Title: Subject independent facial expression recognition with robust face detection using a convolutional neural network

Outline

Neural Networks
Volume 16, Issues 5?6, June?July 2003, Pages 555-559
#
2003 Special issue
Subject independent facial expression recognition with robust face detection
using a convolutional neural network
Author links open overlay panelMasakazu Matsugu, Katsuhiko Mori, Yusuke
Mitari, Yuji Kaneda
Show more

https://doi.org/10.1016/S0893-6080(03)00115-1Get rights and content
Abstract
Reliable detection of ordinary facial expressions (e.g. smile) despite the
variability among individuals as well as face appearance is an important step
toward the realization of perceptual user interface with autonomous perception
of persons. We describe a rule-based algorithm for robust facial expression
recognition combined with robust face detection using a convolutional neural
network. In this study, we address the problem of subject independence as well
as translation, rotation, and scale invariance in the recognition of facial
expression. The result shows reliable detection of smiles with recognition
rate of 97.6% for 5600 still images of more than 10 subjects. The proposed
algorithm demonstrated the ability to discriminate smiling from talking based

on the saliency score obtained from voting visual cues. To the best of our knowledge, it is the first facial expression recognition model with the property of subject independence combined with robustness to variability in facial appearance.

* Previous article in issue
* Next article in issue

Keywords
Convolutional neural network
Face detection
Facial expression

1\. Introduction

Facial expressions as manifestations of emotional states, in general, tend to be different among individuals. For example, smiling face as it appears may have different emotional implications for different persons in that ?smiling face?, perceived by others, for some person does not necessarily represent truly _smiling state_ for that person. To the best of our knowledge, only a few algorithms (Ebine & Nakamura, 1999) have addressed robustness to such individuality in facial expression recognition. Furthermore, in order for facial expression recognition to be used for human?computer interaction, for example, that algorithm must have good ability in dealing with variability of facial appearance (e.g. pose, size, and translation invariance). Most algorithms, so far, have addressed only a part of these problems (Földiák, 1991, Wallis and Rolls, 1997). In this study, we propose a system for facial expression recognition that is robust to variability that originates from individuality and viewing conditions.

Recognizing facial expression under rigid head movements was addressed by Black and Yacoob (1995). Neural network model that learns to recognize facial expressions from an optical flow field was reported in Rosenblum, Yacoob, and Davis (1996). Rule-based systems were reported in Black and Yacoob, 1997, Yacoob and Davis, 1996, in which primary facial features were tracked throughout the image sequence.

Convolutional neural network (CNN) models (Le Cun & Bengio, 1995) as well as neocognitrons (Fukushima, 1980) known as one of biologically inspired models, have been used for pattern recognition tasks such as face recognition and hand-written numeral recognition. The network includes feature detecting (FD) layers, each of which alternated with a sub-sampling layer (pooling layer) to obtain properties leading to translation and deformation invariance. Recently, Fasel (2002) has proposed a model with two independent CNNs, one for facial expression the other for face identity recognition, which are combined by a MLP.

The proposed model in this study turns out to be much more efficient and compact than Fasel's model. In addition, our model comes with the property of subject independence. Specifically, the proposed system can detect smiling or laughing faces based on difference in local features between a normal face and those not.

In Section 2, we introduce a modular convolutional network architecture for robust face detection and facial expression recognition involving module-based learning based on a variant of BP. In Section 3, we propose a rule-based algorithm that utilizes differences of specific local features, extracted in the CNN, between neutral and emotional faces. We show that the proposed scheme attains not only subject independence but also position independence in facial expression recognition. In Section 4, we discuss the properties of proposed scheme that adds to the conventional neural networks for facial expression recognition.

2. Robust face detection using CNN

As in the previously proposed model (Matsugu, Mori, Ishii, & Mitarai, 2002), internal representation of face is provided by a hierarchically ordered set of convolutional kernels defined by the local receptive field of FD neurons. Face model is represented as a spatially ordered set of local features of intermediate complexity, such as eyes, mouth, nose, eyebrow, cheek, or else, and all of these features are represented in terms of a fixed set of lower and intermediate features.

The lower and intermediate features constitute some form of a fixed set of figural alphabets in our CNN. Corresponding receptive fields for the detection of these alphabetical features are learned in advance to form a local template in the hierarchical network, and once learned, they would never be changed during possible learning phase for object recognition in upper layers.

Our CNN model is different from the original model (Le Cun and Bengio, 1995) in three ways. First, training of the proposed model proceeds module by module (i.e. for each local feature class) only for FD _k_ (_k_ >1) layers. Second, we do not train FP (or sub-sampling) layers (FP neurons perform either maximum value detection or local averaging in their receptive fields). Third, we use a detection result of skin color area as input to the face detection module in FD4. The skin area is obtained simply by thresholding hue data of input image in the range of [?0.078,0.255] for the full range of [?0.5,0.5].

The training proceeds as follows. In the first step of training, two layers from the bottom, namely FD1 with eight modules and FD2 with four modules, are trained using standard back-propagation with intermediate local features (e.g. eye corners) as positive training data sets. Negative examples that do not constitute the corresponding feature category are also used as false data. Specifically, we trained the FD2 layer, the second from the bottom FD layer to form detectors of intermediate features, such as end-stop structures or blobs (i.e. end-stop structures for left and right side) and two types of horizontally elongated blobs (e.g. upper part bright, lower part bright) with varying size and rotation (up to 30° with rotation in-plane axis as well as head axis). These data used for training are fragments extracted from face images (Fig. 1).

1. Download: Download full-size image

Fig. 1. Convolutional architecture (feature pooling layers are not shown for simplified illustration) for face detection.

More complex local feature detectors (e.g. eye, mouth detectors, but not restricted to these) are trained in the third (FD3) or fourth (FD4) layer using the patterns extracted from face images with geometric transform as in the FD2 layer. As a result of these training sequences, the top FD layer, FD4, learns to locate faces in complex scenes.

As given in Fig. 2, information concerning locations of eyes and mouth detected by the CNN is fed to the rule-based processing module so that our facial analysis system can deal with variability in position, size, and pose.

1. Download: Download full-size image

Fig. 2. CNN with feedback mechanism for rule-based analysis.

3. Rule-based analysis of facial expressions using local features detected by CNN

In the following, we propose a rule-based processing scheme to enhance subject independence in facial expression recognition. We found that some of lower level features extracted by the first FD layer of CNN were useful for facial expression recognition. Primary features used in the analysis are horizontal line segments and edge-like structures similar to step and roof edges (extracted by the two modules in FP1 layer circled in Fig. 3) representing parts of eyes, mouth, and eyebrows. For example, changes in distance between end-stops (e.g. left-corner of left eye and left side end-stop of mouth) within facial components and changes in width of line segments in lower part of eyes or cheeks are detected to obtain saliency scores of a specific facial expression. Primary cues related to facial actions adopted in our rule-based facial analysis for the detection of smiling/laughing faces are as follows.

* (1)
Distance between endpoints of eye and mouth gets _shorter_ (lip being raised)
* (2)
Length of horizontal line segment in mouth gets _longer_ (lip being stretched)
* (3)
Length of line segments in eye gets _longer_ (wrinkle around the tail of eye gets longer)
* (4)
Gradient of line segment connecting the mid point and endpoint of mouth gets _steeper_ (lip being raised)
* (5)
No. of step-edges in mouth get _increased_ (teeth get appeared)
* (6)
No. of edges in cheeks _increased_ (wrinkle around cheeks gets grown).

1. Download: Download full-size image

Fig. 3. Face detection by proposed convolutional NN with the results of intermediate feature detection.

Each cue was scored based on the degree of positive changes (i.e. designated changes as given above) to the emotional state (e.g. happiness). For example, given a positive change in some cue by _a_ % to the feature value in neutral face, then we give _Ca_ points (_C_ is some constant) to that feature. Saliency score of specific emotional state is calculated using a sort of voting scheme with the summation (with weight _p_ _k_) of scores _s_ _k_ for respective cues given as followsS=?kpkSkAfter the voting process, the score _S_ is normalized and thresholded for judging the facial expression (e.g. smiling/laughing or not).

4\. Results

4.1. Face detection

In the training of our CNN, the number of facial fragment images used is 2900 for the FD2 layer, 5290 for the FD3, and 14,700 (face) for the FD4 layer, respectively. The number of non-face images, also used for the FD4 layer, is 137. Apparently greater number of facial component images as compared with non-face images is used to ensure robustness to varying rotation, size, contrast, and color balance of face images. So, we used a fixed set of transformed images for a given training sample image.

In particular, we used three different sizes of the fragment images ranging from 0.7 to 1.5 times the original image (i.e. fragments of facial components for modules in FD2 and FD3, entire face without background for FD4). The performance for face images other than the training data set was tested for over 200 face images with cluttered background and varying image-capturing conditions.

As shown in Fig. 4, the tested images of face are of different size (from 30×30 to 240×240 in VGA image), pose (up to 30° in head axis rotation and also in-plane rotation), and varying contrasts with cluttered background. The convolutional network demonstrated robust face detection with 1% false rejection rate and 6% false acceptance rate with quite good generalization ability.

1. Download: Download full-size image

Fig. 4. Face detection results in complex scenes.
4.2. Facial expression recognition
Fig. 5 shows a sequence of normalized saliency scores indicating successful detection of smiling faces, around the frame number from 130 to 170, with an appropriate threshold level of 2. The normalization was done by dividing the weighted sum of scores by 10 and subtracting some constant. The weight for each cue was assigned based on individuality factor, tendency of subject dependence. So, cues of larger individuality, found heuristically in advance, was assigned with some smaller weight values (e.g. 40), whereas cues of less individuality was assigned with larger weights (e.g. 45).
 1. Download: Download full-size image
Fig. 5. Normalized saliency score subtracted by constant value for smiling face detection.
As shown in Fig. 6, the proposed system demonstrated the discrimination of smiling from talking state for different persons based on the normalized saliency score. In addition, we obtained results demonstrating reliable detection of smiles with the recognition rate of 97.6% for 5600 still images of more than 10 subjects.
 1. Download: Download full-size image
Fig. 6. Talking faces can be discriminated from laughing face. Note that the second face obtained higher score due to similarity to laughing face.
5\. Discussion
Robust face detection with invariance properties in terms of translation, scale, and pose, inherent in our non-spiking version of CNN model (Matsugu, 2001, Matsugu et al., 2002) brings robustness to dynamical changes both in head movements and in facial expressions. In particular, because of the topographic property of our network preserving positional information of respective facial features from the bottom to top layers, the translation invariance in facial expression recognition is inherent in our convolutional architecture.
The feedback mechanism from the top to intermediate modules is used to identify corresponding facial features between neutral and emotional states. Specifically, face location as detected by FD4 was used to confine the search area of eye and mouth, and those intermediate facial features as detected in FD3 layer were utilized for tracking primitive local features extracted by the bottom layer FD1, which turned out to be useful for facial expression recognition. Location information of eyes and mouth detected in the CNN are thus transmitted, through the feedback loop from the FP3 layer to the rule-based processing module, to confine the processing area of facial feature analysis in the image based on differences in terms of at least six cues shown in Section 3.
A great number of approaches (see Donato, Bartlett, Hager, Ekman, & Sejnowski, 1999 for review) have been taken for robustness in facial expression recognition. However, most existing models do not have all of those properties for robustness stated in the above, and many of them require explicit estimation of motion parameters.
In contrast to a number of conventional methods, the proposed model for facial expression recognition is much more compact and efficient with the following distinct aspects.
 * (1)
Our model requires only differences in local features, extracted by the CNN, between a neutral face and emotional faces, instead of using differences between adjacent frames.
 * (2)

For the facial analysis, we need only a single set of local features from one neutral face as reference data.
 * (3)
Unlike many other approaches, the proposed system requires no explicit estimation of motion parameters over image sequences.
Additionally important feature of our model in terms of compactness is that we need a single system of CNN, whereas the Fasel's model requires two CNNs in tandem to obtain subject dependent facial expression recognition.
Conversely, it turned out that the proposed system is quite insensitive to individuality of facial expressions mainly by virtue of the rule-based facial analysis. Those cues exploited in the analysis undergo voting process with weighted summation of scores for respective cues in terms of differences of facial features in neutral and emotional states. As a result of this, voting individuality is averaged out to obtain subject independence. In practice, this subject independence is quite preferable since we can dispense with a large database that contains individual facial expressions. However, Fasel's model was not endowed with such property. We believe that it is easy to extend the current framework to other tasks for subject independent facial expression recognition. Although our rule-based facial analysis is useful as it is, we may incorporate fuzzy rules to obtain more robust performance.
In conclusion, our model is the first facial expression recognition system with subject independence combined with robustness with regard to variability in face images in terms of appearance and location.
Special issue articlesRecommended articles
References
 1. Black and Yacoob, 1995
M. Black, Y. Yacoob
Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion
Proceedings of IEEE Fifth International Conference on Computer Vision (1995), pp. 374-381
View in ScopusGoogle Scholar
 2. Black and Yacoob, 1997
M. Black, Y. Yacoob
Recognizing facial expressions in image sequences using local parameterized models of image motion
International Journal of Computer Vision, 25 (1997), pp. 23-48
View in ScopusGoogle Scholar
 3. Donato et al., 1999
G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, T. Sejnowski
Classifying facial actions
IEEE Transactions on Pattern Analysis and Machine Intelligence, 21 (1999), pp. 974-989
Google Scholar
 4. Ebine and Nakamura, 1999
H. Ebine, O. Nakamura
The recognition of facial expressions considering the difference between individuality
Transaction on IEE of Japan, 119-C (1999), pp. 474-481
(In Japanese)
CrossrefGoogle Scholar
 5. Fasel, 2002
B. Fasel
Robust face analysis using convolutional neural networks

Proceedings of International Conference on Pattern (2002)
Google Scholar
6. Földiák, 1991
P. Földiák
Learning invariance from transformation sequences
Neural Computation, 3 (1991), pp. 194-200
CrossrefGoogle Scholar
7. Fukushima, 1980
K. Fukushima
Neocognitron: a self-organizing neural networks for a mechanism of pattern recognition unaffected by shift in position
Biological Cybernetics, 36 (1980), pp. 193-202
Google Scholar
8. Le Cun and Bengio, 1995
Y. Le Cun, T. Bengio
Convolutional networks for images, speech, and time series
M.A. Arbib (Ed.), The handbook of brain theory and neural networks, MIT Press, Cambridge, MA (1995), pp. 255-258
Google Scholar
9. Matsugu, 2001
M. Matsugu
Hierarchical pulse-coupled neural network model with temporal coding and emergent feature binding mechanism
Proceedings of International Joint Conference on Neural Networks (2001), pp. 802-807
View in ScopusGoogle Scholar
10. Matsugu et al., 2002
M. Matsugu, K. Mori, M. Ishii, Y. Mitarai
Convolutional spiking neural network model for robust face detection
Proceedings of the Ninth International Conference on Neural Information Processing, Singapore (2002), pp. 660-664
View in ScopusGoogle Scholar
11. Rosenblum et al., 1996
M. Rosenblum, Y. Yacoob, L.S. Davis
Human expression recognition from motion using a radial basis function network architecture
IEEE Transactions on Neural Networks, 7 (1996), pp. 1121-1138
View in ScopusGoogle Scholar
12. Wallis and Rolls, 1997
G. Wallis, E.T. Rolls
Invariant face and object recognition in the visual system
Progress in Neurobiology, 51 (1997), pp. 167-194
View PDFView articleView in ScopusGoogle Scholar
13. Yacoob and Davis, 1996
Y. Yacoob, L.S. Davis
Recognizing human facial expression from long image sequences using optical flow
IEEE Transaction on Pattern Analysis and Machine Intelligence, 18 (1996), pp. 636-642
View in ScopusGoogle Scholar

Advances in biological and medical technologies have been providing us explosive volumes of biological and physiological data, such as **medical images** , electroencephalography, genomic and protein sequences. Learning from these data facilitates the understanding of human health and disease. Developed from artificial neural networks, **deep learning** -based algorithms show great promise in extracting features and learning patterns from complex data. The aim of this paper is to provide an overview of deep learning techniques and some of the state-of-the-art applications in the biomedical field. We first introduce the development of artificial neural network and deep learning. We then describe two main components of deep learning, _i.e._ , deep learning architectures and model optimization. Subsequently, some examples are demonstrated for deep learning applications, including medical image classification, genomic sequence analysis, as well as protein structure classification and prediction. Finally, we offer our perspectives for the future directions in the field of deep learning.

* ### Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order

2017, Pattern Recognition

Show abstract

Facial expression recognition has been an active research area in the past 10 years, with growing application areas including avatar animation, neuromarketing and sociable robots. The recognition of facial expressions is not an easy problem for machine learning methods, since people can vary significantly in the way they show their expressions. Even images of the same person in the same facial expression can vary in brightness, background and pose, and these variations are emphasized if considering different subjects (because of variations in shape, ethnicity among others). Although facial expression recognition is very studied in the literature, few works perform fair evaluation avoiding mixing subjects while training and testing the proposed algorithms. Hence, facial expression recognition is still a challenging problem in computer vision. In this work, we propose a simple solution for facial expression recognition that uses a combination of Convolutional Neural Network and specific image pre-processing steps. Convolutional Neural Networks achieve better accuracy with big data. However, there are no publicly available datasets with sufficient data for facial expression recognition with deep architectures. Therefore, to tackle the problem, we apply some pre-processing techniques to extract only expression specific features from a face image and explore the presentation order of the samples during training. The experiments employed to evaluate our technique were carried out using three largely used public databases (CK+, JAFFE and BU-3DFE). A study of the impact of each image pre-processing operation in the accuracy rate is presented. The proposed method: achieves competitive results when compared with other facial expression recognition methods ? 96.76% of accuracy in the CK+ database ? it is fast to train, and it allows for real time facial expression recognition with standard computers.

* ### Classification of COVID-19 patients from chest CT images using multi-objective differential evolution?based convolutional neural networks

2020, European Journal of Clinical Microbiology and Infectious Diseases

* ### Multi-Classification of Brain Tumor Images Using Deep Neural Network

2019, IEEE Access

* ### Gearbox Fault Identification and Classification with Convolutional Neural Networks

2015, Shock and Vibration

* ### Facial expression recognition in image sequences using geometric deformation features and support vector machines

2007, IEEE Transactions on Image Processing
View all citing articles on Scopus
View Abstract

Part of special issue
Advances in Neural Networks Research: IJCNN '03
Edited by
Donald Wunsch, Michael Hasselmo, Kumar Venayagamoorthy and DeLiang Wang
Download full issue
Other articles from this issue
* ### Adaptive force generation for precision-grip lifting by a spectral timing model of the cerebellum
June?July 2003
Antonio Ulloa, ?, Bradley J. Rhodes
View PDF
* ### Radial basis function neural networks for nonlinear Fisher discrimination and Neyman?Pearson classification
June?July 2003
David Casasent, Xue-wen Chen
View PDF
* ### Intrinsic generalization analysis of low dimensional representations
June?July 2003
Xiuwen Liu, ?, DeLiang Wang
View PDF
View more articles
Recommended articles
No articles found.