Examplar-Based Object Posture Super-Resolution Using Manifold Learning

Chih-Hung Ling[1], Chia-Wen Lin[2], Chiou-Ting Hsu[3], and Hong-Yuan Mark Liao[4]

[1] *Dept. Computer Science, National Chiao Tung University, Hsinchu, Taiwan*

[2]*Dept. Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan*

[3] *Dept. Computer Science, National Tsing Hua University, Hsinchu, Taiwan*

[4]*Inst. Information Science, Academia Sinica, Taipei, Taiwan*

Abstract—**This paper proposes a learning-based approach to increase the temporal resolutions of human motion sequences. Given a set of high resolution motion sequences, our idea is first to learn the motion tendency from this learning dataset and then synthesize new postures for the low-resolution sequence according to the learned motion tendency. We summarize the proposed framework in the following steps: (1) Each motion sequence is first projected into a low-dimension manifold space, where the local distance between postures could be better preserved. We then represent each of the projected motion sequences as a motion trajectory. (2) Next, motion priors learned from the HR training sequences are used to reconstruct the motion trajectory for the input sequence. (3) Finally, we use the reconstructed motion trajectory combined with object inpainting technique to generate the final result. Our experimental results demonstrate the effectiveness of the proposed method, and also show its outperformance over existing approaches.**

I. Introduction

Super-resolution (SR) has attracted much attention for its ability in enhancing the spatial or temporal resolution of low-resolution (LR) images/videos [1]–[6]. While dealing with sequences of human motion, existing SR methods may fail to produce realistic and smooth results if no special efforts are taken to handle the non-rigid human motion.

Since human motion usually contains repeated postures, one may insert interpolated postures into the LR input sequence to increase the temporal resolution. In order to generate postures and animate animal/human motion, Xu *et al*. [7] proposed to animate motions by minimizing a predefined energy function. However, because the energy minimization process did not include human motion model, the performance is unstable and very sensitive to the selected parameters. In [8], Ding *et al*. proposed a rank minimization approach to model and synthesize human motion for video inpainting. They first projected the observed data into a low-dimension manifold and then organized the embedded features to form a Hankel matrix. The missing features in the Hankel matrix are determined by minimizing the rank of Hankel matrix. Finally, they applied the Radial Basis Function (RBF) to inversely transform the embedded features back to the observation domain. This rank minimization approach would usually produce good results as far as the object's motion is periodic. Makihara *et al.* [9] proposed a reconstruction-based method to synthesize periodic human motion with high frame rate from a single periodic motion sequence. The human motion data are first transformed into embedded features in a low-dimension manifold. Then, they iteratively conduct phase registration and motion trajectory reconstruction within an energy minimization process. Under the constraint of periodic motion, their method could also produce good experiment results.

Nevertheless, since human motion is not always periodic, a single motion sequence could provide only limited and insufficient information to generate high quality temporal SR sequences. Therefore, in this paper, we propose using learning-based approach to extract motion tendency from a set of learning sequences and then synthesize interpolated human postures using the learned motion tendency as the prior information. Note that, the extracted motion tendency should preserve only the motion-related information regardless of individual discrepancy in the learning sequences. In [10], Elgammal *et al.* introduced a framework to separate motion data into person and motion factors. However, while this decomposed motion factors can be used to increase the temporal resolution of human motion, we found it difficult to get a stable result. The main reason is because the decomposed person and motion factors are not guaranteed to be orthogonal.

The proposed framework consists of three steps: graphical model construction, manifold-learning-based motion trajectory reconstruction and posture selection. The first step, graphical model construction, projects each input motion sequence into a manifold space and then represent the projected sequence by a motion trajectory. This low-dimensional representation provides a simple and concise representation for human motion. Second, we use the motion priors learned from high-resolution (HR) training sequences to reconstruct the motion trajectory for the input sequence. Finally, we adopt the human object inpainting technique [11] to select interpolated postures based on the reconstructed motion trajectory.

The rest of this paper is organized as follows. Graphical model construction, motion trajectory reconstruction, and posture selection are present in Sections II, III and IV, respectively. Section V shows our experimental results and discussion. Section VI contains the concluding remarks.

II. Graphical Representations of Object Postures

The graphical representation aims to provide a simple and concise representation of a human motion sequence, as shown in Fig. 1. To obtain such a model, we first need to project all the postures onto a manifold space by non-linear dimension reduction, and then link the postures in adjacent frames in the embedded space.

We use the shape context descriptor proposed in [12] to describe human postures and measure the similarity between postures. The silhouette of a posture is represented by a set of sampled points $P = \{p_1, p_2, \dots, p_n\}$. The shape context is then calculated along the silhouette. For each sampled point $p_i \in P$, a corresponding local histogram in the log-polar space is measured to represent the local shape context of p_i. The dissimilarity between two sampled points from two different postures is defined as follows:

$$D(p_i, q_i) = \frac{1}{2}\sum_{k=1}^{N_{\mathrm{bin}}} \frac{\left[h_{p_i}(k) - h_{q_j}(k)\right]^2}{h_{p_i}(k) + h_{q_j}(k)}, \qquad (1)$$

where $h_{p_i}(k)$ and $h_{q_j}(k)$ denote the k-th bin of the two sampled points p_i and q_i, respectively. The value of N_{bin} is empirically set to be 60 for all the sequences. The best match between two different postures is accomplished by minimizing the following total matching cost:

$$H(\pi) = \sum_j D\left(p_j, q_{\pi(j)}\right), \qquad (2)$$

where π is a permutation of 1, 2, …, n. Under the one-to-one matching constraint, the posture matching can be considered as an assignment problem and could be solved using the bipartite graph matching method. Therefore, we could calculate the degree of dissimilarity between two distinct postures P and Q in terms of the shape context distance defined by:

$$D_{SC}(P, Q) = \frac{1}{N_P}\sum_j D\left(p_i, q_{\pi(i)}\right) + \frac{1}{N_Q}\sum_j D\left(p_j, q_{\pi(j)}\right), \quad (3)$$

where N_P and N_Q are the numbers of sample points on the postures P and Q, respectively.

With the dissimilarity measurement as defined in equation (3), we project all the postures in the learning data set using a nonlinear dimension reduction method called isometric feature mapping (Isomap). Note that, the postures are regarded as the input data points in Isomap, and the distance between two data points is equivalent to the dissimilarity between two corresponding postures. The adopted Isomap algorithm is described as follows:

1) Construct a neighborhood graph: Let G denote the neighborhood graph, where an edge is built between two data points x_i and x_j if x_i is one of the K-nearest neighbors (K-NN) of x_j. The edge weight between x_i and x_j is determined in terms of the degree of dissimilarity between their corresponding postures.

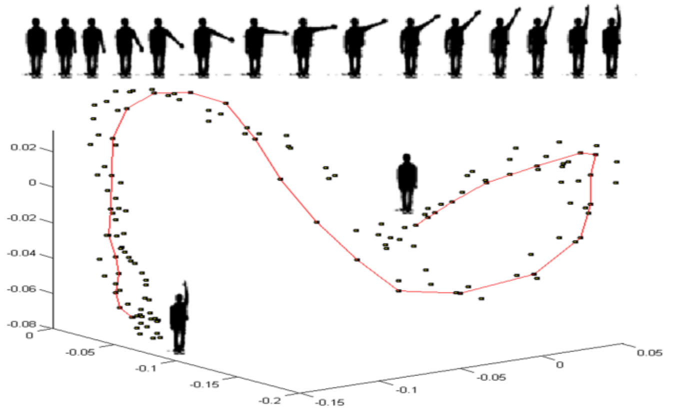


Fig. 1. A graphical model of an object’s motion in a low-dimension manifold. The black points represent the feature points of the postures, and the red lines represent the motion trajectory of input human motion.

2) Compute the shortest paths: Finding the shortest path $d_G(x_i, x_j)$ between each pair of data points x_i and x_j in the graph G.

3) Construct a d-dimensional embedding: Next, we apply the classical Multi-Dimensional Scaling (MDS) to the matrix of graph distances $D_G = \left(d_G(x_i, x_j)\right)$ via eigen-decomposition on the matrix $\Gamma(D_G)$ (The operator Γ is defined as $\Gamma(D_G) = a_{ij} + a_{**} - a_{*j} - a_{i*}$, where $a_{ij} = -\frac{1}{2}d_G^2(x_i, x_j)$, $a_{i*} = \frac{1}{n}\sum_j a_{ij}$, $a_{*j} = \frac{1}{n}\sum_i a_{ij}$, and $a_{**} = \frac{1}{n^2}\sum_i\sum_j a_{ij}$).

The special property of Isomap is its using the geodesic distance between postures to preserve low-dimensional geometry. With this representation, the distribution of motion trajectories in the manifold becomes nearly linear along the time dimension. Hence, we could follow with multi-linear decomposition to decompose motion trajectories into orthogonal factors.

III. Temporal Super-Resolution Using Manifold Learning

After constructing the graphical models, we next wish to transform each human motion sequence into a motion trajectory in the manifold domain. However, since the LR input sequence usually contains poor motion content with low frame rate, its projected motion trajectory in the manifold space would become non-smooth and unreliable. Therefore, we propose to transfer the motion priors learned from the HR training sequences to the input sequence to synthesize the motion trajectory for the input sequence with a high frame rate.

Before learning from the HR training sequences, we will need to arrange motion data in the subspaces of tensor in terms of certain attributes. Since human motion sequences contain no definite labels, we need to take special care to correctly organize motion trajectory data. Below we present our proposed motion data alignment method.

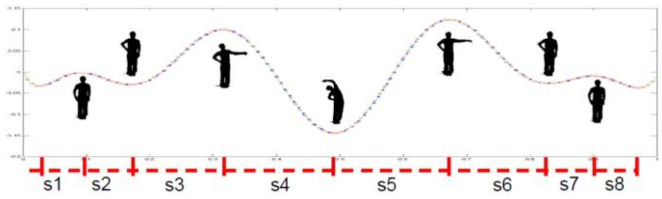
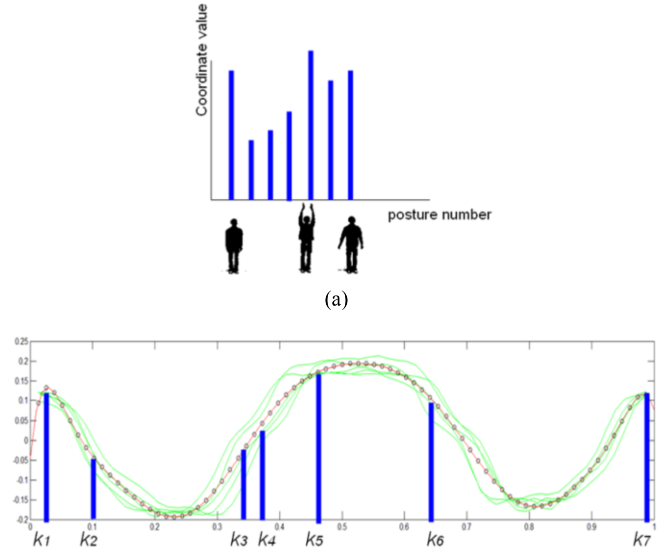


Fig. 2. Illustration of a low-dimensional manifold of a posture sequence and the corresponding postures at the crests and troughs of the manifold.

We first use a continuous motion curve to represent the motion trajectory for each HR training sequence. Each motion trajectory is normalized into the same temporal duration and then mapped into a motion curve by polynomial regression. Next, we find some points with significant motion content along the motion trajectory for data alignment. An example is shown in Fig. 2, where the motion trajectory along the first dimension in the manifold domain has some wave crests and troughs. These wave crests and troughs occur just when the person finishes a previous motion and starts to perform the next motion. The other postures in-between the wave crests and troughs would usually contain slow motion due to the human body constraint. These properties as shown in Fig. 2 are actually invariant to different persons. Therefore, we could sample the points on the wave crests and troughs as the significant points for each motion curve.

In addition, to make sure that the sampled points contain sufficient information to represent the original motion trajectory, we additionally sample *n* points on the motion curve between every two neighboring key points. These additional points are uniformly sampled under the constant motion assumption between two neighboring key points. The number *n* is determined by minimizing the distortion between the original motion trajectory and the reconstructed motion trajectory from the sampled points. The threshold is set as the shape context distance between two continuous postures of human motion with static motion. Finally, a fixed number of *m* sampled points is used to represent the motion trajectory for each training sequence.

For input sequence alignment, since the LR input sequence usually does not contain reliable low-dimensional motion trajectory information, we choose to align the motion data using the raw postures instead of the points along the motion trajectory of test sequence. In order to find *k* postures among the *m* sampled points, we arrange the coordinate value of all postures to form a histogram distribution with *k* bins as shown in Fig. 3. Then, we find *k* out of *m* sampled points along the mean motion trajectory of HR training sequences, where the histogram of *k* sampled points is similar to the histogram of the input sequence. The Bhattacharyya distance is used to calculate the similarity between two histogram distributions.



(a)

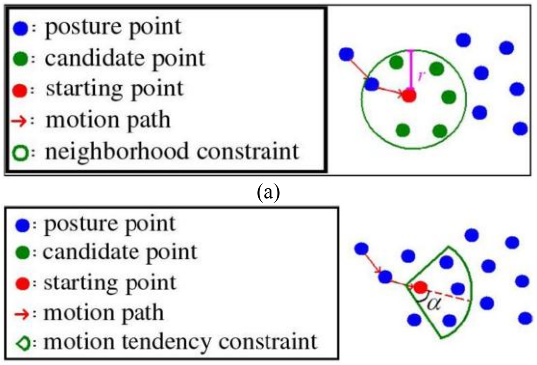
(b)

Fig. 3. (a) The *k* postures coordinate of the LR input sequence. (b) We find *k* reference points among *m* reference points along the mean motion curve of all the HR training sequences. The index of the *k* reference points indicates the suitable position in tensor of the input sequence postures.

Note, the HR training sequences are not necessary to be collected from the same person. If the HR sequences with similar actions performed by different persons are used for training, tensor decomposition [10] can be used to decompose the training data into the person (appearance) factor and the motion factor. By doing so, we will be able to transfer the priors learned from the HR posture sequences of other persons to a person's LR input sequence in the manifold learning stage.

IV. Posture Selection

The reconstructed motion trajectory, obtained via the manifold learning method described in Section III, provides the global motion tendency for the input LR sequence. However, if we include only the global motion tendency but disregard the local motion continuity, the synthesized SR sequence may fail to preserve the original motion characteristics of the input LR sequence. Therefore, we propose to utilize our previous object inpainting method [11], which can effectively preserve local motion continuity, to select postures and then use the selected postures to synthesize the HR sequence. Two constraints are imposed in object inpainting [11] to regulate the motion continuity in the local region of a graphical model. The first constraint limits the



(a)

(b)

Fig. 4. Two motion continuity constraints: (a) the neighborhood constraint, and (b) the motion tendency constraint.

search range to be within a reasonable neighborhood, as shown in Fig. 4(a). The second constraint is introduced to maintain the tendency of an object's motion in each local region as shown in Fig. 4(b). Under the second constraint, the variance of the motion tendency would stay within a reasonable range. Using these two constraints, we then find a number of candidate postures for the upsampled postures and then conduct a two-way (forward–backward) prediction mechanism to further maintain the global motion continuity. Finally, we determine the posture by using the Markov random field to find the one with the highest probability [11].

In our super-resolution application, we determine the values in the above two constraints based on the reconstructed motion trajectory. The number of upsampled postures *p* between every two neighboring postures is first specified by the user. After the value of *p* is determined, we next calculate the possible positions of the upsampled postures in the manifold space. Once we have the coordinate information of all the upsampled and available postures, we could determine the values in the above two constraints for each local region.

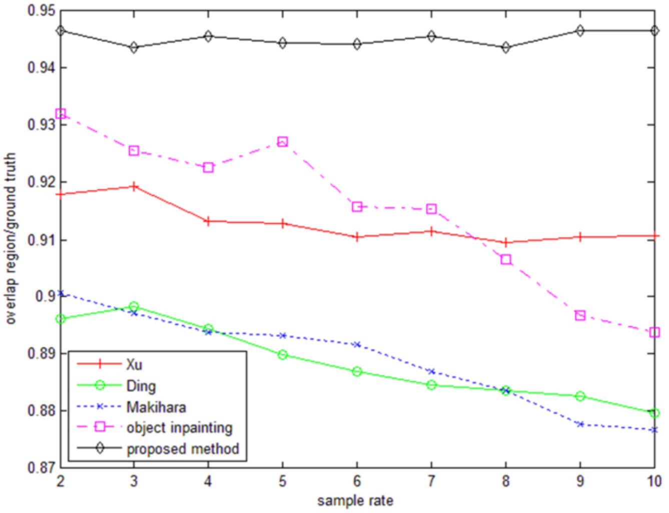


Fig. 5. Comparison of reconstruction accuracy with respect to the ground-truth sequence among five different methods: Xu *et al*'s approach [7], Ding *et al*'s approach [8], Makihara *et al*'s approach [9], object inpainting [11] and the proposed temporal SR approach

V. Experimental Results

To evaluate the effectiveness of the proposed method, we perform experiments on several human object sequences, parts of them were captured with a camcorder and the remaining ones were downloaded from the Weizmann database. In the experiments, we sub-sample each human sequence at different sampling rates to generate the object sequences of low temporal resolutions. Then, we apply the proposed learning-based temporal SR method to synthesize the HR motion sequences. We compare the performance of the proposed method with that of the approaches in [7]–[9], [11]. Due to the space limit, we only show the comparison result for the test sequence #1. Readers can find the complete set of test results

Table I
Comparison of the Ground-truth Postures and the Upsampled Postures Obtained from Different Methods

Ground Truth										Average Accuracy
Xu *et al*'s approach [7]	 90.1%	 91.1%	 89.8%	 90.4%	 89.7%	 94.4%	 85.3%	 88.5%	 87.0%	91.1%
Ding *et al*'s approach [8]	 82.2%	 80.2%	 80.2%	 79.9%	 79.4%	 79.9%	 84.1%	 83.7%	 81.8%	88.0%
Makihara *et al*'s approach [9]	 82.4%	 80.5%	 80.0%	 79.2%	 78.5%	 78.3%	 83.7%	 81.6%	 81.8%	87.7%
Object inpainting [11]	 91.5%	 91.1%	 89.8%	 89.3%	 86.6%	 88.0%	 86.1%	 85.4%	 91.2%	89.4%
Proposed method	 95.5%	 96.1%	 96.1%	 95.8%	 95.3%	 94.4%	 93.1%	 91.6%	 89.5%	94.6%

from our project website [14].

In the first experiment, we down-sampled test sequence #1 with totally 85 frames under different downsampling rates ranging from 2 to 10. Fig. 5 compares the reconstruction accuracies between the ground-truth sequence and the reconstructed sequences obtained using the five different approaches for various down-sampling rates. The result shows that the proposed temporal SR method does not only consistently outperform the other methods, but also achieves stably high accuracy of better than 94% under all the nine downsampling rates. Because the proposed motion synthesis method is more effective in extending frame rate of an object's motion and maintaining motion continuity simultaneously, we conclude that it is more suitable for increasing temporal resolution than the compared methods. On the other hand, the performances of all the other four schemes typically degrade as the downsampling rate increases, since the available information for reconstructing HR sequence becomes fewer and less accurate when the temporal resolution of the input LR sequence decreases. Our previous object inpainting method [11] performs the second best at downsampling rates lower than 8. The motion animation scheme [7] composes a sequence of smooth posture motion from a set of available postures by executing an energy minimization process. Since the performance of motion animation scheme depends mainly on the two postures at both ends and the available posture database, this scheme can also achieve stable performance under different downsampling rates. However, since it does not take into account the low-dimensional manifold prior of human motion, its performance is significantly lower than the proposed method.

Table I illustrates the results of the ground-truth and the five compared methods. The top row shows the sequence of nine ground-truth postures, and the second to sixth rows show the missing frames reconstructed by the methods in [7]–[9], [11] and the proposed method, respectively. The reconstruction accuracy of each posture is also indicated under the posture. From these selected postures, it is obvious that the postures reconstructed by our method are consistently better than those derived by the compared methods both subjectively and objectively.

Note, in the experiments, we compare the performance between the proposed method and our previous object inpainting method [11]. The difference between these two approaches is that the proposed method reconstructs postures based on the rich information in the low-dimensional manifold motion tendency learned from HR training sequences, whereas the object inpainting method [11] utilizes self-contained information in the available LR postures to reconstruct postures without the support of HR training sequences. Since the object inpainting method does not need any HR training sequence, it can be regarded as the baseline mode of the proposed method that can achieve reasonable reconstruction accuracy without the need of HR training sequences. When HR training sequences are available, as an advanced tool, the proposed manifold-learning-based SR scheme can significantly improve the accuracy and stability of reconstructed HR postures.

VI. Conclusion

We proposed a human motion temporal super-resolution method which consists of three steps: (1) graphical models construction; (2) motion trajectory reconstruction; and (3) posture selection. In addition, we also proposed a motion data alignment method to correctly arrange motion data. We transfer the motion priors learned from HR training sequences to reconstruct the motion trajectory for the LR input sequence. Finally, we adopt an object inpainting method on the reconstructed motion trajectory to select interpolated postures. Both global motion tendency and local motion continuity are well preserved in the resultant HR sequence. The experiment results also demonstrate that the proposed approach outperforms existing state-of-the-art approaches.

References

[1] B. Baker and T. Kanade, "Limits on superresolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, Sept. 2002.

[2] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graphics & App.*, vol. 22, no. 2, pp. 56–65, Mar. 2002.

[3] E. Shechtman, Y. Caspi and M. Irani, "Space-time super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 531–544, Apr. 2005.

[4] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 1-8, 2008.

[5] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Trans. Image Process.*, vol.18, no.1, pp.36-51, Jane 2009.

[6] Y. HaCohen, R. Fattal, and D. Lischinski, "Image upsampling via texture hallucination," *in Proc. IEEE Int. Conf. Comput. Photography,* Cambridge MA USA, pp. 20-30, Mar. 2010.

[7] X. Xu, L. Wan, X. Liu, T.-T. Wong, L. Wang and C.-S. Leung, "Animating animal motion from still," *ACM Trans. Graphics,* vol. 27, no. 5, Dec. 2008.

[8] T. Ding, M. Sznaier, and O. I. Camps, "A rank minimization approach to video inpainting," in *Proc. IEEE Conf. Comput. Vis.*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[9] Y. Makihara, A. Mori and Y. Yagi, "Temporal super resolution from a single quasi-periodic image sequence based on phase registration," in *Proc. Asian Conf. Comput. Vis.*, Queenstown, New Zealand, Nov. 2010, pp. 107–120.

[10] Chan-Su Lee and Ahmed Elgammal, "Modeling view and posture manifolds for tracking," in *Proc. Intl. Conf. Comput. Vis.*, Oct. 2007, Rio de Janeiro, Brazil.

[11] C.-H. Ling, Y.-M. Liang, C.-W. Lin, Y.-S. Chen, and H.-Y. M. Liao, "Human object inpainting using manifold learning-based posture sequence estimation," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3124–3135, Nov. 2011.

[12] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[13] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.

[14] "NTHU Object Posture Upsampling Project," NTHU, Hsinchu, Taiwan. [Online]. Available: http://www.ee.nthu.edu.tw/cwlin/posture_SR/posture_SR.htm.