Visual AI

CPSC 532R/533R – 2019/2020 Term 2

Lecture 15. Notes on reviews

Helge Rhodin

Good review structure and length

- structured
- concise
  - more than a line per section
  - about half a page in total

Main Contribution
The authors developed a model that generates plausible 3D human poses representing opportunities for interaction (affordances) in a 3D environment from a single 2D image. This is accomplished in two steps: pose data synthesis (to produce ground-truth poses) and 3D affordance generation (with jointly trained location and gesture prediction components).

Discussion Question
Would trying to model the relationships/hierarchies of poses help improve the results of the gesture prediction component (ie. the *what* module)?

Strength
The authors conduct a detailed evaluation of their model, including a pose authenticity classification score, user studies, ablation studies, and comparisons to state of the art. They show that their model outperforms baseline methods from prior work.

Weakness
The model still has significant failure cases, including cases where the poses generated are unusual (semantic misalignment with the scene) and/or unrealistic (for example, collisions).

Paper Review for Everybody Dance Now

Main Contributions

- They propose a surprisingly simple method for generating compelling results on human motion transfer (having a source video of a person dancing, they transfer the moves to a target person). At first, they extract poses from the source subject and apply the pose-to-appearance mapping to generate the target subject.
- They have a separate pipeline for face synthesis.
- They also release a dataset of videos that can be used for motion transfer.

Strength Point

- Although their method is quite simple, it produces surprisingly high quality and detailed results.
- Their method is the first one that tackles dance motion transfer.

Weakness Point

- The strength of their model highly depends on the strength of the pretained pose estimator. (if it doesn't work well their model fails)
- Their method is only usable in transferring human motion, not anything else. Because they take use of a pretained human pose detector.

Questions

- Can they use multiple GANs each specified to one body part, similar to a specialize GAN for face? e.g. for hands, feet, etc.
- Can they use landmarks instead of poses? In this way their model would be generalized and applicable for objects, animals, etc.

Missing structure

- clearly highlight
  - summary
  - a strength
  - a weakness
  - a question

Everybody Dance Now

Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros

To briefly summarize, this work combines a variety of existing computer vision techniques to achieve very high-quality results in 2D pose and motion transfer. The heart of the method is image-to-image translation using generative adversarial networks conditioned on a simple posed skeleton model as the intermediate representation. A dataset is additionally shared to encourage related work.

My general impression of this paper is that it has a bit of an engineering feeling to it, in that many existing methods (such as pre-trained pose detection, GANs, feature-matching loss, perceptual reconstruction loss, etc.) are combined without much rationalization other than an ablation study. Apart from achieving very visually pleasing results in a specific domain, it's not clear to me what the novel contribution of this paper actually is.

I also found myself wondering why the authors decided to use dense 2D images of posed skeletons rather than a simple vector of joint angles. While the image-based approach is certainly a bit more data intensive, I imagine this could lead to better spatial coherence, as the task of localizing the posed humanoid shape in the image is already done.

I very much appreciate the discussion of ethical concerns about making high-quality fake video available to the public and the creation of a fake-detector is a nice addition. However, I imagine that a knowledgeable reader could extend this method and train the system to outsmart the fake detector with only minor tweaks.

One thing that I wish was clearer is that the generator appears to trained separately for each target subject. There is no mention of how target person's appearance or the background on which they appear are represented, so I'm led to believe that a separate generator is trained for each target person and their background. This does not appear to be clearly stated, and yet it means that this method does not generalize at all to new scenes and new people without first performing costly (and chaotic) neural network training.

Review2: Everybody Dance Now

February 26, 2020

This paper by Chan et al. proposes a method for performing style transfer of a source video of a dancing person, making the person in some target video doing the same movements as the source video. Their architecture is a Generative Adversarial Network (GAN) with a multidimensional discriminator in the loss function. The discriminator exist of a threefold: The pix2pix discriminator, one that takes care of smooth frame transition and one that specializes in faces. The losses are simply summed together. The latter two are added, because temporal artifacts are a strong indicator for the videos being fake, as well as the faces.

I love that they show the effect of the different factors in the loss (usual, adding smoothness and adding face loss) and the effect on the output. You can clearly see the improvement of adding the face GAN.

They compare their method to nearest neighbour (picking the most similar existing frame in the target set), and the PoseWarp method. Nearest neighbour is probably a very bad baseline, assumed that the target dataset is not infinitely big: the result will be not smooth from frame to frame and also the pose probably doesn't match perfectly. However, since these images are real, they score high in the loss for face and overall image. That is why it is interesting that they included this in the comparison too. They use useful metrics to evaluate their work.

In conclusion, I think this is a very interesting and clear paper with a lot of potential applications and further research.

Too long

- Summary should be short

Paper Title: Everybody Dance Now

Summary:
This paper focuses on a motion transfer task: given a source video of a person dancing, the architecture can output a video with a target person dancing in the same way. This work mainly consists of three components: pose detection, global pose normalization, and mapping from normalized pose stick figures to the target subject.

For pose detection, the arthurs simply make use of the state-of-the-art pose detector to generate 2D cartesian coordinates of joint positions. To account for difference of standing position and limb proportions of the source and target persons, global pose normalization is invented to calculate the scale and translation factor of each frame. The major component in this work is the actual pose to video translation, which is composed of two parts. The first part targets learning a mapping from the intermediate pose representation to image with a conditional GAN. Instead of predicting one frame at a time, two frames are predicted with appropriate loss to encourage coherence between images. However, with a single conditional GAN is not sufficient as it is hard to capture the details of human faces. Therefore, another separate GAN focusing on human faces is trained to remedy this. During the training time, the conditional GAN is trained first, and then the face GAN is trained with the weights of the conditional GAN frozen.

Question:
This work predicts two video frames at a time which the second one is conditioned on the first output. However, the first one is conditioned on a zero image, which probably does not contain too much information. Would it be useful if the conditional GAN has a fully recurrent structure, such that the first one can also be conditioned on the previous outputs.

Strength:
This work is capable of generating high-quality video with a person dancing given another target video. There is no need for the target person to do any dancing for the training. Also, the face GAN is applied to produce detailed and natural human facial motion.

Weakness:
If the source motion is very extreme, there could be artifacts in the generated motion. Also, by looking at their video, there exists some footskating artifacts as well, which might be hard to remove.

Everybody dance now

Chan et al.

This paper focuses on "do-as-I-do" motion transfer, i.e. the task of automatically transferring motion from a source to a target subject in videos. Given a source video of a person dancing, the aim of this paper's work is to synthesize a new video consisting of a target subject performing the same source dance. This paper builds upon pose estimation and image-to-image translation, and in particular it exploits GANs for the latter. The paper proposes to solve this video-to-video translation task with a simple yet very effective method that consists of 3 steps. First, it uses a state-of-the-art pose estimation algorithm to extract poses from the source subject. Then, it performs global pose normalization. In this step it modifies the extracted pose in order to adapt it to the target person's body shape and location in the frame. The final step is the actual video synthesis and it relies on an adversarial training process. At training time, it learns the mapping from poses to images of the target person; in particular, it extracts poses from video frames of the target subject and tries to learn the mapping (a generator) together with a discriminator. At transfer time it applies the learned pose-to-appearance mapping to generate the target subject from the normalized pose obtained in the previous step. In order to get better results, two methods are employed in the video synthesis process. Firstly, in order to get temporally coherent video results, the system is built to predict two consecutive frames. This temporal smoothing mitigates temporal incoherences that occur when video translation is performed frame by frame. Furthermore, a specific separate GAN is employed to only work on faces; this additional part allows one to obtain realistic face synthesis. In order to perform evaluation, the method is compared with two baseline methods. Moreover, an ablation study is performed; the performance of the full model is compared to a simple frame-by-frame synthesis method and to a frame-by-frame system enhanced with temporal smoothing. The experimental results show enhancing in performance of the full model compared both to the baselines and to the ablated models.
The paper also presents a fake-detector mechanism that identifies fake video synthetized by the proposed motion transfer system. This method also works on consecutive frame pairs of the video.

Question: Is it possible to exploit the discriminator obtained from the training phase of the video synthesis model, in order to perform fake video detection?

Strength: The work on the paper is based on easy ideas yet very powerful. To train the synthesis model, it does not require the target subject to perform a specific reference dance or acquire specific poses; this facilitates the collection and elaboration of the training data, which can be a difficult process.

Weakness: Each trained synthesis model is specific of a target subject. This means that the model is not general: it is necessary to train a different a model to generate video of each specific target subject.

February 20, 2020

1 Paper Summary and Contributions

- This paper introduces a frame-work for pose translation from one video to the other. Specifically, it provides a framework where source videos of individual A performing basic moves and individual B dancing, it can transfer the dancing style of B onto the semantic scenery and looks of individual A.
- The model is able to achieve this through several novel contributions. The model can be summarized in the following steps: 1. obtain pose information for both target videos, 2. train a GAN for re-generating videos of person A from the poses of that video, 3. obtain frame-by-frame pose information for video of person B, 4. normalize poses of A and B to be of the same scale/joint angels and 5. use this pose information as input to the GAN in step 2, therefore synthesizing a video of person A following poses of person B. The work uses an off-the-shelf method for obtaining poses with a three component (or otherwise variation) GAN to complete step 2 and 5.
- This GAN specifically consists of three modules/extentions: A pose-to-frame generator where given pose, the generator generates an image of that pose and the discriminator not only verifies realness but also correctness of pose to image translation by comparing the synthesized image to the corresponding groun truth frame. This architecture is then extended to account for temporal structure in the video by pair-wise/triplet generation of temporal frames together. Lastly, a Face GAN is trained afterwards for building a residual mask for fine-tuning the faces in the frames.
- The model is able to achieve superior performance as measured by Structural Similarity and Learned Perceptual Image Patch Similarity. Qualitative studies also prove it to produce more realistic videos to an astonishingly degree.

2 Question

- I wonder if with the advent of pose estimators that can to a relatively accurate degree) estimate 3D pose from 2D images/videos, can we see further improvements if we condition the generator on a 3D pose estimation? Additionally, would pose generations on the whole video with intra-frame conditioning also account for temporal dependency within poses?

3 Strength

- The temporal addition to the Pix2Pix GAN architecture along with the Face GAN significantly improves the quality of the synthesized images. The use of pose as an intermediary representation as oppose to vector latent representation as used by others is also very interesting and powerful.

A bit short

Motivate your statements

- why, how, what?
- a single line per question is generally not enouth

Everybody Dance Now

Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros

Main Contribution

The authors developed a method for transferring human motion from a source video to a target person from only a short video of the target moving. This is accomplished by training a pose detector, normalizing produced poses, and learning a mapping from normalized pose stick figures to the target. Additionally, the authors have produced a dataset of videos that can be used for motion transfer learning.

Discussion Question

How would you approach finding the minimal set of target training data poses required for a specific application?

Strength

The authors also trained a fake detector in parallel that is capable of detecting videos that are synthesized using the motion transfer method of the paper with very high accuracy (> 95%).

Weakness

The model tends to produce visual artifacts in complex cases, such as with loose clothing, hair, or occluded limbs.

Everybody Dance Now

A summary of the main contribution:

In this paper, the authors proposed a novel model for motion transfer. The model only requires a few minute motion clip of the target subject, which does not have a synchronized motion to the source motion video. Therefore, the method is unsupervised. They also provided a forensic tool for fake content detection and an open dataset for motion transfer.

One question that is well suited for discussion:

Although the overall videos look reasonable, the shape of the human in the target video deforms very unrealistically. Is it because they did not explicitly model human shape?

One strength:

The method is unsupervised and very simple to implement.

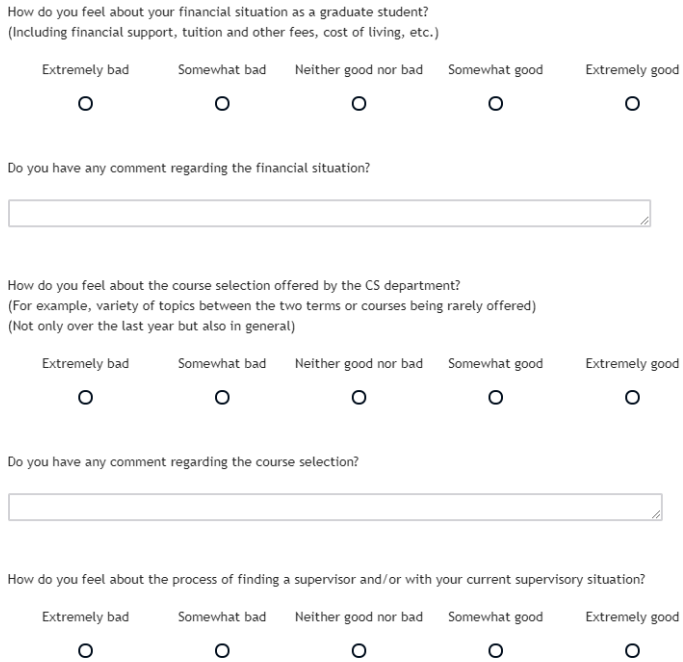
One weakness:

Since their method relies on 2d human pose detection, they are limited by the common problems in 2d keypoint detections, such as occlusion. Also the video background in the paper is static and not clutter.

Grad Student Satisfaction Survey

- by the Grad Affairs Committee
- in preparation of the External Review
  - entire CS department
  - every five years
  - influences how money will be spend in the next 5 years. You can influence it.

- Please let us know how you feel at UBC!
  - https://ubc.ca1.qualtrics.com/jfe/form/SV_9zxybxXiRJT9CHr
  - see mail by Grad Affairs Committee
  - ASAP ;)


How do you feel about your financial situation as a graduate student?
(Including financial support, tuition and other fees, cost of living, etc.)

Extremely bad | Somewhat bad | Neither good nor bad | Somewhat good | Extremely good

Do you have any comment regarding the financial situation?

How do you feel about the course selection offered by the CS department?
(For example, variety of topics between the two terms or courses being rarely offered)
(Not only over the last year but also in general)

Extremely bad | Somewhat bad | Neither good nor bad | Somewhat good | Extremely good

Do you have any comment regarding the course selection?

How do you feel about the process of finding a supervisor and/or with your current supervisory situation?

Extremely bad | Somewhat bad | Neither good nor bad | Somewhat good | Extremely good