AUTOMATIC EMOTION RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

P. Senthilkumar

Department of Instrumentation and Control Engineering, Kalasalingam Academy of Research and Education, India

Abstract

This paper presents a local appearance feature fusion for automatic emotion recognition using Convolutional Neural Network (CNN). The CNN has been known to be a powerful texture feature for facial expression recognition. However, only few approaches utilize the relationship among neighborhood pixels itself. First, CNN is obtained based on two closest vertical and/or horizontal neighborhood pixel relationships. The proposed work is also extended to efficiently handle a large amount of unlabeled data using supervised classification algorithm using modified CNN. At the last stage, ensemble classifiers are trained with a small percentage labeled data and based on the trained model, rest of the unlabeled data is assigned with pseudo-labels.

Keywords:

Cloud Computing, Reliability Assessment, Trust Proof, Personal Opinion, Internet of Things

1. INTRODUCTION

In human interaction, facial intuition creates communication channel and voice that provides vital proof of the person's internal emotional state. The face is a primary communication interface, as reported by [1], which provides multiple messages including emotion, age, intelligence, etc.

Facial expression is a key behavioral action and has become a very important issue in the study of emotions, cognitive processes and social interactions based on computational vision. A range of methods for Automatic Facial Expression recognition for the extraction and recognition of features have been foreseen in the last few decades.

Either the geometrical feature or appearance feature can be analyzed by facial expressions. The main aim of the geometric method is to use geometric links in the absolute and relative position of facial components in the extraction of features. The intensity of expressions varies over time for one person and two people. Then, without a neutral face as a reference image, it is difficult to determine accurate facial expression intensities. A lot of approaches are used to capture geometric changes on the face with a predefined model. Their study shows that some expressions depend exclusively on geometric deformation, regardless of texture characteristics.

The AAM model in [2] takes into account the distance and form signature extraction points selected and only increases its average recognition rate if both characteristics are used together. It includes the LBP feature in its extended version and its effectiveness is checked with a profound algorithm of learning. However, the extraction of geometric features generally requires a precise method of detection and is difficult to implement in real time. In addition, methods based on geometrical features ignore changes in skin texture that are vital in modeling facial expression to capture micro muscle motion [3].

In this case, the proposed facial recognition texture descriptor is trained and tested using a semi-monitored two-stage pseudo-labeling technique. Strong classifiers with small labeled data are provided in the first stage. The early stage prediction model is used to pseudo-label the labeling data in the given partition and then combined in the next stage with labelled data. In the second phase, the most important amount of weight is predicted by the CNN classifier. Final classification accuracy shows that, by using semi-monitored algorithms based on pseudo-labeling and facially unlabeled information, the proposed hybrid texture descriptor recognizes emotion effectively [4].

2. RELATED WORKS

Texture usually contains many pattern types. The combination of many patterns allows an effective texture representation since the information in a different aspect is collectively represented. Therefore, several characteristics of various extraction methods can be used together to improve the discriminating power of the FER facial feature. For instance, a hybrid PCA and LBP approach is introduced, to combine local and global facial expression grayscale features. The histogram of various featured images based on a union vector is concatenated in the serial feature fusion strategy.

The combination of the introduction of SIFT and SURF features uses a space bag of functions to produce a longitudinal feature for each sample. When the number of features increases, multiple feature fusion techniques are calculated intensive. The local ternary gradient pattern combines histograms with positive and negative gradient differences and is used to reduce the spatial dimension of the characteristics to reduce computational costs without losing the informative textures. Recently, the proposed invariant illumination combination of binary pattern coding with PCA has demonstrated that full-face analyzes provide a better rate of identification than zones-based real-time dataset recognition.

In computer-based classification problems the role of machine learning algorithms is significant. This is usually done with training data, which can either be marked or unmarked for the various classes. Supervised learning is the process by which a function is derived from labeled training data. This paper concerns the application of unmarked data in supervised learning and is accompanied by a large number of marked data.

This example of learning is called semi-controlled learning. The reason behind semi-controlled learning is the fact that labeled data are generally far harder to access compared with unlabeled data, e.g. unlabeled data is all images in the data base in object classification, whilst manual labeling of each image into one of the object classes is required for labeled data.

The general assumption for these algorithms is that there are likely to be same class data points in a high-density region and the decision limit lies in the low-density regions. The idea is to use labeled data for creating an initial training model and to determine initial forecasts for the test data (pre-labels).

Combining and retraining the labeled and pre-labeled data, the initial decision edge can change, which can significantly increase the performance. Many semi-monitored approaches to learning were suggested and proved promising. In this study we examine grading performance using semi-supervised learning techniques based on pseudo-labeling. Experimental results confirm the ability to use pseudo labeling methods of the hybrid function proposed in FER.

The data from real-world collection is difficult and expensive as it takes more time and requires domain expert labeling. Furthermore, unlabeled examples with wrong labeling seriously affect system efficiency. The application's semi-controlled learning algorithms benefit from a more natural learning such as human cognition.

Semi-supervised learning in many applications acts as a text-classification [5], spam detection [6], face [7], etc., to combine unlabeled and labeled data in built-in models with promising results. The predominant feature selection method in [8] was done using labeled data against unlabeled data and the results revealed that the system information is affected by neglecting the label information. However, they used only one-stage classification algorithm supervised.

The labeled data gathered from the real world is difficult and costly because it takes time to collect a domain expert for labeling. In addition, unlabeled examples with incorrect labels severely affect system efficiency. The semi-monitored learning algorithms used in application therefore benefit from more natural learning, such as human cognition.

Semi-supervised learning seeks to combine unlabeled and labeled data to integrated models and is used with promising results in many applications such as text classification [9], spam mail detection [10], facial expression [11] etc. The dominant method of selection of features in [12] was used with labeled data to control unlabeled data, and results showed that neglect of label information would impact the system information. But they only have one-stage classification algorithms supervised.

3. PROPOSED METHOD

The details of the proposed approach appear in this section. In general, the FER automatic appearance frame consists of the following steps: face and face component detection, sub-regions identification, extraction of local functions, reduction of dimensionality, and lastly training and testing via an algorithm for master learning. On nearly every step depends the accuracy of the final classification. The different stages of the proposed method are shown in Fig.1.

Facial detection is conducted using Haar Cascade features through a well-known object detection algorithm. Dynamic contrast stretch limits then adjust the illuminating effect and the variation of pixel intensity. The operators LBP are then applied separately and extracted characteristics are concatenated after normalization.

A semi-surveyed technique is used to classify the pseudo-etiquette for unlabeled data, which makes the class of integrated data the weakest classifier by using one of the classifiers CNN. The best base classifier that is best suited to this application is identified from the experimental results.

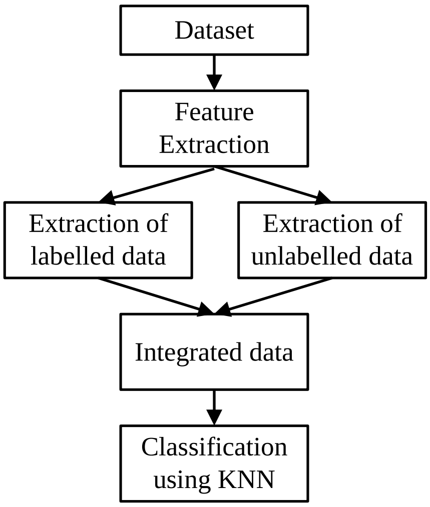

Fig.1. Architecture of proposed classification

The LDP is the binary code for every pixel of an image in the local neighborhood relation. The number of nearest pixels of the current pixel for the given sub-region is less and provides more related information, apart from a center pixel. Unlike LBP, a comparison is made of two adjacent and relevant pixels for each pixel value. Both of these adjacent pixels are either horizontal or vertical and are very near to the pixel in the range. These two adjacent pixels were related to the current pixel to substitute the current pixel value for the binary value.

This texture descriptor is used in AFER, where binary code is computed in a single pass through the XNOR operation in relation to the next pixel to the current pixel. The 8-bit binary pattern is then turned into a decimal to substitute for the central pixel. Finally, the histogram depicts the frequency domain LDP feature image.

3.1 CLASSIFICATION

The CNN consists of eight layers, the first five are convolutional and the other three are completely connected. A 1000 way softmax produces a distribution over the 1000 class labels is used for the output of the last fully connected layer. Our network optimizes the multinomial logistic regression target that corresponds to the maximization of the average log probability under the prediction distribution in training cases.

Only those kernel maps of the previous layer residing in the same GPU are connected to the second, fourth and fifth convolutional layers. The kernels in the third convolutional layer are connected in the second layer to all kernel maps. The neurons are connected to all neurons of the preceding layer in the completely linked layers. The first and second convolutional layers are the response-normalization layers. Both response-normalization layers and the fifth convolutional layer follow the max-pooling layers. For the output of all convolutional and fully connected layers the ReLU non-linearity is applied.

The first convolution layer filters a 4 pixel (this is the distance between receptive field centers of the neighboring neurons in a kernel map) image of 224×223×3 input image with 96 kernels of the size 11×11×3. The second convolutional layer takes the output of the first convolutional layer as an input (response standardized and pooled) to filter it out with 256 kernels size fivefold to four. Without interference in bundling or normalization layers the third, fourth and fifth convolution layers are connected to each other. The third overall layer contains 384 kernels size 3×3×2456 connected with the second convolutional layer's (normalized, pooled) output. There are 384 kernels in the fourth convolutional layer, 3×193 kernels, and the fifth convolutional layer has 256 kernels in the size 3×192 kernels. Each of the fully connected layers consists of 4096 neurons.

The first form of data increase consists of translations of images and horizontal reflections. We take this by removing 224×224 patches from the 256 images and train our network on the extracted patches (and the horizontal reflections). This increases the size of our training by a factor of 2048, although of course, the results are highly interdependent. Without this system, our network is under tremendous overfit, pushing us to use substantially smaller networks. At the time of testing, the network predicts the extraction of five 224×224 patches (four corner patches, middle patches, and ten patches in all) and averages the network softmax layer predictions on the 10 patches.

The second type of data increase is to alter the RGB channels intensity in training pictures. In particular, in the ImageNet training set we perform PCA on the set of pixel values for the RGB. We add multiple of found main components to each training image, with magnitudes proportional to the corresponding own value times a random Gaussian variable with a mean zero and standard 0.1.

Combining predictions for many different models is an excellent way to reduce test errors but for large neural networks which take several days to train, it appears too costly. However, the model combination has a very efficient version, which costs only about a factor of two during training. The newly developed technology, known as dropout, includes setting the output at 0.5 for each hidden neuron to zero. The neurons drop out and are not involved in the forward passage and are not involved in the rear propagation. Thus, the neural network samples a different architecture every time an input is sent, but these architectures share weight.

This technique decreases complex neuronal co-adaptation, since a neuron cannot rely on other neurons' involvement. It is thus forced in tandem with several separate random subsets of other neurons to learn more robust features. At time of the test, all neurons were used, but their outputs multiplied by 0.5, which represents an approximation to the exponential deposition networks' geometric mean of the predictive distributions.

The weight of the zero mean Gaussian distribution with standard deviation 0.01 has been initialized in each layer. In the second, fourth, and fifth convolution layers we initialized the neuron biases and the completely interconnected hidden layers with the constant one 1. This initialization speeds up the early stages of education by providing positive input to the ReLUs. In the remaining layers, we initialized the neuron biases with the constant 0.

4. EVALUATION

The effectiveness of the proposed scheme is assessed in a facial expression database of the extended Cohn Canada (CK+), which contains 593 video sequences, which show 123 individuals with different universal expressions. In the well-organized environment, the facial expression images are taken. The length of the image sequences is between 10 and 60 frames.

Of the whole dataset, 327 images are labeled and grouped from 593 images into the six basic classes of emotion-i.e. rage, disgust, fear, happiness, sorrow and surprise. Each image sequence contains the facial expression pictures that change from the starting point to each person apex (expressive top frame). For further processing, six basic images of the emotions apex state are chosen from the dataset.

After preprocessing, extracted functionality such as LBP are normalized and linked for each observation to the one-dimensional feature vector. Pseudo-labeled semi-supervised algorithm data are divided into 5% of observations as marked and augmented by a maximum of 20% in steps of five and the rest of the data as unlabeled.

A very lesser number of instances are used by many binary semi-supervised learning-based approaches, e.g. 1% of the total instance count. In a multi-class environment, this type of partition is not appropriate as instances from certain classes without a mark cannot be included. The first step involves the assignment of a pseudo label to the unlabeled data on the basis of a predictive models based on small labelled training data and the later phase, the integration of labeled and pseudo-labeled data into the weaker classification.

This performance assessment on a different partition is carried out in two levels. In the first level, a predictive model is used to create pseudo labels using various classifiers. Five distinct classification devices, such as J48, Random Forest (RF), MLP, Radial Base Function Network (RBFN) and M-SVM, are used for the pseudo-labeling of data. At the second level the K-NN is used as the weak classifier, and the next-neighbor value is changed to find the best condition for improving the average recognition rate in order to evaluate the classification accuracy.

Predictive model with a supervised classifier is first developed using either 5%, 10%, 15% or 20% of the pseudo-labeling data. The second stage is completed with labeled and pseudo-labeled data with a different *K*-value for final prediction. MSVM with a poly-kernel performs more effectively in all kinds of partitions with different K-NN values than four other classifiers in the first stage classifier. In the Multiclass RBFN classifier and the Random Forest with confidence factor 0.95, the next two highest recognition rates will be realized.

In contrast, MLP is an activation function of 97.4%, which is 0.8% smaller than M-SVM with two hidden layers and an approximate sigmoid function. It is also observed that the rate of recognition is improved by reducing the tagged data size and increasing the unlabeled data size. It has been demonstrated with different *K* values in all strong classifications.

When the data size of the training increases, the predictive capacity also increases with supervised learning algorithms. But this improvement in this approach costs the unlabeled information by using supervised algorithms at the first stage to assign pseudo labels. This approach thus shows indirectly that the supervised base classifier is robust with the semi-controlled multi-class issue and is verified using KNN in the second phase.

In the second stage, K are calculated, and neighbors' weights for each class are summarized, which are the closest neighbors to each case. The class indicates the predicted label of the observation, with the most considerable weight. The chart shows the error of classification for the various classifier with ANN values 10, 20 and 30 from 5% to 20% for the data partition. Change in the value of KNN parameter is performed in experiments.

The Fig.4 displays misclassification rate of the various KNN value classifiers. From the results, the lowest error rates in all partitions with weight adjustment for pseudo-labeled dates can be seen in the KNN with K=10. Consequently, only the KNN parameter for comparison is considered as 10.

For 5% of the training data with K=10, the predictive ability of the model is high, which confirms that the updated model will work on unseen data because most of the observations in predictive model design are pseudo-labeled based on the error in the first step. The Fig.2 shows that the misclassification rate increases with the different closest neighbor counting as the training data increase. The Fig.2 shows an assessment of the semi-supervised two-stage algorithm with different classifiers under the data 20/80 partition.

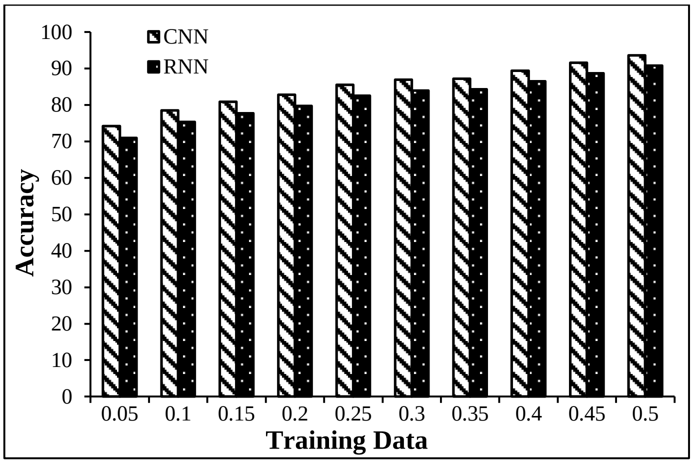

Fig.2. Classification Accuracy

The experimental results show that the accuracy of Random Forest is 0.2% higher than multiclass SVM in 20/80 partitions, whereas the accuracy of both classifiers is equal for 15/85 partitions. The accuracy of the multiclass SVM is 0.5% and 0.3% higher than the random forest and RBFN for the last couple of splits. The misclassified class labels of instances should be modified by a basic classifier in this semi-supervised learning technique. For M-SVM and RBFN classifiers with an average accuracy of recognition of 98.2% and 97.9% each, pseudo-labeling performs better. These two classifications benefit more from the first stage pseudo labeling than the Random Forest, J48 and MLP, which are 97.7%, 97.5% and 97.4% respectively.

In combinations between M-SVM and KNN, all sorts of partitions are best, particularly in classification of happy and sad expression, whereas the expression of surprise and fear sometimes appears confusing due to the very low differences in appearance and the absence of some muscle movements. Furthermore, many different features can contribute to the applied pattern recognition as well as improve the discrimination capabilities of the diverse emotional class by combining multiple features.

Experimental results show that local texture fusion improves the detection accuracy using semi-supervised learning techniques based on pseudo labeling. Training data labels must be carefully applied by the Domain Expert, which is a hard, time consuming and costly process for real-time data, in the supervised machine learning approach. Because only training data depends on predictive models performance. The method proposed uses small labeled training instances with precise class labels and then combines labeled data and huge unmarked data to create a two-stage, pseudo-labeling-based FER system model. As pre-labeling of training data ensures high quality, the semi-monitored algorithm can classify unlabeled instances with a high precision based on a large number of unlabeled data.

With multiple random data partitions, the accuracy of the right labeling of training data is verified to cover any possible combination. The results show clearly that pre-antiquating must be reviewed in order to maintain accuracy in recognition. You can erroneously classify the expressions into another category if it is difficult to distinguish emotions from one extracted feature. This makes it difficult to comply with the high accuracy standards in the supervised learning algorithm classification. The Table.2 present the comparison of the approach proposed with the art status of the CK+ dataset. The results show that our method is best suited to the classification of six facial expressions using fusion and semi-controlled learning in local texture.

The proposed approach has some weaknesses. The main characteristics of the good facial area could be distinguished between classes, while distinct classes are well discriminated against. Second, they could be extracted easily from a raw facial image, thus reducing processing time. The large-scale space analysis increases the computer complexity. The selection or dimensional reduction of the features can therefore be applied to minimize training time, which reduces complexity in turn.

Another problem is that the multi-class problem is class imbalance. The system may lose the performance of one class when dealing with such data, while trying to gain the performance of another. These two problems can be dealt with in numerous ways. We need effective solutions that are appropriate for the semi-supervised learning environment, to be identified or developed. The approach proposed here deals with the proper accuracy of recognition of the features extracted from the images taken from the controlled environment by M-SVM and KNN. However, this must also be extended to handle data in real time.

5. CONCLUSION

This paper presents a descriptor of hybrid local texture to recognize emotions where, in addition to a pixels-center relationship, mutual relationships between local district pixels are considered. The method proposed is the mutual relationship between the nearest neighborhood pixels, whereas LBP codes the nearest pixel with the central pixel. These two features are easy to compute and the most informative feature is their logical fusion.

REFERENCES

[1] J. Chen, Z. Chen and Z. Chi, "Emotion Recognition in the Wild with Feature Fusion and Multiple Kernel Learning", *Proceedings of International Conference on Multimodal Interaction*, pp. 508-513, 2014.

[2] S. Chen and Q. Jin, "Multi-Modal Dimensional Emotion Recognition using Recurrent Neural Networks", *Proceedings of International Workshop on Audio/Visual Emotion Challenge*, pp. 49-56, 2015.

[3] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort and M. Bartlett, "Multiple Kernel Learning for Emotion Recognition in the Wild", *Proceedings of ACM on International Conference on Multimodal Interaction*, pp. 517-524, 2013.

[4] C. Venkata Rami Reddy, K.K. Kishore and D. Bhattacharyya, "Multi-Feature Fusion based Facial Expression Classification using DLBP and DCT", *International Journal of Software Engineering and Its Applications*, Vol. 8, No. 9, pp. 55-68, 2014.

[5] J. Chen and Z. Chen, "Facial Expression Recognition in Video with Multiple Feature Fusion", *IEEE Transactions on Affective Computing*, Vol. 9, No. 1, pp. 38-50, 2019.

[6] T. Senechal, V. Rappa and H. Salam, "Facial Action Recognition combining Heterogeneous features via Multikernel Learning", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Vol. 42, No. 4 pp. 993-1005, 2012.

[7] M. Kaiser, F. Eyben, B. Schuller and G. Rigoll, "LSTM-Modeling of Continuous Emotions in an Audio-Visual Affect Recognition Framework", *Image and Vision Computing*, Vol. 31, No. 2, pp. 153-163, 2013.

[8] X. Wang, C. Jin, W. Liu, M. Hu and L. Xu, "Feature Fusion of HOG and WLD for Facial Expression Recognition", *Proceedings of IEEE/SICE International Symposium on System Integration*, pp. 227-232, 2013.

[9] M. Song, M. You, N. Li and C. Chen, "A Robust Multimodal Approach for Emotion Recognition", *Neurocomputing*, Vol. 71, No. 10-12, pp. 1913-1920, 2008.

[10] Y. Song, L.P. Morency and R. Davis, "Learning A Sparse Codebook of Facial and Body Microexpressions for Emotion Recognition", *Proceedings of ACM International Conference on Multimodal Interaction*, pp. 237-244, 2013.

[11] C.A. Mazefsky, K.A. Pelphrey and R.E. Dahl, "The Need for a Broader Approach to Emotion Regulation Research in Autism", *Child Development Perspectives*, Vol. 6, No. 1, pp. 92-97, 2012.

[12] I. Luengo and E. Navas, "Automatic Emotion Recognition using Prosodic Parameters", *Proceedings of 9th European Conference on Speech Communication and Technology*, pp. 1-8, 2005.