Unsupervised landmark detection for video generation

Sangho Lee, Ayoung Lee, Seoha Baek, Yoonseup Lee
Seoul National University, Seoul, Korea
sangho.lee@snu.ac.kr, ayoung1002@snu.ac.kr, bshfight@snu.ac.kr, navy10021@snu.ac.kr

Abstract

Novel idea to detect landmark from generating motion transfer video

1. Introduction

In these days, computer vision seeks to understand object structures that reflect the physical states and representation of objects. And images and videos are an endless source of data, but many of them is hard to be used because of lack of annotations to train. Manual annotations or designs of object structures (e.g., skeleton, semantic parts) are costly and barely available for most object categories, making the automatic representation learning of object structure an attractive solution to this challenge. So, I propose this subject for this project to solve such challenges with unsupervised methods.

1.1. Key idea

Previous works show limitation in detecting landmarks. Detecting landmarks from each frame shows highly correctness with labelled points. However when detecting from video, it sometimes detects semantically different landmarks or highly escaped from expected key points. So when generating videos with such key points, it fails to generate continuous clip. It comes from not existing regularization for the location of landmark in the model, so we limit the difference of location of each landmark to prevent sudden change of landmark. There is a key idea of regularizing sudden change of location, which is using optical flow compared from source and target frame.

2. Related works

Unsupervised landmark detection. An extensive techniques on detecting landmark with unsupervised learning are exist in the literature. Unsupervised learning of object landmarks by factorized spatial embeddings [1] proposed an unsupervised method that landmarks effectively learn the place where a convolutional neural network detects stable visual patterns, but this method have well-constrained problems not to point landmark at critical location. Unsupervised Learning of Object Landmarks through Conditional Image Generation [2] proposed a method for learning landmark detectors via conditional image translation for visual objects without supervision. However, the landmarks are hard to interpret in the image. For this reason, Unsupervised Discovery of Object Landmarks as Structural Representations [3] discovers landmark with a differential autoencoder structure for informative landmark detection. Expanding this method, Unsupervised Part-Based Disentangling of Object Shape and Appearance [4] suggests disentanglement between object shape and appearance and Unsupervised Human Pose Estimation through Transforming Shape Templates [5] proposed a method for the unsupervised estimation of 2D keypoints requiring only a simple template and an unannotated video of a single human performing actions.

Generative model based landmark detection. Recently, deep generative models based landmark detection have been studied for image animation and video retargeting [6–9]. In these process, Generative Adversarial Networks(GANs) [10] and Variational Auto-Encoders(VAEs) [11] have been used to derive pattern and generate videos. Animating Arbitrary Objects via Deep Motion Transfer [12] have shown Monkey-Net consisting of three networks that predicts sparse key points, motion modeling, motion reconstruction through the self-supervised learning. However, low result can be obtained if input image and frame size are large. To overcome this drawback, First Order motion model for Image Animation [13] extended Monkey-Net by the combination of keypoints and local affine transformation for modeling motion and improved the object appearance when large pose transformations occur.

Predict next frames with optical flow. To find relationships between several image frames, optical flow is a common technique [1]. Unsupervised Discovery of Parts, Structure, and Dynamics [14] recognizes object, predicts hierarchical correspondences, and learns dynamics. Optical flow between two frame images helps to model motion of objects using hierarchical model and dynamics. Unsupervised

Part Representation by Flow Capsules [15] take a similar approach but use a capsule network, which parses the relationship of layers for objects, parts, and relations. Capsule encoder makes flow capsules for single image in preparation that state-of-arts approaches are not able to describe low level part of image. In this process, optical flow takes a role for calculating flow field of images as a self-supervised methodology. It shows better performance in unsupervised segmentation evaluation tasks compared to PSD.

Learning landmarks from video. Learning object landmark from an image is widely studied [2]. However, labeling process into frame-by-frame keypoints is necessary for video prediction and takes time for preprocessing. Unsupervised Keypoint Learning for Guiding Class-Conditional Video Prediction [16] proposed an unsupervised approach to find keypoints of a dynamic object in an image to learn future frames. In terms of video, action conditioned keypoints sequence is generated based on initial keypoints and action. Also, conditional VAEs (CVAEs) [17] framework for learning future distribution and Long-Short Term Memory (LSTM) network [18] for the sequential data are used and shows good performance.

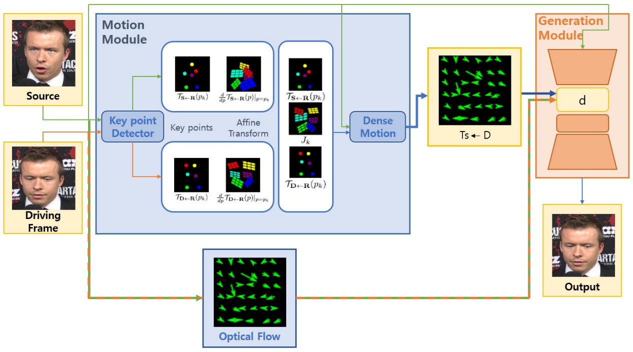
Source
Driving Frame
Motion Module
Key point Detector
Key points
Affine Transform
$\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p_k)$
$\frac{d}{dp}\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p)|_{p=p_k}$
$\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p_k)$
$\frac{d}{dp}\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p)|_{p=p_k}$
J_k
Dense Motion
Ts ← D
Generation Module
d
Output
Optical Flow

Figure 1. Learn representation from video generation

3. Approach

In this section, we will discuss how to learn motion presentation using optical flow. The model learns to predict how this part will proceed in the next step after detecting the motion of key points which is represented as transformation mapping, and this process proceeds without human annotation.The data input output and the entire structure are shown in Figure 1.

3.1. Overview

It is optical flow that shows changes over temporal movement in video data processing. Thus, in this paper, an optical flow that predicts the movement after given steps is generated, and the image is generated based on this . This model first finds the key point in each Source image S and driving frame D through the key point detector. The extracted key point is converted into an affine transformation mapping value to be converted into a point corresponding to each other in S and D. Dense motion network receives this conversion value indicating the movement of each key point, and The movement of each point is output as the movement of the area represented by the point. The motion value of the predicted region, that is, the optical flow, generates an image in region D through synthesis with the original image through the Generation model.

3.2. Network Structures

The structures of each network part is as follows.

The key point detector is the part that receives the images of S and D respectively, and calculates the key points of each image in a self-supervised manner. Key points obtained from each image of S and D are predicted individually through the encoder-decoder network. This key point representation acts as a bottleneck resulting in a compact motion representation. Calculated using key points, this transformation takes place through a reference frame R as one abstract concept before being converted from D to S. The reason why this referenceec frame R is necessary is that the images of S and D are visually very different because they belong to very different areas. Therefore, instead of calculating the conversion from D to S directly, the conversion proceeds through each R frame. This method makes it possible to process D and S separately.

Affine transform. The movement of each key point is newly modeled through local affine transform. Unlike using the aforementioned key point displacement, the reason why this transformation is necessary is that the affine transform calculates the movement of the point including the movement of the neighbor around the point. To this end, Taylor expansion is used to represent $T_{D\acute{\leftarrow}R}$ by a set of keypoint locatins and affine transformations. $T_{S\acute{\leftarrow}D}$ is estimated near the keypoint in D. In order for this, the transformation $T_{R\acute{\leftarrow}D}$ is estimated near the pixel location in the drivng frame D first.Then the transformation $T_{S\acute{\leftarrow}R}$ near the pixel loacation in the reference frame R. $T_{S\acute{\leftarrow}D}$ is obtained as follows :

$$T_{S\acute{\leftarrow}D} = T_{S\acute{\leftarrow}R} \circ T_{R\acute{\leftarrow}D} = T_{S\acute{\leftarrow}R} \circ T_{D\acute{\leftarrow}R-1} \tag{1}$$

After computing the first order Taylor expension of Eq.? the result is below :

$$T_{S\acute{\leftarrow}D} \simeq T_{S\acute{\leftarrow}R} + \acute{J}_k(z - T_{D\acute{\leftarrow}R}) \tag{2}$$

where

$$\acute{J}_k(z - T_{D\acute{\leftarrow}R}) = (\frac{d}{dp}T_{S\acute{\leftarrow}R})(\frac{d}{dp}T_{D\acute{\leftarrow}R})^{-1} \tag{3}$$

$T_{S\acute{\leftarrow}R}$ and $T_{D\acute{\leftarrow}R}$ are predicted by the keypoint detector. K heatmaps are estimated by U-Net architectur for each keypoint.The last layer of the decoder uses softmax activations in order to predict heeatmaps able to be interpreted as keypoint detection confidence map.

Dense motion. Dense motion network combines the local approximations to obtain the resulting dense motion field $T_{s\acute{\leftarrow}d}$. For each keypoint p_k Heatmaps H_k is computed that indicates to the dense motion network where each transformation happens. Each $H_k(z)$ is implemented as the difference of two heatmaps centered in $T_{D\acute{\leftarrow}R}$ and $T_{S\acute{\leftarrow}R}$. The heatmaps and the transformed images S^0, ... S^K are concatenated and processed by a U-Net.

Generation module. The generation module uses the transformed map and renders an image of the source object acting as the driving image. The generation network G warps the source image according to the predicted optical flow, which is the output of dense motion Ts←d. Since the source image S is not pixel-to-pixel aligned with the output image , feature warping method is used. Original feature map $\xi \in R^{\acute{H}\times\acute{W}}$ of dimension $\acute{H} \times \acute{W}$ is warped to $T_{s\acute{\leftarrow}d}$. The transformed feature map is written as :

$$\acute{\xi} = f_w(\xi, T_{S\acute{\leftarrow}D}) \tag{4}$$

where $f_w()$ denotes the back warping operations and ξ represents feature map.

3.3. Training Details

Loss function L consists of two separate components :

$$L = L_{opticalflow} + L_{perceptual} \tag{5}$$

$$\begin{aligned} L_{opticalflow} = \sum_{k=1}^{k} &|\Phi(H_k(u_t + \phi(u_{t+1}, u_t)), H_k(u_t)) \\ &- \Phi(H_k(u_{t+1}, H_k(u_t))| \end{aligned} \tag{6}$$

$$\begin{aligned} L_{perceptual} = \sum_{k=1}^{K} &|VGG(H_k(u_t + \phi(u_{t+1}, u_t)), H_k(u_t)) \\ &- VGG(H_k(u_t))| \end{aligned} \tag{7}$$

The first component is *optical flow loss*, which allows the model to increase the accuracy between the actual optical flow and predict optical flow.

The second component is *perceptual loss*, which encourages the model to accurately generate each parts in the frame at target step based on the predicted optical flow. TBD.

4. Experiments

In this section, we evaluate our model in various tasks including landmark detection and image pose reconstruction. In the first section 4.1. we show the qualitative results of our model for the task of unsupervised landmark detection on Deep-Fasion dataset. In Sect.4.2. we evaluate landmark detection and image reconstruction based on our baseline. Finally, We comparison with Previous Works and report the qualitative and quantitative results in Sect.4.3.

4.1. Landmark Detection on Deep-Fashion

In this section we show the results of part and landmark detection on Deep-Fashion dataset. In our work we only used images of full-body from the front-view. We randomly picked 20 percent of images as the test set. Fig.2. visualizes 10 out of 10 part activation maps of given images, spatially transformed images and appearance transformed images. Activation maps are learned in a self-supervised manner through an image reconstruction task. we can see important keypoints like face, hair, arms, legs, torso, wrists and feet which are detected from the resulting activation maps, even when there is a change in pose and appearance of the object. We can see that model is automatically learned to not overlap, as it leads to lower reconstruction loss and better reconstructed images.

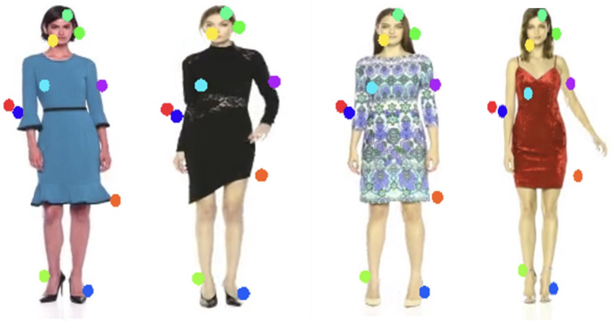

Figure 2. Landmark Detection. Visualization of 10 part activation maps for the given image

Our 10 part activation maps learned 2D Gaussian distributions each acting as a part representation and their corresponding predicted landmarks. We consider center points as part activation maps, which are parameters of Gaussian distributions as our predicted keypoints. Without any labels, our model can detect decent keypoints, especially for arm, leg, hand, feet regions.

4.2. Evaluation based on Generative model

In this section, we introduce a some metrics and present a in-depth evaluation result compared to three techniques tested on Vox dataset. We first evaluate our model landmark on the task of videos reconstruction. This consists in reconstructing the input video from a representation in which motion and content are decoupled. This task is a "proxy" task to image animation and it is only introduced for the purpose of quantitative comparison. In our case, we combine the extracted keypoints of each frame and the first frame of the video to re-generate the input video. Second, we evaluate our approach on the problem of Image-to-Video translation. This problem consists of generating a video from its first frame. Since our model is not directly designed for this task, we train a small recurrent neural network that predicts, from the keypoints coordinates in the first frame, the sequence of keypoints coordinates for the other 32 frames. Additional details can be found in the Supplementary Material A. Finally, we evaluate our model on image animation.

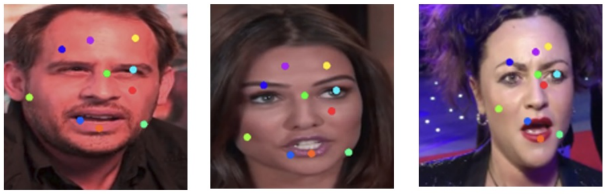

Figure 3. Landmark Detection. Visualization of 10 part Landmarks for the given image

Metrics. We adopt several metrics in order to provide an in-depth comparsion with other methods. We employ the following metrics and attach result of AKD on Vox dataset.

1) AKD. For the Vox dataset we use the facial landmark detector. We compute these keypoints for each frame of the ground truth and the generated videos. From these externally computed keypoints, we deduce the Average Keypoint Distance(AKD), i.e. the average distance between the detected keypoints of the ground truth and the generated video.

2) MKR. The facial-position estimator returns also a binary label for each keypoint indicating whether the keypoints were successfully detected. Therefore, we also report the Missing Keypoints Rate (MKR) that is the percentage of keypoints that are detected in the ground truth frame but not in the generated one. This metric evaluates the appearance quality of each video frame.

3) AED. We compute the feature-based metric employed in that consists in computing the Average Euclidean Distance (AED) between a feature representation of the ground truth and the generated video frames. The feature embedding is chosen such that the metric evaluates how well the identity is preserved.

4) FID. When dealing with Image-to-video translation, we complete our evaluation with the Frechet Inception Distance (FID) in order to evaluate the quality of individual frames.

Table 1. Image reconstruction comparisons on VOX dataset

Model	L_l	AKD	AED
X2Face	0.078	7.687	0.045
Monkeynet	0.049	1.878	0.199
FOMM	0.043	1.294	0.140
Ours	-	1.292	-

4.3. Comparsion With Previous Works

Video Reconstruction. We compare our results with the X2Face model, Monkey-net and FOMM that is closely related to our proposal on the Vox dataset.

5. Conclusion

We presented keypoint detection based on first order motion model. The motion field between two frames is represented by keypoints and local affine transformation and manipulated source image following target image's action is generated based on the motion field. In addition, we suggest optical flow method to solve an inconsistency problem with detecting keypoints. We showed that our approach outperforms state of the art methods in AKD.

References

[1] J. Thewlis, H. Bilen, and A. Vedaldi, "Unsupervised learning of object landmarks by factorized spatial embeddings," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5916–5925. 1

[2] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi, "Unsupervised learning of object landmarks through conditional image generation," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 4020–4031. 1, 2

[3] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee, "Unsupervised discovery of object landmarks as structural representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2694–2703. 1

[4] D. Lorenz, L. Bereska, T. Milbich, and B. Ommer, "Unsupervised part-based disentangling of object shape and appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 955–10 964. 1

[5] L. Schmidtke, A. Vlontzos, S. Ellershaw, A. Lukens, T. Arichi, and B. Kainz, "Unsupervised human pose estimation through transforming shape templates," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2484–2494. 1

[6] O. Wiles, A. Koepke, and A. Zisserman, “X2face: A network for controlling face generation using images, audio, and pose codes,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–686. 1

[7] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, “Recycle-gan: Unsupervised video retargeting,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 119–135. 1

[8] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, “Synthesizing images of humans in unseen poses,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8340–8348. 1

[9] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, “Video-to-video synthesis,” *arXiv preprint arXiv:1808.06601*, 2018. 1

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020. 1

[11] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013. 1

[12] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “Animating arbitrary objects via deep motion transfer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2377–2386. 1

[13] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “First order motion model for image animation,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 7137–7147, 2019. 1

[14] Z. Xu, Z. Liu, C. Sun, K. Murphy, W. T. Freeman, J. B. Tenenbaum, and J. Wu, “Unsupervised discovery of parts, structure, and dynamics,” *arXiv preprint arXiv:1903.05136*, 2019. 1

[15] S. Sabour, A. Tagliasacchi, S. Yazdani, G. Hinton, and D. J. Fleet, “Unsupervised part representation by flow capsules,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 9213–9223. 2

[16] Y. Kim, S. Nam, I. Cho, and S. J. Kim, “Unsupervised keypoint learning for guiding class-conditional video prediction,” *arXiv preprint arXiv:1910.02027*, 2019. 2

[17] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014. 2

[18] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 2