CS 229 Project report: Extracting vital signs from video

D.Deriso, N. Banerjee, A. Fallou

December 7, 2013

Introduction

Cardiovascular health is the *sin qua non* of human life. Early detection of cardiovascular disease is of paramount importance in public health. This project aims to develop a method to visualize the perfusion of blood through the skin via pulse oximetry. Pulse oximetry is a technique that exploits the fact that oxygenated and deoxygenated hemoglobin changes the color of red blood cells. The technique maps these changes in rgb color of the visible skin to the invisible presence of oxygenated vs deoxygenated blood in the local vasculature underneath the skin.

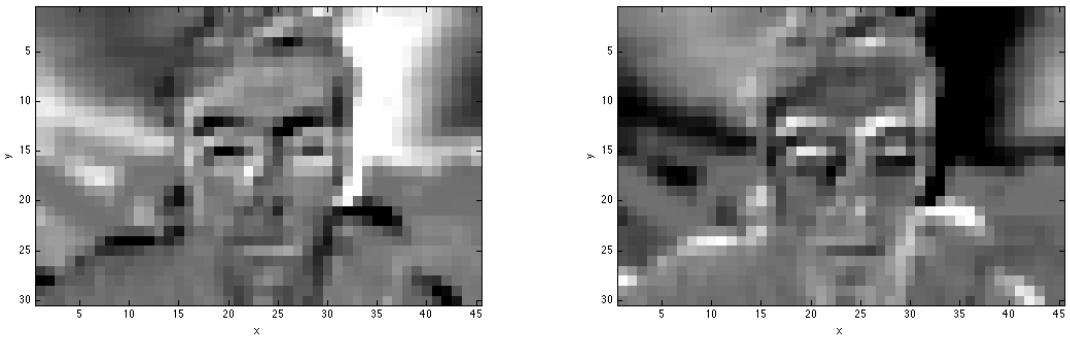
Previous studies have shown that video obtained from an ordinary webcam can be used to visualize perfusion by selectively amplifying temporal frequencies in video [1]. A study by the MIT CSAIL showed that this technique can also be used to infer heart rate from the person being taped. The present project aims to extend this work to detect the relative changes in oxygenated vs deoxygenated blood and reconstruct the pulse oximeter waveform from an ordinary webcam video.

1 Methods

Previous work on Eulerian Video Magnification (EVM), has led to the following processing pipeline, which we later modify for our purpose. The video can undergo different treatments, but the central goal is to amplify spatial or temporal changes that are normally invisible to the naked eye. The process can be summarized as:

- Separate the video into distinct spatial frequency bands by performing a 2D spatial fourier decomposition.
- For each spatial frequency band, blur and downsample several times using Gaussian or Laplacian Pyramids. The first step preserves spatial features (e.g. high frequencies such as edges) through this destructive process.

[1] http://people.csail.mit.edu/mrub/vidmag/



(a) At at a given time

(b) Half a heartbeat period later

Figure 1: Red channel of the EVM output, before recombining with the original video.

- Amplify a pre-selected temporal frequency band.
- Recombine spatial frequencies from step 1, and add the amplified video to the original video.

Figure 1 presents an example for the output of the next-to-last-step.

Strikingly, these two frames showed that the whole video undergoes a periodic color change with a frequency that seems equal to the heartbeat. This did not happen with the data provided in the paper, where in a similar video only the person's face changes color. To extract features relevant to pulse oximetery, we seek out a weighted combination of periodic changes in color space that best reconstruct our training pulse oximetery signal. We therefore undergo the following process:

- Extract each pixels time course from the video.
- For each pixel, separate the colors into R, G, B bands and perform a fourier decomposition where frequency is grouped into n bins.
- Learn weights via linear regression for frequencies within each color band to reproduce the simultaneously recorded pulse oximeter data in the training data set.
- Amplify the temporal frequencies according to the weights of the regression.
- Combine the amplified video with the original video.

2 Preprocessing

2(a) Video data

This new understanding of our problems naturally means that our core set of predictors are the three color channel intensity values through time $I_R(t), I_G(t), I_B(t)$, which are infinite-

dimensional feature vectors, and we want to do a regression of the reading of our pulse oximeter $Ox(t)$ on these intensity values. We then made two additional prior assumptions:

- We are only interested in periodic phenomena.
- These phenomena have a frequency lying in the 0-5Hz range.

The first assumption means we can take Fourier series decomposition of our intensity/oximetry values, the second that we can limit the decomposition to only a small number p of harmonics. We took $p = 10$ as a starting value:

$$I(t)/Ox(t) = \sum_{n=-p}^{p} c_n e^{i\left(\frac{2\pi nt}{T} + \phi_n\right)}$$

Thus, our training output y is a vector of size $2p$, with the coefficients of the Fourier decomposition. Our feature vector contains the $2p$ coefficients of the Fourier decomposition for each of the three color channels. Each pixel is considered a different training example, we have k of them for one video (typically $k = 1280 \times 720$). Thus our parameters matrix, $\theta \in \mathbb{R}^{(6p)\times k}$

2(b) Pulse Oximeter Data

Pulse oximeter data was collected using an Arduino and a Pulse Oximeter connected to a laptop. A program was written that started recording the pulse ox data and simultaneously started taking pictures every 50-80 ms on a webcam. This gave the pulse ox data corresponding to the image/video data.
In order to accurately train our model, we needed to extract from the pulse oximeter data the largest amplitude frequencies and then train these on the video. This required interpolating the pulse ox values over regular time intervals, using a high pass filter and taking into account phase.

3 Initial results

Our initial computations on the pixel-intensity data revealed, unsurprisingly, that pixel intensity variation over time for a reasonably still video was not large. Often a single color channel for a pixel would have a range of 3 (when it could vary from 0-255) over the whole video, even if that pixel was contained in the face. The DFT of the data would produce large peaks at 0 Hz and would rapidly diminish in magnitude. Because of this, we have not had any significant linear regression results. A simple solution is to subtract a mean intensity value of our pixel, so as to increase the relative intensity variations over time. Ultimately, we may have to resort to using EVM-enhanced videos if our training data has signals that are too weak to pick up.