Human motion data refinement unitizing structural sparsity and spatial-temporal information

Zhao Wang
Bournemouth University
Email: zwang@bournemouth.ac.uk

Shuang Liu
Bournemouth University
Email: sliu@bournemouth.ac.uk

Rongqiang Qian
Xi'an Jiaotong-Liverpool University
Email: rongqiang.qian@student.xjtlu.edu.cn

Tao Jiang
Bournemouth University
Email: tjiang@bournemouth.ac.uk

Xiaosong Yang*
Bournemouth University
Email: xyang@bournemouth.ac.uk

Jian J Zhang
Bournemouth University
Email: jzhang@bournemouth.ac.uk

***Abstract*—Human motion capture techniques (MOCAP) are widely applied in many areas such as computer vision, computer animation, digital effect and virtual reality. Even with professional MOCAP system, the acquired motion data still always contains noise and outliers, which highlights the need for the essential motion refinement methods. In recent years, many approaches for motion refinement have been developed, including signal processing based methods, sparse coding based methods and low-rank matrix completion based methods. However, motion refinement is still a challenging task due to the complexity and diversity of human motion. In this paper, we propose a data-driven-based human motion refinement approach by exploiting the structural sparsity and spatio-temporal information embedded in motion data. First of all, a human partial model is applied to replace the entire pose model for a better feature representation to exploit the abundant local body posture. Then, a dictionary learning which is for special task of motion refinement is designed and applied in parallel. Meanwhile, the objective function is derived by taking the statistical and locality property of motion data into account. Compared with several state-of-art motion refine methods, the experimental result demonstrates that our approach outperforms the competitors.**

Keywords: Motion Capture Data, Motion Refinement

Fig. 1: Examples of MOCAP equipments: (1) optical based (upper left), (2) wearable sensors (upper right) and (3) depth sensors (bottom)

Human motion capture (MOCAP) data is now widely used in many areas such as computer animation, digital effect, gaming, physical training, virtual reality and medical rehabilitation. For the film industry, the high quality motion data have been applied to generate the character animation, facial animation and special digital effects in the recent fantastic films e.g. *Avatar, The Avengers, Transformers, Captain America,* and *Warcraft*. The great success demonstrates the importance of MOCAP techniques and data.

These MOCAP data based approaches require high quality raw data as input. Currently, the most popular commercial MOCAP systems are optical-based, such as *Vicon* [1] and *Motion Analysis* [2]. However, even with these professional systems, the acquired raw data still suffers from missing marker problems. For example, the markers may become invisible when they are occluded by other body parts or objects, which could lead to missing data problem. A piece of recorded motion by *Motion Analysis* MOCAP system is shown in Fig 2, where the blanks on the time line denote the occurring of missing marker problem. The process of capturing human motion is usually both expensive and time consuming. Hence, it is essential to refine the captured raw motion data to meet the quality requirement rather than tedious reshooting. In practice, some post-processing tools for cleaning motion data, e.g. filling missing value and removing noise, are provided in the commercial MOCAP systems. However, such tools usually require user to correct the outliers and noise of recorded motion sequence frame-by-frame, which could lead error-prone, tedious and time consuming. In addition, the most often

[1] http://www.motionanalysis.com/

[2] http://www.vicon.com/

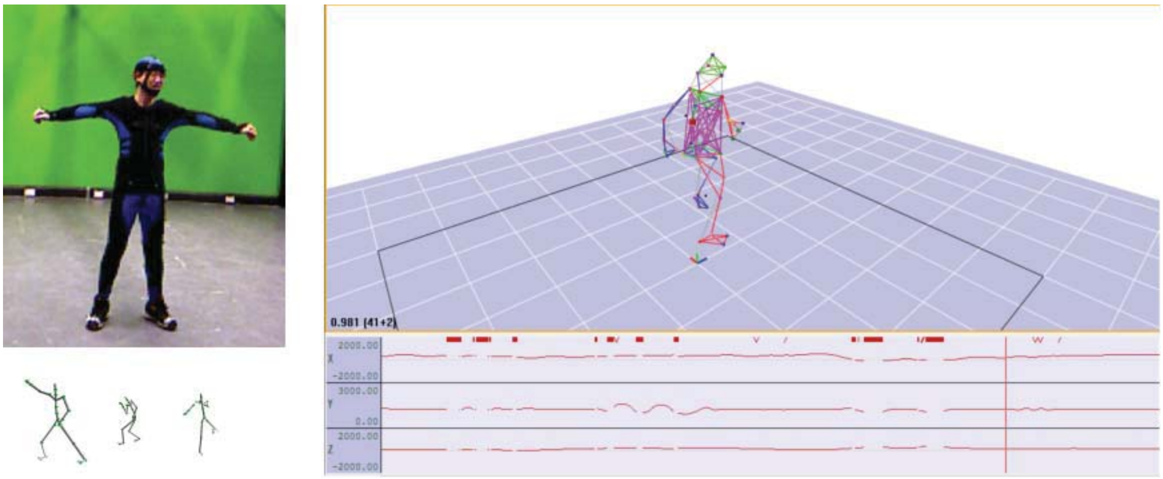


Fig. 2: Examples of MOCAP result recorded by *Motion Analysis System*

used refinement methods used in these commercial MOCAP systems are linear/spline interpolation, which is only effective for simple motion e.g. walking and running. The refinement may fail while dealing with the complex motion. Moreover, the spatio-temporal patterns have been ignored by those methods, which could cause distortion and unrealistic in refinement result. Additionally, the fast developing low-cost depth sensors (e.g. *Microsoft Kinect*, *Google Project Tango*), which are able to acquire a depth stream with acceptable accuracy, can provide new opportunities for accessible motion capture. The motion data derived from the depth stream contains even more noise than the result from current MOCAP system. Although many work have been done on this topic [1]–[3], improving the quality of motion data is still a long uphill journey.

To refine the imperfect motion, a lot of methods have been developed in the literature. The rising of novel motion capture systems and technologies brings explosive growth of motion data in recent years. The data-driven based motion processing methods have attracted many attentions [4]–[8], and achieved many successes for motion denoising. However, in the existing work for filling missing markers, such as Lou et.al. [4] and Xiao et.al. [9], the spatial temporal and kinematic information of human motion haven't been well exploited while training the motion dictionaries. The artifact e.g. dithering could occur in the recovered motion sequence. Therefore, in order to overcome these problems, we propose a novel motion refinement method deriving from sparse coding and dictionary learning in this paper, which focus on solving missing marker problem. The major contributions of our work are

- taking the distribution information of missing marker in motion data into account for deriving the dictionary learning.
- selecting a compact correlated subset of motion bases for the clean motion reconstruction.
- exploiting the spatial-temporal information while learning the motion dictionary to achieve stable and realistic result.
- taking the smooth constraint into account for the motion recovery for ensuring the smooth result. A smooth graph constraint on the sparse representation coefficients matrix is employed in our objective function

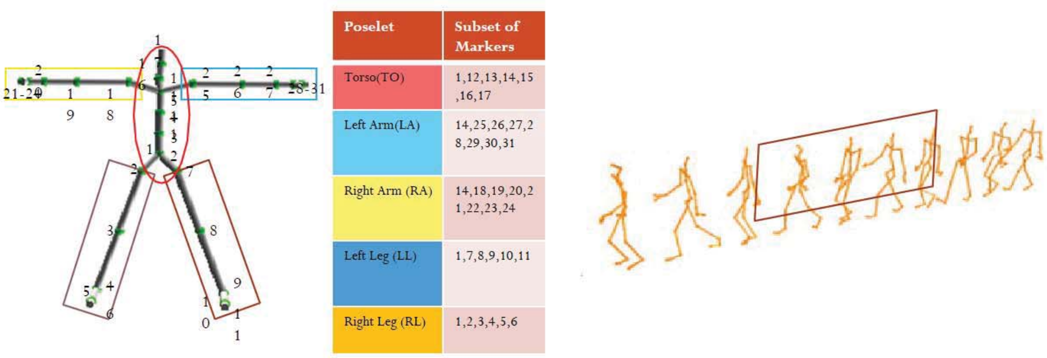
I. Related work

The purpose of human motion refinement is to remove the noise and fill the missing value while preserving the embedded spatio-temporal patterns of motion. Due to the complexity and diversity of MOCAP data, motion refinement is a challenging task, where much effort has been expended on this topic. Generally, the existing approaches of MOCAP data refinement could be divided into three categories: signal processing methods, data-driven methods and matrix completion methods.

A. *Signal processing methods*

In early studies, the classical signal denoising methods such as Gaussian low-pass filter, wavelet transformation, discrete cosine transform (DCT) and Fourier transform have been applied to denoise the motion data [10]. For instance, Hsieh and Kuo have proposed a B-spline wavelet-based method to remove the impulsive noise of body motion data [10]. Another way is to apply linear time-invariant filters (LTI) to refine the noisy motion data [11], [12]. As an improvement, the dynamic system-based methods (DSB) such as Kalman filter and linear dynamic system (LDS) are employed for motion refinement [13], [14]

Dimension reduction (DR) methods have also been applied to motion signal for refinement [15]–[17]. For example, principle component analysis (PCA) can be used to eliminate non-informative components of the motion data by accounting the variance of motion data on some orthogonal directions [15]. Independent component analysis (ICA) is another good choice to reveal the independent latent factors that contribute to generating different kinds of motion [18]. Inspired by the great success of manifold learning on computer vision areas [19], such kind of methods have also been applied to motion



Poselet	Subset of Markers
Torso(TO)	1,12,13,14,15 ,16,17
Left Arm(LA)	14,25,26,27,2 8,29,30,31
Right Arm (RA)	14,18,19,20,2 1,22,23,24
Left Leg (LL)	1,7,8,9,10,11
Right Leg (RL)	1,2,3,4,5,6

Fig. 3: Partial model for CMU motion data. The markers 1, 2, 7, and 14 are the root, the right and left femur markers, and the upper neck marker, respectively., which are used for local coordinate translation. For the grouping operation, with a given S frames sequences and size M window, it will generate $N = S - M + 1$ overlapping clips.

denoising [16] which could also be regarded as a special kind of DR method.

Signal processing methods usually do not require much computational cost and are effective while dealing with simple and short-term motion. However,this kind of methods process each joint degree of freedom (DOF) independently, where the underlying structure correlation between human joints are usually ignored. Hence, the of refining result of signal processing methods on complex motion may be notS sufficient to satisfy the quality requirement .

B. Data-driven methods

The rising of novel motion capture systems and technologies brings explosive growth of motion data in recent years, which facilitate the development of data-driven based methods. [4]–[6], [8], [9], [20]. For example, Lou and Chai [4] have proposed an example based approach to learn a series of spatial-temporal filter bases from pre-captured motion data and use them along with robust statistics techniques to fill in the missing values of motion capture data. In Xiao et al's work [6], [9], they have formulated the predicting missing marker problem as finding spare representation of imperfect pose. They succussed in introducing ℓ_1 sparse representation to solve predicting missing marker of motion data. Hou et al. [20] have provided a method to recover corrupted motion capture data through trajectory-based sparse representation.

The performance of data-driven methods is heavily rely on the training data selection. In addition, many existing data driven methods didn't consider kinematic characteristics and smooth property of human motion in their dictionary learning and pose reconstruction process, which could cause artifact in the recovery result. Additionally, data-driven methods often meet out-of-sample problem, where they are unable to handle the new coming motion sequence when there are no similar motion in the training dataset.

C. Matrix completion methods

Another typical motion refinement method is matrix completion based, which formulates the human motion refinement into a low-rank matrix optimization task [21]–[25]. Lait et al. [21] have noticed the low-rank property of motion matrix has not been exploited explicitly. They reformulate the human motion refinement into a low-rank matrix optimization where singular value thresholding (SVT) is applied to solve the objective function. After that, Feng et al. [22] have proposed a motion data refinement via a matrix completion method using both the low-rank structure and temporal stability properties of the motion data. Liu et al. [23] have presented a MOCAP data denoising approach via filtered subspace clustering and low rank matrix approximation. Recently, Burke and Lasenby [26] have tried to combine the smoothing and low-rank matrix completion by projecting markers into a lower dimensional space learned from the motion sequence, performing Kalman smoothing in this space using and then returning to the original space, using correlated markers to reduce the average error in each marker position estimate.

Matrix completion methods do not require the pre-training, which means that there is no out-of-sample problem. This is the biggest advantage of such kind of methods. However, the matrix completion method may fail when many data entries are badly corrupted,e.g. large amount of missing markers.

Arguably, the human motion refinement is still an difficult problem due to the diversity and complexity of motion data. Inspired by the great success of data driven based methods in computer vision and computer graphic area, we aim to propose a novel human motion refinement method based on sparse representation to overcome the existing the missing marker filling issue.

II. Methodology

A. Data preprocessing

1) Normalization and Coordinate translation: MOCAP data is usually recorded under the real world global coordination.

Even visual-similar motion could have dramatically numerical diversity due to the pose translation and rotation. Therefore, a local coordinate transformation would be applied to the raw data which aims to remove the effect of pose translation and rotation. In addition, we noticed that in various kinds of motion, the torso is usually a rigid part. Hence, we will translate each pose to the local coordinate representation respect to the root marker, i.e. marker 1 for CMU motion data. Then, the local pose frames will be rotated to ensure that the rigid plane, which is consisting of 3 markers, i.e.markers 2 (right femur), 7 (left femur), and 14 (upper neck) for CMU motion data, parallels to the XY plane.

2) Human partial model and grouping: Human motion data intrinsically countains hierarchical spatial-temporal information. In order to better exploit the spatial-temporal relationship, many researches have applied the partial human model while processing motion data [5], [6], [25], [27], [28] and achieved expressive performance. Instead of using whole body model [8], [9], We choose partial human model in this work and divide the whole body into 5 parts [5], [6], which is *Torso(TO), Left Arm (LA), Right Arm (RA), Left Leg(LL) and Right Arm(RL).* On one hand, the partial model could facilitate exploring the hierarchical spatial correlations among the joints. On the other hand, it is helpful for improving the model's generalization ability.

In addition, we chooses using short clips of motion rather than processing the refinement frame by frame, which aims to obtain embedded spatialtemporal patterns and guarantee smoothness for the result motion sequence.

Therefore, for a given motion sequence $X = \{X_1, X_2, \dots, X_S\}$ contains S pose frames, the submatrix X^i will be derived from X to represent each partial motion sequence, as $X^i = \{X^i_1, X^i_2, \dots, X^i_S\} \in R^{d^i \times S}, i = 1, 2, \dots, 5$. With a M length window, it will then generate $N = S-M+1$ overlapping motion clips for each partial motion sequence , that is $X(M)^i_j = [X_{(j-1)\times M+1}, \cdots, X_{j\times M}]$. In each clip, we reshape the M frames into one vector Y^i_j, i.e. $R^{d_i \times M} \rightarrow R^{(d^i)\times 1}$, $d^i = M \times d_i$. Thus, we finally get the groups of partial motion matrixes $Y^i = \{Y^i_1, Y^i_2, \dots, Y^i_N\} \in R^{d^i \times N}, N = S - M + 1, i = 1, 2, \dots, 5$.

B. Motion dictionary learning

Assume that $Y^i = [Y^i_1, Y^i_2, \cdots, Y^i_N] \in R^{d^i \times N}, i = 1, 2, \cdots, 5$. stand for the partial motion group set generated from clean motion clips via the pre-processing operation mentioned in section II-A. A conversation dictionary learning is to solve the following problem to extract the most suitable dictionary $D^i \in R^{d^i \times K^i}$ for the sparse representation of training partial motion group Y^i.

$$\begin{aligned} &\min_{W^i, D^i} \|Y^i - D^i W^i\|^2_F \\ &s.t. W^i_j = [W^i_1, \dots, W^i_N],\ \|W^i_j\|_0 \le t_s, 1 \le j \le N \\ &\quad D^i = [D^i_1, \dots, D^i_K],\ \|D^i_m\|_2 \le 1, 1 \le m \le K^i \end{aligned} \tag{1}$$

In equation 1, W^i is the sparse coefficient, t_s is the target sparsity, D^i is the motion dictionary corresponding to the 5 kinds of human pa4rtial motion. Equation 1 is a non-convex problem and could be solved by some existing methods,e.g. K-SVD [29]. However, the equation 1 is a least square error function which is not stable to the noise and missing value. We will enhance the robustness of equation 1 for dealing with motion data.

The acquired human motion data usually contains only a few of missing markers after post processing. The distribution information of missing markers in motion data will be taken into account for deriving the dictionary learning to improve the objective function. The missing markers are mainly caused by the occlusion and usually last several continuous frames, which are structural sparse. Let's assume that the binary matrix $\Omega_i \in \{0, 1\}^{d^i \times N}, i = 1, 2, \dots S$ denotes the missing feature (i.e. 1 for corresponding marker miss) of a given partial motion clips Y^i. Hence, the missing part could be denoted as $\Omega^i \circ Y^i$ while the observable part is $\overline{\Omega}^i \circ Y^i$.

We relax ℓ_0 pseudo-norm in equation 1 to a ℓ_1 minimization. The objective function is then reformulated and the idea dictionary would provide the sparse representation via satisfying

$$\begin{aligned} &\arg\min_{W_i} \|\overline{\Omega}^i \circ Y^i - \overline{\Omega}^i \circ D^i W^i\|^2_F + \lambda \|W^i\|_1 \\ &s.t. \|\Omega^i \circ Y^i - \Omega^i \hat{Y}^i\|^2_F < \sigma \\ &\quad D^i = [D^i_1, \dots, D^i_K], \|D^i_j\|_2 \le 1, 1 \le m \le K^i \end{aligned} \tag{2}$$

where $\hat{Y}^i = D^i W^i$ is the reconstructed result and the constraint $\|\Omega^i \circ Y^i - \Omega^i \hat{Y}^i\|^2_F$ aims to minimize the difference of the invisible part between the clean motion and reconstructed result.

Equation 2 is actually a nonconvex problem with respect to D^i and W^i jointly, which is difficult to find the global minimum. However, equation 2 is convex with the two variables separately. Hence, the variables D^i and W^i would be optimized alternatively until the convergence is achieved. Finally, five motion dictionaries $D^i, i = 1, \cdots, 5$ can be gotten in the training phase via the proposed dictionary learning algorithm.

C. Motion recovery

1) Trust Data Detection: As mentioned in the previous paragraph, apart from the dictionaries D and parameter λ, our approach also need to specify the missing marker Ω while processing the motion data. Here, we employ a trust data detection (TDD) method [22] to identify the missing data entries.

$$\begin{aligned} &\Omega = \text{TDD}(X, \phi) \\ &s.t. \Omega = [\Omega_1, \dots, \Omega_S] \in \{0, 1\} \end{aligned} \tag{3}$$

where X is the given noisy motion sequence, Ω is the corresponding marker and ϕ is the threshold value which is set as $6cm$ in this work. The detail of TDD implementation is omitted due to paragraph limitation, which is available in [22].

2) Objective function: For a given input imperfect motion sequence, we will firstly take the operations mentioned in section II-A to generate the five partial-group motion matrices, which denoted as $\{Y^i \in d^i \times N, i = 1, 2, \cdots, 5\}$. The corresponding missing mark Ω would be detected via TDD and also be translated to $\{\Omega^1, \cdots, \Omega^5\}$ via similar operations. In order to simplify the problem, the basis number K^i for each partial motion dictionary is all set as K, that is $\{D^i \in d^i \times K, i = 1, 2, \cdots, 5\}$. With the pre-trained five dictionary matrices $\{D^1, \cdots, D^5\}$, the reconstructed result groups $\{Y^i, i = 1, 2, \cdots, 5\}$ could be calculated by solving a $\ell_1 - norm$ minimization framework:

$$\arg\min_{W_i} \|\overline{\Omega}^i \circ Y^i - \overline{\Omega}^i \circ D^i W^i\|_F^2 + \lambda \|W^i\|_1 \tag{4}$$

Since the $\ell_1 - norm$ penalty in equation 4 on the coefficient W^i is not able to promise the smoothness of reconstructed result, a locality-constrained linear (LLC) coding method [30] is used. Hence, we reformulate the objective function as

$$\arg\min_{W_i} \|\overline{\Omega}^i \circ Y^i - \overline{\Omega}^i \circ D^i W^i\|_F^2 + \lambda \|G^i \circ W^i\|_2 \tag{5}$$

where $G^i \in R^{K \times N}$ is the locality adaptor that each column gives the different freedom for each basis vector proportional to its similarity to the input descriptor $Y^i_{:,j}, j = 1, 2, \cdots, N$. Specifically,

$$\begin{aligned} G^i_{:,j} &= \exp(\frac{dist(\tilde{Y}^i_{:,j}, \tilde{D}^i)}{\sigma}) \\ \tilde{Y}^i &= \overline{\Omega}^i \circ Y^i, \ \tilde{D}^i = \overline{\Omega}^i \circ D^i W^i \end{aligned} \tag{6}$$

where $dist(Y^i_{:,j}, D^i) = [dist(Y^i_{:,j}, D^i_{:,1}), dist(Y^i_{:,j}, D^i_{:,2}), \cdots, dist(Y^i_{:,j}, D^i_{:,K})]^T$, and $dist(Y^i_{:,j}, D^i_{:,p})$ is the Euclidean distance between $Y^i_{:,j}$ and $D^i_{:,p}$. Each column of G^i is normalized to be between $(0, 1]$. Note that the LLC code in equation 5 is not sparse in the sense of ℓ_0 norm, but is sparse in the sense that the solution only has few significant values [30]. In practice, a threshold is applied to make those small coefficients be zero.

Solving the $\ell_1 - norm$ problem like equation 4 usually requires optimization procedures, e.g. Feature Sign algorithms [31], which is time consuming. Unlike equation 4, the solution of equation 5 can be derived analytically by:

$$\begin{aligned} \tilde{W^i_{:,j}} &= (C^i_j + \lambda \ diag(G^i)) \setminus 1 \\ W^i_{:,j} &= \tilde{W^i_{:,j}} / 1^T \tilde{W^i_{:,j}} \end{aligned} \tag{7}$$

where $C^i_j = (\tilde{D}^i - \tilde{Y}^i_{:,j} 1^T)^T (\tilde{D}^i - \tilde{Y}^i_{:,j} 1^T)$ denotes the data covariance matrix. Additionally,when the large size dictionary $D^i \in R^{d^i \times K}$ is used, a fast approximated method could be achieved by first performing a k-nearest neighbor $(k < d^i < K)$ search and then solving a small constrained least square fitting problem, bearing computational complexity of $O(K + k^2)$ [30].

Algorithm 1 Sparse based motion refinement

```
Input: motion dictionary matrix $D^i$; the input imperfect motion sequence $X_{global}$; the length of moving window for grouping $M$; the regulation parameter $\lambda$; the threshold value $\phi$.
Output: the refined motion sequence $\hat{X}_{global}$
1: Trust data detection
     generate the missing marker matric $\Omega$ via the TDD method shown in equation 3
2: Pre-processing
     The unperfect motion sequence $X_{global}$ is translated into the local coordinate representation $X_{local}$;generate partial motion group $\{Y^i, i = 1, 2, \cdots, 5\}$ with $X_{local}$ and given window size $M$; generate the corresponding $\{\Omega^1, \cdots, \Omega^5\}$ via similar operations.
3: Motion refinement
     With the trained motion dictionaries $D^i$, calculate $G^i$ according Eq 6, solve Eq 5 via Eq 7 to get the sparse representation $W^i$ and rebuild the refined $\hat{Y}^i$.
4: Decompose groups and reconstruct the refined pose
     decompose the partial groups $\hat{Y}^i$ and reconstruct the local pose frames $\hat{X}^i_{local}$
5: Reconstruct motion sequence and translate back to world coordinate
     form the refined local motion sequence $\hat{X}_{local}$ and translate it back into world coordinate representation motion sequence $\hat{X}_{global}$
```

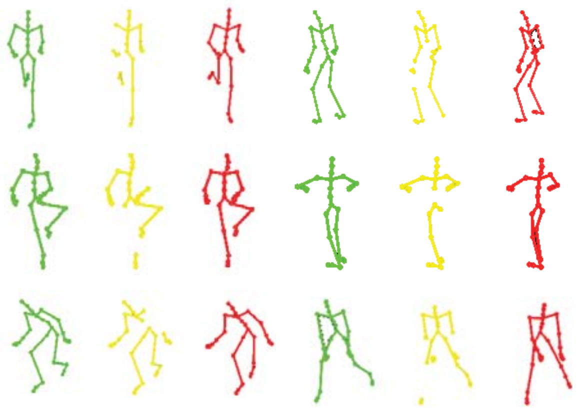
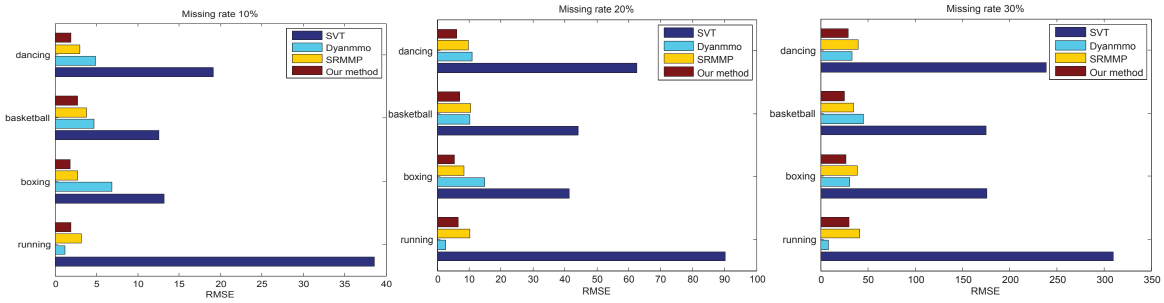


Fig. 4: Results of filling missing value: Original (green), Imperfact (yellow) and Refinement result (red)

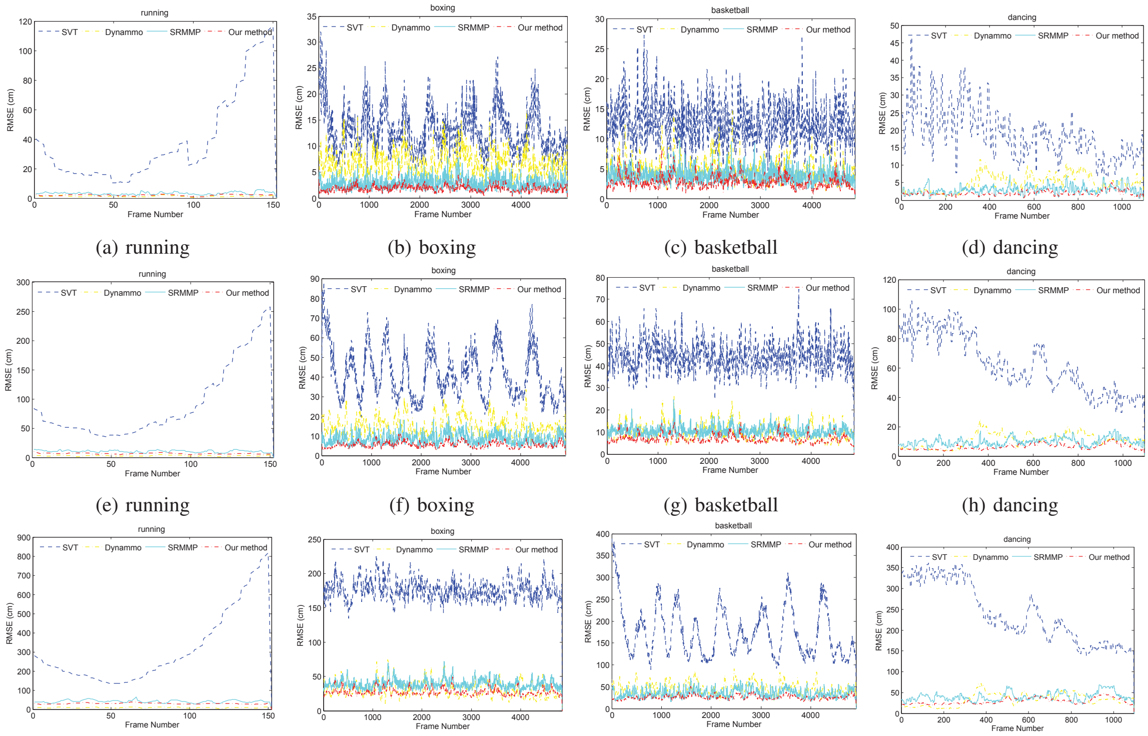


(a) missing rate 10%

(b) missing rate 20%

(c) missing rate 30%

Fig. 5: The Comparisons of our method with other motion refining algorithms on four human motion sequences with missing rate (a) 10%, (b) 20% and (c) 30%. The average RMSE values of each frame(cm/frame) are reported.



(a) running

(b) boxing

(c) basketball

(d) dancing

(e) running

(f) boxing

(g) basketball

(h) dancing

(i) running

(j) boxing

(k) basketball

(l) dancing

Fig. 6: Motion refinement comparisons of different algorithms on four human motion sequences with different missing rates. (a) running (missing rate, 10%), (b) boxing (missing rate,10%), (c) basketball (missing rate 10%), (d) dancing(missing rate 10%), (e) running (missing rate, 20%), (f) boxing (missing rate,20%), (g) basketball (missing rate 20%), (h) dancing (missing rate 20%), (i) running (missing rate, 30%), (j) boxing (missing rate,30%),(k) basketball (missing rate 30%), (l) dancing (missing rate 30%).

III. Results and Discussion

A. Experimental setup

Four representative kinds of actions, i.e., *run*, *dance*, *boxing* and *basketball*, are chosen from CMU human motion database[3] to evaluate the performance of proposed method. Two motion sequences from each category are randomly selected as testing set while others are used for training. Most of CMU motion data are very clear and would be directly used as the training data for our method. For the testing data, we synthesised the noise with missing ratio from 10% to 30% with 10% interval. The size of the moving window M for grouping operation is tuned from $\{2, 4, 8, 16, 30\}$. The parameter λ is tuned from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ and finally set as 10. Finally M is set as 30 and dictionary size K is set as 1024 due to the trade off between the effectiveness and efficiency.

Three methods are implemented as comparison: *Dynammo*

[3] http://mocap.cs.cmu.edu/

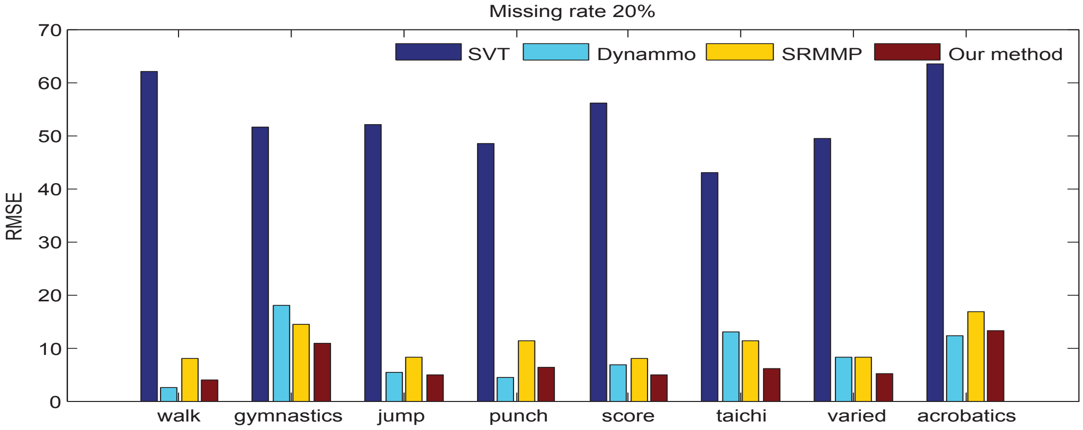


Fig. 7: Experiment on 8 kinds of motion that not contained in the training data: *walk*, *gymnastics*, *jump*, *punch*, *score*, *taichi*, *varied* and *acrobatics*.

[24], [32], a linear dynamic system(LDS) based method; *SRMMP* [9], a sparse coding based method for predicting missing markers; *SVT* [21], a matrix completion based method. In order to make a fair comparison, the parameters for each algorithm are tuned by cross validation.

Additionally, in order to further evaluate the proposed model, we have also taken another experiment on 8 other kinds of actions that not contained in the training data, which are *walk*, *gymnastics*, *jump*, *punch*, *score*, *taichi*, *varied* and *acrobatics*.

B. Experimental Results

Following the work [5], [6], [9], [32], [33], the Root Mean Squared Error (RMSE) measurement is adopted to qualify the refined results:

$$\mathrm{rmse}(X_i, \hat{X}_i) = \sqrt{\frac{1}{n_e}\|X_i - \hat{X}_i\|^2} \tag{8}$$

where X_i is the original pose frame and $\hat{X}_i$ is the recovered one, n_e is the total number of missing markers in X_i. Due to the limited space here, only one motion sequence of each kind of motion is presented, the detail of the refined motion is shown in the demo video.

As shown in Fig 5, our proposed method generally outperforms the competitors in most cases, especially for complex motion. The LTI method *Dynammo* is only effective for the simple motion, e.g. running, while it works not well for the complex motion that contains heterogeneous behaviors. Moreover, the detail result shown in Fig 6 has shown that the recovered result of our proposed method is more stable than the competitors. An example is also shown in the demo video.

C. Discussion

a) Computational complexity analysis.: The computational cost of the proposed method are mainly from two steps operations: Learning dictionaries D^i and the calculation of sparse coefficient W^i. As we know that the dictionary learning just need to be implemented for once time. Hence, the computational cost for refining a motion sequence mainly comes from the sparse coefficient calculation, which is about $O(K + k^2)$ (k nearest neighbor searching, $k < d^i < K$) by using a fast LLC method [30]. In addition, the processing of each partial motion groups is independent, which means that both training and refining could be applied in parallel to increase the time efficiency.

b) Denoising: In this paper, we just focus on solving the missing marker problem. However, the realworld MOCAP data may also contains the noise, e.g. Gaussian noise. To solve this problem, we could follow the strategy in Xiao et al.'s work [6], where the ℓ_2 denoising and missing filling could be combined together. In other words, to refine a piece of imperfect motion data sequence, the missing value will be filled, then the ℓ_2 normalization could be applied to remove the Gaussian noise.

c) Limitations and future work.: The first limitation of our proposed method is that it needs clean motion for training. Hence, both the distribution of missing marker and noise should be considered for dictionary training in the future work to handling the uncleaned training data. Besides, we need an additional optimization step for an ℓ_2 normalization to deal with common gaussian noise. Thus, a new refining objective function will be designed to combine the noise filtering and missing filling. Moreover, our method heavily depends on the missing mask detection. The TDD method we used is based on the assumption that the motion is smooth in the feature space,which may not work well in some extreme cases.

IV. Conclusion

To sum up, human motion refinement is an essential step for MOCAP data based applications. A locality sparse coding based motion refinement method is proposed in this paper. Both hierarchical characteristics and spatial temporal information of motion data are considered while designing the objective function. The LLC coding and grouping operation ensure the smooth property of the recovered result. In addition, the partial model makes our method more robust to the out-of-sample problem. The experimental result shows that our method outperforms the state-of-art method in most cases.

Acknowledgment

The authors would like to thanks the supported from the grant of the Sino-UK project, National Natural Science Foundation of China (Grant No.61572431 and Grant No.51475394), AniNex project (FP7-IRSES-612627), Dr.Inventor project(FP7-ICT-2013.8.1 611383) and Santander PGR Mobility Award.

References

[1] H. P. Shum, E. S. Ho, Y. Jiang, and S. Takagi, "Real-time posture reconstruction for microsoft kinect," *Cybernetics, IEEE Transactions on*, vol. 43, no. 5, pp. 1357–1369, 2013.

[2] L. Zhou, Z. Liu, H. Leung, and H. P. Shum, "Posture reconstruction using kinect with a probabilistic model," in *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*. ACM, 2014, pp. 117–125.

[3] Z. Liu, L. Zhou, H. Leung, and H. P. H. Shum, "Kinect posture reconstruction based on a local mixture of gaussian process models," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2016.

[4] H. Lou and J. Chai, "Example-based human motion denoising," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 16, no. 5, pp. 870–879, 2010.

[5] Y. Feng, M. Ji, J. Xiao, X. Yang, J. J. Zhang, Y. Zhuang, and X. Li, "Mining spatial-temporal patterns and structural sparsity for human motion data denoising," 2014.

[6] J. Xiao, Y. Feng, M. Ji, X. Yang, J. J. Zhang, and Y. Zhuang, "Sparse motion bases selection for human motion denoising," *Signal Processing*, vol. 110, pp. 108–122, 2015.

[7] Z. Wang, Y. Feng, T. Qi, X. Yang, and J. J. Zhang, "Adaptive multi-view feature selection for human motion retrieval," *Signal Processing*, vol. 120, pp. 691–701, 2016.

[8] G. Xia, H. Sun, G. Zhang, and L. Feng, "Human motion recovery jointly utilizing statistical and kinematic information," *Information Sciences*, 2016.

[9] J. Xiao, Y. Feng, and W. Hu, "Predicting missing markers in human motion capture using l1-sparse representation," *Computer Animation and Virtual Worlds*, vol. 22, no. 2-3, pp. 221–228, 2011.

[10] C.-C. Hsieh and P.-L. Kuo, "An impulsive noise reduction agent for rigid body motion data using b-spline wavelets," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1733–1741, 2008.

[11] J. Lee and S. Y. Shin, "General construction of time-domain filters for orientation data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 8, no. 2, pp. 119–128, 2002.

[12] K. Yamane and Y. Nakamura, "Dynamics filter-concept and implementation of online motion generator for human figures," *Robotics and Automation, IEEE Transactions on*, vol. 19, no. 3, pp. 421–432, 2003.

[13] H. J. Shin, J. Lee, S. Y. Shin, and M. Gleicher, "Computer puppetry: An importance-based approach," *ACM Transactions on Graphics (TOG)*, vol. 20, no. 2, pp. 67–94, 2001.

[14] L. Li, J. McCann, N. Pollard, and C. Faloutsos, "Bolero: a principled technique for including bone length constraints in motion capture occlusion filling," in *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. Eurographics Association, 2010, pp. 179–188.

[15] T. Tangkuampien and D. Suter, "Human motion de-noising via greedy kernel principal component analysis filtering," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3. IEEE, 2006, pp. 457–460.

[16] W. Wang and M. A. Carreira-Perpinán, "Manifold blurring mean shift algorithms for manifold denoising," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1759–1766.

[17] Ø. Gløersen and P. Federolf, "Predicting missing marker trajectories in human motion data using marker intercorrelations," *PloS one*, vol. 11, no. 3, p. e0152616, 2016.

[18] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.

[19] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[20] J. Hou, L.-P. Chau, Y. He, J. Chen, and N. Magnenat-Thalmann, "Human motion capture data recovery via trajectory-based sparse representation," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 709–713.

[21] R. Y. Lai, P. C. Yuen, and K. Lee, "Motion capture data completion and denoising by singular value thresholding," *Proceedings of Eurographics, Eurographics Association*, pp. 45–48, 2011.

[22] Y. Feng, J. Xiao, Y. Zhuang, X. Yang, J. J. Zhang, and R. Song, "Exploiting temporal stability and low-rank structure for motion capture data refinement," *Information Sciences*, vol. 277, pp. 777–793, 2014.

[23] X. Liu, Y.-m. Cheung, S.-J. Peng, Z. Cui, B. Zhong, and J.-X. Du, "Automatic motion capture data denoising via filtered subspace clustering and low rank matrix approximation," *Signal Processing*, vol. 105, pp. 350–362, 2014.

[24] C.-H. Tan, J. Hou, and L.-P. Chau, "Motion capture data recovery using skeleton constrained singular value thresholding," *The Visual Computer*, vol. 31, no. 11, pp. 1521–1532, 2015.

[25] S.-J. Peng, G.-F. He, X. Liu, and H.-Z. Wang, "Hierarchical block-based incomplete human mocap data recovery using adaptive nonnegative matrix factorization," *Computers & Graphics*, vol. 49, pp. 10–23, 2015.

[26] M. Burke and J. Lasenby, "Estimating missing marker positions using low dimensional kalman smoothing," *Journal of biomechanics*, vol. 49, no. 9, pp. 1854–1858, 2016.

[27] M. Guay, M.-P. Cani, and R. Ronfard, "The line of action: An intuitive interface for expressive character posing," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 6, p. 205, 2013.

[28] P. Huang, M. Tejera, J. Collomosse, and A. Hilton, "Hybrid skeletal-surface motion graphs for character animation from 4d performance capture," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 2, p. 17, 2015.

[29] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit," *CS Technion*, vol. 40, no. 8, pp. 1–15, 2008.

[30] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[31] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[32] L. Li, J. McCann, N. S. Pollard, and C. Faloutsos, "Dynammo: Mining and summarization of coevolving sequences with missing values," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 507–516.

[33] J. Baumann, B. Krüger, A. Zinke, and A. Weber, "Data-driven completion of motion capture data." in *VRIPHYS*, 2011, pp. 111–118.