International Journal of Computer Vision (2020) 128:1376–1377
https://doi.org/10.1007/s11263-020-01319-w

GUEST EDITORIAL

Special Issue on Generating Realistic Visual Data of Human Behavior

Xavier Alameda-Pineda[1] · Elisa Ricci[2] · Albert Ali Salah[3] · Nicu Sebe[4] · Shuicheng Yan[5]

Published online: 9 March 2020

The fast and broad progress in AI has not only enabled great advances in the analysis of human behavior but has also opened new possibilities for generating realistic human-like behavioral data. This special issue focuses on recent advances and novel methodologies for generating data containing human-like behaviour at multiple scales, and contains eight papers, summarily described. Two types of applications, generation of realistic facial behaviours and body movements, are investigated and represented by four papers each. Methodologically, generation of realistic synthetic data helps with automatically established ground truth labeling and data augmentation, as well as for animation and content generation. The contributions from all of these perspectives are presented in this special issue, which brings together eight contributions, selected from 17 submissions following an open call for papers. Each paper was rigorously peer-reviewed for one or two rounds of revisions according to the journal's high standards.

The first paper in the series, "Generating Human Action Videos by Coupling 3D Game Engines and Probabilistic Graphical Models" by De Souza, Gaidon, Murray, Cabon, and López is on generation of synthetic training data for video based action recognition. With an interpretable parametric model that combines elements of game engines, the proposed approach generates videos for natural and parametrically defined action sequences. These videos are easily employed for training data augmentation, boosting the recognition performance of classifiers they are used with. The paper creates a novel database with this approach, and provides pixel-level semantic segmentation, which is important for creating powerful loss functions.

The problem of semantic segmentation is taken up in more detail in "A Weakly Supervised Multi-task Ranking Framework for Actor-Action Semantic Segmentation" by Yan, Xu, Cai and Corso, who present a new model for weakly-supervised, pixel-level actor-action segmentation, where only video-level tags are given for training samples. While fully supervised approaches to the problem have access to pixel-level segmentation ground truth during training, this is not the case for the proposed algorithm. A robust Schatten p-norm multi-task ranking model is developed to select the most representative supervoxels and action tubes for actor, action and actor-action, respectively.

For human behavior understanding in video, semantic segmentation at the pixel level can be seen as local modeling, whereas aggregation of motion cues is at a more global level of analysis. In "Adversarial Framework for Unsupervised Learning of Motion Dynamics in Videos," Spampinato, Palazzo, D'Oro, Giordano, and Shah propose a generative adversarial network (GAN)-based framework to learn video representations and dynamics through a self-supervision mechanism at both levels. In addition to modeling the static content, object trajectories are learned in a dynamic motion latent space. Using motion masks, the approach relies on self-supervision, and significantly reduces the amount of labeling required to reach a state-of-the-art accuracy, as well as producing realistic synthetic videos.

The idea of using interpretable latent spaces during analysis, and later, during synthesis, is a powerful one. The last paper on body-level synthesis, "DGPose: Deep Generative Models for Human Body Analysis" by de Bem, Ghosh, Ajanthan, Miksik, Boukhayma, Siddarth, and Torr proposes several models to disentangle the body pose and the visual appearance, and to represent the relevant variations in an interpretable latent space. These models implement GAN architectures, where the pose is represented as a heatmap, rather than as a vector of keypoints. The approaches are used to perform cross-domain pose-transfer, for hallucinating peo-

✉ Albert Ali Salah
a.a.salah@uu.nl

1 Inria Grenoble Rhône-Alpes, 38330 Montbonnot Saint-Martin, France

2 FBK and DISI, University of Trento, Via Sommarive 9, 38123 Trento, Italy

3 Utrecht University, Princetonplein 5, 3584CC Utrecht, The Netherlands

4 DISI, University of Trento, Via Sommarive 9, 38123 Trento, Italy

5 Vision and Machine Learning Lab, National University of Singapore, Singapore 117583, Singapore

ple in a large selection of poses, and for image reconstruction conditioned on pose.

The second set of four papers in the special issue deal with generating faces. In "Towards High Fidelity Face Frontalization in the Wild," by Cao, Hu, Zhang, He, and Sun, the authors focus on generating a frontal face from a profile face, which helps with various face processing tasks including face recognition. Their approach uses a novel texture fusion warping procedure and leverages a dense correspondence field to bind the 2D and 3D surface spaces. Like many recent approaches in this area, the paper uses an adversarial loss, but does not extend its methodology to a full GAN. Nonetheless, it provides comparisons with several recent GAN-based algorithms.

In "Masked Linear Regression for Learning Local Receptive Fields for Facial Expression Synthesis" by Khan, Akram, Mahmood, Ashraf and Murtaza tackle a related problem, where they seek to synthesize a novel expression in a given face. The proposed method is a constrained version of ridge regression that exploits the local and sparse structure of facial expressions. This method is also compared with recent GAN-based approaches, and was shown to perform better when the training and testing distributions are not similar.

In a third paper on facial image generation, "Deep Neural Network Augmentation: Generating faces for affect analysis", Kollias, Cheng, Ververas, Kotsia and Zafeirou propose an algorithm that takes a neutral 2D face as input, and synthesizes a single image with target affect label, or a sequence of images that describe an affective change. They use a 3D morphable model to fit the 3D shape on the neutral image, deform the image according to the desired affect, and blend the new face into the old one. The paper shows that faces synthesized with this approach serve well for data augmentation, and when used with state of the art classifiers, they improve the accuracy significantly.

Generating realistic facial motion is difficult, because the human brain is specialized in face processing, and sensitive to minute changes and inconsistencies. In "Realistic Speech-Driven Facial Animation with GANs", Vougioukas, Petridis and Pantic present an end-to-end system to generate videos of a talking face by combining a static image of the face with an audio clip of the speech input that drives the animation. The method is based on a temporal GAN architecture, which has different discriminators to focus on the accuracy of facial details, multimodal synchrony, and realistic expressions. The paper also uses an ablation study to gain insights on different aspects of synthesis, evaluating the videos on sharpness, reconstruction quality, lip-reading accuracy, synchronization, and even natural blinking behavior.

Taken together, these eight contributions provide rich insights into the state-of-the-art for realistic human behavior synthesis. The special issue was preceded by the 9th International Workshop on Human Behavior Understanding, held at ECCV (September 2018 in Munich), organized by the guest editors with the focus theme of the special issue. Three articles are extended versions of contributions in this workshop. We would like to thank all authors for their valuable contributions, and all our reviewers, who devoted time and energy to improve these contributions with their constructive comments.