Deep Learning-Based Human Action Recognition Methods Postprint

Authors: Li Yupeng, Liu Tingting, Zhang Liang

Date: 2018-11-29T00:00:00+00:00

Abstract

To address the problem of action classification confusion that occurs after mapping the four-dimensional information of human action depth videos to two-dimensional space, a deep learning-based human action recognition method is proposed. First, spatial structure dynamic depth maps are constructed to map the four-dimensional information of depth videos to two-dimensional space for information dimensionality reduction; then, a deep convolutional neural network based on a joint cost function is proposed, which combines cross-entropy loss function and center loss function as the joint cost function to guide the convolutional layers to learn more discriminative deep features for more accurate classification. Experimental results on the MSRDailyActivity3D and SYSU 3D HOI datasets demonstrate that, compared with existing methods, the recognition rate of the proposed method is significantly improved, validating its effectiveness and robustness. The proposed method effectively solves the problem of action classification confusion.

Full Text

Preamble

Vol. 37 No. 1
Application Research of Computers (ChinaXiv Cooperative Journal)
Accepted Paper

Human Action Recognition Based on Deep Learning

Li Yupeng, Liu Tingting, Zhang Liang
(Tianjin Key Laboratory of Advanced Signal & Image Processing, Civil Aviation University of China, Tianjin 300300, China)

Abstract: To address the problem of action classification confusion that occurs when mapping four-dimensional depth video information to two-dimensional space, this paper proposes a human action recognition method based on deep

learning. First, we construct spatially structured dynamic depth images to map the four-dimensional information of depth videos into two-dimensional space for dimensionality reduction. Then, we propose a deep convolutional neural network based on a joint cost function that combines cross-entropy loss with center loss to guide the convolutional layers in learning more discriminative depth features for more accurate classification. Experimental results on the MSRDaily-Activity3D and SYSU 3D HOI datasets demonstrate that the proposed method achieves significantly improved recognition rates compared to existing methods, validating its effectiveness and robustness. The method effectively resolves the problem of action classification confusion.

Keywords: depth information; human action recognition; deep learning; spatially structured dynamic depth images; deep convolutional neural network

0 Introduction

Human action recognition has extensive applications in intelligent surveillance, human-computer interaction, video retrieval, virtual reality, and other domains, making it an active research area in computer vision. Previous studies have predominantly focused on traditional RGB videos [1-4]. However, RGB video data presents numerous challenges: lack of viewpoint invariance, sensitivity to illumination and background changes, and vulnerability to noise. Although researchers have made significant progress in recent years, human action recognition remains highly challenging.

The release of Microsoft Kinect has created new opportunities for this field. Kinect devices can capture depth maps in real-time. Compared to traditional color images, depth maps offer several advantages: depth video sequences essentially represent four-dimensional space that can contain richer motion information, they are insensitive to illumination changes, and they enable more reliable estimation of human silhouettes and skeletons [5]. References [6-11] leveraged these characteristics of depth maps to design specialized feature descriptors, which have profoundly influenced the action recognition field to some extent. Liu et al. [12] proposed an enhanced skeleton visualization method that utilizes spatiotemporal sequences of skeleton points for viewpoint-invariant human action recognition, demonstrating broader practicality but remaining limited by the use of skeleton data to construct specific features. Consequently, these methods rely on hand-crafted features that provide shallow descriptions of local or global spatiotemporal information and cannot simultaneously capture important spatiotemporal and structural information in actions.

In recent years, following the tremendous success of deep convolutional neural networks (DCNN) in the ImageNet image classification competition [13], many researchers have applied models trained on ImageNet to tasks such as attribute classification [14], image representation [15], and semantic segmentation [16], achieving excellent results. However, these studies focus on image understanding

tasks for color images. Human action recognition differs from general image understanding tasks, particularly for depth information-based human action recognition, which is represented in the form of four-dimensional depth video space, making it impossible to directly apply DCNNs for recognition as in the aforementioned tasks.

Wang et al. [17] attempted to design weighted depth motion maps as input to DCNNs, converting the human action recognition problem into an image classification problem and pioneering the use of DCNNs for depth map-based human action recognition. However, experimental results indicated that this method lacked robustness. Inspired by the significant success of rank pooling (RP) proposed by Fernando et al. [18-20] in human action recognition based on color images, Wang et al. [21] built upon RP to propose spatially structured dynamic depth images (SSDDI), which overcome the limitation of RP operations suppressing fine-grained spatial motion information in depth maps and achieved higher recognition rates.

The above analysis reveals that current research efforts concentrate on designing effective feature representations that can characterize important action features in two-dimensional space after mapping four-dimensional information, thereby improving action recognition accuracy. However, our research has discovered that after mapping depth information of actions to two-dimensional representations, action classification easily becomes confused, thus limiting the upper bound of recognition rates for such methods.

To address the limitations of existing methods, through research and practice based on reference [22] and inspired by Wen et al.' s approach to solving similar problems in face recognition, we consider the problem from the perspective of feature extraction and classification mechanisms in neural networks. Combining the characteristics of depth map-based human action recognition, we propose a joint cost function-based deep convolutional neural network (JCF-DCNN) for human action recognition to improve classification accuracy and robustness. This method adds a constraint on the distance between training sample feature spaces and class centers during network training to simultaneously consider intra-class compactness and inter-class separability, guiding the DCNN to learn highly discriminative features for more accurate subsequent classification. Figure 1 [Figure 1: see original paper] illustrates the overall flow of the proposed method. Experimental results on the MSRDailyActivity3D and SYSU 3D HOI datasets demonstrate that our method significantly improves the accuracy and robustness of human action recognition.

1 Methodology

1.1 Principles of RP and BRP

An image sequence of T frames can be represented as $X = \langle x_1, x_2, ..., x_t, ..., x_T \rangle$, where $x_t \in \mathbb{R}^d$ denotes the feature vector mapped from each frame t. The average of the first t frames is represented by $V_t = \frac{1}{t}\sum_{\tau=1}^{t} \varphi(x_\tau)$. For any

temporal order r_{ij}, we assign a score, where generally, the later the temporal order, the larger the corresponding score. Therefore, the scoring function r_{ij} satisfies the constraint: $r_{ij} > r_{ji} \iff i > j$.

The purpose of the rank pooling process is to find w^* that satisfies the objective function in Equation (1):

$$w^* = \arg\min_w \frac{\lambda}{2}\|w\|^2 + \sum_{i>j} \xi_{ij}$$

subject to:

$$w^T V_i - w^T V_j \geq 1 - \xi_{ij}, \quad \xi_{ij} \geq 0$$

where ξ_{ij} is a small non-negative value, and parameter w^* can characterize the information corresponding to the image sequence at the beginning when the action has not yet started. This serves as a feature descriptor for the image sequence.

From the above analysis, RP is an unsupervised learning process that can describe an image sequence as a new feature in a two-dimensional space equal in scale to the input image. Since it contains information about spatiotemporal changes throughout the entire action process, it is called a dynamic image (DI). A dynamic image based on depth information is called a dynamic depth image (DDI).

Since the average feature V_t up to time t is used to classify frame x_t during RP, the pooled features are biased toward the starting frames of the image sequence, causing the starting frames to have greater influence on w^*. However, this is clearly unreasonable for action recognition, as there is no prior knowledge indicating which frame is more important for the task.

Bidirectional rank pooling (BRP) can significantly reduce this bias. If the process described above is called forward DDI (DDIF) generation, then reversing the image sequence order before performing RP is the backward DDI (DDIB) generation process. The method that simultaneously generates both DDIF and DDIB is BRP. Thus, each action' s depth image sequence ultimately produces a pair of images (DDIF and DDIB) after BRP processing.

1.2 SSDDI

Research in references [18,21] demonstrates that BRP is limited not only by long-term action dynamics but also by the spatial domain. Due to its unsupervised learning nature, BRP primarily encodes prominent global features in the temporal domain without simultaneously discovering discriminative motion patterns in the spatiotemporal domain. Therefore, directly applying BRP to actions causes fine-grained motion information with high discriminative value

to be suppressed by coarse-grained motion information, particularly for fine-grained actions where local spatiotemporal subspace motion information is more important than global motion information throughout the entire action process.

SSDDI addresses this problem by decomposing depth image sequences in the spatial domain into multiple parts at different granularities, performing BRP operations on each part separately, and then combining them as a new representation. Specifically, after extracting the foreground from depth action sequences and using skeleton data as guidance, the spatial domain is decomposed into three hierarchical levels: whole body region (body), partial region (part), and joint region (joint). The body level treats the entire human body containing 20 skeleton points as a single component, so its DDI after BRP becomes SSDDI. The component composition of the part level is shown in Table 1 , where each component region is determined by the maximum distance among three skeleton points, dividing the body into 9 parts as 9 components covering the entire body. Each component undergoes BRP separately to generate corresponding DDIs, which are constructed into SSDDI as shown in the left side of Figure 2 [Figure 2: see original paper]. At the joint level, each component contains 1 skeleton point. As shown in Table 2 , this level has 16 components in total, with each component region formed by expanding a certain distance from the skeleton point location. The skeleton points used for components are 16 low-noise points selected from all 20 skeleton points, with component distribution shown in the right side of Figure 2 [Figure 2: see original paper].

Figure 1 [Figure 1: see original paper] displays SSDDI at three hierarchical levels for the same action. Compared to body-level SSDDI, part-level and joint-level SSDDI are more discriminative for fine-grained action representations and can more effectively characterize actions from global to local motion and structural information. Training JCF-DCNN separately on these three levels of SSDDI and performing decision-level fusion is beneficial for improving action recognition accuracy.

2 JCF-DCNN

2.1 Network Architecture and Hyperparameter Settings

The significance of JCF-DCNN lies in its strong feature learning and classification capabilities, which can improve classification accuracy for image samples such as SSDDI. In other words, JCF-DCNN can distinguish between two samples with high similarity but belonging to different categories for correct classification, thereby reducing action classification confusion.

As shown in Figure 3 [Figure 3: see original paper], the proposed network has 12 layers, primarily consisting of 5 convolutional layers, 3 fully connected layers, and a classification layer at the backend. The joint cost function of this network's classification layer comprises cross-entropy loss and center loss functions. Table 3 records the relevant hyperparameter settings for the convolutional layers and the first two fully connected layers of this network architecture. The number

of neurons in the third fully connected layer is consistent with the number of sample categories in the corresponding database.

Since existing depth map-based action recognition datasets are generally small in scale, training a deep convolutional neural network with millions of parameters from scratch would result in overfitting. Therefore, this paper employs transfer learning, using parameters pretrained on the large-scale ImageNet dataset to initialize all convolutional layers and the first two fully connected layers of our network. During training, the corresponding network layers are initialized as described above, but the last fully connected layer is randomly initialized using a Gaussian distribution with mean 0 and standard deviation 0.01.

After extensive experimental analysis and comparison, the learning rate is set to 0.001 for the first 3k iterations and 0.0001 for the subsequent 3k iterations, achieving good performance after 6,000 total training iterations. Momentum and weight decay factors use empirical values of 0.9 and 0.0005, respectively. To further avoid overfitting, image samples are scaled to $256\$\times256 before enteringtheDCNN.Then, 224\times\$224$ regions are cropped from the center and four corners as coordinate origins, followed by mirror operations, increasing the actual training samples to 10 times the input sample size. During testing, only the central region of test image samples is cropped without mirror operations. Since only forward propagation is performed during the test phase and center loss is not involved, only the output of the cross-entropy loss function is averaged for fusion.

2.2 Cross-Entropy Loss Function

The cost function in deep networks serves as the "conductor"for the entire model, guiding network parameter learning and representation learning through back-propagation of errors between sample predictions and ground-truth labels. The cross-entropy loss function is currently the most commonly used classification loss function in deep convolutional neural networks, with the form:

$$L_S = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{{W_{y_i}}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}$$

where $x_i \in \mathbb{R}^d$ represents the i-th depth feature belonging to class y_i, W_j is the j-th column of the weight matrix $W \in \mathbb{R}^{d\times n}$ of the last fully connected layer, and $b \in \mathbb{R}^n$ is the bias term. Here, m and n represent the number of training samples per batch and the corresponding number of categories, respectively. Since the bias term has minimal impact on performance, it is often ignored for simplified analysis.

As shown in Equation (2), the cross-entropy loss function offers the advantages of simple structure and low computational cost, leading to its widespread application. However, from a practical perspective, this loss function only focuses

on which class the image to be recognized should belong to—that is, the inter-class separation problem—without considering another equally important issue: whether the space within the classifier' s decision region should uniformly belong to that class. In reality, the distance between two image samples of the same category may be larger than that between samples of different categories. Using cross-entropy loss as the neural network' s cost function in such cases can easily lead to misclassification of highly similar image samples. Notably, the cross-entropy loss function no longer performs gradient computation or back-propagation during the test phase, serving only as a function to calculate the corresponding category probability values.

2.3 Center Loss Function and Joint Cost Function

To address the limitations of cross-entropy loss, the center loss function defines a center point for each class, similar to cluster centers, with the goal of making features computed from the same class data close to their own class center—aggregating intra-class features. The farther the feature is from the center, the greater the penalty. Equation (3) formally characterizes the center loss function:

$$L_C = \frac{1}{2m} \sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2$$

where $c_{y_i} \in \mathbb{R}^d$ represents the class center of the y_i-th class depth feature. The gradient of L_C with respect to x_i can be calculated as $\frac{\partial L_C}{\partial x_i} = x_i - c_{y_i}$. The class center c_{y_i} is updated as the depth features change and can be calculated by Equations (4) and (5):

$$c_{y_i} = \frac{1}{m_{y_i}} \sum_{j=1}^{m} \delta(y_j = y_i) \cdot x_j$$

where $\delta(\cdot)$ is the indicator function. Parameter α is a scaling factor in the range $[0, 1]$ that can further optimize the neural network, making it a hyperparameter. Variables m and x_i have the same meanings as introduced in the cross-entropy loss function.

To integrate the advantages of both cross-entropy and center loss functions while simultaneously achieving intra-class compactness and inter-class separability of deep features, the proposed JCF-DCNN adopts the joint cross-entropy and center loss as the cost function for the network' s classification layer:

$$L = L_S + \lambda L_C$$

where λ is a hyperparameter introduced to control the ratio between the two loss functions. When $\lambda = 0$, the joint cost function degenerates into cross-entropy

loss. The dashed line portion in Figure 3 [Figure 3: see original paper] indicates the specific position of the joint cost function in JCF-DCNN.

3 Experiments and Results Analysis

3.1 Experimental Environment and Methods

The experimental environment uses an NVIDIA Quadro P2000 GPU with Ubuntu 14.04 operating system, configured and compiled with caffe-HAR (https://github.com/liyupeng-ing/caffe-HAR). The caffe-HAR repository, uploaded to facilitate reproduction and verification of our proposed method, contains network model files for JCF-DCNN as described in Section 2.2 and related applications, extending the Caffe deep learning framework [27] with a classification layer module featuring the joint cost function. Hyperparameters for the joint cost function are set based on empirical values, with λ set to 0.5 and α set to 0.003.

The MSRDailyActivity3D and SYSU 3D HOI datasets are used, most of whose actions involve human-object interaction processes and pose significant challenges. Both datasets contain color videos, depth videos, and corresponding skeleton data, but this paper only uses depth images and skeleton data without utilizing color images from the datasets.

The experimental process consists of training and testing phases. During training, three network models are trained separately using body, part, and joint level SSDDI from training samples. During testing, SSDDI from test samples at the three levels are input into the corresponding network models, and the output results from each model' s classification layer are fused. The label corresponding to the maximum fused score is taken as the recognition result. The fusion method used in experiments is average fusion. During the test phase, only cross-entropy loss operates in the classification layer, while center loss is not involved, meaning λ is automatically set to zero, degenerating the joint cost function to cross-entropy loss.

Note that each SSDDI corresponds to a pair of images (DDIF and DDIB). Therefore, fusion must first be performed within each level, followed by fusion across levels. Specifically, DDIF and DDIB corresponding to body-level SSDDI are input into the network model and produce two corresponding results, which are averaged as the output for that level. The same applies to part and joint levels. Finally, the outputs from the three levels are averaged again as the final result.

3.2 Recognition Results and Analysis

The MSRDailyActivity3D dataset, collected by Kinect depth cameras, contains 16 action types performed by 10 subjects, with each subject performing each action twice (once standing, once sitting), totaling 320 files. For fair comparison, training and test sample selection follows reference [7], using actions from performers 2, 4, 6, 8, and 10 for training, and performers 1, 3, 5, 7, and 9 for testing.

Table 4 compares various methods, showing that the human action recognition method based on JCF-DCNN achieves 99.38% accuracy—an improvement of 1.88% over SSDDI.

To demonstrate the role of the joint cost function in JCF-DCNN, we conducted experiments with JCF-DCNN* (center loss removed). As shown in Table 5 , except for lower fusion results at the part level, all other levels and the final result of JCF-DCNN outperform JCF-DCNN. *The lower performance at the part level may be due to lower inter-sample similarity at this level, which does not align with JCF-DCNN' s intra-class aggregation characteristics. Notably, the overall recognition rate of JCF-DCNN shows a substantial improvement of 2.8% over JCF-DCNN*, demonstrating that the joint cost function significantly enhances network performance.

The SYSU 3D HOI dataset contains 480 action files collected by depth cameras from 40 subjects performing 12 actions each. On this dataset, training and test sample selection follows reference [23]. As shown in Table 6 , our proposed method achieves 97.08% recognition rate—1.66% higher than SSDDI.

Our proposed method outperforms existing methods on both datasets, validating its effectiveness and robustness. This is primarily due to the method' s stronger feature learning and discriminative capabilities, which improve classification accuracy for image samples generated by SSDDI that have small inter-class differences but large intra-class differences. This also demonstrates that combining deep convolutional neural networks with traditional action recognition methods offers complementary advantages and holds positive significance for solving human action recognition problems.

4 Conclusion

This paper discusses existing problems in human action recognition. To address the limitations of SSDDI, we propose a human action recognition method based on JCF-DCNN. The main characteristic of this method is the use of a joint cost function combining cross-entropy loss and center loss, which provides strong feature learning and classification capabilities. It simultaneously considers intra-class compactness and inter-class separability of deep features, effectively reducing action classification confusion. Experimental results show that compared to existing methods, the proposed method significantly improves the accuracy and robustness of human action recognition.

In future work, we will conduct research on larger-scale datasets containing more action categories while attempting to design other deep learning models to further improve the accuracy and robustness of human action recognition.

References

[1] Aggarwal J K, Ryoo M S. Human activity analysis: a review [J]. ACM Computing Surveys, 2011, 43 (3): 1-43.

[2] 张良, 鲁梦梦, 姜华. 局部分布信息增强的视觉单词描述与动作识别 [J]. 电子与信息学报, 2016, 38 (3): 549-556. (Zhang Liang, Lu Mengmeng, Jiang Hua. An improved scheme of visual words description and action recognition using local enhanced distribution information [J]. Journal of Electronics & Information Technology, 2016, 38 (3): 549-556.)

[3] 王满一, 宋亚玲, 李玉, 等. 结合区域光流特征的时序模板行为识别 [J]. 系统仿真学报, 2015, 27 (05): 1146-1151. (Wang Manyi, Song Yaling, Li Yu, et al. Behavior recognition combining regional optical flow features and temporal templates [J]. Journal of System Simulation, 2015, 27 (05): 1146-1151.)

[4] 秦华标, 张亚宁, 蔡静静. 基于复合时空特征的人体行为识别方法 [J]. 计算机辅助设计与图形学学报, 2014, 26 (8): 1320-1325. (Qin Huabiao, Zhang Yaning, Cai Jingjing. Human action recognition based on composite spatio-temporal feature [J]. Journal of Computer-Aided Design & Computer Graphics, 2014, 26 (8): 1320-1325.)

[5] Shotton J, Sharp T, Kipman A A, et al. Real-time human pose recognition in parts from single depth images [J]. Communications of ACM, 2013, 56 (1): 116-124.

[6] Li Wanqing, Zhang Zhengyou, Liu Zicheng. Action recognition based on a bag of 3D points [C]// Proc of Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2010: 9-14.

[7] Wang Jiang, Liu Zicheng, Wu Ying, et al. Mining actionlet ensemble for action recognition with depth cameras [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2012: 1290-1297.

[8] Oreifej O, Liu Zicheng. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2013: 716-723.

[9] Yang Xiaodong, Tian Yingli. Super normal vector for activity recognition using depth sequences [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 804-811.

[10] Lu Cewu, Jia Jiaya, Tang Chikeung. Range-sample depth feature for action recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 772-779.

[11] Presti L L, Cascia M L. 3D skeleton-based human action classification: A survey [J]. Pattern Recognition, 2016, 53 (C): 130-147.

[12] Liu Mengyuan, Liu Hong, Chen Chen. Enhanced skeleton visualization for view invariant human action recognition [J]. Pattern Recognition, 2017, 68 (8): 346-362.

[13] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural

Information Processing Systems. Cambridge, MA: MIT Press, 2012: 1097-1105.

[14] Zhang Ning, Paluri M, Ranzato M, et al. PANDA: Pose aligned networks for deep attribute modeling [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 1637-1644.

[15] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 1717-1724.

[16] Girshick R B, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 580-587.

[17] Wang Pichao, Li Wanqing, Gao Zhimin, et al. Action recognition from depth maps using deep convolutional neural networks [J]. IEEE Trans on Human-Machine Systems. Piscataway, NJ: IEEE Press, 2016, 46 (4): 498-509.

[18] Fernando B, Gavves E, Oramas M J, et al. Modeling video evolution for action recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 5378-5387.

[19] Bilen H, Fernando B, Gavves E, et al. Dynamic image networks for action recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 3034-3042.

[20] Fernando B, Anderson P, Hutter M, et al. Discriminative hierarchical rank pooling for activity recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 1924-1932.

[21] Wang Pichao, Wang Shuang, Gao Zhimin, et al. Structured images for RGB-D action recognition [C]// Proc of IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 1005-1014.

[22] Wen Yandong, Zhang Kaipeng, Li Zhifeng, et al. A discriminative feature learning approach for deep face recognition [M]// Computer Vision. Berlin: Springer International Publishing, 2016: 499-515.

[23] Hu Jianfang, Zheng Weishi, Lai Jianhuang, et al. Jointly learning heterogeneous features for RGB-D activity recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 5344-5352.

[24] Zhou Yang, Ni Bingbing, Hong Richang, et al. Interaction part mining: A mid-level approach for fine-grained action recognition [C]// Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 3323-3331.

[25] Wan Jun, Guo Guodong, Li S Z. Explore efficient local features from RGB-D data for one-shot learning gesture recognition [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2016, 38 (8): 1626-1639.

[26] Zhang Yu, Yeung D Y. Multi-task learning in heterogeneous feature spaces [C]// National Conference on Artificial Intelligence. Palo Alto, SF: AAAI Press, 2011: 574-579.

[27] Jia Yangqing, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding [C]// Proc of the 22nd ACM International Conference on Multimedia. New York: ACM Press, 2014: 675-678.

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