ADAPTIVE HUMAN SILHOUETTE RECONSTRUCTION BASED ON THE EXPLORATION OF TEMPORAL INFORMATION

Xue Zhou[⋆] *Xi Li*[†] *Tat-Jun Chin*[†] *David Suter*[†]

[⋆] University of Electronic Science and Technology of China
zhouxue@uestc.edu.cn
[†] School of Computer Science, University of Adelaide, Australia
xi.li03@adelaide.edu.au, {tjchin, dsuter}@cs.adelaide.edu.au

ABSTRACT

Human silhouette reconstruction has a wide range of applications in motion analysis, object segmentation and tracking, etc. In this paper, we propose a human silhouette reconstruction method based on the exploration of temporal information. Given a test silhouette, the proposed method aims to find its reliable templates for reconstruction by using the intrinsic temporal relationship among different frames. To effectively obtain such templates, we propose an adaptive criterion based on the non-negative least square optimization. Experimental results on two challenging datasets demonstrate the effectiveness of our method.

Index Terms— Human silhouette reconstruction, temporal constraint, shortest path searching, level set

1. INTRODUCTION

Typically, three types of models are used to represent the silhouettes, including snakes [1], skeletons [2] and level sets [3]. Due to being numerically stable and capable of handling non-rigid topological shape changes, level sets are widely used to describe human silhouettes. Based on the level set representation, a variety of methods have been proposed to reconstruct a test silhouette given the training data in the presence of noise, occlusion or clutter. However, robust human silhouette reconstruction is actually a challenging task due to limited training data, unpredictable noise disturbance, occlusion and human silhouettes deformation, etc.

Recently, much work has been done in human silhouette analysis [4–7]. Cremers [4] proposes a nonlinear dynamic model for level set segmentation. Based on autoregression, the proposed model is able to approximate a particular silhouette at a given time using the silhouettes observed at previous time instances. To explore the intrinsic nonlinear structures from the training samples, some manifold learning (ML)-based methods [5–7] were proposed. Prisacariu and Reid [5] learn a low-dimensional latent space of training samples using Gaussian Process Latent Variable Models (GPLVM). Elgammal and Lee [6, 7] use locally linear embedding (LLE) method to construct a human silhouette manifold. For each test sample, they find its nearest neighbor in the manifold and project it back to the original sample space for reconstruction. In order to ensure the smoothness of the learned manifold, ML-based methods usually require a large amount of regular training data, which is usually impractical in real applications. Besides, they often ignore the temporal constraint information in finding the nearest neighbor, resulting in inaccurate reconstruction. In this paper, we will show that the temporal constraint information plays an important role in reconstructing a human silhouette.

We propose a temporal constraint-based human silhouette reconstruction (TC-HSR) method which focuses on human silhouette reconstruction in a walking cycle. The proposed method is based on a "shortest paths" searching scheme for finding reliable templates to restore a test sample. Our contributions lie in three aspects. 1) We introduce temporal information into the process of finding the reliable templates for reconstruction. The reliable templates can be robustly found even when the test sample is seriously corrupted. 2) To effectively obtain such templates, we develop a method that searches for the "shortest paths" linking the candidate templates of the consecutive test frames. Our method can also find the reliable templates in different walking cycles. 3) We propose a non-negative least square (NNLS)-based criterion to effectively detect the abnormality of the test sample. Based on the abnormality detection (AD) result, a discriminative criterion is defined to adaptively find the reliable templates for each test sample.

2. TEMPORAL CONSTRAINT-BASED HUMAN SILHOUETTE RECONSTRUCTION

2.1. Problem definition

We first have a training silhouette sequence $\{X_i\}_{i=1}^N$ and their corresponding frame number sets $\{I_{X_i}\}_{i=1}^N$. Given a test sample Y_t at time t which may be seriously corrupted, our objective is

This work is supported by NSFC (No.60905015). Xue Zhou performed this work while visiting the University of Adelaide.

to reconstruct it using its $K < N$ reliable templates X_j^*, where $X_j^* \in \{X_i\}_{i=1}^N,\ j = 1, \ldots, K$. In what follows, each silhouette sample is represented by the level set signed distance function (SDF) [3] which is flattened into a column vector and denoted by Φ. Typically, the test sample can be approximated by a set of its reliable templates:

$$\Phi_{Y_t} \approx \sum_{j=1}^{K} w_j \Phi_{X_j^*} \tag{1}$$

where w_j is the reconstruction weight and $\Phi_{X_j^*}$ is the feature vector of X_j^*. The focus of this paper is on how to find the reliable templates from the training samples. Usually, when the test sample is noisy or seriously corrupted, the templates found simply by K nearest distances are not reliable because the distances do not reflect the true affinity relationship between them. However the sequences often include some useful temporal smoothness constraints among consecutive frames. TC-HSR effectively finds the reliable templates for reconstruction by considering temporal information.

2.2. Adaptive reliable template construction

If the current test sample Y_t is mildly corrupted, templates found by smallest distances can be directly used to restore Y_t. Otherwise, we use TC-HSR to find the reliable templates of an abnormal (very noisy or seriously corrupted) test sample. Thus, we propose a non-negative least square (NNLS)-based [8] criterion for abnormality detection (AD). In principle, NNLS uses a linear combination of all the training samples X to approximately represent the test sample Y_t:

$$\min_{\mathbf{c}} \|\Phi_{Y_t} - B_X \cdot \mathbf{c}\|_2^2 \quad \text{s.t. } \mathbf{c} \geq 0 \tag{2}$$

where $\|\cdot\|_2$ is the L2-norm, $\mathbf{c}$ is the non-negative coefficient vector and $B_X = (\Phi_{X_1}, \Phi_{X_2}, \cdots \Phi_{X_N})$ is the corresponding feature collection of all the training samples $\{X_i\}_{i=1}^N$. The larger the residual $r = \|\Phi_{Y_t} - B_X \cdot \mathbf{c}\|_2$, the more likely the test sample could be abnormal. Motivated by this observation, we use a threshold T to discriminate these two cases.

When $r < T$, i.e. the test sample Y_t is determined to be normal, we find its templates based on the following L2-norm distance directly:

$$d(X_i, Y_t) = \|\Phi_{X_i} - \Phi_{Y_t}\|_2 \tag{3}$$

In other words, we choose the K nearest neighbors as the top K reliable templates $\{X_j^*\}_{j=1}^K$ for reconstructing Y_t.

When $r \geq T$, we concatenate the test sample Y_t with its previous L adjacent frames to form a small test sequence $\mathbb{S}$, and then introduce temporal smoothing to address this case. Without loss of generality, let us take $L = 2$ for example. As a result, we have a test sequence denoted as (Y_{t-2}, Y_{t-1}, Y_t). For convenience, we let $\mathbb{TS}_{(Y_{t-2})}$, $\mathbb{TS}_{(Y_{t-1})}$ and $\mathbb{TS}_{(Y_t)}$ denote the candidate template-sets, each of which is indexed by a collection of frame numbers.

Two steps are taken to obtain these candidate template-sets (i.e. $\mathbb{TS}_{(Y_{t-2})}$, $\mathbb{TS}_{(Y_{t-1})}$, $\mathbb{TS}_{(Y_t)}$) for each frame in $\mathbb{S}$. The first is to collect the candidate templates by 1) selecting the training samples with the top few non-negative reconstruction coefficients in Eq. (2), as well as 2) finding the candidate templates with the smallest distances according to Eq. (3). In addition to computing the candidate templates for each frame in $\mathbb{S}$, the abnormality detection is also performed based on its NNLS residual. The obtained candidate templates of the normal test sample are usually more reliable than the abnormal one, and the consecutive frames often have the similar candidate templates. Thus the second step is to share the candidate template-sets of the normal samples with the candidate template-sets of the abnormal ones. If none of the samples in $\mathbb{S}$ is normal, then candidate template sharing does not occur, and as a result, the candidate template-sets obtained in the first step are used as final.

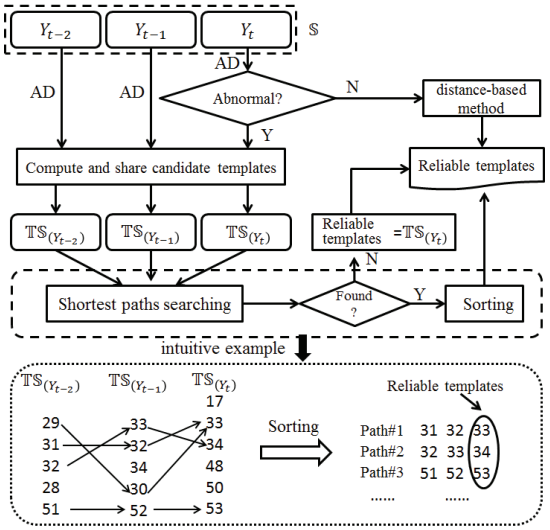


Fig. 1. A flow diagram for finding the reliable templates. The bottom part intuitively shows the "shortest paths" searching.

After obtaining the three final candidate template-sets $\mathbb{TS}_{(Y_{t-2})}$, $\mathbb{TS}_{(Y_{t-1})}$ and $\mathbb{TS}_{(Y_t)}$, we aim to find the shortest frame number paths linking Y_{t-2}, Y_{t-1} and Y_t:

$$\begin{aligned} c^* \rightarrow b^* \rightarrow a^* &= \underset{\{(c,b,a)|c<b<a\}}{\arg\min} \{(b-c) + (a-b)\} \\ &= \underset{\{(c,b,a)|c<b<a\}}{\arg\min} \{a - c\} \end{aligned} \tag{4}$$

where c, b, a are frame numbers belonging to the three candidate template-sets, such that $c \in \{\mathbb{TS}_{(Y_{t-2})}\} \subseteq \{I_{X_i}\}_{i=1}^N$, $b \in \{\mathbb{TS}_{(Y_{t-1})}\} \subseteq \{I_{X_i}\}_{i=1}^N$, $a \in \{\mathbb{TS}_{(Y_t)}\} \subseteq \{I_{X_i}\}_{i=1}^N$. We sort the frame number paths based on Eq. (4) in an ascending order and choose the top K frames associated with Y_t as the final reliable templates (denoted by $\{X_j^*\}_{j=1}^K$). If no such paths could be found, we directly choose the reliable templates from $\mathbb{TS}_{(Y_t)}$. By analogy the other cases of L can be easily generalized. Fig. 1 shows the illustration of finding the reliable templates.

2.3. Silhouette reconstruction using the reliable templates

After obtaining the reliable templates $\{X_j^*\}_{j=1}^K$, we reconstruct the current test sample Y_t by a weighted combination of

$\{X_j^*\}_{j=1}^K$. The reconstruction is formulated as Eq. (1) wherein the weight is calculated as:

$$w_j = \frac{\exp[-d^2(X_j^*, Y_t)]}{\sum_{j=1}^K \exp[-d^2(X_j^*, Y_t)]} \quad (5)$$

and $d(X_j^*, Y_t)$ is computed by Eq. (3). The obtained reconstruction result (denoted by $\Phi_{\bar{Y}_t}$) is re-initialized to stay smooth enough to approximate the signed distance function (SDF) and then added to the test sequence for the reliable templates construction of the subsequent frames instead of Φ_{Y_t}.

3. EXPERIMENTS

To verify our method, we have conducted several experiments on two public gait datasets: OU-ISIR dataset[1] and CASIA dataset[2]. The OU-ISIR dataset comprises gait silhouette sequences of persons walking on a treadmill with varying speed (from 2km/h to 7km/h with 1km/h interval). Each silhouette image is normalized and registered to 138× 88 pixels. The CASIA dataset consists of color image sequences of persons walking outdoor more freely. For each frame, the region of interest surrounded by contour is extracted and normalized to 158×93 silhouette image. Because of the limited number of training samples, CASIA dataset is more challenging.

We evaluate the silhouette reconstruction performance of three methods, including the manifold learning based silhouette reconstruction (MLSR) [7], the single frame based silhouette reconstruction (SFSR) and our TC-HSR method. MLSR [7] finds the nearest neighbor of the test sample in the manifold space and re-maps it to accomplish the reconstruction task. SFSR is based on the current single-frame neighbors search without considering the temporal information. To quantitatively evaluate the performance of these three methods, we introduce a reconstruction score based on the PASCAL VOC overlap ratio (between the ground truth and the reconstruction results).

We show several silhouette reconstruction examples in Fig. 2. Parameter settings of our method are as follows: the threshold T for abnormality detection is 260, the parameter K is chosen as 3. We consider previous two adjacent frames in our experiments (i.e. $L = 2$). For each dataset, three cases are evaluated: 1) the test sample is normal as shown in the top row of Fig. 2 (a) and Fig. 2 (b); 2) the test sample is corrupted moderately (the second row of Fig. 2 (a) and Fig. 2 (b)); and 3) the test sample undergoes the serious deformation (the bottom two rows of Fig. 2 (a) and Fig. 2 (b)). From Fig. 2, we can see that all the three methods achieve good reconstruction performance in cases 1) and 2). However, our method outperforms the other two methods in case 3) with serious corruption and occlusion.

In Fig. 3, we show the reconstruction performance of three methods for CASIA dataset (Fig. 3(a)) and OU-ISIR dataset (Fig. 3(b) and Fig. 3(c)). Two cases are considered. One is the same person with the same speed (SPSS) and the other one is same person with different speed (SPDS). For SPSS case, the first half part of the sequence is used for training and the rest is for testing. For SPDS case, the 4km/h speed sequence is used as training data and the 7km/h speed sequence is for testing. For all the testing sequences, 50% frames are selected randomly and corrupted seriously. From Fig. 3, it's clear that our method performs better than the other two methods on both of the two datasets. We also report the average reconstruction score (ARS) and the standard deviation (SD) among the three methods in the case of serious corruption in Table 1. Note that our method achieves the highest ARS and the lowest SD.

Furthermore, we evaluate the effect of the different corruption rates as shown in Fig. 4. Corruption rate is the proportion of noisy images to the whole testing images. When the test sequence is normal or mildly corrupted, MLSR has a good performance. However, as the corruption rate increases, the average reconstruction score of our method decreases more slowly, which indicates the better stability of our

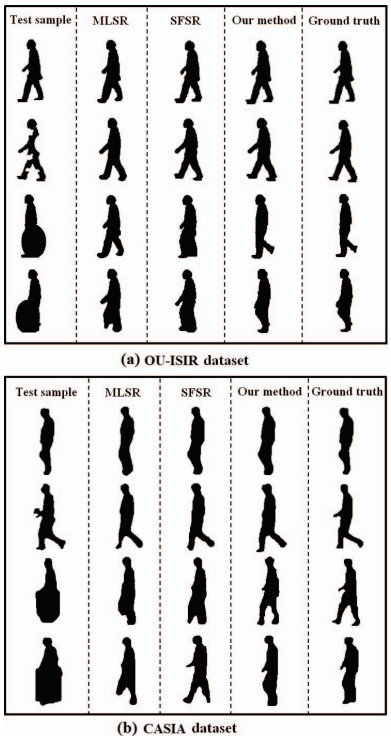
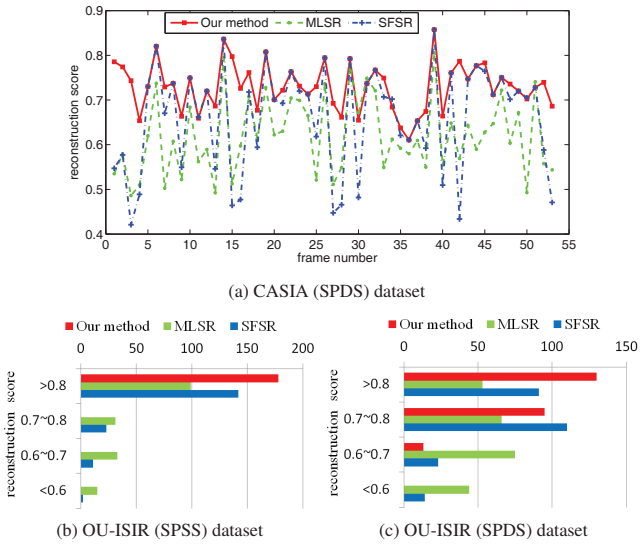


Fig. 2. Silhouette reconstruction results by three different methods: (a) OU-ISIR dataset; (b) CASIA dataset. Each row corresponds to a test case.

Table 1. Quantitative comparison among the three methods in the case of serious corruption

Method	OU (SPSS)		OU(SPDS)		CASIA(SPDS)	
	ARS	SD	ARS	SD	ARS	SD
Ours	**0.90**	**0.02**	**0.80**	**0.05**	**0.74**	**0.05**
SFSR	0.85	0.08	0.77	0.08	0.65	0.12
MLSR	0.80	0.09	0.72	0.07	0.62	0.08

[1]http://www.am.sanken.osaka-u.ac.jp/GaitDB/index.html
[2]http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp



(a) CASIA (SPDS) dataset

(b) OU-ISIR (SPSS) dataset

(c) OU-ISIR (SPDS) dataset

Fig. 3. Illustration of reconstruction performance comparison among three methods. (a) shows the frame-by-frame reconstruction scores for CASIA dataset; (b) and (c) display the distribution of the reconstruction scores for OU-ISIR dataset. The vertical axes of (b) and (c) are the reconstruction score intervals and the horizontal axes correspond to the number of frames fallen into each interval.

method.

In addition, MLSR is susceptible to serious corruption. Fig. 5 gives an intuitive illustration. If the test sample is seriously corrupted, the nearest neighbor obtained by MLSR is extremely unreliable or even wrong, as shown in Fig. 5 (a) and Fig. 5 (b).

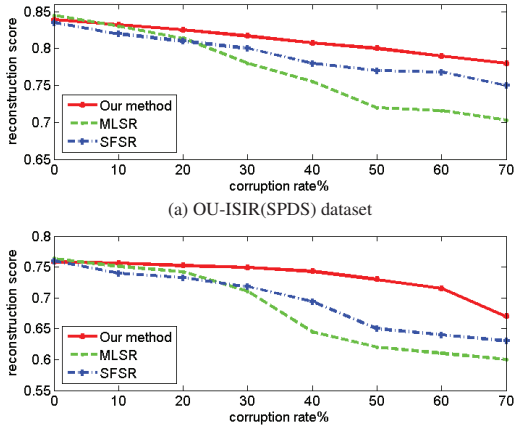
4. CONCLUSION

In this paper, we have presented a temporal constraint-based human silhouette reconstruction method. This method aims to find the reliable templates by considering the temporal constraint among consecutive frames and use them to robustly reconstruct a test silhouette. In order to obtain reliable templates, a NNLS-based criterion is proposed to adaptively determine the status of the current test sample (normal or abnormal). Based on the abnormality detection, the "shortest paths" searching scheme is proposed to obtain the reliable templates. We compare our method with two competing methods on two challenging datasets. Both qualitative and quantitative experimental results demonstrate the effectiveness and robustness of our method.

5. REFERENCES

[1] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: active contour models" in *Int. J. Comput. Vis.*, vol.1, pp. 321-331, 1988.

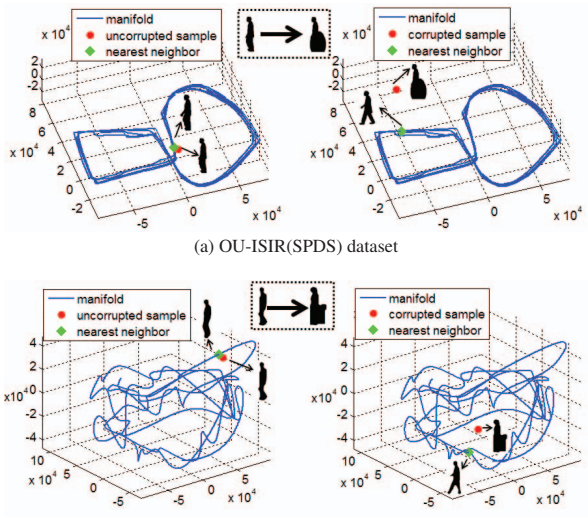
[2] H. Sundar, D. Silver, N. Gagvani and S. Dickinson, "Skeleton based shape matching and retrieval" in *Proc. of the Shape modeling international*, pp. 130-139, 2003.



(a) OU-ISIR(SPDS) dataset

(b) CASIA(SPDS) dataset

Fig. 4. Illustration of the reconstruction performance with the different corruption rates.



(a) OU-ISIR(SPDS) dataset

(b) CASIA(SPDS) dataset

Fig. 5. Illustration of the sensitivity of MLSR to serious corruption.

[3] S. Osher and J. Sethian, "Fronts propagation with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulation" in *Journal of Comput. Phys.*, vol.79, pp. 12-49, 1988.

[4] D. Cremers, "Nonlinear dynamical shape priors for level set segmentation" in *Proc. of CVPR*, pp. 1-7, 2007.

[5] V.A. Prisacariu and I. Reid, "Nonlinear shape manifolds as shape priors in level set segmentation and tracking" in *Proc. of CVPR*, pp. 2185-2192, 2011.

[6] A. Elgammal and C.S. Lee, "The role of manifold learning in human motion analysis" in *Human motion understanding, modeling, capture and animation*, vol. 36(1), pp. 1-29, 2008.

[7] A. Elgammal and C.S. Lee, "Nonlinear manifold learning for dynamic shape and dynamic appearance" in *Comput. Vis. and Imag. Underst.*, vol. 106, pp. 31-46, 2007.

[8] C.L. Lawson and R.J. Hanson, "Solving least-squares problems" in *Prentice-Hall.*, 1974.