Generative Landmarks

D. Ferman and G. Bharaj

AI Foundation, USA

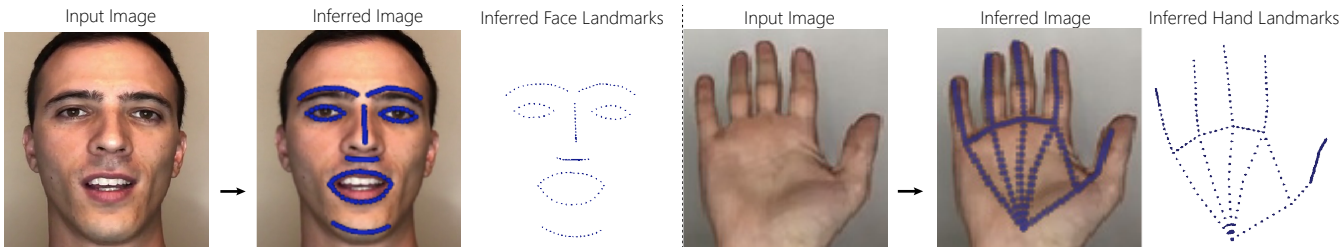


Figure 1: *(Left and right) Unmarked input image, inferred image with markers, and inferred template landmarks.*

Abstract

We propose a general purpose approach to detect landmarks with improved temporal consistency, and personalization. Most sparse landmark detection methods rely on laborious, manually labelled landmarks, where inconsistency in annotations over a temporal volume leads to sub-optimal landmark learning. Further, high-quality landmarks with personalization is often hard to achieve. We pose landmark detection as an image translation problem. We capture two sets of unpaired marked (with paint) and unmarked videos. We then use a generative adversarial network and cyclic consistency to predict deformations of landmark templates that simulate markers on unmarked images until these images are indistinguishable from ground-truth marked images. Our novel method does not rely on manually labelled priors, is temporally consistent, and image class agnostic – face, and hand landmarks detection examples are shown.

CCS Concepts

• ***Computing methodologies*** → ***Interest point and salient region detections; Tracking;***

1. Introduction

Sparse landmarks detection is an important problem for face detection applications [ALS*16], face tracking with landmarks alignment as a sub-task for 3D face model fitting [BBA*07, DBea21] or to guide video synthesis for faces [Wea20], other body parts [CGZE19], among several others. These tasks rely on high-quality and temporally consistent landmarks; however, off-the-shelf landmark detection methods suffer from inconsistencies due to ambiguity in manual landmark annotations as well as temporal imperfections of frame-to-frame labeling, as landmarks are difficult to define precisely, see [Dea18, WQYea18] for a discussion. As a result, landmark detection models suffer from temporal jitters, and sub-optimal personalization. Wu et al. [WQYea18] approach this problem with a focus on the boundary, taking advantage of well defined face boundary lines along which the landmarks reside. Dong et. al. [Dea18] note that frame-to-frame landmark detection should ideally resemble the presence of physical markers and present an approach that uses optical flow, and later a triangulation-based approach [DYea20] that exploits the temporal information inherent in video data.

With the goal of temporally consistent and personalized landmark detection, we propose a method that involves capture of two sets of unpaired videos for a given body region: one in which semantically (e.g. eyes, nose, fingers, etc.) meaningful lines are visibly marked, and the other, unmarked. We predict the landmark deformations for a template, for each unmarked image and render them, such that, it resembles the marked images. Following Zhu et al. [Zea17], we pose this problem as an unpaired image translation problem, with an image generator network that translates from marked-to-unmarked images, while our novel landmark deformation prediction network performs the reverse translation. Thus, our method is capable of learning a set of predefined landmarks in an unsupervised fashion, circumventing the need for laborious and imprecise manual annotations, while providing landmarks that are inherently personalized and temporally stable.

DOI: 10.2312/egp.20211029

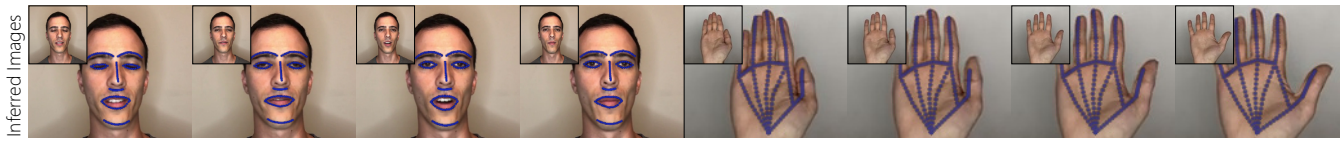


Figure 2: *(Left and right) Temporally consistent landmark predictions for faces and hands.*

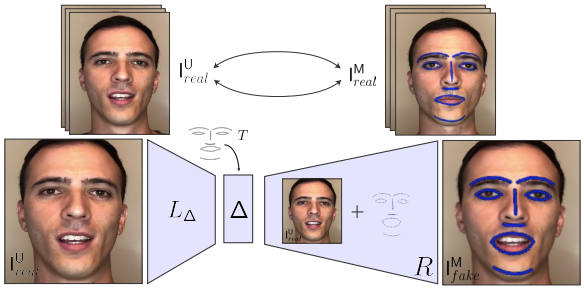
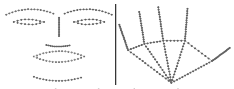


Figure 3: *Top: Cyclic consistency between I^U_{real} and I^M_{real}. Bottom: Unmarked image I^U_{real}, is fed into the encoder, L_Δ, that gives landmark deformations Δ. The deformations are applied to template T, and combined with I^U_{real} via our decoder – differentiable renderer R. The resultant image is treated as a fake marked image I^M_{fake}.*

2. Method: Generative Landmarks

Given two unpaired image sets – marked $\{I^M\}$ and unmarked $\{I^U\}$, our goal is to train a landmark deformation network, $L_\Delta : I^U \rightarrow I^M$, that *takes-in* unmarked images, $i^M_i \in \mathbb{R}^{3\times H\times W}$ and predicts landmark deformations, learning from the marked images in an unsupervised fashion. Rather than predicting these landmarks directly, we use a template $T \in \mathbb{R}^{N\times 2}$ with predefined spatial landmarks t_i, that form lines corresponding to the marked image set. In L_Δ, we intrinsically predict landmark deformations, $\Delta \in \mathbb{R}^{N\times 2}$, that are applied as offsets to the template T (below) and rendered onto the unmarked images.



Synthetically marked images resemble marked images while maintaining the spatial integrity of the template, where template landmarks on the unmarked images are intrinsically inferred. Our formulation is agnostic of the image class, and gives full control over the landmark definitions. Thus, we can even predict landmarks for body parts that previously lacked detailed training data, such as feet.

Landmark Deformation Network L_Δ. Similar to Zhu et. al [Zea17], our network L_Δ learns unmarked-to-marked images, where landmark deformation are used to simulate markers on images, Fig. 3. Cyclic consistency assures consistent learning. For landmark deformation prediction, we use a simple network that consists of 4 convolutional followed by 3 fully connected layers. During training, we employ an off-the-shelf generator network [Zea17], $G : I^M_{real} \rightarrow I^U_{fake}$ that translates marked images into fake unmarked images, and similarly discriminators D^M, D^U. When translating from domains $I^U \rightarrow I^M$, we first predict landmark template deformations, $\Delta = L_\Delta(I^U_{real})$ that are then rendered onto the input image via a differentiable renderer R, that gives us $I^M_{fake} = R(\Delta, I^U_{real})$.

Spring Potential Loss. In addition to CycleGAN loss above, we also employ a spring potential loss [NMea06], that helps maintain spatial consistency of the initial semantic template. As a result, landmarks maintain their spatial structure w.r.t template definitions. This is essential for recovering smooth landmarks while regularizing the GAN loss. Since, we want these separations to be consistent with the original template, we define this loss in terms of the change in deformations between neighbouring pairs of landmarks, $\mathcal{L}_{spring}(t_i) = K\sum_{t_j\in\{T\}} ||\Delta_{ij}||_2^2$, where K defines the spring constant, and Δ_{ij} the change in spring length for a neighbouring pair $\{t_i, t_j\}$ of template landmarks over which the spring loss is defined. Our full objective is given by:

$$\mathcal{L}(L_\Delta, G, D^U, D^M) = \mathcal{L}_{GAN}(L_\Delta, D^U, I^M_{real}, I^U_{fake}) + \mathcal{L}_{GAN}(G, D^M, I^U_{real}, I^M_{fake}) + \mathcal{L}_{cyc}(L_\Delta, G) + \mathcal{L}_{spring}(L_\Delta; T)$$

3. Results

The proposed method is capable of learning landmarks for various body regions, including faces and hands. Our training set consists of roughly 18k frames (about 10 minutes each) for each domain – marked and unmarked, scaled down to resolution 128×128. The landmark template is rendered as 2D points via PyTorch3D [RRN*20]. Training took about 2 hours (wall-clock) on an RTX Titan. Fig. 1 shows results for face and hand body regions.

4. Conclusion

This paper presents a novel method for performing landmark prediction on body regions. While manual annotations suffer inconsistencies due to ambiguities of precise landmark locations, our method uses ground truth-like data for learning landmarks as template-based deformations that match the visible ground truth information when rendered, although in this work we do not model occluded marker regions. In addition, our method is not limited by body regions or landmarks for which there are no available datasets.

References

[ALS*16] AMOS B., LUDWICZUK B., SATYANARAYANAN M., ET AL.: Openface. *CMU School of Computer Science* (2016). 1

[BBA*07] BICKEL B., BOTSCH M., ANGST R., MATUSIK W., EL. AL.: Multi-scale capture of facial geometry and motion. *ACM TOG* (2007). 1

[CGZE19] CHAN C., GINOSAR S., ZHOU T., EFROS A. A.: Everybody dance now. In *Proceedings of the IEEE ICCV* (2019). 1

[DBea21] DIB A., BHARAJ G., ET. AL.: Practical face reconstruction via differentiable ray tracing. *Computer Graphics Forum* (2021). 1

[Dea18] DONG X., ET. AL.: Sbr: An unsupervised approach to improve the precision of facial landmark detectors. In *IEEE CVPR* (2018). 1

[DYea20] DONG X., YANG Y., ET. AL.: Supervision by registration and triangulation for landmark detection. *IEEE TPAMI* (2020). 1

[NMea06] NEALEN A., MÜLLER M., ET. AL.: Physically based deformable models in computer graphics. In *CGF* (2006). 2

[RRN*20] RAVI N., REIZENSTEIN J., NOVOTNY D., GORDON T., LO W.-Y., ET. AL.: Accelerating 3d deep learning with pytorch3d. 2

[Wea20] WANG T.-C., ET. AL.: One-shot free-view neural talking-head synthesis for video conferencing. *preprint arXiv:2011.15126* (2020). 1

[WQYea18] WU W., QIAN C., YANG S., ET. AL.: Look at boundary: A boundary-aware face alignment algorithm. In *IEEE CVPR* (2018). 1

[Zea17] ZHU J.-Y., ET. AL.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE ICCV* (2017). 1, 2