Supplementary material

YoungJoon Yoo[1] Sangdoo Yun[2] Hyung Jin Chang[3] Yiannis Demiris[3] Jin Young Choi[2]
[1]Graduate School of Convergence Science and Technology, Seoul National University, South Korea
[2]ASRI, Dept. of Electrical and Computer Eng., Seoul National University, South Korea
[3]Personal Robotics Laboratory, Department of Electrical and Electronic Engineering
Imperial College London, United Kingdom
[1]yjyoo3312@gmail.com [2]{yunsd101, jychoi}@snu.ac.kr [3]{hj.chang, y.demiris}@imperial.ac.uk

1. Implementation Detail

In the experiments, the encoder $[m_{i,y}, \sigma_{i,y}] = E(y_i; W_E)$, decoder $D(z; W_D)$ and $\sigma(y; W_E)$ of the kernel for GP regression in Figure 2 of the paper are defined as multi-layered perceptrons. The encoder $E(y; W_E)$ is designed with five convolution layers and one fully connected layer (the convolution layers are composed of $16, 32, 64, 128, 256$ channels with filter size 5×5 each). All $y \in \mathcal{Y}$ are resized to three channel 64-by-64 images. We set the dimension of the $m_{i,y}$ and $\sigma_{i,y}$ to 128, and the fully connected layer returns the 256 elements for $m_{i,y}$ and $\sigma_{i,y}$. The former 128 elements are used as $m_{i,y}$, and the latter 128 entries are defined as $\sigma_{i,y}$. For the mapping function $[m_{i,x}, \sigma_{i,x}] = f(x_i, W_x)$, $m_{i,x}$ refers to $x_i \in \mathcal{R}^{n(\mathcal{X})}$ and the additional $n(\mathcal{X})$ outputs in $E(y; W_E)$ indicates $\sigma_{i,x}$, as in Figure 2 of the paper. Therefore, the overall dimension of the final fully connected layer is $256 + n(\mathcal{X}) + 1$; 256 dimensions for $[m_{i,y}, \sigma_{i,y}]$, $n(\mathcal{X})$ dimensions for $\sigma_{i,x}$, and one dimension for σ_k. For the decoding function $\hat{y} = D(z; W_D)$, 6 convolution layers with 2-by-2 upsampling are used to reconstruct the image. The convolution layers have 256, 128, 64, 32, 16, 3 channels with filter size $4 \times 4, 5 \times 5, 5 \times 5, 5 \times 5, 5 \times 5$ and 5×5.

2. Additional Results

To check the validity of the regression procedure conducted in a latent space, we compared the latent vector obtained by regression with the latent vector obtained by reconstruction for an input data pair. For the reconstruction, we used both the joint vector and the corresponding image in the H3.6m dataset [1]. For the regression, only the joint vector was given and the projected point was estimated by the proposed regression method. Since the latent vectors were obtained from the same input data, in ideal conditions the vectors should converge to the same location. Figure 1 shows the qualitative results for the regression and the reconstruction for the same input data pair, where it can be seen that both responses converged to the ground truth image. The graph in Figure 2 indicates the KL-divergence between the two latent vectors.

Figure 1. Regression and reconstruction result from the same data pair. Each image set is composed of three images. Within each set, the leftmost image is generated from a regression; the middle image refers to the result from reconstruction using the joint vector and the corresponding image; and the rightmost image shows the ground truth.

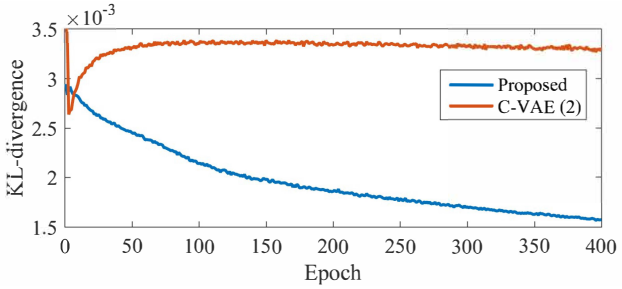

Figure 2. KL divergence between the latent distributions for regression and reconstruction, from the same joint vectors.

Since the latent vector z in the paper is defined by a Gaussian distribution, we used the KL-divergence as distance measure. As seen in the graph, the KL-divergence obtained by the proposed method was gradually decreased. The result demonstrates that the two vectors obtained from both cases converged to the same location, as expected. When tested with the C-VAE (2), the divergence did not

Figure 3. Qualitative results on regression from the baseball swing dataset. The first row in each action represents the proposed method, and the second row shows the result from R-VAE.

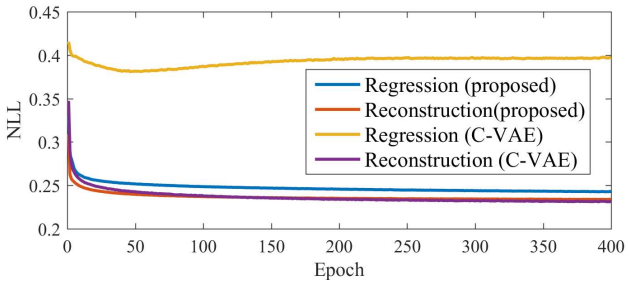

Figure 4. Negative log likelihood ratio for the regressed and reconstructed visual responses, from the proposed method and C-VAE (2).

converge.

As shown in Figure 4, we also measured the changes of the negative log likelihood (NLL), for both the regression and the reconstruction. In the proposed method, we confirmed that the NLL ratio for both cases converged. In C-VAE (2), the NLL ratio converged only when both the joint vector and the images were given, but it did not successfully converge when the regression was applied.

Figure 3, Figure 6 and Figure 7 show additional generation results of sports sequences. The figures describe the regression results of the proposed method and of R-VAE in our work, which are the supplementary results of Figure 7 included in the submitted version. We confirmed that the proposed method achieved a superior regression performance for diverse action sequences compared to R-VAE.

In addition to the image sequence regression and human pose reconstruction examples, we newly performed human joint estimation experiments as shown in Fig. 5. Although our method is not originally designed for the human joint estimation purpose, the proposed method could successfully estimate the human joints by performing regression using an (image - joint) data pair and finding the corresponding joints when a new image is given. This example further shows that the proposed method is applicable to practical applications composed of various data pairs.

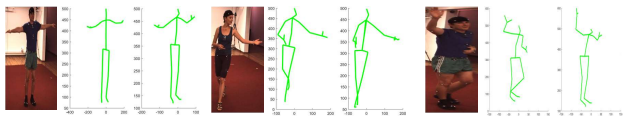

Figure 5. left: input image, center: ground truth, right: regressed result.

References

[1] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 1

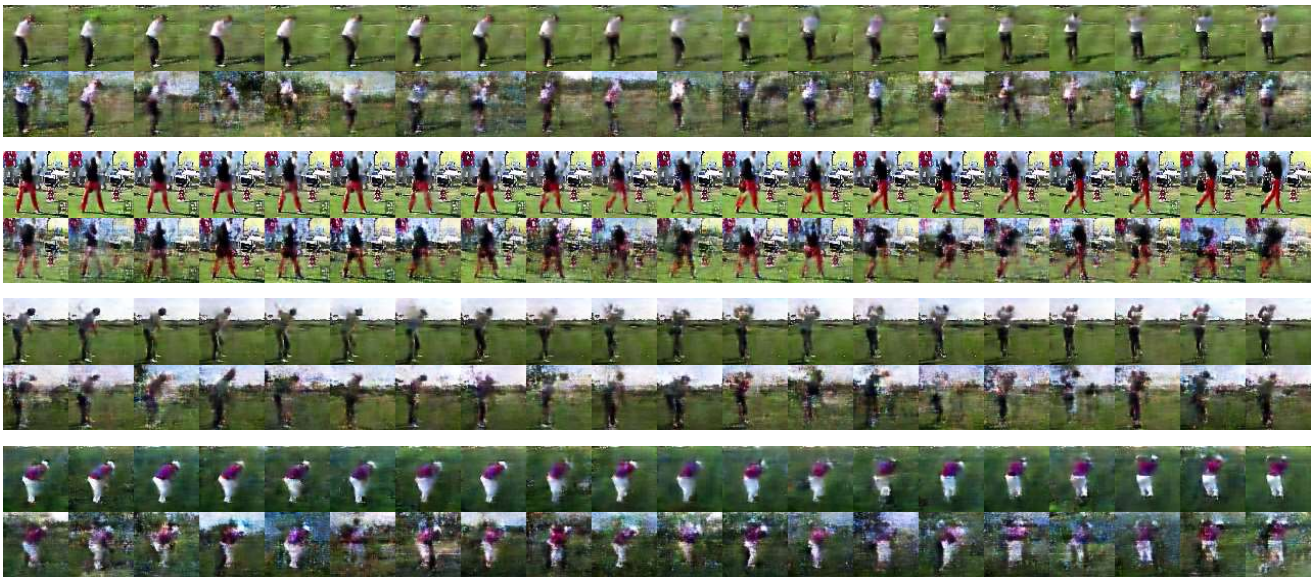

Figure 6. Qualitative results on regression from the golf swing dataset. The first row in each action represents the proposed method, and the second row shows the result from R-VAE.

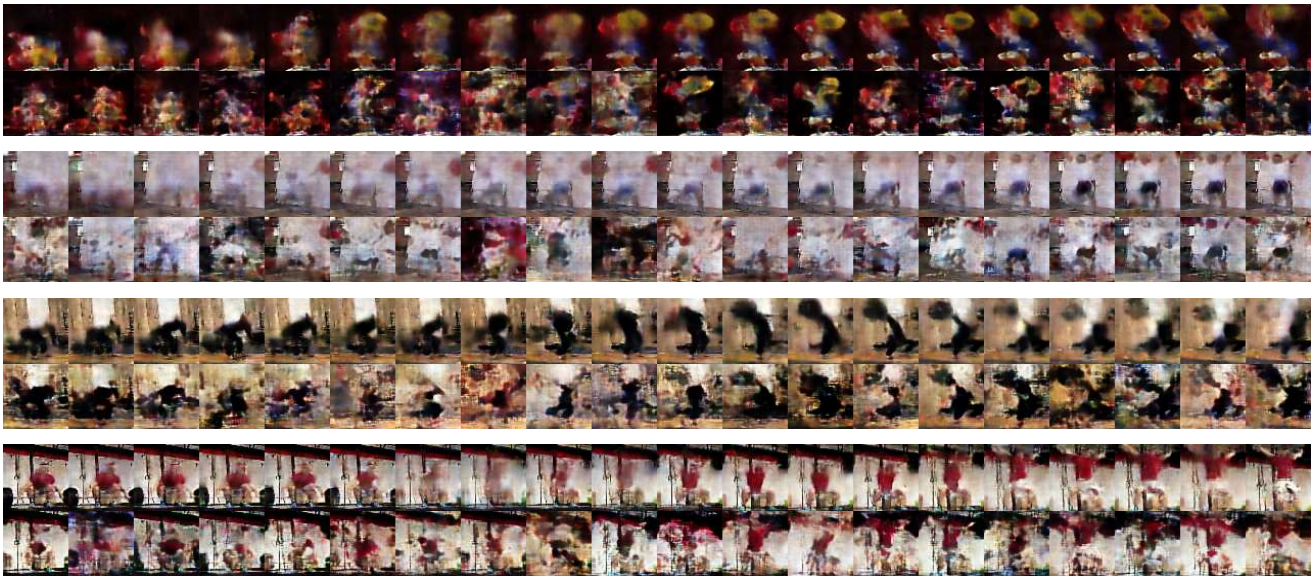

Figure 7. Qualitative results on regression from the weightlifting dataset. The first row in each action represents the proposed method, and the second row shows the result from R-VAE.