Video-Based Face and Facial Landmark Tracking for Neonatal Vital Sign Monitoring

E. Grooby[1,2,3], C. Sitaula[1], S. Ahani[2,3], L. Holsti[2], A. Malhotra[4], G. A. Dumont[2,3], and F. Marzbanrad [1]

[1] Department of Electrical and Computer Systems Engineering, Monash University, Melbourne, VIC, Australia
[2] BC Children's Hospital Research Institute, Vancouver, BC, Canada
[3] Department of Electrical and Computer Engineering, University British Columbia, Vancouver, BC, Canada
[4] Monash Newborn, Monash Children's Hospital and Department of Paediatrics, Monash University, Melbourne, VIC, Australia

Abstract— **This paper explores automated face and facial landmark tracking of neonates for the purposes of vital sign estimation. Utilising a publicly available dataset of neonates in the clinical environment, 25 videos were annotated. Face and facial landmarks (*i.e.* eyes and nose) tracking are then assessed. Additionally, the identification and tracking of the neonate's forehead and cheeks are purposed, as they are ideal regions of interest for vital sign estimation. Tracking of the face produced an average overlap score of 93.0%. Tracking of the eye and nose landmarks produced mean normalised errors of 0.026 and 0.019 respectively. The cheek region of interest could be effectively identified and tracked, whereas the forehead region of interest identification was incorrect 16% of the time.**

Keywords— **vital sign estimation, neonates, face tracking, video monitoring**

I. Introduction

Video monitoring provides a non-contact method for cardiorespiratory monitoring. By providing heart and breathing rate vital signs, it is potentially suitable for assisting clinicians in clinical, home and remote environments [1].

Video-based neonatal vital sign monitoring can be divided into four stages; (1) region of interest (ROI) detection, (2) ROI tracking, (3) photoplethysmogram extraction, and (4) vital sign estimation. In our previous work [2], we developed deep learning-based methods for neonatal face and facial landmark (*i.e.* eyes, nose and mouth) detection to cover step (1). This paper builds upon our previous work for the second stage, ROI tracking.

ROI tracking is an essential step in video-based vital sign monitoring as patient movement, temporary clinical intervention and facial occlusions (*i.e.* from bottle feeding and sleep position) are commonplace in the clinical environment [3]. Past works have either implemented no tracking [4], repeated ROI detection [5], Kanade–Lucas–Tomasi (KLT) feature tracker [3, 6–8], or kernelized correlation filter (KCF) [9–11]. Typically, ROI tracking is a more time-efficient process compared to ROI detection which can be time-consuming and prevent real-time processing [2].

For vital sign estimation, suitable ROIs have included the entire face [3, 5, 11], the forehead [4, 8, 11, 12] and the cheeks [12]. The forehead and cheek ROIs typically only include skin, making them ideal for vital sign estimation. However, because there are no readily identifiable landmarks, tracking of these ROIs has either been non-existent or implemented with poor success [3,4,11].

This paper evaluates face ROI and facial landmark tracking. Utilising the tracked facial landmarks, selection and tracking of the forehead and cheeks ROIs is then proposed and assessed.

II. Methods

A. Data Acquisition

The newborn baby heart rate estimation database (NBHR) [10] is used for this study. NBHR [10] consists of 1,130 videos of 257 neonates at 0-6 days old, totalling 9.6 hours of recordings with synchronous photoplethysmogram and heart rate physiological signals. The newborn infants were recruited from the Department of Obstetrics and Gynaecology, Xinzhou People's Hospital, China. Biological sex was approximately equal (Male 48.6%: Female 51.4%) [10]. Camera angle relative to each baby's location varied, facial occlusion was present in a subset of videos from bottle feeding and sleep position, and a variety of natural hospital room illuminations were obtained [10].

As this is a preliminary study, a subset of 25 videos was randomly selected for manual annotation. The first and final frames of each video were annotated with a rectangular face bounding box and 3 facial landmark positions, namely; the right eye, left eye, and nose (see Figure 1 red annotations).

The face bounding box was defined as the area from the forehead to the chin and between the ears. In cases of partial occlusions, if subsections of the face were still visible/partially

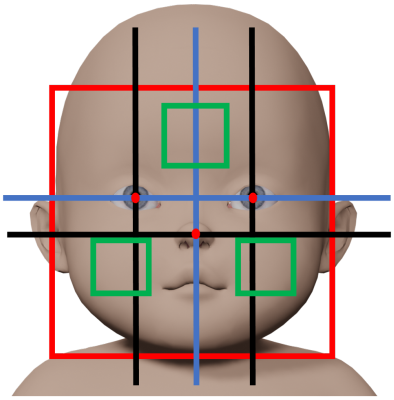

Fig. 1: Annotated Facial ROIs and Landmarks. Red = initial face ROI and associated eye and nose landmarks. Blue = inter-ocular and nose lines. Black = parallel inter-ocular and nose lines. Green = derived forehead and cheek ROIs

visible in the area of occlusion, they were annotated as the face. Otherwise, if a complete region of the face was occluded, it was not included in the bounding box as the size of the face occluded was difficult to infer.

B. Face and Facial Landmark Tracking

Face and facial landmark tracking were implemented together using the KLT algorithm [6, 7].

With reference face ROI inputted as the first video frame, features for tracking are extracted using the minimum eigenvalue algorithm developed by Shi and Tomasi [6]. Additionally, initial rectangular landmark bounding boxes are defined with the landmark position as centerpoint and the size of the bounding box being 20% size of the initial face bounding box. The subset of facial minimum eigenvalue features within these landmark bounding boxes is then identified. These features are then tracked between successive frames using intensity gradient information as proposed by Lucas and Kanade [7].

For overall landmark position tracking, the median change in the position of the features identified in the initial landmark bounding box is used to update the position of the landmark.

C. Forehead and Cheek ROI Identification

The forehead and cheek rectangular ROIs are derived from the relative positions of the facial landmarks and the size of the initial face ROI.

Using the facial landmarks, two lines are derived (see Figure 1 blue annotations). The inter-ocular line is defined as the straight line intersecting both eye landmarks. The nose line is defined as the line that originates at the nose landmark and intersects the inter-ocular line at 90 degrees.

The centre of the forehead is identified as the point along the nose line x pixels past the inter-ocular/nose line intersection, where x is 22.5% of the length of the face ROI. The size of the forehead ROI is 22.5% of the face ROI (see Figure 1 green annotations).

For identifying the cheeks, three additional lines are defined (see Figure 1 black annotations). Namely, two parallel to the nose line that intersects with the eye landmarks, and one parallel to the inter-ocular line that intersects with the nose landmark. The centre of the cheeks is identified as the intersection of these three lines, offset by 10% and 5% of the length of the face ROI parallel to the nose and inter-ocular line respectively. The size of the cheek ROIs is 20% of the face ROI (see Figure 1 green annotations).

D. Forehead and Cheek ROI Tracking

Two methods are proposed for tracking the forehead and cheek ROIs. The first method utilises relative face movements in the face ROI tracking to modify the forehead and cheek ROIs. The second method uses the tracked eye and nose landmark positions to recalculate the forehead and cheek ROIs.

E. Cheek ROI Selection

In a subset of videos, the neonate is sleeping on their side. This means that one side of the face is more readily visible and suitable for cheek ROI identification and tracking. Therefore, based on the first frame, the proximity of the left and eye landmarks to the face bounding box is calculated. The landmark furthest away from the bounding box is the side of the face more available, and the associated cheek is then selected.

F. Evaluation

Results are compared to a baseline if the ROI/landmarks remained stationary. Additionally, two existing KCF trackers are tested for face tracking; the original KCF algorithm [9] and an updated output constraint transfer (OCT) for KCF [13]. The KCF algorithms are implemented and evaluated as they have been used for neonatal ROI face tracking in previous works [10, 11].

Face ROI: For assessment, we used the average overlap score (AOS), which is the mean intersection over union (IoU). IoU is the area of overlap divided by the area of union between the reference and estimated bounding box.

Landmark Tracking: Mean normalised error (MNE) is used for the assessment of landmark tracking between reference (ref) and estimated (est) landmarks for all test samples (N), where the normalisation distance is the reference face bounding box area (ref_bbox) (1),(2).

$$norm\_error(i) = \frac{\sqrt{(ref_x^i - est_x^i)^2 + (ref_y^i - est_y^i)^2}}{\sqrt{ref\_bbox_{width}^i \times ref\_bbox_{height}^i}} \quad (1)$$

$$MNE = \frac{\sum_{i=1}^{N} norm\_error(i)}{N} \quad (2)$$

Execution Time: We evaluated the average execution time per frame within the video. The average execution time per frame was calculated using MATLAB on a MacBook Pro CPU 2.3 GHz 8-Core Intel i9.

III. Results and Discussion

Table 1: Face Tracking Results

Method	Face (AOS*)	Time (ms)
Stationary Baseline	86.0%	NA
KLT [6,7]	93.0%	5
KCF [9]	88.3%	1
OCT KCF [13]	90.2%	1

*AOS= Average Overlap Score

Table 2: Facial Landmark Tracking Results

Method	Eyes (MNE*)	Nose (MNE)
Stationary Baseline	0.256	0.074
KLT [6,7]	0.026	0.019

*MNE=Mean Normalised Error

Tables 1 and 2 present the results for face and facial landmark tracking. As highlighted by the improvement in AOS and reduction in MNE compared to stationary baseline, tracking is essential as movement is commonplace, even when the baby is asleep. The KLT algorithm produced the best results with AOS of 93% and MNE of 0.026 and 0.019 for eyes and nose landmarks, respectively.

Overall, the KLT algorithm was an effective tracker of minor movements, with the most common reason for the decrease in AOS being a change in the size of the tracked bounding box. Whereas, tracking the overall facial movement and facial landmarks was more reliable for finer movements and minor rotations of the head. In two cases, an eye landmark could not be tracked. Case 1 was due to a large sudden movement resulting in the loss of tracking of the right eye. Case 2 was due to the neonate's hand covering the left eye landmark for part of the video. In future work, implementing previously develop face and facial landmark detection [2] as a recovery mechanism could resolve these two cases.

For cheek ROI selection, the simple formula to determine if the left or right cheek was a more suitable ROI was effective for 24 of the 25 recordings. From a qualitative perspective, the formula to identify an appropriate ROI of the cheek was effective. As the KLT algorithm for face and facial landmarks was reliable, cheek ROI could be tracked effectively.

For forehead ROI selection, an appropriate ROI was selected for 21 of the 25 recordings. The recordings with inappropriate forehead ROIs were due to the angle of the face in the videos. This resulted in the false assumption that the nose line was a continuation of the forehead. However, since these forehead ROIs were outside the face bounding box, it is possible to automatically detect these cases. Similarly with cheek ROI tracking, since the KLT algorithm for face and facial landmarks was reliable, forehead ROI could be tracked effectively, even in the cases with incorrect forehead ROI selection.

The average execution time per frame for the KLT and KCF algorithms was 5 ms and 1 ms respectively. This speed is significantly faster than using repeated ROI detection. In our past work [2], our best-performing face and facial landmark detector had an execution time of 1.36 s, and the fastest existing face and facial landmark detector had an execution time of 9 ms per image. Whilst not implemented, a combination of ROI detection and ROI tracking should be utilised when large motion or occlusion is present. ROI tracking is initially used for its speed efficiency. When the quality of tracking deteriorates to a particular threshold, ROI detection should be used to reinitialise the face bounding box and facial landmark locations. One metric presented by Dosso *et al.* [3] is to monitor tracking quality as the number of tracked features. Whereby a substantial loss of tracked features would suggest a large motion or occlusion had occurred.

With regard to the general applicability of video monitoring for neonates, this study only examined a small set of data (25 videos) of neonates aged 0-6 days old from Xinzhou People's Hospital. A larger and more diverse set, including neonates aged 0-28 days, and with varying skin tones and ethnicity would be required to make stronger conclusions. Different ethnicities have different facial structures and features that may be more or less defined which would affect tracking. Whereas varying skin tones make the contrast of facial structures and features different, affecting tracking.

Additionally, considering gestational age for the applicability of video monitoring is important. For instance, there are significant differences between preterm and term facial expressions because of the development of facial muscles and fat deposition. Preterm neonates generally have less defined features, which could make tracking more difficult. Energy levels also differ, with energy levels and corresponding levels of movement typically increasing with gestation and the size of the baby. Higher levels of movement, especially rapid movements, make tracking harder. In future, examination of gestational age and tracker accuracy should be explored.

IV. Conclusion

The KLT algorithm for tracking is only appropriate for face and facial landmark tracking when movements are relatively minor, thus making it suitable only when the baby is asleep or inactive. Whilst ROI tracking is significantly faster than repeat ROI detection per video frame, a combination of periodic ROI detection (*e.g.* per second) in conjunction with ROI tracking, and/or recovery mechanism when tracking quality drops below a certain threshold would be required when the neonate is active.

Based on preliminary results, forehead and cheek ROI identification and tracking were promising, but struggled with particular face angles. Future work to improve forehead and cheek ROI involves either the calculation and factoring in of head orientation, and/or the identification of a larger set of facial features to localise these regions more effectively. Additionally, quantitative assessments of forehead and cheek ROI identification and tracking, similar to face and facial landmark detection and tracking, are required.

Acknowledgements

E. Grooby acknowledges the support of the MIME-Monash Partners-CSIRO sponsored PhD research support program and Research Training Program (RTP). A. Malhotra's research is supported by the Kathleen Tinsley Trust and a Cerebral Palsy Alliance Research Grant. F. Marzbanrad is supported by the Veski Victoria Fellowship 2021. The study is supported by the Monash Institute of Medical Engineering (MIME) and Monash Data Future Institute (MDFI).

References

1. E. Grooby, C. Sitaula, T. Chang Kwok, D. Sharkey, F. Marzbanrad, and A. Malhotra, "Artificial intelligence-driven wearable technologies for neonatal cardiorespiratory monitoring: Part 1 wearable technology," *Pediatric Research*, pp. 1–13, 2023.
2. E. Grooby, C. Sitaula, S. Ahani, L. Holsti, A. Malhotra, G. A. Dumont, and F. Marzbanrad, "Neonatal face and facial landmark detection from video recordings," *arXiv preprint arXiv:2302.04341*, 2023.
3. Y. S. Dosso, S. Aziz, S. Nizami, K. Greenwood, J. Harrold, and J. R. Green, "Neonatal face tracking for non-contact continuous patient monitoring," in *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2020, pp. 1–6.
4. K. Gibson, A. Al-Naji, J. Fleet, M. Steen, A. Esterman, J. Chahl, J. Huynh, and S. Morris, "Non-contact heart and respiratory rate monitoring of preterm infants based on a computer vision system: A method comparison study," *Pediatric research*, vol. 86, no. 6, pp. 738–741, 2019.
5. D. G. Kyrollos, J. B. Tanner, K. Greenwood, J. Harrold, and J. R. Green, "Noncontact neonatal respiration rate estimation using machine vision," in *2021 IEEE Sensors Applications Symposium (SAS)*. IEEE, 2021, pp. 1–6.
6. J. Shi and C. Tomasi, "Good features to track," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.
7. B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679.
8. L. Svoboda, J. Sperrhake, M. Nisser, C. Zhang, G. Notni, and H. Proquitté, "Contactless heart rate measurement in newborn infants using a multimodal 3d camera system," *Frontiers in Pediatrics*, vol. 10, 2022.
9. J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 583–596, 2014.
10. B. Huang, W. Chen, C.-L. Lin, C.-F. Juang, Y. Xing, Y. Wang, and J. Wang, "A neonatal dataset and benchmark for non-contact neonatal heart rate monitoring based on spatio-temporal neural networks," *Engineering Applications of Artificial Intelligence*, vol. 106, p. 104447, 2021.
11. M. Paul, S. Karthik, J. Joseph, M. Sivaprakasam, J. Kumutha, S. Leonhardt, and C. H. Antink, "Non-contact sensing of neonatal pulse rate using camera-based imaging: a clinical feasibility study," *Physiological Measurement*, vol. 41, no. 2, p. 024001, 2020.
12. L. Scalise, N. Bernacchia, I. Ercoli, and P. Marchionni, "Heart rate measurement in neonatal patients using a webcamera," in *2012 IEEE International Symposium on Medical Measurements and Applications Proceedings*. IEEE, 2012, pp. 1–4.
13. B. Zhang, Z. Li, X. Cao, Q. Ye, C. Chen, L. Shen, A. Perina, and R. Jill, "Output constraint transfer for kernelized correlation filter in tracking," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 693–703, 2016.

Corresponding Author: Ethan Grooby
Institute: Monash University
Country: Australia
Email: ethan.grooby@monash.edu