DDP Thesis
Topic: Heart rate Estimation Using Unsupervised learning

Dola Akhila Datta EE18B131

June 16, 2023

0.1 Introduction

Video-based remote physiological measurement utilizes face videos to measure the blood volume change signal, which is also called remote photoplethysmography (rPPG). Supervised methods for rPPG measurements achieve state-of-the-art performance. However, supervised rPPG methods require face videos and ground truth physiological signals for model training.The unsupervised rPPG mea- surement method does not require ground truth signals for training. We use a 3DCNN model to generate multiple rPPG signals from each video in different spatiotemporal locations and train the model with a contrastive loss where rPPG signals from the same video are pulled to- gether while those from different videos are pushed away. We test on five public datasets, including RGB videos and NIR videos. The results show that our method outperforms the previous unsupervised baseline and achieves accuracies very close to the current best supervised rPPG methods on many datasets.

0.2 Method

0.2.1 Prior Knowledge

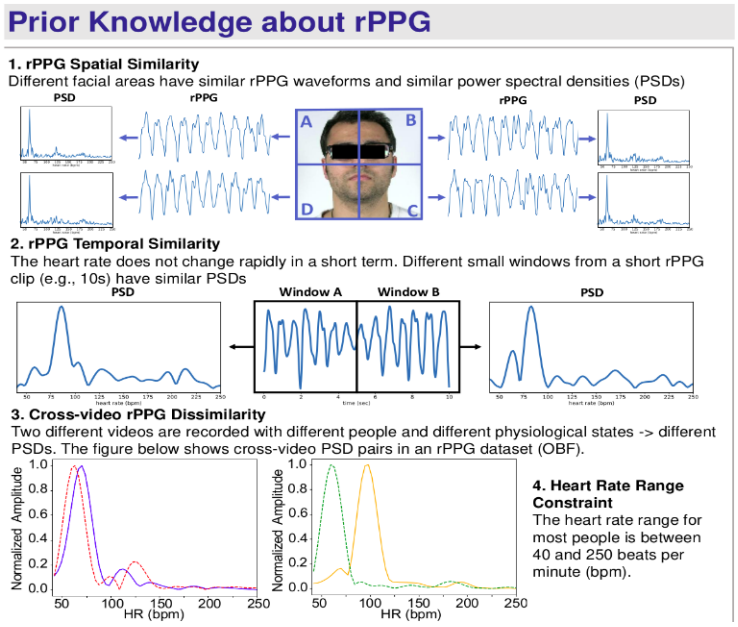

Figure 1: Prior Knowledge used

0.3 Method Used

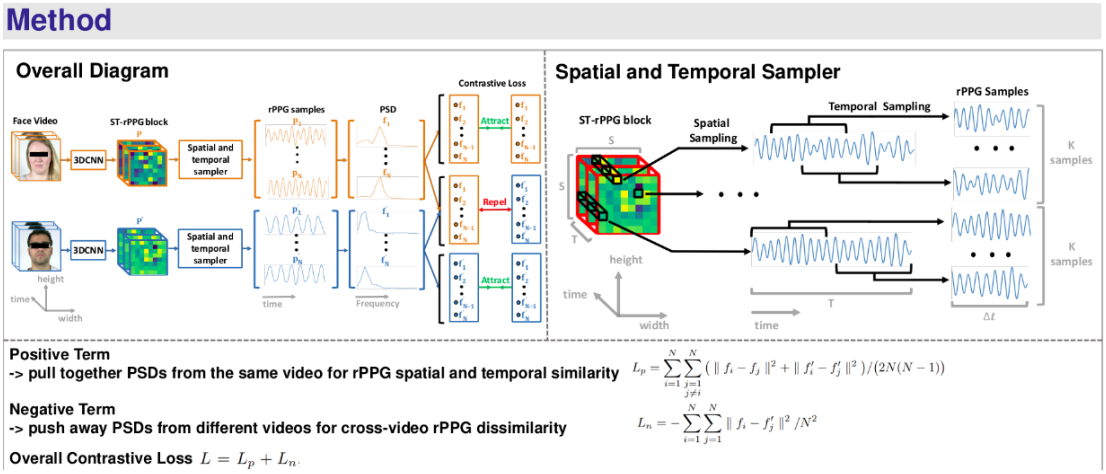

Figure 2: Method Used for training

A pair of videos are fed into the same 3DCNN to generate a pair of ST-rPPG blocks. Multiple rPPG samples are sampled from the ST-rPPG blocks (The spatiotemporal sampler is illustrated in Fig. 2) and converted to PSDs. The PSDs from the same video are attracted while the PSDs from different videos are repelled.

0.4 Preprocessing

The original videos are firstly preprocessed to crop the face to get the face video shown in Fig. 2 left. Facial landmarks are generated using FaceAlignmentNet. We first get the minimum and maximum horizontal and vertical coordinates of the landmarks to locate the central facial point for each frame. The bounding box size is 1.2 times the vertical coordinate range of landmarks from the first frame and is fixed for the following frames. After getting the central facial point of each frame and the size of the bounding box, we crop the face from each frame. The cropped faces are resized to 128 $\times$ 128, which are ready to be fed into our model.

0.5 Spatiotemporal rPPG (ST-rPPG) Block Representation:

We modify 3DCNN-based PhysNet [56] to get the ST-rPPG block representation. The modified model has an input RGB video with the shape of T $\times$ 128 $\times$ 128 $\times$ 3 where T is the number of frames. In the last stage of our model,

we use adaptive average pooling to downsample along spatial dimensions, which can control the output spatial dimension length. This modification allows our model to output a spatiotemporal rPPG block with the shape of T × S × S where S is spatial dimension length as shown in Fig. 5. More details about the 3DCNN model are described in the supplementary material. The ST-rPPG block is a collection of rPPG signals in spatiotemporal dimen- sions. We use T×S×S to denote the ST-rPPG block. Suppose we choose a spatial location (h, w) in the ST-rPPG block. In that case, the corresponding rPPG signal in this position is P (·, h, w) which is extracted from the receptive field of this spatial position in the original video. We can deduce that when the spatial dimension length S is small, each spatial position in the ST-rPPG block has a larger receptive field. The receptive field of each spatial position in the ST-rPPG block can cover part of the facial region, which means all spatial positions in the ST-rPPG block can include rPPG information.For one ST-rPPG block, we will loop over all spatial positions and sample K rPPG clips with a randomly chosen starting time t for each spatial position. Therefore, we can get $S \cdot S \cdot K$ rPPG clips from the ST-rPPG block.

0.6 Why Our Method Works?

Our four rPPG observations are constraints to make the model learn to keep rPPG and exclude noises since noises do not satisfy the observations. Noises that appear in a small local region such as periodical eye blinking are excluded since the noises violate rPPG spatial similarity. Noises such as head motions/-facial expressions that do not have a temporal constant frequency are excluded since they violate rPPG temporal similarity. Noises such as light flickering that exceed the heart rate range are also excluded due to the heart rate range constraint. Cross-video dissimilarity in the loss can make two videos' PSDs discriminative and show clear heart rate peaks since heart rate peaks are one of the discriminative clues between two videos' PSDs.

0.7 Training details

Tried Face alignment net for cropping and alignment of video - better model than openface model - less noise
Due to small size of dataset - validation and test datasets are seeming to not have the same distribution. So had to use Instance Normalization - normalize each batch in inference as well.
Later found out IPR - (Irrelevant Power Ratio) is a better way of checking generalization.. i.e to decide where to stop training.

0.8 Results on UBFC

Paper Results on UBFC - Did 5 fold cross CV - MAE: 1 RMSE: 3
But the paper heart rate evaluation metric seems to be inaccurate - which uses fft!. better is calculating peaks though it is computationally expensive
Improved Paper Results -(Unsupervised) Did 5 fold cross CV - MAE: 0.8 RMSE: 1.2 (After better eval metric and using Instance normalization)
Comparision with Supervised: The same model when trained using supervised labels gives MAE: 0.6 RMSE: 1 Hence, Unsupervised is not bad

0.9 Drawbacks and Experiments

Generated wave sometimes does not resemble a bvp. Solution would be to use GAN.
But Results worsened - probably sensitive training to hyperparameters.
Results - MAE: 1.5, RMSE: 2.5

Flipped videos horizontally lead to no improvement.

As unsupervised results are not bad I tried Fine tuning and using L = L sup + L unsupervised MAE: 0.6, RMSE: 1 - It was not better than supervised. Reason was implicit

0.10 DATASET COLLECTED

Collected 100 Videos with 16 subjects - 12 Male and 4 Females. Consists of Tasks 1 to 7. Mentioned in Figure 3

Task no.	illumination	distance	Task
T1	Normal	Far	Stable
T2	Normal	Far	Talking
T3	Normal	Near	Less Motion
T4	Normal	Far	Large motion
T5	Normal	Near	Stable
T6	Dim	Far	Stable
T7	Normal	Near	Excercise (made optional)

Figure 3: Tasks details

Results on Lab data:

- After Fine- tuning the pretrained model on UBFC with new dataset collected:

  MAE: 7.5, RMSE: 10.25 (Unsupervised)
- For task large motion(largest):

  MAE: 12.7, RMSE: 25
- For stable task (least):

  MAE:2.5, RMSE: 4.1
- Supervised training results:

  MAE: 5.5. RMSE: 9.3

0.11 Conclusion and Future Plans:

Unsupervised is slightly bad than supervised techniques. I collected 30 videos from shutterstock - There are lakhs of facial video clips. Possible that unsupervised method will be much better when trained on large data scraped from internet compared to limited data and supervised training. Also leads to better generalization.
Better tuning of GAN hyperparameters.