Dual CNN-based Face Tracking Algorithm for an Automated Infant Monitoring System

Cheng Li[a], Genyu Song, A. Pourtaherian[b] and P. H. N. de With[c]
Eindhoven University of Technology, Eindhoven, The Netherlands

Keywords: Face Tracking, Online Updating, Infant Monitoring, Real-time Application.

Abstract: Face tracking is important for designing a surveillance system when facial features are used as main descriptors. In this paper, we propose an on-line updating face tracking method, which is not only suitable for specific tasks, such as infant monitoring, but also a generic human-machine interaction application where face recognition is required. The tracking method is based on combining the architecture of the GOTURN and YOLO tiny face detector, which enables the tracking model to be updated over time. Tracking of objects is realized by analyzing two neighboring frames through a deep neural network. On-line updating is achieved by comparing the tracking result and face detection obtained from the YOLO tiny face detector. The experimental results have shown that our proposed tracker achieves an AUC of 97.9% for precision plot and an AUC of 91.8% for success plot, which outperforms other state-of-the-art tracking methods when used in the infant monitoring application.

1 INTRODUCTION

Monitoring the comfort states of young infants for certain disease diagnosis at a hospital is very important, because of their limited verbal ability to express their feelings. For better understanding of infant behavior and moments of pain/discomfort, an automated video-based infant monitoring system can be implemented by analyzing expressions as an auxiliary assessment tool. State-of-the-art infant monitoring systems normally consist of three components, face detection, landmark localization and discomfort detection (Zamzmi et al., 2016) (Sun et al., 2018), which is depicted in Fig.1. Since the system is composed of cascaded stages, the final discomfort detection is strongly dependent on the accuracy of preceding stages, e.g. face detection. However, face detection over the entire image is computationally expensive, which hampers the possibility of a real-time application. Therefore, a face tracking method is commonly used for video analysis when a face is detected.

When using face tracking, several challenges for infant monitoring systems should be considered. First, most state-of-the-art trackers are designed for pedestrians, humans or vehicles, of which textures typically undergo limited deformations and distortions over time. However, infant expressions and their head poses can change dramatically in adjacent frames due to unexpected stimuli and fast head movements, causing significant texture changes compared to the tracking target. Hence, a generic object tracker will fail to track infant faces when such situations occur. Second, infants are normally monitored under a complex environment, such as interacting with their nearby patients. Therefore, multiple object instances (faces of infants and parents) can coexist in the search regions, which increases the ambiguity for the designed object tracker. Last, some of state-of-the-art trackers based on CNNs adopt complex architectures, which require a considerable amount of computations, thereby significantly increasing the implementation burden of a real-time infant monitoring application.

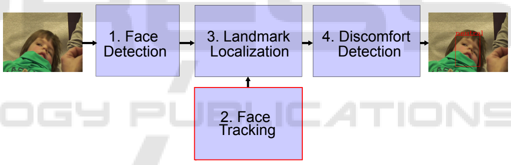


Figure 1: Block diagram of an infant discomfort detection/monitoring system with a face tracking step.

[a] https://orcid.org/0000-0003-2900-637X
[b] https://orcid.org/0000-0003-4542-1354
[c] https://orcid.org/0000-0002-7639-7716

Li, C., Song, G., Pourtaherian, A. and N. de With, P.
Dual CNN-based Face Tracking Algorithm for an Automated Infant Monitoring System.
DOI: 10.5220/0010384308810887
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 881-887
ISBN: 978-989-758-488-6

In order to solve these challenges, a face online-updated tracking algorithm aiming at integrating it into an automated infant monitoring system is proposed. This algorithm is designed based on combining a GOTURN tracker (Held et al., 2016) and a YOLO pre-trained infant face detector (Redmon et al., 2016). Compared to state-of-the-art tracking methods using reinforcement learning (Yun et al., 2017), the architecture of GOTURN is adopted for its simplicity and low computation requirements. The research in this paper presents the following contributions.

1. The proposed tracker applies an online-updating technique which combines GOTURN and the YOLO tiny face detector. This novel combination has proven to outperform the individual single components and other state-of-the-art trackers when being used for the infant monitoring application.
2. In order to thoroughly validate the performance of the proposed tracker, a clinical dataset captured from a local hospital and a consumer-oriented dataset collected from *Youtube* are used for evaluation purposes.
3. The experimental results of the proposed system can achieve an execution speed comparable with state-of-the-art tracking methods based on correlation filtering.

This paper is organized as follows. Section 2 introduces related work on several state-of-the-art tracking methods. Section 3 describes the proposed infant face tracking algorithm. The tracking accuracy and computation costs are evaluated in Section 4. Finally, conclusions are presented in Section 5.

2 RELATED WORK

For decades, researchers have paid significant attention to object and face tracking. Tracking algorithms can be categorized into conventional methods based on template matching (Comaniciu and Meer, 2002), Bayesian inference (Van Der Merwe et al., 2001) (Welch et al., 1995), correlation filter-based tracking (Bolme et al., 2010) (Danelljan et al., 2016a) (Danelljan et al., 2016b) (Henriques et al., 2014), and CNN-based tracking (Nam and Han, 2016) (Li et al., 2018a) (Yun et al., 2017). Conventional methods based on template matching usually heuristically search objects according to pre-defined templates. However, it has been demonstrated that such techniques only perform well for a tracking target that can be represented by a simple feature model (such as a ball with a unified color). Besides, these trackers are likely to fail when obstacles occur near the tracking target, which hampers their application for complex tracking tasks.

Tracking methods based on correlation filtering have become prevalent for addressing the drawbacks of template-matching methods and apply a Fourier Transform (FT) on both the target template (object of interest) and search regions. In (Danelljan et al., 2016b), the target template and search regions are represented by layers of features such as in one or more CNNs. After this, a confidence map is calculated from the FT of the template and the search regions with a convolutional operation. Finally, the tracking target location in the search area is determined with the highest confidence score based on a confidence map. When tracking of the current frame is succeeded, the target template is updated by the current detection. However, this type of tracker lacks the robustness of tracking the deformable objects, thereby being less suited for infant face tracking.

Tracking methods based on CNN features have shown a great success in recent years (Held et al., 2016) (Nam and Han, 2016). The outstanding performance of CNN-based tracking is explained by the strong representation ability of CNN features. These methods can be divided into two categories: (a) tracking in an off-line mode (Held et al., 2016) (Li et al., 2018a) and (b) tracking with an on-line updating scheme (Nam and Han, 2016) (Yun et al., 2017). The framework of these CNN-based trackers accept a target template and search regions as inputs of their CNN architectures. Compared to conventional tracking methods and correlation filter-based trackers, object tracking obtained by CNNs is realized by regression of CNN features instead of optimization. However, the aforementioned online-updated CNN trackers are designed to be generic for various objects, which lack the specificity for tracking infant faces, thereby making it difficult to use directly in an infant monitoring system. This paper proposes a CNN-based online-updating tracking method specifically targeting at infant faces, which combines GOTURN and the YOLO tiny face detector. The CNN features used for tracking can also be shared with infant expression analysis, and are therefore compatible with a general infant monitoring system and other generic human-machine interaction applications.

3 SYSTEM DESIGN

This section discusses the architecture of the proposed online-updating tracking method and the training procedure in more detail.

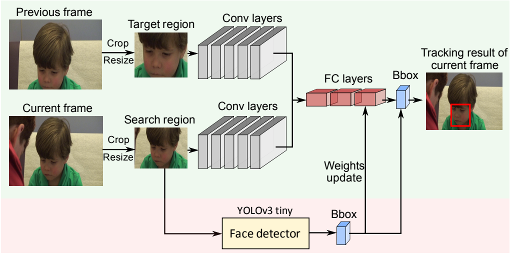


Figure 2: Architecture of the proposed online-updated tracker based on the GOTURN and YOLO tiny face detector.

3.1 Architecture

When differentiating from tracking for a generic purpose, a target-specific tracker can be more robust and become less a complex CNN architecture, provided that it is trained with proper datasets. For this reason, we utilize the same structure as in GOTURN (Held et al., 2016), and combine it with AlexNet (Krizhevsky et al., 2012) because if offers a shallow and fast network. The framework of the target-specific on-line updated tracker is depicted in Fig. 2. As shown, the input of the network consists of a target patch and a cropped patch (search region), which are potentially containing the target to be tracked. Here, the target is an infant face that is detected from the previous frame, whereas the search region is obtained through cropping the next frame. The center of the search region is determined by the target position, while its size is pre-defined according to numerous experiments. To fulfill the input size requirement of AlexNet, both target and search patches are resized, and then go through their individual convolution-layer branch to obtain the respective CNN features. In order to estimate the target in the subsequent frame, CNN feature maps of the target and search patches are concatenated, and are then supplied into Fully Connected (FC) layers. Finally, the target location is computed from the output of the last FC layer.

Probably caused by the computational efficiency of AlexNet, it provides less robustness when infants heavily rotate their head in a video sequence. In this case, the tracker can easily lose the target. To address the non-rigid deformation caused by rapid head movements, a YOLOv3 face detector is added for the potential face detection in the search region to update the network. The bounding box from the face detector is considered as ground truth for fine- tuning the network and therefore adapting the network to the infant facial deformation over time. In this work, only the weights of the last FC layer are fine-tuned to allow online updating.

3.2 Training

3.2.1 Input of Network

To address the challenge of multiple persons presented in video sequences, an image patch for searching the infant face is cropped from the current frame, denoted as F_t. The center of the cropped patch is located at the position of the detected face from the previous frame, denoted as F_{t-1}. Assuming the detected bounding box containing the face has a center $c(x,y)$, then the width and the height of the bounding box are represented as w and h, respectively. Then, in the actual frame F_t, the search region is cropped with k times on the bounding box of the previous frame, where k is a scale factor indicating the size of the search region. Here, the value of k is empirically determined. In order to avoid too much background noise encompassed in the search area, we have set $k = 2$ for balancing the computational overhead and the tracking accuracy.

3.2.2 Loss Function

The loss is computed from the output of the last FC layer and the ground-truth bounding-box coordinates. Here, the L1 loss is chosen, denoted by L_1, because it is easy to compute and also proved to be reliable against outliers. The loss function $\mathcal{L}$ is computed by:

$$\mathcal{L} = \sum_i L_1(m_i - m_i^*), \tag{1}$$

where m_i denotes the coordinates (x_1, y_1, x_2, y_2) of the top-left and bottom-right corners of the predicted bounding box by the tracking network and m_i^* represents the corresponding coordinates of the ground-truth bounding box.

3.2.3 Training Configuration

Because of the limited availability of video sequences for training, the weights of convolutional layers are first trained with ImageNet data (Deng et al., 2009). After that, the whole network is fine-tuned with our training video sequences containing infants, in which the weights of the FC layers are initialized by sampling from a Gaussian distribution with zero mean and standard deviation of 0.01. Furthermore, a Stochastic Gradient Descent (SGD) is used for optimizing the network parameters with a base learning rate of 0.001 and a momentum of 0.9. The final FC layer outputs the coordinates of the bounding box predicted by the tracking network.

3.2.4 On-line Updating Process

In order to address the deformations of infant faces over time, an online adaptation technique is utilized. To maximally reduce the computational requirement, only the weights of the last FC layer are updated. The loss function used for updating the weights is identical to the specification of Eq. (1), where only m_i^* represents the coordinates of the bounding box detected from the YOLO tiny network in this stage.

4 EXPERIMENTS

In this section, we first introduce the databases used for training and testing for the proposed tracking method. Then we describe the applied evaluation metrics. Finally, the experimental results of the infant face tracking are provided.

4.1 Database

The dataset used for this work contains 73 video sequences in total, which can be divided into two subsets. One subset contains 11 video sequences that were recorded at the Maxima Medical Center (MMC), Veldhoven, the Netherlands, whereas the other subset consists of 62 video sequences that were collected from *YouTube*. Video sequences recorded at the MMC comply with the ethical standards, and allow usage for experiments after obtaining a written consent. In this subset, each video sequence lasts at least 2 minutes for meeting the requirements for infant pain assessment. Furthermore, the length of video sequences in the *YouTube* subset lasts from 20 seconds to 2 minutes.

For training the tracking network, 51 video sequences are selected which maximally encompass a variety of challenging situations, such as large head poses and object occlusions. In the training video dataset, 8 videos are selected from the MMC subset and 43 videos are taken from the *YouTube* subset. The rest of the video sequences are then used as testing datasets. In order to thoroughly compare the proposed tracking network with state-of-the-art methods, the testing dataset also contains the challenging cases as earlier mentioned. Table 1 provides the number of frames for the example video sequences in the testing dataset.

4.2 Metrics

In order to evaluate the tracking accuracy, two metrics from the Online Tracking Benchmark (OTB) (Nam and Han, 2016) are used, which are the success plot and the precision plot. These two metrics are specifically designed for evaluating and benchmarking the overall performance of different tracking methods. According to the definition, a success plot presents the successful tracking rate in terms of overlap defined by the Intersection over Union (IoU) ratios between the detected and ground-truth bounding boxes. Alternatively, the precision plot provides the precision of tracking methods in terms of the difference between the center locations of the tracked bounding box and the ground-truth bounding box, which is indicated as a tracking error. In addition to the tracking accuracy, the execution speed is also compared and expressed as frame throughput rate (fps), for indicating the possibility of implementing the tracking methods in a real-time system.

Table 1: Number of frames for the example video sequences for the testing dataset used for evaluation of face tracking.

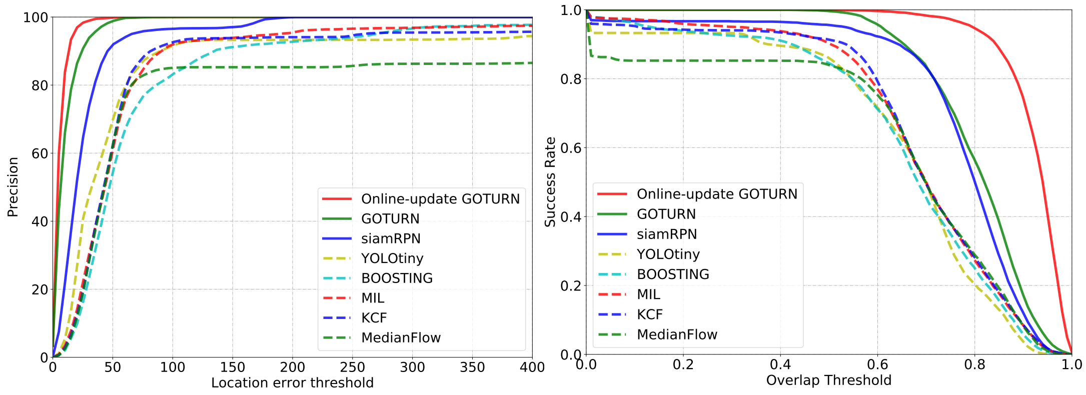
Video Id	No. FRM	Video Id	No. FRM
MMC #001	1,175	*Y-T* #001	329
MMC #002	1,139	*Y-T* #002	1,108
MMC #003	1,752	*Y-T* #003	1,764

4.3 Experimental Results

4.3.1 Tracking Accuracy

In this experiment, both conventional tracking methods (Grabner et al., 2008) (Kalal et al., 2010) (Henriques et al., 2014) (Kalal et al., 2010) and state-of-the-art deep learning-based tracking methods (Li et al., 2018b) (Held et al., 2016), are evaluated and compared with our proposed approach. When applying the conventional methods to the testing video sequences, the tracking target is initialized every 100 frames regardless of the outcome of the tracker, since they are less robust to situations when tracking targets suddenly and significantly deviate from their previous position. Instead, for the deep learning-based method, only the first frame of each sequence is initialized and then tracking is sustained till the end of the sequence.

Fig. 3 shows the success plot and the precision plot for all the utilized tracking methods, evaluated with our infant test video sequences. It can be observed that the correlation filter-based tracker (KCF) (Henriques et al., 2014) outperforms other conventional tracking methods (MIL, Boosting and MedianFlow) (Babenko et al., 2009) (Grabner et al., 2008) (Kalal et al., 2010), due to its online updating process of the target template. However, the CNN-based tracking methods (GOTURN and SiamRPN) (Held et al., 2016) (Li et al., 2018b) out-



(a) (b)

Figure 3: Performance for all the tracking methods used for comparison evaluated with the test infant datasets. (a) Precision plot; (b) Success plot.

Table 2: Area under curve (AUC) of precision plot and success plot evaluated with the testing infant datasets for all the tracking methods. Boldface indicates the highest number.

Tracking methods	Precision	Success
Online-GOTURN	**0.979**	**0.918**
GOTURN	0.969	0.800
SiamRPN	0.927	0.765
YOLOtiny	0.838	0.640
Boosting	0.817	0.654
MIL	0.847	0.682
KCF	0.841	0.680
MedianFlow	0.768	0.630

performs the evaluated conventional methods with a higher accuracy. It should be noted that the proposed on-line updating tracking method achieves the highest performance evaluated with both the success plot and the precision plot. Furthermore, Table 2 shows the Area Under the Curve (AUC) values for both success plot and precision plot. It can be readily noticed that the proposed tracking method achieves an AUC of 97.9% and an AUC of 91.8% for success plot and precision plot, respectively, which performs the best compared with all the evaluated trackers for the infant face tracking application. Particularly, the proposed method is compared with both the GOTURN and the YOLO tiny face detector, where the proposed combined tracking outperforms any of its individual components. Fig. 4 portrays the tracking results obtained by different tracking approaches when being used for infant face tracking.

When implementing different tracking methods for evaluation, we have noticed that KCF and MIL can only be reliable when an infant face is frontal and static. Unfortunately, when the infant heads significantly deviate from the frontal view, such as with a head rotation from frontal to profile view, these trackers are more likely to fail, due to the facial texture changes. This failure can be explained by the aspect that their tracking updating solution is less descriptive for the texture changes between frontal and profile views. However, our proposed on-line updating tracking network is pre-trained with various views. Furthermore, the on-line network updating technique also allows it to model the deformations of the facial appearance between frames, which explains the high performance of the proposed tracking method. As a result, the high accuracy for the tracking of faces makes the proposed tracking method suitable for integrating it in a face-related human-machine-interaction system.

4.3.2 Execution Speed

In addition to the tracking accuracy, the execution speed is also evaluated. Table 3 presents the execution speed of each tracker of interest. It is clear that KCF has the lowest speed as 11 fps, and MedianFlow achieves the highest as 128 fps. Even though the speed of MedianFlow is appealing, it lacks the robustness against large head rotations and oc-

Table 3: Execution speed for each tracker operated on an Xeon(R) CPU E5-2609 v2 @ 2.50 GHz with octal CPU core.

Tracking method	Execution speed (fps)
Boosting	27
MIL	21
KCF	11
MedianFlow	**128**
SiamsRPN	30
GOTURN	50
Online updated tracking	13

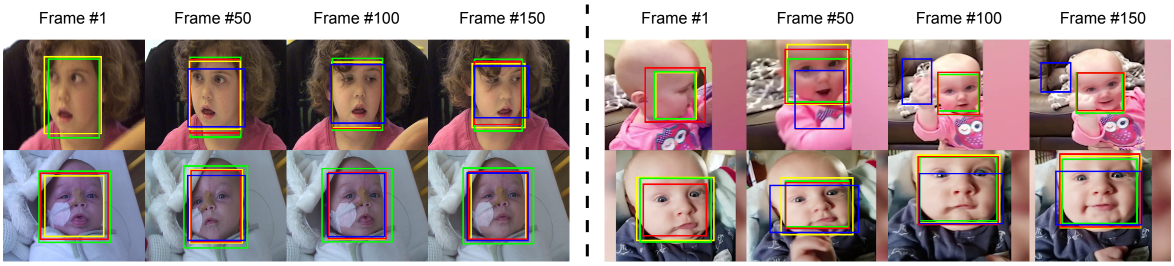


Figure 4: Examples of tracking results on frames within sequences from the testing dataset obtained by three CNN-based tracking methods. Yellow boxes represent the ground truth, green boxes indicate tracking results by GOTURN, blue boxes show the outputs from SiamsRPN and red boxes present the results obtained by our proposed method.

clusions. Compared with the conventional methods, the CNN-based trackers generally require more computation time, due to their high model complexity. However, the CNN-based trackers are demonstrated to provide higher robustness on discontinued texture changes and occlusions compared to conventional methods. Therefore, the frequency of target initialization is much lower than with conventional methods. As shown, the speed of the proposed on-line updating tracking is slower than SiamsRPN and GOTURN, due to the overhead computation of online weight-factor updating and the parallel structure of the YOLO tiny face detection. Therefore, the proposed tracker obtains high tracking accuracy at the cost of a lower execution speed. Nevertheless, the speed of the proposed tracking method is still feasible for implementing an infant monitoring system. Moreover, the convolutional layers in the tracking network can also be shared with the expression detection when tracking and detection are trained in an end-to-end fashion. It is assumed that this reduces the computation overhead induced by the additional face detector.

5 CONCLUSIONS

In this paper, we have proposed an on-line updating tracking method aiming at infant face tracking, based on combining the architecture of the GOTURN and a YOLO tiny face detector. The proposed solution can also be reused for other human-machine interaction applications. The tracking position of an object is obtained by analyzing two neighboring frames through a deep neural network, where the network is on-line updated by comparing the tracking output and a detection provided by the YOLO tiny face detector. The experimental results have shown that our proposed tracker achieves an AUC of 97.9% for precision plot and an AUC of 91.8% for success plot, which outperforms not only its single components, but also other state-of-the-art tracking methods when used in this application. Although the execution speed is less competitive than other evaluated trackers, the overall computation complexity can be further reduced when using this tracker in an infant monitoring system or other face-related applications, since the parameters of the network can be shared for other detection purposes. For future work, we will integrate this tracking method in an facial surveillance application, and jointly train the face tracking and recognition in one framework.

REFERENCES

Babenko, B., Yang, M.-H., and Belongie, S. (2009). Visual tracking with online multiple instance learning. In *2009 Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 983–990. IEEE.

Bolme, D. S., Beveridge, J. R., Draper, B. A., and Lui, Y. M. (2010). Visual object tracking using adaptive correlation filters. In *2010 Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2544–2550.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):603–619.

Danelljan, M., Bhat, G., Khan, F. S., and Felsberg, M. (2016a). ECO: efficient convolution operators for tracking. *CoRR*, abs/1611.09224.

Danelljan, M., Robinson, A., Shahbaz Khan, F., and Felsberg, M. (2016b). Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *ECCV*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Grabner, H., Leistner, C., and Bischof, H. (2008). Semi-supervised on-line boosting for robust tracking. In *ECCV*, pages 234–247. Springer.

Held, D., Thrun, S., and Savarese, S. (2016). Learning to track at 100 fps with deep regression networks. In *ECCV*.

Henriques, J. F., Caseiro, R., Martins, P., and Batista, J. (2014). High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intel*, 37(3):583–596.

Kalal, Z., Mikolajczyk, K., and Matas, J. (2010). Face-tld: Tracking-learning-detection applied to faces. In *2010 IEEE International Conference on Image Processing*, pages 3789–3792. IEEE.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Li, B., Yan, J., Wu, W., Zhu, Z., and Hu, X. (2018a). High performance visual tracking with siamese region proposal network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 8971–8980.

Li, B., Yan, J., Wu, W., Zhu, Z., and Hu, X. (2018b). High performance visual tracking with siamese region proposal network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*

Nam, H. and Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In *2016 Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4293–4302.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 779–788.

Sun, Y., Shan, C., Tan, T., Long, X., Pourtaherian, A., Zinger, S., and de With, P. H. N. (2018). Video-based discomfort detection for infants. *Mach. Vis. Appl.*

Van Der Merwe, R., Doucet, A., De Freitas, N., and Wan, E. A. (2001). The unscented particle filter. In *Advances in neural information processing systems*, pages 584–590.

Welch, G., Bishop, G., et al. (1995). An introduction to the kalman filter.

Yun, S., Choi, J., Yoo, Y., Yun, K., and Choi, J. Y. (2017). Action-decision networks for visual tracking with deep reinforcement learning. In *2017 Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1349–1358.

Zamzmi, G., Pai, C., Goldgof, D., Kasturi, R., Ashmeade, T., and Sun, Y. (2016). An approach for automated multimodal analysis of infants' pain. In *2016 23rd Int. Conf. Pattern Recognit. (ICPR)*, pages 4148–4153.