Method

전체적인 pipeline은 landmark parametric model(LPM), LPMNet, 그리고 DPM 으로 크게 3가지로 구성되어 있습니다. 첫째로, LPM은 parameters의 linear combination으로 arbitrary face landmark를 reconstruction 할 수 있으며 face landmark에 parametric control을 가능하게끔 합니다. 한가지 중요한점은 face landmark는 head pose나 눈깜박임, 입벌림과 같은 facial expression들을 visualize 할 수 있기 때문에 LPM이 intuitive pose control의 source 역할을 한다는 점입니다. 그렇기에 head pose와 facial expression을 parameter로 control 할 수 있게끔 하는 것이 LPM의 주 목적입니다.

두번째로, LPMNet은 주어진 face image로부터 LPM의 parameter를 예측하는 function 입니다. Arbitrary face landmark는 LPM을 통해 만들어 질 수 있으나 random face landmark를 generation 하는 것이 아닌 특정 face landmark를 reconstruction 하기 위해서 어떤 parameter가 조합되어야 하는지는 parameter estimation이 필요합니다. LPMNet의 역할은 입력받은 face image의 ground truth face landmark를 reconstruction 하기 위한 parameter를 예측 하는것입니다.

마지막으로, LPM에서 보여준 intuitive pose control를 neural-talking-head model에서도 가능하게끔 하기 위해서 parameter space와 latent space간의 matching system이 필요하게 됩니다. DPM은 parameter space에서 latent space로 transform 시켜주는 역할을 수행하며 특정 parameter를 조절 하였을 때 landmark의 pose가 조절되는 것처럼 generated face image에서도 semantically 일치하는 pose가 control 되도록 하는 것이 목적입니다. 예를들어, LPM의 첫번째 parameter 조절하면 yaw가 조절 된다고 할 때 이와 똑같이 neural talking head model의 output face image의 yaw를 조절 하고 싶은 경우 facial expression이 변화하는 것과 같은 side effect 없이 yaw만 control 할 수 있는 latent space의 combination of elements를 찾아야 했습니다. StyleGAN 연구**[ref]** 에서는 control 가능한 layer나 element를 직접 찾아 보고하는 경우도 있으나 이는 상당한 노고가 들어갑니다. DPM은 이러한 parameter 변화에 의한 pose control 결과가 LPD**[ref]**에서도 동일하게 일어나기 위해서 latent vector를 어떻게 조작해야 하는지 mapping 시켜주는 역할을 수행합니다.

이어지는 section에서는 pipeline의 각 part의 detail 부분을 설명합니다.

Landmark Parametric Model(LPM)

3DMM is based on a data set of 3D face, assuming that all exemplar faces are in full correspondence~\cite{blanz1999@3dmm}. 그렇기에 3DMM과 같은 parametric model을 만들기 위해서는 3D scan이 필요하며 이는 2D face landmark를 detect 하는것 보다 더 큰 노고가 필요합니다. Still, learning 3DMM and fitting a learned 3DMM almost invariably involves detecting landmarks, thus inheriting many of the landmark features. (LPD에서 발췌)

본 논문에서는 intuitive pose control을 위해 Face landmark

$L=(x_{1},y_{1},x_{2},y_{2},.....,x_{n},y_{n})\in\mathbb{R}^{2n}$ coordinates consists of $n=68$ that mean each facial component such as eyes, nose, mouth, etc . 로부터 추출한 landmark parametric model을 제안합니다. Landmark는 hair style이나 피부와 같은 identity-specific feature를 포함하고 있지 않고 사람의 눈,코,입등이 abstract 하게 표현되어 있는 형태 이기에 face landmark를 이용하여 identity-agnostic 하고 pose-specific한 abstract representation이 가능하다고 생각 하였습니다. 그렇기에, 우리는 exemplar face landmarks의 linear combination으로 arbitrary head pose and expression을 generate 할 수 있다고 가정하였습니다.

Face landmark는 off-the-shelf 2d facial landmarks detector[ref:face_landmark] 를 사용하였기에 annotation과 같은 additional effort가 필요하지 않습니다.

우리는 intuitive pose control을 위한 landmark parametric model을 도출하기 위하여 PCA[ref pca]를 활용 하였습니다. PCA는 data compression이나 dimension reduction을 위해 주로 사용되며 parametric model을 만들 때[ref 3DMM series] 사용 되기도 합니다. PCA의 main algorithms은 covariance matrix의 eigen decomposition이기 때문에 data의 variation에 크게 영향을 미치는 요소들부터 차례대로 도출이 되며 각각의 eigen vector들은 서로가 orthogonal 하기 때문에 linearly independent 합니다. 그렇기에 face landmarks에 PCA를 수행하게 되면 face landmark coordinates의 변화에 크게 영향을 미치는 요소들부터 도출이 될것이라 생각했고 face landmark coordinates 전체에 영향을 미치며 direction이 서로가 orthogonal한 yaw, pith, roll 3개의 headpose가 먼저 도출이 될 것이다 라고 가정하였습니다. Because as blinking or speaking, the variation of face landmark corresponding with each facial components(e.g. eye, mouth, etc.) are relevantly smaller than when head poses are changed. 이렇게 구성된 parametric model은 yaw,pitch,roll에 대해 rig-like control을 가능하게 하며 later parameters는 facial expression을 control 할 수 있게 될것입니다.

Landmark parametric model은 set of face landmarks로 이루어져 있으며 이는 parameterized by the coefficients p. New arbitrary face landmark can be generated by summation of average face landmark $\bar{L}$ and linear combination of the parameters p and eigen vectors $e\in\mathbb{R}^{2n}$. In this case the maximum value of k is $2n$, because the data for landmark parametric model are much bigger than $2n$.

We fit PCA to Voxceleb1 dataset~\cite{Nagrani17@vox1} pre-processed all frames to be contain one main face is aligned at center using face detector~\cite{zhang2017s3fd}, so that certain parameter which affecting translation of landmark not to be extracted. (Figure.~\ref{fig:lpm}) shows that average landmark $\bar{L}$ visualizing neutral expression facing front. As our assumption, 1st, 2nd, 5th eigen

vector의 parameter를 interpolation 시켯을 때 각각 yaw, roll, ptich가 independent 하게 control 되는것을 확인 할 수 있었으며 각각의 headpose를 control 할 때에 facial expression과 disentangle 되어 있는 것을 확인 할 수 있었다. (Figure)

Landmark Parametric Model Net(LPMNet)

LPMNet의 기능은 face landmark detector와 유사하지만 key difference 는 LPMNet은 주어진 이미지로부터 face landmark의 coordinates를 detection 하는것이 아닌 face landmark를 reconstruction 하는 parameters를 estimation 한다는것입니다.

Face landmark detector가 detection한 face landmark와 일치하는 coordinates를 reconstruction 하는 parameter p를 예측한다.

Given an image $I\in\mathbb{R}^{H\times W\times3}$, LPMNet predict the parameters $p^{(I)}=(p^{(I)}_{1},p^{(I)}_{2},.....,p^{(I)}_{k})$ which reconstruct the face landmark $\hat{L}_{I}$ corresponds to an face landmark L_{I} from face landmark detector~\cite{bulat2017@face_landmark_detector}

LPMNet은 eccoder, decoder 형태로 이루어져 있으며 encoder는 주어진 이미지로부터 parameter p를 예측하고 decoder는 예측된 p 로부터 LPM을 이용하여 face landmark를 reconstruction 합니다. LPMNet encoder는 MobileNetV2 구조를 사용하였으며 LPMNet을 학습 시킬때에는 pred, gt 사이의 L1 loss를 사용 하였습니다. Decoder는 parameters를 linear combination 하는 역할을 수행함으로 학습은 encoder만 이루어집니다.

LPMNet의 주 목적은 intuitive pose control을 위해 parameter space와 latent space 사이의 bridge 역할을 하는 것이다. 그렇기에 face landmark reconstruction performance가 landmark detector 만큼 정교할 필요는 없다. 또한, face landmark 자체는 visualization 용도로만 사용될뿐 DPM을 학습하거나 inference 단계에서 사용되지 않는다는 점이 face landmark를 input으로 사용하는 FSAL과 다른 점이다.

하지만 parameter를 이용하여 head pose나 facial expression 과 같은 pose information을 전달해야 함으로 LPMNet으로 reconstruction 된 face landmark가 이러한 detail pose information의 embedding이 잘 되어 있는지가 중요하다. 우리는 k를 increasing 하면서 LPMNet의 performance를 측정해 보았고 landmark parametric model의 PCA variance ratio의 결과와 LPMNet의 face landmark reconstruction performance를 NME으로 measure하여 K=40으로도 충분히 arbitrary face pose를 representation 할 수

있음을 발견하였습니다. (결과는 experiments section ref)

DPM

마지막으로, LPM에서 보여준 intuitive pose control를 neural-talking-head model에서도 가능하게끔 하기 위해서 parameter space와 latent space간의 matching system이 필요하게 됩니다. DPM은 parameter space에서 latent space로 transform 시켜주는 역할을 수행하며 특정 parameter를 조절하였을 때 landmark의 pose가 조절되는 것처럼 generated face image에서도 semantically 일치하는 pose가 control 되도록 하는 것이 목적입니다. 예를들어, LPM의 첫번째 parameter 조절하면 yaw가 조절된다고 할 때 이와 똑같이 neural talking head model의 output face image의 yaw를 조절 하고 싶은 경우 facial expression이 변화하는 것과 같은 side effect 없이 yaw만 control 할 수 있는 latent space의 combination of elements를 찾아야 했습니다. StyleGAN 연구**[ref]** 에서는 control 가능한 layer나 element를 직접 찾아 보고하는 경우도 있으나 이는 상당한 노고가 들어갑니다. DPM은 이러한 parameter 변화에 의한 pose control 결과가 LPD**[ref]**에서도 동일하게 일어나기 위해서 latent vector를 어떻게 조작해야 하는지 mapping 시켜주는 역할을 수행합니다.

LPMnet을 training 시킨 이후, pre-trained된 LPMNet을 이용하여 DPM이 parameter space를 latent space로 transform 시킵니다. Image generation과 pose vector embedding을 위하여 pretrained neural talking head model LPD의 pose encoder와 generator를 사용합니다. Pose encoder는 주어진 이미지로부터 identity-agnostic pose vector를 embedding 하며 generator는 주어진 이미지의 pose를 따라하는 specific identity의 face를 generation하고 이 identity는 LPD의 finetuning stage에서 결정됩니다. DPM의 목적은 LPM에서 가능했던 parameter를 이용한 intuitive pose control을 generator에 부여하여 generated face image의 pose를 intuitively and directly control 하는것입니다.

우리는 landmark parameters를 이용하여 pose를 조작할 때 output 이미지의 identity specific features(e.g. hairstyle, eye colors, etc.)가 consistent 하기를 바랬습니다. Previous works~\cite{zakharov2019@fsal, burkov2020@lpd} discussed that using face landmark on neural talking head tasks may induce \textit{identity-bleeding} problem which output image follows person identity from pose source not from identity source. 우리 방식의 경우 face landmark를 input으로 활용하지 않으며 training pipeline에서 identity embedding은 fixed하여 사용하기에 any complex module or loss for identity-pose disentangle problem like cycle-consistent per-pixel editing loss~\cite{tewari2020@stylerig} to maintain the consistency of identity during controlling the pose. 없이도 DPM은 pose latent space에만 focusing 할 수 있습니다. 그리고 DPM의 pose controllability는 학습에 사용된 specific identity에만 limited 되지 않고 does not needed to be re-trained when attached to other generator which output different person identity.

Dataset

우리는 landmark parametric model 과 전체 학습 pipe line에 Voxceleb1 dataset[vox1참조]을 사용하였습니다. Voxceleb1 dataset은 celebrities의 interview 영상의 youtube 모음이며 video 내에 하나의 main face가 존재하는 dataset 입니다. 우리는 video의 모든 frames을 sampling 하고 each frame 에서 face landmark detector[landmark detector]가 동작하지 않는 frame들을 drop하여 총 4.2M images를 수집하였습니다. Face image를 샘플링 할 때에 face detector[s3fd참조]가 처음으로 caputuring 한 bounding box를 사용하여 face image를 crop하여 each sample의 face가 image의 center에 위치하도록 하였고,(centered-aligne?) 이는 LPM을 만들시 translation에 의한 coordinate variation으로 인해 PCA 결과에 translation parametric control이 포함되지 않도록 하기 위함 입니다. 이후의 pre-processing은 LPD[lpd참조]의 방식을 따랐습니다.

Model

LPM : LPM은 parametric model로 학습이 필요한 neural network를 포함하지 않습니다.

LPMNet : 우리는 LPMNet encoder의 구조로 scratch 상태의 MobileNetv2[mobilenetv2 참조] 사용하였습니다. LPMNet의 decoder는 encoder로 estimation된 parameter를 linear combination하는 function으로 학습이 필요한 neural network가 아닙니다.

We use only pose encoder and generator, except identity encoder and each of modules are fixed when we train DPM.We use finetuned generator which generate specific person. We test several generator which generate different identity for training DPM but there is no big difference in results.

At first, we consider that DPM learned poses of one specific person if we use one finetuned

DPM : DPM은 3개 layer의 MLP 구조를 사용한 비교적 가벼운 network이며 DPM을 학습할 때에 pre-trained된 LPMNet encoder와 LPD의 pose encoder, generator를 사용합니다. 또한, DPM을 학습 할 때에 DPM을 제외한 모든 network는 freeze하여 사용합니다. LPD의 generator의 경우 specific person을 생성하는 identity embedding이 finetune 되어 있는 generator를 사용하였습니다. LPD를 finetuning 시킬때는 meta train이 완료된 LPD 모델을 이용하여 생성하고자 하는 특정 identity의 few shot image로

finetuning을 진행하였습니다. 그렇기에 본 논문의 학습 pipe line에서 LPD의 identity encoder는 사용되지 않습니다. 처음에는 LPD를 finetune 할 시 network에 포함되어 있는 모든 batch normalization layer가 training data distribution의 statistics를 tracking 하지 않도록 하였습니다. 이는 finetuning으로 인한 batch normalization layer의 statistics 변화로 pose vector의 latent space가 shift 되는것을 방지하기 위함 입니다. 하지만 실험을 해 보았을 때 batch normalization의 statistics에 따른 변화가 DPM을 학습 시키는것에 큰 영향을 미치지 않았습니다.

하나의 finetuned generator를 사용하게 되는 경우 DPM이 학습할 때에 하나의 specific person의 pose image 만을 학습하기 때문에 학습 이후 DPM을 different identity를 생성하는 finetuned generator에 적용시킬 때에 parametric control이 잘 반영되지 않을것을 우려하였습니다. 하지만 하나의 generator만 학습에 사용하여도 parametric pose control을 하는데에 충분하다는 사실을 발견하였습니다. 이는 LPD pose encoder의 pose encoding이 facial expression의 detail을 잘 반영하기 때문이라 생각하며 다양한 pose variation을 포함하고 있는 millions of data를 사용하기 때문에 하나의 identity로 학습을 하여도 충분히 많은 pose variation에 대한 학습이 가능했기 때문이라 생각합니다.

Implementation Details

Pose vector의 dimension(d_v)은 256 이며 parameter의 개수 k는 5,10,20,40,136 으로 실험을 하였습니다.

Evaluation

LPM

LPMNet

DPM

A more classic approach is to model face/head pose in the 3D morphable model (3DMM) framework [1] or using a similar approach in 2D (e.g. an active appearance model) [6]. Still, learning 3DMM and fitting a learned 3DMM almost invariably involves detecting landmarks, thus inheriting many of the

landmark deficiencies. Alternatively, a dataset of 3D scans is required to build a model for pose/identity disentanglement in 3DMM framework.

적고싶은말

landmark parametric model은 dataset을 어떻게 구성하는가에 따라 control 할 수 있는 pose의 요소들을 달리 할 수 있다. (e.g. facial expression) landmark detection의 경우 3D scanning에 비해 큰 노고없이 수행 가능한 task임으로 dataset을 모으고 parametric modeling을 explicitly design 하는 것이 3DMM에 비해 쉽다.

또한, 본연구를 통해 Identity bleeding 이슈 없이 face landmark를 이용하여 pose만 targeting하여 control 할 수 있다는 것을 보였기에 이는 face landmark를 활용한 다른 연구 혹은 application에도 적용시킬 수 있는 bridge 역할을 해 줄 수 있다는 점에서 큰 잠재가치를 지니고 있다.