Facial Boundary Recognition and Heart Rate Extraction From Remote Video Source

Vlad Slyusar

ECE 5831: Pattern Recognition & Neural Networks
University of Michigan - Dearborn, Fall 2016

December 16, 2016

0.1 Introduction

The goal of this paper centers on investigating techniques for heart rate extraction from an ordinary video source. While such biometric information is theoretically present within any uncovered region of the body that contains a pulse, the implementation focuses on attempting to extract the heartbeat specifically from a human face. A forward-facing head region provides less of a challenge in detection and tracking throughout the video than an arbitrary body part with exposed skin. The first frame of the video is used for initial face detection with user prompt for choice in the case of multiple targets. The selected face is tracked sequentially across all frames via its representative feature vectors. The tracking produces coordinates for a facial boundary for each individual frame. The identified boundaries are extracted, rotated upright, and scaled to the same size producing a cutout of the face adjusted for rotation and distance variation. In order to reduce noise from movement or breathing, the tracked facial boundary can also be optionally stabilized using a phase-based approach[WRDF13]. Next, the Eularian video magnification technique can be applied to amplify frequencies within a selected range corresponding to a valid heartbeat[WRS$^+$12]. The amplified video is processed to determine the color variations within linear progression of the frames and the most prominent frequency is identified as the candidate heart rate and adjusted from Hz to Beats/Minute. In the face of significant background noise, the process can be repeated with a user specified region that does not contain faces or exposed body parts. The results are used as a baseline in an attempt to remove the bias frequencies extracted from the face and isolate the heartbeat.

0.2 Problem Statement

While camera based heart rate detection systems already exist, the most popular implementation of the technology requires direct contact to skin, generally by placing a finger directly over the camera lens. Such an approach has become quite common with the relative abundance of personal devices encompassing both a camera and an LED within relative proximity. By applying a strong illumination to the skin pressed firmly against the camera, the color intensity fluctuations induced by blood circulation can be tracked to extrapolate an approximate rate of heartbeat[Mel13]. While most applications with such functionality strongly discourage using results for medical work, the simple and unobtrusive nature of the method provides a good approximation with little special equipment or complicated procedures. Since the heartbeat can be extracted in this way using only the input data from a camera, the approach should also apply to a video stream without direct skin contact with the lens. Skin contact is only required in the first place as a means of noise rejection by filling up entire camera view with the region of interest. If skin region can be identified and properly tracked throughout all frames, even a non-contact video will contain such biometric data. Furthermore, even if the boundary can be properly extrapolated from the video, the region of interest will encompass a fraction of the total pixel count of the lens and will contain a significant amount of noise from the environment. The challenge lies within amplifying the frequencies corresponding to the potential heartbeat signal and rejecting the frequencies introduced by noise within the video.

0.3 Related Work

While video-based heart rate detection systems are not yet widely implemented, the concept is not new. In November 2012, a collective team of MIT and Cambridge researchers demonstrated the capability to amplify sub-threshold frequencies within a video using a technique they called Eularian Video Magnification[WRS$^+$12]. The technique decomposes the video into the spacial domain and applies a temporal bandpass filter to the frames. The filter provides amplification of a specified range of frequencies within the spatio-temporal domain. Their work was used as one of the main foundations for the implementation in this paper. While the demonstrated method can extract imperceptible frequencies from a video, this project expands on the idea by first extracting only the face to be amplified and then comparing to the amplification of non-facial region within the same video. Furthermore, this paper focuses on amplifying a wide frequency range encompassing resting and active heart rate, in comparison to the relatively narrow-band filters specifically selected for each implementation by Wu et al.

In addition to amplifying color changes, a similar approach can be applied to movement amplification. In this case, the goal was to reduce small motion across the video frames. While the Eularian Video Magnification technique works in this case, a more efficient solution to the problem was proven using a Phase-Based approach. Another team at MIT Computer Science and Artificial Intelligence Lab introduced a method of achieving this through analysis of phase variations of the coefficients of a complex-valued steerable image pyramid over time, corresponding to motion[WRDF13]. Their work was used within an optional intermediary step of the overall process and provided a more stable and clean final output image. However, outside of the visual results, the extra step provided little improvement in heart rate prediction. Since this step incurs a significant computation cost, the option to skip such movement attenuation altogether is provided within the code.

0.4 Method

0.4.1 Processing Stage 1 - Facial Boundary Recognition and Tracking

This stage of processing takes in the raw video directly from recording device, and produces an output of just the face adjusted for rotation and distance from camera. In order to achieve this, the Viola-Jones Object Detection Framework is applied to the first frame[Wan14]. All potential faces are identified via detecting prominent features and if more than 1 face is identified the user is prompted to pick one of their choice. An example of the feature set and facial boundary is provided in figure 2. Once a face is selected, the Kanade-Lucas-Tomasi(KLT) algorithm is applied to track the face over the entire video based on the eigenvectors of the selected face within the first frame[Bir07]. This produces a boundary region of the face for each frame, the coordinates are stored in an array for generating a new video. The coordinates from KLT execution scale with head rotation and adapt to size differences that result from changes in distance to the camera. In order to compile a final output video, the image within the coordinate box in each frame must be cropped, rotated, and re-sized. The rotation relied on simple geometry to determine angle between the selected region and horizontal with respect to the image. Cropping, however required a more advanced approach when the image was rotated. One approach would rely on extracting all the pixels within the box and applying a transformation, but this would produce jagged edges without additional methods for compensation. Instead, the implemented method first computes a region for initial crop which perfectly encompasses the rotated region. Figure 3 illustrates the crop region in red, and facial boundary in yellow. The outer coordinates of the face ROI lie on exactly on the lines of the selected box for cropping. After performing the first crop, the new image is rotated, and finally cropped to eliminate the dead regions.