Using RGB Videos to Predict ECG

Jacob Ouyang **Kevin Fang** **Christopher Ho**
University of California, Irvine
(jnouyang,kpfang,christh9)@uci.edu

Abstract

Despite the many machine learning research projects using physiological measurements as an input to predict the risk of medical diseases, there has been limited progress in the actual prediction of an individual's physiological measurements. In addition to the prediction of diseases, physiological measurements can allow healthcare providers to quickly provide online diagnostics through patient portals. In this paper, we perform a comparative study of using several algorithms to predict the ECG (electrocardiogram) of an individual from an RGB video, as well as experimenting with a variety of data preprocessing techniques. We also attempt to predict the peaks of an ECG which can be used to measure an individual's heart rate.

1 Introduction

In this paper, we attempt to create a machine learning algorithm where an individual can simply record a video of themselves and receive an approximate ECG. Normally, the process of recording a patient's ECG typically requires electrodes to be placed on the body. The benefits of a virtualized ECG appointment could extend beyond providing patients with more information about their body. Between 2008 and 2017, 53 research publications were made using a variant of ECG data to identify or predict diseases in individuals [1]. According to a 2017 survey by the Office of the National Coordinator for Health Information Technology, 52% of individuals have access to online patient portals, which doctors could use to monitor patient ECGs for general health statuses. Improvements could also be made in emotion detection, an expanding field [2][3] with applications that expand beyond healthcare into almost every field, most notably marketing, psychology, and human-computer design.

An ECG maps the electric potential of the heart to a particular time. The resulting graph is periodic; each oscillation consists of smaller segments that signify activity in specific chambers in the heart. By tracking the time between successive segments, we can easily calculate the heart rate from an ECG.

The primary issue with creating machine learning models generated off of facial features is the sheer amount of noise in each sample. In our dataset with 22 individuals, each individual has many noticeable features that have little to no impact on the output such as facial hair, glasses, or ethnicity. The features that allow us to identify ECG are far more subtle, such as small head movements caused by the cardiovascular pulse or small flushes in the skin, and these already subtle features can be further affected by personal habits, small movements like swallowing or fidgeting, or the lighting of the video. Furthermore, within the healthcare industry, it is difficult to garner large datasets, as each hospital has a responsibility to provide their patients with privacy, preventing them from cooperating in creating large datasets for researchers.

With these limitations in mind, our approaches heavily involved transfer learning, a technique that has already shown success in medical imaging [4]. Using transfer learning, our model is able to learn features consistent in nearly all computer vision tasks, most notably edge detection. Our approaches include a multi-headed 2D convolutional neural network, a 2D convolutional neural network with an LSTM cell, and a 3D convolutional neural network. From these approaches, we create a set of baseline models with minimal preprocessing and state of the art algorithms to serve as a point of comparison for future models.

1.1 Previous Work

Past papers have mostly focused on the prediction of heart rate rather than the entire ECG graph. Špetlík et al. [5] proposed a two step process, the first CNN generating a scalar per image in a video clip, and another generating the heart rate from these scalars. However, the image-scalar reduction approach lacks transparency. The lack of ability to train image to scalar directly hinders researcher's abilities to understand what such dimensionality reduction is doing. Furthermore, the paper's wildly varying results between datasets with a correlation coefficient ranging from 0.29 to 0.98 show the model's inability to generalize.

Chen et al. [6] proposed DeepPhys, an end-to-end convolutional attention network made to detect motion between subsequent frames. However, this network only has two inputs at once, which leads to two shortcomings: the network lacks valuable sequential information, and is easily affected by larger, irrelevant movements.

EVM-CNN by Qiu et al. [7] proposes a multi-stage preprocessing technique that feeds into a vanilla convolutional neural network. The Fast Fourier Transform is first applied to the input video, which is then decomposed into frequencies through spatial decomposition, amplified to make color changes more prominent, and finally reconstructed back into video. While this preprocessing step may have shown results, as said in Chen et al. [6], multi-stage processes have recently been outperformed by end to end neural networks, perhaps due to neural networks having the ability to customize these processes directly to the training data.

These papers show that there is potential for deep learning approaches to predict physiological signals from raw video. Although this field is still in its infancy, its practicality in today's pandemic stricken society makes this a problem of utmost importance.

1.2 Relevance

In our approach, we decided to focus on directly predicting an ECG signal instead of focusing on heart rate. Asides from providing heart rate information, ECG signals can be analyzed to detect arrhythmia, blocked arteries, and previous heart attacks. A model that can predict such a signal can be further fine-tuned to detect any one of these diseases [8]. Furthermore, disease detection models built from physiological models can provide significantly more transparency in the model's decision making process.

2 Dataset

Our dataset provided by AICure includes 56 RGB videos from 22 participants who recorded their physiological data (ECG, skin conductance, and respiration force) at rest and after exercise. The data was recorded every 1/2000 seconds with 12 frames per second, with a total of 174,381 frames across all videos.

This particular framerate, combined with the given sampling rate, means that approximately 166 readings correspond with each frame. As a result, no single label schematic can capture the full detail of the ECG, so we constructed several schematics:

- Average all the readings corresponding to 1 frame
- Take the maximum of all the readings corresponding to 1 frame (and normalizing them)
- Take the closest reading to each frame (i.e. label for frame 1 is the first label in the set of labels corresponding to that frame)
- Record peaks in activity (label '1' for each peak and '0' everywhere else)

We found that peak detection preserved the most important part of the graph (the peaks represent the heartbeat), while reducing the amount of noise caused by the low frame rate. Other labeling strategies introduced fake peaks and trends caused by sampling from noisy and modulating sections of the graphs.

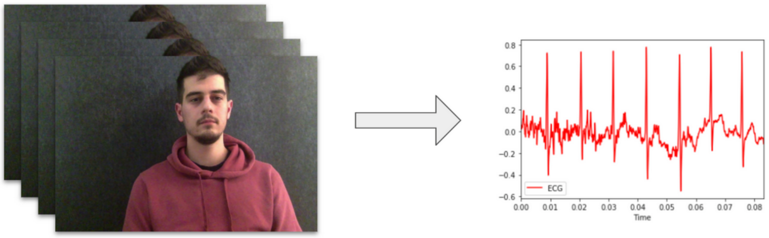

Figure 1: Flow of data, with ECG visualization.

We used DLib and OpenCV's Facial ROI models to extract the cheek from each image, as it was shown by Špetlík et al. [5] that their HR-CNN focused primarily on the cheek and lip area.

3 Methods

Because the ECG is a (somewhat) periodic signal, we deemed analysis of sequences of images to be appropriate instead of singular images. Both of our architectures used an image recognition model as a base to process each frame. We employed two different pre-trained networks: MobileNet, a fast yet accurate 2D CNN trained to classify images, and I3D, a 3D CNN trained to classify moving objects in videos.

As an initial model, we constructed a Siamese 2D convolutional neural network. Two consecutive frames are fed into the MobileNet network, and its output is concatenated and fed into a fully connected layer before outputting the predicted ECG value of the first frame. This architecture was an attempt to emulate HR-CNN [5]; we replaced their extractor network with MobileNet, and their predictor network with a simple FC layer.

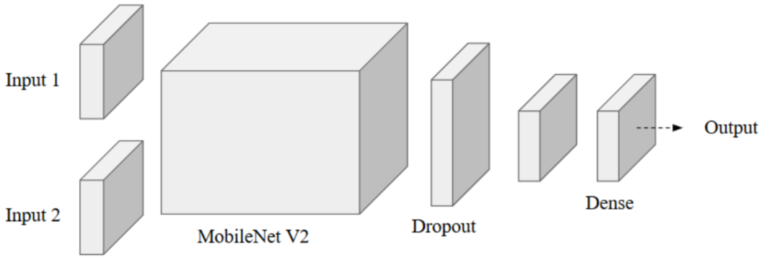

Figure 2: Initial Siamese 2D convolutional architecture.

After our initial model, we switched from 2 input frames to 12. Since our videos were 12 FPS, we thought it would be a good unit of measurement to use 1 second of video per sample. These models also output 12 numbers to predict the sample's entire ECG signal. Our models were trained for 50 epochs using MSE loss, and an Adam optimizer with a decaying learning rate starting at 1×10^{-4}.

Our highest-performing approaches included an LSTM 2D convolutional neural network transferred from MobileNet [9], as well as a 3D convolutional neural network transferred from the I3D Kinetics model [10]. I3D was used since it was trained to pick up small motion blurs. In some 3D CNN experiments, we left certain layers in the base models frozen, as according to Yosinski et.al [11], in experiments with relatively smaller datasets and large features, overfitting can often occur.

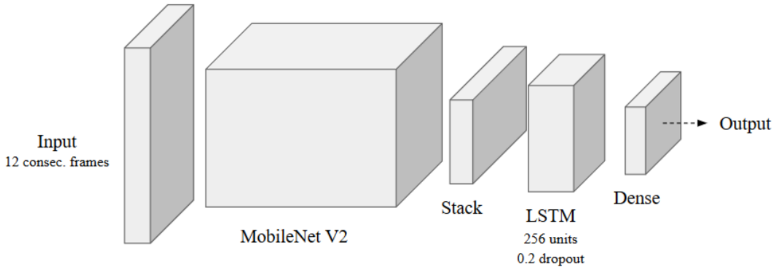

Figure 3: LSTM 2D convolutional architecture.

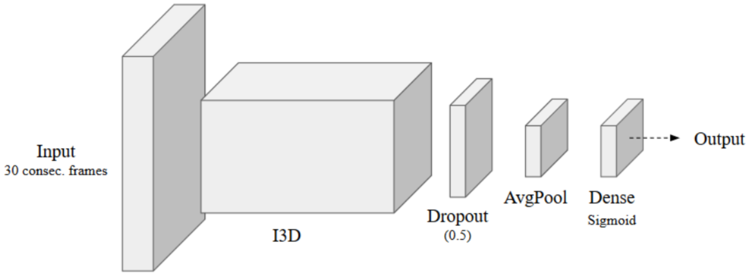

Figure 4: I3D convolutional architecture.

4 Results

All models were trained on all variations of the labeling scheme. We split the dataset into a 70/30 training/validation split.

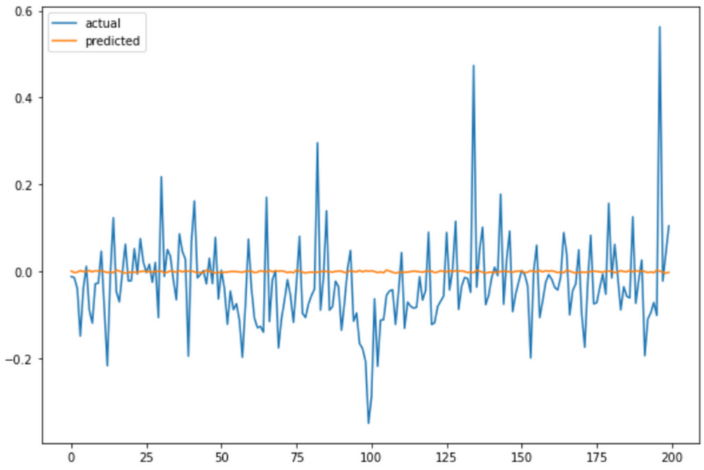

Figure 5: Siamese 2D network predictions (orange) and ECG ground truth (blue).

Regardless of the labeling scheme, the Siamese 2D CNN predicted zeroes with little variation. The model was unable to learn any trends in the data. For the peak detection scheme, non-peaks constituted >90% of labels. For the other labeling schemes, 0 was the approximate mean of the dataset. This model shows that the data is hard to pick up on with the current complexity of the model, and more information is needed.

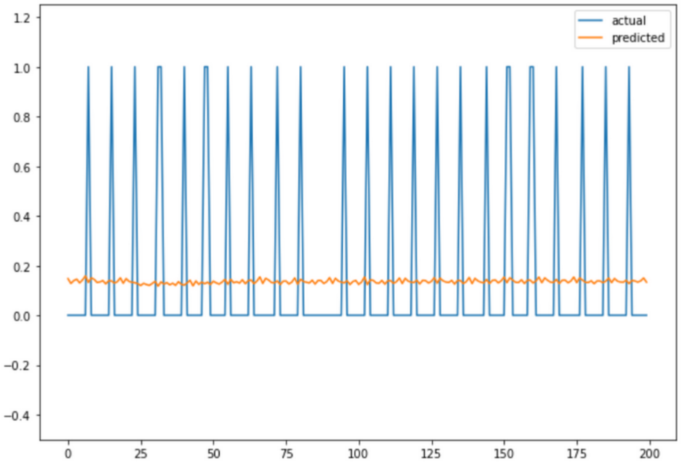

Figure 6: Siamese 2D network predictions (orange) and ECG peaks ground truth (blue).

The 2D CNN+LSTM predicted similar results to the baseline model for all datasets except peak detection. Instead of predicting strictly 0, it predicted the mean of the peak dataset. Although this model showed slight improvement, its lack of variance in prediction strategy shows that it was still unable to learn the dataset. Inferior results on our 2D CNN trials may also be attributed to a lack of customization in our feature vectors. Using a vanilla ImageNet CNN without any modifications to its layers weakened its ability to detect small motions between frames. Furthermore, the softmax layer in MobileNet's architecture may have reduced the feature vector's ability to exaggerate differences between subtle changes.

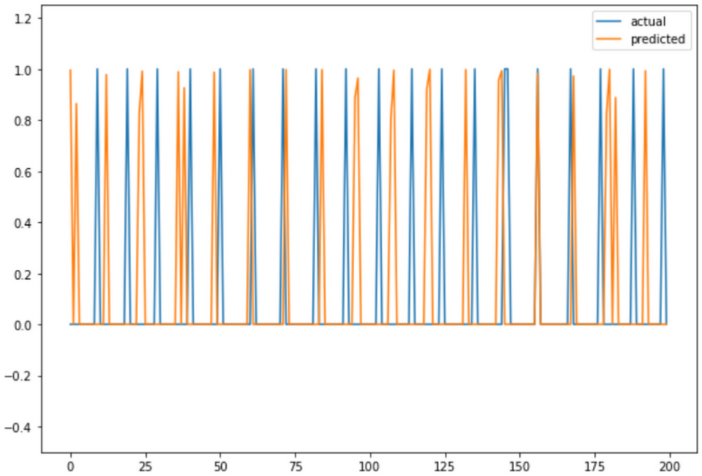

Figure 7: I3D network predictions (orange) and ECG peaks ground truth (blue).

The 3D CNN shows significant improvement from the past 2 models. Even though it is unable to completely follow the original ECG graph, it shows significant variation and a 97.09% reduction in loss from the LSTM model. Our trials using MSE loss had a substantial increase in performance compared to in our MAE loss trials. However, one drawback of this is that the peaks and valleys in the actual ECG are not predicted.

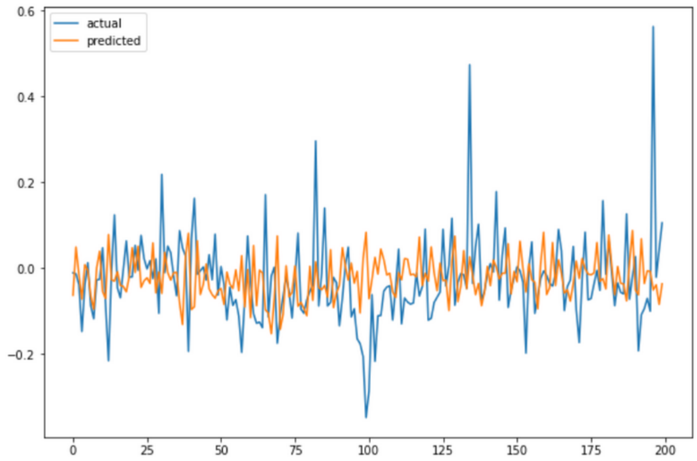

Figure 8: I3D network predictions (orange) and ECG ground truth (blue).

5 Conclusion

Our 3D CNN outperformed any of our other models by a considerable margin. As its original purpose was to analyze motion in video, it's understandable that it would be better at capturing and understanding the subtle changes in movement and color that our project demands.

Compared to our other models, I3D's 3D convolutions allow time-based comparisons between frames to occur before individual image analysis. Rather than comparing the feature vectors of each frame like our 2D CNN architectures did, I3D can find patterns between frames in its initial layers.

Even though our model is far from complete, we've shown that the ability for models to highlight motions between frames is extremely important. Whether that approach is taken through including custom features to exaggerate differences using Fourier transforms or fine-tuning feature vectors, detail magnification is imperative for a successful model.

As far as we can find, this is the first end-to-end neural network that attempts to trace the ECG waveform. While not entirely accurate, our network creates a rough estimation of the waveform's shape, which can easily be improved upon with future work.

6 Future Work

With our success using I3D, future models could use similar models such as a 3D residual attention network [6] in combination with our current I3D network. For preprocessing, proven techniques for heart rate such as Fourier transforms and temporal filtering could be applied.

Data augmentation is also pivotal for improving model performance. Due to the sensitivity of the model to face position and lighting, generic data augmentation techniques such as image flipping, random cropping, and Gaussian noise can corrupt a dataset instead of augmenting it. Generative adversarial networks that can artificially generate new faces and data from pre-existing data may significantly boost performance, and avoid some of the aforementioned privacy issues associated with real data.

Using a 3D residual attention network architecture could also improve results, combining the attention in DeepPhys [6] and our I3D transfer learning approach. While this network has not been used in the healthcare industry, it has shown extremely promising results in Res3ATN [12], a neural network made for hand gesture recognition.

7 Contributions

Jacob Ouyang: Procured dataset and initial reading material. Wrote skeleton code and initial versions of preprocessing/model training/data loading. Actively contributed to model architectures and preprocessing procedures. Gave class presentation. Wrote initial draft of results/conclusion and worked/proofread other parts of the paper.

Kevin Fang: Researched model architectures. Wrote the LSTM and I3D Model architectures. Contributed to model preprocessing procedures. Helped write class presentation slides. In charge of the introduction, past papers and abstract sections of the paper.

Christopher Ho: Researched model architectures. In charge of hyperparameter tuning and model training on GPU. Designed visuals and wrote class presentation slides. Wrote all visualization and modifying model structure to fit various label structures. Proofread, organized, and visualized results for paper and presentation.

References

[1] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, 161:1–13, jul 2018.

[2] Masih Aminbeidokhti, Marco Pedersoli, Patrick Cardinal, and Eric Granger. Emotion recognition with spatial attention and temporal softmax pooling. *Image Analysis and Recognition*, page 323–331, 2019.

[3] Armin Seyeditabari, Narges Tabari, Shafie Gholizadeh, and Wlodek Zadrozny. Emotion detection in text: Focusing on latent representation, 2019.

[4] Hariharan Ravishankar, Prasad Sudhakar, Rahul Venkataramani, Sheshadri Thiruvenkadam, Pavan Annangi, Narayanan Babu, and Vivek Vaidya. Understanding the mechanisms of deep transfer learning for medical images. *CoRR*, abs/1704.06040, 2017.

[5] Radim Spetlik, Jan Cech, Vojtěch Franc, and Jiri Matas. Visual heart rate estimation with convolutional neural network. 08 2018.

[6] Weixuan Chen and Daniel J. McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. *CoRR*, abs/1805.07888, 2018.

[7] Ying Qiu, Y.-H. Liu, Juan Sebastian Arteaga-Falconi, Haiwei Dong, and Abdulmotaleb El Saddik. Evm-cnn: Real-time contactless heart rate estimation from facial video. *IEEE Transactions on Multimedia*, 21:1778–1787, 2019.

[8] Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. Ecg heartbeat classification: A deep transferable representation. *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, Jun 2018.

[9] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[10] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017.

[11] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014.

[12] Naina Dhingra and Andreas Kunz. Res3atn - deep 3d residual attention network for hand gesture recognition in videos. *2019 International Conference on 3D Vision (3DV)*, Sep 2019.