4 August 2013

Abstract 3D human pose estimation is a challenging but important research topic with abundant applications. As for discriminative human pose estimation, the main goal is to learn a nonlinear mapping from image descriptors to 3D human pose configurations, which is difficult due to the high-dimensionality of human pose space and the multimodality of the distribution. To address these problems, we propose a novel motionlet LLC coding in a discriminative framework. A motionlet consists of training examples covering a local area in terms of image space, pose space and time stream. We first group most informative and helpful training examples into motionlets, then perform LLC Coding to learn the nonlinear mapping and get candidate poses, and finally choose the most appropriate pose as the result estimate. To further eliminate ambiguities and improve robustness, we extend our framework to incorporate multiviews. We conduct qualitative evaluation on our Taichi data set and quantitative evaluation on HumanEva data set, which show that our approach has gained

L. Sun (✉) · M. Song · J. Bu · C. Chen
Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science,
Zhejiang University, Hangzhou 310027, China
e-mail: lsun@zju.edu.cn

M. Song
e-mail: brooksong@zju.edu.cn

J. Bu
e-mail: bjj@zju.edu.cn

C. Chen
e-mail: chenc@zju.edu.cn

D. Tao
School of Electronic and Information Engineering, South China University of Technology,
GuangZhou 510640, China
e-mail: dapeng.tao@gmail.com

the-state-of-the-art performance and significant improvement against previous approaches.

Keywords Pose estimation · Multimodality · Motionlet · LLC coding · Multiview

1 Introduction

3D human pose estimation from images is a challenging but important research topic with applications in many areas including Human-Computer Interaction [26], robotics, surveillance, computer graphics and sport science. Recent approaches to 3D human pose estimation can be roughly classified into two categories, generative and discriminative. Generative approaches explicitly model human body appearance and kinematic constraints and usually concentrate on development of efficient inference methods that are able to handle the high dimensionality of human pose. Discriminative approaches directly learn the mapping from image space to pose space.

Generative approaches estimate the pose by building a geometric model associated with human body pose and evaluating how much the model agrees with human body appearance in the image. However, recovering the pose involves heavy inference because pose estimation has been transformed into a complex optimization problem [5, 9, 16, 20, 28, 31].

Discriminative approaches are popular due to their flexibility of choosing image descriptors, easy adaptation to different learning methods, no need for initialization, and most importantly, the ability of fast inference in real-world databases [27]. The main goal of discriminative 3D human pose estimation is to learn a nonlinear mapping from image descriptors to 3D human pose configurations. This is challenging due to high-dimensionality and multimodality of the mapping. Moreover, the mapping is highly noisy because of image ambiguities and subject variations.

In this paper we present a novel discriminative framework that can learn a complex mapping from image descriptors to 3D human pose configurations. We propose a local online approach to select most informative and helpful training examples for the query frame, and then group them into motionlets. As depicted by Fig. 1, every

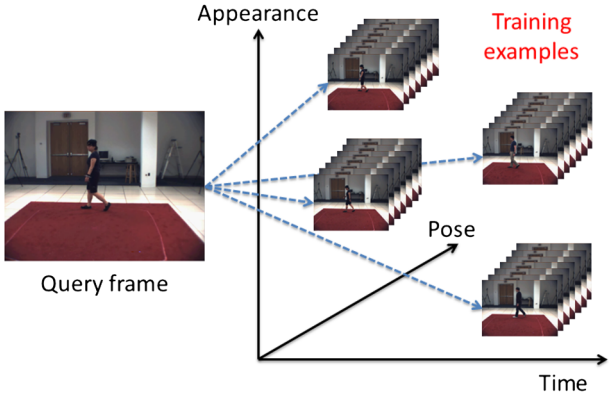

Fig. 1 Motionlets for a query frame. Each motionlet consists of training examples that cover a local region in terms of appearance space, pose space and time stream

motionlet consists of training examples that covers a local area with respect to image space, pose space and time stream. The concept of motionlets is a natural embodiment of the local motion similarity of human motion, which is the basis assumption of discriminative human pose estimation. We take advantage of Locality-constrained Linear Coding (LLC) algorithm [13] to reconstruct 3D human poses using motionlets as codebooks. LLC offers an efficient local smooth sparse projection of an image descriptor into its local-coordinate system with good reconstruction. Each motionlet contributes a candidate pose. We handle the problem of multimodality through selecting the most appropriate pose from these candidate poses. To further eliminate inference ambiguities, we extend our framework to incorporate multiviews and retain an accurate and robust inference from image descriptors to 3D human poses.

This work is an extension of our previous research in [29] with more technical details, experimental result comparison and analysis. In the following sections, we first review related work, and then present our online framework of motionlet LLC coding. We define local neighborhoods for a query frame, and then show that multimodality of the mapping is mainly caused by the multiple instances of motionlets. We demonstrate how to choose from candidate poses recovered by LLC coding and how to incorporate multiviews into our framework. Finally, we show qualitative results on our Taichi data set and quantitative results on the HumanEva-I data set [23].

2 Related work

Discriminative approaches to human pose estimation provide fast inference because they directly learn the mapping from image descriptors to human pose configurations, avoiding heavy likelihood inference. There are various image descriptors that can be flexibly incorporated into discriminative human pose estimation. These image descriptors are usually based on silhouettes [1, 2, 6, 7, 11], gradients [4, 18], or edges [3, 17, 22, 25], with different computation complexity, descriptor dimensionality, robustness against clutter, discriminating power and generalization ability. Notably, many of these image descriptors requires accurate location of the subject or background segmentation. As a consequence, the result of segmentation has substantial influence on the performance of a pose estimation algorithm. Alternatively, hierarchical multi-level image descriptors such as HMAX [14, 21], spatial pyramids [14], and vocabulary trees [14] can be used with no need for localization or segmentation. In this work, we choose HMAX [14, 21] as image descriptor based on two considerations. First, incorporating background segmentation or bounding box of the subject as a preprocessing step will cost more computation and the overall pose estimation performance will decay as the result of background segmentation or human detection deteriorates. Thus, descriptors that don't require accurate location of the subject or background segmentation are preferred. Second, hierarchical multi-level image descriptors like HMAX can provide some robustness against background clutter. In the proposed framework, HMAX serves as an appearance clue of body pose, working together with other clues of spacial and temporal constraints.

In the existing literature, various methods can be adopted to learn the mapping from image descriptors to human pose configurations, ranging from nearest-neighbor retrieval [22] and manifold embedding [7, 14] to linear/nonlinear regression [1, 33] and probabilistic mixture of predictors [15, 24]. As we mentioned in the introduction,

discriminative approaches have to model multimodality of the high-dimensional nonlinear appearance-to-pose mapping. Usually, multimodality of the mapping is represented by mixture of models, such as Bayesian mixture of experts (BME) [15, 24], mixture of probabilistic PCA [10], and mixture of multi-layer perceptrons [19]. In [8, 30, 34, 35], mixture of local gaussian process experts was used, where multimodality was handled by expert selection. Different from previous strategies, we propose to model multimodality of the mapping by motionlets, each of which contains training examples that covers a local area with respect to image space, pose space and time stream. As will be shown in following section, the main cause of multimodality of the appearance-to-pose mapping is that there usually exist multiple instances of motionlets for the query frame. Our solution directly and efficiently deal with the multimodality by selecting candidate poses contributed by these motionlets.

Recently, Local Coordinate Coding (LCC) has demonstrated promising results on learning the local geometry of data points [32]. As a variant of LCC, Locality-constrained Linear Coding [13] offers an efficient local smooth sparse projection of an image descriptor into its local-coordinate system with good reconstruction. In our discriminative pose estimation framework, we take advantage of LLC to reconstruct 3D human poses for the query frame using motionlets as codebooks.

3 Motionlet LLC coding for human pose estimation

3.1 Local motion similarity

There are three aspects of human motion that should be taken into consideration for human pose estimation: appearance, pose, and time. One key property of human motion is local motion similarity with respect to appearance, pose and time, which is the basis assumption of all discriminative human pose estimation methods. From this point of view, the improved accuracy and efficiency of recent local approaches [8, 30, 34] should be owed to their effective and efficient use of local motion similarity of human motion.

An embodiment of local motion similarity is showed by Fig. 2. Local motion similarity is represented by the similarity among training/validation frames within an local region in terms of image space, pose space and time stream, which is reflected by dark strips along inclined downward 45° in the affinity matrices for HMAX image descriptors and pose vectors. Note that several dark strips along inclined upward 45° appear in the affinity matrices for HMAX image descriptors, which are caused by ambiguities of HMAX image descriptors. Moreover, the affinity matrices for pose vectors are clean and smooth while the affinity matrices for HMAX image descriptors are noisy and jittery. We believe that it is interesting to develop a criterion of good image descriptors for discriminative human pose estimation based on this observation. Intuitively speaking, the more the affinity matrix for image descriptors resembles the corresponding affinity matrix for pose vectors in appearance, the better are the image descriptors for pose estimation. We leave it here for future study.

It also worths mentioning that there are multiple dark strips along inclined downward 45° in the affinity matrices, which will cause the problem of multimodality when one tries to learn the mapping from appearance space to pose space. Usually, training examples of a data set for human pose estimation (like HumanEva [23]) contains

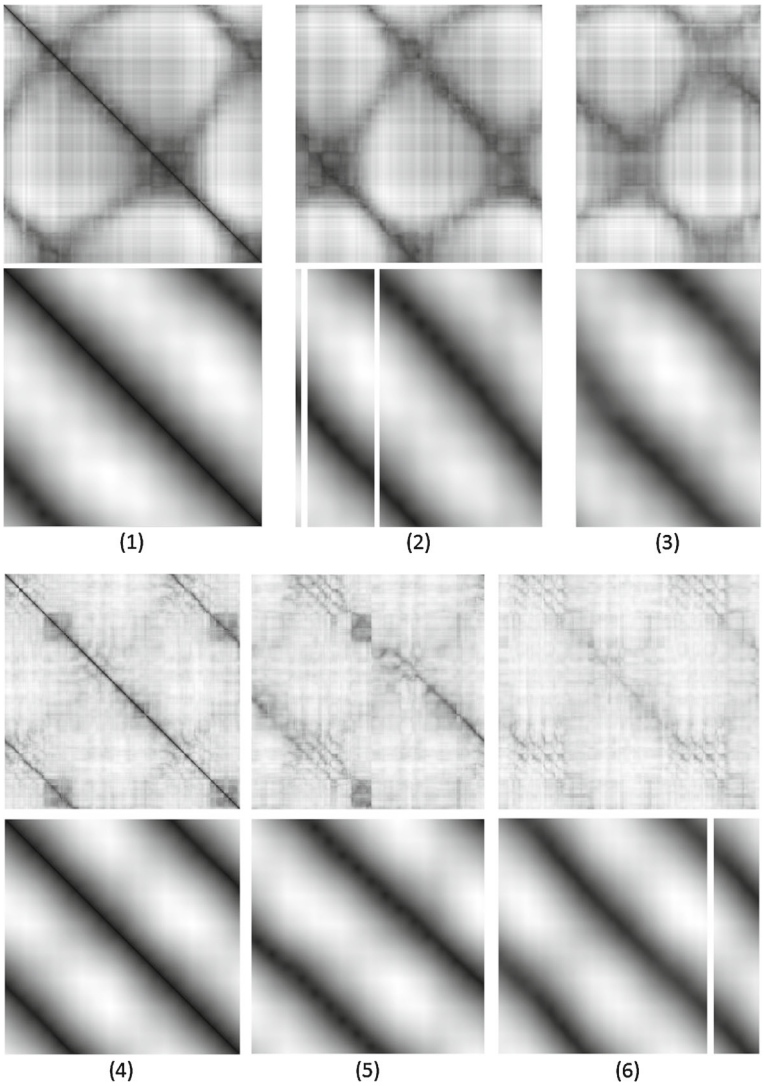

Fig. 2 Affinity matrices for camera 1 of (*1*) subject 1 training walking sequence, (*2*) subject 1 training walking sequence v.s. validation walking sequence, (*3*) subject 1 training walking sequence v.s. subject 2 training walking sequence, (*4*) subject 2 training jog sequence, (*5*) subject 2 training jog sequence v.s. validation jog sequence, (*6*) subject 2 training jog sequence v.s. subject 3 training jog sequence of HumanEva-I data set. *Odd rows* show affinity matrices of HMAX image descriptors, and even rows show affinity matrices of ground truth poses represented as vectors. *Dark values* stand for small distances. The *white bar* area in the affinity matrices is caused by absence of mocap data. Note that subject 1–3 are persons of different genders with visually different dressing

samples of multiple subject performing the same action and/or one subject performing the same action multiple times and/or multiple actions sharing some similar poses. All these contribute to multiple modes of the conditional distribution when estimating the pose from an image, which are one big source of ambiguities. In the following text, we will introduce a novel concept of motionlets to address this problem.

3.2 Motionlets for human pose estimation

As discussed previously, local motion similarity plays a very important role in discriminative human pose estimation. The concept of motionlets for human pose estimation is a natural embodiment of local motion similarity of human motion. We now formulate the definition of motionlets in the context of discriminative human pose estimation.

Let $X = (F, P)$ be a training sequence, where $F = [f_1, f_2, \cdots, f_N]$ consists of image descriptors (e.g. HMAX) of the image sequence of length N, and $P = [p_1, p_2, \cdots, p_N]$ contains corresponding ground truth poses. Given a query frame with its image descriptor f_q, there exists one or more motionlets, denoted by $M = [M_1, \cdots, M_T]$ where $T \geq 1$. As shown by Fig. 1, each motionlet covers a local region of training examples in terms of appearance space, pose space and time stream, which is given by

$$M_i = (F_i, P_i), i \in \{1, \cdots, T\} \tag{1}$$

$$F_i = [f_{a_i}, \cdots, f_{b_i}] \tag{2}$$

$$P_i = [p_{a_i}, \cdots, p_{b_i}]. \tag{3}$$

a_i and b_i are head index and tail index of the motionlet respectively, which should hold the following conditions,

1. $a_i < b_i,\ a_i, b_i \in \{1, 2, \cdots, N\}$;
2. $\forall j \in \{a_i, \cdots, b_i\},\ d(f_q, f_j) < \delta$;
3. $a_i < 1 \parallel d(f_q, f_{a_i-1}) > \delta$;
4. $b_i > N \parallel d(f_q, f_{b_i+1}) > \delta$,

where d is a distance function and δ is a threshold determining how close the image descriptors of the motionlet are with that of the query frame. Here we use Euclidean distance.

Combining the definition of motionlets and local motion similarity of human motion, it's easy to understand that there are usually multiple motionlets for a query frame. This is because there usually several training sequences, each of which usually contains several motionlets for the query frame. As mentioned in the introduction, the multiple instances of motionlets is the main cause of multimodality of the mapping in discriminative human pose estimation. By making use of the concept of motionlets, multimodality of the mapping can be handled more directly.

As depicted by Fig. 3, multiple motionlets can be extracted from training data for a query frame. Specifically, the query frame is from subject 1 test walking sequence and the motionlets are from subject 1 training walking sequence, subject 1 validation walking sequence, subject 2 training walking sequence and subject 3 training walking sequence respectively. We can see that the motionlets are subsequence of training samples possessing close appearance with the query frame, which give more valuable information about the true pose for the query frame compared with the rest training samples.

Organizing training data into motionlets can actually be beneficial in two ways. First, it naturally encodes time-sequential prior of human motion. This is helpful even we are recovering pose from a single image, because the result estimate is guaranteed

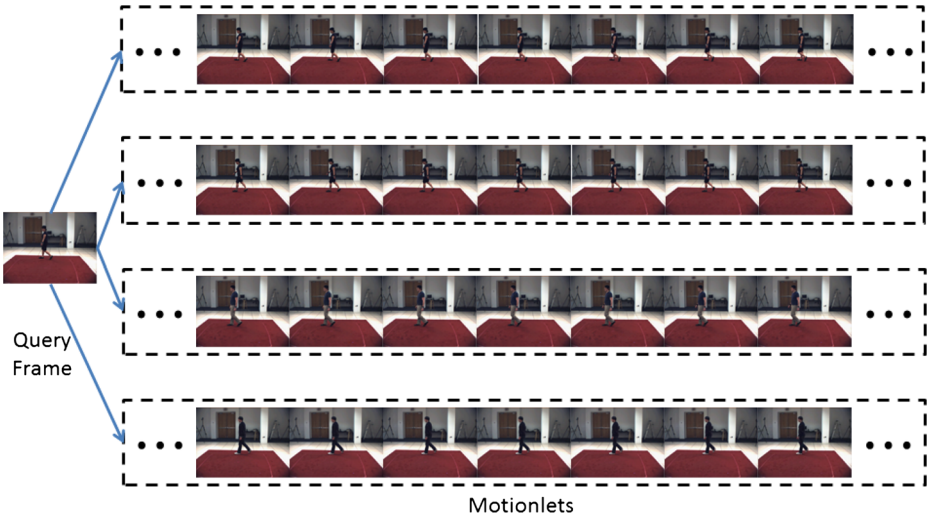

Fig. 3 Example motionlets extracted for a query frame. Better view zoomed in to see small body pose differences among each motionlet such as the distance between subject's two feets

to be based on sets of coherent training samples, not groups of irrelevant ones. Second, it reduces the number of training samples to be considered for recovering the pose, thus the computational expenses are reduced which is good for inference on large data sets.

We've formulated the concept of motionlets, explained how it relates to multi-modality of the mapping from appearance space to pose space in discriminative human pose estimation, and analyzed the benefits of incorporating the idea of motionlets. In the following text, we'll show how to integrate motionlets into a discriminative pose estimation framework.

3.3 Locality-constrained linear coding with coupled codebooks

In order to learn the high dimensional nonlinear mapping from appearance space to pose space, it is essential to capture the relation between the two spaces. Recently, Local Coordinate Coding (LCC) has shown promising results on learning the local geometry of data points [32]. Yu et al. [32] confirm that locality is essential when fitting a nonlinear function on the manifold. They find that a high dimensional nonlinear function can be approximated by a global linear function, and proposes a new method called Local Coordinate Coding. The points on the manifold can be expressed as coordinates with respect to a set of locally anchor points, which have a lower dimensional. Jinjun et al. [13] present a fast implementation of LCC called Locality-constrained Linear Coding(LLC) which introduces a locality penalty item that participates in the coordinate computation.

Assuming a dictionary with N bases $B \in \mathbb{R}^{Q \times N}$ is known, for a given data point $x \in \mathbb{R}^{Q}$, local-constrained linear coding finds a best coding $w \in \mathbb{R}^{N}$ for the input patch which minimizes the reconstruction error and the violation of the locality

constraint. Formally, this process can be formulated as optimizing the following objective function:

$$\min_w \left||x - B \cdot w|\right|^2 + \lambda \sum_{i=1}^{N} Dist_i * w_i,$$
$$s.t. \sum_{i=1}^{N} w_i = 1 \tag{4}$$

where $Dist_i = \exp\left(\frac{\left||x-B_i|\right|^2}{\sigma}\right)$ and it is the locality adaptor that gives different freedom for each basis vector proportional to its similarity to the input descriptor x. And the solution of LLC can be derived analytically as follows:

$$c^* = Norm\left(C_i + \lambda * diag\left(\exp\left(\frac{\left||x - B_i|\right|^2}{\sigma}\right)\right)\right)$$

where $C_i = (B - 1x)(B - 1x)^T$ denotes the data covariance matrix.

The data manifold in appearance space possesses similar local geometry with that in pose space. Given a query point in appearance space, we can recover the corresponding pose data using the coefficients of query point in the learned appearance subspace with respect to the pose dictionary B^P. For a query point f in appearance space, the coefficients w^* with respect to appearance dictionary B^F can be obtain by (4). Then the recovered pose p^* can be obtained by

$$p^* = \sum B^P * w^*. \tag{5}$$

In all the above discussions, the appearance dictionary B^F and the corresponding pose dictionary B^P are assumed to be known. Here we concatenate the appearance data and its corresponding pose data to get a coupled appearance-pose dictionary, where each dictionary entry B_i is can be separated into appearance part B_i^F and pose part B_i^P. Using the coupled appearance-pose dictionary, we can get a high dimensional nonlinear mapping from appearance space to pose space through LLC coding.

3.4 Motionlet LLC coding

In our local discriminative framework of human pose estimation, Local Coordinate Coding (LCC) is adopted to reconstruct 3D human poses for the query frame using motionlets as codebooks. Let $M' = \{M'_1, M'_2, \cdots, M'_{T'}\}$, where $M'_i = (F'_i, P'_i)$ and $i \in \{1, \cdots, T'\}$, be all the motionlets for the query frame from training sequences. A series of LLC coding coefficients of f_q, denoted by $C^M = \{c_1, \cdots, c_{T'}\}$, are computed by performing LLC coding on the image descriptor f_q with each of the motionlets as codebook. Then the reconstructed image descriptors and corresponding 3D human poses are given by

$$F^M = \{f'_1, \cdots, f'_{T'}\} \tag{6}$$
$$f'_i = F'_i c_i \tag{7}$$
$$P^M = \{p'_1, \cdots, p'_{T'}\} \tag{8}$$
$$p'_i = P'_i c_i, \tag{9}$$

where $i \in \{1, \cdots, T'\}$.

Now P^M contains candidate poses, each of which is contributed by one of the motionlets. The most appropriate candidate pose is selected as result estimate given by

$$p^* = p'_\theta \tag{10}$$

$$\theta = \arg\min_i [d(f_q, f'_i) + \lambda dist(f_q, F'_i)], \tag{11}$$

where λ is relative weight of the image descriptor distance term over the image descriptor reconstruction error term and the distance between a image descriptor and those of a motionlet is defined by

$$dist(f, F') = \min_{f_i \in F'} d(f, f_i). \tag{12}$$

This is much like a Nearest-Neighbor strategy. One may deem that the image descriptor reconstruction error term is positively correlated with the image descriptor distance term and one of them can be omitted. But in fact there exist cases where the former is relatively small with the latter being relatively big and vice versa.

At this end, we can inference 3D human pose from monocular image sequences. The following text will show how to exploit multiview image sequences to achieve more accurate and robust human pose estimation.

3.5 Multiview integration

For monocular human pose estimation, there are considerable ambiguities in the mapping from image space to pose space, not mentioning the ambiguities caused by image descriptors (e.g. dark strips along inclined upward 45° in the affinity matrices for HMAX image descriptors shown by the first row of Fig. 2). To account for these inference ambiguities, we extend our framework to incorporate multiviews.

When multiview sequences are available, we first estimate 3D human pose from each single view. This provides several candidate poses, each of which corresponds with one view. Again, candidate pose selection is applied to get final estimate. Unlike the method in [18] combining all views into one descriptor, our method is more robust against estimation failure occurs in one of the views, while in [18] the occurrence of inaccurate background segmentations in one or more views will always generate bad estimation result.

4 Experimental evaluation

4.1 HumanEva-I data set

To quantitatively evaluate our method, we conduct experiments on the HumanEva-I data set [23]. The HumanEva-I data set contains multiview video sequences that are synchronized with 3D body poses obtained from a motion capture system. The database contains sequences of different subjects performing several predefined actions (e.g. walking, jogging, gesturing, etc.), which are originally partitioned into training, validation, and testing sets. In this experiment, we use the original training and validation sets as training set and the original testing set for testing. 3D body pose is represented by joint positions and HMAX image descriptors are used. Except for

Table 1 Quantitative results on HumanEva-I data set: Mean 3D error in mm for HumanEva-I testing set of 3 subjects performing various actions, evaluated with a single view (C1) and multiviews (C1,BW1-BW4)

	Subject1		Subject2	
Action	Single-View	Multiview	Single-View	Multiview
Walking	43.80(19.84)	34.89(10.56)	53.66(35.39)	40.44(22.66)
Jog	70.70(37.35)	50.90(18.16)	51.59(27.06)	35.55(10.95)
Throw/Catch	NA	NA	81.69(34.23)	61.75(25.38)
Gestures	26.40(4.69)	22.14(4.71)	95.96(34.50)	68.05(15.89)
Box	92.64(28.01)	64.88(12.42)	110.98(43.73)	87.25(31.36)
Average	49.90(33.00)	37.76(19.20)	77.57(41.57)	57.68(28.52)
	Subject3		Average	
Action	Single-View	Multiview	Single-View	Multiview
Walking	50.80(40.33)	36.54(20.96)	49.80(32.28)	37.87(19.22)
Jog	61.10(31.45)	43.30(18.86)	60.13(32.41)	42.46(17.36)
Throw/Catch	NA	NA	81.69(34.23)	61.75(25.38)
Gestures	73.70(16.91)	54.96(7.56)	61.56(38.91)	45.66(23.62)
Box	121.50(68.64)	81.79(20.45)	110.60(53.19)	79.82(25.40)
Average	79.59(53.01)	55.77(25.29)	70.18(44.48)	51.52(26.82)

monocular pose recovery, we also evaluate our 3D human pose estimation method with multiviews incorporated on the data set.

We report mean 3D difference errors between estimated joint positions and ground truth joint positions in mm, relative to the pelvis (torsoDistal) joint. In the experiments, we remove frames with invalid ground truth poses from the training set.[1] With ground truth poses of testing set withheld, we use the on-line evaluation system of HumanEva project [12]. The results are reported in Table 1. Note that training sequences and testing sequences are originally from different sequences, but our method accurately infers the 3D body poses even though the poses and appearances in training and testing data might have been relatively different. In the single-view case, the errors are relatively big. This should be due to the lack of discriminating power of HMAX descriptors and ambiguities caused by them (see Fig. 2).

It worths mentioning that for different query point location in appearance space, the number and the size of motionlets vary. When the query point falls on the manifolds of all subjects performing Throw/Catch and Box actions, the motionlets for the query point tend to be small. This is because those actions are relatively violent and the distribution of training samples on the manifolds is sparse. Small motionlets will lead to ambiguity in capturing local geometry of the manifolds. As a result, the reconstructed pose by LLC coding may not be accurate. In addition, when the query point falls on the manifold of subject 1 performing Jog action, the number of motionles decreases due to that plenty macap data for the training/validation sequence is invalid. This again hurts the result estimate because less candidate poses are available.

[1] Most of the ground truth poses of Subject1 Throw/Catch sequence are invalid and those of Subject3 are unavailable.

Figure 4 shows mean 3D error plots of subject 2 performing various actions with blue thin plots for single view and red thick plots for multivew. The estimation result of single view suffers from a high degree of ambiguity. This can be understood that we lost much information by projecting 3D scene onto 2D image. For example, there is forward-backward ambiguity when the subject is walking towards/away from the camera, which caused some peaks in the plot of single view. Fortunately,

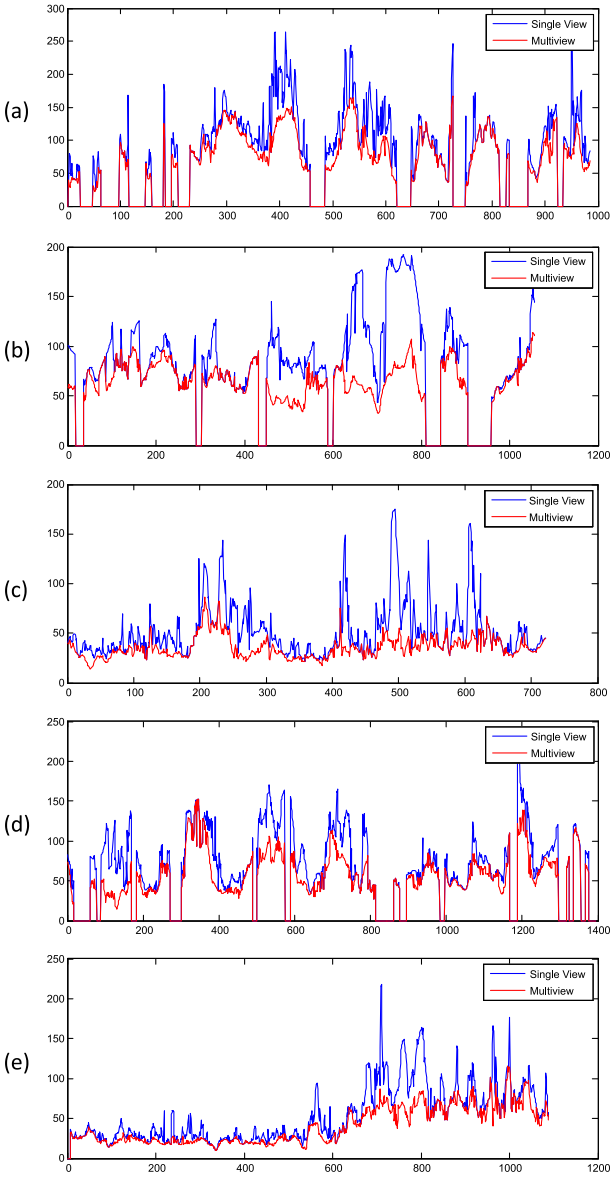

Fig. 4 Mean 3D error (in mm) plots for subject 2 **a** Walking, **b** Jog, **c** Throw/catch, **d** Gestures and **e** Box test sequences. Areas with zero error contain invalid mocap data

incorporating multiviews can resolve the ambiguity and help bring down the error. The plots for multiview are less jittery than those for single view.

Figure 5 makes a comparison of out method with Poppe's [18]. Note that in [18] foreground HOG is used as image descriptors, which relies on good background segmentation, which is a strong assumption. In the single view setting, the error of our results is larger. However, our method gets comparable or even better results in subject 1 Gestures, subject 3 Walking and subject 3 Gestures. In addition, our method outperforms [18] in the multiview setting. By incorporating multiviews, our

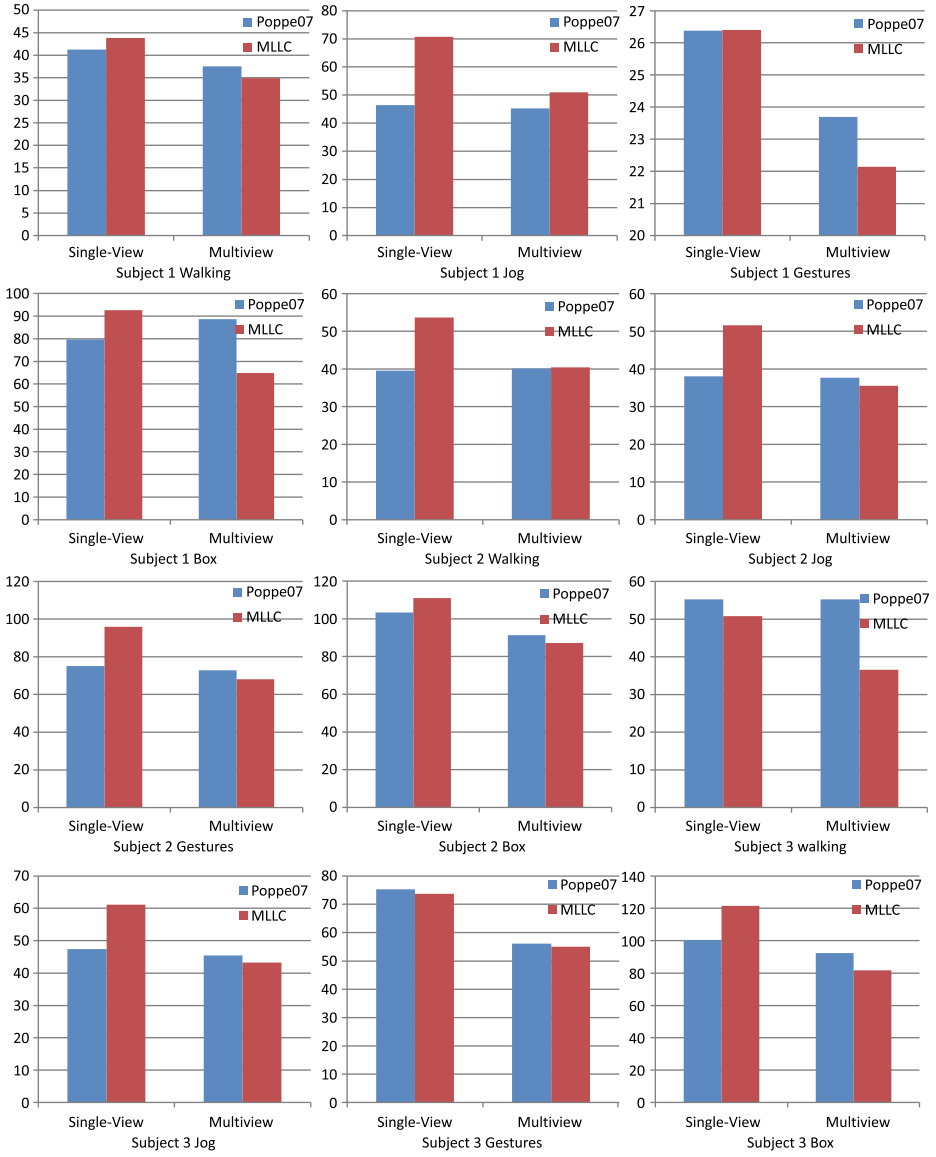

Fig. 5 Mean 3D error (in mm) of the proposed method and Poppe's method

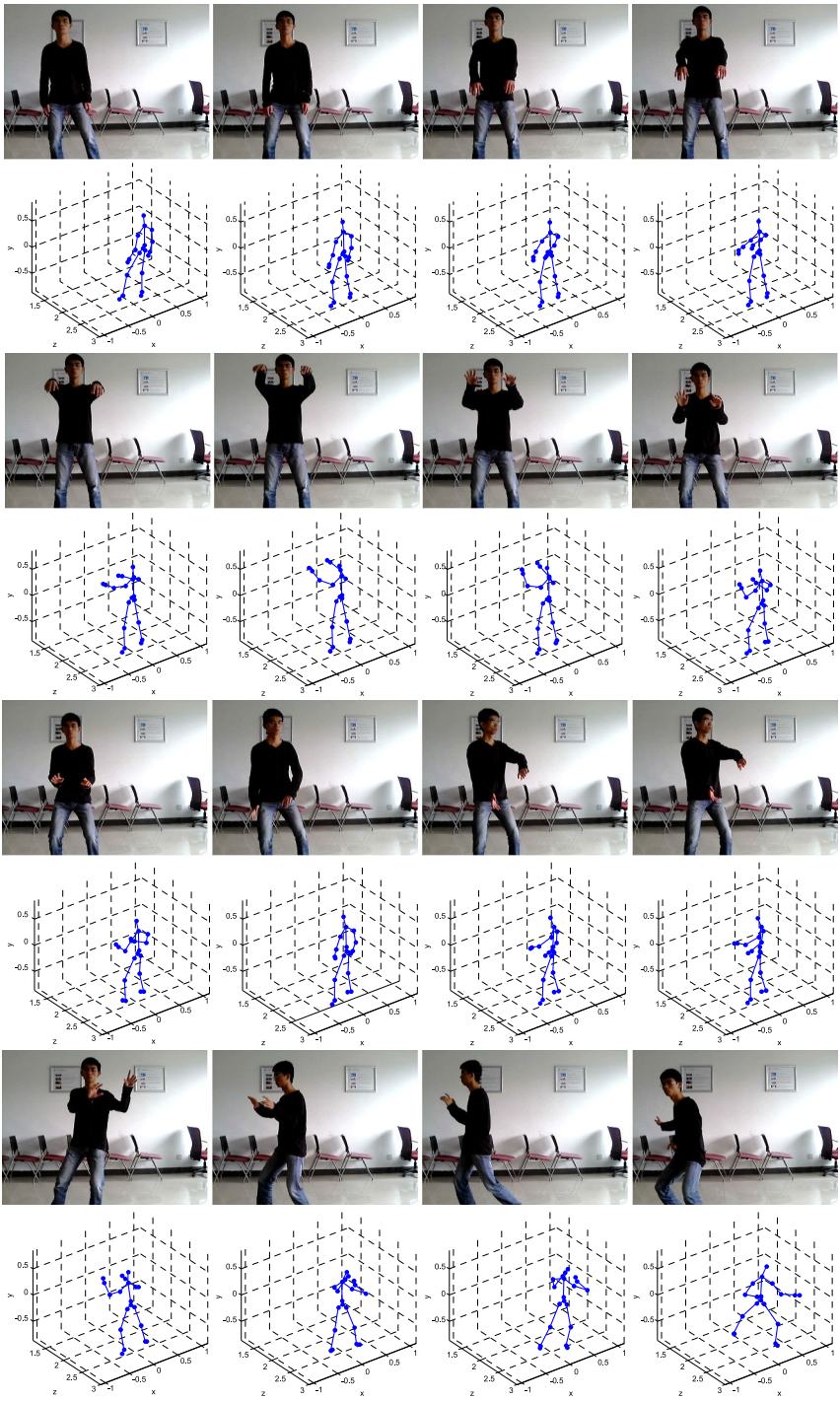

Fig. 6 Qualitative results from monocular 3D human pose estimation of our proposed method on our Taichi data set

method becomes more robust, while in [18] estimation failure in one or more views will always ruin the result estimate.

4.2 Taichi data set

In this experiment, we evaluate our method qualitatively on our Taichi data set. For the Taichi data set, we collect monocular image sequences of a subject exercising Taichi with synchronized ground truth pose captured by two Microsoft Kinect sensors. 3D pose is represented as a vector of 20 concatenated 3D joint positions and is estimated from HMAX image descriptors.

Figure 6 depicts several test frames of Taichi data set and their corresponding estimated 3D human poses. Note that in some cases it's hard to distinguish arms from torso due to very dark clothes and self occlusion, but our method can accurately infers 3D poses under such condition.

5 Conclusions

In this paper we presented a local online framework for 3D human pose estimation, which is able to learn a complex, high-dimensional, and multimodal nonlinear mapping from image descriptors to 3D human poses. We have formulated the concept of motionlets and showed that multimodality of the mapping is mainly caused by the multiple instances of motionlets for each query frame. We directly handle multimodality of the mapping by first group most informative and helpful training examples into motionlets, then perform LLC Coding to learn the nonlinear mapping and get candidate poses, and finally choose the most appropriate pose as the result estimate. To improve accuracy and robustness, we extend our framework to incorporate multiviews. We conducted the experiments on our Taichi data set and the real HumanEva-I data set to evaluate our proposed method qualitatively and quantitatively, and achieved accurate results. In future work, we plan to develop a method to assess image descriptors and to find good image descriptors for discriminative human pose estimation.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (61170142), National Key Technology R&D Program (2011BAG05B04), International Science & Technology Cooperation Program of China (2013DFG12840), and the Fundamental Research Funds for the Central Universities.

References

1. Agarwal A, Triggs B (2004) 3D human pose from silhouettes by relevance vector regression. In: CVPR
2. Agarwal A, Triggs B (2006) Recovering 3D human pose from monocular images. PAMI 28(1):44–58
3. Agarwal A, Triggs B (2006) A local basis representation for estimating human pose from cluttered images. In: ACCV
4. Bo L, Sminchisescu C (2010) Twin gaussian processes for structured prediction. In: IJCV

5. Duan K, Batra D, Crandall D (2012) A multi-layer composite model for human pose estimation. In: BMVC
6. Elgammal A, Lee C (2004) Infering 3D body pose from silhouettes using activity manifold learning. In: CVPR
7. Elgammal A, Lee C-S (2007) Nonlinear manifold learning for dynamic shape and dynamic appearance. CVIU 106(1):31–46
8. Fergie M, Galata A (2010) Local Gaussian processes for pose recognition from noisy inputs. In: BMVC
9. Felzenszwalb PF, Huttenlocher DP (2005) Pictorial structures for object recognition. IJCV 61(1):55–79
10. Grauman K, Shakhnarovich G, Darell T (2003) Inferring 3D structure with a statistical image-based shape model. In: ICCV
11. Howe NR (2007) Silhouette lookup for monocular 3D pose tracking. Image Vis Comput 25(3):331–341
12. HumanEva project (2007) http://vision.cs.brown.edu/humaneva/
13. Jinjun W, Jianchao Y, Kai Y, Fengjun L, Huang T, Yihong G (2010) Locality-constrained linear coding for image classification. In: CVPR
14. Kanaujia A, Sminchisescu C, Metaxas D (2007) Semi-supervised hierarchical models for 3D human pose reconstruction. In: CVPR
15. Ning H, Wei X, Gong Y, Huang T (2008) Discriminative learning of visual words for 3D human pose estimation. In: CVPR
16. Lee MW, Chohen I (2004) Human upper body pose estimation in static images. In: ECCV
17. Ong E-J, Micilotta AS, Bowden R, Hilton A (2006) Viewpoint invariant exemplar-based 3D human tracking. CVIU 104(23):178–189
18. Poppe RW (2007) Evaluating example-based pose estimation: experiments on the Humaneva sets. Tech. Report TR-CTIT-07-72, University of Twente
19. Rosales, R, Sclaroff S (2002) Learning body pose via specialized maps. In: NIPS
20. Sapp B, Toshev A, Taskar B (2010) Cascaded models for articulated pose estimation. In: ECCV
21. Serre T, Wolf L, Poggion T (2005) Object recognition with features inspired by visual cortex. In: CVPR
22. Shakhnarovich G, Viola PA, Darrel T (2003) Fast pose estimation with parameter-sensitive hashing. In: ICCV
23. Sigal L, Black M (2006) Humaneva: synchronized video and motion capture dataset for evaluation of articulated human motion. Tech. Report CS-06-08, Brown University
24. Sminchisescu C, Kanaujia A, Li Z, Metaxas D (2005) Discriminative density propagation for 3D human motion estimation. In: CVPR
25. Sminchisescu C, Kanaujia A, Metaxas D (2006) Learning joint top-down and bottom-up processes for 3D visual inference. In: CVPR
26. Song M, Tao D, Liu Z, Li X, Zhou M (2010) Image ratio features for facial expression recognition application. TSMCB 40(3):779–788
27. Song M, Tao D, Li X (2010) Visual context boosting for eye detection. TSMCB 40(6):1460–1467
28. Stenger B, Thyananthan A, Torr PHS, Cipolla R (2006) Model-based hand tracking using a hierarchical Bayesian filter. PAMI 28(9):1372–1384
29. Sun L, Song ML, Bu JJ, Chen C (2012) Pose estimation with motionlet LLC coding. In: PCM
30. Urtasun R, Darrel T (2008) Local probabilistic regression for activity-indenpendent human pose inference. In: CVPR
31. Yang Y, Ramanan D (2011) Articulated pose estimation with flexible mixture-of-parts. In: CVPR
32. Yu K, Zhang T, Gong Y (2009) Nonlinear learning using local coordinate coding. In: NIPS
33. Zhao X, Ning H, Liu Y, Huang T (2008) Discriminative estimation of 3D human pose using Gaussian processes. In: CVPR
34. Zhao X, Fu Y, Liu Y (2009) Temporal-spatial local Gaussian processes experts for human pose estimation. In: ACCV
35. Zhao X, Fu Y, Liu Y (2011) Human motion tracking by temporal-spacial local Gaussian process experts. TIP 20(4):1141–1151

Li Sun is currently a PhD student in the College of Computer Science, Zhejiang University, Zhejiang, China. His research interests include computer vision and machine learning.

Mingli Song received the Ph.D. degree in computer science from Zhejiang University, Zhejiang, China, in 2006.

He is currently an Associate Professor with the College of Computer Science and the Microsoft Visual Perception Laboratory, Zhejiang University. His research interests include visual perception analysis, image enhancement, and face modeling.

Dapeng Tao received the B.S. degree in Electronics and Information Engineering from Northwestern Polytechnical University, Xi'an, China. He is currently a Ph.D. candidate in Information and Communication Engineering at South China University of Technology, Guangzhou, China. His research interests include machine learning, computer vision and cloud computing.

Jiajun Bu is currently a Professor with the College of Computer Science, Zhejiang University, Zhejiang, China. His research interests include information retrieval, computer vision, and embedded system.

Chun Chen is currently a Professor with the College of Computer Science, Zhejiang University, Zhejiang, China. His research interests include computer vision, computer graphics, and embedded technology.