Master Thesis

Cross-view Non-local Neural Networks for Joint Representation Learning between First and Third Person Videos

（クロスビュー・非局所ニューラルネットワークによる
自己視点映像と固定視点映像間の共通特徴量の学習）

Zhehao Zhu
Advisor: Professor Yoichi Sato
Submission Date: Jan 28th, 2021

Department of Information and Communication Engineering
Graduate School of Information Science and Technology
The University of Tokyo

Advisor
Prof. Yoichi Sato

Abstract

A first-person video captured by a wearable camera provides observation from the egocentric perspective for video understanding of human activities. In contrast, a video captured by a fixed camera observes the same activities from the third-person perspective, i.e., outside views. Since first and third person videos provide complementary information, jointly using such videos may contribute to better understanding of human activities. In order to conjointly analyze first and third person videos, learning a joint representation of both views which could describe both views with a unified model and transfer knowledge crossing the views is necessary.

The key challenge of learning a joint representation first and third videos is how to find correspondence of common objects appearing in both viewpoints and how to learn a joint representation of the objects that can share information across the views. In this paper, we propose a cross-view non-local neural network to learn joint representation from first and third person videos. The core of our method is a non-local model to extract and enhance the global visual feature similarity between both views while reducing dissimilarity. We also introduce hierarchical average pooling and zero-centered correlation matrix to the typical non-local modules which may prove the performance of non-local block from different aspects.

Our method was evaluated on an action recognition benchmark dataset from cross-view videos. We execute multiple experiments and the proposed model achieve overall state-of-the-art performance both qualitatively and quantatively.

Contents

List of Figures 1

List of Tables 3

1 Introduction 5
- 1.1 Background and Motivation . . . 5
- 1.2 Challenges and Contributions . . . 6
- 1.3 Thesis Outlines . . . 8

2 Related Work 11
- 2.1 Cross-view Representation Learning . . . 11
  - 2.1.1 Joint Attention Guided Representation Learning . . . 12
- 2.2 Non-local Neural Networks . . . 13
  - 2.2.1 Non-Local Modules . . . 14

3 Proposed Method 17
- 3.1 Model Architecture . . . 17
  - 3.1.1 Feature Extraction . . . 17
  - 3.1.2 Cross-view Non-local Module . . . 17
- 3.2 Unbiased Hierarchical Non-local Module . . . 19
  - 3.2.1 Hierarchical Pooling . . . 20
  - 3.2.2 Zero-centered Correlation Matrix . . . 20
- 3.3 Loss Function . . . 21
  - 3.3.1 Frame Screening . . . 22
- 3.4 Implementation Details . . . 22

4 Experiments 25
- 4.1 Experimental Settings . . . 25
  - 4.1.1 Dataset . . . 25
  - 4.1.2 Tasks and Evaluation . . . 25
- 4.2 Quantitative Analysis . . . 26
- 4.3 Qualitative Analysis . . . 27
  - 4.3.1 Visualization of Feature Activation . . . 27
  - 4.3.2 Visualization of the Correlation Maps . . . 29
- 4.4 Experiments on Downstream Tasks . . . 30
  - 4.4.1 Video Action Recognition . . . 30
  - 4.4.2 Gaze Prediction . . . 31

5 Discussion **33**
5.1 The Dependencies that Non-local Modules Concern about 33
5.2 Limitations of the Proposed Method 34

6 Conclusion and Future Work **35**

Acknowledgments **39**

References **41**

List of Figures

1.1 Overview of our motivation 6
1.2 Challenges of matching common regions. 7
1.3 Dependencies captured by non-local networks 8

2.1 The overall structure of non-local block 14

3.1 Overview of proposed model. 18
3.2 Improved non-local block 20
3.3 Examples for good and bad frames. 22

4.1 Data screening. 26
4.2 Visualization of feature activation. 28
4.3 Visualization of correlation maps 29

5.1 High-level dependencies captured by non-local networks 34

List of Tables

4.1 Quantitative comparisons on two tasks 27
4.2 Action recognition task. 30
4.3 Gaze prediction task. 31

1 Introduction

1.1 Background and Motivation

Recently, with the development of digital camera technology, a massive number of digital videos and images are captured every day. Especially after the advent of smartphones and high-speed mobile Internet, people may record their lives and upload videos or images to the Internet anytime and anywhere. These rich visual data contain a large amount of human behavior information, which makes it possible to analyze human behavior based on deep learning technology. Video-based human action analysis has become an important research direction in the field of computer vision and artificial intelligence. The analysis of human behavior has prerequisite significance for subsequent semantic segmentation, human-computer interaction, and augmented reality.

Videos captured from third-person viewpoint take a dominant position among all the videos. Considering the classic way of holding a camera, whether it is handheld or fixed shooting, it is difficult for the camera operator to perform other actions and have a high degree of interaction with the environment or other people. The common feature of these third-person videos is that the photographer of these videos is a independent object with no connections with other objects or people, and the video mainly records the interaction between other people and the environment. The advantage of third-person videos is that they can record human behavior and movements from a more objective and holistic perspective. This feature provides great convenience for early video action analysis.

However, with the increasing popularity of wearable cameras, such as Google glass [1], videos captured from first-person viewpoint provide a new perspective to observe human actions. Contrary to third-person videos, first-person videos take the camera wearer as the subject and mainly record the behaviors and scene interaction of the camera wearer. Due to the limitations of the perspective and field of view of the wearable cameras, first-person videos cannot record the complete movement of the camera wearer (actor), and the recording range of the scene is also narrower. However, since the first-person video directly reflects what the actor sees, the information such as camera ego-motion which is hidden in first-person videos can more accurately imply the actor's attention and thought tendency, which are not

[1] https://www.google.com/glass/start/

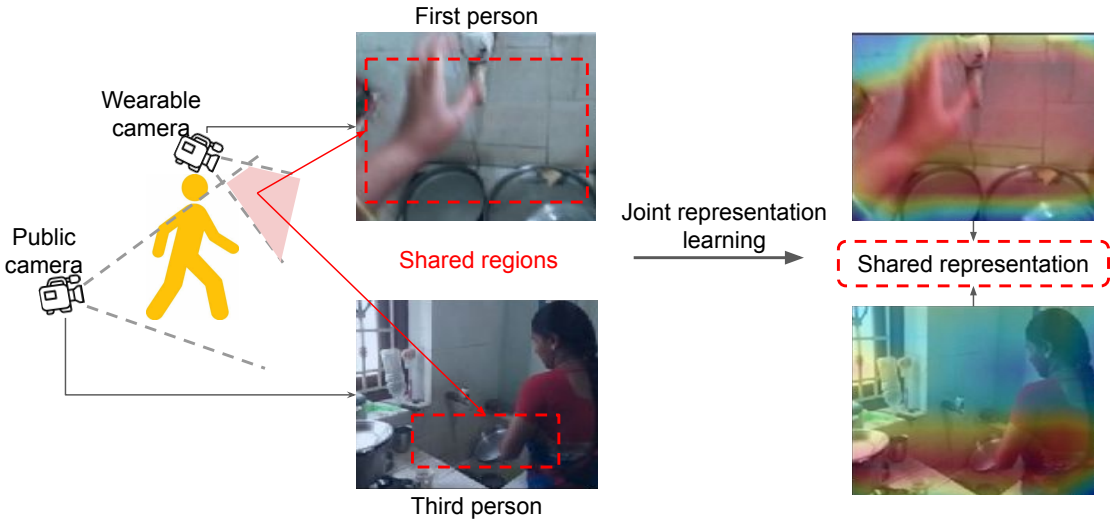

Figure 1.1: Our model learns a joint representation from the correspondence between first person and third person videos. The representation transfers information from the bystander's to the actor's perspective.)

available in the third-person video. Researches on first person action understanding [JG15, LGG12, LYR15, PR12, RK17, RM13] have been opened the door recent years.

Since the first-person and third-person videos capture the same world from different perspectives, both of them are important. Considering a scenario that an actor is taking actions in a scene while a first and a third-person video are captured simultaneously recording the actor's actions, there are a part of the same objects and human body in these two videos, and each video will also record some unique details. By observing and matching the common parts of the two videos and combining the complementary information, we humans can build up a more comprehensive description of the relationship between people and scenes. This means that it is also possible to jointly analyze videos from two perspectives through computer. In order to conjointly analyze first and third person videos, learning a joint representation of both views is necessary. The joint representation learning means to build up relationship between the common entities appearing in both views and establish a unified model describing the visual information of those entities. Through the unified model, knowledge should be able to transfer crossing the views. (see Figure 1.1)

1.2 Challenges and Contributions

The first fundamental challenge of joint representation learning is how to find out correspondence between first and third person views. Since the captured regions

and perspectives from the two views are in huge difference, it is challenging to localize the common regions and then build up relationship between them. First, in most cases, the common objects that appears in both videos may have different relative positions, shapes and sizes (Figure 1.2(a)). In addition, due to different light conditions, objects may also have different color and texture information. This means that some low-level feature matching algorithms based on geometry, such as SIFT[Low99], almost doesn't work. We need to explore more high-level correspondence, Secondly, because objects may be occluded from each other in space, some common areas may be overlapped by images of other objects (Figure 1.2(b)). This requires that the model not only pay attention to the visual information of objects themselves, but also pay attention to the connection between objects and objects.

(a) differant shape, size and location (b) overlapping

Figure 1.2: Challenges of matching common regions.

After determining the correspondence between first and third person views, The second challenge comes to how to extract a cooperative model that can share information crossing views. As mentioned above, because the appearance of an object changes significantly according to the perspective, feature extraction models based on limited perceptual fields such as traditional CNN cannot effectively describe cross-view features. We need a model that can effectively perform feature matching and joint feature presentation over all features equally.

Faced with mentioned challenges, a model called non-local network [WGGH18] shows superiority compared to other feature extraction structure. Convolutional and recurrent operations depends on a local perception window, therefore the long-range dependencies can only be detected by propagating signals iteratively in network. In this case, the weight will be gradually diluted with the process of propagation. Instead, non-local networks take all inputs as equal status and capture global dependencies directly by computing interactions between any two positions of inputs, regardless of their positional distance (See in Figure 1.3). However, original non-local network wasn't designed to deal with objects in different size, which will be briefly explained in the following chapter. In order to solve this problem, we propose a hierarchical structure inside non-local block which may handle the issues of object size difference on the basis of ensuring the original functions of non-local networks.

Overall, in this paper, our main contributions are summarized as follows:

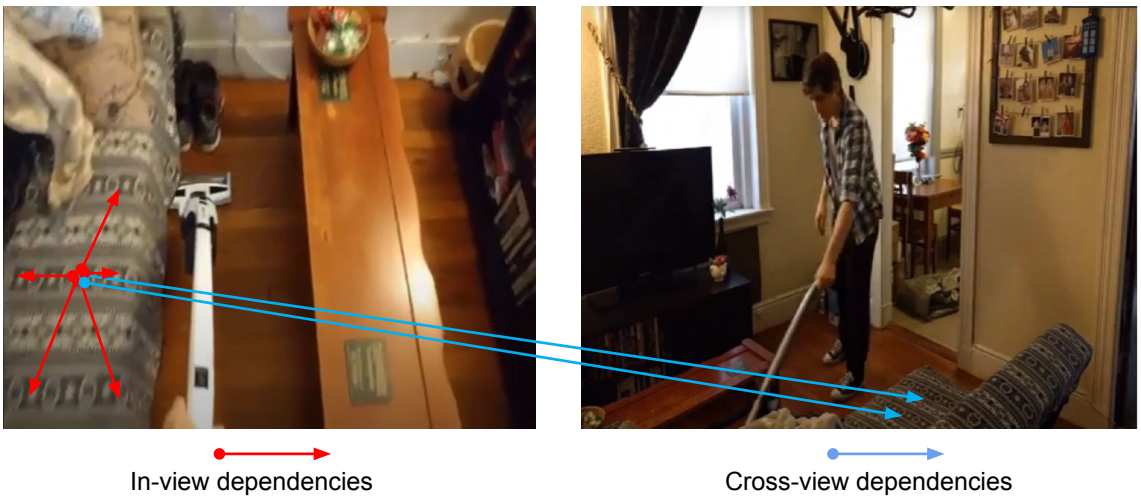

Figure 1.3: The arrows in the figure intuitively show what long-range dependencies refers to. Red arrows refer to in-view dependencies and blue arrows refer to cross-view dependencies.

- Firstly, We propose a novel method based on non-local networks for learning joint representation between first and third-person videos. To our best knowledge, our model is the first work to introduce non-local operations in this area. Our model addresses the challenges that are mentioned: 1) difficulties of establishing connections between obviously different object projections; 2) weak ability of traditional network to deal with long-range dependencies.
- Secondly, We make two improvements on the structure of non-local module. 1) We introduce hierarchical average pooling to the correlation matrix computing. This allows the non-local model to deal with multiple sizes of perception fields and thus solving the problem that objects may show totally different size; 2) We introduce the zero-centered correlation matrix to guarantee that the features are normalized and zero-centered . The two novel change improve the performance of non-local block from different aspects.
- Thirdly, we evaluate our model on public dataset with multiple tasks. The proposed model outperforms the state-of-the-art both qualitatively and quantitatively on the joint representation learning task.

1.3 Thesis Outlines

The rest of this thesis is organized as follows. In Chapter 2, we first provide an overview of recent related works on cross-view representation learning and non-local neural networks. After that, three closely related methods are described in detail. We then propose our method in Chapter 3, our method includes the base model

for joint representation learning, and details of improved of non-local model. In Chapter 4, we design multiple experiments to evaluate our method and show its superiority over other baseline methods. In Chapter 5, how non-local operations actually work is qualitatively analyzed. Current limitations are also presented and possible solutions and other modifications are discussed. Finally, in Chapter 6, we summarizes this thesis.

2 Related Work

The main goal of our paper is to realize joint representation between first and third-person view videos, therefore papers relating to cross-view representation learning [RB18, RB19, YKS15, AB16, FLX+17, SGS+18a, YCLL19] are highly related to our task. Besides, in our model, the kernel method serving the goal is based on non-local neural networks, thus papers about non-local neural networks [WGGH18, BCM11], their improved version and extended applications[ZXB+19, YYC+20, MFZ+20] are also related to our work.

2.1 Cross-view Representation Learning

Cross-view representation learning is a general expression for a large class of research. It is committed to establishing a unified model for videos or pictures from two or more perspectives that are related in reality, and uses the model to transfer information between perspectives. It is apparent that our task is a specific case. In this field of researches, two types of structures are currently widely used. One structure is based on encoder-decoder structure to encode visual information, and attempts to reconstruct the image of other views by using the encoded features of one view, in order to obtain a compressed encoded model applied to multiple views. The other type of method is based on the self-supervised learning model of the siamese structure. The methods artificially introduce non-correlated negative samples in the correlated multi-view image sample pairs. By designing loss functions, models are constrained to strengthen the similarity between correlated samples and amplify the differences between non-correlated samples, and finally enable the model to describe the commonalities between pairs of corresponding samples.

The encoder-decoder model is widely used in cross-view synthesis task. The task is defined as generating the simulated image of one view from the other view which picture the same scene. Regmi *et al.* [RB18] took two images of aerial view and street view showing the same scenery as input, and applied a two-step Generative Adversarial Networks (GANs) [GPAM+14], which is a typical encoder-decoder based image synthesizing structure. [RB19] further introduced attention mechanism based on previous work which was able to extract information more efficiently.

In cross-view matching tasks, the siamese structure is widely considered as backbone of models. Yonetani *et al.* [YKS15] proposed a novel face detection algorithm which

was based on motion correlations between actor (first-person) and observer (third-person) videos. Ardeshir *et al.* [AB16] studied a human-human matching task which matched humans appearing in two third-person view videos. Fan *et al.* [FLX+17] studied a similar task that matched multiple actors (first-person camera wearers) appearing in a single third-person video to multiple first-person videos. Both [AB16] and [FLX+17] were based on an improved version of siemese structure and depended on the groundtruth of human bounding box. Sigurdsson *et al.* [SGS+18a] firstly studied frame-frame matching problem, which entirely depended on self-supervision and didn't need any kinds of object level groundtruth. While [SGS+18a] proposed a pure CNN siemese structure, [YCLL19] noticed the limitations of CNN structure and tried to resolve them by introducing attention module. Our work is mostly related to [YCLL19] and reached better performance on dealing with cross-view representation than it.

In the following subsection, we introduce the most important related study with us. This work cared about similar challenges we faced and proposed an attention-based solution.

2.1.1 Joint Attention Guided Representation Learning

Yu et al. [YCLL19] studied a frame-frame matching task. The task is defined as follows: There are three video frames taken as input, two of which are first-person view frames and the other is third-person. One of the first-person frame is taken simultaneously with the third-person frame (i.e. corresponding frame pairs), while the other one is non-corresponding. The task is to determine which first-person frame is corresponding to the third person frame. They noticed that the common area between corresponding image pairs was generally salient area that the actor paid more attention to and fixed gaze on. Therefore, They applied a attention-based module to focus on those salient areas.

As a result, they proposed a Joint Attention Guided Representation Learning Network. In their model, they took Cbam [WPLK18] model as the attention feature extraction module. After extracting deep visual features separately from three frames, the Cbam module generated three channel weight vectors for feature maps, which weighted the importance of different high-level features represented by different channels. After that, the loss function forced the channel weight vectors of corresponding frames to tend the same, and also forced the feature maps weighted by vectors of corresponding frames to tend the same.

In experiment section, they evaluated their model with frame alignment experiment. The baselines were pure CNN backbones and AONet [SGS+18a]. Comparing to the baselines, their method achieved to the best both quantitatively and qualitatively. However, their attention module assumes that each features extracted from different objects must appear in different channels of feature maps so that they can be

weighted by channel vectors. The assumption wasn't convincing enough when the layers of deep network wasn't deep enough.

2.2 Non-local Neural Networks

The key idea in our paper of resolving the cross-view feature dependencies is the non-local neural networks. The concept of non-local neural networks was firstly proposed by Wang et al. [WGGH18]. They were inspired by a traditional denoising algorithm called non-local means [BCM11]. As the name suggests, non-local means is a non-local average algorithm. Different from local average filtering algorithms that smoothly average the area around a target pixel, non-local mean filtering means that it uses all pixels in the image for filtering, and these pixels are weighted and averaged according to a certain degree of similarity. The key idea should be that more areas with the similar properties contribute more robust denoising effect.

[WGGH18] borrowed the idea of non-local means and expands it to deep networks. They pointed out that both convolution and recurrent are operations performed on a local area, so that they are typical local operations. Instead, they proposed a non-local operation to capture long-range dependencies, that is, how to establish the connection between two pixels with a certain distance in the image, how to establish the connection between two frames in the video, how to establish the connection between different words in a paragraph, and so on (See in Figure 1.3). Similar to non-local means, non-local operations take the weighting of all spatiotemporal features as account when computing the feature of each single pixel. They evaluated the non-local network on video action classification tasks.

Given that non-local neural networks demonstrate excellent performance on dealing with comprehensive correlations, it was quickly applied in other computer vision tasks. [ZXB+19, YYC+20] applied non-local operations in semantic segmentation. non-local module was used to deal with the encoded features in encoder-decoder structure. Zhu et al. [ZXB+19] noticed that the non-local operations are time-expensive when taking large-size feature maps as input. They solved the problem by down-sampling the input. Yin et al. [YYC+20] further improved the model by decoupling the constant component and normalized local component in non-local correlation matrix. Mei et al. [MFZ+20] applied non-local blocks on image super resolution. We observe that the application of non-local networks focused on problems that have needs for transferring knowledge spatially or temporally, which is highly related to our task.

In the following subsection, we will briefly introduce the structure of non-local modules for convenience of introducing our proposed method in Chapter 3.

2.2.1 Non-Local Modules

Let us start from a non-local means filter. Given a noisy image $v = v(i)|i \in I$, the estimated value of pixel i, $\hat{v}(i)$ is computed as :

$$\hat{v}(i) = \frac{1}{Z(i)} \sum_{j \in I} w(i,j)v(j), \tag{2.1}$$

where the w worked as weight and $Z(i)$ is the normalization factor.The weights are measured by the Gaussian similarity between the neighbourhood pixel sets of i and j, i.e.:

$$w(i,j) = e^{-\frac{\|v(N_i)-v(N_j)\|^2}{h^2}}, \tag{2.2}$$

where $v(N_k)$ refers to a vector containing all pixel values of a square neighborhood of fixed size which centered at pixel k and h acts as degree factor.

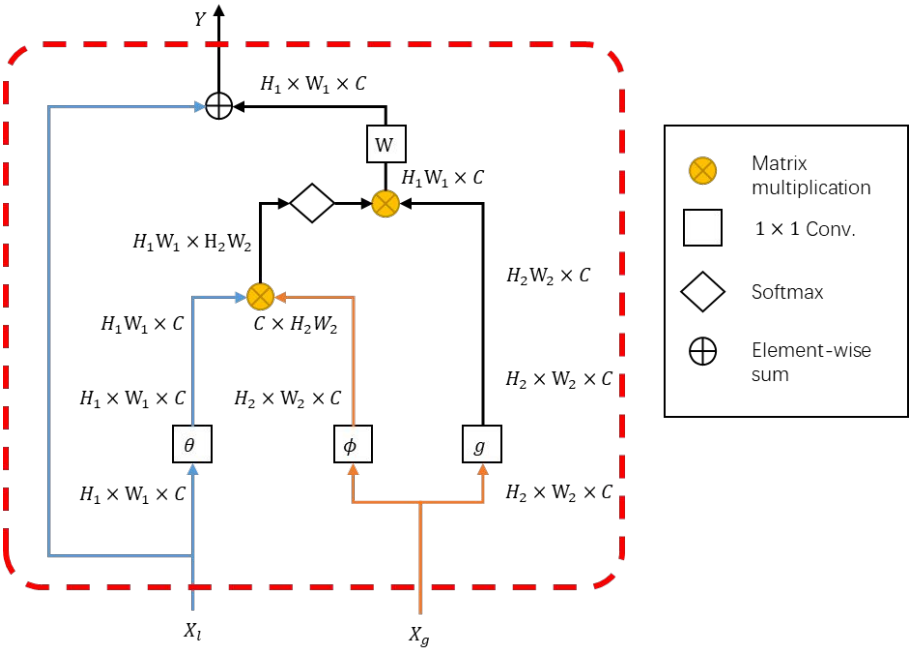

Figure 2.1: The overall structure of non-local block.

When it comes to non-local neural networks, similarly, the operation is defined as:

$$y(i) = \frac{1}{Z(\mathbf{x})} \sum_{\forall j} f(\mathbf{x_i}, \mathbf{x_j}) g(\mathbf{x_j}) \tag{2.3}$$

where $\mathbf{x}$ is input feature map and $\mathbf{y}$ is output map. $Z(\mathbf{x} = \sum_{\forall j} f(\mathbf{x_i}, \mathbf{x_j}$ works as normalizing factor. g is a 1×1 convolutional layer. $f(\mathbf{x_i}, \mathbf{x_j})$ works as similarity

function between features x_i and x_j. Multiple available functions were given in the paper and the most practical of all is given as:

$$f(\mathbf{x_i}, \mathbf{x_j}) = e^{\mathbf{x_i}^T \mathbf{x_j}}, \tag{2.4}$$

After turning the equations into matrix formula and introducing residual connections [HZRS16], the overall structure is shown as Figure 2.1.

They attached the non-local blocks inside the traditional CNN networks and evaluated their performance on video action recognition tasks. They analysed the results and observed that only several non-local layers significantly improved the overall performance of the network.

3 Proposed Method

3.1 Model Architecture

We propose a self-supervised joint representation learning model based on non-local operations. The overview architecture of the model is presented as Figure 3.1. The framework is composed by a multi-branch neural network and takes a triplet of frames (x, y, z) as input. x, y, z refer to corresponding first-person, third-person, non-corresponding first-person frames separately. It consists of two modules: feature extraction and cross-view non-local module.

3.1.1 Feature Extraction

The function of this module is to first extract regional basic deep visual features from the triplet inputs. These features will contain the semantic, visual texture and other information hidden in the images. There is a backbone CNN model which extracts feature maps $(\mathbf{F}_x, \mathbf{F}_y, \mathbf{F}_z)$ for each input branch. These visual features will be utilized to calculate the global correlations between features by non-local operations in the next module.

3.1.2 Cross-view Non-local Module

The typical non-local blocks take a single feature map $\mathbf{x}$ as input. Each estimated output $\mathbf{y_i}$ is computed as the weighted average of all input features from $\mathbf{x}$, shown as:

$$\mathbf{y_i} = \frac{1}{Z(\mathbf{x})} \sum_{\forall j} f(\mathbf{x_i}, \mathbf{x_j}) g(\mathbf{x_j}). \tag{3.1}$$

where $\mathbf{x}$ is input feature map and $\mathbf{y}$ is output map. $Z(\mathbf{x} = \sum_{\forall j} f(\mathbf{x_i}, \mathbf{x_j}$ works as normalizing factor. g is a 1×1 convolutional layer. The similarity function f is the same as 2.4. In this case, each feature vector $\mathbf{x_i} \in \mathbf{x}$ actually acted as two roles: 1) when $\mathbf{y_i}$ is computed, $\mathbf{x_i}$ acts as a target feature, which is anchored to compute similarities with all other features; 2) when $(\mathbf{y_j}, j \neq i)$ is computed, $\mathbf{x_i}$ acts as a key value, constituting a part for computing the weighted mean. Therefore, the input of

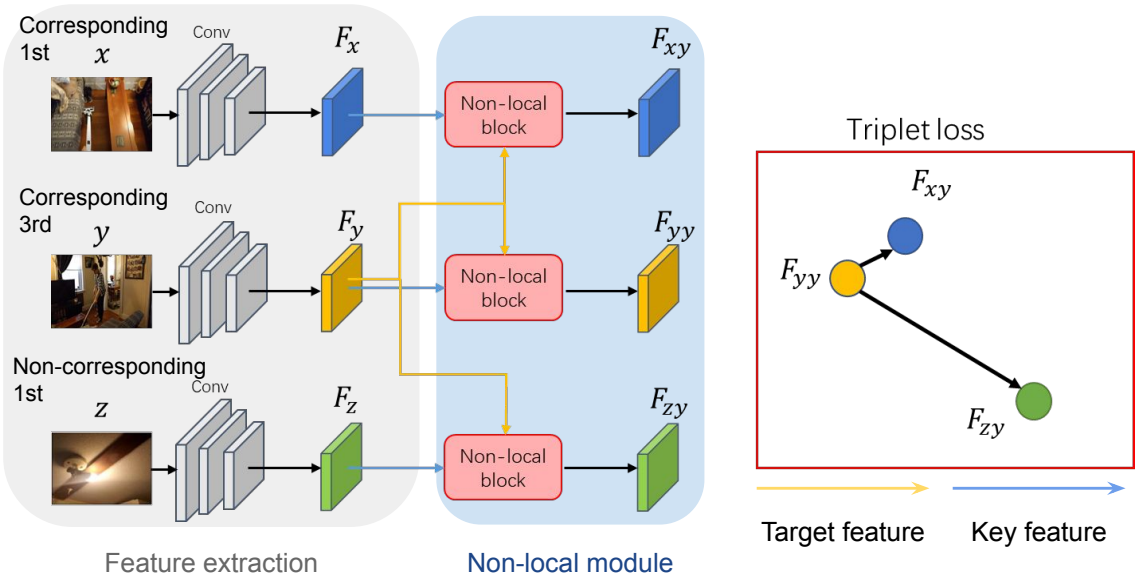

Figure 3.1: Overview of our proposed model. The network takes a triplet input (x, y, z). The deep feature maps $(\mathbf{F}_x, \mathbf{F}_y, \mathbf{F}_z)$ are extracted by three independent branches of convolutional networks. The three feature maps work as key features and are fed to non-local blocks as input, used for building up correlation matrix with third-person feature matrix. With the assumption that there are higher similarities between corresponding pairs, the correlation matrix built by $(\mathbf{F}_x, \mathbf{F}_y)$ and $\mathbf{F}_y$ itself should also have higher similarities than that between $(\mathbf{F}_z, \mathbf{F}_y)$ and $\mathbf{F}_y$ itself.

non-local blocks can also be considered as two branches: target feature input $\mathbf{x}$ and key feature input $\mathbf{x}$. The function of non-local block can be described as: the output feature $\mathbf{y}$ is estimated as the weighted mean of all key feature $\mathbf{x}$, while the weight is computed by the similarity between all key features and each specific target feature.

Through the above understanding, we discover that the typical non-local block actually works as a spatial autocorrelation estimator. Therefore, if we change the target feature and key feature to two different inputs from different images, we may get the cross correlation estimator, which we call cross-view non-local module. The formula of cross-view non-local module is defined as Cross-NL:

$$y(i) = \left[\text{Cross-NL}(\mathbf{x^t}, \mathbf{x^k})\right](i) = \frac{1}{Z(\mathbf{x^t})} \sum_{\forall \mathbf{j}: \mathbf{x_j^k} \in \mathbf{x^k}} f(\mathbf{x_i^t}, \mathbf{x_j^k}) g(\mathbf{x_j^k}), \tag{3.2}$$

where $\mathbf{x^t}, \mathbf{x^k}$ are target input and key input separately.

After extracting deep feature maps $(\mathbf{F}_x, \mathbf{F}_y, \mathbf{F}_z)$ from feature extraction module, we feed the maps into three cross-view non-local blocks. The three blocks take $\mathbf{F}_y$ as target input in common and take $(\mathbf{F}_x, \mathbf{F}_y, \mathbf{F}_z)$ as key input separately. The formulations are presented as:

$$\begin{aligned}
\mathbf{F}_{xy} &= \text{Cross-NL}(\mathbf{F}_y, \mathbf{F}_x) \\
\mathbf{F}_{yy} &= \text{Cross-NL}(\mathbf{F}_y, \mathbf{F}_y) \\
\mathbf{F}_{zy} &= \text{Cross-NL}(\mathbf{F}_y, \mathbf{F}_z)
\end{aligned}$$

The function of three blocks is to estimate the third person feature map by the key features from feature maps extracted by three inputs. Naturally, the estimation by third person feature map $\mathbf{F}_y$ itself (i.e. $\mathbf{F}_{yy}$) should be most successful. Afterward, since we have an assumption that there are some common regions shared by the corresponding image pairs, there should be much more key features in $\mathbf{F}_x$ similar to $\mathbf{F}_y$ than those in $\mathbf{F}_z$. Therefore, based on our assumption, the similarity between $\mathbf{F}_{yy}$ and $\mathbf{F}_{xy}$ should apparently higher than that between $\mathbf{F}_{yy}$ and $\mathbf{F}_{zy}$. The loss function is designed based on this synthesis.

3.2 Unbiased Hierarchical Non-local Module

By analyzing the typical non-local block, we find out there are drawbacks that limit the performance of non-local operations. In following subsections, we will propose our unbiased hierarchical non-local module. We mainly make two modifications (Figure 3.2 (a) and (b)) which improve the performance of the non-local block from different aspects.

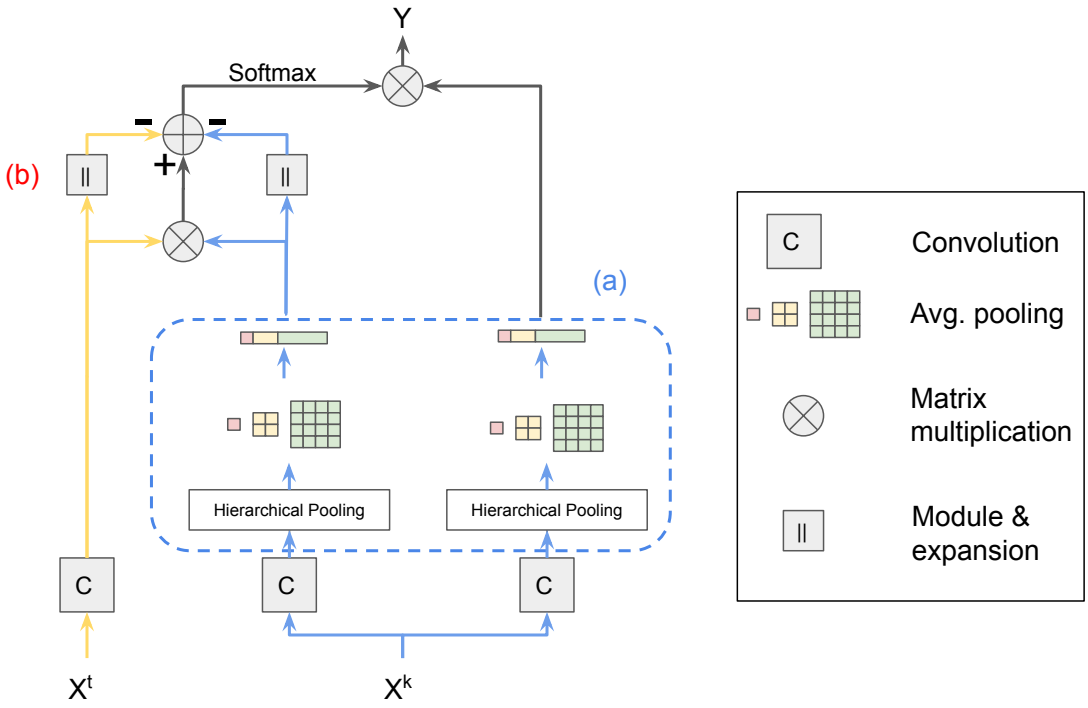

Figure 3.2: Two modifications compared to typical non-local block. (a) Hierarchical Pooling. (b) Zero-centered Correlation Matrix.

3.2.1 Hierarchical Pooling

Motivation. By inspecting the correlation computation process in non-local blocks, one may clear find that even if the computation of correlation matrix does treat all global dependencies equally, the perception field of each feature, no matter key feature or target feature, is still fixed. This means we may only compute the similarity of features which represent the same size of regions in input images. It is still a challenge for block to deal with the different sizes of common objects appearing in first and third-person view.

Therefore, our proposed solution is to apply average pooling operation with multiple window sizes, practically $1 \times 1, 2 \times 2, 4 \times 4, 8 \times 8$. With larger window size, the output vectors of pooling may perceive a larger region. After pooling operation, all the output over the pooling layers are reshaped linearly and concatenated together.

3.2.2 Zero-centered Correlation Matrix

Due to the convenience of implementation, the most generic correlation function applied in typical non-local blocks is the **simplified Gaussian kernel function**:

$$f(\mathbf{x_i}, \mathbf{x_j}) = e^{\mathbf{x_i}^T \mathbf{x_j}}, \tag{3.3}$$

However, if we concern the original similarity function applied in non-local Means, the function is presented as standard Gaussian kernel function:

$$\begin{aligned} w(i,j) &= e^{-\frac{\|v(N_i)-v(N_j)\|^2}{h^2}} \\ &= e^{-\frac{\|v(N_i)\|^2+\|v(N_i)\|^2-2v(N_i)^T v(N_j)}{h^2}} \\ &= e^{\frac{2v(N_i)^T v(N_j)}{h^2}-\frac{\|v(N_i)\|^2+\|v(N_j)\|^2}{h^2}} \end{aligned}$$

We may find that the standard Gaussian kernel function is zero-centered with the module of two vectors are subtracted, which is not implemented in typical non-local block. The bias of input features may lead to inaccurate estimation of the similarity between two features. In order to eliminate this bias, given target feature matrix $X^t \in \mathbb{R}^{N^t \times C}$ and key feature matrix $X^k \in \mathbb{R}^{N^k \times C}$, where N^t, N^k, C means the number of target feature vectors, target feature vectors and vector dimensions, we derive the function to zero-centered correlation matrix as follows:

$$\begin{aligned} \mathbb{F}[\mathbf{X^t}, \mathbf{X^k}] &= \exp\left(\begin{vmatrix} \cdots & \cdots \\ -\frac{\|\mathbf{x_i^t}-\mathbf{x_j^k}\|_2}{2} & \\ \cdots & \cdots \end{vmatrix}_{N^t \times N^k}\right) \\ &= \exp\left(\begin{vmatrix} \cdots & \cdots \\ \mathbf{x_i^t}^T\mathbf{x_j^k} - \frac{\|\mathbf{x_i^t}\|_2+\|\mathbf{x_j^k}\|_2}{2} & \\ \cdots & \cdots \end{vmatrix}_{N^t \times N^k}\right) \\ &= \exp\left(\mathbf{X^t}^T\mathbf{X^k} - \frac{1}{2}\left(\begin{vmatrix} \cdots \\ \|\mathbf{x_i^t}\|_2 \\ \cdots \end{vmatrix} \cdot \mathbf{1}^T + \mathbf{1} \cdot \begin{vmatrix} \cdots & \|\mathbf{x_j^k}\|_2 & \cdots \end{vmatrix}\right)\right) \end{aligned} \tag{3.4}$$

The flow chart is shown as Figure 3.2(b).

3.3 Loss Function

Here we describe the loss function used to train our proposed network. The loss function is a triplet loss used in siamese structure to learn shared representation from corresponding pairs.

The triplet loss is denoted as $L_{TL}(x, y, z)$, which enforces similarity between corresponding feature representations $\mathbf{F}_{xy}$ and $\mathbf{F}_{yy}$, and penalizes similarity between non-corresponding feature representations $\mathbf{F}_{zy}$ and $\mathbf{F}_{yy}$. The cost function is formulated as [HHA16]:

$$L_{TL}(x, y, z) = -\log\left(\frac{e^{-\|\mathbf{F}_{xy}-\mathbf{F}_{yy}\|_2}}{e^{-\|\mathbf{F}_{xy}-\mathbf{F}_{yy}\|_2} + e^{-\|\mathbf{F}_{zy}-\mathbf{F}_{yy}\|_2}}\right) \tag{3.5}$$

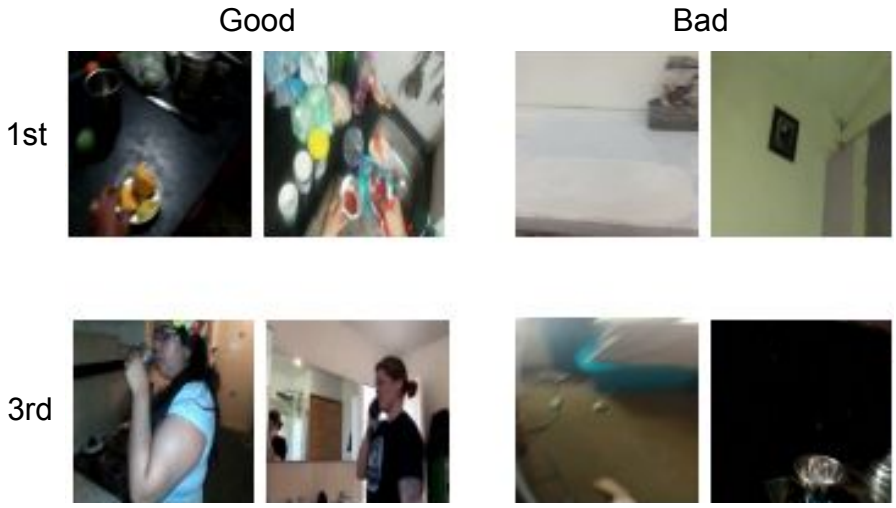

Figure 3.3: Examples for Good and Bad Frames. Left two columns are examples for informative frames. Right two columns are examples for meaningless frames

3.3.1 Frame Screening

In practice, we find that there is a huge gap in the effective information contained between different frames in each video (seen in Figure 3.3. Some frames contain more objects or segments with clear boundaries, or contain rich character actions, which are conducive to feature presentation. On the contrary, there are also some frames that have a lot of blur due to fast motion, or capture the plain background, from which it is difficult to extract meaningful features. Therefore, we need to filter the data according to the quality of the sample frames to minimize the impact of low-quality frames on model training.

Here we use the same method as [SGS+18a], which introduce a selector w to give each set of samples a weight $w(x, y, z)$. This weight will make the model pay more attention to informative sample triplets. The final overall loss function comes to:

$$L(x, y, z) = L_{TL}(x, y, z) \cdot w(x, y, z) \tag{3.6}$$

3.4 Implementation Details

Our framework is implemented by using PyTorch [HHA16], and input frames are cropped into 224×224. We apply a ResNet-152 architecture [HZRS16] as the base CNN model, which is pretrained on ImageNet dataset [SVW+16][1]. The first two

[1]The pretrained model can be downloaded from: https://download.pytorch.org/models/resnet152-b121ed2d.pth

convolutional layers of ResNet-152 are used to extract feature maps. The channels of feature maps and the dimension of channel attention vectors are both 256. The spatial size of feature maps as well as that of attention maps is 56×56. The input features of non-local blocks are based on the conv2 features of ResNet-152. SGD[Bot12] is used to train the whole model, with the learning rate of 1e-4, exponent learning rate decay of 0.95, and batch size of 16. The model is trained on single V100 machine for 50 epochs.

4 Experiments

In this chapter we evaluate our proposed model on public first and third person video dataset. We design two basic experiments, in which we compare and analyze the experimental results with state-of-the-arts both quantitatively and qualitatively. By visualizing the correlation matrix calculated by non-local modules, we analyze the actual performance of non-local operations and how they seek global dependencies. In addition, we also designed other additional experiments to further illustrate the applications of our model.

4.1 Experimental Settings

4.1.1 Dataset

We use the Charades-Ego dataset [SGS+18b]. This dataset contains 3930 sets of synchronized first and third-person videos, recording a large amount of human behavior information. However, there are some invalid data pairs in this dataset, such as videos pairs that are both third-person views (See in Figure 4.1). By combining the settings in [YCLL19] and our own manual screening, 313 pairs of invalid videos were finally eliminated from the dataset.

4.1.2 Tasks and Evaluation

The first basic experiment is to discriminate the corresponding frame. The experimental design and model structure are basically the same, and each data sample comes from the same pair of videos. First, select two frames at the same time point from the first-person and third-person videos to form a positive sample, and then randomly select one frame as a negative sample from the first-person video. In order to prevent the positive sample and the negative sample from getting closer and increasing the difficulty of the experiment, we artificially set the time difference between the positive and negative samples to be more than 5s (120 frames). The result is evaluated by the discrimination accuracy (in %).

The second experiment is to quantitatively evaluate the matching time. We send all frames in the first person and a certain frame in the third person into the model, calculate their feature distances and sort them. The frame with the smallest feature

Figure 4.1: Invalid data pairs (first row) and valid data pairs (second row).

distance is regarded as the prediction result and the time difference is calculated with groundtruth. The result is evaluated by the average time difference (in seconds).

During the experiment, we select AONet [SGS+18a] and Joint Attention Network [YCLL19] as state-of-the-arts. The both models are re-trained with same hyperparameter settings as our proposed method.

Also, considering that we make two modifications on the non-local blocks, We will conduct multiple sets of controlled experiments by controlling the variables to verify the effects of each of them on the performance of the model when they are introduced separately and the two are introduced together: (a) pure non-local blocks (Abbreviated as NL); (b) non-local blocks with hierarchical pooling (NL+Pool); (c) non-local block with zero-centered correlation matrix (NL+Center); (d) non-local blocks, with both hierarchical pooling and zero-centered correlation matrix (NL+Both).

4.2 Quantitative Analysis

The quantitative results of different models in the cross view discrimination task and best matching time task are shown in the table 4.1. It can be seen that the distribution accuracy and the time error show a relatively obvious negative correlation. This is reasonable because the process of image changing frame by frame over time is smooth and continuous, and the result proves that the model's perception of fea-

Method	Discrimination accuracy (%)	Average time error (s)
AONet[SGS+18a]	50.9	7.0
Attention[YCLL19]	86.2	5.0
NL	85.7	5.9
NL+Pool	88.1	4.6
NL+Center	87.8	5.7
NL+Both	**89.0**	**4.5**

Table 4.1: Quantitative Comparisons on Discrimination Task and Time Error Task.

ture changes is also smooth. It can be seen that AONet [SGS+18a] is significantly behind AttentionNet [YCLL19] and our model in both tasks. AONet [SGS+18a] simply extracts the features from images and calculates the distances, and does not adopt a mechanism to evaluate the importance of features, which means the model is easy to be disturbed by redundant information. Therefore, it cannot obtain a good performance on joint feature presentation.

Our pure non-local model performs similarly to the AttentionNet [YCLL19]. However, after introducing the two improvements of hierarchical pooling and zero-centered matrix, the discrimination accuracy rate has been significantly improved (2.4% and 2.1%), which proves that the two improvements bring improvement to model from different perspectives. Compared with the AttentionNet [YCLL19], the overall models has achieved a larger accuracy rate improvement (2.8%), which proves that the joint presentation learnt based on the assumption of global similarity of common regions is more suitable for cross-perspective discrimination tasks than that based on attention and saliency detection.

4.3 Qualitative Analysis

4.3.1 Visualization of Feature Activation

We first conduct quantitative analysis by visualizing the visual features extracted by CNN from the image. The visualized result is shown in the figure 4.2. The visualization method is to average the output of the Resnet152 conv2 layer along channel axis and then apply upsampling to the feature maps upon the size of the original image. Since non-local blocks calculates the similarity between features and backpropagates, those vectors with more similar features will have greater weight in backpropagation, thereby activating the convolution kernel related to the generation of these features in the CNN. In the visualization, the activation value corresponding to this part of the feature will be higher.

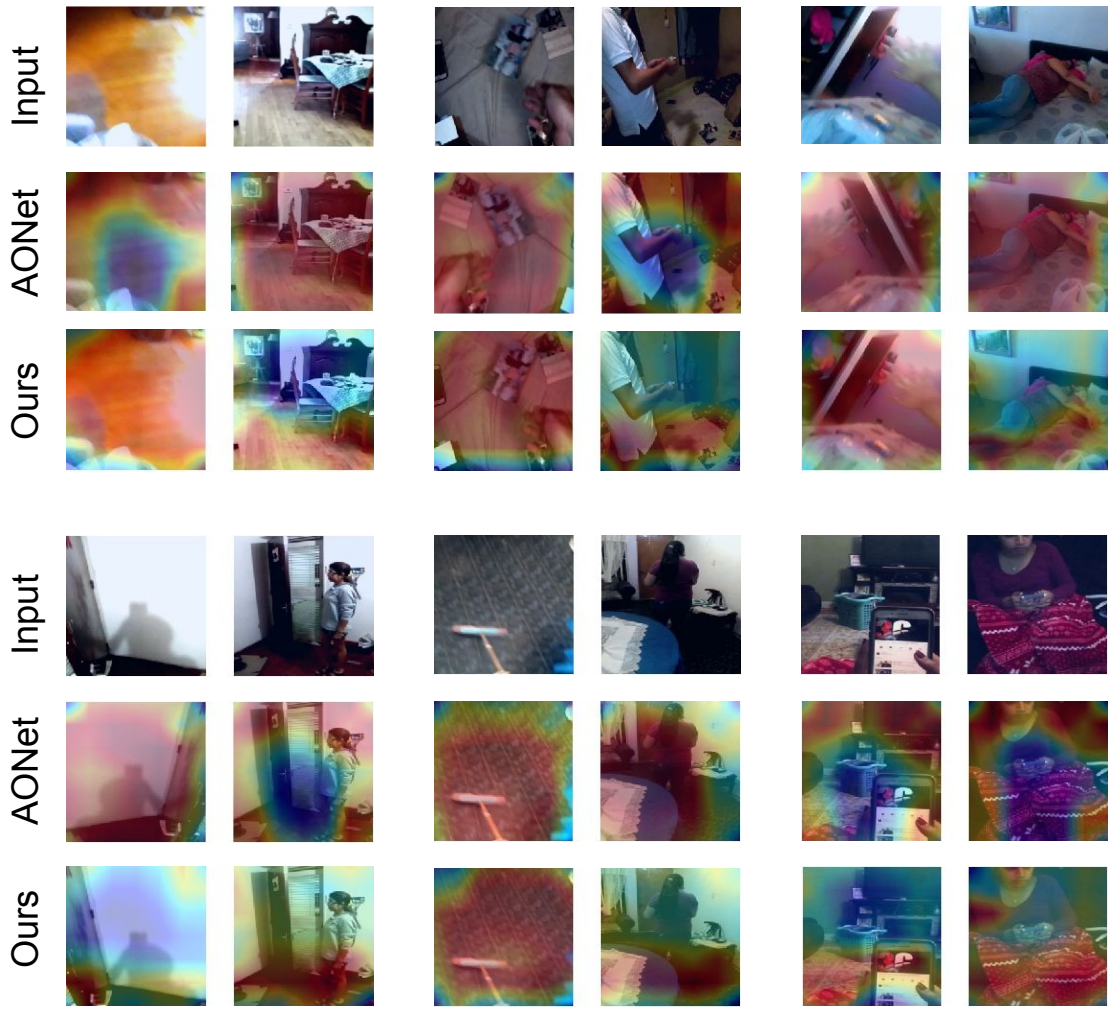

Figure 4.2: Visualization of feature activation. the activation values are visualized with heatmaps on input images. The colour ranges from red to blue referring to high to low activation.

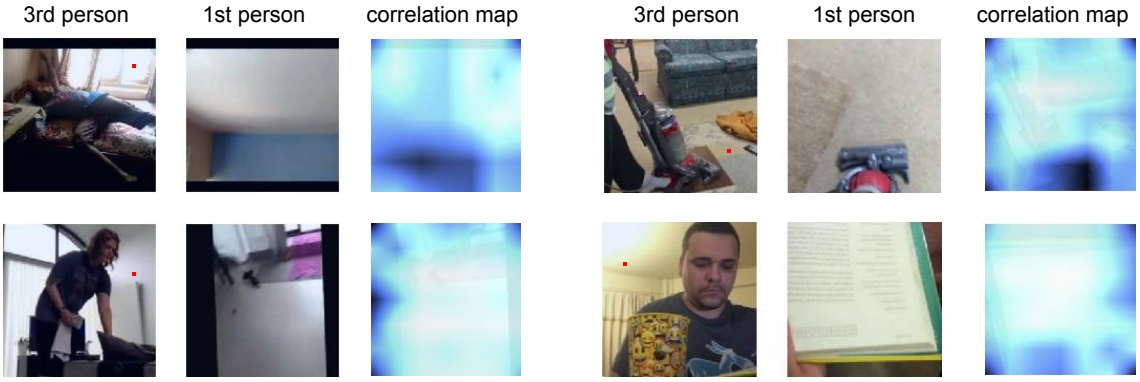

Figure 4.3: Visualization of correlation maps. The target feature pixel is marked as red point in third person view input. The correlation maps are shown as grayscale images, ranging from black to white representing low to high correlation.

By comparing the heat maps of activation features, we can find that compared to AONet [SGS+18a], the activated features in our model are more concentrated. Moreover, in the corresponding first-person and third-person images, the main activated features of the two are generally distributed in similar visual features, such as color, texture, etc., and these features are often the common area between the two perspective images. For example, in the example above in the first column, we can clearly see that the red area in the heat map, which are the main activated features from CNN, is concentrated on the floor in both images, and the floor is the common area of both images. In the example above in the third column, the activated features are concentrated on the pattern of the bed sheet. These examples fully prove that our model will pay attention to areas with similar visual features crossing the views.

4.3.2 Visualization of the Correlation Maps

Next, we visualize the internal layers of non-local network to get a more intuitive experience of how non-local network works. The core structure used in non-local network to represent the dependency between features is the correlation matrix. However, the correlation matrix expresses the correlation pixel-pixel in the feature map, thus the size of its horizontal and vertical coordinates is the total number of feature vectors. That is to say, if the feature matrix is converted into the size of the input feature map, it is actually a four-dimensional tensor, which is very inconvenient for visualization. But we can target a certain feature in the third-person feature map, and specifically visualize and target the first-person correlation map with this feature. This visualized structure is shown in the Figure 4.3.

As we analyzed above, the correlation map for the target pixel does reflect the similarity with the target pixel. It can be seen that areas with similar colors to the

target point have higher correlation. This nature does help the model identify the common area. But when we look at the example in the lower right, the target point is near the light, and the color is close to white. The corresponding first-person picture is a book, and the subject is also in white. Although the two are completely different objects, the feature vector corresponding to the book in the correlation map still has a high correlation. This shows that in this example, non local network does not explore the deep visual features very well, but is disturbed by the shallow color information. We will discuss this issue further in Chapter 5.

4.4 Experiments on Downstream Tasks

Through the above experiment, we use the non-local based model to obtain a joint presentation of the first and third person images. This joint presentation proved its effectiveness in the discrimination experiment. However, by the analysis of some specific samples, we also find its limitations. In this section, we hope to further test the model through some other simple experiments to evaluate the model's performance of extracting visual features and transferring knowledge crossing the perspectives.

4.4.1 Video Action Recognition

In this part, we show that joint representation learnt from first and third-person videos could be exploited to apply action recognition. CharadesEgo dataset provides action annotations for the behavior of actors in the video, which makes it possible for us to conduct the experiment of action recognition. Traditional video action recognition work [SZ14, CZ17, FPZ16] is mainly for a single video, and we want to see whether the feature extraction network learned through the cross-view model could help improve the performance of action recognition. We pretrain the model in the discrimination task, and then connect the CNN branch of third-person view to a fully connected layer for action classification. AONet also has motion recognition experiments, so we chose AONet as the baseline. The results are shown in the table.

The results show that our model performs better than AONet.

	AONet	Ours
Accuracy	23.1	25.8

Table 4.2: The results of action recognition task. The accuracy is measured by mAP (%). Higher is better.

	AttentionNet	Ours
Pridiction error	0.22	0.30

Table 4.3: The results of gaze prediction task. The accuracy is measured by L_2 distance. Lower is better.

4.4.2 Gaze Prediction

Gaze prediction is also a typical task to evaluate the feature extraction of a model. Firstly, the actors' head coordinates are estimated based on the existed method [CSWS17]. Secondly, we compute the mean of values in correlation map of each target feature. Finally the coordinate of the target feature in third person feature map which has the largest mean is considered as prediction. In this experiment we take AttentionNet as baseline.

We do the experiment on gaze prediction dataset GazeFollow[RKVT15]. The prediction accuracy is measured by average Euclidean distance (We assume each image is of size 1×1) between predictions and annotations. The result is shown in table:

Our result doesn't perform as good as AttentionNet. We observe that the center of common area and human attention area are not the same in most conditions. The possible reason should be our model focus on common region instead of saliency and attention, while the latter factors mainly affects the gaze prediction.

5 Discussion

5.1 The Dependencies that Non-local Modules Concern about

In the experimental part of the above chapter, we qualitatively and quantitatively proved the feasibility of applying the non-local network to the common feature extraction problem of cross-view images through multiple experiments. By visualizing the internal matrix of non-local network, we do observe that non-local network shows higher sensitivity to similar colors, shapes and textures in the two viewing angles. But are these all the feature dependencies that non-local network focuses on? In this chapter, I would like to further discuss the specific characteristics of global dependencies that non-local network is concerned about.

Since in non-local network, the dependence between features is achieved by calculating the distance between two feature vectors, the features that non-local network tends to pay attention to must also have a relationship with the attributes of the feature itself. Several works committed to the interpretability of deep learning models [EBCV09, MV15, SCD+17] have pointed out that as the number of model layers deepens, the features extracted by the model gradually shift from low level knowledge such as color, edge, and texture to high-level knowledge such as overall shape and even semantics, interaction, and logic.

In this paper, the vision features utilized in non-local blocks are primarily low-level knowledge. That's the reason why two totally distinct objects are decided as high dependencies only if they share similar color. Nonetheless, When Wang et al.[WGGH18] applied non-local blocks to the video action recognition task, the result showed that high-level dependencies such as the relationship between a football and people's body when one bounds the ball can also be detected (See in Figure 5.1), which means it is also possible for us to build up joint representation between first and third-person view videos based on more high-level features. In this scenario, high-level features could refer to spatial relative positions between objects, human-object interaction, and so on.

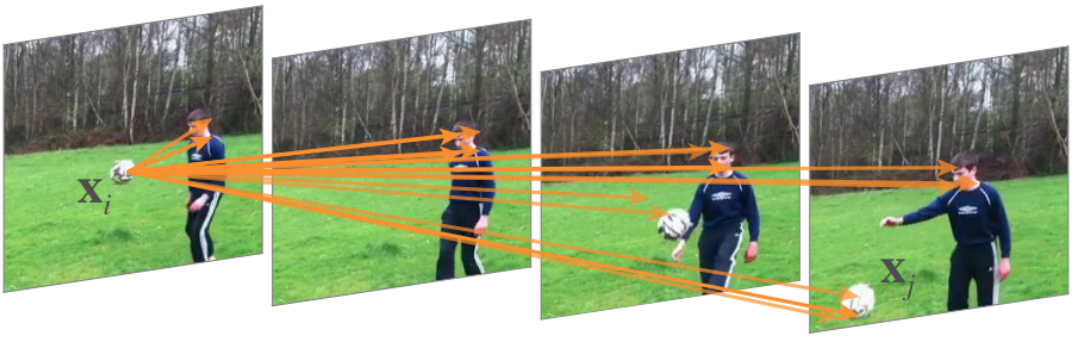

Figure 5.1: High-level dependencies can be detected by non-local networks even if there is no visual similarity.

5.2 Limitations of the Proposed Method

Even if the proposed method has proved is effectiveness on establishing joint representation that is capable of transfer knowledge crossing the first and third-person view videos, there are limitations of our model. The limitations can be summarized as our ignorance of the actors. Among all corresponding first and third-person videos, actors, i.e. camera wearers, are always crutial entities worth to be studied. The actions and interactions of actors take an irreplaceable status in analysis. However, in our paper, the actors are not paid enough attention to. They only act as an ordinary part of environment and participate in calculation of feature dependencies. This means that the joint representation learnt in this paper is not able to represent human's tendency when needed. This is part of reason why the model doesn't perform well on gaze prediction task.

6 Conclusion and Future Work

In this thesis, we propose a novel method of joint representation learning between corresponding first and third-person view videos. The research on cross-view joint representation learning should contribute to further study on co-analysis between multiple view videos, such as co-segmentation. This requires a more precise model that may build up a pixel to pixel correspondence between first and third-person views. The main challenges to be gotten over is how to understand the connections between different perspectives in same scene from a higher level, just as human beings do.

The clue idea of our approach presented in this thesis is to introduce the non-local neural network in order to learn global dependencies from multiple viewpoints. Since the first and third person videos are captured synchronously in single locale, there should be some common regions shared by both viewpoints. These universal regions could take rolls like bridges between two videos. Since non-local networks capture global dependencies directly by computing interactions between any two positions of inputs regardless of their positional distance, they are selected to deal with the correspondence establishment issue. We adopt a self-supervised structure which takes triplets of images as input, with both positive and negative samples. Deep features are extracted separately and global dependencies is built up by non-local modules. Since there are more in common between the corresponding pairs, more stable dependencies should exists and the features of them should be more similar. A triplet loss is applied to force that. The performance of model is evaluated on open dataset. Fundamental experiments include frame discrimination tasks and matching time error tasks. The experimental results demonstrate that the model manages to build up clear joint representation between both views. Two additional experiments including action recognition and gaze estimation are conducted so as to further check out the effectiveness of model. While in action recognition task, the model performs better than the baseline, the model seems not to be competent in gaze prediction task.

The failure in gaze prediction task and some analysis expose the limitations of our model. The limitations are analyzed and will be tackled in the future. Some limitations are caused by the inner drawbacks of the non-local operations. The performance of non-local operations are based on the quality of features fed into the block. Some other limitations are caused by the overall structure which ignores the significance of human actors in the videos. In the future, human motion and attention should be specifically considered as a influential factor of joint representation

learning.

Based on the analysis of limitations and insights we obtained from the results, The following directions are considered as future directions of improving the current model.

Taking hierarchical features as non-local input As already mentioned in Section 5.1, the current non-local model mainly take low-level features as input. In the future, we consider to introduce multiple levels of features which are activated in different layers of CNNs to the non-local modules. Not only are similar color, appearance and texture considered as highly dependent, but also those features that exist logical connections.

Human pose and gaze estimation The gaze and pose information (i.e. what he/she is looking at and what he/she is doing) of human actors indicate actors' insight tendencies of interaction with environment. We need to introduce a effective human analysis model and take the relationship with actor into consideration when computing the dependencies.

Applying graph models Another available solution to summarizing high-level dependences is based on graph model. Graph model turns convolutional features into some abstract node features and the analyze them by graph algorithms. Since graph models are good at abstract logical correspondence, we may first extract multiple independent semantic entities from original input and the seeking for relevance between entities by graph models.

Introducing temporal motion information Current works on joint representation learning between first and third person view videos are almost taking single frame and input. However, there are even more correspondence information existing in the dynamic change of views. For instance, if the model takes a series of frames as input, it may perceive a wider range of view for each perspective. Besides, the future action may also predicted. Overall, those dynamic information could contribute to joint representation learning.

3D reconstruction This may provide an alternative train of thought. Since our goal is to learn a joint representation that may describe the both views uniformly, the most integrated and robust presentation should of course be the precise 3d model of scene. However, current works [IKH+11, MLD+06, GZS11] aimed to 3D reconstruction most need camera calibration or depth sampling. This should be very challenging.

Further applications Also, besides considering how to improve presentation model, study on some further applications based on the joint representation should also be interesting. For example, **video co-segmentation**. If a precise pixel-wise joint representation is available, segments from each single view could build up stable semantic connections. In conclusions, there are a lot of interesting topic related to the co-analysis of multiple view videos waiting for people to discover.

Acknowledgments

First of all, I want to express my sincere gratitude to my advisor Prof. Yoichi Sato for the continuous support of my master's study and research, for his patience, motivation, enthusiasm, and immense knowledge. Even if I faced great difficulties on my master research, he still encouraged me and helped me change to more proper topic. His guidance helped me in all the time of research and writing of this thesis. I really appreciate his suggestions.

I would also like to thank Prof. Sugano for his professional advice on my research. As my second supervisor, he gave me many precious advice on how to make innovations and details of model implementation.

I thank all the members in Sato & Sugano Lab, for the regular scholar discussion and sharing. Those discussions gave me a lot of inspirations. Furthermore, I thank my labmates for the fun and meaningful life we share in last two years.

I thank the Chinese Students and Scholars Association of UTokyo for their enthusiastic help for my life as a foreign student in Tokyo. Becoming a member of it was one of most important decisions I made. The activities we held together made my off-campus life colorful.

Last but not the least, I would like to thank my family: my parents Yang Wang and Wantong Zhu. They are always the most stable supporters to me and give me great spiritual encouragement.

References

[AB16] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *Proc. of European Conf. Computer Vision*, pages 253–268. Springer, 2016.

[BCM11] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-local means denoising. *Image Processing On Line*, 1:208–212, 2011.

[Bot12] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[CSWS17] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[CZ17] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[EBCV09] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[FLX+17] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 5125–5133, 2017.

[FPZ16] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. of Int. Conf. Neural Information Processing Systems*, pages 2672–2680, 2014.

[GZS11] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *Proc. of IEEE Conf. Intelligent Vehicles Symposium*, pages 963–968, 2011.

[HHA16] Elad Hoffer, Itay Hubara, and Nir Ailon. Deep unsupervised learning through spatial contrasting. *arXiv preprint arXiv:1610.00243*, 2016.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[IKH+11] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proc. of the 24th Annual ACM Symposium on User Interface Software and Technology*, pages 559–568, 2011.

[JG15] Dinesh Jayaraman and Kristen Grauman. Learning image representations tied to ego-motion. In *Proc. of IEEE Int. Conf. Computer Vision*, pages 1413–1421, 2015.

[LGG12] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 1346–1353. IEEE, 2012.

[Low99] David G Lowe. Object recognition from local scale-invariant features. In *Proc. of IEEE Int. Conf. Computer Vision*, volume 2, pages 1150–1157. Ieee, 1999.

[LYR15] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 287–295, 2015.

[MFZ+20] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 5690–5699, 2020.

[MLD+06] Etienne Mouragnon, Maxime Lhuillier, Michel Dhome, Fabien Dekeyser, and Patrick Sayd. Real time localization and 3d reconstruction. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, volume 1, pages 363–370, 2006.

[MV15] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 5188–5196, 2015.

[PR12] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 2847–2854. IEEE, 2012.

[RB18] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 3501–3510, 2018.

[RB19] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding*, 187:102788, 2019.

[RK17] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proc. of IEEE Int. Conf. Computer Vision*, pages 3696–3705, 2017.

[RKVT15] Adria Recasens*, Aditya Khosla*, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Proc. of Int. Conf. Neural Information Processing Systems*, 2015. * indicates equal contribution.

[RM13] Michael S Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 2730–2737, 2013.

[SCD+17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. of IEEE Int. Conf. Computer Vision*, pages 618–626, 2017.

[SGS+18a] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 7396–7404, 2018.

[SGS+18b] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

[SVW+16] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proc. of European Conf. Computer Vision*, pages 510–526, 2016.

[SZ14] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Proc. of the 27th Int. Conf. Neural Information Processing Systems*, 2014.

[WGGH18] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[WPLK18] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proc. of European Conf. Computer Vision*, pages 3–19, 2018.

[YCLL19] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *Proc. of the 27th ACM International Conference on Multimedia*, pages 1358–1366, 2019.

[YKS15] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Ego-surfing first-person videos. In *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, pages 5445–5454, 2015.

[YYC+20] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *Proc. of European Conf. Computer Vision*, pages 191–207, 2020.

[ZXB+19] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proc. of IEEE Int. Conf. Computer Vision*, pages 593–602, 2019.