A Vision Transformer Model with Compressive Sensing for crowd density level classification

1st Dan Sun
Soochow University
Suzhou, China
20225246029@stu.suda.edu.cn

2nd Jin Zhang
Soochow University
Suzhou, China
zhangjin1983@suda.edu.cn

3rd Jie Sheng
Soochow University
Suzhou, China
jsheng@suda.edu.cn

4rd Cheng Wu
Soochow University
Suzhou, China
cwu@suda.edu.cn

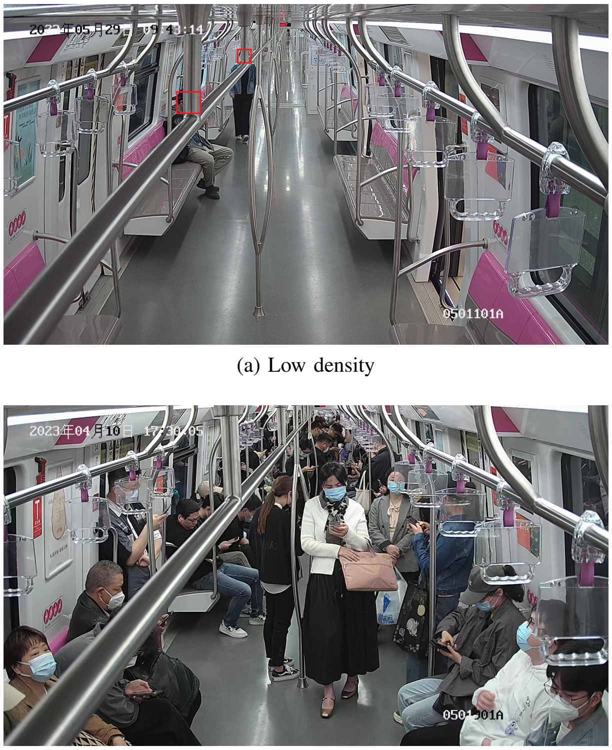
***Abstract*—In this paper, a Vision Transformer Model with Compressive Sensing for crowd density level classification is proposed. Crowd density level classification is an important crowd monitoring task that is widely used in public places. However, the performance of existing methods degrades when dealing with heavily occluded scenes because they have difficulty in extracting complete and accurate crowd features. To solve this problem, we perform compressed perception on selected image blocks after removing occlusions, and then feed the results into the transformer backbone network, which outputs classification results from the density classification task head. We conducted experiments in a typical occlusion scenario of subway cars, and the results show that our approach achieves relatively good results.**

***Index Terms*—vision transformer, compressive sensing, crowd density classification**

I. Introduction

In recent years, video surveillance systems in many places have provided a good way to collect data for crowd density monitoring and analysis. Analyzing crowd density, as an important task of crowd monitoring, is widely used in public places such as subways, shopping malls, stations and squares. In large-scale crowd flow in public places, counting by hand is slow and inaccurate, which can easily lead to dangerous situations like crowding and trampling. Therefore, the use of intelligent video surveillance technology to monitor crowd density is a necessary choice.

In the actual crowd density classification application, crowd density estimation also faces many challenges, such as the occlusion problem. The occlusion problem occurs when in an image or video, objects or people block part or all of the human body, making it hard to extract and estimate crowd features. The subway car, shown in Figure 1, is a typical scene with serious occlusion problems. Subway cars have limited space and high frequency of people flow. This often causes a large number of occlusions between people and objects such as handrails and railings. These occlusions make images or videos show only partial or scattered features such as heads, bodies, and postures. Therefore, complete and accurate crowd information cannot be obtained.



(a) Low density

(b) High density

Fig. 1: Occlusion in the subway car scene. The head information is obscured by the railing in (a); the crowd density increases and the crowd is heavily obscured in (b).

There are two main categories of existing crowd density classification methods: detection-based methods and regression-based methods. The detection-based methods refer to the method of locating each person in an image or video, and then counting the number of people and calculating the density. This type of method performs better in sparse or medium-density crowd scenes, but in high-density or severely occluded scenes, they are prone to missed or false detection due to the difficulty of the detector to locate each head or body, which leads to inaccurate estimation results. Regression-based methods are methods that learn the mapping relationship between image or video features and the number of people or density, and then directly output the estimation results. This type of method performs better in high-density or heavily occluded scenes, but there are some problems. For example, how to deal with crowd features of different scales and distributions, how to reduce the complexity and resource consumption of the model, etc.

To address the above problems, this paper proposes an image density classification method based on compression perception and ViT (Vision Transformer). As a pre-processing method for downscaling and sparsification of input images, compression perception has also been introduced into the study of crowd density estimation in recent years. The core idea of compressive perception is to exploit the sparsity of the signal in a certain transform domain, and then to randomly subsample the signal at a density sparser than the sampling density required by the Nyquist sampling frequency. By using a special reconstruction method, the original signal can be recovered. This can realize the process of data compression during the sampling process and break the golden rule in signal processing - Nyquist's law of sampling. We use ViT [1] as the backbone model for density classification to further improve the effectiveness of crowd density classification. Because CNN relies on local texture information and ignores global and local contextual relationships, existing methods usually use convolutional neural networks as the backbone model, but the performance of CNN degrades when processing images with occlusions and texture changes. For the occlusion problem, [2] found that the Transformer-based model is significantly more robust than the CNN-based model. Our approach requires detection and removal of occluded regions from the image. Then, we segment the removed image into multiple sub-blocks and perform feature extraction and classification for each sub-block. ViT with its powerful self-attention mechanism and multi-headed attention mechanism can fuse the features between different sub-blocks to improve the accuracy and robustness of classification. This is because the removed images may have missing or discontinuous information.

II. Method

The method proposed in this paper has three main modules: compressive sampling, ViT encoding, and a density-level classification task head. Firstly, the occluded objects in the image are removed, and then the removed image is chunked, and each sub-block is input into the compressive sampling module for processing to obtain the measurement value of each sub-block. Secondly, the obtained measurements are input into ViT backbone network for feature extraction and classification, and the final classification results are obtained after the density level classification task head.

The compressive sampling module is divided into two main parts, sampling and reconstruction. In the traditional compressed sensing method, assuming a one-dimensional sparse signal $x \in \mathbb{R}^N$ and a measurement matrix $\Phi \in \mathbb{R}^{M \times N}$, the measurement value $y = \Phi x$, where $y \in \mathbb{R}^m (M \ll N)$. To recover the signal x from the measurement value y is to solve the l_1 parametric optimization problem. However, two-dimensional images are more informative, and obtaining the whole projection for the whole image with the measurement matrix will make the projected data larger and require more storage space. This also leads to an exponential increase in the computational complexity of image reconstruction. The block-based compressive sensing (BCS) method is proposed to solve this problem. BCS is a lightweight compressive sensing method that divides the image into equal-sized, non-overlapping image blocks instead of processing the whole image. Then, each image block is sampled and reconstructed using a small measurement matrix.

The image X of size S is divided into N image blocks of size $B \times B$. The i-th image block is x_i, then $X = [x_1, x_2, x_3, ..., x_N]$, the block-based measurement matrix is Φ_B, and the measurement value of each block is $y_i = \Phi_B x_i$. where $\Phi_B \in \mathbb{R}^{M_B \times B^2}$, $M_B = cs \times B^2/S$ is the number of samples per sub-image block, and cs is the compression-aware ratio. Finally the compressed sampling module gets the measurement $Y = [y_1, y_2, y_3, ..., y_N]$ of the whole image, i.e. the output of compressed sampling is $Y = \Phi_B X$.

The measurements obtained by compressive sampling are used as input to the ViT encoding module, and information features are extracted from the serialized measurements. the ViT encoding module contains a linear projection layer and multiple transformer encoder layers. Each Transformer Encoder layer contains two sub-layers: a multi-headed self-attentive mechanism and a multi-layer perceptron. Each sub-layer is followed by a residual join and a layer normalization. The output of the image measurements input to the linear projection layer is $Z^0 = Q + P = W_B Y + P$. The transformer encoder layer takes Z^0 as input, and each layer contains a multi-headed self-attentive mechanism and a feedforward neural network, as well as residual connections and layer normalization, as shown in the following equations:

$$Z'^l = MSA(LN(Z^l)) \tag{1}$$

$$Z^l = MLP(LN(Z'^l)) + Z'^l, l = 0, 1, 2, ..., K-1 \tag{2}$$

Where K is the number of layers of the Transformer layer, the output matrix Z of the last Transformer Encoder layer is used as the output of the whole Transformer encoder for the subsequent density level classification task.

The density level classification task header contains two linear projection layers and a layer normalization. The input is a 768-length class token $Z^K = [Z_1, Z_2, Z_3, ..., Z_N]$, where N is the number of classes, and the final output is the density rank of the judgment.

We collected and compiled a dataset of subway car scenes and experimentally showed that with a compression ratio of 10%, an accuracy of 95.969% could still be achieved.

References

[1] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[2] Naseer M M, Ranasinghe K, Khan S H, et al. Intriguing properties of vision transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 23296-23308.