Monocular 3D human pose estimation with a semi-supervised graph-based method

M. Abbasi [1,2], H. R. Rabiee[2], C. Gagné[1]

[1]Computer Vision and Systems Laboratory
Université Laval, Quebec City, QC, Canada
[2]Digital Media Lab, Department of Computer Engineering
Sharif University of Technology, Tehran, Iran

1 Introduction

Due to the existence of depth ambiguity in silhouettes, it is more suitable to use geodesic distance than Euclidean distance for monocular 3D human pose estimation . To achieve this, a manifold which contains geodesic distance can be approximated by a graph using the k-NN method. However the depth ambiguity causes the occurrence of shortcut edges within the graph.

As input data is a sequence of images, temporal information is used to identify and remove these shortcut edges by measuring the similarity of each pair of connected vertices through the use of sliding temporal windows. Furthermore, by exploiting the relationships between labeled and unlabeled data, the proposed method can estimate the 3D body poses with a small set of labeled data.

2 Goal

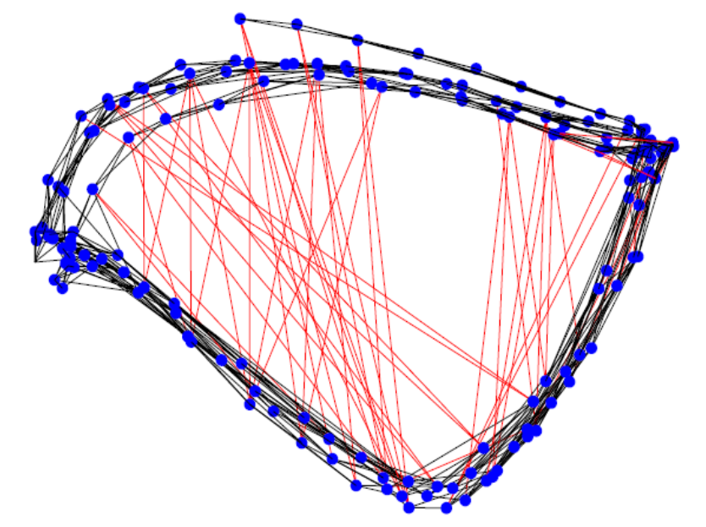


Figure 1: Constructed graph ($G_f^{(i,j)}$)

Remove shortcut edges

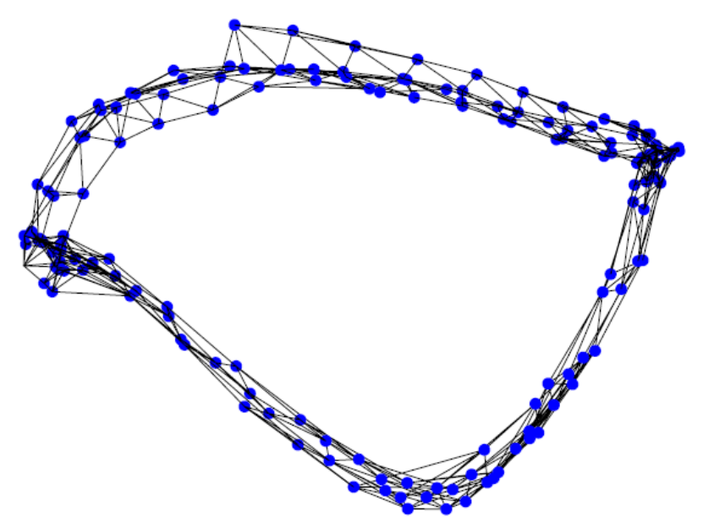


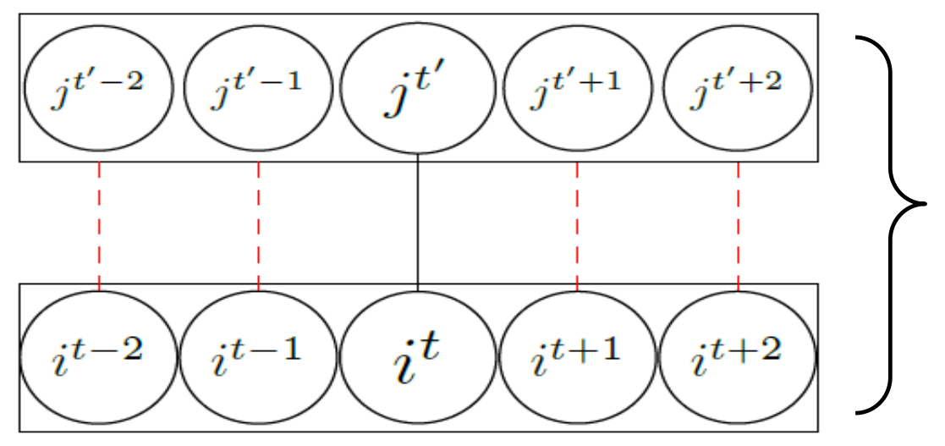
Figure 2: True Manifold

3 Methodology

3.1. Assumption

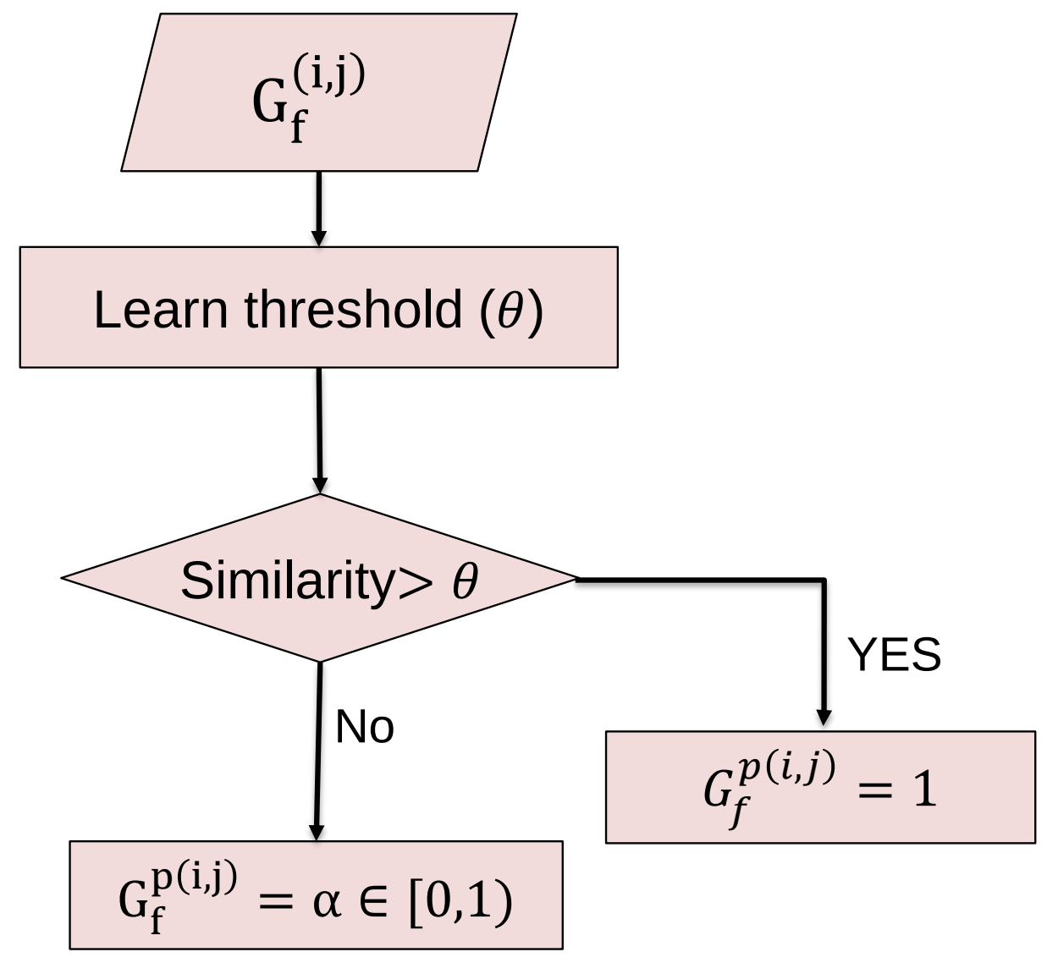
If two silhouettes are truly similar enough to share an edge, then their temporal neighbors obtained from two temporal windows should be similar too.

3.2. Similarity measurement of temporal windows



Similarity = $\sum_{k=-2}^{2} \mathrm{G_f}^{(\mathrm{j}^{t'+k}, i^{t+k})}$

Figure 3: Two temporal windows


$\mathrm{G_f^{(i,j)}}$
Learn threshold (θ)
Similarity> θ
YES
No
$G_f^{p(i,j)} = 1$
$\mathrm{G_f^{p(i,j)}} = \alpha \in [0,1)$

α set by grid search and 5-fold cross validation

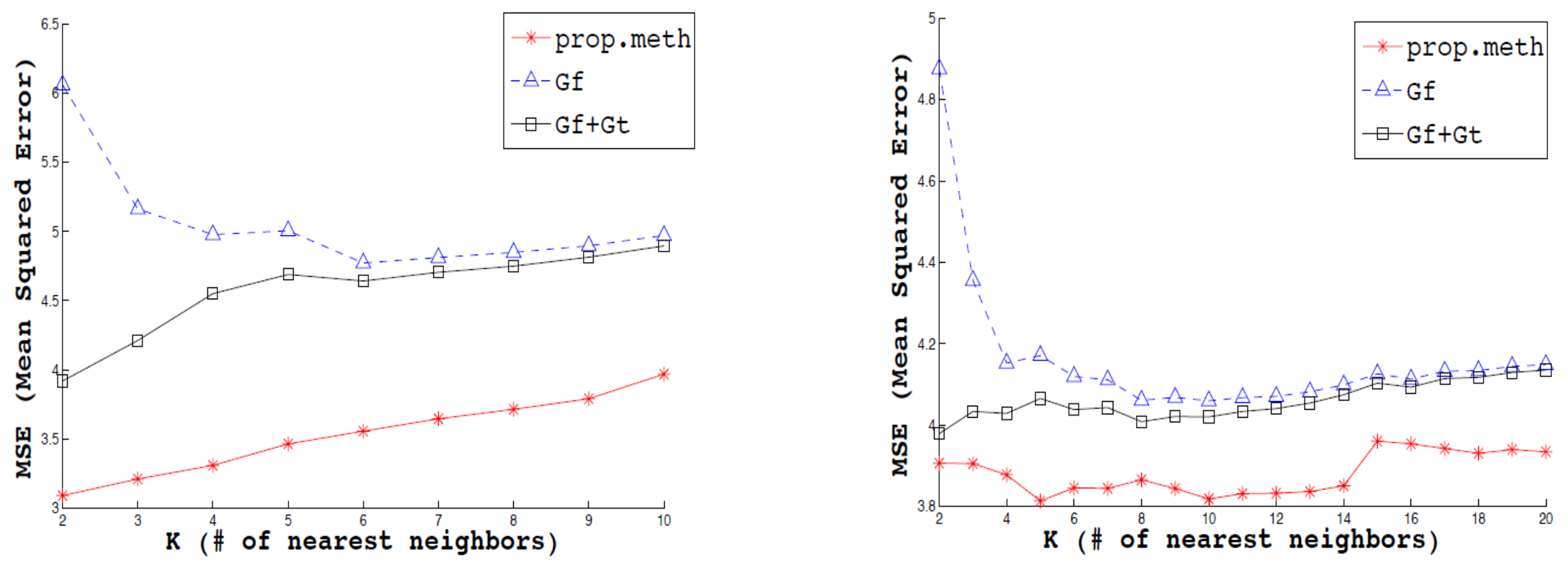
Algorithm 1 Learning of parameter θ.

```
Require: $g_p(Y^l, E_l)$, $g_f(X^l, E_l')$
  $\mathcal{S} = 0$                ▷ Sum of similarity from shortcut edges
  $C = 0$                          ▷ Number of shortcut edges in $g_f$
  for all $g_f^{i,j} = 1$ do
      if vertex $j$ not reachable from vertex $i$ by at most by
  4 hops in graph $g_p$ then
          $\mathcal{S} \leftarrow \mathcal{S} + \sum_{k=-2}^{2} g_f^{(i^{t+k}, j^{t'+k})}$
          $C \leftarrow C + 1$
      end if
  end for
  $\theta = \mathcal{S}/C$
```

4 Quantitative results

4.1. Comparison of Graphs

The comparison of MSE curves for G_f (base graph) and $G_f + G_t$ (base graph with temporal edges) shows that as the value of K increases, the impact of temporal edges becomes less on their MSE performance. Moreover, the comparison of MSE curves for $G_f + G_t$ and $G_f^p + G_t$ (the proposed method) reveals the significant influence of removing shortcut edges.



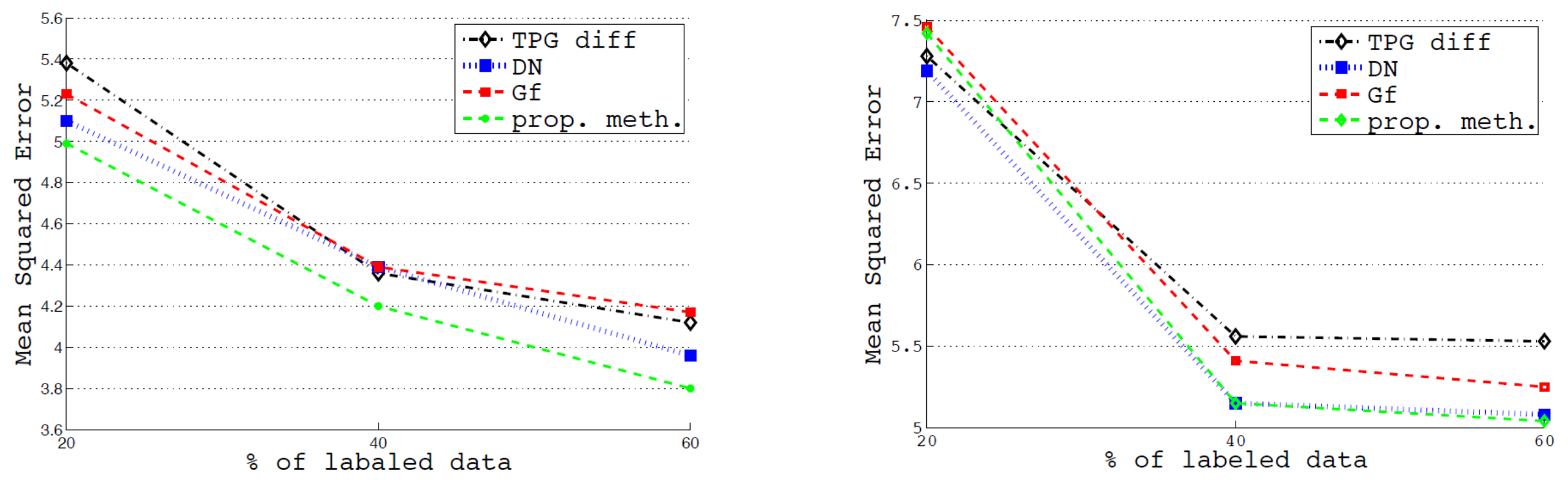
(a) Swimming

(b) Walking

Figure 4: The MSE curves as a function of K on swimming (a) and walking (b) datasets, where k indicates the number of nearest neighbors in K-NN.

4.2. Comparison with recent graph construction method

The proposed method was compared with TPG (Tensor Product graph) diffusion and Dominate Neighbor (DN) method, two recent state of the art methods.



(c) Navigate (walking)

(d) Circular walking

Figure 5: The MSE curves as a function of percentage of labeled data on navigate (c) and circular walking (d)

4.3. Comparison with recent semi-supervised method

Activity (# all data)	% labeled data	TGP	GC+RT	G_f	Proposed method
Circular walking (1961)	60 %	7.28	5.30	5.25	**5.04**
	40 %	8.54	5.37	5.41	**5.15**
	20 %	21.42	7.63	7.46	**7.42**
Boxing (1400)	60 %	12.01	10.51	10.00	**9.34**
	40 %	17.91	12.04	11.69	**10.87**
	20 %	18.95	12.18	11.75	**10.96**
Swimming (1202)	60 %	5.03	4.91	4.77	**3.55**
	40 %	5.75	5.38	5.30	**4.54**
	20 %	7.10	6.67	6.65	**6.57**
Walking (1000)	60 %	4.80	4.36	4.17	**3.80**
	40 %	5.26	4.57	4.39	**4.20**
	20 %	8.59	5.65	5.23	**4.99**

5 Qualitative results

- Detected shortcut edges by sliding temporal windows
  - Nodes: pose data points
  - Edges: connection of feature data points
  - Blue nodes: corresponding silhouettes connected by the detected shortcut (red) edges
- For visualization purposes, the dimensionality of pose data points was reduced to 3 by kernel PCA

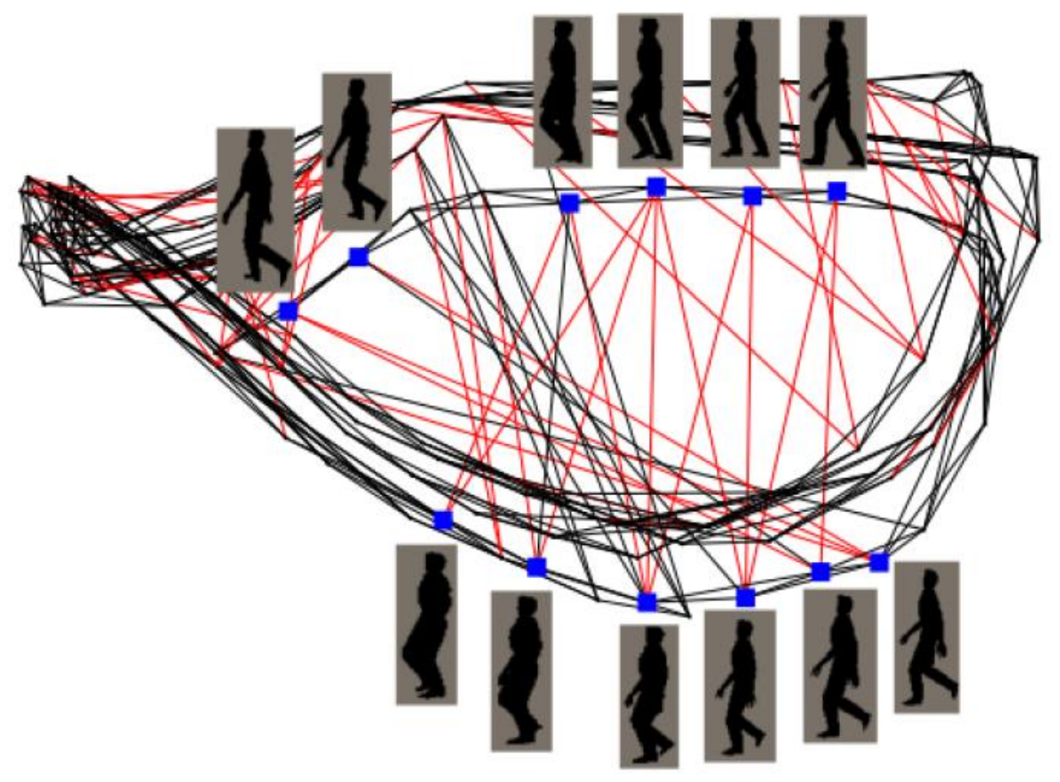
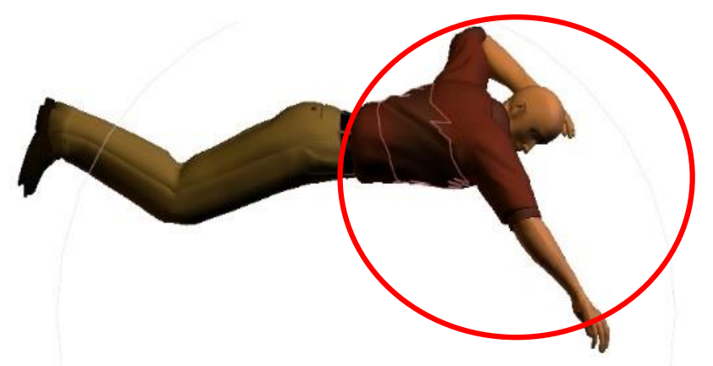
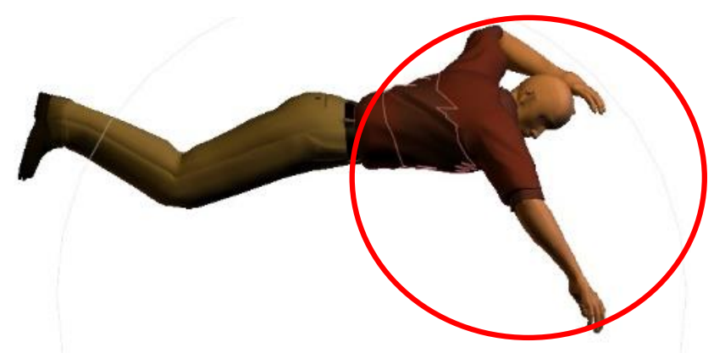
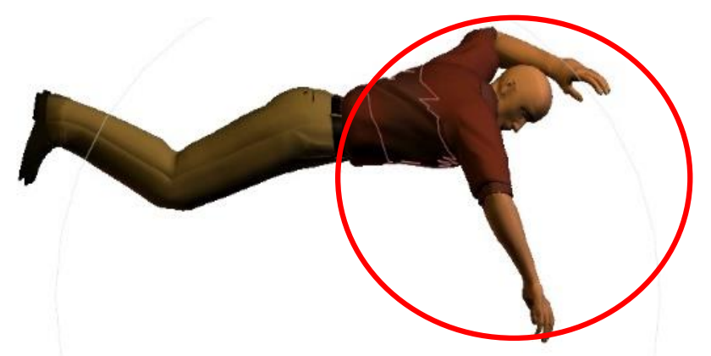
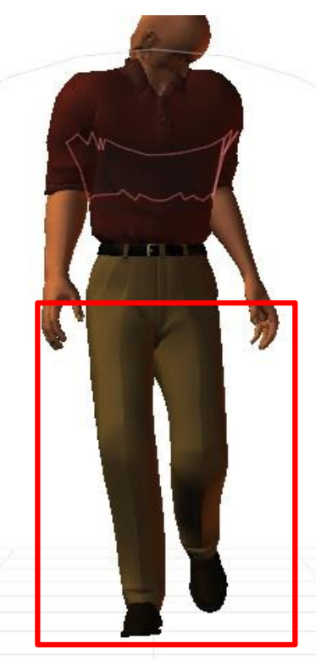
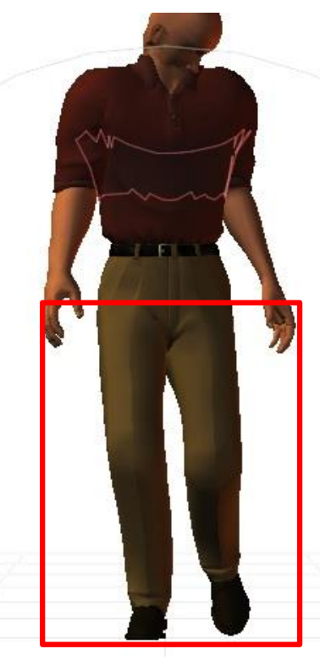
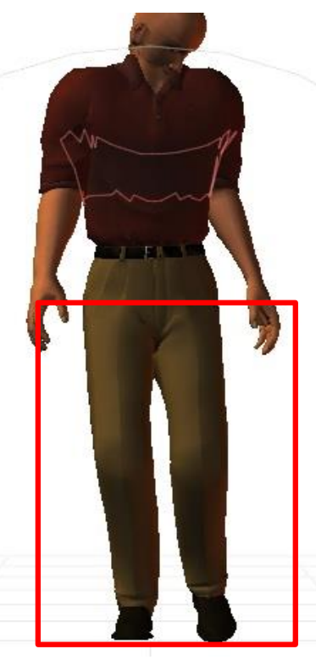
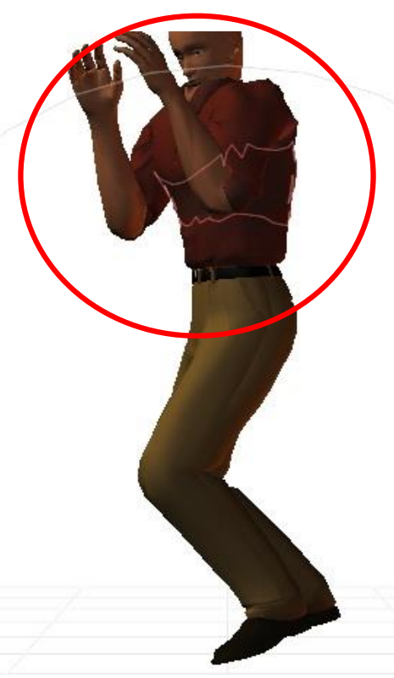
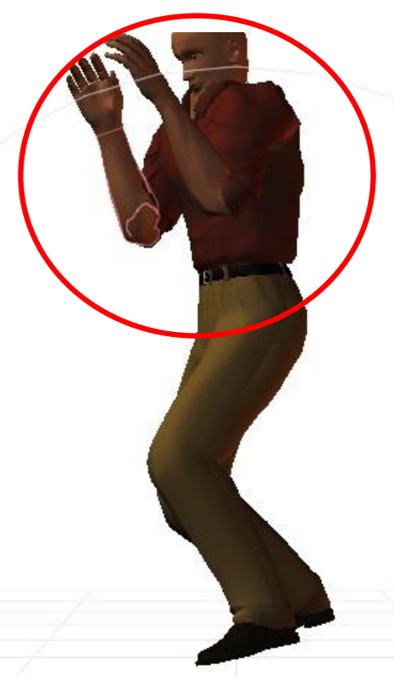
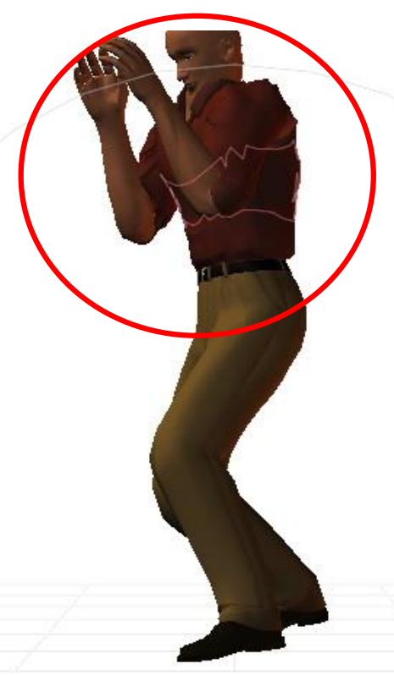


Figure 6: Red edges as shortcut edges

Activity	Ground truth	Our method (G_f^p+G_t)	Base graph (G_f)
Swimming			
Navigate			
Boxing			

6 Conclusion

- Increase in the pose estimation performance by constructing a more dependable graph
- Due to the removal of shortcut edges, the graph becomes more dependable
- Elimination of the requirement for a large number of labeled data with semi-supervised learning

7 Acknowledgments

This work was funded through grants from NSERC (Canada). The first author (Abbasi) would like to thank Digital Media Lab, AICTC Research Center (Sharif University of Technology) for supporting her during her Master's thesis.