Everybody Dance Now

2018-08-28

1. どんなもの？

- モーション変換の手法を提案．
- 踊っている人のビデオ(source video)があれば，標準的な動きをしているターゲットにその動きを転送することができる．
- 時空間の滑らかさを用いてフレームごとの画像から画像への変換としてこの問題を扱う．

2. 先行研究と比べてどこがすごいの？

- ビデオ間の人のモーション変換に関する学習ベースのパイプラインを提案したこと
- 現実的で詳細なビデオにおいて，複雑なモーション変換できるという結果

3. 技術や手法の“キモ”はどこにある？

- ゴールは，ソースと同じ動きをするターゲットの新しいビデオを生成すること．
- 3つのステージからなる．

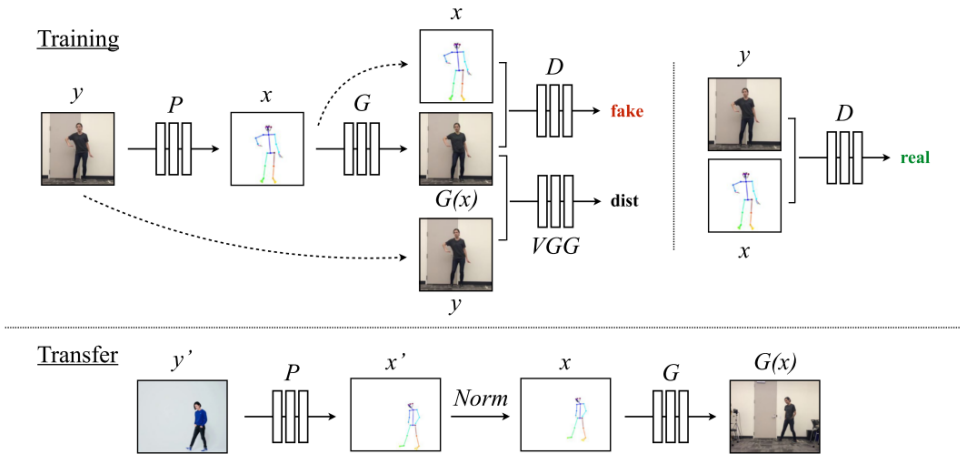

Fig. 3. (Top) **Training**: Our model uses a pose detector P to create pose stick figures from video frames of the target subject. During training we learn the mapping G alongside an adversarial discriminator D which attempts to distinguish between the "real" correspondence pair (x, y) and the "fake" pair $(G(x), y)$. (Bottom) **Transfer**: We use a pose detector $P : Y' \rightarrow X'$ to obtain pose joints for the source person that are transformed by our normalization process $Norm$ into joints for the target person for which pose stick figures are created. Then we apply the trained mapping G.

Figure 1

- pose detection
  - pretrainingされたSOTAのpose detectorを使って，pose stick figuresを作成
- global pose normalization
  - ソースとターゲットの体の形と位置間の違いを説明する．
  - 高さと足首の位置を分析することによって，この変換を探す．
- mapping from normalized pose stick figures to the subject
  - adversarial trainingを用いて，normalized pose stick figuresからターゲットの画像への変換（写像）を学習する．
  - 独立したフレームを生成する代わりに，２つの連続フレームを生成する．
  - 一つは，$G_{x_{t-1}}$．これは，対応するpose stick figure x_{t-1} とゼロ画像に条件付けられている．
  - 2つ目の出力は，G_{x_t} は対応するpose stick figure x_t と最初の出力$G_{x_{t-1}}$ に条件つけられている．
  - Discriminatorは，fake sequence $(x_{t-1}, x_t, G_{x_{t-1}}, G_{x_t})$ とreal sequence $(x_{t-1}, x_t, y_{t-1}, y_t)$ 間の現実と時間一貫性の違い両方の差を決定することを任されている．

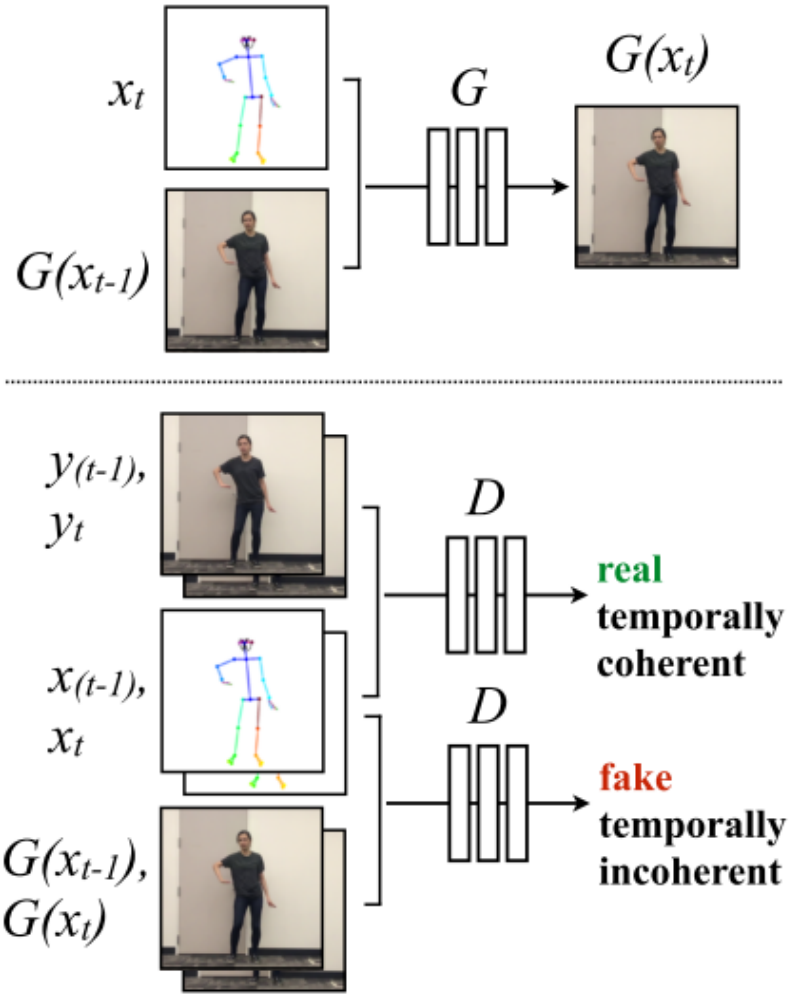

Fig. 4. Temporal smoothing setup. When synthesizing the current frame $G(x_t)$, we condition on its corresponding pose stick figure x_t and the previously synthesized frame $G(x_{t-1})$ to obtain temporally smooth outputs. Discriminator D then attempts differentiate the "real" temporally coherent sequence $(x_{t-1}, x_t, y_{t-1}, y_t)$ from the "fake" sequence $(x_{t-1}, x_t, G(x_{t-1}), G(x_t))$.

Figure 2

- 最初にメインGeneratorとDiscriminatorを訓練する.

$$\min_G \left(\left(\max_{D_1,D_2,D_3} \sum_{k=1,2,3} \mathcal{L}_{\text{smooth}}(G, D_k) \right) + \lambda_{FM} \sum_{k=1,2,3} \mathcal{L}_{\text{FM}}(G, D_k) + \lambda_{VGG} \left(\mathcal{L}_{VGG}(G(x_{t-1}), y_{t-1}) + \mathcal{L}_{VGG}(G(x_t), y_t) \right) \right) \tag{5}$$

Figure 3

$$\mathcal{L}_{\text{smooth}}(G, D) = \mathbb{E}_{(x,y)}[\log D(x_{t-1}, x_t, y_{t-1}, y_t)] + \mathbb{E}_x[\log(1 - D(x_{t-1}, x_t, G(x_{t-1}), G(x_t))] \tag{3}$$

Figure 4

- generatorとdiscriminatorの重みは固定し，face GANの訓練を行う．

$$\min_{G_f}\left(\left(\max_{D_f}\mathcal{L}_{\text{face}}(G_f,D_f)\right)+\lambda_{VGG}\mathcal{L}_{VGG}(r+G(x)_F,y_F)\right) \quad (6)$$

Figure 5

4. どうやって有効だと検証した？

- ablation study
  - temporal smoothingとface GANをいれたとき
- 評価値
  - SSIM
  - Learned Perceptual Image Patch Similarity(LPIPS)
  - キーポイントの平均の距離
- 結果

Loss	SSIM mean	LPIPS mean
pix2pixHD	0.89564	0.03189
T.S.	0.89597	0.03137
T.S. + Face **[Ours]**	**0.89807**	**0.03066**

Table 1. Body output image comparisons - result cropped to bounding box around input pose. For all tables, T.S. denotes a model with our temporal smoothing setup, and T.S. + Face is our full model with both the temporal smoothing setup and Face GAN.

Loss	SSIM mean	LPIPS mean
pix2pixHD	0.81374	0.03731
T.S.	0.8177	0.03662
T.S. + Face **[Ours]**	**0.83046**	**0.03304**

Table 2. Face output image comparisons - result cropped to bounding box around input face

Loss	Body (23)	Face (70)	Hands (21)	Overall (135)
pix2pixHD	**2.39352**	1.1872	3.86359	2.0781
T.S.	2.63446	1.14348	3.76056	2.06884
T.S. + Face **[Ours]**	2.56743	**0.91636**	**3.29771**	**1.92704**

Table 3. Mean pose distances, using the pose distance metric described in Section 7. Lower pose distance is more favorable.

Loss	Body (23)	Face (70)	Hands (21)	Overall (135)
pix2pixHD	0.17864	0.77796	1.67584	2.63244
T.S.	0.15989	0.56318	1.76016	2.48323
T.S. + Face **[Ours]**	**0.15578**	**0.47392**	**1.66366**	**2.29336**

Table 4. Mean number of missed detections per image, fewer missed detections is better.

Figure 6

Fig. 8. Face image comparison from different models on the validation set. T.S. denotes a model with our temporal smoothing setup, and T.S. + Face is our full model with both the temporal smoothing setup and Face GAN. Details improve and distortions decrease upon the additions of the temporal smoothing setup and the face GAN.

Figure 7

5. 議論はあるか？

- 計算コストが高そう．
- 顔の構造を学習させた方が，いい形状になっている

6. 次に読むべき論文はあるか？

- Pose detector[5, 27, 35]
    - Zhe Cao, Tomas Simon, Shih-EnWei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In CVPR.
    - German KM Cheung, Simon Baker, Jessica Hodgins, and Takeo Kanade. 2004. Markerless human motion transfer. In 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on. IEEE, 373–378.
    - Shih-EnWei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In CVPR.

- pix2pix
    - Ting-ChunWang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2017. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. arXiv preprint arXiv:1711.11585 (2017).
- OpenPose
    - ネットワークアーキテクチャ
    - Zhe Cao, Tomas Simon, Shih-EnWei, and Yaser Sheikh. 2017. Realtime Multi- Person 2D Pose Estimation using Part Affinity Fields. In CVPR.
    - Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. HandKeypoint Detection in Single Images using Multiview Bootstrapping. In CVPR
    - Shih-EnWei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convo- lutional pose machines. In CVPR.
- Learned Perceptual Image Patch Similarity
    - Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and OliverWang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In CVPR.

論文情報・リンク

- Chan, C., Berkeley, U. C., Ginosar, S., Zhou, B. T., & Efros, A. A. (2018). Everybody Dance Now,