Group Emotion Recognition Based on Multi-Stream CNN-LSTM Network (Postprint)

Authors: Qing Linbo, Xiong Wenshi, Zhou Wenjun, Xiong Shanshan, Wu Xiaohong

Date: 2018-05-20T00:00:00+00:00

Abstract

Group emotion recognition is a frontier research topic in the field of human-computer interaction. To address the accuracy issue in group emotion recognition, this paper proposes a multi-stream CNN-LSTM network model that integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM) to learn both static and dynamic features of group emotions. The original images from video sequences, visual saliency maps, and stacked optical flow images are employed as inputs for three distinct channels. The CNN network analyzes spatial features and local motion features, with the resulting feature maps being directly fed into the LSTM network for learning global motion features. Finally, a Softmax classifier is appended, and the Softmax outputs from the three channels are subjected to weighted fusion to produce the classification results. Experimental results indicate that the proposed model can effectively recognize four typical categories of group emotions, achieving a recognition rate superior to existing algorithms, with accuracy (ACC) and macro-average precision (MAP) reaching up to 82.6% and 84.1%, respectively.

Full Text

Abstract

Crowd emotion recognition is a frontier topic in human-computer interaction. To address the accuracy of group emotion recognition, this paper proposes a multi-stream CNN-LSTM network model that combines convolutional neural networks (CNN) with long short-term memory networks (LSTM) to learn both static and dynamic features of crowd emotions. The model uses original video frames, visual saliency maps, and stacked optical flow images as inputs for three separate channels. CNNs analyze spatial features and local motion characteristics, with the resulting feature maps fed directly into LSTM networks to learn

global motion patterns. Finally, Softmax classifiers are connected to each channel, and weighted fusion of the three Softmax outputs yields the classification result. Experimental results demonstrate that the proposed model can effectively identify four typical crowd emotions with higher recognition rates than existing algorithms, achieving maximum accuracy (ACC) and macro average precision (MAP) of 82.6% and 84.1%, respectively.

Keywords: crowd emotion recognition; convolutional neural network; long short-term memory network; multi-stream

0 Introduction

Currently, numerous traditional methods have been investigated for crowd emotion recognition, focusing on extracting motion features between video sequence frames for classification. Urizar et al. [1] proposed a hierarchical Bayesian model for crowd emotion recognition that infers crowd emotional states by mining relationships between behavior and emotion. Rabiee et al. [2] trained an emotion-based SVM classifier for abnormality detection in surveillance videos. Patwardhan [3] performed edge detection on entire video sequences and extracted features through grid-based linear superposition for SVM classification. Zhang et al. [4] utilized structured trajectory learning to detect coherent motion patterns in crowds, mapped these patterns to an emotional plane, and classified features using a classifier. Although these methods have achieved certain effectiveness in crowd emotion recognition, the manually selected features vary across different environments due to the complexity of real-world scenarios, resulting in poor generalization performance of model parameters.

In 2006, Hinton et al. [5] introduced deep learning theory, which employs hierarchical feature extraction to replace manual feature selection, thereby eliminating human-induced feature variations and enabling automatic feature learning. The typical deep learning model, convolutional neural network (CNN), takes image pixel values as input and simulates human brain analysis and learning, demonstrating strong robustness and achieving remarkable recognition rates in image recognition. For video content recognition, Simonyan et al. [6] argued that both static and motion features should be considered, proposing a two-stream CNN combining original static images and optical flow images. Additionally, Zhao et al. [7] utilized skeleton and raw image information for individual action recognition. Yi et al. [8] established four multi-layer CNNs to learn features from gradient images in four different directions, fusing them to obtain classification results. Long short-term memory network (LSTM) [9] is a novel recurrent neural network that can selectively memorize current inputs and feed outputs back to subsequent inputs, functioning as a dynamic time-delay network with significant advantages in processing temporal inputs. Donahue et al. [10] combined CNN with LSTM to propose Long-term Recurrent Convolutional Networks (LRCN) for video recognition and description, achieving favorable results. Cai et al. [11] combined CNN with bidirectional RNN for facial expression recognition. Qin et al. [12] proposed a fusion model combining 3D CNN and LSTM networks for

action recognition.

Deep learning demonstrates high recognition rates and generalization capabilities in video analysis, yet no scholars have applied it to crowd emotion recognition. Therefore, this paper proposes a multi-stream CNN-LSTM network model for crowd emotion recognition. First, video raw image sequences, saliency map sequences, and optical flow sequences are fed into three CNN networks for feature learning to obtain visual features. Then, the visual features from the three channels are input into LSTM networks to model global motion information and classified using Softmax layers. Finally, the outputs from the three Softmax layers are averaged and fused.

1 Multi-Stream CNN-LSTM Network for Crowd Emotion Recognition

1.1 Video Feature Selection for Crowd Emotion

Static image recognition only requires learning features among pixel values within an image, whereas video-based crowd emotion recognition necessitates learning motion features between video sequence frames. Therefore, both spatial correlations among pixel values and temporal correlations between frames must be considered. Video information can be naturally decomposed into spatial and temporal components. The spatial component consists of video texture features that describe the appearance of scenes and objects, while the temporal component manifests as motion features between frames that describe object movements.

As shown in Figure 1 [Figure 1: see original paper], this paper employs original video images, visual saliency images, and optical flow images to learn spatiotemporal features of crowd emotions in videos. For the spatial component, scene and crowd appearance features are described using original video frames. In recent years, visual saliency maps have been widely applied in image data processing, as they highlight regions of interest to the human visual system. In crowd video images, they precisely reflect the saliency degree of crowds in scenes, as illustrated in Figure 1(b). This paper adopts the typical multi-channel saliency map fusion method from [13] to generate visual saliency maps. From a temporal correlation perspective, crowd emotions are also related to crowd motion information: smooth motion indicates a relaxed state, while intense motion suggests an excited state or anomaly occurrence. Therefore, this paper utilizes optical flow images [14] to describe crowd motion features in the temporal component.

1.2 Multi-Stream CNN-LSTM Network Model

To fully leverage the video features described in Section 1.1 for crowd emotion recognition, this paper proposes the multi-stream CNN-LSTM network model shown in Figure 2 [Figure 2: see original paper], which consists of three-channel CNN-LSTM networks. The two spatial channels use original images and vi-

sual saliency images as inputs, primarily for learning static image information. Original images describe scene and crowd static appearance information, while visual saliency maps represent crowd saliency degrees. Since optical flow images are generated based on correlations between adjacent frames, they are used to represent local motion information in videos as input for the temporal stream. Reference [6] proposed that continuously stacking optical flow images within a short time window can more compactly represent video motion information, yielding better recognition results than raw optical flow images. Therefore, this paper adopts stacked optical flow images as input, with a stack size of 10 frames.

CNN networks not only possess strong image feature learning capabilities but also reduce computational complexity during training. Consequently, this paper employs three CNN networks to model static image information from spatial streams and local motion information from motion streams separately. The primary difference between LSTM and CNN networks lies in LSTM' s ability to continuously retain information, infer subsequent states from previous ones, and thereby learn global motion features from videos. To model static features, local motion features, and global motion features of video sequences, this paper integrates CNN and LSTM networks based on [10], connecting LSTM networks after the first fully connected layer of CNN networks. Finally, the outputs from the three Softmax layers are averaged and fused to obtain the final classification result.

1.3 Network Selection and Training

To improve the accuracy of crowd emotion features, this paper adopts two CNN network models: VGG-19 [15] and AlexNet [16]. The two spatial channels utilize the VGG-19 network model for automatic feature learning. First, the VGG-19 model is pre-trained on the ImageNet dataset [13]; then, the pre-trained model is fine-tuned using training data. During fine-tuning, input images are resized to 224×224, learning rate is set to 10 , and momentum is set to 0.9. To prevent overfitting, Dropout layers are added after each fully connected layer with a Dropout ratio of 0.5.

Reference [18] trained both VGG-19 and AlexNet models on stacked optical flow images, finding that AlexNet demonstrated stronger learning capability for stacked optical flow images. Therefore, this paper adopts the AlexNet model for the temporal stream CNN. Following the pre-trained optical flow model provided in [10], the model is fine-tuned using training stacked optical flow sequences. Unlike the spatial stream VGG-19 network, the AlexNet network for the temporal stream accepts input images of size 227×227, and the Dropout ratio after fully connected layers is set to 0.7 to avoid overfitting.

The connection between CNN and LSTM networks is illustrated in Figure 3 [Figure 3: see original paper]. The number of video frames T for LSTM network input is fixed at 16 frames, with each LSTM network containing 512 memory units. Learning rate is 10 and momentum is 0.9. The training process is shown

in Figure 4 [Figure 4: see original paper], demonstrating that the CNN-LSTM network successfully learns spatiotemporal features representing crowd emotions from sample data shortly after training begins. The training loss converges well, and test accuracy reaches a plateau at approximately 80% after 30,000 training iterations.

2 Experimental Testing

2.1 Experimental Method

Experiments are conducted using the Caffe deep learning framework based on Python. The experimental environment consists of: 1. Intel i5 2.4 GHz 2 Cores 2. NVIDIA GeForce GTX 1070 3. 8 GB RAM 4. Ubuntu 14.04×64

To evaluate the performance of the proposed multi-stream CNN-LSTM crowd emotion recognition network, the following experiments are conducted on the three-channel network: 1. Using only the original image channel 2. Using only the visual saliency image channel 3. Using only the stacked optical flow image channel 4. Averaging and fusing Softmax outputs from original and saliency image channels 5. Averaging and fusing Softmax outputs from original and optical flow image channels 6. Averaging and fusing Softmax outputs from all three channels

2.2 Dataset and Evaluation Metrics

Since existing crowd datasets primarily target group behavior analysis and lack standard crowd emotion labels, this paper constructs a crowd emotion-labeled dataset by combining the CUHK crowd dataset [19], UCF dataset [20], Web dataset [21], and PET2009 dataset [22]. Crowd emotions are categorized into four classes: Bored, Excited, Frantic, and Relaxed. Typical video scenes are shown in Figure 5 [Figure 5: see original paper]. The dataset is augmented through rotation and noise addition, resulting in 863 videos for training, 142 for validation, and 86 for testing (the test set follows [4]). The validation set is used for testing during training to prevent overfitting, while the test set evaluates the trained model' s accuracy.

Accuracy (ACC) and macro average precision (MAP) are adopted as evaluation metrics, calculated as follows:

$$ACC = \frac{\sum_{i=1}^{s} TP_i}{\sum_{i=1}^{s} (TP_i + FN_i)}$$

$$MAP = \frac{1}{s} \sum_{i=1}^{s} P_i, \quad P_i = \frac{TP_i}{TP_i + FN_i}$$

where s represents the number of crowd emotion categories, P_i denotes the precision of class i, and TP_i and FN_i indicate the numbers of correct and

incorrect predictions for class i, respectively.

2.3 Experimental Results and Analysis

To demonstrate the effectiveness of the proposed model, comparative experiments are conducted with [4]. Reference [4] employs traditional algorithms for crowd emotion recognition, using classifiers to categorize motion features learned from structured trajectories. Table 1 presents the experimental comparison between the proposed multi-stream CNN-LSTM crowd emotion recognition network and the algorithm from [4].

Table 1 ACC and MAP Results for Crowd Emotion Recognition

Method	ACC (%)	MAP (%)
Algorithm from [4]	74.9	74.3
Stacked optical flow only	76.7	78.1
Original + Saliency	79.4	80.9
Original + Optical flow	81.4	82.7
Three-channel fusion	**82.6**	**84.1**

The results show that when training models separately on original images, saliency maps, and stacked optical flow images, the original image channel achieves the best performance with ACC and MAP of 80.2% and 82%, respectively, surpassing [4]. Fusing the original and saliency image channels decreases ACC and MAP because both channels are suitable for videos with smooth motion and cannot improve recognition rates for videos with intense motion. However, fusing the original and optical flow channels improves ACC and MAP, as the original image channel performs well on smooth-motion videos (e.g., Relaxed class) while the optical flow channel excels at intense-motion videos (e.g., Frantic class). Weighted fusion enhances the model' s learning capability for videos with varying motion intensities. Ultimately, fusing all three channels achieves the highest ACC and MAP of 82.6% and 84.1%, respectively, representing improvements of 7.7% and 9.8% over [4]. This proves that the proposed model yields more accurate classification results than traditional algorithms and demonstrates deep learning' s effectiveness for crowd emotion recognition.

Figure 6 [Figure 6: see original paper] presents confusion matrices for the proposed multi-stream CNN-LSTM network and [4], showing results from fusing original, saliency, and optical flow channels. The proposed model achieves higher recognition rates for Bored, Frantic, and Relaxed classes compared to [4], with identical performance for Excited. Although the typical scenes in Figure 5 show that Bored and Relaxed classes are not highly distinguishable, the proposed model achieves the highest recognition rates for both, thereby validating its effectiveness and accuracy along with superior generalization capability compared to [4].

3 Conclusion

To address crowd emotion recognition in computer vision, this paper proposes a multi-stream CNN-LSTM network model that uses original video image sequences, saliency map sequences, and stacked optical flow sequences as inputs to three-channel CNN-LSTM networks. This architecture learns static features of scenes and crowds, local motion features, and global motion features from videos. Compared with existing algorithms, the proposed multi-stream CNN-LSTM network model achieves higher crowd emotion recognition rates, with ACC and MAP reaching up to 82.6% and 84.1%, respectively. Moreover, the entire deep network-based model requires no prior information and exhibits good generalization performance.

References

[1] Urizar O J, Baig M S, Barakova E I, et al. A hierarchical bayesian model for crowd emotions [J]. Frontiers in Computational Neuroscience, 2016, 10 (63): 1-15.

[2] Rabiee H, Haddadnia J, Mousavi H, et al. Emotion-based crowd representation for abnormality detection [J]. arXiv preprint arXiv: 1607.07646, 2016.

[3] Patwardhan A. Edge based grid super-imposition for crowd emotion recognition [J]. International Research Journal of Engineering and Technology, 2016, 5.

[4] Zhang Y, Qin L, Ji R, et al. Exploring coherent motion patterns via structured trajectory learning for crowd mood modeling [J]. IEEE Trans on Circuits & Systems for Video Technology, 2017, 27 (3): 635-648.

[5] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313 (5786): 504.

[6] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [C]// Advances in Neural Information Processing Systems. 2014: 568-576.

[7] Zhao R, Ali H, Smagt PVD. Two-stream RNN//CNN for action recognition in 3D videos [C]// Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems. 2017.

[8] Yi C, Deng Y. Multi-channel convolutional neural network image recognition method [J]. Journal of Henan University of Science and Technology: Natural Science Edition, 2017, 38 (3): 41-44.

[9] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 2012, 9 (8): 1735-1780.

[10] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2016, 39 (4): 677.

[11] Cai Y, Zheng W, Zhang T, et al. Video based emotion recognition using CNN and BRNN [C]// Proc of Chinese Conference on Pattern Recognition. 2016: 1-6.

[12] Qin Y, Mo L, Guo W, et al. Combination of 3D CNNs and LSTMs in action recognition and its application [J]. Measurement & Control Technology, 2017, 36 (2): 28-32.

[13] Borji A, Cheng M M, Jiang H, et al. Salient object detection: a benchmark [J]. IEEE Trans on Image Processing, 2015, 24 (12): 5706-5722.

[14] Brox T, Bruhn A, Papenberg N, et al. High accuracy optical flow estimation based on a theory for warping [C]// Proc of European Conference on Computer Vision. 2004: 25-36.

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014: 1-14.

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

[17] Berg A, Deng J, Li Feifei. Large scale visual recognition challenge [EB/OL]. (2010). http://www.image-net.org/challenges/LSVRC/2010/.

[18] Wu Z, Jiang Y G, Wang X, et al. Multi-stream multi-class fusion of deep networks for video classification [C]// Proc of ACM on Multimedia Conference. 2016: 791-800.

[19] Shao J, Chen C L, Wang X. Learning scene-independent group descriptors for crowd understanding [J]. IEEE Trans on Circuits & Systems for Video Technology, 2017, 27 (6): 1290-1303.

[20] Ali S, Shah M. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.]: IEEE Computer Society, 2007: 1-6.

[21] Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model [C]// Proc of Computer Vision and Pattern Recognition. 2009: 935-942.

[22] Ferryman J, Shahrokni A. PETS2009: dataset and challenge [C]// Proc of the 20th IEEE International Workshop on PERFORMANCE Evaluation of Tracking and Surveillance. 2009: 1-6.

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