The Journal of Supercomputing (2022) 78:14343–14361
https://doi.org/10.1007/s11227-022-04439-x

Full-convolution Siamese network algorithm under deep learning used in tracking of facial video image in newborns

Yun Wang[1] · Lu Huang[2] · Austin Lin Yee[3]

Accepted: 10 March 2022 / Published online: 1 April 2022

Abstract

This study was carried out with the aim of exploring the full-convolution Siamese network (SiamFC) in the application of neonatal facial video image tracking, achieving accurate recognition of neonatal pain and helping doctors evaluate neonatal emotions in an automatic manner. The current technology shows low accuracy on facial image recognition of newborns, so the SiamFC algorithm under the deep learning was optimized in this study. Besides, a newborn facial video image tracking model (FVIT model) was constructed based on the SiamFC algorithm in combination with the attention mechanism with face tracking algorithm, and the facial features of newborns were tracked and recognized. In addition, a newborn face database was constructed based on the adult face database to evaluate performance of the FVIT model. It was found that the accuracy of the improved algorithm is 0.889, higher by 0.036 in contrast to other models; the area under the curve (AUC) of success rate reaches 0.748, higher by 0.075 compared with other algorithms. What's more, the improved algorithm shows good performance in tracking the facial occlusion, facial expression changes, and scale conversion of newborns. Therefore, the improved algorithm shows higher accuracy and success rate and has good effect in capturing and tracking the facial images of newborns, thereby providing an experimental basis for facial recognition and pain assessment of newborns in the later stage.

Keywords Full-convolution Siamese Network · Newborns · Image Tracking Technology · One Pass Evaluation · Facial Recognition

✉ Austin Lin Yee
Austinlinyee@pku.org.cn

Extended author information available on the last page of the article

1 Introduction

Under the rapid development of scientific technology, the human daily life has changed day by day. Face detection technology has been widely applied in different scenarios, which has set off a great mass fervour of "face recognition", such as mobile payment, attendance clocking, and criminal investigation. Face recognition is an important part of computer vision. It is a prerequisite and basis for the research on the recognition of face attributes and emotions. For this reason, the precision of recognition is directly related to the effectiveness and reliability of subsequent operations [1]. However, the accuracy of the algorithm related to face recognition has dropped sharply in the actual environment, which puts forward higher requirements on the algorithm technology related to face detection. Hence, related research on face recognition and detection has become the focus of scientific research scholars.

Human and computer usually can achieve the information transmission with a dialogue language in a certain interactive manner, among which human facial expression is a common interactive information. Facial expression is the most directly external manifestation of human inner thoughts and emotions, as well as an important source of emotional information. In addition, it involves psychology, physiology, pattern recognition, artificial intelligence (AI), and other fields. Thus, research on the human facial recognition appears rapidly in recent years. Human face is a main biological feature that distinguishes each person. Compared with biometrics such as fingerprints and iris, collection methods and equipment of human face are relatively simple, and no human behavior is required [2]. Some information characteristics, such as gender, age, skin colour, and emotions, can be collected by observing the human face, so it has gradually become an important research object in the field of vision [3]. Besides, databases corresponding to human faces become increasing more and standard. The research on face recognition and detection has gradually transitioned from previous static images to video sequences. Facial expression recognition based on video sequences is widely used in daily life of human beings. For example, it can assist the public security organs in solving crimes and arresting criminals in the security field, help patients with disease treatment and achieve good doctor-patient communication in the medical field, and improve the efficient in the classroom by showing the recognition and feedback results of facial expressions of students in the remote education field [4, 5]. In short, effective facial expression recognition on real-time video sequences can make human–computer interaction smarter and more natural, so that computers can better serve humans.

At present, there are many technologies related to facial recognition. Visual target tracking is one of the important algorithms under the computer vision, which integrates multi-disciplinary technologies such as image processing, pattern recognition, automatic control, machine learning, and deep learning. Thus, it is the basis for many complex computer vision tasks [6–11]. Visual target tracking can predict the position and size of the target in subsequent frames given the position and size of the target in the initial frame of a video sequence. For a moving target, it will constantly change during the movement, and its moving scene is usually very complicated and

accompanied by various interferences. The deep learning algorithm has been widely applied in many fields, such as energy regeneration [12], machine translation [13], medicine and healthcare [14], prediction on air concentration [15], and analysis on air pollution [16]. It can effectively identify and extract the features with the assistance of independent learning. With the continuous development of deep learning, convolutional neural network (CNN) has been widely used in the computer vision. Therefore, it is of extremely important significance of applying the deep learning algorithms in face recognition and visual target tracking.

In summary, the research on facial recognition has gradually become mature, and most of the current facial recognition testing instruments with better results are designed based on adult samples. Nevertheless, the facial features of newborns are quite different from those of adults. For example, the skin of newborns is redder and the eyes are often closed. During the growth of newborns, different pain stimuli may have a series of adverse effects on the newborns, threatening their healthy growth, and even endangering their life in severe cases. Therefore, this study was developed to realize a more accurate recognition effect on the expression of newborns pain and to help medical staff automatically evaluate the pain degree of newborns, thereby reducing the interference caused by the subjective judgment of medical staff. The full-convolution Siamese network (SiamFC) algorithm in deep learning was improved based on the adult face recognition algorithm, and the attention mechanism was integrated to build a newborn facial video image tracking model (FVIT model). The performance of the constructed model is analysed through simulation, which will provide an experimental reference for the pain assessment and facial recognition of neonates in the later period.

The overall structure of this work is as follows. Chapter II is literature review, in which the research status of face recognition and tracking was analysed, and the shortcomings and advantages of artificial intelligence algorithm in the application and research of face recognition were expounded, so as to highlight the focus of this study. Chapter III introduces the methodology, in which the full-convolution twin network was introduced into the newborns, the algorithm model of newborns face video image tracking was built based on the full-convolution twin network, and its simulation was carried out. Chapter IV gives the results and discussion, in which the proposed algorithm was compared with the algorithms in related fields through simulation. In Chapter V, the advantages and achievements of this research were described, and the existing shortcomings and prospects of the relevant content were explained.

2 Literature review

Face recognition and face video tracking is one of the important research directions in the field of computer vision because of its wide application field and practical value. Many scholars in related fields have studied its application status and development trend.

2.1 Current research status of facial detection and recognition technologies

Face detection technology can judge whether there are human faces in the picture through facial feature extraction, classification, and processing by using the regression model. If any human face is found, its position and size will be generated. Research on face detection has been gradually developed for half a century since the semi-automatic face detection and recognition system was first implemented in the 1960s and 1970s. Al-Janabi et al. (2016) proposed a security method that uses genetic algorithm to generate keys and select the optimal mixing matrix value to hide one or more images in the cover image of the same size. The results showed that this method improves the ability to hide multiple images in the cover image and enhances the hiding capacity, security, and robustness against specific attacks [17]. Omer (2019) put forward the non-face objects, which could produce strong facial perception (namely hallucinations). It turned out that face detection relied on specific facial features, eyes, and mouth by evaluating the appearance, local and global facial features of this group of inanimate images [18]. Al-Janabi et al. (2020) established a new tool to estimate the ability of missing values in various image recognition data sets, namely the developed random forest and local least squares (DRFLLS), estimated the missing values using local least squares (LLS), and measured the accuracy of the results by normalized root-mean-square error (NRMSE) and Pearson correlation [19]. Al-Janabi et al. (2021) proposed a smart data analysis model to find the best mode of human activity based on the biological characteristics obtained by four sensors installed on smart phones and smart watch devices, that is, the scheduling activities of smartphones and smartwatches based on the optimal pattern model (SA-OPM). The analysis results from the four stages showed that the proposed SA-OPM model generates robust and real human activity patterns [14].

2.2 Current research status of facial tracking technology

Target tracking is an important research branch and direction in computer vision and machine learning. Its development includes three stages: traditional target tracking algorithms as the mainstream, detection-based target tracking algorithms as the mainstream, and correlation filtering and deep learning-based target tracking algorithms as mainstream. With the rapid development of scientific information technology, more and more scholars in related fields have researched on it. Al-Janabi et al. (2016) proposed a video compression technology to obtain the highest compression ratio and high-quality video compression. It was found that the proposed method divides the video sequence file according to the tilt measurement value and the specific threshold determined by the embedded zerotree wavelet (EZW) algorithm, thus confirming the effectiveness of the method [20]. Chrysos et al. (2018) firstly adopted the recently introduced 300 VW benchmark for the most advanced deformable face tracking pipeline for a comprehensive evaluation. It was compared with the universal face detection + universal facial landmark positioning, universal model-free tracking + universal facial landmark positioning + the most advanced facial detection, and the model-free tracking + facial landmark positioning technology, to reveal its future

use and availability [21]. Sonkusare et al. (2019) developed a new semi-automatic heat signal extraction method based on the deep learning algorithms to identify the facial landmarks and found that this method could capture the signals of manual intervention and manual pre-processing with minimal changes in facial physiology, which was a sensitive and robust tool. The proposed method was expected to promote the use of thermal imaging (infrared imaging) in social and emotional neuroscience as an ecologically effective technology [22]. Low et al. (2020) proposed a new framework that combined a digital signage with a depth camera to track multiple faces in a three-dimensional environment. The proposed framework extracted the facial centroid position (x, y) and depth information (z) of audience and drew them to an aerial map to simulate the audience movement corresponding to the real environment [23].

In summary, analysis of the research of scholars in the above-mentioned related fields reveals that many scholars adopted artificial intelligence technologies such as deep learning and genetic algorithms to image analysis such as face recognition, but most of them are based on adults for design analysis, and the methods used are relatively simple, and no clearer results are obtained. Therefore, in this study, newborns are undertaken as the research objects, the basic deep learning technology is improved and deepened, and corresponding model is constructed to study and research on newborns face detection, which is of extremely important value.

3 Methods

Facial image recognition is a popular field with a wide range of applications. Further accurate recognition of pain and expression of newborns will provide great convenience to the work of medical staff. Newborns are recruited as the research object. The theoretical knowledge related to deep learning and face recognition is explained, and the full-convolution twin network is designed by introducing the deep learning algorithm and further improving it. A newborn face video image tracking model based on the algorithm is constructed, and its performance is evaluated through simulation experiments.

3.1 Limitations of newborn facial video image recognition

With the rapid development of science and technology in recent years, theoretical research based on neural networks has gradually improved and has achieved good results in areas such as face recognition since the introduction of deep learning algorithms. However, the current face detection methods are generally applicable to adult face detection, and the missed detection and misdetection are more serious for children, especially newborns. The main reason is that the current face detection method is designed for the facial features of adults, but there are big differences between the facial features of newborns and adults in actual situations, such as closed eyes and narrow nasal cavity. Therefore, it is necessary to combine the characteristics of the newborns' face and improve the existing algorithm, train the deep learning model by

using the newborns face image data set, and adaptively learn the feature expression suitable for the newborns' face, to avoid the limitations for manually extracting the facial features of newborns.

Therefore, the commonly used face detection methods are introduced firstly, and then, the convolutional neural network (CNN) in deep learning is analysed and improved and then applied to the facial recognition of newborns, to understand the newborns' pain and what it wants to express, which is of great significance to the healthy growth of newborns.

3.2 Human facial detection algorithm

Face detection is one of the pre-processing works of facial video recognition as well as the basis of facial recognition. The quality of face detection algorithms directly affects the effectiveness and reliability of subsequent face tracking and recognition work.

As a key technology in the field of face information processing, face detection has now become an independent research direction, and its research results have also been widely used in the security monitoring and vision-related human–computer interaction. Many current face detection methods have harvested good results for adult face detection, but there are still serious mistaken detection and missed detection in facial detection of newborns. This may be because the existing face detection algorithms are mainly designed by taking adults as samples. However, the facial features of newborns and adults are quite different. For example, eyes of newborns are often closed, which is a typical difference. Therefore, it is necessary to combine the characteristics of the newborns to improve the existing algorithm by increasing its effectiveness and robustness, to improve the effect of newborn facial detection. At present, the technology of face detection and recognition include the knowledge-based, features-based, template matching-based, and statistics-based (Fig. 1) [24–26].

In the facial recognition algorithm, the statistical analysis or machine learning is adopted regarding the statistics-based face detection to learn the statistical features of face and non-face samples from data samples, aiming to build a classifier to realize the human face detection. The discriminant rules of this method are obtained by statistical learning from massive samples, which requires no prior knowledge of the researcher, so it is relatively simple and can reduce a series of detection problems caused by inaccurate prior knowledge [27]. In other algorithms, the manual calculations or manual selection methods are adopted to obtain the facial discrimination models, while the machine learning methods are adopted to mine the deep features of human faces from abundant data samples and discover the key features to distinguish the non-human faces. Thus, good detection results can be obtained, so this method is increasingly being used in this field.

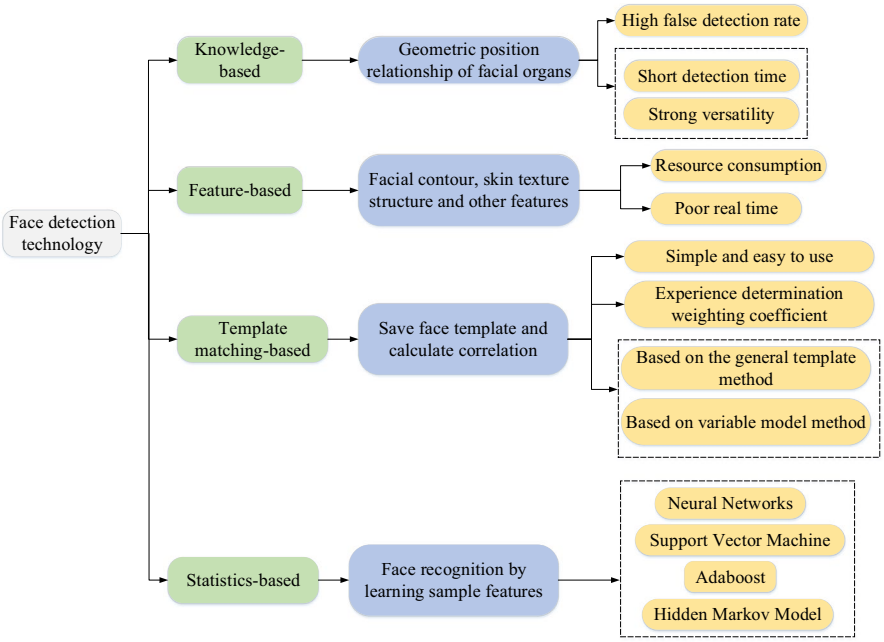

Fig. 1 Classification of human facial recognition algorithms

3.3 Analysis on human facial tracking algorithm

As one of the hot spots in the field of vision, target tracking has achieved great progress in recent years. In the face tracking algorithm, the features of the current frame image were firstly extracted. Then, the target feature corresponding to the target frame area in the template frame was measured for similarity (or other evaluation methods) with the features of many candidate frame areas in the current frame, so as to obtain the score (probability) of the candidate frame. Finally, the candidate frame with the highest score was the location of the tracking target.

3.3.1 Shortcomings of human facial tracking technology

Great challenges are faced in the research of face tracking technology. How to balance the speed and precision under unconstrained conditions has become an urgent problem to be solved. The main reasons are summarized as follows. Firstly, face poses are changeable. The complex and diverse facial expression changes distort the facial features to varying degrees. Secondly, environmental factors play a great effect. Various occlusions (such as occlusion by the target itself or by other target) and background interference make face tracking more difficult. Thirdly, the face state changes greatly. The fast movement speed brings the face image blur, deflection, field of view, and scale change, which may cause template drift or even tracking failure. Fourth, image quality is poor in image exposure, resolution, colour,

noise, and blur. Therefore, the realization of a real-time face tracking system with good precision and robustness still must be continuously researched in deep.

3.3.2 Convolutional neural network

As one of the deep learning algorithms, the CNN is a hierarchical model stacked through operations such as convolution, pooling, and nonlinear activation function mapping [28]. In this model, the original red, green, and blue (RGB) image and audio data are inputted, and then, the original data are extracted layer by layer and abstracted into high-level semantic information. Finally, the last layer of the CNN formalizes the target tasks such as classification and regression into the objective function. The error between the predicted value and the true value is calculated, the error or loss is transferred from the last layer to the forward, and the parameters of are updated in each layer. The feed-forward is conducted again after the parameters are updated. The feed-forward and feedback process is repeated continuously until the network model converges, thereby fabricating a CNN model with trained parameters. The classical CNN models include AlexNet [29] and visual geometry group network (VGGNet) [30]. These network models have greatly improved the performance of related tasks. Figure 2 shows the network structure of VGG-16.

3.3.3 SiamFC for human facial tracking

An important branch of target tracking is matching-based tracking. Using the matching ability of the SiamFC in the tracking task is a better solution to the similarity learning, and the SiamFC has received great attention in the tracking field. SiamFC refers to a neural network architecture that includes two identical sub-networks with the same network framework and the same parameters and weights [31]. In the SiamFC, there are two inputs (X_1 and X_2). The sub-network maps the input to the target space and then calculates the similarity distance between the two inputs in the target space. The mapping method of the sub-network is regarded as a function $G_W(X)$, which takes W as a parameter. The learning purpose of the entire SiamFC is to find the appropriate parameter W by training the sub-network. When the inputs X_1 and X_2 are classified as the same category, the calculated similarity is larger, and vice versa. The network framework of the SiamFC gives it unique advantages in

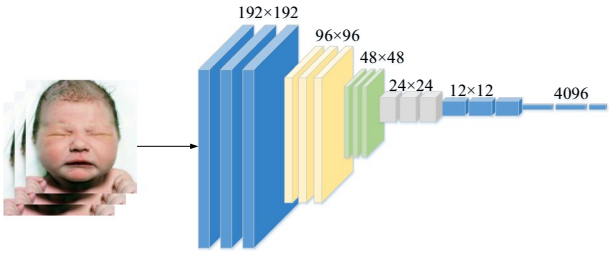

Fig. 2 Structure diagram of VGG-16 network

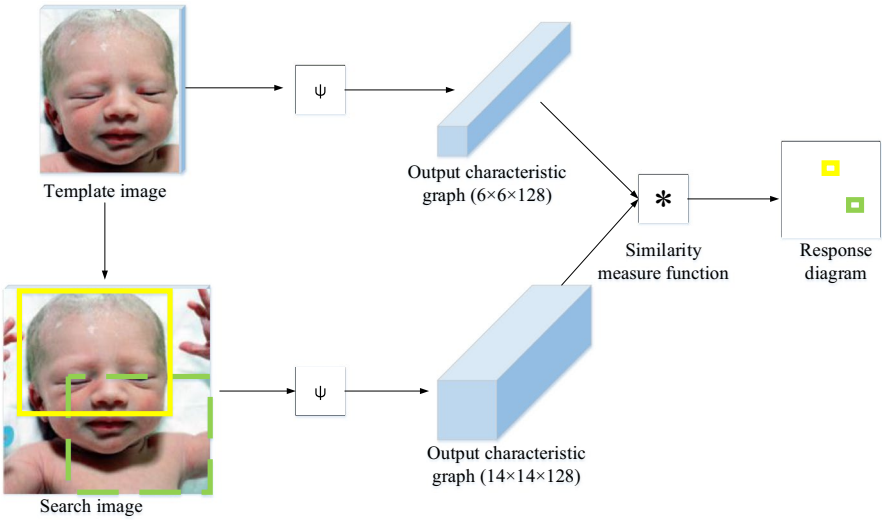

Fig. 3 Application of SiamFC in target tracking

solving similarity, and the sub-network sharing weights only require training fewer parameters, which means that the training data are less and it is not easy to suffer overfitting [32]. In SiamFC, the convolutional network is undertaken as the sub-network, which is trainable, multi-levelled, and nonlinear. In addition, it realizes the pixel-level processing and the learning of the shallow surface features and deep semantic representation of the image. The framework of SiamFC for tracking tasks is shown in Fig. 3.

The network mainly consists of two branches. The template branch extracts the features of target template, and detection branch extracts the features of subsequent frame images. The feature extraction network is a fully CNN, which can translate the input image without any change, so the network can receive a large input image and obtain more information about the target object and the background object. For a translation operation L_τ and an input image x, the translation operation of sub-region u of the image satisfies below equation:

$$L_\tau(x(u)) = x(u - \tau) \tag{1}$$

Then, if a CNN a parameter θ is a fully CNN, the mapping relationship of any translation operation τ in the process of mapping the input image x to the feature map ϕ_θ can be expressed as Eq. (2):

$$\phi_\theta(L_{n\tau}x) = L_\tau\phi_\theta(x) \tag{2}$$

where n refers to the total step length of the CNN. For SiamFC, the detection branch can provide a larger search image for input instead of the image with the same size as the template image and measure the similarity between each cyclically shifted

sub-window and the template image in a new frame, of which the similarity measurement is completed through the cross-correlation layer, and the similarity measurement relationship is given as follows:

$$g(z, x) = \phi_\theta(z) \cdot \phi_\theta(x) + b \tag{3}$$

where $\phi_\theta(z)$ and $\phi_\theta(x)$ refer to the feature map of the template image and the search image, respectively, and b refers to the offset term. $g(z, x)$ refers to the similarity score response graph of the two, and the template image and search image shift sub-window feature corresponding to each value are compared in terms of similarity, to obtain the position with the highest similarity score, which is the predicted position of the tracking target.

3.4 Construction of newborns FVIT model based on siamFC

There are some deficiencies for SiamFC when it is applied in target tracking. For example, the template branch is only performed in the first frame, so that the template features are unable to adapt to changes in the target. When the target changes significantly, the features from the first frame cannot characterize features. In addition, the network can only obtain the centre position of the tracked target, but its size can't be estimated, so additional techniques are required to solve the scale changes. Therefore, the initial SiamFC framework is improved, and attention mechanism is combined with the face tracking algorithm to track and recognize newborn facial features, to solve the above shortcomings and improve the tracking performance. Figure 4 discloses the framework of newborn FVIT model based on the SiamFC.

In the above algorithm model, a method is designed by combining the first frame and the previous frame of the current frame to update the target template in real

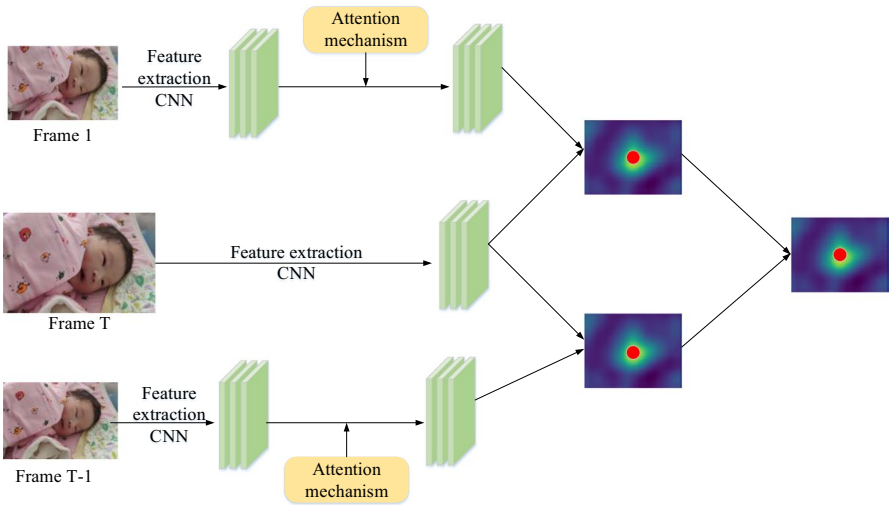

Fig. 4 Framework of newborn FVIT model based on SiamFC

time. At the same time, the discriminant and motion characteristics of the target are considered to avoid the template drift problem while updating the target template in real time. The input of the model is composed of three images, including the images in target template area cropped in the first frame, the target template area cropped in the previous frame (the T-1th frame) of the current frame, and the current frame (the Tth frame), which are represented by x_{first}, x_{latest}, and z, respectively. The two tracking branches of the model take (x_{first}, z) and (x_{latest}, z) as input, respectively. The feature of x_{first} can be expressed as $w_{first}.\varphi\left(x_{first}\right)$, and the feature of x_{latest} can be expressed as $w_{latest}.\varphi\left(x_{latest}\right)$, where the dimension of w_{first} and w_{latest} is the same as the channel number of the target template feature, and· refers to element-level multiplication.

The target template x_{first} of the first frame and the multi-layer convolution feature of the current frame z are, respectively, cross-correlated to obtain the corresponding position response graph of the predicted target, as shown in Eq. (4).

$$f\left(z, x_{first}\right) = corr\left(\varphi(z), w_{first}, \varphi\left(x_{first}\right)\right) \tag{4}$$

where $corr()$ refers to convolution cross-correlation operation. The feature extraction process of the current frame image z is performed only once. Similarly, the target template w_{latest} of the previous frame of the current frame and the multi-layer convolution feature of the current frame w_{latest} are, respectively, cross-correlated to obtain the corresponding position response graph of the predicted target, as shown in Eq. (5).

$$f\left(z, x_{latest}\right) = corr\left(\varphi(z), w_{latest}, \varphi\left(x_{latest}\right)\right) \tag{5}$$

Then, the response graph of the target position predicted in the current frame is weighted and averaged by the response graph of the two tracking branches, as shown in Eq. (6).

$$f\left(z, x_{first}, x_{latest}\right) = \lambda f\left(z, x_{first}\right) + (1 - \lambda) f\left(z, x_{latest}\right) \tag{6}$$

where λ denotes the hyperparameter that balances the importance of two trace branches. It is known from the experiment that when the λ is 0.3, the model achieves the optimal effect.

The T-1th frame is added as the template image in the template branch of the SiamFC, so it is no longer possible to simply compare the loss of the first frame and the current frame to optimize the SiamFC. Thus, the loss function of the neural network must be adjusted. SiamFC can cross-correlate the template feature $\phi_\theta(z)$ after the attention mechanism and the search feature $\phi_\theta(x)$ in the sliding window area of the search image to obtain the similarity score response graph g. Each position on the score map satisfies $u \in D$, and the positive and negative sample label y_u must be set as follows:

$$y_u = \begin{cases} 1 & \|u - c\| \le d \\ -1 & otherwise \end{cases} \tag{7}$$

In Eq. (4), c refers to the centre of the template feature map $\phi_\theta(z)$, and d refers to a pre-set distance threshold. When each sample in $\phi_\theta(x)$ is performed with the matching

operation with the sample $\phi_\theta(z)$ with the same size during the training process, it can be defined as a positive sample if the Euclidean distance between $\phi_\theta(x)$ of the position u and centre position c of $\phi_\theta(z)$ in each score map is less than d; otherwise, it is defined as a negative sample. The logistic loss used for each position of each score map in the sample is given as follows.

$$l(y, g) = \log(1 + e^{-yg}) \tag{8}$$

Then, the average of all scores is calculated as the loss function of each score map.

$$L(y, g) = \frac{1}{|D|} \sum_{u \in D} l(y_u, g_u) \tag{9}$$

When the training data set is adopted for offline training, the SiamFC parameter γ can be optimized with the stochastic gradient descent method for all score maps.

$$\arg\min_{\gamma} EL(y, g(z, x; \gamma)) \tag{10}$$

The feature extraction of multiple convolutional layers of the input image combines the apparent information and semantic information of the target. The feature extraction network used in this study is to load the pre-trained AlexNet framework on ImageNet. The input image is extracted through the AlexNet network to extract features, and then, the model extracts three types of features, which are conv3, conv4, and conv5 output features, to mainly express different levels of the image features. Finally, $\phi(\cdot)$ is allowed to refer to the integration of the extracted three-layer convolutional features. The convolutional feature extracted from x_{first} can be represented by $\phi(x_{first})$, the feature extracted from x_{latest} can be represented by $\phi(x_{latest})$, and the feature extracted from z can be represented by $\phi(z)$. Finally, combination of multi-layer convolutional features increases the diversity of features to a reliable extent and improves the robustness of the model.

The channel attention mechanism is combined with the multi-layer convolution feature of the target template, and the channel feature that have a larger impact on the tracking target is given a higher weight, to improve the discrimination of the target template features. During the combination, the model characterizes the widely combined target in multiple levels and learns the importance of different channel features in a more granular manner, which effectively improves the performance of newborn facial tracking.

3.5 Simulation experiment

The research objects are the facial video images of newborns, and the neural network involved refers to the CNN. From the speed and convenience of the framework, the TensorFlow [33] is selected. The experimental hardware environment includes a personal computer (PC) with the operating system of Ubuntu 14.04.4 and s equipped with a Linux kernel. In addition, the graphics processing unit (GPU) is

Table 1 Experimental configuration

	Name and version	
Hardware	Processor	Intel Broadwell E5-2650 v4 2.2G/32 M/Tray
	Display card	Nvidia GeForce GTX 1080Ti
	Storage	126 GB
	Disc	1.2TR SSD + 4TR SATA3
Software	Matrix transportation	Numpy1.12.6;Pandas 0.23.0
	Programming language	Python 3.2
	Development platform	TensorFlow 1.4.0
	System	Ubuntu 14.04.4

applied to accelerate the deep learning. The experimental configurations are shown in Table 1.

The database for newborn facial video image is based on the newborn facial tracking algorithm. At present, there are many public facial detection databases, such as facial detection database (FDDB) and CelebFaces Attributes Dataset (CelebA). All of them are samples of adult faces, and there are few data for the newborns. The facial features of newborns are greatly different from those of adults, and standardization of the database plays an extremely important role in the research of facial detection and tracking of newborns in the later stage. Thus, taking the establishment of a public adult face database as the standard, a more standard newborn image database with richer data is built. The original data of the database come from the pain expression video of newborns collected jointly by the research group and Hubei hospital. About 86 segments of facial expression changes of newborns under different external stimuli were collected from 40 newborns (20 boys and 20 girls) with the age of no more than 7 days. Then, the collected videos are pre-processed, including extraction of video key frame and selection of the key frame images with angle changes, background changes, and facial state changes of the newborns. In addition, coordinates for facial area of the newborns are calibrated in the key frame images, including the classification label of the facial area and the coordinates of the upper left vertex and the lower right vertex of the face position in the image. Finally, the image with the marked coordinate position information is amplified and normalized to complete the establishment of the database.

The proposed newborn FVIT model was compared with the algorithms with good results (which are widely adopted in recent years) to verify its performance. The algorithms for comparison include efficient convolution operators (ECO) [34], SiamFC [35], correlation filter network (CFNet) [36], Staple [37], and kernel correlation filter (KCF) [38]. The comparison results were obtained from other related studies. The one pass evaluation (OPE) was undertaken as the evaluation standard, including two evaluation curves: precision plot and success plot.

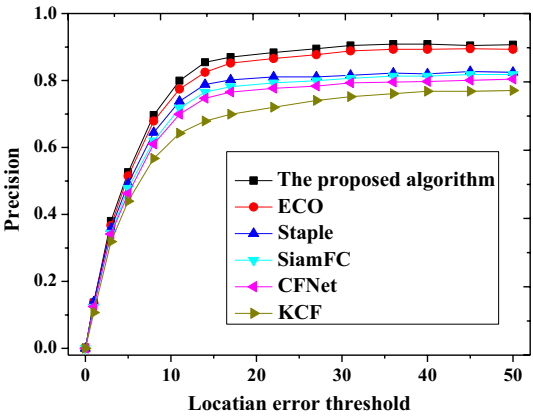

Fig. 5 Precision curve evaluated by OPE

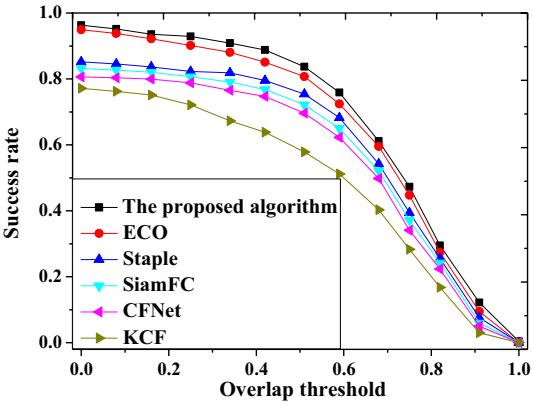

Fig. 6 Success rate curve evaluated by OPE

4 Results and discussion

4.1 Performance analysis of the system model compared with different algorithms

The proposed tracking algorithm is compared with ECO, SiamFC, CFNet, Staple, and KCF algorithms. Figures 5 and 6 show the precision and success rate evaluated by the OPE.

The precision analysis shows that the precisions of the proposed algorithm, ECO, Staple, SiamFC, CFNet, and KCF are 0.889, 0.863, 0.782, 0.778, 0.743, and 0.707, respectively. The algorithm of this study shows the best precision (as shown in Fig. 5). The analysis of the success rate discloses that the AUC values of the proposed tracking algorithm, ECO, Staple, SiamFC, CFNet, and KCF are 0.748, 0.673, 0.592, 0.587, 0.574, and 0.479, respectively. The success rate of the tracking algorithm in this study is higher (as illustrated in Fig. 6). The tracker based on CFNet slows down after the deep learning features are adopted, so that its original

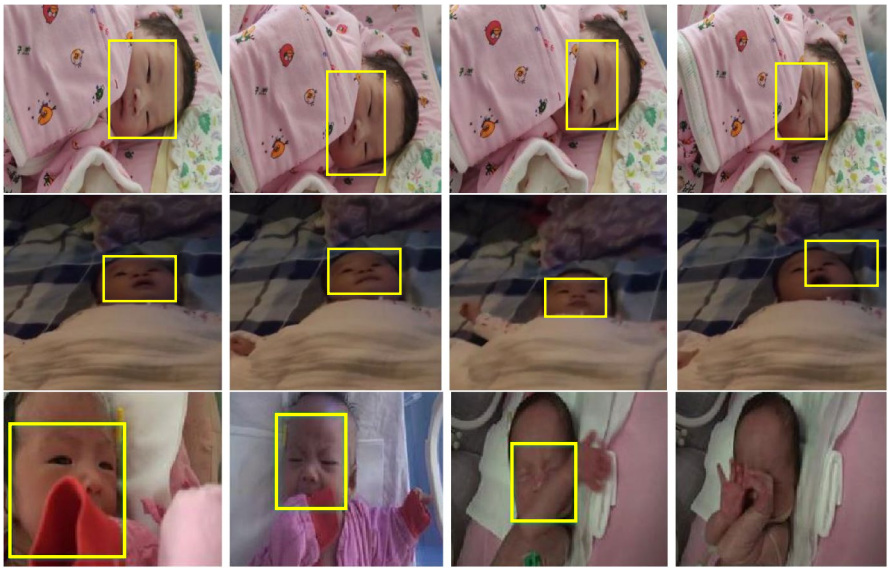

Fig. 7 The specific tracking effect based on improved SiamFC

advantage in speed is weakened. The method based on the SiamFC also shows a very favourable tracking speed and a good tracking precision. In addition, its end-to-end training method successfully utilizes large video sequence data sets to enable the network model to capture important information, thereby effectively improving the tracking precision. Therefore, evaluation and analysis based on the OPE standard prove that the improved tracking algorithm has achieved good performance on both the centring error and zone overlapping standards.

4.2 Analysis on facial tracking effect of the constructed model

The optimized SiamFC was applied for facial tracking in videos of newborns further. The specific tracking effect is shown in Fig. 7.

Figure 7 indicates that the tracker adopted in this study has a good tracking effect on facial expression changes, scale changes, and partially occluded face of newborns. However, SiamFC target tracking achieves target tracking through offline similarity learning, so the model cannot be updated online. Thus, the tracking frame will shift and the border position is not precise enough if there is another object similar to the target in the background. The improved target tracking algorithm can better adapt to the facial occlusion, facial expression changes, and scale transformation of newborns, so it can meet the real-time requirements.

4.3 Discussion

The performance of the proposed model algorithm was analysed through simulation experiments. Firstly, the newborn image database used in the experiment was introduced in detail, including the establishment process of the newborn image database and the related steps of the pre-treatment operation. At the same time, the evaluation standard of newborns face detection was expounded. Then, face detection and tracking experiments were carried out on the database of newborn images. The model algorithm proposed in this study was compared with the model algorithm of scholars in related fields. According to the characteristics of newborn facial features, a detection method suitable for newborns face was proposed to verify the effectiveness of using full-convolution twin network for newborns face tracking. This algorithm has strong robustness and can meet the real-time requirements of video sequences compared with other algorithms, highlighting the advantages of the algorithm constructed in this study in the recognition of newborn facial images.

5 Conclusion

The real-time pain evaluation system for newborns enables clinical medical staff to effectively evaluate the pain suffered by the newborns, and the face detection and tracking are the premise and foundation of the pain evaluation system. In this study, the SiamFC algorithm was improved and integrated with the attention mechanism, and a newborn FVIT model was constructed based on SiamFC. Simulation experiment reveals that the improved algorithm shows better precision and success rate as well as satisfied effect of capturing and tracking the facial images of newborns. Therefore, this study provides an experimental reference for the facial recognition and pain assessment of the newborns in the later period. However, there are some shortcomings in this study. For example, the amount of data in the newborn image database is not very large, and it must amplify the image on the data set during the experiment. At present, there is no standard for newborns face in the world, so it lacks clear and objective evaluation criteria for the effect of facial detection of newborns. Therefore, in the follow-up research, the algorithm can be further optimized based on the model algorithm built by this research firstly to improve its real-time performance and recognize the emotions of newborns in real time. Secondly, it can combine with other auxiliary information such as crying and body movements for multi-modal classification and recognition when the facial expressions of newborns are collected to build a standard newborn image database with high quality and large data volume. The research results of this study provide important practical significance for the related exploration of facial emotion recognition of newborns in future.

Acknowledgements This research was supported by the following projects: 1. Research on Publicity Channels of Traditional Chinese Medicine Culture in Primary and Middle Schools in Ethnic Minority Areas, a Project of Collaborative Development and Research Center for Sichuan Traditional Chinese Medicine Culture, Project No. ZYYWH1813. 2. Research on Home Protection Methods of Multi-Dimensional Linkage for Tibetan and Yi Infants in Major Public Health Emergencies—Exemplified by

COVID-19 Pandemic, a Project of Sichuan 0-3 Years Old Infants' Early Development and Education Research Center, Project No. SCLS20-13. 3. Research on AIDS Prevention Publicity Channels for Medical Students of Yi Ethnic Group to Serve the Hometown, a project of Sichuan Sex Sociology and Sex Education Research Center, Project No. SXJYB1927.

References

1. Dang LM, Hassan SI, Im S et al (2019) Face image manipulation detection based on a convolutional neural network. Expert Syst Appl 129:156–168. https://doi.org/10.1016/j.eswa.2019.04.005
2. Deffo LL, Fute ET, Tonye E (2018) CNNSFR: a convolutional neural network system for face detection and recognition. Int J Adv Computer Sci Appl 9(12):240–244. https://doi.org/10.14569/IJACSA.2018.091235
3. Brumancia E, Samuel SJ, Gladence LM et al (2019) Hybrid data fusion model for restricted information using Dempster-Shafer and adaptive neuro-fuzzy inference (DSANFI) system. Soft Comput 23(8):2637–2644. https://doi.org/10.1007/s00500-018-03734-1
4. Kusiak A (2020) Convolutional and generative adversarial neural networks in manufacturing. Int J Prod Res 58(5):1594–1604. https://doi.org/10.1080/00207543.2019.1662133
5. Chen J, Lv Y, Xu R et al (2019) Automatic social signal analysis: Facial expression recognition using difference convolution neural network. J Parallel Distrib Comput 131:97–102. https://doi.org/10.1016/j.jpdc.2019.04.017
6. Islas MA, Rubio JJ, Muñiz S et al (2021) A fuzzy logic model for hourly electrical power demand modeling. Electronics 10(4):448. https://doi.org/10.3390/electronics10040448
7. de Jesús RJ, Lughofer E, Pieper J et al (2021) Adapting H-infinity controller for the desired reference tracking of the sphere position in the maglev process. Inf Sci 569:669–686. https://doi.org/10.1016/j.ins.2021.05.018
8. Chiang HS, Chen MY, Huang YJ (2019) Wavelet-based EEG processing for epilepsy detection using fuzzy entropy and associative petri net. IEEE Access 7:103255–103262. https://doi.org/10.1109/ACCESS.2019.2929266
9. de Rubio JJ (2020) Stability analysis of the modified Levenberg-Marquardt algorithm for the artificial neural network training. IEEE Trans Neural Netw Learn Syst 32(8):3510–3524. https://doi.org/10.1109/TNNLS.2020.3015200
10. Meda-Campaña JA (2018) On the estimation and control of nonlinear systems with parametric uncertainties and noisy outputs. IEEE Access 6:31968–31973. https://doi.org/10.1109/ACCESS.2018.2846483
11. Soriano LA, Zamora E, Vazquez-Nicolas JM et al (2020) PD control compensation based on a cascade neural network applied to a robot manipulator. Front Neurorobot 14:577749. https://doi.org/10.3389/fnbot.2020.577749
12. Al-Janabi S, Alkaim AF, Adel Z (2020) An Innovative synthesis of deep learning techniques (DCapsNet & DCOM) for generation electrical renewable energy from wind energy. Soft Comput 24(14):10943–10962. https://doi.org/10.1007/s00500-020-04905-9
13. Wang C, Han D, Liu Q et al (2018) A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. IEEE Access 7:2161–2168. https://doi.org/10.1109/ACCESS.2018.2887138
14. Al-Janabi S, Salman AH (2021) Sensitive integration of multilevel optimization model in human activity recognition for smartphone and smartwatch applications. Big Data Mining Anal 4(2):124–138. https://doi.org/10.1007/978-3-030-23672-4_23
15. Al-Janabi S, Alkaim A, Al-Janabi E et al (2021) Intelligent forecaster of concentrations (PM2. 5, PM10, NO2, CO, O3, SO2) caused air pollution (IFCsAP). Neural Comput Appl. https://doi.org/10.1007/s00521-021-06067-7
16. Al-Janabi S, Mohammad M, Al-Sultan A (2020) A new method for prediction of air pollution based on intelligent computation. Soft Comput 24(1):661–680. https://doi.org/10.1007/s00500-019-04495-1

17. Al-Janabi S, Al-Shourbaji I (2016) A hybrid image steganography method based on genetic algorithm. In: 2016 7th international conference on sciences of electronics, technologies of information and telecommunications (SETIT). IEEE, pp. 398–404. https://doi.org/10.1109/SETIT.2016.7939903
18. Omer Y, Sapir R, Hatuka Y et al (2019) What is a face? Critical Features Face Detect Percep 48(5):437–446. https://doi.org/10.1177/0301006619838734
19. Al-Janabi S, Alkaim AF (2020) A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation[J]. Soft Comput 24(1):555–569. https://doi.org/10.1007/s00500-019-03972-x
20. Al-Janabi S, Al-Shourbaji I (2016) A smart and effective method for digital video compression. In: 2016 7th international conference on sciences of electronics, technologies of information and telecommunications (SETIT). IEEE, pp. 532–538. https://doi.org/10.1109/SETIT.2016.7939927
21. Chrysos GG, Antonakos E, Snape P et al (2018) A comprehensive performance evaluation of deformable face tracking "in-the-wild." Int J Comput Vision 126(2–4):198–232. https://doi.org/10.1007/s11263-017-0999-5
22. Sonkusare S, Ahmedt-Aristizabal D, Aburn MJ et al (2019) Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking. Sci Rep 9(1):1–11. https://doi.org/10.1038/s41598-019-41172-7
23. Low CC, Ong LY, Koo VC et al (2020) Multi-audience tracking with RGB-D camera on digital signage. Heliyon 6(9):e05107. https://doi.org/10.1016/j.heliyon.2020.e05107
24. Yang A, Yang X, Wu W et al (2019) Research on feature extraction of tumor image based on convolutional neural network. IEEE Access 7:24204–24213. https://doi.org/10.1109/ACCESS.2019.2897131
25. Rajan AP, Mathew AR (2019) Evaluation and applying feature extraction techniques for face detection and recognition. Indonesian J Elect Eng Inform (IJEEI) 7(4):742–749. https://doi.org/10.52549/ijeei.v7i4.935
26. Tao X, Zhang D, Ma W et al (2018) Automatic metallic surface defect detection and recognition with convolutional neural networks. Appl Sci 8(9):1575. https://doi.org/10.3390/app8091575
27. Jangid M, Srivastava S (2018) Handwritten devanagari character recognition using layer-wise training of deep convolutional neural networks and adaptive gradient methods. J Imaging 4(2):41. https://doi.org/10.3390/jimaging4020041
28. Yuan F, Zhang L, Wan B et al (2019) Convolutional neural networks based on multi-scale additive merging layers for visual smoke recognition. Mach Vis Appl 30(2):345–358. https://doi.org/10.1007/s00138-018-0990-3
29. Ashwin TS, Guddeti RMR (2020) Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. Educ Inf Technol 25(2):1387–1415. https://doi.org/10.1007/s10639-019-10004-6
30. Saeedimoghaddam M, Stepinski TF (2020) Automatic extraction of road intersection points from USGS historical map series using deep convolutional neural networks. Int J Geogr Inf Sci 34(5):947–968. https://doi.org/10.1080/13658816.2019.1696968
31. Jumani SZ, Ali F, Guriro S et al (2019) Facial expression recognition with histogram of oriented gradients using CNN. Indian J Sci Technol 12(24):1–8. https://doi.org/10.17485/ijst/2019/v12i24/145093
32. Achour B, Belkadi M, Filali I et al (2020) Image analysis for individual identification and feeding behaviour monitoring of dairy cows based on Convolutional Neural Networks (CNN). Biosys Eng 198:31–49. https://doi.org/10.1016/j.biosystemseng.2020.07.019
33. Rauber J, Zimmermann R, Bethge M et al (2020) Foolbox Native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX. J Open Source Softw 5(53):2607. https://doi.org/10.21105/joss.02607
34. Bendjillali RI, Beladgham M, Merit K et al (2019) Improved facial expression recognition based on DWT feature for deep CNN. Electronics 8(3):324. https://doi.org/10.3390/electronics8030324
35. Zhu R, Gong X, Hu S et al (2019) Power quality disturbances classification via fully-convolutional Siamese network and k-nearest neighbor. Energies 12(24):4732. https://doi.org/10.3390/en12244732
36. Yang L, Jiang P, Wang F et al (2018) Robust real-time visual object tracking via multi-scale full-convolution Siamese networks. Multimed Tools Appl 77(17):22131–22143. https://doi.org/10.1007/s11042-018-5664-7

37. Li D, Yu Y, Chen X (2019) Object tracking framework with Siamese network and re-detection mechanism. EURASIP J Wirel Commun Netw 2019(1):261. https://doi.org/10.1186/s13638-019-1579-x
38. Nguyen TL, Han DY (2020) Detection of road surface changes from multi-temporal unmanned aerial vehicle images using a convolutional Siamese network. Sustainability 12(6):2482. https://doi.org/10.3390/su12062482

Authors and Affiliations

Yun Wang[1] · Lu Huang[2] · Austin Lin Yee[3]

Yun Wang
shxdx_wy168@163.com

Lu Huang
huanglu21@mails.ucas.ac.cn

[1] Department of Computer Engineering, Shanxi Polytechnic College, Taiyuan 030006, China

[2] Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100029, China

[3] Department of Oral Biology, Division of Orthodontics, Harvard School of Dental Medicine, Harvard University, Boston 02115, USA