Learning vital signs from face videos

[redacted]

Introduction and background

Camera-based health monitoring is an emerging field of research, mainly owing to the advantage of being a non-contact modality. Research has shown that many health parameters like heart rate, can be accurately measured from a human face video, by using just an off-the-shelf camera as a detector. This has opened a sea of opportunities since a camera can be easily accessible by anyone- whether it be a webcam while working on a laptop, or the front camera while surfing on the phone.

However, extracting this bio-signal from a camera is very challenging due to several factors, apart from the fact that this bio-signal is very weak in magnitude. One of the main factors is human motion - any movement of the person results in pixel intensity changes that completely dominates the desired biosignal, thus making it more challenging to extract heart rate.

Some of the prior methods model RGB channel pixel intensity measurements as the summation of coherent biosignal and additive motion artifacts which are uncorrelated to the biosignal. This is a a blind source separation problem, and ICA [1] or PCA is generally used to extract the biosignal. However, the performance is not satisfactory when the motion is large. A recent paper [2] uses Convolutional Neural Network with consecutive frames as input. It also uses an additional "Attentive Network" that assigns weights on each pixel on the face. The weights place importance on regions where the signal-to-noise ratio of the biosignal is high. Inspite of superior performance on public datasets, it still does not address the issue of motion artifacts, and therefore fails under heavy motion. The motion issue is somewhat addressed in [3], where the authors use Face Tracker coordinates to model motion distortions. However, we believe the motion distortion cancellation process can be improved using a deep net.

Project goal

For our project, we will use a Deep Convolution Network to extract a time-series biosignal from face video. We will use [2] as reference and improve on it by additionally simulating motion artifacts. We will use a separate architecture that simulates motion distortions from the video. The generated motion artifacts/noise will then be used in parallel to extract a distortion-free biosignal.

One primary idea is that we will use videos of mannequin as a training dataset for generating motion signals. A mannequin is devoid of any bio-signal. Under random movements of the mannequin, the intensity fluctuations of a pixel on the mannequin in the video is purely due to motion distortion (due to changes in surface BRDF). We will use a Convolutional Neural Network to learn the relation between natural movements and the resulting intensity change and use this relation to simulate motion distortions in actual human face video. Secondly, in some basic experiments that we tried with the mannequin, we observed that using the coordinates of a 3D Face Tracker is correlated with pixel intensity fluctuations. We want to further develop on this and learn the relation between the Tracker and the intensity changes. We plan to explore the usage of OpenPose[4] to track facial features in real-time. Moreover, incorporating details about illumination, geometry, and reflectance will be useful. [7] has some ideas about learning the same from a single snapshot, and we will try to investigate this and try to incorporate the same as well. We will distribute the workload evenly among our members.

Dataset and feasibility

As mentioned earlier, we plan to first render videos of mannequin used to learn motion signals. Motion artifacts can be generated, for example, via Blender animation (https://docs.blender.org/manual/en/latest/render/cycles/render_settings/motion_blur.html). In this case we also have control over different lighting and surface normal conditions, which could potentially allow our network to learn a good range of motion features.

We also need datasets to include the bio-signals. The publicly available datasets consist of high-resolution human face videos, which is a necessity in our case since extracting the weak bio-signal requires uncompressed videos. Example datasets include RGB Video I [5] and II [6]. The only challenge we have is the huge size of datasets. We will be using short clips (~ 5 seconds) to make the computation feasible.

[1] Mansouri, “Remote photoplethysmography with constrained ICA using periodicity and chrominance constraints”, 2018

[2] McDuff, “DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks”, 2018

[3] Wang, “Discriminative signatures for remote-ppg”, 2019

[4] Cao, Zhe, et al. "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields." *arXiv preprint arXiv:1812.08008* (2018).

[5] Estepp, Justin R., Ethan B. Blackford, and Christopher M. Meier. "Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography." 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2014.

[6] Chen, Weixuan, and Rosalind W. Picard. "Eliminating physiological information from facial videos." 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017.

[7] Tewari, Ayush, et al. "Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.