JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 12, pp. 1-7, December 2023
https://doi.org/10.9708/jksci.2023.28.12.001

Attention-Based Heart Rate Estimation using MobilenetV3

Yeo-Chan Yoon*

*Professor, Dept. of Artificial Intelligence, Jeju National University, Jeju, Korea

[Abstract]

The advent of deep learning technologies has led to the development of various medical applications, making healthcare services more convenient and effective. Among these applications, heart rate estimation is considered a vital method for assessing an individual's health. Traditional methods, such as photoplethysmography through smart watches, have been widely used but are invasive and require additional hardware. Recent advancements allow for contactless heart rate estimation through facial image analysis, providing a more hygienic and convenient approach. In this paper, we propose a lightweight methodology capable of accurately estimating heart rate in mobile environments, using a specialized 2-channel network structure based on 2D convolution. Our method considers both subtle facial movements and color changes resulting from blood flow and muscle contractions. The approach comprises two major components: an Encoder for analyzing image features and a regression layer for evaluating Blood Volume Pulse. By incorporating both features simultaneously our methodology delivers more accurate results even in computing environments with limited resources. The proposed approach is expected to offer a more efficient way to monitor heart rate without invasive technology, particularly well-suited for mobile devices.

▸ **Key words**: rPPG, Mobile AI, Heart rate Estimation, Healthcare, Video Analysis

[요 약]

딥러닝의 발전은 의료 분야에서도 다양한 응용을 가능하게 하고 있으며 이러한 애플리케이션 중에 심박수 측정은 개인의 건강을 관리하기 위한 필수적인 아이템이라 할 수 있다. 광혈류 측정을 이용한 기존 방법의 경우 스마트워치 같은 장비의 착용이 필수적이다. 그러나 최근 딥러닝 기술의 발전은 비침습식으로 원격에서 사용자의 얼굴 이미지를 분석하여 심박수를 높은 성능으로 측정가능하게 한다. 본 연구에서는 모바일 환경에서 사용 가능한 경량화된 심박수 추정 방법론을 제안한다. 이 방법론은 2D 컨볼루션에 기반한 특화된 2채널 네트워크 구조를 사용하여, 혈류와 근육 수축으로 인한 얼굴의 미세한 움직임과 색상 변화를 고려한다. 제안하는 네트워크 구조는 이미지 특성을 분석하는 인코더와 혈류량 파동을 예측하는 회귀 레이어로 구성되어있다. 이러한 복합적인 특성을 동시에 분석함으로써, 제한된 컴퓨팅 리소스를 가진 환경에서도 심박수를 정확하게 추정할 수 있다. 이 연구의 접근 방식은 침습적인 기술 없이도 심박수를 효과적으로 모니터링 할 수 있는 새로운 경로를 제공할 것으로 예상한다.

▸ **주제어**: rPPG, 모바일 AI, 심박수측정, 헬스케어, 비디오분석

• First Author: Yeo-Chan Yoon, Corresponding Author: Yeo-Chan Yoon
*Yeo-Chan Yoon (ycyoon@jejunu.ac.kr), Dept. of Artificial Intelligence, Jeju National University
• Received: 2023. 09. 26, Revised: 2023. 11. 23, Accepted: 2023. 11. 23.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 딥러닝 기술의 발전은 다양한 의료 응용 프로그램의 개발을 이끌어냈으며, 이로써 의료 서비스가 더욱 편리하고 효과적이게 되었다. 이러한 응용 프로그램 중 하나인 심박수 측정은 개인의 건강 상태를 측정하는 가장 필수적인 방법의 하나로 간주된다. 심박수가 100 bpm을 초과하는 경우 타키카르디아(tachycardia)로 불리며 심장부하를 나타내는 지표로 사용되며, 60 bpm 미만이면 브라디카르디아(bradycardia)로 불리며 감상선 기능 저하증, 약물 부작용, 심장 조직 증상 등 여러 건강 문제의 신호로 간주될 수 있다. 과거에는 스마트워치 등의 센서를 이용한 광혈용적 측정법으로 심박수를 측정했으나, 최근에는 딥러닝을 활용한 얼굴 이미지 분석을 통해 카메라만으로도 비접촉 심박수 측정이 가능하게 되었다. 딥러닝 기반의 얼굴 이미지 분석 방법을 사용하여 디지털 카메라만을 사용하여 비접촉 방식으로 심박수를 측정하면, 별도의 장비 없이 위생적이고 편리하게 심박수를 파악할 수 있다. 이는 침습적 기술 없이 환자의 심박수를 효율적으로 관찰할 수 있는 방법을 의료 분야에 제공하는 것으로 큰 의미를 가진다.

한편, 스마트폰으로 대표되는 모바일기기의 보급은 다양한 디지털 기술을 손쉽게 적용할 수 있는 창구가 되었다. 그러나 모바일 장치의 급속한 발전에도 불구하고, 제한된 연산 능력을 갖춘 환경에서의 높은 성능 요구는 여전히 도전적인 문제로 남아있다. 특히 딥러닝 방법의 경우 많은 연산량으로 인하여 높은 사양의 GPU 장치가 필요하다. 따라서 최근에는 모바일기기의 제한된 성능의 연산 장치에서도 딥러닝이 효과적으로 작동하도록 다양한 연구가 진행되고 있다. 모바일넷(MobileNet)[1] 은 경량화된 네트워크 방법으로 적은 연산량으로 높은 성능을 제공할 수 있어 모바일에 특화된 네트워크로, 모바일넷 이외에도 다양한 방법론이 등장하고 있다. 본 연구는 모바일 환경에서 심박수를 정확하게 측정할 수 있는 새롭고 경량화된 방법론을 제시한다. 제안하는 방법론은 2D 컨볼루션에 기반하여 움직임과 색상 변화를 동시에 고려하는 특화된 2채널 네트워크 구조를 사용한다. 심장이 수축하고 이완할 때마다 혈류량과 혈색소의 농도가 변화하게 되어 피부 아래에서 광학적 변화를 초래한다. 이러한 미세한 움직임과 색상 변화를 탐지하면, 심박수의 주기적인 변동을 정확하게 파악하는 데 도움이 된다. 또한 심장박동에 따라 피부에 미세한 떨림이 발생하는데, 떨림의 주기를 측정하면 심박수를 예측할 수 있게 된다. 이러한 원리에 근거하여, 본 연구는 딥러닝을 활용해 이 두 가지 정보를 동시에 분석하고 정확한 심박수를 측정하는 방안을 탐구한다. 본 연구에서 제시하는 방법론의 주요 장점은 제한된 연산 능력을 갖춘 환경에서도 높은 성능의 심박수 측정이 가능하다는 것이다. 본 논문이 제안하는 방법은 이미지의 특성을 분석하는 엔코더(Encoder)와 혈류량 파동(Blood Volume Pulse)을 예측하기 위한 회귀분석 레이어 두 단계로 이루어져 있다. 기존의 방식들과 비교하여 본 방법론은 근육의 미세한 움직임과 혈류량 변화로 인한 얼굴의 미세한 색상 변화 정보를 동시에 고려함으로써 더욱 정확한 결과를 얻을 수 있게 된다. 이를 위하여 이미지의 변화량을 분석하고 원시 이미지와 어텐션(Attention)을 수행하여 원시 이미지와 얼굴 모션의 변동량을 동시 고려할 수 있도록 한다.

본 논문은 다음과 같이 구성된다. 관련 연구에서는 딥러닝 기반의 비침습적 심박수 측정 모델의 기존 연구에 대하여 소개하고, 딥러닝 경량화를 위한 기존 방법에 대하여 분석한다. 3장에서는 제안하는 방법이 어떻게 구성되었는지 설명하며 4장에서 실험을 통하여 제안한 방법의 효용성을 분석한다. 5장에서는 제안한 방법에 대하여 정리하고 향후 연구 방향을 제안하도록 한다.

II. Related works

2.1 Deep Learning-Based Heart Rate Measurement Model

최근 몇 년 동안 원격 광학혈류계측법(rPPG, Remote Photoplethysmography)은 심박수, 호흡수, 혈압 등의 생리적 신호를 디지털 카메라를 사용하여 측정하기 위한 유망한 비침습적 기술로 부상하였다. rPPG 동영상에서 정확하고 견고한 생리 신호를 추출하기 위해 여러 딥러닝 기반 방법들이 제안되었다.

Liu[1]는 다중 작업 시간적 이동 합성곱 어텐션 네트워크 (MTTS-CAN)를 포함하는 rPPG 도구 상자를 제안했다. 이 도구를 사용하면 전문 장비 없이도 카메라로 얼굴을 촬영한 후에 이를 분석하여 심박수와 호흡 속도를 측정할 수 있다. MTTS-CAN은 멀티테스크 러닝을 통하여 심박수 측정과 동시에 산소포화도 분석을 측정하여 성능을 향상시켰다. MetaPhys[2]는 비접촉식 맥박 및 심박수 모니터링을 위한 개인화된 비디오 기반 심장 측정을 위한 새로운 메타 학습 접근법을 제안하였다. 이를 통해 라벨이 없는 소량의 샘플에서 개인화 또는 맞춤형 모델을 학습하여 개인별로 편향된 결과가 나타나는 것을 방지할 수 있도록 하였다. Yu[3]는 트랜스포머 기반의 심박수 측정 방법

인 PhysFormer를 제안하였다. 제안한 방법을 통하여 시간 축에 따라 여러 프레임 간의 어텐션 메커니즘을 분석하여 효과적으로 심박수를 측정할 수 있었다. Lee[4]는 학습 집합과 다른 테스트 셋에서의 분포의 변화에 대응하기 위하여 레이블이 없는 데이터를 활용한 메타학습 방법을 제안하였다. 이를 이용하여 실제 배포 단계에서 분포가 다른 데이터에 손쉽게 대응할 수 있도록 하였다. 기존의 방법들은 높은 성능으로 심박수 측정이 가능하였으나, 연산량이 많아 모바일 장치에 적합하지 않다. 따라서 본 논문에서는 모바일 장치에 적합한 경량화된 모델을 제안하여 온디바이스(On-Device) 환경에서 동작할 수 있도록 한다.

2.2 Mobile-Based Deep Learning Research

딥러닝은 일반적으로 높은 연산량을 처리하기 위하여 고가의 GPU 장치를 요구한다. 그러나 대다수의 모바일 장치는 이러한 고도의 연산 처리에 제한이 있으며 GPU 장치를 탑재하고 있지 않다. 이러한 제약 때문에 낮은 컴퓨팅 파워의 장치에서 효과적으로 작동할 수 있는 경량 딥러닝 네트워크 개발에 관한 연구가 활발히 진행되고 있다. MobileNets [5, 6, 7]는 깊이별로 이어지는 컨볼루션을 사용하여 지역 속성을 능숙하게 포착하는 방법을 사용하여 네트워크를 경량화하였다. ShuffleNet [8, 9]은 포인트별 컨볼루션을 간소화하기 위해 그룹 컨볼루션과 채널 셔플을 사용하여 연산량을 획기적으로 줄였다. MicroNet [10]은 노드 연결을 줄여 네트워크의 넓이를 확장함으로써 최소한의 FLOPs를 관리하는 미세 팩터화된 컨볼루션을 제시한다. 최근 어텐션 기반의 트랜스포머가 높은 성능을 보임에 따라 트랜스포머를 경량화하는 시도도 이어지고 있다. Mobile-Former[11]는 MobileNet과 트랜스포머 디자인을 중간에 양방향 다리로 병합한다. 이 구조는 MobileNet의 지역 특징 처리와 트랜스포머의 전역 상호작용의 장점을 활용한다. EfficientFormer[12]는 모바일 환경에서 최고의 성능과 최소 지연을 위해 트랜스포머를 미세 조정하는 모델의 집합을 제시한다. 특히, EfficientFormer-L1은 iPhone 12에서 MobileNetV2의 속도와 동일하게 작동하며 ImageNet-1K에서 79.2%의 정확도를 달성하였다. Sinha[13]은 MobileNet v1과 비교하여 향상된 정확도와 축소된 지표를 목표로 하는 세 가지 수정된 MobileNet 아키텍처를 제안한다. 이 논문에서 제안한 아키텍쳐 중 메모리 제약이 있는 MCU(Microcontroller Unit)에 배포하기 위해 맞춤 제작된 가장 컴팩트한 모델은 Thin MobileNet이며, 크기가 9.9 MB로 매우 작은 편이다. 이 모델의 개선 사항으로는 ReLU 대신 Drop Activation의 채택, 드롭아웃 대신 랜덤 지우기 정규화 전략의 통합, 그리고 깊이별 분리 컨볼루션 대신 분리 컨볼루션의 채택이 있다.

본 논문에서는 2채널의 입력에 대하여 컨볼루션 연산을 적용한 후에 경량화 네트워크로 대표적인 모바일넷v3를 적용하여 모바일 장치에서 효율적으로 심박수를 측정할 수 있는 네트워크를 제안한다.

III. The Proposed Scheme

모바일 장치는 빠른 연산 속도와 높은 성능을 보이며 다양한 응용 분야에서 활용되고 있다. 이러한 기술적 발전에도 불구하고, 연산 능력이 제한된 환경에서 높은 성능을 요구하는 애플리케이션들은 여전히 도전 과제로 남아있다.

본 연구에서는 이러한 제한적인 환경을 극복하고 심박수를 정확하게 측정할 수 있는 새로운 방법론을 제안한다. 심박수 측정의 기본 원리는 광적맥혈류량 측정에 근거한다. 이 방법은 피부 아래 혈류의 미세한 색상 변화를 감지하여 심박수 변화를 파악하는 원리로 작동한다. 여기서 주요한 데이터 소스로는 연속적인 이미지 프레임, 즉 동영상이 활용된다. 동영상 데이터는 시간에 따른 피부의 색상 변화를 3D 이미지로 나타낼 수 있다. 그러나, 3D 컨볼루션은 그 연산량이 크기 때문에 모바일 장치와 같은 연산 능력이 제한된 환경에서는 적합하지 않다. 이 문제를 해결하기 위해, 2D 컨볼루션에 중점을 둔 접근 방식을 채택하되, 움직임과 색상 변화를 동시에 고려할 수 있도록 특화된 2채널 네트워크 구조를 설계하였다. 그림 1은 제안하는 방법의 시스템 구조도를 보여준다. 제안하는 방법은 엔코더와 회귀분석 모듈로 구성되어 있으며, Liu[2]가 제안한 모양 엔코더(Appearance Encoder)와 모션 엔코더(Motion Encoder)를 활용하여 엔코더의 출력값에 어텐션을 적용한 후에 모바일넷 V3를 이용하여 혈류량 펄스를 예측할 수 있도록 구성된다. 입력값으로 원시 이미지 시계열과 움직임 정보를 활용하는데, 움직임 정보는 연속된 이미지 프레임 간의 차이를 계산함으로써 얻을 수 있다. 이 차이 정보는 움직임과 관련된 주요 특징들을 잡아낼 수 있게 해준다. 실제로 움직임 정보는 맥박에 의한 미세한 근육의 움직임을 인식하는 데 큰 도움을 제공한다. 두 프레임 간의 차이를 분석하는 방법은 간단히 n+t 시점 후의 프레임에서 t 시점의 프레임의 차이 값을 이용한다. 본 논문에서는 실험적인 방법을 통해서 n을 3으로 설정하여 3프레임 차이가 나는 이미지 둘을 뺀 값을 입력으로 주었다.

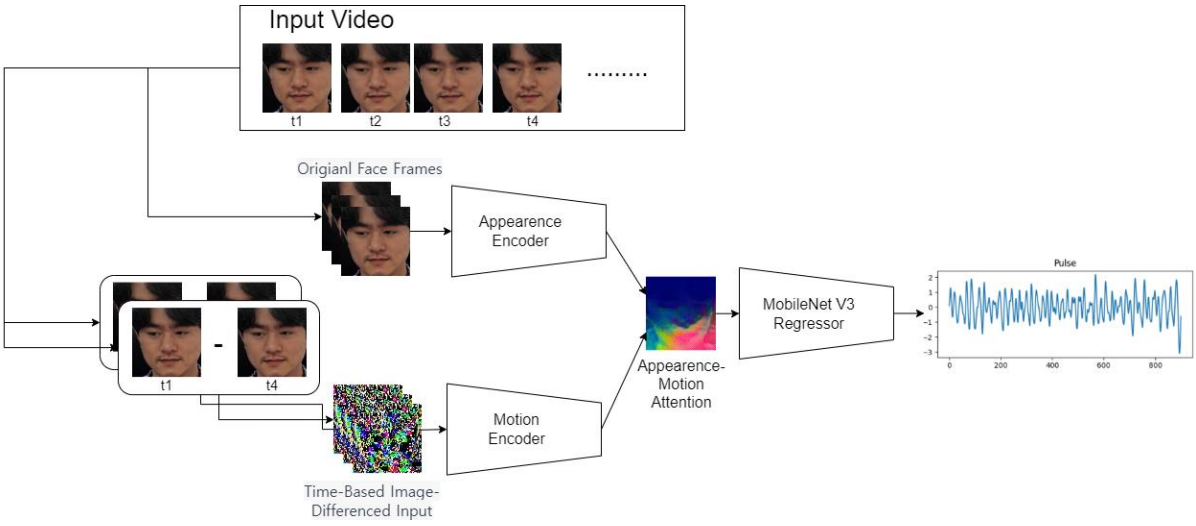

Fig. 1. System Architecture

이렇게 하여 간단히 움직임 정보를 입력으로 사용하였다. 2채널 네트워크는 움직임 정보와 색 변화를 독립적으로 처리할 수 있도록 설계되었다. 하나의 채널은 움직임 정보를 중점적으로 분석하며, 다른 하나는 색상 변화를 중점적으로 분석한다. 이 두 채널의 컨볼루션 결과는 행렬곱을 통해 통합된다. 이때, 어텐션 메커니즘을 도입하여 맥박에 의한 움직임과 혈류량 변화의 상관성을 분석하며, 실제 심박수 측정에 중요한 부분에 집중하도록 한다.

Table 1. The Layer Manifest of Motion Encoder

Layer	Output Shape	Param #
Conv2d	(32,178,70)	896
Dropout	(32,178,70)	0
Conv2d	(32,176,68)	18,496
tanh	(32,176,68)	0
Conv2d	(32,176,68)	36,928
tanh	(32,176,68)	0

테이블 1은 Motion Encoder의 레이어 구조를 보여준다. 이미지 시계열 분석을 위해서는 일반적으로 3D 컨볼루션을 사용하는데, 이 경우 2D 컨볼루션 보다 계산량이 매우 커 속도가 느려지는 문제가 발생한다. 본 논문의 목표가 상대적으로 연산속도가 느린 모바일 장치를 대상으로 하기에 이는 치명적인 문제로 작용한다. 따라서 본 논문에서는 3D 컨볼루션을 사용하는 대신, 비디오를 시퀀스를 하나의 이미지처럼 처리하되, 배치로 나눠서 처리하는 방법을 사용하였다. B가 batch size, T를 frame의 수, C를 채널수, W, H를 각각 width, height라고 할 때 비디오 입력은 아래와 같이 처리된다.

Table 2. Tensor Shapes Results of each Process.

Process	Result Shape
Input	(B,T,C,W,H)
Reshape the Input for Motion Encoder	(B×T,C,W,H)
Output of the Motion Encoder	(B×T,32,176,68)
Reshape the output for Attention	(B,T,32,176,68)

입력으로 주어진 5차원의 입력 텐서는 2D CNN의 입력으로 처리하기 위하여 4차원의 텐서로 변형된다. 이후 어텐션에서 시계열 특성을 반영하기 위하여 다시 5차원으로 변경된다. 이러한 방법을 사용하여 연산량을 줄여 속도를 높일 수 있었다.

테이블 3은 Appearance Encoder의 레이어 구조를 보여준다. 혈류량 변화를 인식하기 위한 Appearance Encoder는 Motion Encoder에 비하여 더 복잡한 네트워크를 구성하는 것이 성능에 도움이 되었다. Appearance Encoder 역시 속도 개선을 위하여 2D Convolution을 사용하였다.

본 논문에서는 회기분석기로 모바일넷v3를 사용한다. 모바일넷v3는 적은 수의 파라미터에서 효율적으로 작동하도록 제안된 네트워크이다. 이전 버전인 모바일넷v1, v2에서 사용된 인셉션 모듈을 개선하고, 중간 채널수를 감소시키며, 각 채널의 중요성을 고려하여 설계되었다. 얻어진 어텐션 메커니즘 결과는 이 모바일넷 v3의 입력으로 제공되어, 최종적으로 심박수를 예측하게 된다.

Table 3. The Layer Manifest of Appearance Encoder

Layer	Output Shape	Param #
Conv2d	(32,178,70)	896
tanh	(32,178,70)	0
Conv2d	(32,176,68)	1,056
sigmoid	(32,176,68)	0
Conv2d	(64,176,68)	18,496
tanh	(64,176,68)	0
Conv2d	(64,176,68)	36,928
tanh	(64,176,68)	0
Conv2d	(1,176,68)	65
sigmoid	(1,176,68)	0

IV. Experimental Results

본 논문에서는 세 개의 벤치마크 데이터셋, UBFC2[29], COHFACE[30], PURE[31]를 이용해서 rPPG 기반의 심박수 측정에 대한 실험을 수행하였다.

4.1 Dataset and Performance Metrics

UBFC2 데이터셋은 640x480 해상도와 초당 30 프레임의 비디오를 웹캠으로 캡처하였다. PPG 신호는 62Hz의 주파수로 투과성 펄스 산소계를 통해 얻어진다. 촬영세팅 문제로 피촬영자가 정면보다 낮은 각도를 바라보고 있으며, 영상에 따라 머리 이동 영상이 있는 것이 특징이다. 동일한 장소에서 촬영하였기에 조명과 배경의 차이가 없다. COHFACE는 40명의 개인의 얼굴 비디오 클립을 포함하며, 640x480의 해상도와 초당 20 프레임의 프레임 레이트로 캡처되었다. 다양한 장소와 조명환경에서 촬영되어 일반 PPG 신호 역시 256Hz의 주파수로 동시에 획득되었다. 참가자들은 녹화 동안 카메라 앞에서 앉아 있고 움직이지 않도록 지시 받았다. PURE 데이터셋은 10명의 참가자(남성 8명, 여성 2명)로 구성되며, 각 주제별로 6가지 다른 설정으로 총 60개의 비디오 시퀀스가 있다. 각 주제는 총 6분 동안 녹화되었다. 이 데이터셋은 고정, 말하기, 느린 및 빠른 변환, 작은 회전, 중간 회전 등 여섯가지 다른 설정을 반영하였다. 세가지 데이터셋 모두 비디오 영상의 길이는 1분 내외이다. 각 데이터셋별로 학습 및 평가 데이터의 개수는 테이블 3과 같다. 학습셋과 테스트셋의 숫자는 데이터셋 배포 내용을 따랐다.

Table 4. Number of data of each method using the UBFC2, COHFACE, and PURE datasets.

Method	TRAIN	TEST
UBFC2	323	84
COHFACE	487	323
PURE	514	133

모델의 성능을 평가하기 위해 회귀분석에서 일반적으로 많이 활용되는 평가 지표로, 피어슨 상관 계수 (r), 제곱근 평균 제곱 오차 (RMSE), 오차 표준편차 (STD) 등이 포함된다. 평균제곱 오차는 정답값과 예측값의 차이를 계산하여 정확성을 확인하고 오차 표준편차는 표준편차를 이용하여 정답의 평균과 예측치의 분포 차이를 분석하여 성능을 나타낸다. 평균제곱오차와 오차표준편차는 0에 가까울수록 좋은 성능을 나타낸다.

Table 5. Performance of each method using the UBFC2, COHFACE, and PURE datasets.

Method	Dataset	STD	RMSE	r
Physformer	UBFC2	**0.6984**	**0.6988**	**0.6974**
	COHFACE	0.932	0.9324	0.3904
	PURE	0.8627	0.8628	0.5195
Resnet18	UBFC2	0.9814	0.9818	0.3843
	COHFACE	0.9977	0.998	0.1329
	PURE	1.038	1.041	0.3168
TSCAN	UBFC2	0.7499	0.7502	0.6252
	COHFACE	0.9552	0.9555	0.3044
	PURE	0.8373	0.8377	0.5532
Proposed Method	UBFC2	0.7363	0.7402	0.6489
	COHFACE	**0.8494**	**0.8545**	**0.5242**
	PURE	**0.8015**	**0.8048**	**0.6008**

피어슨 상관계수는 이미지 시퀀스에 대하여 결과가 일관성 있게 나타나는지를 확인하기 위하여 자주 사용되는 지표로서, 오차의 절대적 값 대신 예측의 일관성을 분석한다. 정답의 심박수가 낮은 구간에서 낮은 예측값을 보이고, 높은 구간에서는 높은 예측값을 보이는지를 확인하는 방식으로 일관된 결과를 나타내는지 분석한다. 피어슨 상관계수는 1에 가까울수록 좋은 값을 나타낸다. 본 논문에서 이들 지표를 사용하여 성능을 평가하도록 한다.

4.2 Performance Evaluation

테이블 4는 제안하는 방법과 기존 방법의 성능을 측정한 결과이다. 비교 모델로 Physformer[3], TSCAN[1], Resnet 18을 사용하였다. Physformer와 TSCAN 모두 attention을 사용한 방법으로 심박수 측정 관련하여 높은 성능을 보여준다. Physformer는 transformer 기반의 네트워크이며 TSCAN은 본 논문과 같이 CNN과 attention을 같이 사용한 모델이다. 다만 본 논문과 달리 모바일 장치를 대상으로 하지 않아 높은 연산량을 보이며 파라미터 크기도 훨씬 크다. Resnet18은 가장 일반적인 CNN 모델로 베이스라인으로서 최소의 성능치를 확인하기 위하여 비교하였다. 테이블 4는 제안하는 모바일넷v3와 어텐션 엔코더를 사용한 모델의 성능을 보여준다. 제안하는 방법

은 다른 방법들을 이용하여 혈류량 펄스를 예측한 방법보다 COHFACE, PURE 셋에서 높은 성능을 보여줬다. 제안하는 방법이 기존 방법보다 파라미터 수도 적고 연산량도 적은데도 높은 성능을 보인 것은 제안한 기존의 방법이 데이터셋의 규모에 비해 상대적으로 더 복잡했기 때문으로 보인다. 기존 두 개의 엔코더와 어텐션 모듈로 중요한 특징들과 관계를 충분히 표현하였고 상대적으로 간단한 모바일넷v3을 회귀분석기로 사용한 것이 데이터 규모에 적합한 결과를 도출한 것으로 보인다. Resnet18이 낮은 성능을 보인 것은 어텐션 연산이 적용되지 않아 시계열 데이터의 관계성 분석에 실패한 것이 원인으로 보인다. 데이터셋 중 UBFC2의 경우만 Physformer가 가장 높은 성능을 보여주었는데, 데이터셋의 특성상 정면부 얼굴이 아니어서 참고 가능한 자질이 다소 제한적이고, 머리 이동이 있는 영상이 포함되어 트랜스포머를 사용한 깊은 어텐션을 적용한 방법이 효과를 보인 것으로 보인다. 다만 제안하는 방법과 큰 차이를 보이지는 않았다. 테이블 5는 기존 방법과 제안된 방법의 파라미터 크기 및 연산량(FLOP)을 보여준다. 파라미터 크기가 클수록 심박수 측정 모델을 저장할 장치에 더 많은 용량이 필요하다. 따라서 모델 파라미터 수를 줄이는 것이 저장용량에 한계가 있는 모바일 장치에 탑재하기 위해선 필수적이다. 제안하는 방법의 경우 파라미터 크기가 0.24 메가바이트로 다른 모델에 비하여 월등히 작아 모바일 장치의 저장용량 한계를 극복하기 수월하다. 반면 Resnet18과 Physformer의 경우 파라미터 크기가 각각 7.38 메가바이트, 11.69 메가바이트로 실제 저장 장치에 저장되는 모델 출력물의 경우 부가 정보 때문에 파라미터 크기의 50~100배 정도의 저장용량을 필요한 것을 고려하면 모바일 장치 저장공간에 큰 부담을 줄 수 있다. 제안하는 모델은 연산량 또한 다른 방법에 비해 크게 개선되었다. 제안하는 방법은 TSCAN에 비하여 약 42% 정도 속도가 계산되어 연산장치의 능력이 제한적인 모바일 장치에서 효과적으로 사용될 수 있음을 확인하였다.

Table 6. Parameter size and FLOPs

Method	Param Size (M)	FLOPs (G)
Physformer	7.38	35.32
Resnet18	11.69	41.75
TSCAN	2.23	40.89
Proposed Method	0.24	28.77

V. Conclusion

본 논문에서는 모바일 장치에 적합한 경량화된 심박수 측정 방법을 제안하였다. 본 논문에서는 심박수 측정 모델을 엔코더와 회귀분석기 두 가지로 나누고, 회귀분석기를 위하여 MobileNetV3를 사용하여 연산량과 파라미터 크기를 효과적으로 줄였다. 엔코더의 경우 모션 엔코더와 모양 엔코더간의 어텐션을 적용하여 주어진 이미지 시계열을 잠재벡터로 변환하여 사용하여 성능을 높일 수 있었다. 향후 연구로, 어텐션을 효과적으로 개선하여 얼굴 움직임이 큰 영상에서도 심박수 측정이 효과적으로 작동하도록 네트워크 구조를 개선할 계획이다.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. RS-2023-00245316)

REFERENCES

[1] Liu, Xin, et al. "Multi-task temporal shift attention networks for on-device contactless vitals measurement." Advances in Neural Information Processing Systems 33 (2020): 19400-19411.

[2] Liu, Xin, et al. "MetaPhys: few-shot adaptation for non-contact physiological measurement." Proceedings of the conference on health, inference, and learning. 2021.

[3] Yu, Zitong, et al. "Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection." IEEE Signal Processing Letters 28 (2021): 1290-1294.

[4] Lee, Eugene, Evan Chen, and Chen-Yi Lee. "Meta-rppg: Remote heart rate estimation using a transductive meta-learner." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. Springer International Publishing, 2020.

[5] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019

[6] 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. arXiv preprint arXiv:1704.04861, 2017

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4510–4520, 2018.

[8] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In The European Conference on Computer Vision (ECCV), September 2018.

[9] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018

[10] Yunsheng Li, Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Lu Yuan, Zicheng Liu, Lei Zhang, and Nuno Vasconcelos. Micronet: Improving image recognition with extremely low flops. In International Conference on Computer Vision, 2021

[11] Chen, Yinpeng, et al. "Mobile-former: Bridging mobilenet and transformer." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[12] Li, Yanyu, et al. "Efficientformer: Vision transformers at mobilenet speed." Advances in Neural Information Processing Systems 35 (2022): 12934-12949.

[13] D. Sinha and M. El-Sharkawy, "Thin MobileNet: An Enhanced MobileNet Architecture," 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2019, pp. 0280-0285, doi: 10.1109/UEMCON47517.2019.8993089.

[14] Sabour, Rita Meziati, et al. "Ubfc-phys: A multimodal database for psychophysiological studies of social stress." IEEE Transactions on Affective Computing (2021).

[15] Tsou, Yun-Yun, et al. "Siamese-rPPG network: Remote photoplethysmography signal estimation from face videos." Proceedings of the 35th annual ACM symposium on applied computing. 2020.

[16] Stricker, Ronny, Steffen Müller, and Horst-Michael Gross. "Non-contact video-based pulse rate measurement on a mobile service robot." The 23rd IEEE International Symposium on Robot and Human Interactive Communication. IEEE, 2014.

Authors

Yeo-Chan Yoon received BS, MS and Ph.D degrees in computer science and engineering from Korea University, Seoul, Rep. of Korea, in 2004, 2007 and 2020 respectively. Currently, he is an assistant professor in the Department of Artificial Intelligence at Jeju National University, Jeju-si, Rep. of Korea. His research interests include Deep Learning, Vision, Natural Language Processing and machine learning.