3D Res-Inception Network Transfer Learning for Multiple Label Crowd Behavior Recognition

Hao Nan[1], Min Li[1], Lvyuan Fan[1] and Minglei Tong[1*]
[1] EE School, Shanghai University of Electric Power, Shanghai 200082, P.R. China
[e-mail: tongminglei@gmail.com]
*Corresponding author: Minglei Tong

Received July 4, 2018; revised September 17, 2018; accepted October 22, 2018; published March 31 2019

Abstract

The problem towards crowd behavior recognition in a serious clustered scene is extremely challenged on account of variable scales with non-uniformity. This paper aims to propose a crowed behavior classification framework based on a transferring hybrid network blending 3D res-net with inception-v3. First, the 3D res-inception network is presented so as to learn the augmented visual feature of UCF 101. Then the target dataset is applied to fine-tune the network parameters in an attempt to classify the behavior of densely crowded scenes. Finally, a transferred entropy function is used to calculate the probability of multiple labels in accordance with these features. Experimental results show that the proposed method could greatly improve the accuracy of crowd behavior recognition and enhance the accuracy of multiple label classification.

Keywords: Densely crowed group, 3D Convolutional Neural Network (3D CNN), 3D Res-Inception, Transfer Learning

Minglei Tong: received his Ph.D. from Shanghai Jiao Tong University in 2008. His research interests focus on computer vision, pattern recognition and 3D reconstruction from video. He is vice Professor in the EE School of Shanghai University of Electric Power; and a leader of Image Processing and Artificial Intelligence groups. This paper is sponsored by NSF of Shanghai (No, 16ZR1413300).

http://doi.org/10.3837/tiis.2019.03.019 **ISSN : 1976-7277**

1. Introduction

1.1 Motivation

Crowded scene analysis has attracted a considerable amount of attention of researchers in computer vision has brought great challenges to the local infrastructure and public transport facilities. To tackle these problems, a great many algorithms were proposed for crowd behavior understanding including groups dividing, small group interaction and abnormal activity detection such as riots in large crowds in order to get more accurate and comprehensive information, which could be critical for making correct decisions in high-risk environments. These algorithms have better results in some specific scenarios, such as low crowd density and fixed monitoring perspective. However, it is not suitable for complicated scenes with high-density population due to different movements of different individuals and variable environmental factors. The example of crowed scene is shown in **Fig. 1**.

Fig. 1. Example for crowed people

1.2 Recent Work

In the past decade, crowd scene understanding or analysis has already attracted much research attention on the computer vision community [1-3]. Jodoin et al. [4] utilized optical flow [5] proposed by Horn and Schunck to obtain spatio-temporal motion features for crowd movement detection, in which the particle flow is based on the fluid flow integral of the fluid dynamics. Commencing with the optical flow field estimation adopted from [6], Loy et al. [7] presented a global motion saliency detection framework. Wu et al. [8] applied the particle flow to the abnormal population behavior detection. Yang C et al. [9] raised a new feature descriptor called multi-scale optical current histogram (MHOF) to retain continuous spatial information and motion information. Wang C et al. [10] proposed to learn trajectory clustering of semantic regions, extract trajectories from dense feature points, and then use special models to enhance the spatio-temporal correlation between trajectories to detect pedestrian behavior patterns in crowded scenes. Moore et al. [11] posed the opinion that people appear as particles in a fluid in certain aspects. They used aerodynamics at both the macroscopic scale for crowd segmentation and the microscopic scale for behavior detection. Social force model (SFM) is first proposed by Helbun et al. [12]. It is assumed that the interaction force between pedestrians is an important characteristic of analysing crowd behavior. The social force model

(SFM) has been successfully employed in research fields as simulation and analysis of crowds. Framework proposed by Ali et al. [13] is utilized to compute particle flows, and their interaction forces are estimated using SFM. Shao J et al. [14] put forward collective transition (CT) based on scene-independent group descriptor used for crowd behavior recognition and pattern segmentation of crowd movement.

Recently, deep learning has further made significant progress especially in the field of video analytics and it can automatically learn images features with strong generalization ability to various pattern recognition tasks. It has achieved a series of breakthroughs, such as image classification, object detection, semantic segmentation and face recognition. The convolutional neural network [15] combines feature extraction and classification into one with local connections, weight sharing, and pooling operation reduced the complexity of the network. It also has strong robustness and fault tolerance ability. Alex created a "large, deep convolutional neural network" that was used to win the 2012 ILSVRC (ImageNet [16] Large-Scale Visual Recognition Challenge), their network "Alex Net" [17] has been successfully applied to various computer vision tasks such as target detection [18], video classification [19] and target tracking [20]. Typical CNN like methods emerged constantly, such as 3D CNN [21] network and LSTM [22]. 3D CNN network extends 2D convolution and pooling to 3D, so that neural networks have the ability to deal with spatial-temporal features. The two stream neural network [23] divides the CNN into two branches and merges the extracted features. The spatial-temporal perception ability of deep neural networks provides a new platform for the research of behavior analysis on crowd video. Ng et al. [24] combined LSTM and convolutional neural network models to extract features first from CNN frame-by-frame and then process spatio-temporal information via LSTM networks. In addition, deep learning was also applied to crowd analysis. The multi-scale convolutional neural network proposed by Zeng L et al. [25] was applied to population counting. Kang K et al. [26] applied the full convolutional network to population segmentation. Shao J et al. [27] put forward a slicing CNN for crowd scene understanding.

In order to reduce notable consideration on memory and power usage in training a new model, transfer learning was proposed to predict in a target domain acquiring knowledge from a source domain. To address the inductive transfer learning problem, Dai et al. [28] proposed TrAdaBoost, a boosting algorithm, which is an extension of the AdaBoost algorithm. Tzeng et al. [29] proposed a new CNN architecture to exploit unlabelled and sparsely labelled target domain data, this approach simultaneously optimizes for domain invariance to facilitate domain and distribution matching loss to transfer information between tasks, which exceeds previously published results on two standard benchmark visual domain adaptation tasks. Huang et al. [30] proposed a shared-hidden-layer multilingual DNN (SHL-MDNN), show that the learned hidden layers sharing across languages, while through different SoftMax layers controlled different learning tasks can be transferred to improve recognition accuracy of new tasks. Long et al. [31-32] proposed a new Deep Adaptation Network (DAN) architecture, which transfers deep convolutional neural network to the adaptation scenario. DAN can enhance the general network's feature transfer capabilities at specific task levels. They also proposed a new joint release distance measurement relationship, using this relationship to generalize deep learning ability to adapt data distribution in different fields. The DAN model is trained by fine-tuning from the AlexNet model pre-trained on ImageNet. Comprehensive empirical evidence demonstrates that the proposed architecture outperforms state-of-the-art results evaluated on the standard domain adaptation benchmarks. Oquab M et al. [33]

proposed used internal layers of the CNN can act as a generic extractor of mid-level image representation, which can be pre-trained on ImageNet, then modified last fully connected layer as an adaptive feature, only the adaptive feature layer is trained in the training process.

1.3 Proposed method

At present, most related work on crowd behavior classification is mainly focused on the task of single-label crowd classification, each of which belongs to only one category. However, many real-world perceptive tasks requires assigning more than one label to each instance such as scene spots, behavior and population groups.

In this paper, a hybrid network is firstly proposed with the use of combining 3D res-net with inception-v3 to accurately recognize the crowd behavior from a small segment of video. After introducing the network structure of 3D res-inception and loss data pre-processing, we established a transferring learning model delivering the knowledge from pre-trained network. In training process, only higher-level portion of the network is fine-tune and loss function of output layer is designed to adapt multi-label video classification. The framework of this paper is shown in **Fig. 2**. Experiments on both two datasets show that the proposed approach outperforms existing methods. In comparison to the latest work, our approach contributions can be summarized as follows: 1) A novel 3D res-inception network. 2) A transfer learning framework of our proposed model.

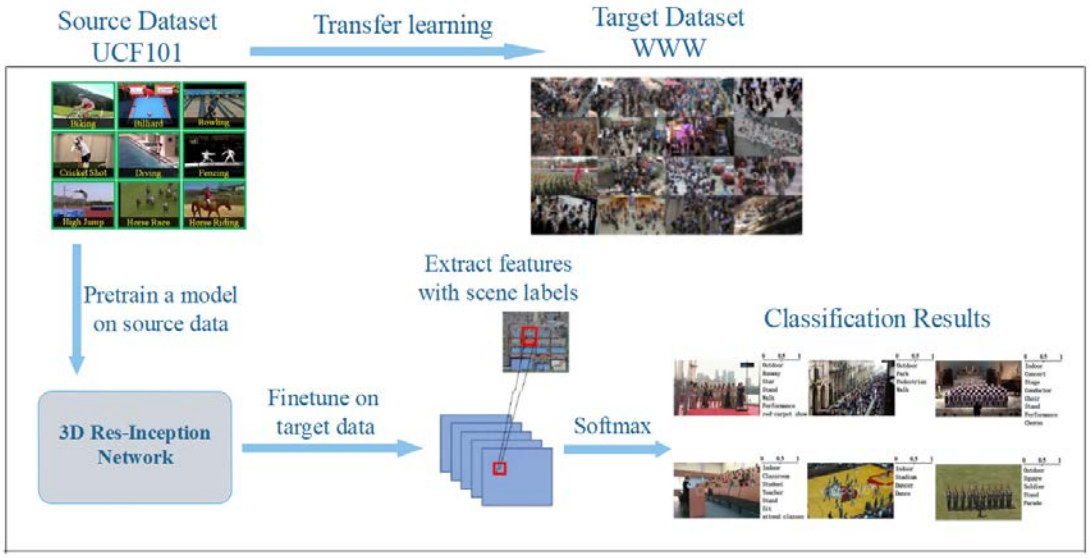

Fig. 2. Pipeline of this paper

The rest of this paper is organized as follows. Section 2 presents the approaches to 3D res-inception network. Section 3 introduces the transfer learning framework of our model. Section 4 reports the experimental results. Section 5 concludes this paper.

2. Related Work

2.1 3D res-Inception network

Building a spatio-temporal network for video processing can start with the three-dimensional expansion of a two-dimensional convolutional neural network, that is, all the convolutions and pooling are dimensionally extended. For example, N×N convolution kernel is promoted to N×N×N in which first dimension is the time dimension. Like the Inception architecture, which comprises four basic branches with different convolutions from each other. The proposed 3D Inception is shown in **Fig. 3**. Through four kinds of convolution methods, the width and depth of the network structure is expanded via four types of convolution methodology, thereby improving the network performance and enhancing the adaptability of the network to features in different scales.

$$x_{n+1} = f\left(x_n + F(x_n; W)\right) \tag{1}$$

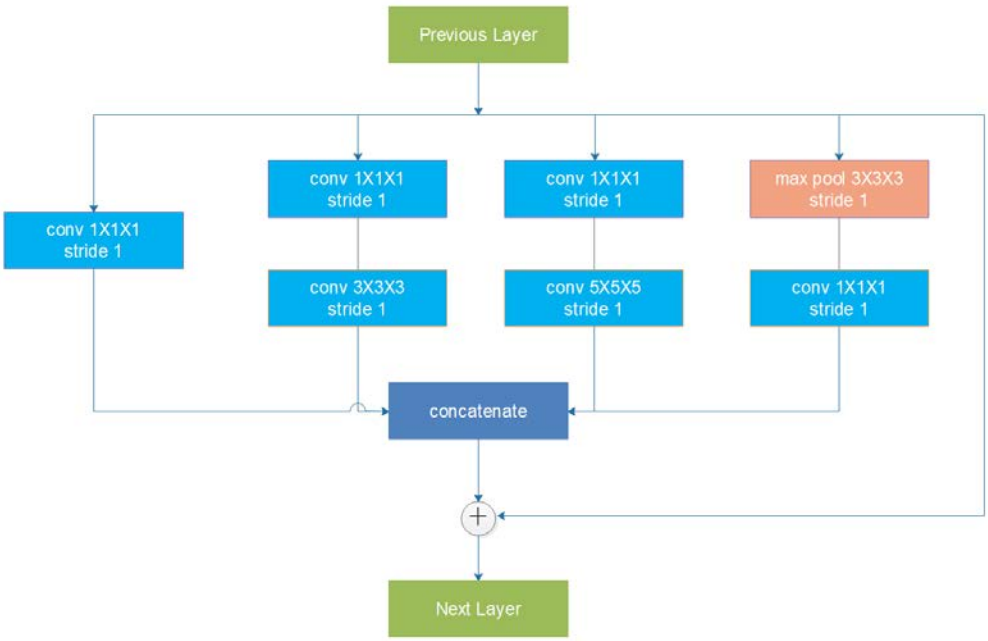

Fig. 3. Spatio-temporal residual unit

Where x_n and x_{n+1} are the input and output of the n-layer, f represents the activation function ReLU, F is the residual equation, and W represents the weight. Every 3D Inception structure adds a residual link to make up the spatio-temporal residual units.

The 3D Res-Inception network model is shown in **Fig. 4**. A total of 9 spatio-temporal residual units, 4 max-pooling levels, and 1 average-pooling level are used. A large-scale 3×7×7 convolution was employed at the beginning, and a small-scale 1×1×1 convolution was used at the end of the network. Finally, the features are flattened and output through Softmax. In addition, Batch Normalization can renormalize the activation value of the previous layer so that the average value of its output data is close to 0, and its standard deviation is close to 1. For the purpose of accelerating training and complex deep network structures, the three-dimensional convolution in each layer network structure uses batch normalization after the convolution, which can ensure that the distribution of input data at each layer is stable.

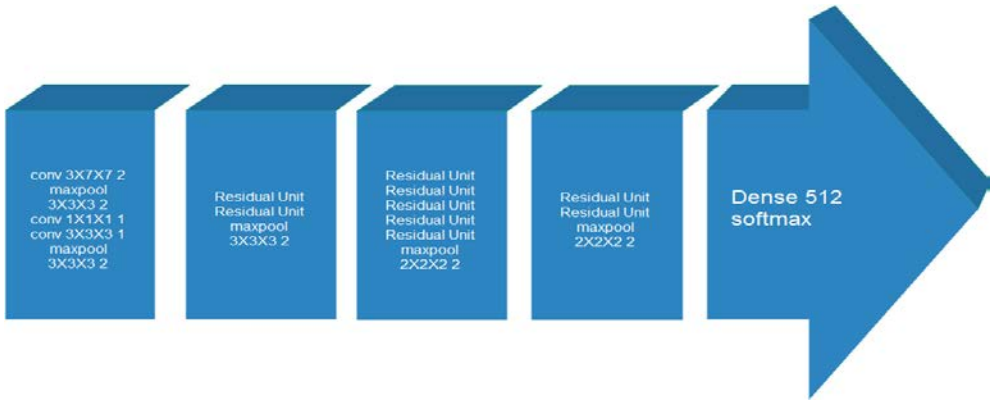

Fig. 4. 3D Res-Inception network architecture

2.2 Double –branch 3D res-inception network

In the face of dense crowd videos, one of the major difficulties is how to obtain better time and space features. The convolutional neural network can extract features, while with the deepening of the network, the obtained features are more abstract and tend to higher-level semantics. The original image sequences are all RGB image sequences while the advanced features obtained only from a single neural network which mainly includes the appearance information of the crowd, furthermore, the included motion information are still not comprehensive enough. Therefore, motion information of network is implemented inspired by the following steps. Firstly, computer vision method is used to extract low-level motion features, and then obtain its high-level features through the neural network. Additionally, the feature fusion method is used to fuse motion and appearance features to expand the overall feature dimension so that the neural network obtains sufficient spatio-temporal information. As can be seen from the following picture, we use concatenation fusion. It stacks the two feature maps at the same spatial locations i,j across the feature channels d as above :

$$y_{i,j,2d}^{cat} = f^{cat}\left(x_{i,j,d}^{a}, x_{i,j,d}^{b}\right) \tag{2}$$

where $y^{cat} \in R^{H\times W\times 2D}$, $x^{a} \in R^{H\times W\times D}$, $x^{b} \in R^{H\times W\times D}$

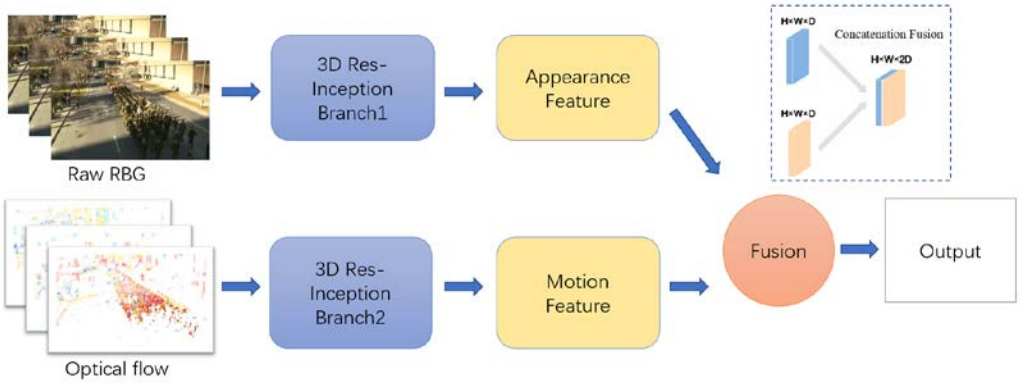

Fig. 5. Structure for crowd behavior recognition

Many methods can represent motion information. This article mainly utilizes optical flow image sequences as the input of motion branches for the following reasons: 1) Optical flow is used for calculating the crowd motion of adjacent frames. Optical flow image sequence contains information on movements of people in a relatively long period. 2) Continuous optical flow images can represent dense trace information in the image. 3) The image information obtained by the optical flow method in the edge and displacement parts are helpful for the recognition of crowd behavior.

The double-branch joint network is as shown in **Fig. 5**, two identical 3D Res-Inception networks replace the softmax layer. One branch is applied for extracting appearance features (input original RGB image sequence), the other one is used for motion features extraction. (input optical stream image sequence). The single and dual-branch networks are tested separately in sec.4 while compare with each other in order to verify the superior performance of the dual-branch joint network.

3. Domain Adaptation

3.1 Self-Organizing Map

Transfer learning involves domains and tasks, a domain D consists of a feature space X and a marginal probability distribution $P(X)$ on the feature space, $X=\mathrm{x}_1,...,\mathrm{x}_n$. Given a domain $D=\{X,\ P(X)\}$, a task T consists of a label space y and a conditional probability distribution $P(Y \mid X)$, this conditional probability distribution is usually learned from training data consisting of "feature-tag" pairs $(\mathrm{x}_i \in X, y_i \in Y)$. Given a source domain D_s , a corresponding source task T_s , a target domain D_t , and a target task T_t , the purpose of transfer learning is to learn the conditional probability distribution $P(Y_t \mid X_t)$ in the target domain D_t when $D_\mathrm{s} \neq D_\mathrm{t}$, $T_\mathrm{s} \neq T_\mathrm{t}$, and information from D_s and T_s are available. In most cases, it is assumed that the available labeled target samples are limited and far less than the source samples.

Since the 3D res-Inception network is trained on the UCF 101 dataset to perform single-label video classification training, the single-label can be converted to a multi-label classifier by modifying the training script. The SoftMax function compresses all the vector values to the range [0,1] and the sum of all its elements is 1. While in the case of multiple-label, the probability of getting a class is more than a label, each attribute should be independent of each other. Therefore, SoftMax is not suitable for multi-label classification. For instance, in a scene of a concert some people could be represented with a certain probability, and there is also a certain probability that some people stand. The following modifications can be applied: First, we take advantage of the sigmoid function instead of the SoftMax function, the sigmoid function obtain a probability value for each scalar element of each tensor, and returns the resulting probability value. Second, we choose the corresponding loss function: binary cross entropy. It is defined in the following equation:

$$L=-\frac{1}{n}\sum_{x}\left[y\ln y'+(1-y')\ln(1-y)\right] \quad (3)$$

Where y is ground truth, y' is estimation.

The representation of the mark vector of the loss function could be modified so that it can be used for both single-label and multi-label classification, our algorithm is described in Algorithm 1.

Algorithm 1: The transfer training procedure of the 3D Res-inception

Use pre-trained network 3D Res-inception model to extract feature, for every video convert into single frame do

1 Single frame converted into a frame sequence, and pass the sequence to the MLP.

2 Use pre-trained model extract feature of each frame.

3 Enter eigenvectors into a single-layer fully connected neural network.

4 Train label video and obtained the final classification result.

End for

4. Experimental Classification Results and Analysis

4.1 Datasets and Pre-processing

In this paper, we employed two crowd datasets to test our model. The CUHK Crowd Dataset (CUHK)[12] of the Chinese University of Hong Kong: It contains 474 surveillance videos of 215 scenes. According to different crowd behaviors, the dataset is divided into eight categories. There are no duplicate samples in the dataset to ensure the validity of the experiment. The WWW Crowd Dataset [15] of the Chinese University of Hong Kong which is a comprehensive crowd dataset that collects videos from films, surveillance and networks. It extracts 10,000 videos of large-scale crowd data from 8257 dense scenes. The set defines 94 crowd-related attributes, including where, who, and why, and annotates each video. In accordance with the settings in the original data set, 7220 videos for training, a set of 936 videos are used as validation, and the results of the remaining 1844 videos were tested simultaneously. These collections do not overlap in the scene to ensure that the attributes are learned independently.

In the beginning, expanding training data is a direct way when the data volume is limited. Mirroring, flipping and random clipping is effective data augmentation technique. By amplifying the data, the training sample is expanded to increase the richness of the data and the over fitting is reduced, so that the neural network learns the transformation invariance. CUHK Crowd Dataset contains 474 video samples about 6 GB are not enough. Therefore, we amplify the dataset by steps of data augmentation as follows:

1) Decompose the video data into an image sequence, and then amplify the original image sequence by 2 times the original image,

2) Since the original samples have different resolutions, all the image sequences need to be adjusted to the same size 160×240, and finally cut to 128×128,

3) Because group motions in each segmented clip remain similar across its whole length. We decompose each training sample to new with 25 frames. The data augmentation case is shown in **Fig. 6**.

Fig. 6. Data augmentation case. From the left to right, original, mirror, crop size.

First, each frame of each video is run through 3D Res-Inception, saving the output of the final pooling layer of the network in order to get a new (2048 × L) input vector is passed to the fully connected network AKA multi-layer perceptron. The MLP can infer the temporal characteristics organically from the sequence without having to know that it is a sequence. It was found that the best performing MLP is a simple two-layer network with 512 neurons per layer.

We will now have some precision values for each class, take an average of these values. This average value is called the Average Precision (AP) of the class. If it has 20 classes in whole set. For every class we will follow the same approach of calculating Average Precision. So we will now have 20 different values of Average Precision. Using these values of Average Precision, we can easily judge the performance of our model for any given class. To represent the performance of a model in one single number, (The One metric to rule them all), we take an average value (mean) of all the class as Average Precision value (MAP).

4.2 3D res-inception single branch vs two branches

Our first experiment is a comparison between single branch and two branches. Experiment results are shown on **Table 1**, we can draw a conclusion that continuous optical flow images represent dense motion information which contributes better performance on crowd behavior recognition.

Table 1. comparison between single branch and two branches on CUHK Crowd dataset

Architecture	One branch(raw RGB)	One branch(optical flow)	Two branch
AP	92.16%	87.5%	95.48%

Collective Transition (CT) based descriptors is the best algorithm without neural network in crowd behavior recognition. This method constructs a group detector, and then quantifies a set of descriptors for group state analysis and crowd behavior understanding. CNN-LSTM flattens the optical flow sequence and RGB image sequence into CNN to extract features, and then handle spatio-temporal information through LSTM. 3D CNN extends two-dimensional convolution to three dimensions. The results in **Table 2** when compared with available methods show that significant benefits can be derived from the actual application of our

proposed network.

Table 2. comparison with other methods on CUHK Crowd dataset

Method	AP
Collective Transition Based descriptors	70%
3D CNN[21]	86.05%
CNN-LSTM[24]	88.65%
Two branch 3D CNN	90.62%
3D Res-Inception	92.16%
Two branch 3D Res-Inception	95.48%

4.3 Transfer learning with 3D res-inception

In pre-trained 3D Res-inception, the output of the bottleneck layer can be divided into multiple classes of images by a single layer of fully connected layers. Therefore, the node vector output by the bottleneck layer can be used as a simpler, more expressive feature vector of any image. So on the new dataset, this trained neural network can be directly used to extracted feature vector is used as input to train a new single-layer full connection neural to handle the new classification problem.

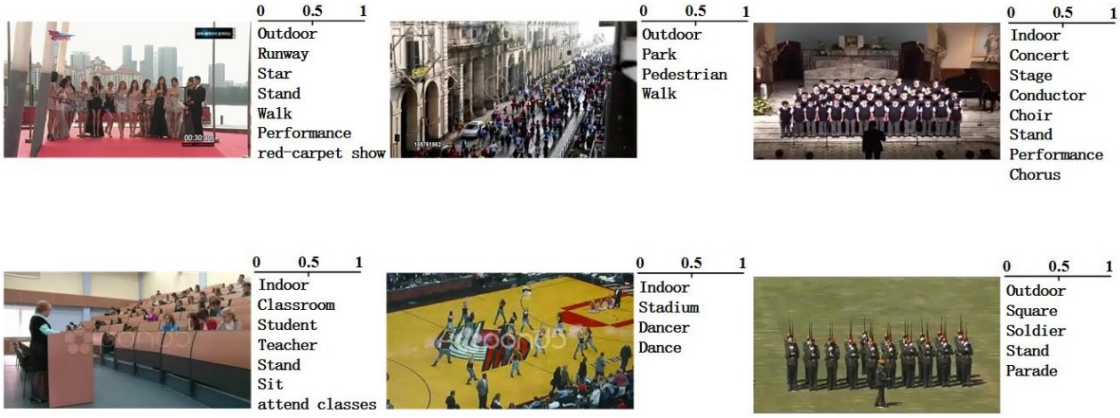

Fig. 7. Samples of MLP classification by transfer learning

Table 3. The comparison table of recognition accuracy on WWW dataset

Method	AUC	mAP
3D CNN[21]	0.67	0.26
CNN-LSTM[24]	0.71	0.33
Two branch 3D CNN	0.73	0.32
3D Res-Inception	0.83	0.41
2 Branch 3D Res-Inception Transfer + Finetune	0.91	0.52

In comparison, 3D CNN optimizing the number of iterations of the top-level fully-connected layer, and finding that the top-level fully-connected layer is fine-tuned through experiments, which can reach a maximum testing accuracy of 67%. By using 3D res-inception, an accuracy of 83% is obtained. In our proposed method, we used the bottleneck feature of the pre-training network, but we made the final of the further network. One layer was fine-tuned, weights were adjusted, and the last layer of the network was retrained. The best accuracy achieved was 91%. Qualitative recognition results on ground truth attribute annotated for the given examples are shown in **Fig. 7**.

Because there are more than one tag in pictures of multi-label image classification, this experiment adopts mean average precision (mAP). This standard is suitable for calculating actual classification tags and sorting them in predictive classification tags. The average value, the larger the better the evaluation index. Comparing the three methods, the final conclusions are shown in **Table 3**. Our method could perform better compared to the regular 3D CNN like method.

5. Conclusion

This paper presents a 3D residual unit used to extract the spatial-temporal features of crowd video. Based on this, a two-branch 3D Res-Inception deep convolution neural network is designed. The two branch combines the appearance attributes and motion features of the crowd to solve the behavior recognition problem of dense crowd. Through the experiments, the double-branch 3D Res-Inception network achieved 95.48% experimental results, which greatly improved the traditional crowd descriptor algorithm, CNN-LSTM and 3D CNN network. A transferred entropy function combining the 3D Res-Inception model is used to calculate the probability of multiple labels in accordance with these features. The proposed transfer learning enhances the accuracy of multiple label classification task to 91%.

References

[1] Saleemi, I., Hartung, L., Shah, M., “Scene understanding by statis- tical modeling of motion patterns,” in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2069–2076, 2010. Article (CrossRef Link).

[2] Yang, Y., Liu, J., Shah, M., “Video scene understanding using multi- scale analysis,” in *Proc. of IEEE International Conference on Computer Vision*, pp. 1669–1676, 2009. Article (CrossRef Link).

[3] Zhou, B., Wang, X., Tang, X., “Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents,” in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2871–2878, 2012. Article (CrossRef Link).

[4] Jodoin, P. M., Benezeth, Y., Wang, Y., “Meta-tracking for video scene understanding,” in *Proc. of IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 1-6, 2013. Article (CrossRef Link).

[5] Horn, B. K., Schunck, B. G., “Determining optical flow,” *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185-203, 1981. Article (CrossRef Link).

[6] Liu, C., “Beyond pixels: exploring new representations and applications for motion analysis,” *Ph.D. dissertation, Massachusetts Institute of Technology*, 2009. Article (CrossRef Link).

[7] Chen, C. L., Xiang, T., Gong, S., “Salient motion detection in crowded scenes,” in *Proc. of 5th Int. Symposium on Communications Control and Signal Processing*, pp. 1–4, 2012. Article (CrossRef Link).

[8] Wu, S., Moore, B. E., Shah, M., "Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes," in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 2054-2060, 2010. Article (CrossRef Link).
[9] Cong, Y., Yuan, J., Liu, J., "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognition*, vol. 46, no. 7, pp. 1851-1864, 2013. Article (CrossRef Link).
[10] Chongjing, W., Xu, Z., Yi, Z., Yuncai, L., "Analyzing motion patterns in crowded scenes via automatic tracklets clustering," *China* Communications, vol. 10, no. 4, pp. 144-154, 2013. Article (CrossRef Link).
[11] Moore, B. E., Ali, S., Mehran, R., Shah, M., "Visual crowd surveillance through a hydrodynamics lens," *Communications of the Acm*, vol. 54, no. 12, pp. 64-73, December 2011. Article (CrossRef Link).
[12] Helbing, D., Molnar, P., "Social force model for pedestrian dynamics," *Physical Review E*, vol. 51, no. 5, pp. 4282–4286, 1995. Article (CrossRef Link).
[13] Ali, S., Shah, M., "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 1–6, 2007. Article (CrossRef Link).
[14] Shao, J., Change Loy, C., Wang, X., "Scene-Independent Group Profiling in Crowd," in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 2227-2234, 2014. Article (CrossRef Link).
[15] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, November 1998. Article (CrossRef Link).
[16] Russakovsky, O., Deng, J., et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, December 2015. Article (CrossRef Link).
[17] Girshick, R., Donahue, J., Darrell, T., & Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 580-587, 2014. Article (CrossRef Link).
[18] Long, J., Shelhamer, E., & Darrell, T., "Fully convolutional networks for semantic segmentation," in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015. Article (CrossRef Link).
[19] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., "Large-scale video classification with convolutional neural networks," in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014. Article (CrossRef Link).
[20] Wang, N., Yeung, D. Y., "Learning a deep compact image representation for visual tracking," in *Proc. of Advances in Neural Information Processing Systems*, 2013. Article (CrossRef Link).
[21] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., "Learning Spatiotemporal Features with 3D Convolutional Networks," in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 4489-4497, 2015. Article (CrossRef Link).
[22] Hochreiter, S., Schmidhuber, J., "Long short-term memory." *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, November 1997. Article (CrossRef Link).
[23] Simonyan, K., Zisserman, A., "Two-Stream Convolutional Networks for Action Recognition in Videos," in *Proc. of Advances in Neural Information Processing Systems*, vol. 1, no. 4, pp. 568-576, 2014. Article (CrossRef Link).
[24] Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S. et al., "Beyond short snippets: Deep networks for video classification," in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 4694-4702, 2015. Article (CrossRef Link).
[25] Zeng, L., Xu, X., Cai, B., Qiu, S., Zhang, T., "Multi-scale convolutional neural networks for crowd counting," in *Proc. of IEEE Conference Image Processing (ICIP)*, pp. 465-469, 2017. Article (CrossRef Link).
[26] Kang, K. and Wang, X., "Fully Convolutional Neural Networks for Crowd Segmentation," *Computer Science*, vol. 49, no. 1, pp. 25-30, 2014. Article (CrossRef Link).

[27] Shao, J., Loy, C. C., Kang, K., Wang, X., "Slicing Convolutional Neural Network for Crowd Video Understanding," in *Proc. of IEEE Conference Computer Vision and Pattern Recognition*, pp. 5620-5628, 2016. Article (CrossRef Link).
[28] Dai, W., Yang, Q., Xue, G. R.,Yu, Y. "Boosting for transfer learning," in *Proc. of International Conference on Machine Learning ACM*, pp. 193-200, 2007. Article (CrossRef Link).
[29] Tzeng, E., Hoffman, J., Darrell, T., Saenko, K., "Simultaneous Deep Transfer Across Domains and Tasks," in *Proc. of IEEE International Conference on Computer Vision*, pp. 4068-4076, 2015. Article (CrossRef Link).
[30] Huang, J. T., Li, J., Yu, D., Deng, L., Gong, Y., "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7304-7308, 2013. Article (CrossRef Link).
[31] Long, M., Cao, Y., Wang, J., Jordan, M. I., "Learning Transferable Features with Deep Adaptation Networks," in *Proc. of International Conference on Machine Learning*, pp. 97-105, 2015. Article (CrossRef Link).
[32] Long, M., Wang, J., Ding, G., Sun, J., Yu, P. S., "Transfer feature learning with joint distribution adaptation," in *Proc. of the IEEE international conference on computer vision*, pp. 2200-2207, 2013. Article (CrossRef Link).
[33] Oquab, M., Bottou, L., Laptev, I., Sivic, J., "Learning and transferring mid-level image representations using convolutional neural networks," in *Proce. of the IEEE conference on computer vision and pattern recognition*, pp. 1717-1724, 2014. Article (CrossRef Link).

Hao Nan He is currently pursuing a M.S. degree in Shanghai University of Electric Power, Shanghai, China. His academic interests include computer vision and deep learning.

Min Li received her B.S. degree from Shanghai University of Electric Power She is currently pursuing a M.S. degree in Shanghai University of Electric Power, under the supervision of Dr. Minglei Tong. Her research interests include deep learning and video prediction.

Lyuyuan Fan received her B.E degree from Henan Polytechnic University in 2015. She is a master student at Shanghai University of Electric Power from 2016. Her research interest includes computer vision, and artificial intelligence.

Minglei Tong received his B.E. and M.S. degrees both from Shandong University, in 1998 and 2001, respectively, and the Ph.D. degree from Shanghai Jiao Tong University in 2008. He is a vice professor of Shanghai University of Electric Power. His research interest includes computer vision, and artificial intelligence.