# A Two-Stage Voting Mechanism for ROI Regions in Facial Expression Recognition Postprint

**Authors:** Wen Yuanmei, Ouyang Wen, Ling Yongquan

**Date:** 2018-05-24T00:00:00+00:00

## Abstract

To investigate how to more effectively utilize the distributed features learned by convolutional neural networks (CNNs) from training images, this paper proposes a region-of-interest (ROI)-based two-level voting mechanism for facial expression recognition. First, the image is divided into a series of ROI images, which are input into the CNN for training. Then, the ROI images of the test image are fed into the CNN, and the classification results of all ROI images are aggregated. Finally, a two-level voting mechanism is employed to determine the final category of the test image, yielding the ultimate classification result. Additionally, to address the inability of CNNs to learn spatial position information such as rotation from facial images, a spatial transformer network (STN) is introduced to enhance the algorithm' s capability in tackling expression recognition problems under complex conditions. Experimental results demonstrate that the ROI-based two-level voting mechanism can more effectively leverage the distributed features learned by CNNs from training images, achieving a 1.1% accuracy improvement over the method that directly employs ROI images for voting. The introduction of the STN network can effectively improve the robustness of CNNs, yielding a 1.5% accuracy improvement over the method without STN.

## Full Text

## Preamble

**Title:** An Expression-Oriented ROI Region Secondary Voting Mechanism for Facial Expression Recognition

**Authors:** Wen Yuanmei, Ouyang Wen, Ling Yongquan
**Affiliation:** School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China

**Abstract:** This paper addresses the challenge of effectively utilizing distributed features learned by convolutional neural networks (CNNs) from training images, proposing a Region of Interest (ROI) region secondary voting mechanism for facial expression recognition. The approach first divides images into a series of ROI images for CNN training. During testing, ROI images from test samples are fed into the CNN to collect discriminative results from all ROI regions. A secondary voting mechanism then determines the final category of the test image. Additionally, to overcome the limitation of CNNs in learning spatial position information such as rotation, we incorporate Spatial Transformer Networks (STN) to enhance the algorithm' s capability for handling complex expression recognition scenarios. Experimental results demonstrate that the ROI region secondary voting mechanism more effectively leverages distributed features learned by CNNs, achieving a 1.1% accuracy improvement over direct ROI voting methods. The introduction of STN networks significantly enhances CNN robustness, yielding a 1.5% accuracy gain compared to methods without STN.

**Keywords:** convolutional neural network; expression recognition; STN network; secondary voting mechanism

## 0 Introduction

Since Krizhevsky et al. [4] demonstrated that convolutional neural networks (CNNs) could outperform hand-crafted features in the ILSVRC-2012 image recognition competition, CNNs have attracted widespread attention. Facial expressions, generated by muscle deformations around the eyes, nose, mouth, and eyebrows, represent one of the most powerful, natural, and direct means of emotional communication in humans, accurately reflecting an individual' s current state. Facial expression recognition enables computers to perceive human emotions and provide more personalized services based on expressive information, with broad applications in safe driving, human-computer interaction, lie detection, and numerous other domains, making it a prominent research focus in computer vision [1–3].

Researchers have conducted extensive explorations of CNN-based expression recognition [5,6]. For instance, Hamester et al. [7] proposed a dual-channel CNN comprising a standard CNN channel and a CAE channel for expression recognition. Liu et al. [8] combined CNN-extracted features with hand-crafted Centralized Binary Patterns (CBP) features, using an SVM classifier. Meng et al. [9] fused identity features with expression features extracted from CNNs to improve recognition accuracy. However, these methods primarily focus on learning global features directly from expression images, failing to fully exploit local distributed expression features that could guide CNNs to concentrate on critical regions of expression variation.

ROI (Region of Interest) regions refer to areas of particular interest that vary

across different computer vision tasks. Defining ROI regions can actively guide algorithms to focus on key areas, thereby improving recognition accuracy and accelerating processing speed. In expression recognition, introducing ROI regions can direct CNNs to attend to expression-relevant key areas, enhancing recognition precision. Previous researchers have incorporated ROI regions into CNN-based expression recognition. For example, Vo et al. [10] trained CNNs separately on global and local images to learn both types of information, adjusting global image probability distributions using local image test results during inference. Sun et al. [11] proposed dividing faces into a series of ROI regions for CNN training, employing ROI image voting during testing and selecting the category with the most votes as the final result. These methods utilize ROI images in both training and testing, demonstrating that auxiliary discrimination using ROI images during the test phase effectively improves CNN recognition accuracy.

While Sun et al. [11] achieved superior results compared to traditional CNN-based expression recognition methods through ROI voting, local ROI images contain limited information, making them prone to misjudgment. To fully utilize the distributed expressive features learned by CNNs during training while mitigating the impact of insufficient information in local ROI images, this paper proposes an expression-oriented ROI region secondary voting mechanism. This approach assigns greater influence weights to global images while reducing the impact of local ROI images on final discrimination results. Additionally, to address the rotation invariance limitation identified in [11], we incorporate Spatial Transformer Networks (STN) [12] into expression recognition, enabling CNNs to learn spatial position information from expression images and enhancing system robustness.

## 1 Model Improvement Methods

Building upon the model in [11], this paper investigates improvements to ROI image auxiliary discrimination methods and model rotation invariance. The following sections detail these two aspects.

### 1.1 ROI Auxiliary Discrimination Method

Introducing ROI images during CNN training not only expands the dataset to prevent overfitting but also enables CNNs to learn distributed expressive features from expression images. To fully leverage these distributed features while reducing misjudgment caused by insufficient information in local ROI images, we propose an ROI region secondary voting mechanism that improves expression recognition accuracy.

Following [11], we define nine distinct ROI regions based on facial structure through segmentation, occlusion, flipping, and center-focusing operations. During segmentation, we focus on changes in the eye, nose, and mouth regions, extracting four ROI images (ROI0, ROI1, ROI2, ROI3). Occlusion operations

block the upper and lower face portions, yielding two ROI images (ROI4, ROI5). Flipping addresses varying camera angles by performing horizontal flips, producing one ROI image (ROI6). Center-focusing eliminates hair and other noise effects by concentrating on key facial expression areas, generating one ROI image (ROI7). These eight processed ROI images plus the original image (ROI8) constitute nine total ROI images, as illustrated in [Figure 1: see original paper].

Typically, neutral expressions show no significant changes in the eyes, mouth, or nose. Happy expressions feature widened or upturned mouth corners, narrowed eyes, and raised nose wings. Sad expressions exhibit downward-sloping eyebrows and eye corners, with widened mouths or downturned corners. Angry expressions display raised eyebrows, downturned mouth corners, furrowed brows, and raised nostrils, sometimes accompanied by open mouths. Surprise expressions involve wide-open mouths and eyes with raised eyebrows. However, each expression varies in amplitude, and different amplitude levels produce different manifestations. For example, when expressing happiness with a widely open mouth, the mouth region ROI image becomes highly similar to that of surprise expressions. Direct discrimination of such ROI images thus risks misclassifying them as surprise. Therefore, during voting, we should reduce the influence of local-information ROI images while enhancing the impact of global-information ROI images on final results.

Inspired by decision tree multi-level decision-making, we propose a secondary voting mechanism. Decision trees typically use information gain for attribute selection, calculated as:

$$Gain(D, V) = Ent(D) - \sum_{v \in V} \frac{|D_v|}{|D|} Ent(D_v)$$

where $D$ represents the sample set and $V$ denotes a sample attribute. Higher information gain indicates greater attribute influence, leading to priority node selection. Consequently, we prioritize discriminative nodes using global-information ROI images.

Among the nine ROI region images, ROI6 and ROI8 contain complete expression information. Therefore, when voting with ROI images, we increase the influence of ROI6 and ROI8 while reducing the impact of other ROI images. The specific steps of our secondary auxiliary discrimination are:

a) Divide test images into a series of ROI region images.

b) Input the divided ROI images into a trained CNN and record discrimination results for each ROI image.

c) Compare and verify consistency between ROI6 and ROI8 discrimination results. If consistent, merge their result as the test image' s final discrimination result. If inconsistent, apply the ROI-KNN method for voting,

selecting the category with the most votes as the test image result.

### 1.2 Rotation Invariance Study

In practical applications, facial images often contain rotation angles, and captured images vary in shooting distance. Enhancing system capability to handle rotation and scaling thus improves practicality.

To address CNN limitations in learning spatial information such as rotation and scaling, we introduce STN networks to solve rotation invariance issues. Proposed by Google's Jaderberg et al. [12] in 2015, STN networks consist of three modules: Localisation Network, Grid Generator, and Sampler, as shown in [Figure 2: see original paper].

The STN learning process for expression image position information comprises forward propagation and backward adjustment phases.

**Forward propagation:**
a) Input expression images into the Localisation Network, which outputs transformation parameters $\theta$ through fully connected layers. Assuming $\theta$ is:

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix}$$

b) Input transformation parameters $\theta$ into the Grid Generator module to obtain coordinate mapping $T_\theta$ between generated and original images:

$$\begin{bmatrix} x_i^t \\ y_i^t \end{bmatrix} = T_\theta \begin{bmatrix} x_i^s \\ y_i^s \\ 1 \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{bmatrix} x_i^s \\ y_i^s \\ 1 \end{bmatrix}$$

where $(x_i^t, y_i^t)$ are coordinates in the generated grid and $(x_i^s, y_i^s)$ are coordinates in the original image.

c) Perform interpolation using the coordinate mapping, applying bilinear interpolation to pixel values from the original image coordinates in the generated grid:

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|)$$

where $U_{nm}^c$ represents pixel values at channel $c$ and coordinates $(n, m)$ in the original image.

d) Input the interpolated generated image into the CNN for feature extraction.

**Backward adjustment:** The backpropagation process allows error to continue forward through the STN network. Assuming the error from the previous CNN layer is $\frac{\partial loss}{\partial V_i^c}$, the STN backpropagation process is:

$$\frac{\partial loss}{\partial U_{nm}^c} = \sum_i \sum_{h=1}^{H} \sum_{w=1}^{W} \frac{\partial loss}{\partial V_i^c} \cdot \frac{\partial V_i^c}{\partial U_{nm}^c}$$

The parameter $\theta$ adjustment process in STN is:

$$\frac{\partial loss}{\partial \theta_{11}} = \sum_i \sum_{c=1}^{C} \frac{\partial loss}{\partial V_i^c} \cdot \frac{\partial V_i^c}{\partial x_i^s} \cdot \frac{\partial x_i^s}{\partial \theta_{11}}$$

Similar calculations apply for other parameters.

## 2 Experiments

To validate the effectiveness of our proposed ROI region secondary voting mechanism and whether STN introduction enables rotation invariance, we compare our method (Ours+CNN) against ROI-KNN (ROI-KNN+CNN) and direct image classification (ROI+CNN). We also conduct comparative experiments on CNNs with and without STN networks under rotated sample conditions.

### 2.1 Sample Selection and Evaluation Metrics

Our dataset combines the CK+ (Extended Cohn-Kanade) dataset [13] with Wild facial expression data collected from the internet, forming a new dataset. It contains 700 images each of happy, sad, surprised, and angry expressions from CK+ mixed with 200 internet-collected images per category, plus 900 neutral images from laboratory conditions. This yields five categories with 900 images each, totaling 4,500 images. The test set comprises 300 internet-collected images per category (excluding neutral) mixed with 300 laboratory neutral images, totaling 1,500 images across five categories, designated as Dataset I.

Building upon Dataset I, Sun et al. [11] generated nine ROI images per image through cropping, flipping, occlusion, and center-focusing, as shown in [Figure 1: see original paper], resulting in five categories with 4,500 images each, with unchanged test images, designated as Dataset II. To investigate whether injecting rotated samples enables rotation invariance, Sun et al. [11] generated rotated samples from Dataset I' s regular data and mixed them with Dataset II, creating 83,500 images across five categories, designated as Dataset III. Our experiments utilize Datasets II and III.

We employ accuracy as the evaluation metric:

$$Accuracy = \left(1 - \frac{\text{Total misclassified samples}}{\text{Total samples}}\right) \times 100\%$$

### 2.2 CNN Structure and Parameter Settings

Our CNN architecture follows [11], using a 9-layer network with three convolutional layers, three max-pooling layers, one fully connected layer, one dropout layer [14], and one softmax layer, as detailed in .

**Table 1: CNN Architecture Used in This Paper**

| Layer | Output Feature Map | Convolution Kernel Size | Pooling Kernel Size |
|---|---|---|---|
| Input | 32$×32×1\| − \| − | - | - |
| | $\|\|Conv1\|30×30×64\|3×3\|−$ | | |
| | $\|\|Pool1\|15×15×64\|−$ | | |
| | $\|2×2\|\|Conv2\|12×12×64\|4×4\|−$ | | |
| | $\|\|Pool2\|6×6×64\| −$ | | |
| | $\|2×2\|\|Conv3\|2×2×128\|5×5\|−$ | | |
| | $\|\|Pool3\|1×1×128\|−$ | | |
| | $\|2×2\|\|FullyConnected\|1×1×300\|−$ | | |
| | $\| −$ | | |
| | $\|\|Dropout\|1×1×300\|−$ | | |
| | $\| −$ | | |
| | $\|\|Softmax\|1×1×$5 | | |

The first convolutional layer uses $3$×3kernels, outputting64featuremapsofsize30×30.Thefirstpoolinglayerus
The fully connected layer outputs 300 features, followed by a dropout layer with probability 0.5, and finally a softmax layer. All convolutional layers use ReLU activation functions [15].

Weight and bias initializations follow a standard normal distribution with mean 0 and standard deviation 0.1. During training, we randomly sample 100 samples per batch for 50,000 iterations. The initial learning rate is 0.01 with momentum 0.09. After every 5,000 iterations, we validate the model; if accuracy stagnates or declines, the learning rate decreases by one order of magnitude. Training stops when the learning rate reaches 0.0001.

### 2.3 Experimental Procedures

**2.3.1 ROI Auxiliary Discrimination Experimental Steps** Using Dataset II, we validate our proposed ROI auxiliary discrimination method through three stages: CNN training, ROI image testing, and ROI region auxiliary discrimination.

**Training Stage:** Divided ROI images are normalized using:

$$\text{train\_image} = \frac{\text{image} - (255/2.0)}{255}$$

Normalized images are input to the CNN for training to obtain the final model.

**Testing Stage:** Test image ROI images are normalized and input to the trained CNN, and discrimination results for all ROI images are recorded.

**Final Discrimination Stage:** Our proposed method processes the ROI test results to obtain final discrimination outcomes.

**2.3.2 Rotation Invariance Experimental Steps** Using Dataset III, we conduct rotation invariance studies. First, Dataset III is input to the CNN for training; after training completes, the test set is evaluated and accuracy recorded. Next, an STN network is introduced at the first CNN layer. Dataset III is input to this STN-enhanced CNN for training, followed by testing and accuracy recording.

## 3 Experimental Results and Analysis

### 3.1 ROI Auxiliary Discrimination Results and Analysis

To verify our method' s effectiveness, we compare it against ROI-KNN+CNN and ROI+CNN methods, with results shown in .

**Table 2: Accuracy Comparison of Different Auxiliary Discrimination Methods**

| Method | Accuracy |
|---|---|
| Ours+CNN | 78.0% |
| ROI-KNN+CNN | 76.9% |
| ROI+CNN | 73.2% |

Our method achieves 78.0% accuracy, representing a 1.1% improvement over ROI-KNN+CNN and a 4.8% improvement over ROI+CNN, demonstrating superior performance.

To analyze the reasons for this improvement, we examine confusion matrices for each method, shown in [Figure 3: see original paper], [Figure 4: see original paper], and [Figure 5: see original paper]. The confusion matrices have predicted categories on the horizontal axis and actual categories on the vertical axis, with categories ordered as neutral, happy, sad, surprised, and angry. Diagonal values indicate per-category accuracy.

Our method achieves per-category accuracies of 0.98, 0.71, 0.56, 0.89, and 0.76 for neutral, happy, sad, surprised, and angry expressions, respectively. ROI-KNN+CNN achieves 0.98, 0.69, 0.54, 0.88, and 0.75, while ROI+CNN achieves 0.96, 0.66, 0.49, 0.83, and 0.68. Our method outperforms or matches the others across all categories.

To understand why our method surpasses ROI-KNN, we analyze samples correctly classified by our method but misclassified by ROI-KNN+CNN, as shown in [Figure 6: see original paper]. We visualize the category probability distributions for these samples' ROI images in .

In , ROI0, ROI2, and ROI4 show maximum probabilities for surprise (0.900, 0.855, 0.844), leading to three ROI images being classified as surprised. Similarly, ROI3, ROI5, and ROI8 show maximum probabilities for happiness (0.561, 0.629, 0.992), resulting in three ROI images being classified as happy. This demonstrates how local ROI images with limited information can be misjudged due to similarity across expression categories. When multiple local ROI images are misclassified, they may equal or outnumber correctly classified ROI images, causing ROI-KNN to fail. Our secondary voting mechanism first compares results from two global-information ROI images (ROI6 and ROI8), applying voting only when they disagree, thereby reducing local ROI image influence and achieving better results.

While our method and ROI-KNN require nine discriminations per test image versus one for ROI+CNN, the time overhead is minimal. Testing 1,500 images takes 0.125s for ROI+CNN and 0.680s for our method—only 0.555s longer—but yields a 4.8% accuracy improvement. Thus, our method achieves superior discrimination with only slight time increase.

### 3.2 Rotation Invariance Results and Analysis

Using Dataset III, we conduct six experiments: three with STN and three without. Results appear in .

**Table 4: Comparison of STN Introduction Results**

| Method | Without STN | With STN |
|---|---|---|
| ROI+CNN | 73.5% | 75.0% |
| Ours+CNN | 76.7% | 78.4% |
| ROI-KNN+CNN | 74.8% | 76.9% |

Introducing STN improves accuracy by 1.5% for ROI+CNN, 1.7% for Ours+CNN, and 2.1% for ROI-KNN+CNN. Under rotated sample conditions, STN consistently enhances expression recognition accuracy, demonstrating that STN enables CNNs to learn spatial information and achieve higher recognition rates. Moreover, STN introduction does not degrade accuracy for neutral expressions (laboratory conditions without rotation), confirming that STN does not adversely affect non-rotated images.

Since STN is introduced at the first CNN layer, it adds 32$×32×$6 = 6,144 connections compared to the baseline CNN. Training time increases from 928s to 1,383s (455s additional), while test time increases modestly from 0.148s to

0.229s (0.081s additional). Thus, STN effectively solves rotation invariance with minimal inference time overhead, meeting real-time processing requirements.

To further observe STN' s effect, we examine confusion matrices for ROI+CNN with and without STN, shown in [Figure 7: see original paper] and [Figure 8: see original paper]. With STN, per-category accuracies are 0.98, 0.73, 0.43, 0.84, and 0.77 versus 0.96, 0.73, 0.45, 0.82, and 0.71 without STN. STN improves accuracy across most expressions, confirming its effectiveness.

## 4 Conclusion

In facial expression recognition tasks, to fully utilize distributed features learned by CNNs during training while reducing misjudgment impact from local ROI images with limited information, this paper proposes a secondary voting mechanism for auxiliary discrimination of expression images. Comparative experiments against ROI-KNN and data augmentation using ROI images demonstrate that our secondary voting mechanism achieves superior results. Additionally, we investigate enabling CNNs to learn spatial position information from expression images by introducing STN networks, conferring rotation invariance and enhancing system robustness. While our method improves expression recognition accuracy, its performance under complex conditions involving varying illumination and angles remains limited. Establishing an expression recognition system capable of accurate recognition under such complex conditions represents our future research focus.

## References

[1] Hernandez-Matamoros A, Bonarini A, Escamilla-Hernandez E, et al. Facial expression recognition with automatic segmentation of face regions using a fuzzy based classification approach [J]. Knowledge-Base Systems, 2016, 110: 1-14.

[2] Shan Caifeng, Gong Shaogang, McOwan P W. Facial expression recognition based on Local Binary Patterns: A comprehensive study [J]. Image and Vision Computing, 2009, 6 (27): 803-816.

[3] Luo Yuan, Zhang Ling, Chen Yunhua, et al. Facial expression recognition based on hierarchy structured dictionary learning [J]. Application Research of Computer, 2017, 34 (11): 3514-3517.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. 2012: 1097-1105.

[5] Shin M, Kim M, Kwon D S. Baseline CNN structure analysis for facial expression recognition [C]// Proc of the 25th IEEE International Symposium on Robot and Human Interactive Communication. 2016: 724-729.

[6] Perveen N, Singh D, Mohan C K. Spontaneous facial expression recognition: a part based approach [C]// Proc of IEEE International Conference on Machine

Learning and Applications. 2017: 819-824.

[7] Hamester D, Barros P, Wermter S. Face expression recognition with a 2-channel convolutional neural network [C]// Proc of International Joint Conference on Neural Networks. 2015: 1-8.

[8] Liu Yize, Chen Yixiang. Recognition of facial expression based on CNN-CBP features [C]// Proc of the 29th Chinese Control And Decision Conference. 2017: 2139-2145.

[9] Meng Zibo, Liu Ping, Cai Jie, et al. Identity-aware convolutional neural network for facial expression recognition [C]// Proc of the 12th International Conference on Automatic Face & Gesture Recognition. 2017: 558-565.

[10] Vo D M, Sugimoto A, Le T H. Facial expression recognition by re-ranking with global and local generic features [C]// Proc of the 23rd International Conference on Pattern Recognition. 2017: 4118-4123.

[11] Sun Xiao, Pan Ting, Ren Fuji. Facial expression recognition using ROI-KNN deep convolutional neural networks [J]. Acta Automatica Sinica, 2016, 42 (6): 883-890.

[12] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2015: 665-673.

[13] Lucey P, Cohn J F, Kanade T, et al. The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression [C]// Proc of Computer Vision and Pattern Recognition Workshops. 2010: 94-101.

[14] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing coadaptation of feature detectors [J]. Computer Science, 2012, 3 (4): 212-223.

[15] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks [C]// Proc of the International Conference on Artificial Intelligence and Statistics. 2012: 315-323.

*Note: Figure translations are in progress. See original paper for figures.*