Image Semantic Segmentation Approach for Studying Human Behavior on Image Data

ZHENG Zhan¹, CHEN Da²†, HUANG Yanrong³,⁴

1. School of Communication, Wuhan Textile University, Wuhan 430073, Hubei, China;
2. Walnut Street (Shanghai) Information Technology Co., Ltd., Shanghai 200051, China;
3. College of Economics & Management, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, Zhejiang, China;
4. Research Center for Digital Economy and Sustainable Development of Water Resources, Hangzhou 310018, Zhejiang, China

Abstract: Image semantic segmentation is an essential technique for studying human behavior through image data. This paper proposes an image semantic segmentation method for human behavior research. Firstly, an end-to-end convolutional neural network architecture is proposed, which consists of a depth-separable jump-connected fully convolutional network and a conditional random field network; then a jump-connected convolution is used to classify each pixel in the image, and an image semantic segmentation method based on convolutional neural network is proposed; and then a conditional random field network is used to improve the effect of image segmentation of human behavior and a linear modeling and nonlinear modeling method based on the semantic segmentation of conditional random field image is proposed. Finally, using the proposed image segmentation network, the input entrepreneurial image data is semantically segmented to obtain the contour features of the person; and the segmentation of the images in the medical field. The experimental results show that the image semantic segmentation method is effective. It is a new way to use image data to study human behavior and can be extended to other research areas.

Key words: human behavior research; image semantic segmentation; hop-connected full convolution network; conditional random field network; deep learning

CLC number: TP751

Introduction

The research on human behavior is based on the big data method. It includes three aspects: numerical data analysis, text data analysis, and image data analysis. With the rapid development of mobile devices and Internet technology, the image has become an essential carrier for more and more people to record their behavior. It records the moment of behavior and is an important information source for analyzing behavior character-
istics. With the iterative innovation of artificial intelligence technology, image semantic segmentation technology provides us with a creative means to extract and understand the information in the image. The development of this technology can be roughly divided into three stages\(^1\): ① the method of traditional image semantic segmentation technology stage can only achieve image segmentation, not image semantic segmentation; ② The semantic segmentation stage combining deep learning and traditional technology — using the Convolutional Neural Networks (CNN) algorithm to achieve the effect of semantic segmentation; ③ The Fully Convolutional Networks (FCN) for semantic segmentation is the mainstream image semantic segmentation technique nowadays.

CNN have shown extraordinary ability in image data processing. As the task of image semantic segmentation needs to classify each pixel of the input image, the general CNN cannot do it, and it needs to be designed manually, but it has the disadvantages of long training time and low accuracy.

Aiming at the above problems, this paper proposes an end-to-end convolutional neural network architecture to realize the semantic segmentation of human behavior images. The network consists of a deep Separator Stride Fully Convolutional Network (SSFCN) and a Dense Conditional Random Field (DCRF). The algorithm is named "SSFCN-DCRF image semantic segmentation deep learning algorithm for human behavior research".

This article has the following innovations:

1) An end-to-end convolutional neural network architecture (Separator Stride Fully Convolutional Network and Dense Conditional Random Field, SSFCN-DCRF) for semantic segmentation of human behavior images is proposed, consisting of a deeply separable SSFCN and DCRF.

2) A skip-connected fully convolutional network SSFCN is proposed, which combines convolutional and pooling layers with skip-connected convolution to classify each pixel in the image. A semantic segmentation method based on convolutional networks is proposed.

3) Using the DCRF network to improve the image segmentation effect of human behavior. In this paper, we propose to implement linear modeling based on semantic segmentation of conditional random field images by using Gibbs distribution and introducing energy function.

We use fully connected conditional random fields and introduce two random variables — hidden and observed — to implement nonlinear modeling based on semantic segmentation of conditional random field images.

### 1 Related Work

#### 1.1 Traditional Image Semantic Segmentation Algorithms

The traditional method of image semantic segmentation is to design some features manually, and then use the classifier in machine learning to complete the semantic segmentation task. According to the different classification algorithms, it can be divided into three categories: ① Image semantic segmentation based on threshold segmentation. The image is converted to a gray image. Then several segmentation thresholds are determined, and then the pixels in the gray image are divided into corresponding categories according to each defined threshold. Each segmentation threshold corresponds to a segmentation map. ② Image semantic segmentation based on the clustering algorithm. The image semantic segmentation problem is transformed into a clustering problem in machine learning, and the segmentation graph can be output after the iteration of the clustering algorithm. The common clustering algorithms include K-means clustering\(^5\), mean shift clustering\(^9\), Gaussian mixture model clustering\(^4\) and agglomerative hierarchical clustering\(^8\). Different clustering targets can be divided into RGB color value clustering\(^6\), gray value clustering\(^7\) and pixel spatial position clustering\(^9\). ③ Image semantic segmentation based on graph theory is to treat the problem of image semantic segmentation as the minimum cut problem in graph theory. Think of an image as an undirected weighted graph, expressed by \(G=(V, E)\), where \(V\) represents each pixel in the image, \(E\) represents the connection between pixels, and the weight of edges represents the difference in the correlation between adjacent pixels. A partition graph of the image corresponds to a partition \(S\) in the undirected weighted graph, and each sub-region \(C\) \((C, S)\) in the partition corresponds to a sub-graph in the graph. Common segmentation methods based on graph theory include GraphCut\(^9\), GrabCut\(^9\) and Random Walk\(^4\).

#### 1.2 State-of-the-Art Research on Image Semantic Segmentation Based on CNN

With the success of CNN in some fields of computer vision (such as face recognition\(^12\), object detec-
tion\textsuperscript{[13]}, pedestrian recognition\textsuperscript{[14]}, etc.), more and more researchers have applied them to image semantic segmentation\textsuperscript{[15]-[17]}.

The network architecture based on full convolutional networks\textsuperscript{[18]} predicts a single pixel directly, which is an end-to-end training method. The network architecture of FCN\textsuperscript{[19]} is in the form of a coder-decoder. The encoder, namely the feature extraction module, is used to extract the image of the features. The decoder, that is, the upsampling module, is used to output the final split image. The decoder needs to expand the size of the feature map, such as linear interpolation and bilinear interpolation. Decoders can also accomplish sampling in the form of an inverse convolution, such as SegNet\textsuperscript{[20]}, DeconvNet\textsuperscript{[21]}, and Cipola SegNet\textsuperscript{[22]}.

1.3 Human Behavior Research by Image Semantic Segmentation Method

1.3.1 Human behavior research

The use of image semantic segmentation to study human behavior can be divided into three levels. The first level of low-level vision is human detection, which extracts the category of "human" in the picture. The second level of intermediate vision is human tracking, which targets the person in the picture and recognizes the movement characteristics of the person. The third level, high-level vision, is behavioral understanding, the process of assigning meaning to human actions.

According to the complexity of the technology used, the behavior of using image semantic segmentation to study can be divided into three categories. ① Recognition and detection of individual human behavior, such as Zhang et al\textsuperscript{[23]} who built a video-based human abnormal behavior judgment process, Shao et al\textsuperscript{[24]} proposed a key-frame cascade recognition network and space-time map convolution AST-GCN (Attention ST-GCN). ② Multi-person interaction behavior recognition and detection. Wang et al\textsuperscript{[25]} pointed out that multi-person interaction behavior recognition and detection need to pay attention to the relationship between people, which carries the key information to interpret group behavior. The group behavior recognition methods are summarized and classified into: conventional non-interaction type, model based on interaction relationship, model based on key person interaction and multiple decision fusion type\textsuperscript{[26]}. ③ Detection of behavior tracks. Behavior is often not an instantaneous event, but a process composed of multiple nodes. Based on deep network learning, Hu\textsuperscript{[26]} extracted the internal characteristic law from the complex pedestrian movement path and predicts people's movement path in the next stage.

1.3.2 Dynamic scenario research in time and space dimensions

At present, it is a hot topic to consider the real dynamic situation in time and space dimensions. Ji et al\textsuperscript{[27]} proposed to capture spatial and temporal information from video and Wald et al\textsuperscript{[28]} presented a new neural network architecture of 3D data to explore the relationship between entities, and returned semantics from a given 3D scene through learning. Fernando et al\textsuperscript{[29]} combined motion dynamics into the image and feed the image into any standard CNN for end-to-end learning. Bilen et al\textsuperscript{[30]} proposed a dynamic method to represent the motion of image sequence by time sequence. Khowaja et al\textsuperscript{[31]} performed local sparse segmentation using global clustering to construct semantic images.

To summarize, the method of image semantic segmentation based on the convolutional neural network has the following two problems: first, image semantic segmentation needs to classify each pixel of the input image, which is generally not possible with a convolutional neural network and is mainly designed manually; second, the image segmentation algorithm has the disadvantage of long training time and low accuracy. To address these two problems, this paper proposes a deep learning algorithm for image semantic segmentation for human behavior study.

2 Image Semantic Segmentation Method Based on Convolutional Neural Networks (CNN)

2.1 End-to-End CNN Architecture

To achieve image semantic segmentation of human behaviors, this paper proposes an end-to-end convolutional neural network architecture consisting of a depth-wise separable skip connected fully convolutional network (SSFCN) and a conditional random field network (DCRF). SSFCN classifies each pixel in the image, and DCRF improves human behavioral image segmentation.

2.2 Basic Building Blocks of CNN

2.2.1 Convolution

The convolution layer is composed of multiple convolution cores. The input of the convolution layer is a multi-channel image. The convolution kernel performs a convolution operation on the input image in a specific
The process of extracting different levels of feature information with convolution kernels of different sizes.

Convolution neural network is a multi-layer network structure. The convolution cores of different layers receive the output of the previous convolution core, and take the output as the input of the next convolution core. Generally, the lower convolution kernel detects the more basic visual features (such as horizontal lines, vertical lines, etc.), and the higher convolution kernel detects the more specific visual features (such as circles, boxes, etc.). The convolution kernel of different layers can be combined to extract most features. The final result of the convolution operation is a characteristic graph.

2.2.2 Pool layer

The role of pooling layers is to solve the problem of parameter explosion caused by simply stacking convolution layers. It compresses the input characteristic image by sampling. Common pooling compression methods used include maximum pooling and average pooling.

The pooling layer has two characteristics: (1) feature invariance. That is, when an image undergoes simple transformations such as flipping, translation, rotation and scaling, it can also extract the same features at the same location. (2) Feature dimension reduction. That is, after the pooling layer, the feature map is reduced, reducing the size of the input of the next layer, and reducing the amount of calculation and parameters of the entire network structure. For example, the input image is 224×224×3, and after the maximum pooling operation with a step of 2, its feature map is compressed to 112×112×3.

2.2.3 Full connection layer

The role of fully connected layers (FC) is "classifier". By introducing non-linear transformation, it maps the learned "distributed feature representation" to the sample marker space and transmits it to the classifier.

The feature map is mapped to probability to classify the whole network at the full connection layer. As shown in Fig. 2, a feature map of the size of 3×3 is expanded into three one-dimensional vectors of the size of 1×3 by row, and finally spliced into a one-dimensional vector of 1×9 in sequence providing input for the classifier.

After the full connection layer, the \((-\infty, +\infty)\) fraction \(z\) of \(K\) categories is obtained. In order to obtain the probability belonging to each category, the fraction is first mapped to \((0, +\infty)\) through \(e^z\), and then normalized to \((0, 1)\). The output of the full connection layer cannot directly represent the probability that the image belongs to a category. Therefore, the softmax \((1)\) formula must be used for conversion, see formula (1).

\[
a(Z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}
\]

where \(j=1, \ldots, K\), \(K\) represents the total number of targets to be classified. In this way, the output of each neuron is mapped to the probability of belonging to a specific cat-
category and meets the requirement that the sum of the mapping values of all neurons is 0.

The full connection Layer+softmax views the picture from three perspectives: ① Weighting, which takes weight as the importance of each dimension feature; ② Template matching to help understand the visualization of parameters; ③ From a geometric point of view, the feature is regarded as a point in a multi-dimensional space. The properties of different types of points can help understand the design idea behind some loss functions.

2.3 Image Semantic Segmentation Based on FCN

The image semantic segmentation algorithm based on a convolution neural network is completed by FCN. The convolution layer is used to replace the final full connection layer to achieve the classification of each pixel in the image. The network structure is divided into two parts: feature extraction and upper sampling. First, the combination of the convolution layer and pooling layer is used to extract the image’s visual characteristics. Then the deconvolution method is used to restore the size of the feature image to the size of the original image, and the final segmented image is generated.

3 Deep Learning Algorithm for Image Semantic Segmentation in Conditional Random Field Networks

The key to image recognition of human behavior is semantic segmentation of the image. To solve the problem of poor image edge segmentation caused by the direct application of a full convolution neural network, this paper proposes a dense conditional random field network (DCRF) to optimize the segmentation results output by a full convolution neural network SSFCN.

3.1 Segmenting Image Semantics with DCRF Networks

Because the classical conditional random field model has a large number of connection edges, directly applying it to image semantic segmentation will cause too much computation due to too many image pixels. The mean field approximation algorithm in variational inference will be used for approximate calculation. One iteration of the mean field approximation algorithm will be designed to represent the convolution layer. Meanwhile, the multiple iterations of the algorithm will be expressed as the form of the cyclic neural network to form the conditional random field network DCRF. The whole network will be trained using the backpropagation algorithm. The deep learning algorithm of human behavior image segmentation proposed in this paper is to combine conditional random field network DCRF and full convolution network SSFCN to form an end-to-end image semantic segmentation network SSFCN-DCRF, which is called SSFCN-DCRF image semantic segmentation deep learning algorithm. The whole network can be trained using a backpropagation algorithm.

The key of this algorithm is image semantic segmentation modeling based on DCRF, including linear and nonlinear parts.

3.2 Linear Modeling for Semantic Segmentation of Images in DCRF Networks

Gibbs distribution is used for linear modeling of image semantic segmentation based on the conditional random field \(^\text{[15]}\). Gibbs distribution is the probability distribution of the undirected graph model expressed by factor. The specific expression of Gibbs distribution is shown in formula (2).

\[
P(X_1, X_2, \cdots, X_n) = \frac{1}{Z(X)} \tilde{P}(X_1, X_2, \cdots, X_n) \tag{2}
\]

where

\[
\tilde{P}(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} \phi_i(X) \tag{3}
\]

\[
Z(X) = \sum \prod \phi_i(X) \tag{4}
\]

Equation (2) is the normalization coefficient, and equation (3) is the factor function. In order to make this model convenient for image semantic segmentation, the implementation form of the factor function is redefined, as shown in (5).

\[
\phi_i(X) = \exp(-\xi(X)) \tag{5}
\]

where \(\xi(X)\) is the energy functions.

Finally, the linear model of image semantic segmentation based on a conditional random field is obtained, as shown in formula (6).

\[
P(X_1, X_2, \cdots, X_n) = \exp \left( \sum_{i=1}^{n} \xi_i(X) \right) \tag{6}
\]

It can be seen that due to the introduction of the energy function, the elements in equation (3) are multiplied by the natural logarithm to become additive. That is, the multiplication relationship between the elements becomes the addition relationship.
3.3 Nonlinear Modeling of Semantic Segmentation of DCRF Images

The nonlinear modeling of image semantic segmentation based on a conditional random field will use the fully connected conditional random field (DCRF) to model the semantic segmentation problem. Because DCRF meets the Gibbs distribution, as shown in formula (7):

\[ P(Y|X) = \frac{1}{Z(X)} \hat{P}(Y, X) \]  \hspace{1cm} (7)

where

\[ \hat{P}(Y, X) = \exp \left( \sum_i w_i \psi_i(Y_i, X) \right) \]  \hspace{1cm} (8)

\[ Z(X) = \sum_Y \exp \left( \sum_i w_i \psi_i(Y_i, X) \right) \]  \hspace{1cm} (9)

In equations (7), (8), and (9), \( Y \) is called the hidden variable and \( X \) is called the observed variable. In image semantic segmentation, \( Y \) represents the category label to which the pixel belongs, and \( X \) represents the information that each pixel can be directly observed. In the process of modeling with DCRF, it is necessary to calculate the joint probability density of two random variables \( Y \) and \( X \), which is equivalent to the energy function. In this section, we designed several energy functions to obtain the final modeling form of DCRF for the image semantic segmentation task, that is, nonlinear modeling based on conditional random field image semantic segmentation, as shown in formula (10):

\[ \hat{P}(Y, X) = \exp \left( \sum_i \psi_i(X_i, Y_i) + \psi_j(Y_i, Y_{i+1}) \right) \]  \hspace{1cm} (10)

In formula (10), \( \psi_i(X_i, Y_i) \) is a first-order energy function, representing the information entropy brought by assigning a category label \( i \) to a pixel point \( i \); \( \psi_j(Y_i, Y_{i+1}) \) is a second-order energy function, which is the case that two-pixel points are assigned at the same time.

When all pixels in the image are connected to each other in pairs. The second-order energy function can be expanded; see formula (11).

\[ \psi_i(x_i, x_j) = u(x_i, x_j) \sum_{k=1}^{K} w^k \psi^k(f_i, f_j) \]  \hspace{1cm} (11)

Set the kernel function \( k(f_i, f_j) \) to:

\[ k(x_i, x_j) = \mu(x_i, x_j) \]

\[ w_1 \exp \left(-\frac{(r - \mu)^2}{2\sigma^2} - \frac{1}{2\sigma^2} \right) + w_2 \exp \left(-\frac{(r - \mu)^2}{2\sigma^2} \right) \]

\[ \mu(x_i, x_j) \]

\[ \sigma \]

In formula (12), \( \mu(x_i, x_j) \) represents pixel \( x_i \) and pixel \( x_j \). The information entropy between \( j \) is \( \mu(x_i, x_j) = 1 \) only if \( x_i \neq x_j \), otherwise \( \mu(x_i, x_j) = 0 \). The above conditions mean that only nodes with different labels will generate information entropy, and the information entropy between nodes with the same label is 0. In formula (12), two different Gaussian kernel functions are also defined according to the difference of feature space: the first Gaussian kernel will consider the pixel position expressed as \( \rho \), the pixel gray value expressed as \( i \), and the second Gaussian kernel will only consider the spatial position of the pixel expressed as \( i \). Hyperparameters \( w_1 \) and \( w_2 \) control the size of two Gaussian kernels. Hyperparameters \( \sigma_\rho \), \( \sigma_i \), and \( \sigma_\gamma \) are used to control the weight of position and color information in each Gaussian kernel. In terms of effect, the first Gaussian kernel will encourage similar category labels to be assigned to pixels with similar color and position. In contrast, the second Gaussian kernel only considers the spatial correlation between pixels.

4 Case

This section presents two cases of using the proposed method. Case 1 is used for entrepreneur image segmentation and case 2 is used for medical image segmentation, which is an extension of human behavior research.

4.1 SSFCN Image Segmentation of Entrepreneurial Images

4.1.1 Data set — Entrepreneur image

The data set used in this paper is the entrepreneur image crawled from the government website. By defining different fields, 39 entrepreneurs representing various industries were selected. There are 600 original images in total. The original images are expanded to 1 950 by using data enhancement methods (random rotation, rotation, cropping, random setting of image brightness and contrast, and random left and right rotation in training to generate more images), with 50 for each enterprise. 1 500 of them are used as training sets and 450 as test sets.

4.1.2 Image segmentation based on SSFCN

Entrepreneur data sets are trained using the full convolution neural network structure. The structure of the full convolution neural network is shown in Fig. 3.
As shown in Fig. 3, the full convolution neural network structure network is composed of a feature extraction path (left side) and an upper sampling path (right side). The feature extraction path is composed of three feature extraction modules. Each feature extraction module contains three deep separable jump concatenated convolutions with different convolution core sizes (3×3, 5×5, 7×7). In order to strengthen the feature extraction capability of the network, the number of convolution cores in the three feature extraction modules is set to 64, 128, 256. It should be noted that the output of the merge layer in each feature extraction module is saved as an intermediate result. In the upper sampling path, these intermediate results will be used again. In the up-sampling path, each up-sampling module consists of a common convolution layer with a convolution core size of 3×3 (step side=1) and a transposed convolution layer with a convolution core size of 2×2 (step side=2). After each upper sampling module, the size of the feature map will be enlarged to twice the original size. Before entering the next upsampling module, the output feature map of each upsampling module will be fused with the corresponding intermediate results saved in the feature extraction path (add pixel by pixel) to generate a new feature map. For example, the output of the first upsampling module will be fused with the output of the third feature extraction module, and the new feature map generated after the fusion will be used as the input of the second upsampling module. At the network’s last layer, this case uses the convolution layer with the convolution core size of 1×1 to map the 64-dimensional feature vector to the required number of classes (here we have two classes, and the number of convolution cores is set to 2).

In this case, the image segmentation network proposed in this paper can be used for semantic segmentation of the input image, and the segmentation results will show the contour features of the characters.

Using the segmentation model proposed in this paper can not only support the study of human behavior, but also has good performance for image segmentation in the medical field, which shows that the method has some generalization and universality.

### 4.2 Segmentation of Images in the Medical Field

The dataset used in this case is the Kaggle lung CT dataset. This dataset is a set of two-dimensional CT scans of the lungs of cancer patients and the corresponding segmentation maps after manual calibration. The CT scans and the corresponding segmentation maps are shown in Fig. 4 (a) and (b), respectively. The number of original images is 58 and the image size is 512×512 pixels. To better utilize the dataset, we used the data enhancement method to expand the data of the original im-
ages (58 images), and finally get 500 images. We divided the enhanced dataset into two parts, where 80% of the images (400 images) are used for the training set and 20% (100 images) are used for the validation set. For the test set, we used 58 images from the original dataset. It can be seen that our segmentation model has good segmentation performance for images in the medical domain as well, with some generalization.

4.3 Comparison of Segmentation Effects

The comparison of segmentation performance (accuracy) between the method proposed in this article and current mainstream methods is shown in Table 1. The current mainstream methods include: IDSIA [31] (inverse distance spatial interpolation algorithm), SegNet [20] (a deep convolutional encoder-decoder architecture for image segmentation), U-Net [34] (convolutional networks for biomedical image segmentation), DeepLabV3 [25] (deep lab semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs).

<table>
<thead>
<tr>
<th>Method</th>
<th>Entrepreneur</th>
<th>Kaggle lung</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDSIA</td>
<td>88.0</td>
<td>61.0</td>
</tr>
<tr>
<td>SegNet</td>
<td>93.6</td>
<td>85.2</td>
</tr>
<tr>
<td>U-Net</td>
<td>95.9</td>
<td>87.0</td>
</tr>
<tr>
<td>DeepLabV3</td>
<td>96.5</td>
<td>90.5</td>
</tr>
<tr>
<td>Our method</td>
<td>98.2</td>
<td>94.1</td>
</tr>
</tbody>
</table>

5 Conclusion

In this paper, for the problem of studying human behavior with image data describing human behavior, we propose an image semantic segmentation method for human behavior study. Specifically, the following researches are done and innovated.

1) An end-to-end convolutional neural network architecture (Separator Stride Fully Convolution Network-Dense Conditional Random Field, SSFCN-DCRF) is proposed, which consists of a depth-separable jump-connected fully convolutional network SSFCN and a conditional random field network DCRF.

2) The jump-connected convolution is used to classify each pixel in an image, and a convolutional neural network-based semantic segmentation method for images is proposed.

3) A Conditional Random Field Network (Dense Conditional Random Field (DCRF)) is used to improve the effect of image segmentation of human behavior, and linear modeling and nonlinear modeling based on semantic segmentation of Conditional Random Field images are proposed.

4) Full convolutional network image segmentations are implemented for entrepreneurial images and images in the medical field.

However, in the experimental part, there are some limitations in our research. In this paper, the segmentation of full convolutional network images of entrepreneur images, only made the results, has not yet carried out a comparative analysis of the effect of image segmentation. In the future, we will continue to collect entrepreneur image data, conduct a more comprehensive comparative analysis of image segmentation effects, and further expand into other areas.

References


